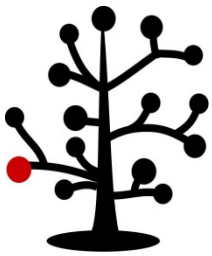


N.i.D.S.

NATIONAL INCOME DYNAMICS STUDY

National Income Dynamics Study
Wave 3 User Manual

Edited by Louise de Villiers, Michael Brown, Ingrid Woolard, Reza Daniels and Murray Leibbrandt

Contents

List of Contributors

This document was created by the NIDS team. For the correct citation method, see section 1.2 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

Poverty and inequality are complex issues that have many causes and possible solutions. The paths out of poverty are manifold. Panel surveys have been used the world over to study the dynamics that influence households' and individuals' ability to move out of poverty. Panel surveys also give the opportunity to study respondents as they move through life-phase transitions, such as school to work. They provide the data to determine the factors that facilitate or impede progress through education and transitions into the labour market.

NIDS as a panel is maturing, with three waves of data available for the analysis of South Africa's current social dynamics. The process began in 2005 with the realisation that South Africa would benefit from a longitudinal survey that would shed light on the realities of living in South Africa. In 2006 SALDRU was appointed as the implementation agency of this important project. Wave 1 was implemented in 2008, Wave 2 followed in 2010/2011. During this wave the use of Computer Assisted Personal Interviewing (CAPI) was introduced. Wave 3 was collected during 2012 and built very successfully on the lessons learned in Wave 2; indeed, Wave 3 had a negative attrition rate when compared with Wave 2.

Important improvements were made to the quality of the data for each of the three waves during the production of Wave 3. The linking of respondents across waves and across households was especially improved. This, together with the longer period covered by the panel, creates important and exciting research opportunities.

This manual serves as an overview to help users understand the methodology employed to collect the data and some of the technicalities regarding the more complex aspects of the data.

The use of NIDS by national and international researchers for policy analysis is growing. There is important work on the impact of social grants, on progress through school, on child poverty and many other important issues. This is the reason why government invested in NIDS and we encourage users to use the new Wave 3 data together with the latest releases of Waves 1 and 2.

We wish you all the best with your research.

NIDS Team

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 document accompanies the release of the Wave 3 data. As with any new wave data release there have been updates to the data of previous waves. Please refer to the latest documentation for previous waves if merging to this dataset. 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 2.2.67 on merging datasets within and between Waves.
- The fieldwork schedule: see section 3.3.
- Non-response and attrition in Wave 3: see section 3.4.
- Updated weights for Wave 1 and Wave 2 and new weights for Wave 3, see section 4.9.
- Links to examples of how to correctly merge NIDS data using Stata: see section 5.1.
- Links to deflate the financial data: see section 5.1

1.2 Citation Of NIDS Data And Documentation

Users wishing to cite the Wave 3 data should use the following reference:

Data Citation:

Southern Africa Labour and Development Research Unit. National Income Dynamics Study 2012, Wave 3 [dataset]. Version 1.2. Cape Town: Southern Africa Labour and Development Research Unit [producer], 2013. Cape Town: DataFirst [distributor], 2013

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

Documentation Citation:

De Villiers, L., Brown, M., Woolard, I., Daniels, R.C., & Leibbrandt, M, eds. 2013, "National Income Dynamics Study Wave 3 User Manual", Cape Town: Southern Africa Labour and Development Research Unit

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). 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 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 website.

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, 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.

2.2 Data Formats

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

2.3 Data Structure

Every resident[1] individual (CSM[2] or TSM[3]) 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 (*w3_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 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 the details of their 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

[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
[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.

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 all waves.

2.4 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 = 41,307).

HHQuestionnaire: One record per household with data from the household questionnaire, excluding the household roster. Unique identifier: *w3_hhid* (n= 10,236).

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: *w3_hhid* (n = 10,236), non-unique identifier for individual: *pid* (n= 42,230). The combination of *w3_hhid* and *pid* is unique per person within each wave.

Adult: One record per entry from the Adult[6] questionnaire. Unique identifier for household: *w3_hhid* (n=9,983), unique identifier for individual: *pid* (n=22,481); 3,771 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. The non-response records have a value greater than one in the *w3_a_outcome* variable. Polygamists in the sample appear only once in the adult file. This is in the household in which their individual interview was conducted.

Proxy: One record per entry from the Proxy[7] questionnaire. Unique identifier for household: *w3_hhid* (n=2,071), unique identifier for individual: *pid* (n=2,720).

Child: One record per entry from the Child questionnaire. Unique identifier for household: *w3_hhid* (n=5,615), unique identifier for individual: *pid* (n=12,235); 1,028 observations have no data beyond Section A as these individuals refused to

[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 2 individuals who are 14 years old in the Adult file and 62 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.

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 the *w3_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: *w3_hhid* (n=10,236). 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: *w3_hhid* (n=10,130), unique identifier for individual: *pid* (n=37,436).

See section 4 - Derived Variables and section 5 - Program Library for more information.

2.5 Identifiers

Individuals can be identified across wave by their unique identifier *pid.* Households are identifiable within wave by their unique identifier *wx_hhid*. Different household identifiers are assigned each wave as NIDS is a panel of individuals, and the household identifier is simply a tool to connect each individual to their household within 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.

2.6 Merging Datasets Within & Between Waves

Since the release of Wave 2 the longitudinal dimension of NIDS can be explored and with the Wave 3 release new opportunities open up. 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 and any babies born to CSM females after 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. Links to examples of the Stata code to merge within and between waves are provided below in Section 5 - Program Library.

From 2013 releases onwards, non-resident household members on the Wave 1 roster have also been assigned *pid's*. Previously they were system missing on that variable. This means that where users previously dropped those with missing *pid* to identify Wave 1 CSMs in the Wave 1 Household Roster file, they will now have to use the *w1_r_pres* and *w1_r_csm* variables to identify original CSMs and to identify where they are resident in Wave 1. The residency criteria are important as there are 3 identified polygamists in the wave 1 dataset. Now that we know that these are the same individuals they have been assigned the same *pid* in both households. They are, however, only resident in one household.

The same principle is carried in subsequent 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 edit the data to ensure that he/she is recorded as "resident" only in the household where their individual interview was conducted. He/she is marked as non-resident in all other households. In the unlikely event that a person had an individual questionnaire completed in more than one household, we will randomly assign him/her as resident in only one household. In summary, 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.

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.

2.6.1 Merging within Wave 3

We recommend that the merging within wave should be done using w3_*hhid* and *pid*. The exception to the rule would be when specifically looking for people who are resident in more than one household. The roster is the only file where merging with *pid* only will yield different results to merging on *pid* and w3_*hhid*.

2.6.2 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 on the roster 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 deletes all records for non-residents from the roster file. 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 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 all waves (*w1_hhid, w2_hhid, w3_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 all 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 files 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.

2.7 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_, w2_, w3_

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:

2.7.1 Wave

The wave prefix indicates in which wave the data was collected.

Wave indicator	***Meaning***
w1	Wave 1
w2	Wave 2
w3	Wave 3

2.7.2 Source

The source indicates which dataset the variable belongs to.

Source indicator	***Meaning***
A	Adult file
C	Child file
P	Proxy file
H	Household file
R	Household roster file

2.7.3 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:

Section Leader	**Meaning**	**Section Leader**	**Meaning**
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

2.7.4 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	em1	Self-employment	ems
Secondary employment	em2	Casual employment	emc

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	hllf
Recent consultations	hlcon	Smoker	hllfsmk
Vision	hlvis	Difficulty of activities	hldif

2.7.5 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.

2.7.6 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".

2.8 Non-Response Codes

Non-response codes are usually indicated by negative numbers. The only exception is dates where four digits are 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.

2.9 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).

2.10 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 Centre 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.

2.11 Program Library

NIDS makes several Stata Programs 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 of this User Guide for a detailed list of these files.

3. Data Collection

Wave 3 saw an extension of the Computer Assisted Personal Interviewing (CAPI) and in-house systems. Every effort has been made to be consistent in the methodology applied across waves, while also paying attention to being more efficient in field operations. Increased use of paradata on interviewer performance was made to improve the quality of data collected and so reduce interviewer effects. This section first describes the field processes followed and then gives more detail on the increased monitoring of fieldworker behaviour during field operations and other quality control measures taken.

3.1 Data Collection Process

As in previous waves, four types of questionnaires were administered:

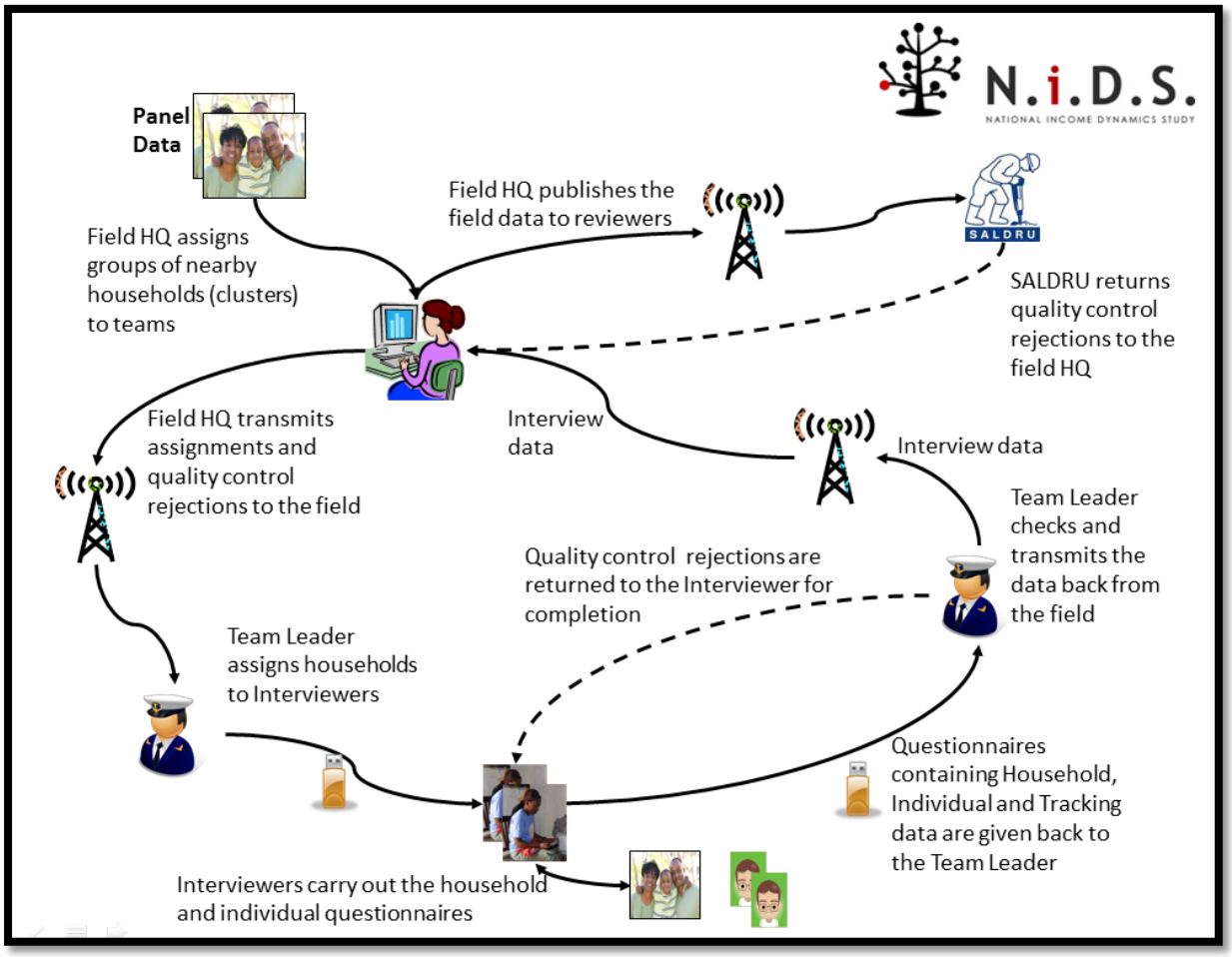
- **Household questionnaire**: One household questionnaire was completed per household by the oldest woman in the household or another person knowledgeable about household affairs and particularly household spending. Household questionnaires took approximately 39 minutes in non-agricultural households and 50 minutes in agricultural households to complete.
- **Adult questionnaire**: The Adult questionnaire was applied to all present Continuing Sample Members and other household members resident in their households that are aged 15 years or over. This questionnaire took an average of 38 minutes per adult to complete.
- **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 12 minutes to complete. 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 16 minutes per child to complete.

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 so that 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 dataset.

3.1.1 Overview of CAPI cycle

The CAPI cycle is illustrated below. This is almost the same cycle as applied in Wave 2.

Figure 1: The CAPI Cycle



Listing data (PSUs, household addresses, contact details, roster make up and individual contact details) drawn predominantly from Wave 2 were pre-loaded into the CAPI system. Some respondents who were not located during Wave 2 were listed with their Wave 1 information in order to allow fieldworkers to reattempt to gather information about them from the area or household where we last observed them. This process allowed a number of CSMs to re-enter the sample when they would have been lost due to insufficient information collected during Wave 2. Listing data was centrally distributed via modems to field teams on a cluster by cluster basis prior to their arrival.

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). 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 pre-loaded information was sometimes not displayed, but was used by the CAPI system to challenge inconsistent answers (e.g. attendance at school during

Wave 2). Where Wave 3 answers were inconsistent with data previously collected, the interviewer was challenged to confirm the answer and enter substantiating notes for the change.

Certain pre-populated data were used to skip questions if valid and consistent answers had been discovered in Wave 1 and Wave 2, an example being head circumference of a child at birth.

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.

3.1.2 Overview of the tracking process

An essential part of the panel aspect of the survey is to track CSMs as they move within the borders of South Africa. CSMs could either be in the same location as they were in Wave 2 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:

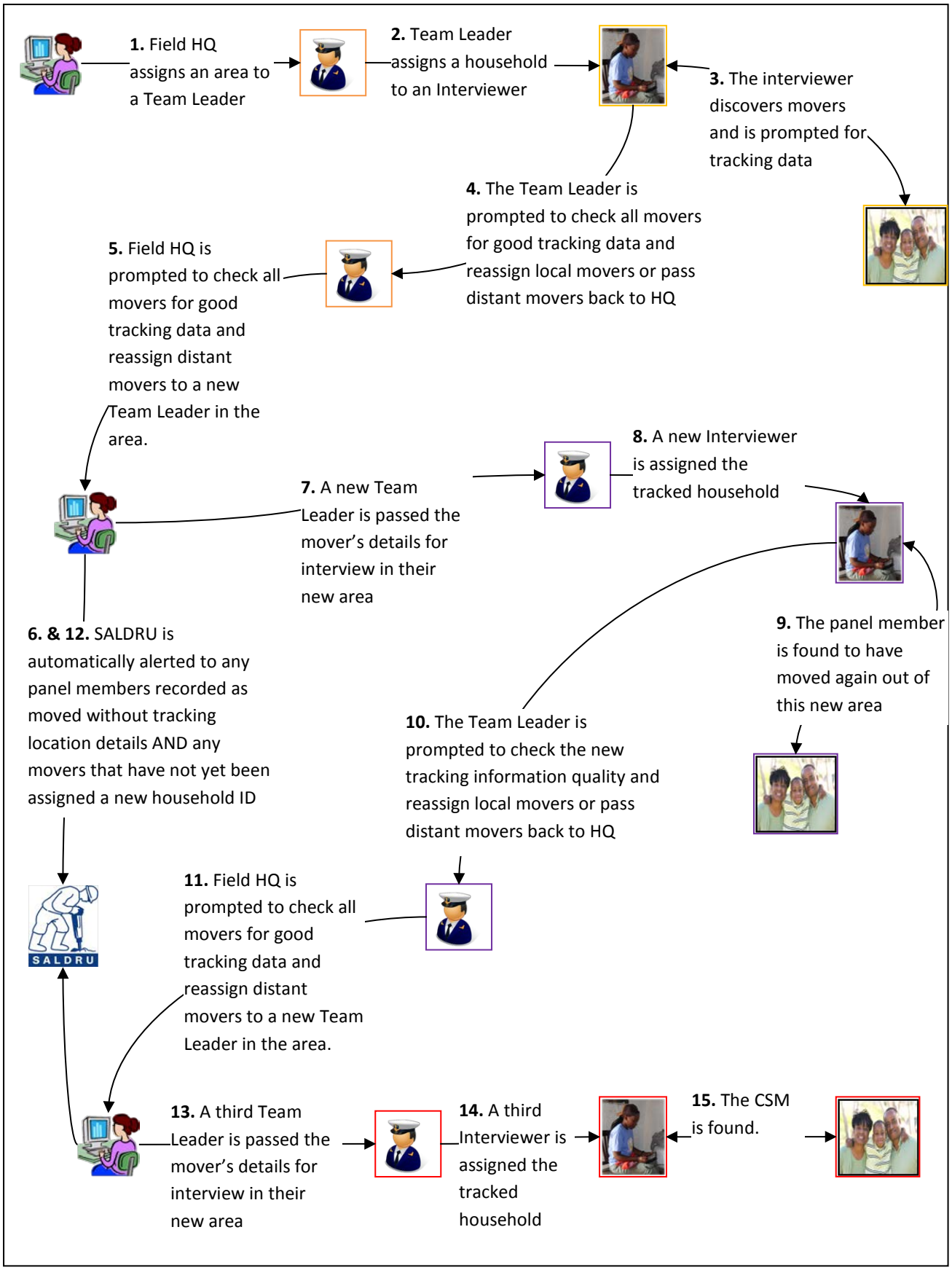
1. Generate new household IDs locally containing the movers to be dealt with by that team; or
2. Transmit the location details back to field control to generate household identifiers for movers and assign them to the relevant team on a geographical level.

Households were created around these location details which were indexed and linked to respondents. A household ID was generated for each location with new CSM records linked to that household ID for all CSMs identified as having moved to that location. 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. By making use of the extensive family networks now represented in the Panel Maintenance System the SALDRU office team was often able to locate respondents and in this way help improve the response rate of the field team.

The process is illustrated in the following diagram:

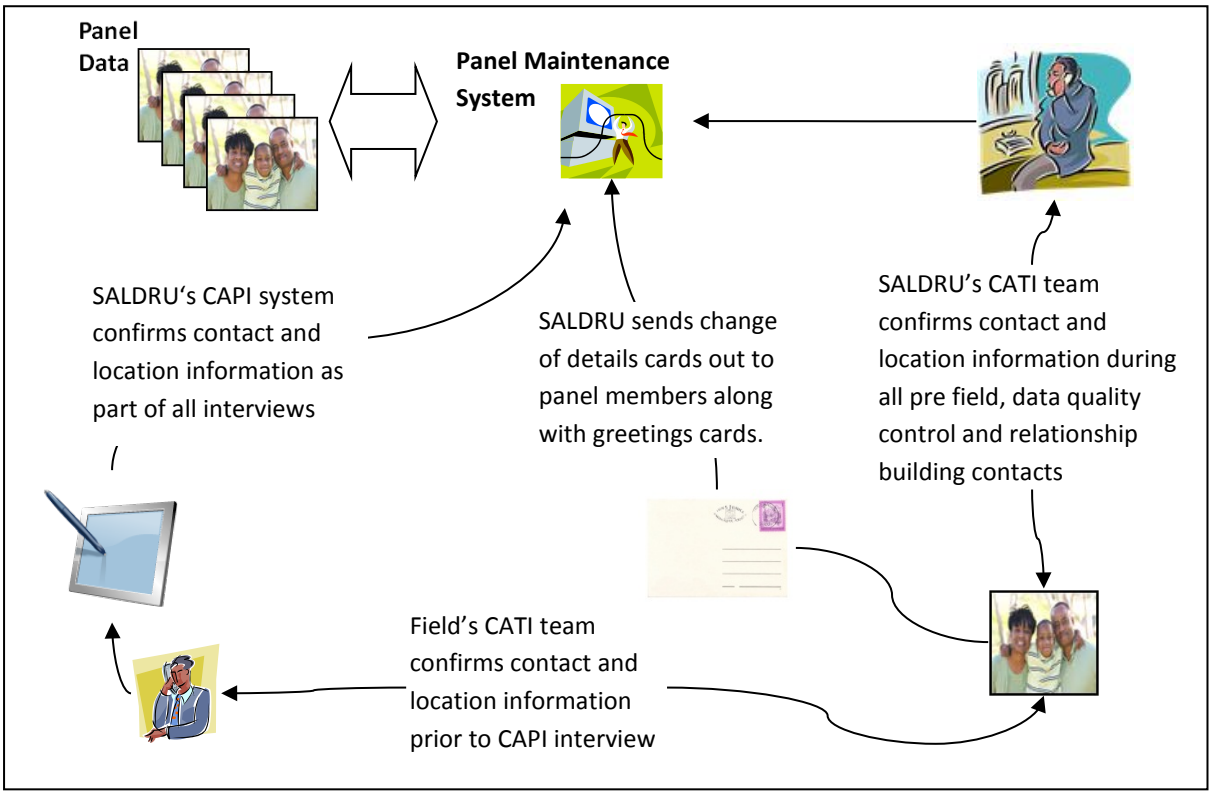
Figure 2: Tracking movers



3.1.3 Contacting respondents

A Panel Maintenance System integrated into a Computer Assisted Telephonic Interviewing (CATI) Call-Centre 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:

Figure 3: 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).

3.2 Data Quality Issues And Data Collection

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

3.2.1 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. **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 address and telephone details taken for movers. This was also done in an effort to minimise non-contact.
3. **New field status for temporarily away respondents:** Wave 3 added a new status for households, that of "temporarily away". This caught instances where no one was at a dwelling but it was discovered that they would return within the fieldwork period (but not while the team was currently in the relevant cluster). These dwellings would then be revisited later in the fieldwork period to "catch" the respondents at a later date. In Wave 2 these respondents would have been missed and recorded as "no one at home" after the mandated three attempts on differing days and times when the field team was in that cluster. The result is that more temporarily absent respondents were interviewed in Wave 3 than in Wave 2 and the number of "no one at home" respondents in Wave 3 contains a smaller proportion of these respondents than is the case for Wave 2.
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:** Field company bonus schemes and targets were restructured in Wave 3 to encourage better completion and lower attrition during fieldwork. Wave 3 saw negative attrition (see the attrition section of this document); however no claim is made for any causal link to field contract structures.
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.

3.2.2 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 second wave with CAPI, a stricter policy was in place than in previous waves and data was accepted from field only if all sections had been completed. A system for accepting exceptions was created, but each exception had to be approved by SALDRU staff. Any questionnaires submitted that were not completed correctly and which did not have an exception raised were returned to field for completion.

3.2.3 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 waves. 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. **Use of paradata on interviewer performance:** In order to improve the quality of data collected, certain key indicators were closely monitored during field. This would also reduce the interviewer effects. The following areas were examined, by interviewer:
    - Questionnaire duration
    - Numbers of non-resident roster members added
    - Refusal rates achieved by interviewer
    - Magnitude of anthropometric measurement differences between current waves and previous waves, as well as flags for extreme BMI measures
    - Individual questionnaires reporting subsistence agriculture, but households not reporting agriculture

- Item level non-response.

These checks were taken periodically from mid-August (approximately halfway through fieldwork). Where interviewers' performance measures lay outside of ±50% of mean they were investigated, retrained, moved to differing teams for closer supervision or removed; in some cases the households were re-interviewed to include hitherto missed respondents. The nature of the measures used and their commencement from August may therefore need to be considered when addressing issues of interviewer effect.

4. **Within wave and across wave consistency checks in office:** SALDRU carried out a range of pattern searches and consistency checks on the date during field to identify interviewer effects and possible miscapture. When areas of concern were found, the respondents / households were contacted to ensure that the data was correct. If a call-back was successful the data collected during the call-back were used to correct the information collected infield. If the query was across wave it could result in a change of data for a previous wave. If the call was unsuccessful the conflicting information was left 'as is' in the data. A number of key variables (sex, race, age, education, mother and father) have "best" variables created for them in the *indderived* file to indicate what the best estimate of the variable is given the information collected across the waves. Less than 1% of respondents have unresolved conflicts.
5. **Live behavioural correction:** The use of CAPI 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.

3.2.4 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.

3.2.4.1 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.
- A new feature in Wave 3 was CSMs who had split from their Wave 1 household in Wave 2, returning to the Wave 1 household. In the CAPI system a new record would have been created for the returned CSMs. Through careful identification of likeness within household

dynasties such cases could be identified. Sometimes the identification took place before the fieldwork company attempted to track the original CSM and they could be informed that it was no longer necessary to track that respondent.

- Conversely, there was the need to identify people who were incorrectly identified as a CSM when in fact the wrong person was interviewed. Where these cases were identified during field they were returned to the fieldwork company to attempt to interview the right person.

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 (Date of birth, 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.

3.2.4.2 Returning incorrect data "Back To Field"

New controls in Wave 3 included a "status" visible on the CAPI systems used by interviewers and through all management layers. This status system transferred a large proportion of the Wave 2 SALDRU quality control office checks to the CAPI system itself. This meant that in Wave 3 new and more sophisticated checks could be carried out by the SALDRU quality control office which could result in a questionnaire being rejected (see above section).

The Wave 3 CAPI status system would automatically reject questionnaires where:

- Not all individuals in the household were attempted.
- No GPS coordinates were collected for households successfully interviewed or households found but with valid non-response outcome[8].
- 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.
- Validations not having been run.
- Validation errors having occurred.
- The questionnaire does not have a final outcome (e.g. "complete", "now refusing" etc.)

Having met these criteria, SALDRU would then check for other invalidities:

- Incorrect person interviewed.
- Aberrant field behaviour (for example clear evidence of invention of data, unfeasible numbers of proxies rather than direct interviews etc.).
- Non-receipt of the paper consent form.
- Mismatches between household rosters and individual birth histories.
- Unlisted household members identified through follow up calls.
- Invalid non response

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

"Invalid non-response" was where 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. 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.

If a questionnaire was deemed invalid by SALDRU's data quality checks , it was marked as rejected in the CAPI systems and therefore sent "back to field" and a further in-person interview was required (i.e. telephonic interviews were also not permitted in resolving "back to field" issues). SALDRU and the field company met twice a week to review any outstanding "back to fields".

3.3 Fieldwork Schedule

3.3.1 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 the pre-test in Wave 1 and Wave 2, all aspects of the panel and pre-population could be tested. The pre-test tracks 586 individuals from 160 households. 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 3 pre-test fieldworker training and fieldwork was conducted in February 2012.

3.3.2 Main data collection

In contrast to previous waves, all fieldworker training was conducted at the same time. This allowed training to be very consistent and for the fieldwork to start in earnest as soon as possible. In total there were 136 fieldworkers who operated in teams of 4 comprised of 1 team leader and 3 interviewers. Occasionally team sizes varied depending on the region and/or typical household characteristics for that area.

Wave 3 fieldwork was completed within one calendar year whereas the previous wave saw respondents interviewed either in 2010 and 2011. For Wave 3 all questions relating to the current year all referred to 2012 and so on for references to previous years.

3.4 Response Rates & Attrition

Wave 3 saw an improvement in the overall number of Wave 1 CSMs that were interviewed relative to Wave 2. The table below presents the figures of CSMs and TSMs successfully interviewed in each wave.

Table 1: CSMs and TSMs successfully interviewed by wave

		Wave 1	Wave 2	Wave 3
Wave 1	CSM	26776	22058	22375
Wave 2	CSM		908	887
	TSM		5585	3223
Wave 3	CSM			1067
	TSM			5081
Total successful individual interviews		26776	28551	32633

The non-response rate from Wave 1 to 2, when excluding those that moved out of scope or died between waves, is 19%. The equivalent non-response rate from Wave 2 to 3 for CSMs only (excluding deceased and those moved out of scope but including new CSMs from wave 2), is 16%.

The non-response rate for Wave 2 TSMs is significantly higher than the CSMs at 43%. This is expected as TSMs are not followed if they move out of a CSM household or if the CSM(s) leave them. The reasons for Wave 3 non-response can be seen in the table below (analysis compares Wave 2 individual outcomes to Wave 3 individual outcomes):

Table 2: Wave 2 and Wave 3 individual outcomes

Wave 2	Wave 3					
	Success	Refused/Not available	Household Level Non Response	Moved outside SA	Dead	Not co-resident with CSM
Success	23604	257	1944	1	547	2198
Refused/Not available	562	43	164	0	2	80
Household Level Non Response	2313	78	2078	13	153	0
Moved outside SA	6	0	3	42	0	0
Not in Wave 2	6148	169	11	0	0	0

Very encouraging is the significant number of individuals (2875) that were individual or household non-response in Wave 2, but who participated in Wave 3.

As is expected the biggest single reason for not re-interviewing those successfully interviewed in Wave 2 is due to TSMs no longer living with CSMs. When considering the biggest reason for not interviewing, regardless of Wave 2 outcome, household level non-response is once again the major factor. The specific reasons for household level non-response in Wave 3 are presented below:

Table 3: Reasons for household non-response at the individual level

	Number	Percent
Refused/Not Available	2054	48.67
Not Located	2129	50.45
Not Tracked	37	0.88
Total	4220	100

Two things are important to note here: Firstly the proportion of individuals not tracked is significantly lower than in Wave 2, where it was 13%. Secondly, the reasons for non-response are now almost equally distributed between refusal and no contact. It is a credit to the fieldwork company and the SALDRU CATI team that the number of individuals not located in Wave 3 is less than Wave 2. As time passes one would expect it to be harder to track respondents, but through their collective effort this trend was reversed.

Additionally, Wave 3 saw the addition of a new field new status for households, that of "temporarily away". This caught instances where no one was at a dwelling but it was discovered that they would return within the fieldwork period (but not while the team was currently in the relevant cluster). These dwellings would then be revisited later in the fieldwork period to "catch" and successfully interview the respondents at a later date. In Wave 2 these respondents would have been missed and recorded as "no one at home" after the mandated three attempts on differing days and times when the field team was in that cluster (and thus appear in the data reported as a "Not Available" household level non-response). The reasons for attrition between Waves 2 and 3 include:

Table 4: Reasons for Attrition

Reason	Number	Percent
Refusal	2405	44.21
Non-contact	2279	41.89
Deceased	756	13.90
Total	5440	100.00

The table shows three categories of attrition: "Refusals" are attritees who were not interviewed in Wave 3 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.

Attrition rates by province and income decile are not shown as the province and income decile are not known for those that were not interviewed in Wave 2 either. The racial distribution of attrition is presented below.

Table 5: Attrition by Racial Group

Pop. Group	Refusal	Non-contact	Deceased	Total	Attrition Rate
African	1300	1748	628	3676	13.39%
Coloured	480	282	97	859	18.20%
Asian/Indian	122	41	5	168	36.36%
White	503	208	26	737	50.31%
Total	**2405**	**2279**	**756**	**5440**	**15.95%**

In a pattern consistent with Wave 2, we see that non-contacts are the dominant reason for attrition among African respondents, while refusals dominate for White, Asian/Indian and Coloured respondents. The population groups with the highest attrition rates are Whites and Asian/Indian respondents.

4. Derived Variables

Certain variables in the derived datasets are created by the NIDS team. 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 variables, expenditure variables 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 rosters 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

The GPS information was used to determine the characteristics such as Main Place, District Council and Province for each dwelling. If the household could not be found and no GPS reading was taken then the geographical variables are empty. All successfully interviewed households had more accurate GPS readings taken during Wave 3.

For Wave 2 and Wave 3 a variable was defined *(wx_stayer)* at the individual level for respondents that remained in the same dwelling unit between waves. This variable identifies three types of respondents ((0) movers, (1) stayers and (2) new respondents) and refers in each wave to the individual's status relative to the previous wave.

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 as part of the Secure Data release.

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 Public Release data. Disaggregated occupational codes are available as part of the Secure Data release.

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, by distinguishing those who desired a job and were actively searching for work from those not actively searching.

4.6 Income

Total household income (*w3_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, *w3_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, we used the so-called "one-shot" income variable *w3_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 *w3_hhimprent* was added to all households, irrespective of the method of aggregation, where appropriate.

Table 6: Sources of Aggregation

Source of HH Income	Number of HHs	Percent
Individual Aggregation	6817	84.58
One-shot	1243	15.42
Total	**8060**	**100**

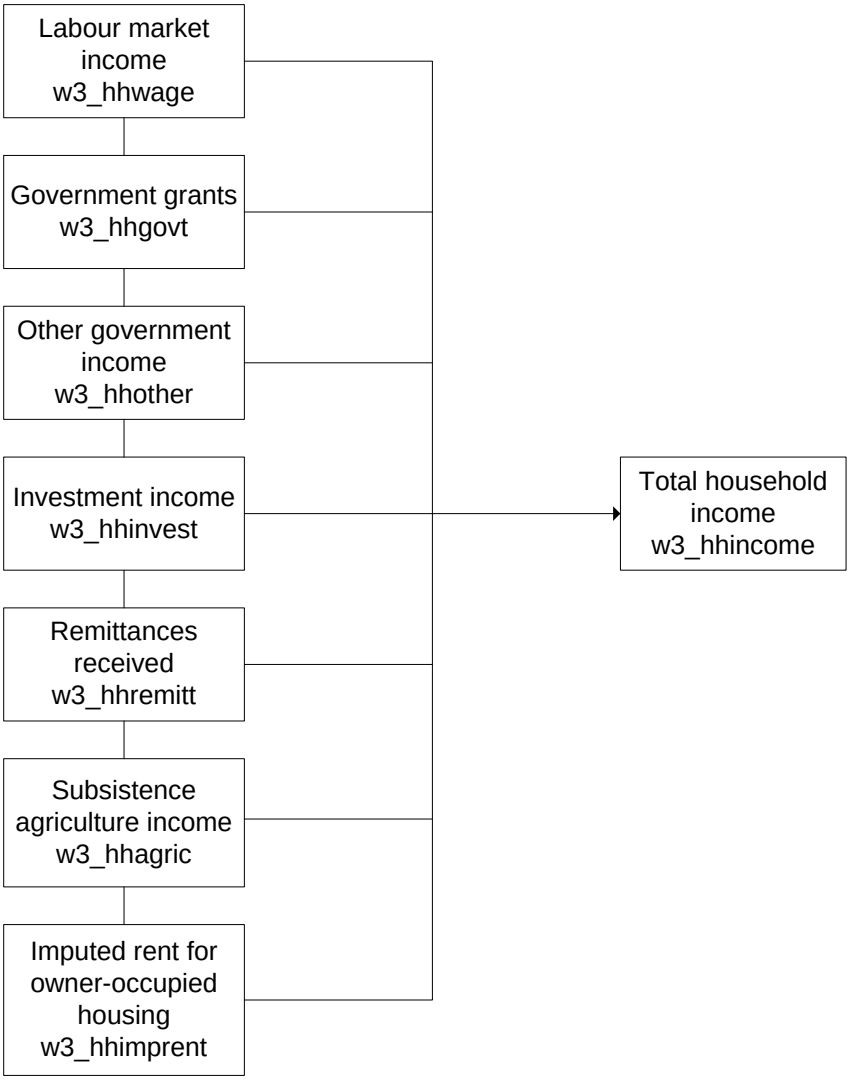
The table below lists the variables that make up each component of total household income.

Table 7: Components of aggregate household income

Household-level Variable	Individual-level Variable	Variable Name
Labour Market Income w3_hhwage	Main and second job	w3_fwag
	Casual wages	w3_cwag
	Self-employment income	w3_swag
	13th cheque	w3_cheq
	Bonus payment	w3_bonu
	Profit share	w3_prof
	"Help friends" income	w3_help
	Extra piece-rate income	w3_extra
Government Grant Income w3_hhgovt	State old age pension	w3_spen
	Disability grant	w3_dis
	Child support grant	w3_chld
	Foster care grant	w3_fost
	Care dependency grant	w3_cdep
Other Income from Government w3_hhother	Unemployment Insurance Fund	w3_uif
	Workmen's compensation	w3_comp
Investment Income w3_hhinvest	Interest/dividend income	w3_indi
	Rental income	w3_rnt
	Private pensions and annuities	w3_ppen
Remittance Income w3_hhremitt	Remittances received	w3_remt
Subsistence Agricultural Income w3_hhagric	N/A	N/A
Imputed Rental Income w3_hhimprent	N/A	N/A

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

Figure 4: Components of aggregate household income



4.6.1 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 upper bound of the top bracket[9].

[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.

4.6.2 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 and Wave 2 (see Finn et al, 2009; Brown et al 2011), 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.

Table 8: Income variable item non-response

Variable	Description	Obs	Achieved	% Missing	Imputation
w3_fwag	Main and secondary wages	5548	5272	4.97	Regression
w3_cwag	Casual wages	682	664	2.64	Regression
w3_swag	Self-employment income	831	665	19.98	Regression
w3_cheq	13th cheque	82	69	15.85	None
w3_prof	Profit share	9	9	0.00	None
w3_extr	Extra payment	6	6	0.00	None
w3_bonu	Bonus income	33	31	6.06	None
w3_othe	Other income	36	36	0.00	None
w3_help	Help friend income	48	47	2.08	None
w3_spen	State pension	2467	2466	0.04	Rule
w3_ppen	Private pension	341	321	5.87	Regression
w3_uif	UIF income	54	48	11.11	None
w3_comp	Workmen's compensation	15	14	6.67	None
w3_dis	Disability grant	722	719	0.42	Rule
w3_chld	Child support grant	4822	4820	0.04	Rule
w3_fost	Foster care grant	305	298	2.30	Rule
w3_cdep	Care dependency grant	104	103	0.96	Rule

w3_indi	Interest/dividend income	43	38	11.63	None
w3_rnt	Rental income	134	132	1.49	Regression
w3_remt	Remittances	1308	1128	13.76	Regression
w3_hhimprent	Imputed rental income	6922	4938	28.66	Regression

4.6.3 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.

The Wave 3 Household questionnaire differed from the Wave 2 questionnaire by asking for the rand values accruing to the household from the sale of agricultural produce and livestock. 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. The process used was similar to that applied in wave 1. This was deemed as the best estimated for household level agricultural income.

See the program library files on http://www.nids.uct.ac.za/documents/program-library/151-wave-3-income-dofiles for details on how agriculture income was calculated.

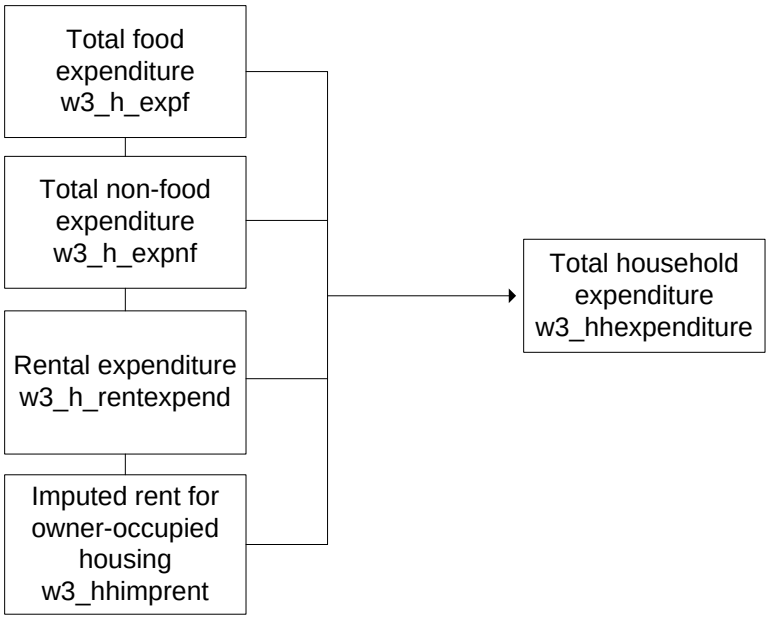
4.6.4 Bonus payments

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. Wave 3 asked for both annual and monthly amounts, and the latter was chosen so as to be consistent with Wave 2.

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 (*w3_h_expf*) and total non-food expenditure (*w3_h_expnf*) respectively. These two components were added to total rental expenditure (*w3_h_rentexpend*) and imputed income from owner occupied housing[10] (*w3_hhimprent*) to constitute aggregated total household expenditure (*w3_h_expenditure*).

Figure 5: Components of aggregate household expenditure



4.7.1 Imputations

4.7.1.1 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 previous waves. 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).

4.7.1.2 Non-food

If a respondent indicated that the household purchased one of the 54 non-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.

4.7.1.3 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).

4.7.1.4 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 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 BMI for age were calculated using the 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 Stata macros igrowup and who2007 were used to calculate the z-scores and are available for downloaded from www.who.int/childgrowth/software/en/.

The following variables were created:

w3_zhfa - height or age for individuals up to 19 years of age
w3_zwfa - weight for age for individuals up to 10 years of age
w3_zwfh - weight for height for individuals up to 5 years of age
w3_zbmi - 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 zscores, 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. Age in days was used to calculate the z-scores.

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 were 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.

4.8.1 Important note about using the publically released NIDS data to create your own z-scores

NIDS has received a number of queries from users who have created their own z-scores using the publically released data sets and noticed substantial discrepancies with the z-scores released by NIDS. Most queries are from researchers who have used the zanthro macro. There are a number of reasons why z-scores created by zanthro differ from those released by NIDS. First and most important, is the precision of the age variable. The zanthro macro expects an exact age variable and the default unit for age is age in years. This means that a 2 year old child is considered to be 2 years and 0 days old. In the NIDS sample, on average, we would expect 2 year olds to be 2 years and 6 months old. When the zantrho macro is used with age measured in years, children are being compared to a reference population that is on average 6 months and in some cases as much as 364 days younger than they are. This results in substantially inflated z-scores and under-estimates of the proportion of children who are stunted or underweight for age. The problem is particularly severe at younger ages when velocity of growth is high. The table below illustrates just how misleading estimates of stunting calculated using zanthro can be. The prevalence of stunting among children aged 2 to 10 years is estimated at 17% using the WHO macros with age measured in days. The corresponding estimates using the zanthro macro with age measured in years is only 8%. The underestimation from using zanthro is most pronounced at the youngest ages.

Table 9: Comparison of the proportion of children who are stunted (z-score < 2) by calculation method

	Calculation method	
	WHO macros with age in days	zanthro with age in years
Age		
2	0.311	0.054
3	0.266	0.083
4	0.177	0.052
5	0.131	0.068
6	0.142	0.073
7	0.140	0.102
8	0.129	0.111
9	0.101	0.080
10	0.115	0.067
Total	0.166	0.077

Adding 0.5 to the age in years variable and re-running the zanthro macro produces estimates for mean z-scores and prevalence of stunting and underweight for age that are in line with the WHO estimates using age in days. The problem with this approach is that, while averages will be correct, z-scores for individual children can be substantially over or under estimated.

Running the zanthro macro using age in days produces very similar results to the WHO macros, both on average and at the individual level. There are other reasons for minor discrepancies between results using the WHO and zanthro macros. The cut-offs for biologically implausible values are

slightly different. For example, zanthro sets z-scores for height for age to missing if they are below -5 or above 5. Note that for comparison purposes in the table above, the WHO z-scores were restricted to be between -5 and 5. The reference populations for the two macros are also different. The zanthro macro uses either the 2000 CDC Growth Reference or the 1990 British Growth Reference as the reference population. In practice, these differences have very little impact on the calculated z-scores.

The publically released datasets allow one to create a variable for age in months. Using this variable with the WHO macros or zanthro will produce similar results to the publically released z-scores.

4.9 Weights[11]

4.9.1 What is new?

Together with Wave 3 of the National Income Dynamics Study, updates to Wave 2 and Wave 1 have been released. Since the information on the sample for these waves has changed a little (e.g. age information has been improved, some households have been removed) it has been necessary to recalculate **all** the weights previously released as well. Indeed since a few households have been removed from Wave 1 even the "design weights correcting for nonresponse" will be slightly different in the affected clusters.

Nevertheless the **methods** used, i.e. the algorithms underpinning the calculations, have not been changed. This means that the revised weights will be very similar in most cases to the ones released previously. Indeed because the algorithms have not been changed, the documentation released with previous weights should be consulted as well for further information.

The **calibrated weights**, however, have changed in that all calibration has happened to the revised mid-year population estimates as released by Statistics South Africa in 2013. This was necessary to ensure that the population totals (and totals within particular provinces and age groups) did not jump discontinuously as a result of the upward revision of South Africa's overall population size. In practice this means that the calibrated weights for 2008 and 2010 will now gross up to slightly larger totals than before.

4.9.2 The relationship between the different weights

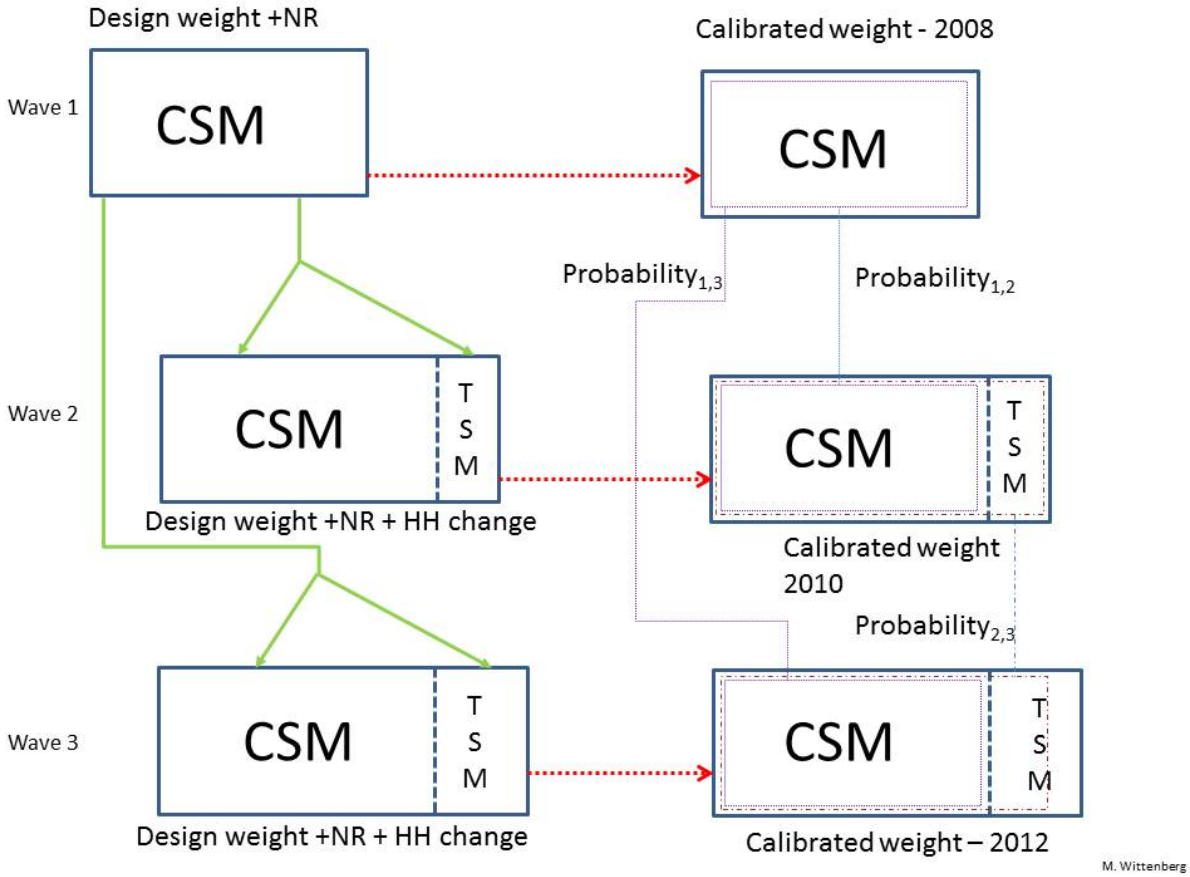
It can be rather difficult to keep track of all the different types of weights that there are in the National Income Dynamics Study. Figure 1 presents the relationships in diagrammatic form.

Fundamentally there are **three** types of weights:

a) Design weights (correcting for nonresponse)
b) Calibrated weights
c) Panel weights

[11] This section was drafted by Martin Wittenberg

Figure 6: The relationship between the different weights in NIDS



The design weights released with Wave 1 are fundamental to every other weight released with NIDS[12]. They are used to calculate the corresponding design weights for waves 2 and 3 (the green arrows in Figure 1).

Each of the waves, treated as a cross-section of the South African population, has been separately **calibrated** to the corresponding population totals as given in the mid-year population estimates released in 2013. This process is indicated in the diagram by the red arrows.

In order to work with changes over time we need to work with individuals that we observe at least twice. This in turn means that we need to correct for attrition. In order to do this, the probability of observing the individual again is calculated. There are three such probabilities shown in Figure 1:

- **Probability$_{1,2}$ –** This is the probability of observing an individual from Wave 1 (i.e. one of the CSMs) again in Wave 2
- **Probability$_{1,3}$** – This is the probability of observing an individual from Wave 1 (i.e. one of the CSMs) again in Wave 3
- **Probability$_{2,3}$** – This is the probability of observing an individual from Wave 2 (CSM or TSM) again in Wave 3

[12] As the technical document released with wave 1 indicates (Wittenberg 2009), calculating appropriate design weights is not straightforward. The weights released for waves 2 and 3 are based on the weights ignoring replacement.

Given one of these probabilities, one could calculate either panel versions of the design weights, i.e. design weights correcting for attrition, or panel versions of the calibrated weights, i.e. panel weights correcting for attrition. As shown in Figure 1 (by the purple connecting lines in the right hand side of the Figure) the panel weights released with NIDS are based on the calibrated weights.

We now turn to a more detailed discussion of the different types of weights.

4.9.3 Design weights

The individuals interviewed in waves 2 and 3 included both household members in the original sample (CSMs) as well as some new individuals who were now co-resident 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 co-resident. The most straightforward procedure (used to calculate the NIDS cross-sectional weights) is to average out sample weights within the Wave 2 or Wave 3 households, assigning TSMs a weight of zero.

The case of new-born 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 new-borns are CSMs, they should be thought of as indirectly sampled through their mothers, i.e. their mothers weight should be assigned to the new-born 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_hhweight1). The resultant weights (w2_hhweight1 and w3_hhweight1) 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 co-resident 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.

4.9.4 The calibrated weights

All waves were calibrated to provincial totals and to sex-race-age group cell totals (with the oldest three age categories for Indian males and Indian females collapsed, as noted in the release notes accompanying the Wave 2 release). The calibration was done using the Stata ***maxentropy*** add-in (Wittenberg 2010). All individuals within the same household were constrained to get the same weight.

4.9.4.1 Why is there a need to calibrate the weights

The "design weights" have solid theoretical credentials. Nevertheless there are also good reasons for using the calibrated weights. Even when we adjust the design weights for household nonresponse we find that the realised (weighted) sample differs from the national population in systematic ways. For instance old Africans (male and female) are overrepresented, while African males and females aged 25 to 39 are relatively underrepresented, which suggests that households with pensioners were more readily enumerated (probably because there was somebody home when the survey teams called) than households in which there were neither younger children or pensioners. Any statistics which are correlated with the age-sex-race or provincial breakdowns are likely to be measured more accurately with the calibrated weights.

4.9.4.2 Issues to take note of when using the calibrated weights

Nevertheless getting the sample aligned with the national demography comes at a cost. It is much harder to find weights to align certain "cells" of the age-sex-race cross-tabulation with the national distribution than others. One measure of how far the weights had to be pushed from their baseline is given by the Lagrange multipliers that the ***maxentropy*** command returns. Values close to zero indicate that the constraint did not bind[13]. The following cells gave difficulties (taking a λ value in excess of four as a sign that the constraint gave problems):

Wave	Constraint	λ	Constraint	λ	Constraint	λ
1	African Male 80+	-4.9	Col Male 75-79	-4.9	Indian Male 25-29	11.0
	Indian Male 30-34	4.2	Indian Male 50-54	8.5	Indian Male 55-59	-4.9
	Indian Female 0-4	-5.2	Indian Female 40-44	-4.3	Indian Female 45-49	-9.3
2	African Male 80+	-4.3	African Female 80+	-4.2	Indian Male 40-44	-4.4
	Indian Male 45-49	10.5	Indian Male 50-54	5.8	Indian Male 65-69	18.3
	Indian Female 0-4	4.7	Indian Female 5-9	6.1	Indian Female 15-19	-4.1
	Indian Female 30-34	4.9	White Female 25-29	4.4		
3	African Female 80+	-5.0	Col Male 80+	-4.6	Indian Male 5-9	-6.0
	Indian Male 30-34	10.4	Indian Male 45-49	7.8	Indian Male 50-54	9.7
	Indian Male 55-59	7.0	Indian Male 65-59	9.6	Indian Male 70+	5.3
	Indian Female 20-24	-4.9	Indian Female 25-29	-9.3	Indian Female 45-49	-4.0
	Indian Female 50-54	-6.1				

It should be noted that the sign of the multiplier is an indication whether the weight associated with that group had to be increased (positive multiplier) or decreased (negative). As noted earlier, the sample shows a clear excess of old Africans and, indeed, Coloured males. It is also evident that the calibration had great difficulty with the Indian subpopulation. The general picture is that there seem to be relatively too few prime-age males and too many women. The fact that we also constrained weights to be common within household would have made this problem much more difficult, hence some of the rather large Lagrange multipliers.

The main lesson to be drawn from this is that **great caution should be exercised if the Indian subsample is analysed by itself**. The raw sample shows curious relative deficits and surpluses. The

[13] If all weights have to be scaled up by the same ratio then the multiplier will also be zero. It will only be nonzero if the *relative* weights have to be changed.

calibrated weights will smooth those over – but because they have been heavily adjusted they might introduce unexpected effects in turn.

It might also be observed that the pattern seems to have become worse over time. This is probably due, in part, to differential attrition.

4.9.5 Panel weights

The individuals who were successfully re-interviewed in waves 2 and 3 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.

All of the probabilities shown in Figure 1 were estimated using probit models using the baseline characteristics of the individual. 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.

4.9.5.1 Wave 1 to Wave 2 attrition

As shown in Table 1 the pattern of probabilities varies quite strongly with race and age[14]. And unfortunately this pattern of attrition is correlated quite strongly with the initial pattern of nonresponse.

Table 10 Average probabilities of successful re-interview - Wave 1 to Wave 2 – by age, gender and race

	Best gender and best race							
Age Intervals	Male				Female			
	African	Coloured	Asian/Indian	White	African	Coloured	Asian/Indian	White
-9					0.021			
0-1	0.855	0.803	0.883	0.781	0.859	0.811		0.838
1-4	0.857	0.788	0.782	0.718	0.848	0.782	0.808	0.665
5-9	0.873	0.794	0.645	0.617	0.862	0.776	0.740	0.580
10-14	0.866	0.765	0.524	0.438	0.853	0.771	0.640	0.414
15-19	0.827	0.719	0.486	0.344	0.825	0.734	0.540	0.328
20-24	0.783	0.679	0.478	0.306	0.805	0.720	0.555	0.320
25-29	0.746	0.645	0.522	0.367	0.802	0.716	0.586	0.370
30-34	0.726	0.653	0.609	0.352	0.810	0.735	0.598	0.420

[14] Observe that if the probit had been uninformative, i.e. all coefficients equal to zero, then these probabilities would all be the same.

35-39	0.714	0.660	0.643	0.428	0.817	0.757	0.626	0.486
40-44	0.712	0.671	0.664	0.465	0.834	0.764	0.659	0.481
45-49	0.734	0.679	0.718	0.502	0.843	0.778	0.704	0.514
50-54	0.744	0.710	0.714	0.504	0.850	0.788	0.700	0.537
55-59	0.757	0.714	0.663	0.562	0.862	0.772	0.680	0.553
60-64	0.772	0.693	0.616	0.543	0.861	0.766	0.646	0.551
65-69	0.789	0.688	0.444	0.511	0.856	0.762	0.585	0.498
70-74	0.785	0.643		0.485	0.846	0.717	0.442	0.433
75-79	0.747	0.533	0.289	0.470	0.817	0.672	0.410	0.424
80-84	0.681	0.498		0.376	0.777	0.615		0.373

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, 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 baffled by weights. In order to prevent avoidable errors we decided to trim the weights to the 1st and 99th percentiles of the weight distribution.

4.9.5.2 Wave 1 to Wave 3

The table of average probabilities by age-sex-race cells is given in Table 2.

Table 11 Average probabilities of successful re-interview - Wave 1 to Wave 3 – by age, gender and race

	Best gender and best race							
Age Intervals	Male				Female			
	African	Coloured	Asian/Indian	White	African	Coloured	Asian/Indian	White
-9		0.020			0.066			
0-1	0.884	0.863	0.905	0.859	0.889	0.860		0.839
1-4	0.884	0.850	0.812	0.776	0.876	0.842	0.832	0.718

5-9	0.894	0.862	0.676	0.660	0.885	0.843	0.752	0.599
10-14	0.884	0.843	0.565	0.487	0.869	0.831	0.645	0.457
15-19	0.847	0.806	0.536	0.390	0.845	0.804	0.574	0.357
20-24	0.796	0.766	0.585	0.355	0.824	0.787	0.586	0.330
25-29	0.754	0.738	0.603	0.379	0.808	0.775	0.609	0.376
30-34	0.712	0.731	0.647	0.416	0.811	0.782	0.657	0.429
35-39	0.706	0.720	0.713	0.447	0.810	0.789	0.686	0.472
40-44	0.693	0.721	0.726	0.498	0.823	0.790	0.716	0.506
45-49	0.701	0.717	0.714	0.515	0.824	0.806	0.737	0.519
50-54	0.708	0.722	0.689	0.517	0.831	0.809	0.727	0.544
55-59	0.714	0.714	0.604	0.527	0.841	0.796	0.699	0.567
60-64	0.733	0.672	0.570	0.515	0.837	0.792	0.665	0.547
65-69	0.738	0.640	0.414	0.472	0.821	0.782	0.580	0.542
70-74	0.714	0.557		0.448	0.800	0.728	0.446	0.515
75-79	0.657	0.397	0.266	0.473	0.754	0.649	0.415	0.512
80-84	0.549	0.223		0.402	0.681	0.616		0.550
85+	0.305			0.454	0.507	0.443		0.527

It seems noteworthy that some of the probabilities are actually higher for a reinterview in Wave 3 than was the case for Wave 2. This suggests that the survey team was more successful in tracing some of the individuals first interviewed in 2008. The age profile is as expected, with fewer survivors at high baseline ages.

4.9.5.3 Wave 2 to Wave 3

It should be noted that the CSMs form the core of the panel, i.e. dynamic questions should preferably be investigated using only the sample of CSMs. Nevertheless many TSMs who were seen for the first time in Wave 2 were again interviewed in Wave 3. It is to be expected that analysts that wish to focus on the changes that occurred between 2010 and 2012 will probably want to use as many individuals that they see twice as they can.

Assuming that at least some analysts will want to do these kinds of studies, we have released a set of weights (based on the "Probability$_{2,3}$" shown in Figure 1). Nevertheless these weights come with a strong warning: attrition of TSMs between Wave 2 and Wave 3 is a very different type of process

than attrition of a CSM. Besides all the different ways in which a CSM might be lost to the study (death, migration with no forwarding address, refusal to participate again) TSMs will drop out of the study the moment that they cease to co-reside with a CSM. The "attrition weights" for the change in sample between Wave 2 and Wave 3 are therefore conceptually much more messy than the corresponding weights for CSMs[15] .

4.9.6 A final comment on the weights

If any of these details look unappealing, it is possible to re-do any of these weights according to the logic outlined in Figure 1. With the exception of the original Wave 1 design weights (corrected for nonresponse), none of the other steps require "insider" information. Every subsequent step is simply a transformation of those original weights.

Should one use these weights? For most purposes it would be simply inappropriate to do unweighted analyses. Multivariate regressions that control for many of the same variables that are used in the sampling or that are important for nonresponse may be one exception. But then one would need to be confident that one has adequately controlled for the sampling design.

It is true that in some cases one gets "nice" results with unweighted data and strange ones with weights. In those cases one should investigate why the weights produce strange results. A good starting point would be to exclude a handful of observations with the largest weights. If the weighted results are driven by one or two individuals then one would be entitled to be sceptical of the weighted results. More typically one may find that one is asking questions that the data are simply not capable of answering. As noted above (in the case of the Indian sub-sample) analysing any sub-sample that is too small is probably inviting trouble.

[15] Note that if one wanted to restrict the analysis of changes between waves 2 and waves3 only to CSMs then the "wave 1 to wave 3" transition weights would still be appropriate.

5. Program Library

Stata syntax files (do-files) compressed into Zip format can 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.

5.1 Data Manipulation

5.1.1 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 merging

Program 3a - Merging individual across Wave 1 and Wave 2 (Balanced panel)

Program 3b - Merging Individuals across Wave 2 and Wave 3 (Balanced panel)

Program 3c - Merging Individuals across Wave 1, Wave 2 and Wave 3 (Balanced panel)

Program 4 - Merging Households across Wave 1, Wave 2 and Wave 3

5.1.2 Reshaping data

Program 5a - Wave 1 Reshaping the birth history and merging in the Child questionnaires

Program 5b - Reshaping the birth history and merging in the Child questionnaires

Program 6a - Reshaping the mortality section

Program 6b - W2 Reshaping the mortality section

5.2 Derived Variables

5.2.1 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. The following do files shows exactly how the derived income variables were created. In order to replicate results they have to be run as a set.

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

5.2.2 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. The following do files shows exactly how the derived expenditure variables were created. In order to replicate results they have to be run as a set.

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

5.2.3 Deflator

Because fieldwork for each Wave of NIDS takes place over at least one calendar year, all financial data need to be deflated.

Program 10a – Deflators W2_BaseMonth_Sep2010

Program 10b – Deflators W1_BaseMonth_Dec2012

Program 10c – Deflators W2_BaseMonth_Dec2012

Program 10d – Deflators W3_BaseMonth_Dec2012

5.2.4 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

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

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, 2009, “Weights: Report on NIDS Wave 1”, NIDS Technical Paper 2, available at http://www.nids.uct.ac.za/home/index.php?option=com_docman&task=doc_download&gid=106&Itemid=19

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