



National Income Dynamics Study Wave 2 User Manual

Edited by Michael Brown, Reza C. Daniels, Louise De Villiers, Murray
Leibbrandt and Ingrid Woolard

Contents

List of Contributors	iv
Foreword.....	v
1. Using This Manual.....	1
1.1 What All Users Have to Know	1
2. The NIDS Data	2
2.1 Citation of NIDS data and documentation.....	2
2.2 Process to download the data	2
2.3 Data Formats.....	3
2.4 Data Structure.....	3
2.5 File Structure.....	4
2.6 Identifiers	6
2.7 Merging Datasets Within & Between Waves.....	6
Merging Within Wave 1 or Within Wave 2	7
Merging Between Waves	7
2.8 Variable naming convention	7
2.9 Non-Response Codes	9
2.10 Anonymisation	10
2.11 Secure data	10
2.12 Program Library	10
3. Data Collection.....	11
3.1 Overview of CAPI Cycle	12
3.2 Overview of the Tracking Process.....	13
3.3 Listing and Panel Data.....	15
3.4 Contacting Respondents	15
3.5 Data Quality Issues and Data Collection	16
Unit Non-response	16
Item Non-Response.....	17
Data Consistency.....	17
The Mechanics of Data Quality Checks.....	18
3.6 Fieldwork schedule	20
Pre-test.....	20
Main data collection	20

Phase 2	20
Questionnaire Differences between W2 Phase 1 & W2 Phase2	21
3.7 Response Rates & Attrition	21
4. Derived Variables	26
4.1 Best Variables.....	26
4.2 Geography.....	26
4.3 Occupation	27
4.4 Industry	27
4.5 Employment Status.....	27
4.6 Income	28
Bracket Responses	30
Item Non-Response and Imputation.....	30
Income From Subsistence Agriculture	31
Bonus Payments.....	31
4.7 Expenditure.....	32
Imputations.....	32
4.8 Wealth.....	33
Wealth in the NIDS Wave 2 Household and Adult Questionnaires	34
Imputation	35
Aggregating Household Net Worth and Including One-Shot Measures Where Appropriate	37
4.9 Anthropometric Z-Scores.....	37
4.10 Weights	38
Updated Wave 1 Weights	38
Panel Weights	39
Cross-Sectional Weights for Wave 2.....	40
5. Further Resources	42
5.1 Program Library	42
Data manipulation	42
Derived variables	43
6. References	45

List of Contributors

This document was created by the NIDS team. For the correct citation method, see section 2.1 of this document. Authors in alphabetical order include:

- Cally Ardington
- Timothy Brophy
- Michael Brown
- Michelle Chinhema
- Reza C. Daniels
- Louise De Villiers
- Arden Finn
- Murray Leibbrandt
- Sibongile Musundwa
- Martin Wittenberg
- Ingrid Woolard

Foreword

In any society, data from socio-economic panel surveys are essential in order to describe and explain unfolding social dynamics and patterns of social mobility. Worldwide, it is now widely recognised that it is exactly this research agenda that is required to understand who is getting ahead in society, who is falling behind, whether younger generations are better-off than their parents, who is trapped in poverty, who is escaping poverty and what are the policies and social forces driving all of these changes? These are key questions for any country with a policy agenda that includes poverty and inequality alleviation and inclusive economic growth and it is no surprise that almost all developed countries and an increasing number of developing countries have embarked on such panel studies.

The post-apartheid project in South Africa is all about social transformation and the above questions are at the heart of our policy agenda. It is for this reason that in 2005 the Presidency decided that South Africa needed its own national panel study to provide the information base to benchmark our progress and assist in assessing the effectiveness of policies to promote positive social mobility. In line with this decision, in 2006 a service provider was selected to undertake the Wave 1 of South Africa's National Income Dynamics Study (NIDS) and, in 2008, this base Wave of NIDS was implemented across the length and breadth of South Africa. Between 2006 and the fieldwork in 2008, the infrastructure to design such a national survey was set up and a set of questionnaires were designed to give effect to the dimensions of well-being that were to be tracked over time in the survey. At the broadest level, these dimensions were:

- Wealth creation in terms of income and expenditure dynamics and asset endowments;
- Demographic dynamics as these relate to household composition and migration;
- Social heritage, including education and employment dynamics, the impact of life events (including positive and negative shocks), social capital and intergenerational developments; and
- Access to cash transfers and social services

The Wave 1 questionnaires contain the detailed information that was gathered on our national sample at the first visit. This information provided the basis from which to track their progress. In 2010/2011 Wave 2 of NIDS re-interviewed these sampled people, gathering information on developments in their lives since they were interviewed first in 2008. As such, the comparison of Wave 1 and Wave 2 information provides a detailed picture of how South Africans have fared over two years of very difficult socio-economic circumstances.

It has been SALDRU's privilege to undertake the first two waves of NIDS on behalf of the Presidency. It is important for the Presidency that the data contained in NIDS becomes available for policy analysis as soon as possible after it has been collected. In addition, from the outset the Presidency committed itself to speedy public release of the NIDS data and to an ambitious Program to encourage analysis of the data by the broader research community. To this end, the Wave 1 data were released in early July 2009 and Wave 2 data in early 2012. NIDS is a powerful research resource for the nation and is seen as a partnership between the government and the research community directed at better understanding of our contemporary social dynamics and at better policy making.

Hopefully, you are reading this User Guide as you are about to embark on the analysis of NIDS data for the first time or to begin an analysis of Wave 2 data having used the Wave 1 data already. This guide is a practical document designed to help you to understand two waves of NIDS data and to use them correctly and appropriately.

1. Using This Manual

The NIDS survey is a face-to-face longitudinal survey of individuals living in South Africa as well as their households. This User Manual has been designed to assist users of the data to understand the operation of the survey and the resulting structure of the datasets.

The User Manual is a reference tool for users. As such, it is unlikely that it will be read from cover-to-cover. Rather, the detailed contents page can be used as an index to guide users to appropriate pages for themes of interest. This Manual will be updated with each wave of NIDS, though additional, supplementary reports should also be consulted by users. These are available on the NIDS website: www.nids.uct.ac.za

1.1 What All Users Have to Know

It is recommended that all users familiarise themselves with at least the following sections of this document:

- The structure of the data: see section 2. This entire section should be read, especially subsection 2.7 on merging datasets within and between Waves.
- The fieldwork schedule: see section 3.6.
(Note well that the Wave 2 Phase 2 questionnaire is a shortened version of the Wave 2 Phase 1 questionnaire.)
- Non-response and attrition in Wave 2: see section 3.7.
- Updated weights for Wave 1 and Wave 2: see section 4.10.
- Examples of how to correctly merge NIDS data using *Stata*: see section 5.1.
- How to deflate the financial data: see section 5.1

2. The NIDS Data

The National Income Dynamics Study (NIDS) uses a combination of household and individual level questionnaires. The data from the different questionnaires are recorded in separate data files with one row per record (individual or household). The data can be exported into most standard statistical packages. A set of files is released for each Wave, but they can be combined across Waves using the unique identifier for the individual, variable name “*pid*”.

2.1 Citation of NIDS data and documentation

Users wishing to cite the data should use the following reference:

Data Citation:

Southern Africa Labour and Development Research Unit. National Income Dynamics Study 2010-2011, Wave 2 [dataset]. Version 2.2. Cape Town: Southern Africa Labour and Development Research Unit [producer], 2012. Cape Town: DataFirst [distributor], 2012

Readers wishing to cite this document should use the following reference:

Documentation Citation:

Brown, M., Daniels, R.C., De Villiers, L., Leibbrandt, M., & Woolard, I., eds. 2012, “National Income Dynamics Study Wave 2 User Manual”, Cape Town: Southern Africa Labour and Development Research Unit

2.2 Process to download the data

The NIDS data can be downloaded from the DataFirst website:

<http://www.datafirst.uct.ac.za/dataportal/index.php/catalog/central/about>

The steps to follow to gain access to the data are:

Step 1: **Register as a user on the DataFirst website.** Once you have registered on the DataFirst website the registration details can be used to access datasets from the site.

Step 2: **Complete a short online *Application for Access to a Public Use Dataset for the NIDS datasets*.** On the form you will need to provide a short description of your intended use of the data. The information provided here helps us to understand how NIDS data is being used by the research community. The form also asks you to agree to Terms and Conditions related to the use of the NIDS data. You will need to agree to the Terms and Conditions, namely:

- a) The data provided by DataFirst will not be redistributed or sold to other individuals, institutions, or organisations without the written agreement of DataFirst.
- b) The data will be used for statistical and scientific research purposes only. They will be used solely for reporting of aggregated information, and not for investigation of specific individuals or organisations.

- c) No attempt will be made to re-identify respondents, and no use will be made of the identity of any person or establishment discovered inadvertently. Any such discovery would immediately be reported to NIDS at the following address: nids-survey@uct.ac.za
- d) No attempt will be made to produce links among datasets provided by DataFirst, or among data from DataFirst and other datasets that could identify individuals or organisations.
- e) Any books, articles, conference papers, theses, dissertations, reports, or other publications that employ data obtained from DataFirst will cite the source of data in accordance with the Citation Requirement provided with each dataset.
- f) A digital copy of all reports and publications based on the requested data will be sent to DataFirst.
- g) The original collector of the data, DataFirst, and the relevant funding agencies bear no responsibility for use of the data or for interpretations or inferences based upon such uses.

Step 3: **Download the data.** Selected coding and syntax files can also be downloaded at this stage.

2.3 Data Formats

The data are available in the following formats: R, S-Plus, SPSS, Stata. Please contact DataFirst to obtain the data in other formats.

2.4 Data Structure

Every *resident*¹ individual (CSM² or TSM³) is allocated an individual identifier (*pid*). Individual interview records are created for all resident household members. The data file in which the record can be found is dependent on age at interview and type of interview conducted. Deceased CSMs do not have individual interview records as no interview was conducted. A record of all deceased individuals is contained in the “*Link File*”.

Each individual questionnaire maps uniquely to a household questionnaire and household roster file using the household identifier (*w2_hhid*). This is the household in which the person is resident at the time they were interviewed. Individual identifiers on their own merge non-uniquely to the household roster file. This lists all the rosters on which they are considered *household members*⁴. An individual can be a household member of more than one household because of the nature of familial relationships. However, they can only be resident, as defined in NIDS, in one household in each Wave of the survey.

The household roster file for each household includes the details of all household members, even if they are not all resident at that household. Those that are non-resident may be resident in another

¹ Residency: Usually resides at the house for more than four nights a week.

² Continuing Sample Member: All resident members of the original selected Wave 1 households (including children) and any children born to or adopted by female CSMs in subsequent waves

³ Temporary Sample Member: A person who is not a CSM but is co-resident with a CSM at the time of the interview

⁴ Household membership: Defined as spending more than 15 days in the last 12 months at the household and sharing food and resources when staying at that household

household, deceased or living in an institution such as a prison, hospital, university residence or boarding school. The following interview and data rules apply to non-residents:

- If a person left the household more than 12 months ago and subsequently died we record their death and details about the death in their last known household. The deceased person will stay on that household's roster even if they were not strictly speaking a household member at the time of their death. However, no individual questionnaire record exists for them in the data because no individual interview was conducted.
- If a person lived in an institution at the time of interview a proxy questionnaire was completed for them in their last known household although they are not strictly speaking a household member. This is the same methodology as was followed in Wave 1 and allows information to be collected for household members who are *out of scope*⁵.

If a respondent moved outside the borders of South Africa to a private dwelling they are assigned their own household identifier which links to a household questionnaire record in the household roster and individual questionnaire files. Out-of-scope households are identified in the "*Link File*" with the household and individual outcome identifier variables.

If the household refused to participate or there is some other type of non-response (e.g. the household could not be located), the individual questionnaires will still appear in the data files but the outcome will indicate that it was household level non-response. The individual and household outcome variables in the "*Link File*" (see below) identify the outcomes of respondents in both Wave 1 and 2.

2.5 File Structure

The data files that make up the NIDS dataset are as follows:

Link File: One record per individual. It lists the individual identifiers and the household identifier for each Wave in which that person is resident. The link file also has other pertinent information such as if the individual is a CSM or TSM, in which individual questionnaire file their record can be found for that Wave, and the original Wave 1 cluster of the household. Household and individual outcomes are also provided for each Wave. Unique identifier: *pid* (n = 34979).

HHQuestionnaire: One record per household with data from the household questionnaire, excluding the household roster. Unique identifier: *w2_hhid* (n= 9134).

HouseholdRoster: One record per person for every household of which they are a household member. Because one person can be a member of more than one household, duplicate *pid*'s are present in this dataset. Unique identifier for household: *w2_hhid*

⁵ Out of scope: A person residing outside of the sampling frame and who has a zero probability of being interviewed. Examples include people living in institutions (such as hospitals, prisons and boarding schools) and those that moved outside of South Africa.

(n = 9134), non-unique identifier for individual: *pid* (n= 36181). The combination of *w2_hhid* and *pid* is unique per person within each wave. The *difference* between the count of *pid*'s between *HouseholdRoster* and the *Link File* is due to non-resident TSMs, individuals that are members of more than one household.

Adult: One record per entry from the adult⁶ questionnaire. Unique identifier for household: *w2_hhid* (n=8851), unique identifier for individual: *pid* (n=21880); 4224 observations have no data beyond Section A of the questionnaire as these individuals refused to participate in the survey either at a household level or at an individual level or moved outside of South Africa. These records are recorded in the *w2_a_outcome* variable. The one polygamist in the sample appears only once in the adult file. This is the household in which their individual interview was conducted.

Proxy: One record per entry from the proxy⁷ questionnaire. Unique identifier for household: *w2_hhid* (n=898), unique identifier for individual: *pid* (n=1124).

Child: One record per entry from the child questionnaire. Unique identifier for household: *w2_hhid* (n=5037), unique identifier for individual: *pid* (n=11094); 1278 observations have no data beyond Section A as these individuals refused to participate in the survey either at a household level or at an individual level or moved outside of South Africa. The non-response records have a value greater than one in *w2_c_outcome* variable.

Derived variables are variables that were not asked directly of the respondent, but which were calculated or imputed from other information. For example, aggregate income and expenditure variables were constructed. Most of the derived variables are in the individual derived or household derived files. The following derived data files are part of the NIDS Public Release for each Wave:

hhderived: One record per household. Unique identifier for household: *w2_hhid* (n=9134). Geographic information of the current location of households and the weights variables are included in this file.

indderived: One record per resident person. Deceased and non-resident household members are not included in this file. Unique identifier for household: *w2_hhid* (n=9023), unique identifier for individual: *pid* (n=34098).

See section 4 on Derived Variables and section 5.1 Program Library for more information.

⁶ A person is defined as an adult if they were 15-years old or older on the day of the interview. Unfortunately due to inaccuracies in date of birth information there are 43 individuals who are 14 years old in the Adult file and 26 individuals who are 15 years old in the Child file.

⁷ Proxy questionnaires were completed where possible for adults that were unavailable or unable to answer their own adult questionnaire. Proxy questionnaires were also completed for individuals that were out-of-scope at the time of the interview.

2.6 Identifiers

Individuals can be identified across Wave by their unique identifier *pid*. Households are identifiable within Wave by their unique identifier *w2_hhid*. Different household identifiers are assigned to each Wave as NIDS is a panel of individuals, and the household identifier is simply a tool to connect each individual to their household in each Wave. Households are not identifiable across waves except insofar as they are made up of the same individuals across waves. The *Link File* provides the information necessary to identify co-resident individuals across waves.

2.7 Merging Datasets Within & Between Waves

With the release of Wave 2, the longitudinal dimension of NIDS can now be explored. It is important to remember that NIDS is a survey of continuing sample members (CSMs), i.e. all persons that were resident in participating households in Wave 1. This has a particular consequence for the data structure and merging operations required to generate a panel dataset. This section is designed to provide users with the necessary information to understand how to merge within and between Waves. It also highlights important features of the data that can affect merges. Examples of the *Stata* code for how to merge within and between waves are provided below in Section 5.1 - the Program Library.

Note that for Wave 1, an important difference was introduced in the treatment of a polygamist individual in the dataset between Versions 3 and 4. Prior to Version 4.0, this polygamist had been recorded as two different people in two different households. It was only during Wave 2 that it became apparent that the same person had been recorded on the roster of two independently sampled households. This meant that this individual had two different personal identifiers (*pid*) in previous versions of the data. Now that we know that this is the same individual he has been assigned the same *pid* in both households. He is, however, only resident in one household.

This principle will be carried forward into future waves, i.e. a person can appear on multiple rosters, but can only be resident (usually sleep 4 nights a week) in one household. We accept that this might be difficult for some individuals (such as polygamists) to self-identify. In cases where a person is recorded as resident in two households we make him/her “resident” in the household in which he/she was actually interviewed and non-resident in all other households. In the unlikely event that a person is actually interviewed in more than one household, we will randomly assign him/her as resident in only one household. In sum, individuals with multiple memberships retain the same *pid* in all households in which they appear on the roster but are resident in one household only. This principle now applies to any individual that has multiple household memberships.

In NIDS Wave 1, resident CSMs were given a *pid* but non-resident household members were not; they have a missing value for *pid*. From Wave 2, TSMs are allocated a *pid* in order that we can match them across waves.

These features of the data have important implications for merging the datasets. We discuss these and make recommendations separately for merges within Waves and merges between Waves.

Merging Within Wave 1 or Within Wave 2

1. In general, we recommend that the principle when merging within Wave is that it should always be done on both *hhid* and *pid*.
2. When merging to the Household Roster file, which has duplicate records of *pid* in both Wave 1 and Wave 2, we recommend the following process:
 - a. In Wave 1 from Version 4.0 onwards, when merging files to the Household Roster, *pid* will no longer be unique for the one polygamist in the dataset. Therefore, we recommend that a merge on both *w1_hhid* and *pid* be performed.
 - b. In Wave 2, both resident and non-resident members now have a *pid*. Note that this does not affect the need to merge within Wave 2 on both *w2_hhid* and *pid* when merging to the Household Roster. This is of particular importance because an individual can be listed on more than one household roster. In other words, individuals can be resident in only one household, but members of more than one household.

Merging Between Waves

There are two ways to think about merging between waves:

1. NIDS is a panel of individuals. Therefore the person identifier (*pid*) is central to merging across waves. Within a given wave, a given *pid* will not be unique if the same individual is a member of more than one household. This prevents a simple merge across waves by *pid*. However, each individual can be resident in only one household. Therefore, before merging across waves a temporary version of the data from each wave can be created that sets the *pid* for non-residents to missing. These temporary data sets will be unique on *pid* within each wave, enabling cross-wave merging to take place on *pid*.
2. Merging between waves can also be done by firstly merging an existing Wave to the Link File using both *pid* and the relevant household identifier. The Link File contains the person identifier (*pid*) and household identifiers for both Wave 1 and Wave 2 (*w1_hhid*, *w2_hhid*). It also contains variable identifiers for CSMs and TSMs, and individual and household interview outcomes. Because the household identifier differs between Waves, the Link File plays an important role in mapping individuals to households in both Waves. Once the first merge from an initial Wave to the Link File has been made, the remaining merges to the datasets of interest in the alternative Wave can be performed.
 - Note that the Link File contains only resident household members (including deceased members). The Household Roster file in Waves 1 & 2 contain resident and non-resident household members (including deceased members). Caution therefore needs to be applied when merging the Link File to the Household Roster file.

2.8 Variable naming convention

Variables are named consistently across waves for ease of reference. Where questions are the same across waves the core of the variable name will be the same. If the question is slightly different a

different name will be given. Each variable, except unique identifiers, is prefixed with the appropriate Wave identifier, e.g. w1_ or w2_.

The naming convention used by NIDS is made up of several naming components and is constructed as follows:

Wave _ source _ section - subsection - main_descriptor - extension / subquestion

Details of each component are described below:

Wave

The Wave prefix, indicates which Wave the data was collected.

<i>Wave indicator</i>	<i>Meaning</i>
w1	Wave 1
w2	Wave 2

Source

The source indicates which dataset the variable belongs to.

<i>Source indicator</i>	<i>Meaning</i>
a	Adult file
c	Child file
p	Proxy file
h	Household file
r	household roster file

Section Leaders

Many of these follow a mnemonic convention using two or three letters. The conventions are not unique to sections in the questionnaires; rather, they are unique to the major topic that is covered.

Examples of significant section leaders are:

<i>Section Leader</i>	<i>Meaning</i>	<i>Section Leader</i>	<i>Meaning</i>
Em	Employment	inc	Income sources
Unem	Unemployment	mth	Mother
Noem	No employment (voluntary)	fth	Father
Ed	Education	agr	Agriculture
Hl	Health	fd	Food Expenditure
Bh	Birth History	nf	Non-food expenditure
Brn	Born	gr	Grant information
Lv	Living place	mrt	Mortality

Subsections

The subsections are used for grouping similar questions. There are a number of sub-sections to many of the main sections. Some of these are outlined below.

Within Employment:

Primary employment	emp1	Self employment	emps
Secondary employment	emp2	Casual employment	empc

Within Education:

School education(achieved)	edsch	Tertiary education (achieved)	edter
Repetition of grades	edrep	Education: literacy	edlit
Current Education	edcur	Education: intentions	edint
Education in 2010	ed10		

Within Health:

Ailments in last 30 days	hl30	Lifestyle	hl1f
Recent Consultations	hlcon	Smoker	hl1fsmk
Vision	hlvis	Difficulty of activities	hldif

Descriptors

The descriptors are the main part of the name which differentiates the question from the others in its section and subsection. These are usually one or two (appended) mnemonics formed from the most important descriptive parts of the question.

Sub-questions

Note that the sub-question is *not* a descriptor. Sub-questions *only* qualify a previous question, with a finite number of qualifying properties, such as location, value or explanation. A sub-question differs from an extension because it qualifies directly from a previous question. For instance where the question asks if the respondent sells the produce produced on their small-holding, that question is followed by an additional question asking the monetary value of the produce sold (e.g. w2_a_empsll_v). This variable is classified as a sub question of the "Do you sell produce?", and receives the suffix "_v".

2.9 Non-Response Codes

Non-response codes are usually indicated by negative numbers. The only exception is dates where four digits were used for years and two digits for months. Specifically the following non-response codes are used in NIDS:

Type of item non-response	Non-response code	Year	Month
Don't know	-9	9999	99
Refused	-8	8888	88
Not applicable	-5	5555	55
Missing*	-3	3333	33
Not asked in Phase 2 of Wave 2	-2	2222	22

*Missing (-3) indicates that a question was supposed to have been answered, but was not. A system missing (.) indicates that a skip pattern was enforced and that no data had to be collected.

2.10 Anonymisation

In order to protect the identity of our respondents every effort is made to remove personal information that could be used to identify them. Names and contact details are kept separately from the public release dataset and certain variables that are collected in field are not released or are only released at an aggregated level (e.g. occupation and migration data).

2.11 Secure data

In addition to the public release dataset, SALDRU also prepares an internal dataset that includes the full geo-coding, employment coding and PSU information. The Secure Datasets include text variables as they are captured in the questionnaire. Where possible, coded or aggregated information is released as part of the public release dataset, e.g. employment and sector codes to the one-digit level.

The purpose of the Secure Datasets is to allow users the opportunity to compare the NIDS data with administrative or other external data sources in an environment where the confidentiality of respondent information can be respected while allowing important data linkages to happen. The NIDS Secure Datasets only include information as collected infield. Special releases are made from time to time of Administrative data that has been matched to NIDS data.

Access to the Secure Datasets is only granted at the DataFirst's Secure Research Data Center in the School of Economics Building, Middle Campus, University of Cape Town, Cape Town. Secure data may not leave the premises.

Users wishing to access the Secure Datasets at NIDS are requested to complete a NIDS Accredited Researcher Application. If you are a student your application has to be counter-signed by your supervisor. The application will be reviewed by the NIDS management committee within two weeks of submission and you will receive feedback on the success of your application. If you are successful you will also be required to sign a NIDS Secure End-user Agreement. Both documents can be downloaded from the DataFirst website <http://www.datafirst.uct.ac.za/services/secure-data-services>.

Applications must be made by emailing the NIDS Accredited Research Application to: nids-survey@uct.ac.za.

2.12 Program Library

NIDS makes several Stata Programmes available to users to assist them in understanding how to use and manipulate the NIDS datasets. Also, we provide users with the Stata do-files used to create derived variables. See section 5.1 of this User Guide for a detailed list of these files.

3. Data Collection

In this section we describe the data collection process. There were two major changes in data collection methodology from Wave 1 to Wave 2:

1. The introduction of Computer Assisted Personal Interviewing (CAPI) as the means of data collection. This allowed us take advantage of a range of data assurance and quality checks.
2. Tracking of CSMs to new addresses. In addition to in-field information gathering on CSMs that had moved, NIDS also uses an in-house call-centre to assist with tracking.

These methodological changes required careful pre-testing (over and above the changes made to the questionnaire) to ensure that the systems and field protocols functioned correctly. At the level of interviewing, the CAPI system followed the paper instruments as closely as possible.

Paper consent forms were issued in all languages and the informed consent process was conducted in the respondent's language of choice. For each questionnaire, two consent forms were signed. One signed copy remained with respondents and the other was returned to SALDRU. These forms carried unique bar-coded numbers that were entered into the CAPI system; similarly the household and person level IDs were displayed on the CAPI system and written onto the consent forms to cross-referencing was possible. Data coming in from the field were accepted as valid only if SALDRU had a signed consent form for each interview that produced the data. If signed consent forms were not located, the associated interviews were deleted from the data set.

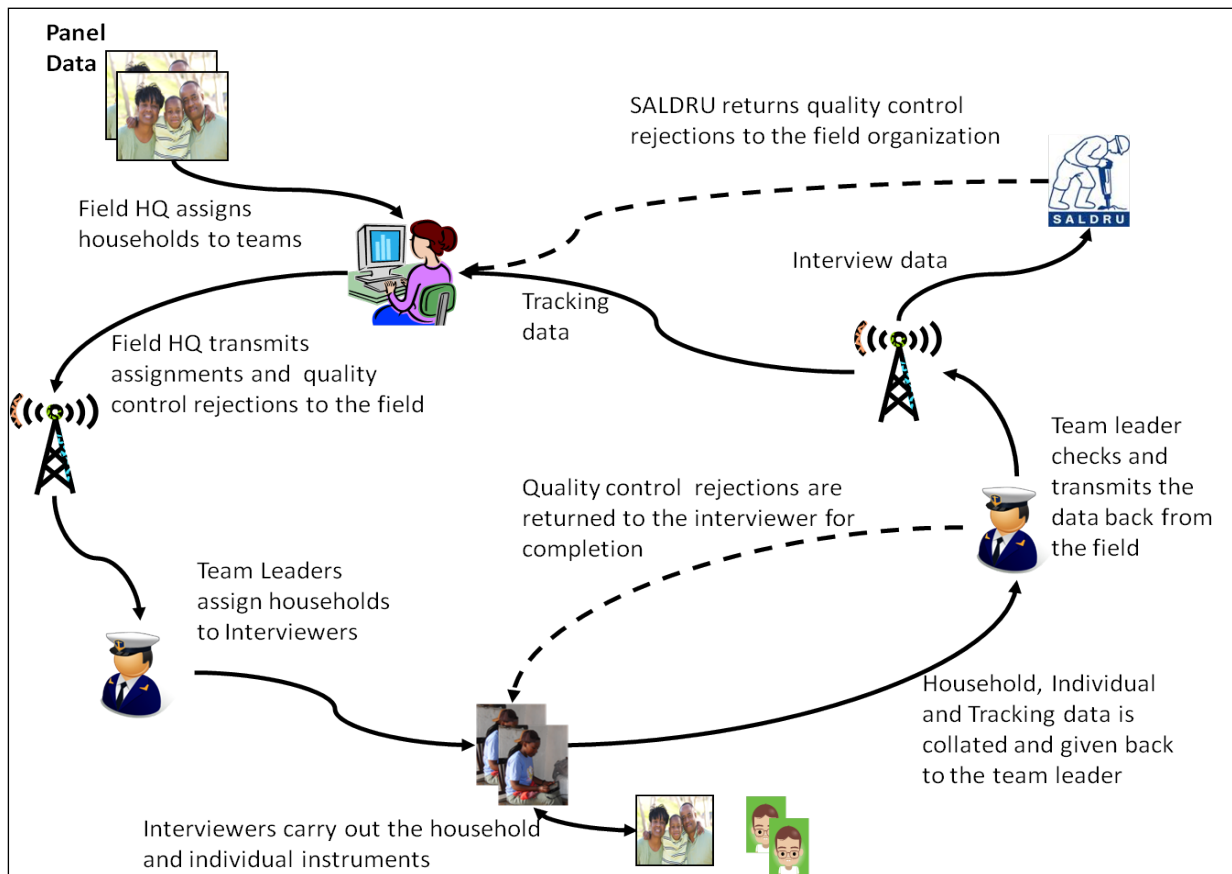
As in Wave 1 four types of questionnaires were administered:

- **Household questionnaire:** One household questionnaire was completed per household by the oldest woman in the household or another person knowledgeable about household affairs and particularly household spending. Household questionnaires took approximately 45 minutes in non-agricultural households and 70 minutes in agricultural households to complete.
- **Individual Adult questionnaire:** The Adult questionnaire was applied to all present Continuing Sample Members and other household member's resident in their households that are aged 15 years or over. This questionnaire took an average of 45 minutes per adult to complete.
- **Individual Proxy Questionnaire:** Should an individual qualifying for an Adult questionnaire not be present then a Proxy Questionnaire (a much reduced Adult Questionnaire using third party referencing in the questioning) was taken on their behalf with a present resident adult. On average a Proxy questionnaire took 20 minutes. Proxy Questionnaires were also asked for CSMs who had moved out of scope (out of South Africa or to a non-accessible institution such as prison), except if the whole household moved out of scope, and could therefore not be tracked or interviewed directly.
- **Child questionnaire:** This questionnaire collected information about all Continuing Sample Members and residents in their household younger than 15. Information about the child was gathered from the care-giver of the child. The questionnaire focused on the child's educational history, education, anthropometrics and access to grants. This questionnaire took an average of 20 minutes per child to complete.

3.1 Overview of CAPI Cycle

The CAPI cycle is illustrated below.

The CAPI cycle



Listing data (PSUs, household addresses, contact details, roster make up and individual contact details) drawn from Wave 1 was pre-loaded into the CAPI system. Also included were panel data on individuals covering items not expected to change (e.g. birth date and preferred language), or to change within a predictable range (e.g. highest level of education attained). Panel data was used within the CAPI system as a form of quality control and to ensure the respondent was the correct one. Listing data were centrally distributed via modems to field teams on a cluster by cluster basis prior to their arrival.

Using handheld devices (Ultra Mobile PCs or UMPCs) the fieldworkers conducted the surveys and validated the content. Field Team Leaders then re-validated the fieldworker data prior to transmission back to NIDS (SALDRU in the diagram above).

The data arrived at NIDS in the form of a relational database that was then merged into flat Stata files matching the instrument's uses (Household, Adult, Child and Proxy). These flat files were then validated again, with any data inconsistency or non-response issues returned to the field company directly, or checked via calls to the respondents.

3.2 Overview of the Tracking Process

An essential part of the Wave 2 data collection process was to track CSMs from Wave 1. CSMs could either be in the same location as the dwelling unit of interview in Wave 1 or they could have moved. Interviewers used the CAPI system to load address and contact details for movers (either "Whole Household Moved" or "Household Splitters"). The field team leader would then assess these details to:

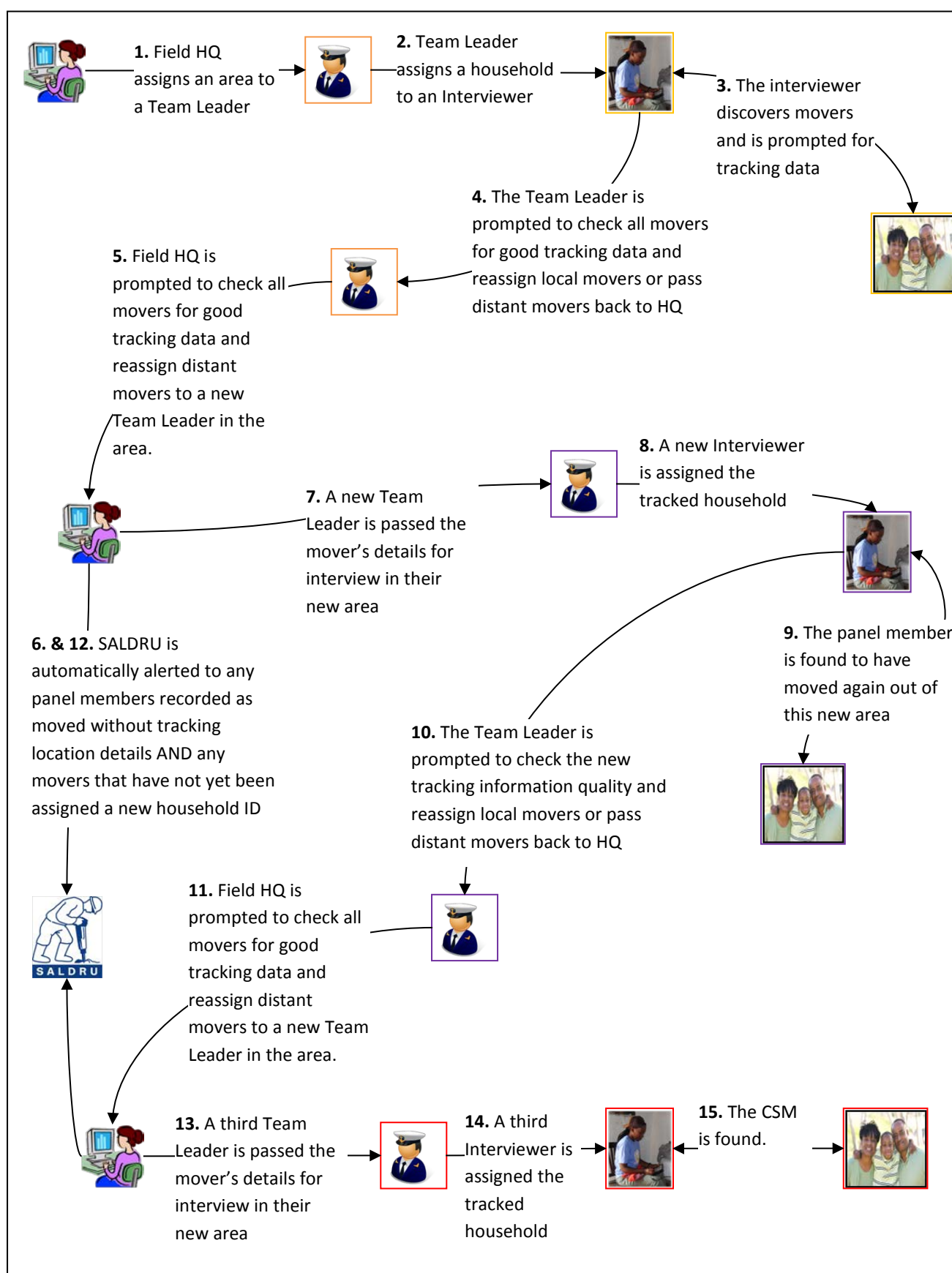
1. Generate new household IDs locally containing the movers to be dealt with by that team; or
2. Transmit the location details back to field control to generate household identifiers for movers and assign them to the relevant team on a geographical level.

Households were created around location details. These location details were indexed and linked to respondents. A household ID was generated for each location with new CSM records linked to that household ID and corresponding to the indexed location details. These identifiers were finalised only after the location of the CSM was confirmed.

Where no useable data was available for movers, household and person records were moved to a dummy "PSU" signifying lost in tracking. In these cases SALDRU examined the location information available and the contact details of the originating household in an attempt to improve or verify the mover details. Where this was successful, these households were sent "back to field" for completion.

The process is illustrated below:

Tracking movers



3.3 Listing and Panel Data

The listing data for household location was drawn from Wave 1 information. All members of the original selected households (including children) were selected to be tracked indefinitely (known as continuing sample members or CSMs). In addition the panel data was used to verify respondent information which included birth dates, gender, preferred language, and schooling in Wave 2.

Listing data and additional information were pre-populated onto the CAPI device screens to aid with household and person identification (e.g. gender and birth dates on the household roster) and facilitate data entry. Other Wave 1 information was sometimes not displayed, but was used by the CAPI system to challenge inconsistent answers (e.g. attendance at school during Wave 1).

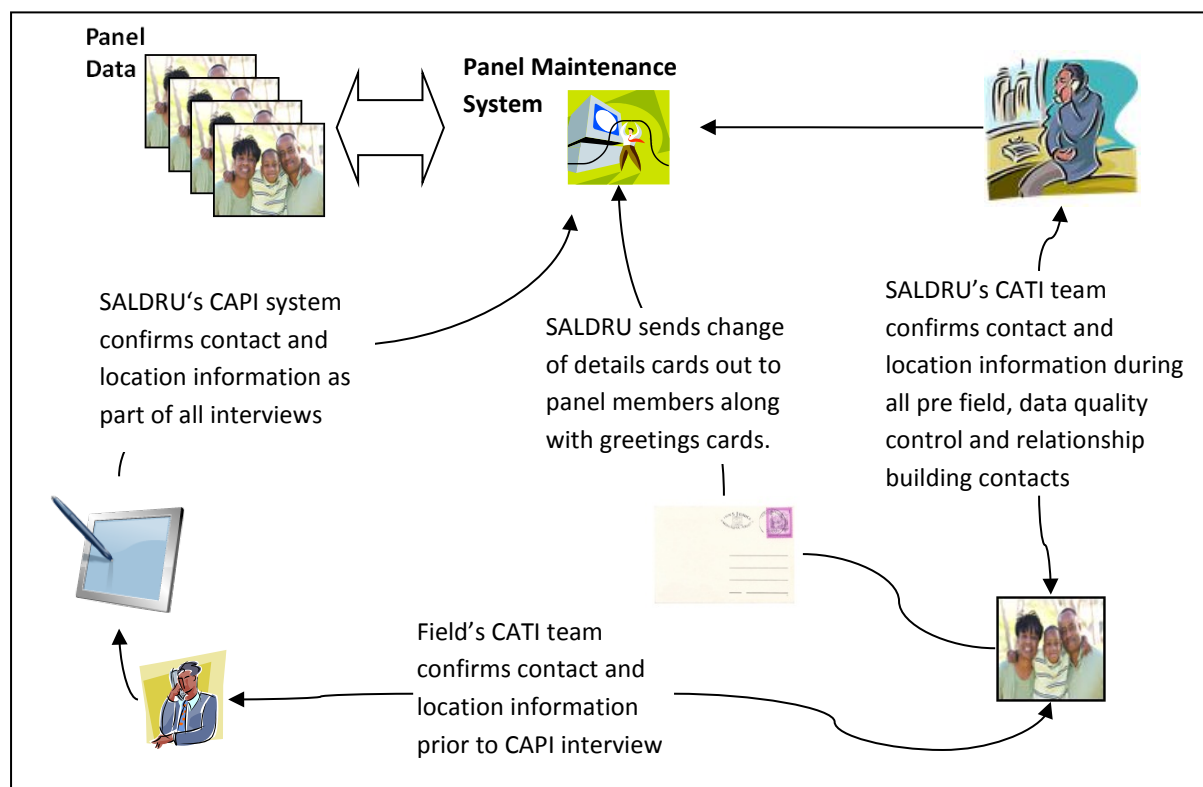
Certain pre-populated data were used to skip questions if valid answers had been discovered in Wave 1, an example being head circumference of a child at birth.

Where Wave 2 answers were inconsistent with Wave 1, the interviewer was challenged to confirm the answer and enter any substantiating notes for the change.

3.4 Contacting Respondents

A Panel Maintenance System integrated into a Computer Assisted Telephonic Interviewing (CATI) Call-Center at SALDRU's offices at the University of Cape Town plays a major role in how SALDRU interacts with panel members. The diagram below provides a schematic overview of the process:

Contact Procedures



The reasons for contact with respondents often differ – from arranging a time for an interview to checking the veracity of information through telephonic follow-ups post-interview. The contact details for all respondents are maintained centrally and updated by (1) the upload of CAPI field data, (2) post-interview “call backs” through a Call Centre System, and (3) through the post (a postcard and change of address card was sent out between Waves 1 and 2 to maintain contact with panel members and allow them to inform us of any address changes).

3.5 Data Quality Issues and Data Collection

Data quality issues that arose and were mitigated in the data collection process included the following:

Unit Non-response

Unit non-response was minimized through a series of measures:

1. **Valuing Panel Members:** Along with the unconditional gifts given to respondents, information pamphlets about NIDS translated into all eleven official South African languages re-explained what the survey was about and the value of respondent’s contribution. Similarly written records were left with respondents about their anthropometric data including whether to seek medical advice over their blood pressure readings; anecdotal evidence is that this information was highly prized by respondents. SALDRU also carried out random call backs to respondents to ensure that they were treated courteously and to collect any respondent feedback on their experience. In this way, survey participation was encouraged as much as possible.
2. **Multiple Listing Information:** The original Wave 1 maps and listings were issued along with the captured Wave 1 address data (including GPS coordinates). Interviewers were required to retake GPS coordinates of Wave 1 households (as well as new ones for movers) to enhance listing for Wave 3. In this way, non-contact was minimised amongst CSMs.
3. **Tracking systems:** The CAPI devices carried a search function to search on town or local area to identify the mover location from province down to Main Place level to further support the test address and telephone details taken for movers. This was also done in an effort to minimise non-contact.
4. **Household Level Non-response Call Backs:** Households may have come back from field as a refusal, dwelling-unit vacant or un-locatable / un-traceable. Households that came back from field as refused were contacted by SALDRU to confirm this refusal and attempt to overturn it; where refusal was overturned these would be returned to the field company for re-interview. Where the field organisation failed to track individuals, SALDRU would further investigate using the history of co-residents and alternative contacts for movers. Operationally, this was done through the SALDRU call-centre with the Panel Maintenance System.
5. **Individual Level Non-response Call Backs:** SALDRU attempted to contact all individual level refusals to confirm this refusal and attempt to overturn it; where refusal was overturned these would be returned to the field company for re-interview.
6. **Field organizations rewards:** The contract with the field organization contained a large variable component that encouraged maximum effort to keep attrition low.
7. **CAPI pre-population:** Pre populating the CAPI roster along with the automatic insertion of the relevant names into individual’s questions ensured easy monitoring that all CSMs were being

approached and that the correct roster members were being referred to in their individual questionnaires.

8. **No one at home policy:** Should there be no one at a dwelling, the interviewer was required to visit no less than 3 times at three different times of day, on at least two different days before recording a household as non-respondents.
9. **Phase Two of Wave 2:** In June 2011 NIDS commissioned a Phase Two of Wave 2 as a Non-Response Follow-Up from Phase 1 of Wave 2. Household included in this subsample where those that refused and those that could not be located or tracked in Phase 1. Out of a total of 1064 households attempted, an additional 389 households were successfully interviewed in Phase Two.

Item Non-Response

Item non-response can arise for different reasons, for example when a respondent refuses to answer a question or doesn't know the answer, or if the interviewer mistakenly skips over a question. "Don't Know" and "Refuse" response options are coded accordingly, allowing users to estimate item non-response rates for relevant questions.

The use of CAPI radically reduces the instances of interviewer-induced item non-response because CAPI automates the skip pattern for the interviewer and prompts them if a question in each section of the questionnaire has been left blank. Since this was the first trial run with CAPI, a cautious approach was taken and data was accepted from field if all sections were more than 50% complete. Any instruments submitted as finished from field that had more than 50% of data missing was returned to field for completion.

Data Consistency

Over and above the issue of item and unit non-response is the internal consistency of the data: within instrument, across instrument, and across Wave. Data collection involved several checks and mitigations:

1. **Translation, Respondent understanding and Measurement Error:** the CAPI system held all questions, prompts and pre-coded responses in all 11 official South African languages. Translations were outsourced to a translation company before loading to CAPI. However, some translation error was picked up in the field, though the magnitude of this error is likely to be very small since the overwhelming majority of interviews took place in English. To reduce interviewer effects SALDRU made some use of the context sensitive help afforded by the use of CAPI.
2. **CAPI consistency checks:** the CAPI system had a range of within questionnaire consistency checks such as feasible height weight ratios, birth rates, age versus date of birth etc. In addition cross questionnaire checks were also built in such as cross checks between the roster data and individual questionnaires (for example consistency between children on the roster and the birth details given by a mother). Panel data is also used for cross-Wave CAPI validation, an example of which was prompting the interviewer if schooling appeared to have advanced too far between waves. All of these checks were carried out on a screen by screen basis by

interviewers (during the interview), on a household basis by their Team Leaders (as a monitoring process at the close of each day) and at a cluster (PSU) level by field controllers (as a monitoring process several times a week) using the CAPI system.

3. **In Wave pattern searches:** SALDRU carried out a range of pattern searches on the data during field to identify interviewer effects and mis-capture. When areas of concern were found, the respondents / households were contacted to ensure that the data were correct. Where call backs did not allay concerns about mis-capture, action was taken in line with the “cross Wave consistency checks” detailed below.
4. **Cross-Wave consistency checks:** SALDRU used information gathered in Wave 1 to create a set of cross-Wave consistency checks. Where discrepancies arose, households were contacted again to verify the information.
5. **Live behavioural correction:** the use of CAPI in Wave 2 allowed live checking of data quality from the commencement of field. Through returning data “back to field” for recollection in a timely fashion, NIDS was able to mitigate and normalise the most obvious interviewer effects.

The Mechanics of Data Quality Checks

In this section we discuss three main data quality checks that were run concurrently or after the fieldwork process, including (1) early identification of identifier mismatches; (2) returning information back to field; and (3) correcting data issues with call backs. Since CAPI allowed the interviews to be downloaded by SALDRU in real time, the data quality process could commence in real time.

Early Identification and Cleaning Of Identifier Mismatches

As part of cleaning the NIDS dataset, we performed basic cleaning of the data in its raw relational data form, before the data was converted to the five flat files, namely the Adult, Child, Proxy Household questionnaire and Household roster data files.

The cleaning at this level consisted of ensuring identifiers for these files were correct and consistent. Identifier mismatch typically arose from:

- Erroneous moving of households, which created new household identifiers when in fact the household remained intact and at their original physical address. In these cases the household identifiers were returned to their original household ID.
- Mover CSMs splitting from differing households but moving in together, which created the situation of one CSM being recorded as a TSM (the new household having been created around the other splitter). This happened very infrequently.

Identification of these problems occurred through:

- Automatic checks built into the flat file creation process that highlighted interview data from households not appearing in the same location.

- Queries raised through data consistency checks on the flat files such as pattern matching on key variables (DOB, name, gender etc.) indicating that a TSM in a mover household was likely a splitter CSM from a third household.
- System merge error detection during flat file production.

Following telephonic investigation to confirm the existence and nature of an identifier problem, automatic identifier fixes were built into the flat file production code for the next daily CAPI data upload.

Returning Incorrect Data “Back To Field”

SALDRU reserved the right to reject questionnaires as invalid on evidence of:

- Aberrant field behaviour (for example clear evidence of invention of data, unfeasible numbers of proxies rather than direct interviews etc.).
- Less than 50% of any section in the questionnaires completed.
- Not all individuals in the household attempted.
- No consent forms collected for respondents.
- No GPS coordinates were collected for households successfully interviewed or households found but with valid non-response outcome⁸.
- Invalid “No one at home”. Field teams had to demonstrate that they had visited the households and individuals on at least two different days at three different times.
- Invalid non-response.

If a questionnaire was deemed invalid by SALDRU’s data quality checks , it was sent “back to field” and a further in-person interview was required (i.e. telephonic interviews were also not permitted to resolve “back to field” issues).

SALDRU reported back to field instances twice a week to the field company who were required to rectify them within the pre-set field period.

The SALDRU team attempted to call all non-response households to ensure that the field teams had tried enough times to get hold of the respondents, refusals were genuine or that households could really not be contacted or physically located (i.e. not “invalid non-response” from the above list). If the SALDRU team got in contact with the respondents and they were willing to participate in the survey then these were returned as “back to fields” to the field company in the form of an exception report.

⁸ Valid unit non-response outcomes – Refused, No one at Home.

3.6 Fieldwork schedule

Pre-test

As part of the preparations for fieldwork a full system pre-test was conducted that acted as a trial run for all the components of NIDS fieldwork; training fieldworkers, locating and tracking respondents, administering the questionnaires, etc. By using the same sample as pre-test in Wave 1, all aspects of the panel and pre-population can be tested. CAPI was introduced for the first time in Wave 2 pre-test. In this Wave, the pre-test contributed significantly to our understanding of how an electronic data collection exercise differs from a paper based system. The pre-test tracks 586 individuals from 160 households where they were resident in Wave 1 pre-test. These households originated in 8 clusters (4 in Kwa-Zulu Natal, 3 in Gauteng, and 1 in North West province). The distribution of the clusters is aimed at covering a range of demographic and geographic scenarios. As with the main survey all resident CSMs are tracked when they move within South Africa. For Wave 2 pre-test fieldworker training was conducted in January 2010 and the fieldwork in February 2010.

Main data collection

A staggered fieldwork approach was used to roll out across the country. There were four training sessions with subsequent training sessions building on lessons learnt in previous training sessions and realities of what other teams struggled with in field. In total there were 120 fieldworkers who operated in teams of 4 – 1 team leader and 3 interviewers. Occasionally team sizes varied depending on the region and/or typical household characteristics for that area.

Fieldwork ran over into 2011. This necessitated a few small changes to the questionnaires, especially the education sections to ensure that comparative data was collected.

Phase 2

In mid-2011 it was decided to exercise the option to implement a “Phase 2” for Wave 2. Internal data checking revealed that there were a variety of households that we believed could be successfully interviewed through a focused mop-up phase. The focus of Phase 2 was:

- Movers that had not been tracked by the end of Phase 1;
- Re-attempting to locate households that had been labelled as Not Located in phase 1;
- Re-attempting households that had been unavailable during Phase 1, but which now indicated their availability; and
- To overturn household level refusals from Phase 1.

It was decided not to re-attempt individual refusals that remained after Phase 1 as this would have placed an undue burden on those households that had already participated. Through the efforts of Phase 2, fieldwork was completed and response rates improved. Due to the dispersed smaller sample, more mobile fieldworker teams were deployed. There were 40 fieldworkers working in teams of two.

Fieldwork for Wave 2 (including both Phase 1 & Phase 2 fieldwork) commenced in May 2010 and concluded in September 2011. There were breaks in fieldwork from 15 December 2010 to 3 January 2011 and again from 9 May to 1 August 2011.

Questionnaire Differences between W2 Phase 1 & W2 Phase2

There are two important methodological differences between Phase 1 and Phase 2:

1. Not all sections of the original Wave 2 questionnaires were asked. This reduced respondent burden and the time required for fieldworker training. Questions NOT asked in Phase 2 are indicated in the data with the non-response code “-2”. Core modules such as household composition and income were still asked. Consult the Wave 2 Phase 2 questionnaires for more details of these differences.
2. Movers out of Phase 2 dwelling units were not tracked further. Address information was collected for this sub-sample and they will be tracked as part of the Wave 3 fieldwork exercise. These individuals are classified as “Not tracked” in the Wave 2 dataset.

3.7 Response Rates & Attrition

In Wave 1, 10367 dwellings were selected to be approached to take part in NIDS. Of those dwelling units, 491 (4.5%) were found to be multi-household dwellings. Of the 10858 eligible households, 7296 agreed to participate (these numbers are taken from Version 4.1 of Wave 1 and Version 1.0 of Wave 2).

Wave 1 Household Responses

	Number	Percent
Total dwelling units sampled	10367	
Plus multi-households added to sample	491	
Total Potential Sample	10858	100%
Less Vacant and out-of-scope Dwelling Units	523	5%
Less No-Access Areas	119	1%
Less Non-contacts	1214	11%
Less Refusals	1698	16%
Participating Households (as % potential sample)	7301	67%
Participating Households (as % potential sample less vacant and out-of-scope)	7301	71%

Within the participating households, 31163 individuals were identified as household members. However, 2916 people were not resident members and were thus excluded from the study (so as to avoid double counting, as they had a chance of being selected for the study at their “usual” place of residence). A resident member was defined as a person who usually resides at the dwelling four nights a week. In addition, we included in the sample non-residents who were currently residing in institutions that are regarded as “out-of-scope”, such as a hospital, prison or student hostel. All these sample members, including children, are continuing sample members and will be re-interviewed in subsequent waves.

Wave 1 Individual Responses

	Number	Percent
Household members	31163	
Less Non-resident members	2916	
Continuing Sample Members	28247	100%
Adult Questionnaire	15633	55%
Proxy Questionnaire	1753	6%
Child Questionnaire	9408	33%
Individuals refusal or unavailability	1453	5%

Of the possible 28247 CSMs from Wave 1, 22050 were re-interviewed in Wave 2. When excluding those that moved out of scope or died between waves, the attrition rate is 19%. The table below summarises individual outcomes between Wave 1 and 2.

Wave 1 and Wave 2 Individual Outcomes

Wave 1	Wave 2					
	Success	Refused/Not available	HH NR	Moved outside SA	Dead	Total
Success	21 098	520	4 279	51	846	26 794
Refused/Not available	952	91	365	2	43	1 453
Not in W1	6 591	209	23	0	140	6 963

It is important to note that non-respondents in Wave 2 (the sum of refusals, household non-response and those that moved outside of SA) are not lost to the panel in perpetuity. Their names and contact details remain on record and they will be attempted again in future waves. We also keep the contact details of up to three additional friends or family members of each respondent on record. These are used to help find CSMs as part of panel maintenance.

The reasons for household non response include:

Reasons for household non-response at the individual level

	Number	Percent
Refused/Not available	1814	38.87
Not located	2224	47.65
Not tracked	629	13.48
Total	4667	100

The biggest reason for individual non-response is household level non-response and the major reason for household level non-response (48%) is that the household could not be located. This refers to households that moved but insufficient or incorrect information was collected about their new address and the actual dwelling unit could not be found. Also included in this subset are 254 original dwelling units from Wave 1 that could not be found in Wave 2. Some of this was due to significant redevelopment or relocation of informal settlements.

The reasons for attrition between Waves 1 and 2 include:

Reasons for Attrition

Reason	Number	Percent
Refusal	2136	37.50
Non-contact	2714	47.65
Deceased	846	14.85
Total	5696	100.00

The table shows three categories of attrition: “Refusals” are attriters who were not interviewed in Wave 2 because of an individual or household refusal. “Not contacted” individuals consist of respondents who were not tracked, not located or moved outside South Africa. Finally, there are respondents who died between waves.

The rate of attrition by province is:

Attrition by Wave 1 Province

Decile	Number of Refusals	Number of Non-Contacts	Number of Deceased	Total Number of Individuals	Attrition Rate (%)
Western Cape	594	215	70	879	28.34
Eastern Cape	208	412	152	772	22.55
Northern Cape	166	112	51	329	18.98
Free State	149	58	64	271	18.78
KZN	281	439	250	970	13.21
North West	180	112	79	371	16.63
Gauteng	322	211	62	595	25.27
Mpumalanga	116	65	55	236	13.94

Limpopo	105	97	60	262	10.77
Outside RSA	0	51	0	51	63.75
Missing	15	942	3	960	100.00
Total	2136	2714	846	5696	21.26

The table shows that the number of people that refused compared to those that could not be contacted and those who died. Provinces where the ratio of refusals exceeded non-contacts were the Western Cape, the Northern Cape, the Free State, North West, Gauteng, Mpumalanga and Limpopo. Of the 5 696 attriters, 960 do not have a province in Wave 2. These are movers for whom we have no location information.

Of the nine provinces, the highest attrition rate is in the Western Cape at 28%. Gauteng is next at 25%, followed by the Eastern Cape at 23%. Limpopo, with 11%, had the lowest rate of attrition.

Further insight into the incidence of non-response is presented below, where we disaggregate attriters by income decile.

Attrition by Wave 1 Income Decile

Decile	Number of Refusals	Number of Non-Contacts	Number of Deceased	Total Number of Individuals	Attrition Rate (%)
1	125	322	67	514	18.61
2	124	266	78	468	17.17
3	79	228	81	388	14.39
4	158	230	70	458	16.98
5	122	212	93	427	16.07
6	148	251	116	515	19.14
7	174	260	91	525	19.79
8	191	271	122	584	22.05
9	310	327	77	714	27.34
10	705	347	51	1103	41.59
Total	2136	2714	846	5696	21.26

The table shows that non-contact, rather than refusal, is the main reason for attrition for deciles 1 to 9. Interestingly, however, this is reversed for the top decile where the ratio of refusal to non-contact is about 2:1.

Analysing the attrition rate by Wave 1 income deciles show that the richest 10% attritted at a far higher rate than those in deciles 1 to 9. 41.59% of the top decile were not successfully re-interviewed in Wave 2. Attrition rates in deciles 1-7 are in the 14% to 20% range.

The racial distribution of attrition is presented below.

Attrition by Racial Group

Pop. Group	Refusal	Non-contact	Deceased	Total	Attrition Rate
African	1054	2067	710	3831	18.05%
Coloured	480	427	100	1007	26.10%
Asian/Indian	112	30	8	150	39.47%
White	490	190	28	708	53.11%
Total	2136	2714	846	5696	21.26%

Here we see that Non-contacts are the dominant reason for attrition among African respondents, while Refusals dominate for White, Indian and Coloured respondents. The population groups with the highest attrition rates are Whites and Indian respondents. These are also expected to be in the highest income decile.

4. Derived Variables

As with the first Wave of NIDS, certain variables in the derived datasets are created at NIDS. These variables appear in the household derived and individual derived datasets. Derived variables are created for:

- Any variable that is finalised after field through a post-coding exercise;
- Any variable that is the result of a combination of other variables;
- Any variable that is imputed and that is part of public release data.

Examples of derived variables include “best” variables, geographical variables, employment variables, income, expenditure and wealth variables. The process leading to the creation of the variable or variable groups is discussed below.

4.1 Best Variables

Certain information should remain unchanged or at least internally consistent for individuals across the waves. Examples include education, gender, population group, date of birth and age. We might get better information in a subsequent wave or we may get no information if they are a non-response. In order to present what we estimate to be the best known information for each of our respondents the relevant variables from the individual questionnaires and roster(s) for all the waves are compared for consistency. Naturally, item non-responses are excluded from the comparison. In the few cases (typically around 1% of cases) where there are inconsistencies, best is set to the answer that has appeared most often across the waves. If there is no mode or more than one mode then best is set to the answer from the last individual questionnaire. This is done for every respondent that has been resident in a surveyed household. The result is that best may not be calculated within wave, but it is consistent across waves. Where necessary additional calculations are done within wave for the *indderived* file, for example *best_age* is calculated within each wave using the best date of birth and the date of interview for that wave.

4.2 Geography

Household addresses and GPS coordinates were used to assign geographical codes for Province, District Council and Geographical Type (urban, rural). The GPS coordinates were used together with an algorithm placing the points within relevant borders (defined by province, district council and enumerator area shape files). When there were missing GPS coordinates (e.g. when the information was not recorded by the interviewer), the household address information was used to assign GPS coordinates using Google Earth and/or Maps, and fuzzy string matching was also used to retrieve Main Place level geographical information to aid in the assignment of Province codes when any ambiguity arose (e.g. when tracking mover individuals and households).

Wave 1 GPS coordinates and geographical data was integrated into this process where relevant (e.g. for those who remained in the same dwelling unit). The Wave 1 GPS coordinates underwent their own cleaning process. The most important of these were standardising the format in decimal degrees (as opposed to degrees-minutes-seconds) and replacing GPS coordinates to the centroid of

the PSU when the point location of the household was clearly outside a five kilometre circumference of the known PSU border.

For Wave 1 and Wave 2 a variable was defined (*w2_stayer*) at the individual level for respondents that remained in the same dwelling unit between waves. This variable identifies three types of continuing sample members: stayers, movers and new respondents. It was created in two steps: firstly from a variable identifying the migration of individuals from one house to the next, and secondly from the GPS coordinates for Wave 1 and Wave 2.

4.3 Occupation

Occupation was coded in two parts. Firstly, occupations were automatically grouped together based on the descriptions given to us by respondents into a list of occupational codes found in the International Standard Classification of Occupations (ISCO) code list. This grouping process was initially done and quality controlled electronically using a fuzzy string matching algorithm, which grouped similar words together and matched words incorrectly spelt by the interviewer into likely alternatives. The second part involved hand-coding those descriptions that the algorithm could not identify. This meant providing NIDS survey assistants with the occupation descriptions and ISCO codes, as well as the work description data given to us by respondents. A manual matching process was then performed.

These codes were then truncated down to the one-digit level and included in the Public Release data. Disaggregated occupational codes are available at NIDS on the secure data computer.

4.4 Industry

Industry coding was done in two parts, similar to occupational coding. Part one also involved an automated computer process using a fuzzy string matching algorithm to link the main goods or services provided by the employer to the industry description found in the International Standard Industrial Classification (ISIC) code list. The second part involved hand coding the descriptions that the algorithm could not identify.

These codes were then truncated down to the one-digit level and included in the data.

4.5 Employment Status

Employment Status was coded using the International Labour Organization's definitions to assign respondents to one of the following categories - Employed, Unemployed (strict definition), Unemployed (broad definition) and Not Economically Active.

The respondent was determined to be employed if they were economically active and reported having any form of employment, including a primary job, secondary job, self-employment, paid casual work, personal agricultural work, or if they assisted others in business activities. Unemployment was differentiated into broad and narrow unemployment as per the definitions, viz. by distinguishing those who desired a job and were actively searching for work from those not actively searching.

We also developed an employment variable for proxy respondents in a similar manner.

4.6 Income

Total household income (*w2_hhincome*) was derived from variables in the adult, proxy and household datasets. The variable reflects regular income received by the household on a monthly basis, net of taxes, as well as imputed rental income from owner-occupied housing.

The aggregate measure was derived in one of three ways. If all adult household resident members were successfully interviewed, *w2_hhincome* is the aggregation of all income sources for all individuals in the household. If, however, an adult respondent refused to be interviewed or was not available (partial unit non-response), we used the so-called “one-shot” income variable *w2_hhq_incb* as the measure of household income. Finally, in households where there was partial unit non-response and one-shot income was missing, we aggregated any income data we had from the remaining responding household resident members. Imputed rental income from owner-occupied housing *w2_hhimprent* was added to all households, irrespective of the method of aggregation, where appropriate.

Sources of Aggregate Household Income

Source of HH Income	Number of HHs	Percent
Individual Aggregation	5738	83.89
One-shot	1100	16.08
Aggregation with PUNR	2	0.03
Total	6840	100

The table below lists the variables that make up each component of total household income.

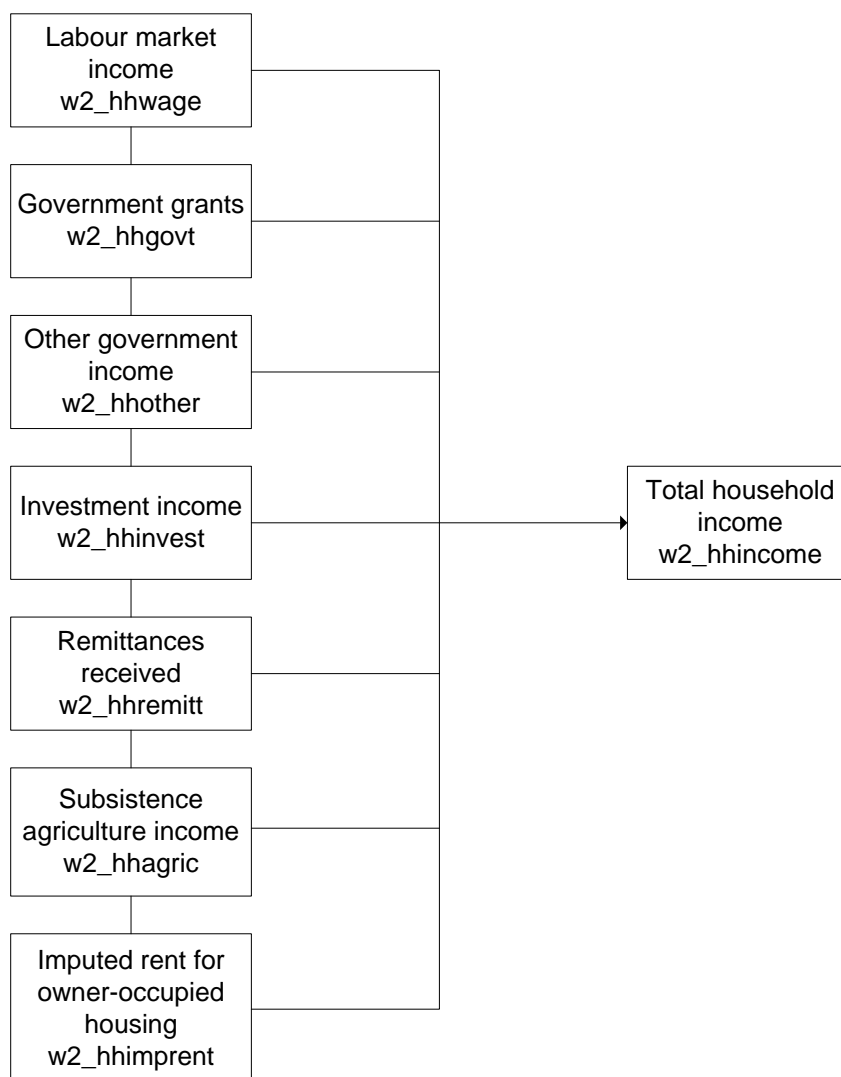
Components of Aggregate Household Income

Household-level Variable	Individual-level Variable	Variable Name
Labour Market Income <i>w2_hhwage</i>	Main and second job	<i>w2_fwag</i>
	Casual wages	<i>w2_cwag</i>
	Self-employment income	<i>w2_swag</i>
	13th cheque	<i>w2_cheq</i>
	Bonus payment	<i>w2_bonu</i>
	Profit share	<i>w2_prof</i>
	"Help friends" income	<i>w2_help</i>
	Extra piece-rate income	<i>w2_extra</i>
Government Grant Income <i>w2_hhgovt</i>	State old age pension	<i>w2_spen</i>
	Disability grant	<i>w2_dis</i>
	Child support grant	<i>w2_chld</i>
	Foster care grant	<i>w2_fost</i>
	Care dependency grant	<i>w2_cdep</i>

Other Income from Government w2_hhother	Unemployment insurance Fund	w2_uif
	Workmen's compensation	w2_comp
Investment Income w2_hhinvest	Interest/dividend income	w2_indi
	Rental income	w2_rnt
	Private pensions and annuities	w2_ppen
Remittance Income w2_hhremitt	Remittances received	w2_remt
Subsistence Agricultural Income w2_hhagric	Income from subsistence agriculture	w2_plot
	Value of own production consumed	w2_opro
Imputed Rental Income w2_hhimprent	N/A	N/A

The seven variables in the first column in Table 2, above, were summed to create aggregate household income.

Components of Aggregate Household Income



Bracket Responses

For certain variables, if respondents were not able to provide a point estimate for the amount of income from a particular source, a response was elicited through a series of unfolding brackets. Where respondents indicated that they fell inside a bracket, the mid-point of the interval was assigned. Those who indicated that they received income above the value of the highest bracket were assigned twice the value of the bracket⁹.

Item Non-Response and Imputation

Item non-response occurs when the respondent refuses to answer a particular question in the survey or states that they “Don’t Know” the answer. In these circumstances, imputation can be performed on the individual variables affected. This was conducted only once a few qualifying conditions were satisfied. Single imputation regressions were run only when there were a) 100 or more “valid” responses for a variable and b) the extent of missingness did not exceed 40%. Pre-imputation, post-imputation and imputation flags are available in the individual derived and household derived datasets for each variable that was imputed.

A rule-based imputation process was followed for the state old age pension, child support grant, disability grant, care dependency grant and foster care grant. Respondents acknowledging receipt of one of these grants, but failing to provide an amount, were assigned the maximum value of the grant for the month in which the interview took place. This is because individuals receiving one of the state grants rarely receive less than the full amount.

The table below describes the extent of missingness for each component of income, as well as the imputation method used to impute for item non-response. As in Wave 1 (see Finn, Franklin, Keswell, Leibbrandt & Levinsohn, 2009), imputed rental income from owner-occupied housing posed the largest problem. The value of imputed rental income from owner-occupied housing come from the question “What is the value of monthly rent you *would* pay if you had to pay to stay here?” which is asked in the household questionnaire. The question is relevant to those households that own the primary dwelling unit (whether or not the mortgage is fully paid off) and those who don’t own and don’t rent the dwelling unit, and are living in it free of charge. As in Wave 1, we ignored our rule-of-thumb of 40% missingness or less, and imputed for the 41.91% of households that were missing on this question. This represents the only case where the 40% rule was not followed.

⁹ Note that this practise is associated with estimating a Pareto Index for the upper tail of the distribution (see Cowell, 2000 for motivation). Wittenberg (2011) estimated the Pareto Index for the individual income distribution for multiple survey years for South Africa from 1995-2007. We utilise this source for our imputation procedures for the highest income category.

Income Variable Item Non-response

Variable	Description	Obs	Achieved	% Missing	Imputation
w2_fwag	Main and secondary wages	4321	4009	7.22	Regression
w2_cwag	Casual wages	541	528	2.40	Regression
w2_swag	self-employment income	648	478	26.23	Regression
w2_chcq	13th cheque	227	154	32.16	Regression
w2_prof	Profit share	31	19	38.71	None
w2_extr	Extra payment	73	63	13.70	None
w2_bonu	Bonus income	82	62	24.39	None
w2_othe	Other income	123	121	1.63	Regression
w2_help	Help friend income	57	51	10.53	None
w2_spen	State pension	2146	2138	0.37	Rule
w2_ppen	Private pension	362	334	7.73	Regression
w2_uif	UIF income	61	49	19.67	None
w2_comp	Workmen's compensation	5	5	0.00	None
w2_dis	Disability grant	599	592	1.17	Rule
w2_chld	Child support grant	3446	3439	0.20	Rule
w2_fost	Foster care grant	238	230	3.36	Rule
w2_cdep	Care dependency grant	59	58	1.69	Rule
w2_indi	Interest/dividend income	26	23	11.54	None
w2_rnt	Rental income	84	82	2.38	None
w2_remt	Remittances	679	534	21.35	Regression
w2_plot	Subsistence agriculture income	49	33	32.65	None
w2_opro	Own production consumed	122	92	24.59	None
w2_hhimprent	Imputed rental income	5910	3433	41.91	Regression

Income From Subsistence Agriculture

In Wave 1, income from subsistence agriculture was calculated from the household questionnaire. The aggregated value of all crops and/or animals harvested or consumed by the household formed the measure of this income source. In the second Wave, however, we calculated this value from the adult questionnaire. The Wave 2 adult questionnaire included the question “Think about all the produce that you consumed from your own production last month. How much would it cost to buy all of this at the market?”. This question was not asked in Wave 1. The answer to this, plus the answer to “Please estimate how much you earned from [subsistence agricultural activities] during the past 30 days” were summed to provide an individual-level value of agricultural income. Individual incomes were then aggregated up to the household level.

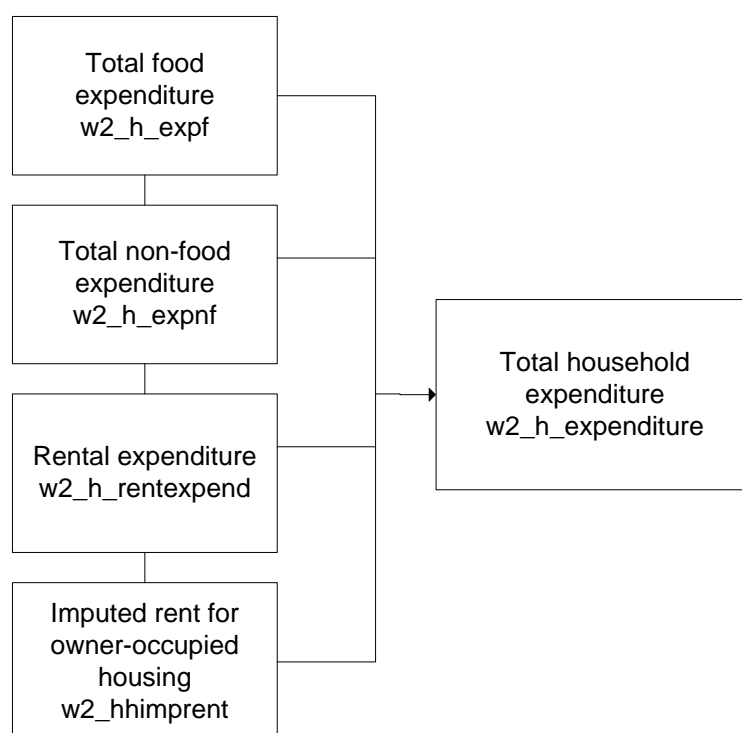
Bonus Payments

In the first Wave, respondents were asked about the value of 13th cheques, profit shares and bonus payments received in the past 12 months. This amount was then divided by 12, to reflect an “average” monthly amount. In the Wave 2 adult questionnaire, respondents were asked about receiving these sources of income in the last 30 days, rather than in the last 12 months. Therefore, in

constructing labour market income for individuals for Wave 2, we did not divide these monthly amounts by 12.

4.7 Expenditure

All expenditure data come from the household questionnaire. The respondent answering the household questionnaire was asked about total household expenditure in the last 30 days for each of 32 food items and 54 non-food items. These were summed to provide total food expenditure (*w2_h_expf*) and total non-food expenditure (*w2_h_expnf*) respectively. These two components were added to total rental expenditure (*w2_h_rentexpend*) and imputed income from owner occupied housing¹⁰ (*w2_hhimprent*) to constitute aggregated total household expenditure (*w2_h_expenditure*).



Imputations

Food

If a respondent indicated that the household purchased one of the 32 food items in the last 30 days, but could not give an expenditure amount, this value was imputed using the same single regression imputation approach as was used in Wave 1. If a household was unable to provide a value for any of the food items, the “one-shot” food expenditure was used, rather than an aggregation over the 32 line items. We maintained the rule-of-thumb that imputation only took place when there were at least 100 recorded observations and missingness did not exceed 40%.

¹⁰ Imputed rental income from owner-occupied housing was added to both income and expenditure in order to avoid underestimating household welfare by selecting one measure of welfare (for example income) over another (expenditure).

Non-food

In this section of the questionnaire, unlike in the food section, there was no leading question for each line item that asked the respondent whether or not the household consumed the item in question. Instead, non-consumption was recorded as a zero, and “don’t know” or “refuse” were recorded as -9 and -8, respectively. Therefore, imputations were done for each line item for which the response was -9 or -8.

There were also some households with a high number of non-food items recorded as negative values. For these households we followed a slightly different imputation process, as it was impossible to distinguish between a legitimate non-response and what was a true expenditure value of zero. For households with missing values for 40 or more of the 54 non-food items, we imputed an aggregate non-food expenditure measure, after obtaining an aggregate measure for all other households.

Rental Expenditure

Missing values for households that rent the dwelling unit that they live in were imputed using a single imputation approach identical to Wave 1 (see Finn, et al, 2009).

Imputed Rental Income for Owner-occupied Housing

This is the same variable that was outlined in the income section of the user document, to which readers are referred.

4.8 Wealth

The first Wave of NIDS allowed us to measure certain components of wealth on an individual and household level, but did not include enough questions to calculate a full measure for either individuals or households. Wealth was one of the special themes of Wave 2 and the questionnaires included components that allowed us to calculate a more refined and accurate measure of household net worth. Given this special status we describe the derivation of household wealth in some detail here.

We define household net worth as household assets less household debts. This concept of household net worth is spread over six different asset types, namely: net financial wealth, net business equity, net real estate equity, the value of vehicles, the total value of pension/retirement annuities and livestock wealth. A broader definition of each of these terms as relevant to NIDS Wave 2 is provided below.

Net financial wealth: the total value of interest-bearing assets held in banks and other institutions, stocks and mutual funds, life insurance funds, trust funds and collectibles minus the total value of unsecured debts (which also includes car loans).

Net business equity: the net value of all business shares owned by all household members.

Net real estate equity: the net value of all properties owned by the household including principal home, holiday and other properties.

Value of vehicles: the total value of all vehicles owned by household members including all transport and recreational (boats/caravan) vehicles.

Pension/retirement annuities: the total amount of pension/retirement capital owned by all household members. The strict definition of these assets requires that they need to be funds in an account that grows without any tax implications until retirement or withdrawal. For example, this could be something like an organisational/company pension plan for the benefit of employees.

Livestock assets: the total value of all livestock in the household's possession at the time of interview.

Wealth in the NIDS Wave 2 Household and Adult Questionnaires

Questions relating to household net worth were asked in both the household and the adult questionnaires. These questions, in addition to other portfolio composition questions, allow us to estimate individual and household net worth.

Wealth is particularly challenging to measure in household interview surveys because of its social sensitivity and the difficulties associated with obtaining accurate estimates of the market value of different asset types (whether physical or financial). Each component of the overall measure of household wealth is provided below and is followed by a flowchart that maps the construction of the total net worth variable.

- Household Questionnaire
 - F2.1 established whether the household would be in debt, breakeven or have something left over if the home and all major possessions were sold, all investments were turned into cash and all debts were paid off.
 - If something would be left over, then F2.2 asked for the Rand value. If respondents refused or did not know, then a series of unfolding brackets from F2.3.1 to F2.3.5 kicked in.
 - If the household would be in debt, F2.4 asked for the Rand value of that debt. Once again, if the respondent refused or did not know, a series of unfolding brackets kicked in from F2.5.1 to F2.5.5.
 - H8.4.1 to H8.4.7 asks about the value of livestock in the household's possession, over seven categories of animals.

The household questionnaire also contained questions about the market value of all properties owned by members of the household, as well as the outstanding amount owing on bonds attached to these properties.

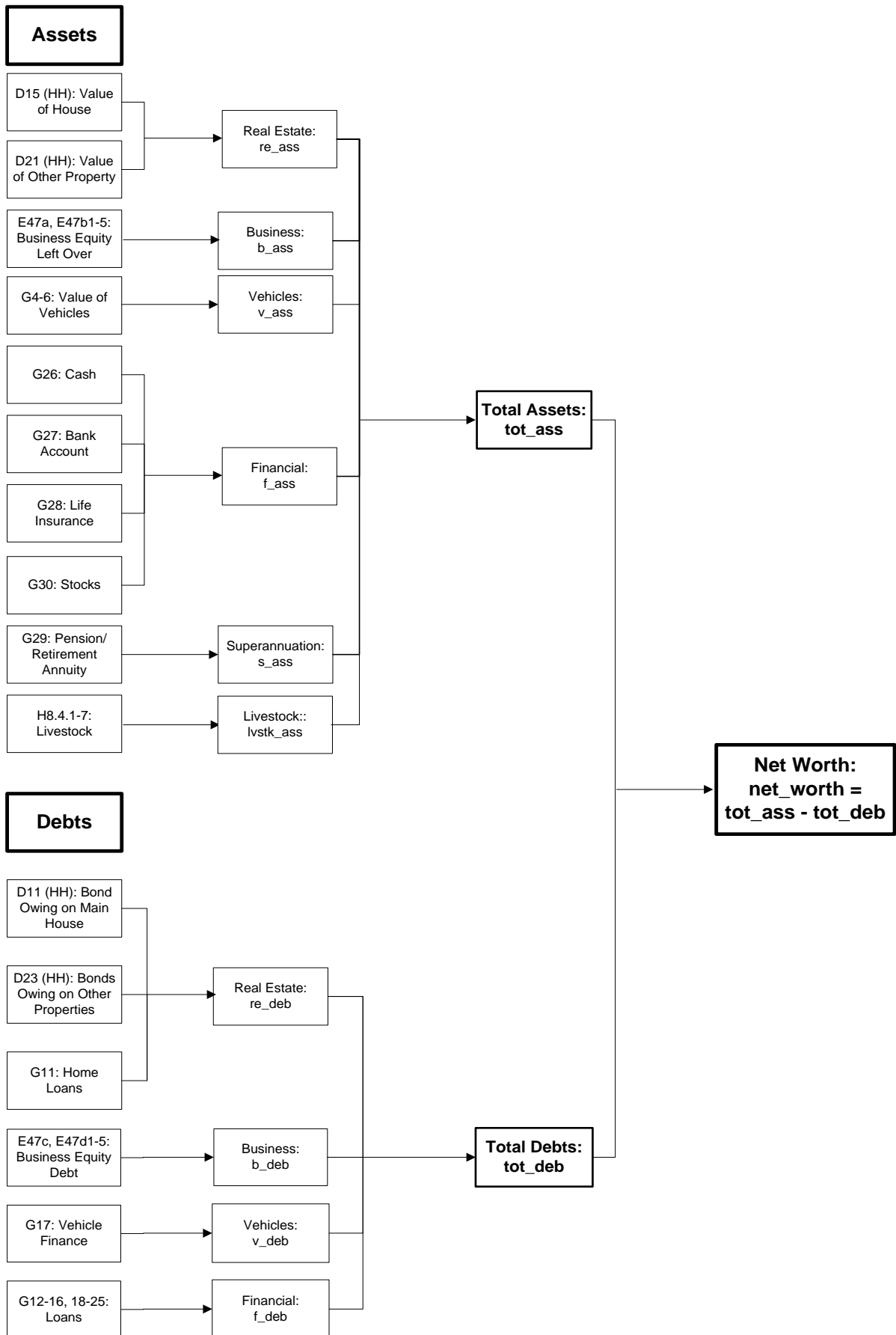
- D11 asked for the amount of bond still owing on the property if it is owned by a member of the household.
- D15 asked for a reasonable value for which the household could be sold.
- D21 asked about a reasonable market value for which all other properties owned by the household could be sold for.

- D23 asked about the total value of bonds that were still owing on all other properties owned by resident household members.
- Adult Questionnaire
 - E46 established whether the respondent would be in debt, breakeven or have something left over if all business assets and investments were turned into cash and all debts were paid off.
 - E47a (point estimate) and E47b1-5 (unfolding brackets) asked about how much money would be left over.
 - E47c (point estimate) and E47d1-5 (unfolding brackets) asked about how much debt would be left over.
 - G4, G5 and G6 asked about the value of all motor vehicles, bakkies/trucks and motorbikes owned by the respondent.
 - G11 asked about home loans/bonds.
 - G12 to G16 and G18 to G27 asked about a variety of other assets and debts, such as personal bank loans, store cards and study loans.
 - G17 asked about vehicle finance.
 - G28 and G30 asked about life insurance and unit trusts/stocks/shares respectively.
 - G29 asked about pensions/retirement annuities.

Imputation

Where a household acknowledged an asset or a debt, but was unable to provide a value, we imputed using a single equation imputation regression approach. As usual, our rule-of-thumb required the number of reported observations to be 100 or more, and for missingness to be at 40% or below.

The diagram below outlines how the final net worth for each household was calculated.



Aggregating Household Net Worth and Including One-Shot Measures Where Appropriate

The quality of the aggregated measure of household net worth is superior if we can add up the various debts and losses reported by all adults in the household. However, in some cases, this was impossible because of non-response (both item and partial-unit). The rule used in this case was consistent with what took place in the aggregation of Wave 2 income. If wealth was missing for an individual in a household (item non-response for the each question in the section or unit non-response for the individual), then we used the one-shot measure for household net worth. If an individual's wealth section was missing and the household one-shot question was also missing, then we used whatever we could from the remaining adults in the household. Finally, if all adults recorded non-responses to wealth question and the household one-shot was also non-response, household net worth was set to missing.

4.9 Anthropometric Z-Scores

For children up to the age of 5 years z-scores for height for age, weight for age, weight for height and body mass index (BMI) for age were calculated using the World Health Organisation (WHO) international child growth standards as the reference (WHO 2006). For individuals older than 5 years the WHO growth standards for school-aged children and adolescents (de Onis et al. 2007) were used as a reference in the calculation of z-scores for height for age, BMI for age and weight for age.

The following variables were created:

- w1_zhfa and w2_zhfa - height for age for individuals up to 19 years of age
- w1_zwfa and w2_zhfa - weight for age for individuals up to 10 years of age
- w1_zwfh and w2_zhfa - weight for height for individuals up to 5 years of age
- w1_zbmi and w2_zhfa - BMI for age for individuals up to 19 years of age

Using the WHO guidelines we set biologically implausible z-scores to missing as follows:

- zhfa<-6 or zhfa >6
- zwfa<-6 or zwfa>6
- zwfh<-5 or zwfh>5
- zbmi<-5 or zbmi>5

In calculating the weight for height z-scores, we assumed that the child was measured in the recumbent position if the child's age is below 24 months (731 days). If the child is aged 24 months or above, we assumed that the measured height is standing height.

NIDS fieldworkers were instructed to take two height measures and then a third if the first two measures were more than one centimetre apart. Similarly, a third weight measure was required if the first two weight measures were more than one kilogram apart. In practice, the third measures were very seldom taken. For calculating z-scores, we used the average of the first two measures. In

instances where these first two measures differed by more than one centimetre in the case of height and one kilogram in the case of weight, we used the third measure if it was available.

4.10 Weights

The section below was written for Version 1.0 of Wave 2 and Version 4.1 of Wave 3. See the Wave 3 User Manual for further discussion of the weights calculation for Version 2.0 of Wave 2 and Version 5.0 of Wave 1.

Three sets of weights have been released together with the second Wave:

- a) An update to the weights to be used with the Wave 1 dataset
- b) Weights to be used for panel analyses of the changes between Wave 1 and Wave 2
- c) Cross-sectional weights for analysing the Wave 2 data as a cross-section of the South African population

Updated Wave 1 Weights

It was necessary to update the Wave 1 weights for two reasons. Firstly work subsequent to the initial release of the NIDS data led to three households being removed from the dataset. The weights were, however, not updated at the time of version 3 being released. Another household has been deleted with the release of Wave 1 version 4.1. Secondly in the process of creating weights for the second Wave it became clear that the age-gender-race classification used for calibrating was overly detailed for the Indian subsample. The issue can be seen in the following table of cell counts for Wave 1:

Age group	Indian male	Indian female
60-64	7	11
65-69	4	8
70-74	4	4
75-79	1	5
80+	1	1

Such small cell sizes imply that the constraints on the calibration are very tight. In the case of Wave 2 some of these cells ended up empty, making the calibration unfeasible. In any case it seems undesirable to calibrate on such small numbers. Consequently for the Indian subsample the top three age brackets were collapsed. In all other respects the rereleased Wave 1 weights were calculated according to the procedures laid out in the Wave 1 technical documentation.

The rereleased weights are in practically all cases very close to the original weights. The “design weights” versions of the weights show a correlation coefficient of .9999 with the original weights and in all cases the maximum deviation is only 6.7%. The new calibrated weights have a correlation

coefficient of 0.9876 with the originally released calibrated weights. Virtually all weights (ranging from the 5th to the 95th percentile) lie within 15% of the original weights. There are some isolated cases where the final weights differ by larger amounts.

In short virtually every analysis done with the old weights should look the same with the new ones. If the results are different, this is an indication that one or other individual in the old Indian subsample is having a disproportionate impact on the outcome. Such analyses should be treated with suspicion.

Panel Weights

The individuals who were successfully reinterviewed in the second Wave of NIDS are not a random subset of all the individuals surveyed in the first Wave. The panel weights are intended to correct for this attrition bias. The probability of being successfully interviewed in Wave 2 given the Wave 1 characteristics of the individual was estimated using a probit model. The explanatory variables used in this regression were race-gender specific quartics in age, dummies for provincial location, marital status and educational attainment. The reason for using age quartics rather than age dummies is to allow the probability to vary smoothly with age, which given the nature of age related mortality is more appropriate.

One of the regrettable features of the pattern of attrition is that particular categories of individuals who had a relatively lower probability of being interviewed in Wave 1 also showed much higher rates of attrition. In the table below we record the probability of being successfully interviewed, according to the probit model. It is evident that Whites, particularly in their twenties and late in life had much lower probabilities of being reinterviewed than their African counterparts.

Age Group	Male				Female			
	African	Coloured	Indian	White	African	Coloured	Indian	White
-9	0.063				0.069			
-3					0.067			
0-1	0.854	0.797	0.881	0.772	0.857	0.803		0.836
1-4	0.852	0.777	0.781	0.707	0.841	0.773	0.808	0.659
5-9	0.882	0.800	0.649	0.628	0.869	0.787	0.739	0.588
9-14	0.863	0.764	0.523	0.441	0.849	0.768	0.637	0.410
15-19	0.830	0.723	0.497	0.350	0.824	0.738	0.549	0.327
20-24	0.784	0.677	0.474	0.304	0.802	0.720	0.554	0.319
25-29	0.749	0.647	0.515	0.381	0.801	0.718	0.577	0.366
30-34	0.728	0.654	0.599	0.359	0.808	0.739	0.600	0.425
35-39	0.719	0.667	0.635	0.430	0.818	0.762	0.626	0.479
40-44	0.718	0.685	0.662	0.470	0.835	0.767	0.665	0.488
45-49	0.739	0.689	0.725	0.509	0.845	0.781	0.705	0.517
50-54	0.745	0.717	0.719	0.504	0.853	0.788	0.713	0.542
55-59	0.759	0.727	0.670	0.565	0.863	0.770	0.681	0.551
60-64	0.771	0.704	0.620	0.553	0.862	0.755	0.628	0.547
65-69	0.786	0.703	0.448	0.518	0.857	0.749	0.593	0.499

70-74	0.781	0.665	0.490	0.844	0.709	0.440	0.428
75-79	0.742	0.568	0.300	0.466	0.811	0.645	0.412
80-84	0.675	0.512	0.380	0.775	0.593		0.369
85+	0.467	0.195	0.361	0.662	0.501		0.421

The panel weights are the inverse of the probability of appearing in the sample. This probability is the product of the probability of being interviewed in Wave 1, times the probability of being successfully reinterviewed in Wave 2, conditional on appearing in Wave 1. The panel weights are therefore the product of two weights: the weight corresponding to appearing in Wave 1 (as represented by the calibrated weight) and an attrition weight, i.e. the inverse of the conditional probability of being reinterviewed.

Given that some individuals with a high weight in Wave 1 also carried a high attrition weight, this led to some extreme weights. Provided that end users are sufficiently cautious in working with the weights there would have been nothing intrinsically wrong with releasing such weights. Our experience, however, has been that the bulk of users are not experienced when it comes to working with weights. In order to prevent avoidable errors we decided to trim the weights to the 1st and 99th percentiles of the weight distribution.

Cross-Sectional Weights for Wave 2

The individuals interviewed in Wave 2 included both household members in the original sample (CSMs) as well as some new individuals who are now coresident with them (new birth CSMs or TSMs). The theory for how to weight such cases is discussed by Rendtel and Harms (2009) and Deville and Lavallée (2006). In brief, the idea is that individuals who were part of the original universe covered by the Wave 1 sample (but did not get sampled themselves) get allocated a share of the sampling weight attached to the individuals with whom they are now coresident. The most straightforward procedure (used to calculate the NIDS cross-sectional weights) is to average out sample weights within the Wave 2 households, assigning TSMs a weight of zero.

The case of newborn CSMs has to be tackled differently. They are a subpopulation that was not part of the original frame. If households did not get reshuffled they should get the same weight as other members of their household and the overall increase in the sum of the weights would give an unbiased estimate of the total population increase. Given the NIDS definition of which newborns are CSMs, they should be thought of as indirectly sampled through their mothers, i.e. their mothers weight should be assigned to the newborn CSMs.

The Wave 1 household weights that were used as inputs for the “generalised share method” were the design weights corrected for non-response (i.e. *w1_dwgt* in Wave 1 version 4.1). The resultant Wave 2 weight (*w2_dwgt*) should be thought of as design weights corrected for non-response and for the reshuffling of household membership. Theoretically, use of these weights should give unbiased estimates of the population defined by the sampling rules, i.e. individuals who could have been sampled in Wave 1 and individuals who come to be coresident with individuals who could have been sampled in Wave 1. Two categories of individuals are excluded: immigrants who form their own

separate households and people who emigrated and who therefore no longer form part of the South African population.

The *w2_dwgt* weights were then calibrated to the mid-2010 population estimates, released by Statistics South Africa. The calibration occurred to sex-race-age group cell totals (with the oldest three age categories for Indian males and Indian females collapsed, as noted before) and provincial totals. The calibration was again done using the Stata ***maxentropy*** add-in (Wittenberg 2010). Again individuals within the same second Wave household were constrained to get the same weight. The resultant weights are contained in the variable *w2_wgt*.

5. Further Resources

5.1 Program Library

Stata syntax files (do-files) compressed into Zip format can be found on the Data First website:

<http://www.datafirst.uct.ac.za/dataportal/index.php/catalog/central/about>

They can also be found on the NIDS website:

<http://www.nids.uct.ac.za/nids-data/program-library>

There are generally two kinds of coding files that we provide: (1) those that assist with data manipulation of the panel, and (2) those that give insight into derived variables.

Data manipulation

Merging datasets

It should be noted that, in general, merges to the household roster and across Waves should always be done on both *hhid* and *pid*, the combination of which is unique.

Within Wave merging

[Program 1 - Merging the Adult, Child and Proxy datasets to the Household Roster](#)

[Program 2 - Merging the Household questionnaire to the individual datasets](#)

Across Wave 1 and Wave 2 merging

[Program 3 - Merging Individuals across Wave 1 and Wave 2](#)

[Program 4 - Merging Households across Wave 1 and Wave 2](#)

Reshaping data

[Program 5 - Reshaping the birth history and merging in child questionnaires](#)

[Program 6 - Reshaping the mortality section](#)

Derived variables

Income

As explained above in section 4.6, NIDS has constructed a derived variable as a measure of total regular household income received in the 30 days prior to the interview taking place.

[Program 7 - Master Income do file](#)

[Program 7.1 - Income - Merging datasets to create income variables](#)

[Program 7.2 - Income - Preparing variables for imputation](#)

[Program 7.3 - Income - Performing Imputations for missing data on Income variables](#)

[Program 7.4 - Income - Aggregation of pre-imputation variables](#)

[Program 7.5 - Income - Aggregation of post-imputation variables](#)

[Program 7.6 - Income - Variables for public release](#)

Expenditure

As explained above in section 4.7, NIDS constructed a derived variable as a measure of total household expenditure in the 30 days preceding the interview taking place.

[Program 8 - Expenditure - Master expenditure do file](#)

[Program 8.1 - Expenditure - Merging datasets to create expenditure variables](#)

[Program 8.2 - Expenditure - Preparing variables for imputation](#)

[Program 8.3 - Expenditure - Performing Imputations on Expenditure variables](#)

[Program 8.4 - Expenditure - Aggregation of imputation variables](#)

[Program 8.5 - Expenditure - Variables for public release](#)

Wealth

As explained above in section 4.8, NIDS constructed a variable as a measure of household net worth by subtracting total household debts from total household assets.

[Program 9 - Wealth - Master wealth do file](#)

[Program 9.1 - Wealth - Merging datasets to create wealth variables](#)

[Program 9.2 - Wealth - Preparing variables for imputation](#)

[Program 9.3 - Wealth - Performing Imputations on Wealth variables](#)

[Program 9.4 - Wealth - Aggregation of imputation variables](#)

[Program 9.5 - Wealth - Variables for public release](#)

Deflator

Because fieldwork for Wave 2 took place over a period that lasted more than twelve months, all the financial data needs to be deflated before it can be analysed.

[Program 10 – Deflator](#)

Employment Status

NIDS constructed a derived variable using the International Labour Organization definitions to assign respondents to one of the following categories - Employed, Unemployed (strict definition), Unemployed (broad definition) and Not Economically Active.

[Program 11 – Employment Status](#)

6. References

- Cowell, F.A., 2000, "Measurement of inequality", in Atkinson, A.B. and Bourguignon, F. (eds), *Handbook of income distribution, Volume One*, New York: Elsevier
- de Onis, M., A. Onyango, E. Borghi, A. Siyam, C. Nishida and J. Siekmann "Development of a WHO growth reference for school-aged children and adolescents." *Bulletin of the World Health Organization* 85: 661-668.
- Deville, Jean-Claude and Lavallée (2006), "Indirect sampling: The Foundations of the Generalized Weight Share Method", *Survey Methodology*, 32(2): 165-176
- Finn, A., Franklin, S., Keswell, M., Leibbrandt, M. & Levinsohn, J., 2009, "Expenditure: Report on NIDS Wave 1", Technical Paper no. 4, Cape Town: National Income Dynamics Study
- Rendtel, Ulrich and Harms, Torsten (2009), "Weighting and Calibration for Household Panels", in P. Lynn (ed) *Methodology of Longitudinal Surveys*, Wiley, Chapter 15. Working paper version available at <http://www.iser.essex.ac.uk/files/survey/ulsc/methodological-research/mols-2006/scientific-social-program/papers/Rendtel.pdf>
- Wittenberg, Martin (2010), "An introduction to maximum entropy and minimum cross-entropy estimation using Stata", *Stata Journal*, 10(3):315-330.
- Wittenberg, M., 2011, "Fat tales of South Africa's income distribution", Mimeo, Cape Town: University of Cape Town
- World Health Organization (2006) WHO Child Growth Standards: Length/height-for-age, weight-for-age, weight-for-length, weight-for-height and body mass index-for-age: Methods and development. Geneva: World Health Organization