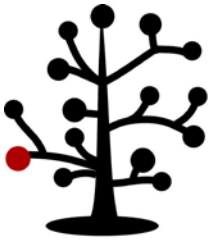


N.i.D.S.

NATIONAL INCOME DYNAMICS STUDY

National Income Dynamics Study
Panel User Manual

Edited by Michelle Chinhema, Timothy Brophy, Michael Brown, Murray Leibbrandt, Cecil Mlatsheni and Ingrid Woolard

Read Me

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 documentation accompanies the release of the Wave 4 data, together with updated versions of Wave 1, 2 & 3 data sets. Highlights in this data are as follows:

- The inclusion of Police Station District data
- Coding of migration variables using Statistics South Africa's Census 2011 Geographical data
- Use of International Standard Classification of Occupations (ISCO) codes to consistently classify employment codes across all waves
- Addition of education progression questions in Wave 4 to correct inconsistent historic education data
- Allocation of identifiers (PID's) to all children on the birth history
- Improvement of parental data
- For panel consistency, pcodes have been removed in Wave 1 data
- Consistently low attrition rates

Contents

List of Contributors

This document was created by the NIDS team. For the correct citation method, see section 2.2 of this document. Authors in alphabetical order include:

- Cally Ardington
- Lydia Boateng
- Timothy Brophy
- Michael Brown
- Michelle Chinhema
- Reza C. Daniels
- Louise De Villiers
- Arden Finn
- Amy Kahn
- Murray Leibbrandt
- Zvikomborero Madari
- Cecil Mlatsheni
- Sibongile Musundwa
- Adeola Oyenubi
- Martin Wittenberg
- Ingrid Woolard

1. Summary Figures

This section presents the total number of observations in each dataset for each wave, the response rate for each wave and finally attrition between waves.

1.1 Number of Observations

Table 1.1 below shows the total number of observations in each dataset for each wave.

Table 1.1: Summary of n-values across waves

File Name	Identifiers*	n			
		w1	w2	w3	w4
Link File	Pid	-	34961	41250	50379
HHQuestionnaire	w**X**_hhid	7296	9127	10219	11895
HouseholdRoster	w**X**_hhid	7296	9127	10219	11895
	pid	31141	35216	40643	47059
Adult	w**X**_hhid	7289	8845	9967	11611
	pid	16871	21880	22466	26819
Proxy	w**X**_hhid	1375	898	2068	1385
	pid	1750	1124	2715	1600
Child	w**X**_hhid	4328	5031	5606	6303
	pid	9605	11081	12216	13918
Hhderived	w**X**_hhid	7296	9127	10219	11895
Indderived	w**X**_hhid	7296	9016	10114	11732
	pid	28226	34085	37397	42337

* **X** represents the wave number i.e. w1

1.2 Response Rates

Table 1.2 presents the numbers of CSMs and TSMs successfully interviewed in each wave as well as the number of CSMs and TSMs that were added each wave. 78% of the individuals who were interviewed in Wave 1 were successfully interviewed in Wave 4. Out of the 1858 CSMs who were either added to the study in Wave 2 or not successfully interviewed in Wave 1, 84% were successfully interviewed in Wave 4 and 92% of the CSMs who were added in Wave 3 were successfully interviewed in Wave 4. It can be seen that the percentage of successfully interviewed individuals is much larger for the CSMs than for the TSMs because TSMs are not followed if they move out of a CSM household or if the CSMs leave the household.

Table 1.2: CSMs and TSMs successfully interviewed by wave

		Interviewed in Wave 1	Interviewed in Wave 2	Interviewed in Wave 3	Interviewed in Wave 4
First Present in Wave 1	CSM	26776	21112	21391	20774
First Present in Wave 2	CSM		1858	1598	1558
	TSM		5568	3136	2268
First Present in Wave 3	CSM			1345	1232
	TSM			5127	2535
First Present in Wave 4	CSM				1704
	TSM				7325
Total successful individual interviews		**26776**	**28538**	**32597**	**37396**
CSMs attempted		28226	29222	29449	30478
TSMs attempted			5739	8656	12742

A comparison on individual outcomes across waves is presented in Table 1.3, Table 1.4 and Table 1.5. The most common reason individuals were interviewed in one wave but not the next is TSMs who no longer live in a household with any CSMs. Since TSMs are not tracked should they move out of a household of CSMs, they will not be re-interviewed.

Table 1.3: Wave 4 and Wave 3 individual outcomes

Wave 4	**Wave 3**						
	Successfully Interviewed	Refused/ Not Available	Household Level Non-Response	Moved Outside of SA	Deceased This Wave	Deceased in a Prior Wave	Not Co-resident with any CSMs
Successfully Interviewed	26534	305	1383	0	0	0	290
Refused/ Not Available	271	61	160	0	0	0	6
Household Level Non-Response	1444	72	1074	0	0	0	20
Not Tracked in Wave 4	91	18	1401	56	0	0	0
Moved Outside of SA	8	1	13	0	0	0	0
Deceased this Wave	770	16	93	0	0	0	4
Deceased in a Prior Wave	0	0	0	0	708	876	0
Not Co-resident with any CSMs	3479	89	58	0	0	0	1949

Table 1.4: Wave 3 and Wave 2 individual outcomes

Wave 3	Wave 2				
	Successfully Interviewed	Refused/Not Available	Household Level Non-Response	Moved Outside of SA	Deceased this Wave
Successfully Interviewed	23609	560	2309	6	0
Refused/Not Available	263	50	82	0	0
Household Level Non- Response	1932	164	2074	3	0
Moved Outside SA	1	0	13	42	0
Deceased this Wave	543	12	153	0	0
Deceased in a prior wave	0	0	0	0	876
Not co-resident with any CSMs	2190	79	0	0	0

Table 1.5 below examines the interview outcomes for individuals between Wave 1 and Wave 2. As Wave 1 was the baseline study, only two outcomes were used in field, namely "Successfully Interviewed" or "Refused/Not Available".

Table 1.5: Wave 2 and Wave 1 individual outcomes

Wave 2	Wave 1	
	Successfully Interviewed	Refused/Not Available
Successfully Interviewed	21112	947
Refused/Not Available	532	94
Household Level Non- Response	4249	365
Moved Outside SA	49	2
Deceased this Wave	834	42

The reasons for individual household-level non-responses are given in Table 1.6. Household non-responses were not specified in Wave 1 and therefore there are no reasons for non-responses available for this wave.

Table 1.6: Reasons for household non-response at the individual level

		Refused / Not Available	Not Located	Not Tracked	Whole HH Dead	Moved Outside SA	Total
Wave 4	Number	1958	817	1546	189	38	4548
	Percent	43.05	17.96	33.99	4.16	0.84	100
Wave 3	Number	2049	2113	41	133	117	4453
	Percent	46.01	47.45	0.92	2.99	2.63	100
Wave 2	Number	1805	2200	625	158	82	4870
	Percent	37.06	45.17	12.83	3.24	1.68	100

Wave 4 sees an apparent spike in "Not Tracked" outcomes, this inflation was artificially created by removing multiple wave on wave "Refusers" and "Not Located" from the Wave 4 listing that went to fieldwork.

1.3 Attrition

Attrition between waves is defined by comparing a wave to its proceding waves. For example, the number of successful interviews in Wave 3 is compared to that of Wave 2, providing us with the Wave 3 attrition rate. The sample used to determine attrition contains those respondents that are present in both waves and alive at the beginning of the wave of interest. For example, a respondent must be alive in Wave 3 but can be deceased at the end of Wave 4.

Table 1.7: Reasons for attrition

	Reason	Refusal	Non-Contact	Deceased	Total
Wave 4	Number	2293	2397	883	5573
	Percent	41	43	16	100
Wave 3	Number	2418	2267	708	5393
	Percent	45	42	13	100
Wave 2	Number	2427	2893	876	6196
	Percent	39	47	14	100

Table 1.7 shows three categories of attrition: "Refusals" are attritees who were not interviewed across the panel 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, "Deceased" are those respondents who died between waves.

The racial distribution of attrition is presented below.

Table 1.8: Wave on wave attrition by race

	Pop. Group	Refusal	Non-Contact	Deceased	Total	Attrition Rate
Wave 4	African	1410	1487	718	3615	10.97
	Coloured	418	368	120	906	16.41
	Asian/Indian	117	86	10	213	42.94
	White	348	456	35	839	53.47
	Total	2293	2397	883	5573	13.75
Wave 3	African	1309	1737	581	3627	13.22
	Coloured	483	281	97	861	18.21
	Asian/Indian	122	41	5	168	36.36
	White	504	208	25	737	50.31
	Total	2418	2267	708	5393	15.82
Wave 2	African	1200	2189	738	4127	18.59
	Coloured	554	465	102	1121	26.93
	Asian/Indian	135	32	8	175	40.79
	White	538	207	28	773	53.94
	Total	2427	2893	876	6196	21.95

As shown in Table 1.8, 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.

It is important to note that this wave on wave attrition does not reflect previously attrited respondents that were successfully interviewed in subsequent waves. As such from a panel perspective, attrition rates will be lower than reflected on those wave on wave attrition rates. E.g. an African respondent who refused in Wave 2 but was successfully interviewed in Wave 3. This negative attrition is not reflected in Table 1.8.

2. Using This Manual

The National Income Dynamics Study (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.

This document accompanies the release of the Wave 4 data. As with any new wave data release, there have been updates to the data of previous waves. Please refer to the latest documentation of changes between waves if merging to this dataset. These are available on the NIDS website: www.nids.uct.ac.za

2.1 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 3. This entire section should be read, especially subsection 3.6 on merging datasets within and between waves.
- The fieldwork schedule: see section 4.3.
- Weights. See section 6.11.
- Correctly merge NIDS data using Stata: see section 8.1.1.
- Deflate financial data: see section 8.2.4.

2.2 Citation of NIDS Data and Documentation

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

Data Citation:

Wave 4:

Southern Africa Labour and Development Research Unit. National Income Dynamics Study 2014 - 2015, Wave 4 [dataset]. Version 1.1. Cape Town: Southern Africa Labour and Development Research Unit [producer], 2016. Cape Town: DataFirst [distributor], 2016. Pretoria: Department of Planning Monitoring and Evaluation [commissioner], 2014

Wave 3:

Southern Africa Labour and Development Research Unit. National Income Dynamics Study 2012, Wave 3 [dataset]. Version 2.1. Cape Town: Southern Africa Labour and Development Research Unit [producer], 2016. Cape Town: DataFirst [distributor], 2016

Wave 2:

Southern Africa Labour and Development Research Unit. National Income Dynamics Study 2010 - 2011, Wave 2 [dataset]. Version 3.1. Cape Town: Southern Africa Labour and Development Research Unit [producer], 2016. Cape Town: DataFirst [distributor], 2016

Wave 1:

Southern Africa Labour and Development Research Unit. National Income Dynamics Study 2008, Wave 1 [dataset]. Version 6.1. Cape Town: Southern Africa Labour and Development Research Unit [producer], 2016. Cape Town: DataFirst [distributor], 2016

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

Documentation Citation:

Chinhema, M., Brophy, T., Brown, M., Leibbrandt, M., Mlatsheni, C., & Woolard, I., eds. 2016. "National Income Dynamics Study Panel User Manual", Cape Town: Southern Africa Labour and Development Research Unit.

3. The NIDS Data

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.*

3.1 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

See the "how to register' video can be viewed by clicking here or follow steps below.

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 should 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.

3.2 Data Formats

The data sets are available in Stata format only. Previously, the data was made available in R, SAS and SPSS. However all these programs now have a functionality to import Stata files.

3.3 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*, w**X**_hhid* (where ***X*** denotes the wave[4]). 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*[5]*.* 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 who 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, even though they are not strictly speaking a household member. This allows information to be collected for household members who are *out of scope*[6]*.*

[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 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] This notation is used throughout this document.
[5] 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.
[6] 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.

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

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

3.4 File Structure

The data files that make up the NIDS dataset in each wave 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*.

HHQuestionnaire: One record per household with data from the household questionnaire, excluding the household roster. Unique identifier: *w**X**_hhid*.

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: *w**X**_hhid*, non-unique identifier for individual: *pid*. The combination of *w**X**_hhid* and *pid* is unique per person within each wave.

Adult: One record per entry from the Adult[7] questionnaire. Unique identifier for household: *w**X**_hhid*, unique identifier for individual: *pid*. Observations with no data beyond Section A of the questionnaire are individuals who 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 *w**X**_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[8] questionnaire. Unique identifier for household: *w**X**_hhid*, unique identifier for individual: *pid*.

Child: One record per entry from the Child questionnaire. Unique identifier for household: *w**X**_hhid*, unique identifier for individual: *pid*. Observations with no data beyond Section A are individuals who refused to participate in the survey either at a

[7] A person is defined as an adult if they were 15 years old or older on the day of the interview.

[8] 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.

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 *w**X**_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: *w**X**_hhid*. 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: *w**X**_hhid*, unique identifier for individual: *pid*.

Admin: One record per entry from the Admin data. Unique identifier for household: *w**X**_hhid*, unique identifier for individual: *pid*.

3.5 Identifiers

Individuals can be identified across waves by their unique identifier *pid.* Households are identifiable within waves by their unique identifier *w**X**_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.

3.6 Merging Datasets Within and Between Waves

Since the release of Wave 2 the longitudinal dimension of NIDS can be explored and with each subsequent wave's 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. A link to examples of the Stata code to merge within and between waves is provided below in Section 8.

3.6.1 Identifying CSMs and Residents

The variable *wX_r_csm* in each wave's Household Roster file can be used to identify CSMs. All original CSMs can be identified by using the *w**X**_r_csm* variable in the Household Roster file. Note that only *resident* household members in Wave 1 were selected to be CSMs; however all household members in all waves have been assigned a *pid*, regardless of their CSM or residency status.

The variable *w**X**_r_pres* in each wave's Household Roster file can be used to identify residents. The residency criteria is important as 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.

Figure 3.1 below shows how the NIDS sample of CSMS's and TSM's has grown over time. In Wave 1 a total of 28 226 respondents were resident in the households selected to participate in the survey. These respondents became continuing sample members (CSMs). Between Wave 1 and Wave 2, 876 CSMs died and 996 children who were born to CSM mothers were added. The total number of CSMs in Wave 2 was 28 295. Almost 40 percent of the 5 739 TSMs added to the sample in Wave 2 were no longer residing with a CSM in Wave 3 and these respondents were therefore not tracked. The number of TSMs grew to 12 604 in Wave 4. Moving from Wave 4 into Wave 5, NIDS will strive to track the lives of 29 733 CSMs.

Figure 3.1: CSMs and TSMs across waves

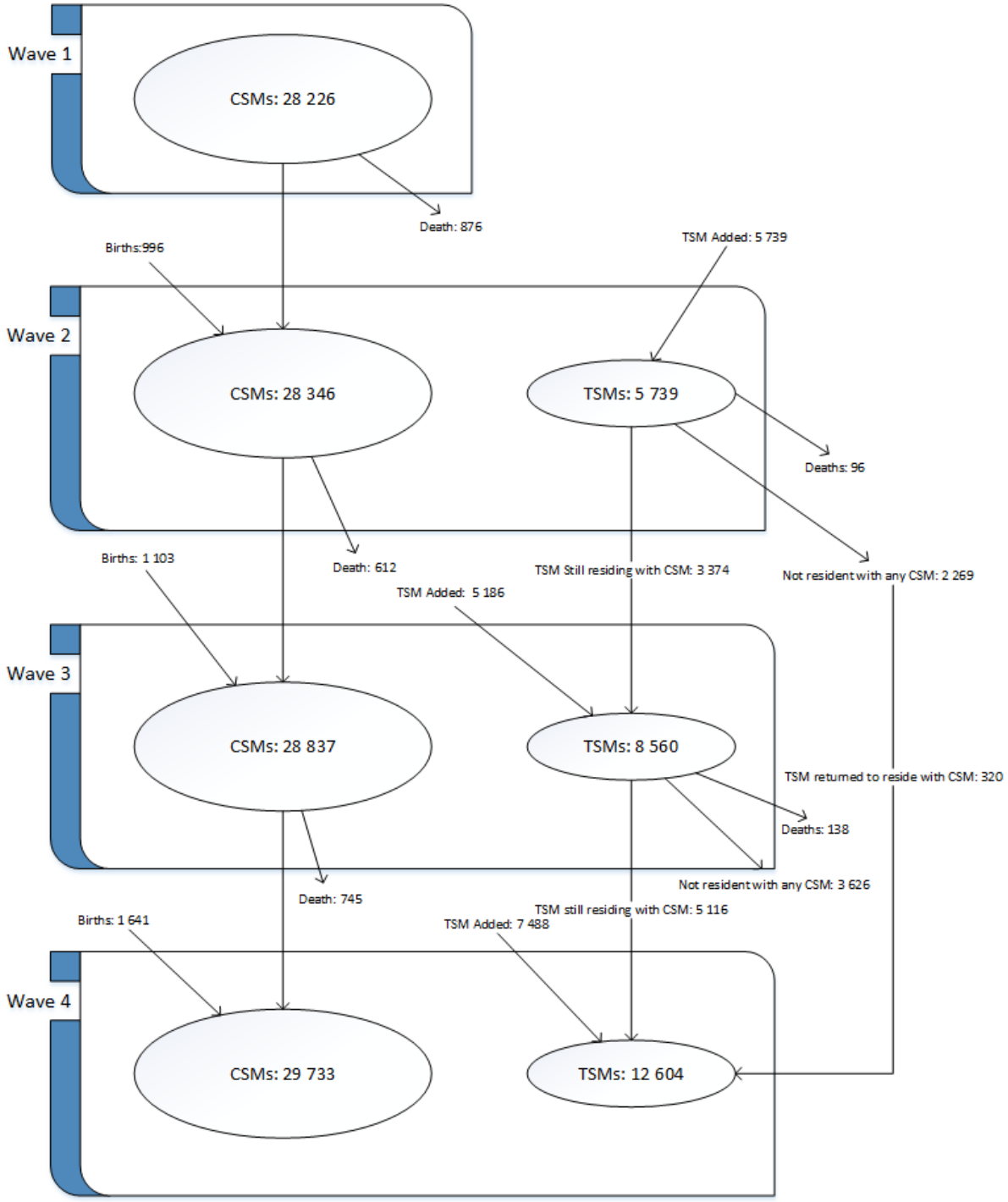


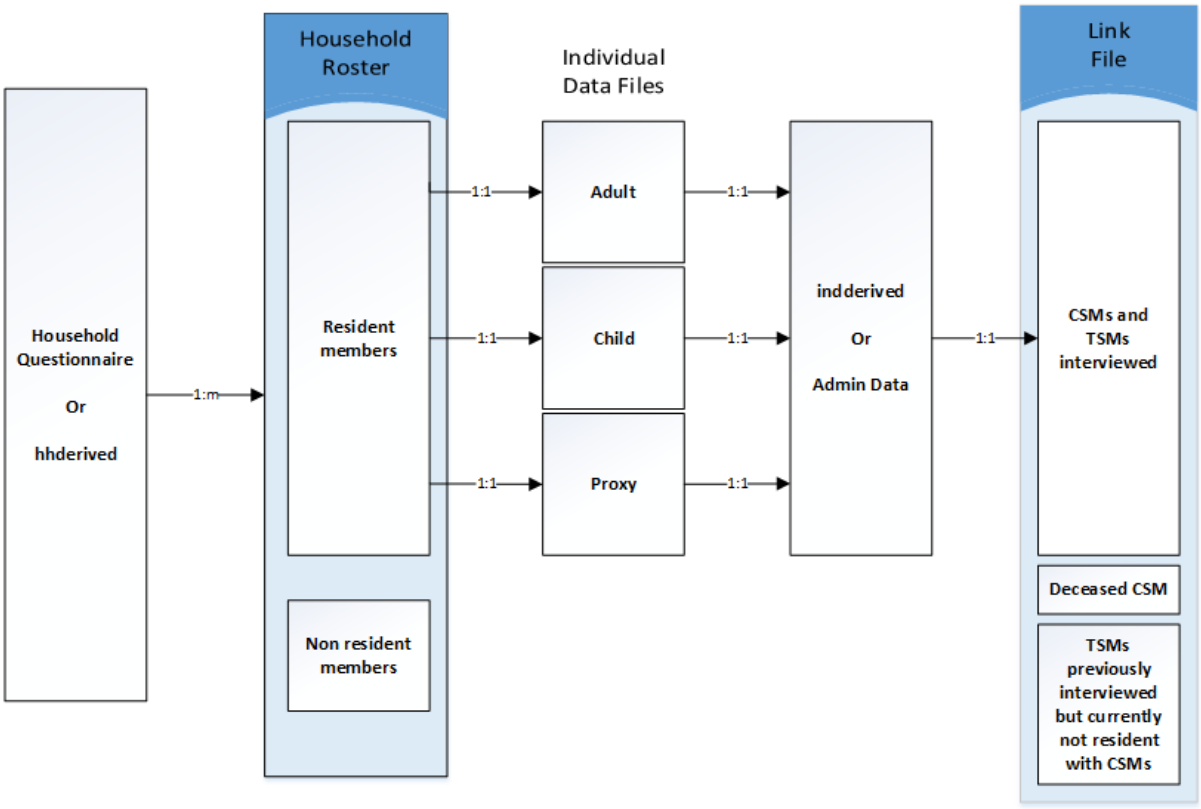
*Diagram adapted from the HILDA User Manual – Release 14

These numbers and 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.

3.6.2 Merging within Waves

We recommend that merging at the individual level within a wave is done using both *w**X**_hhid* and *pid*. The exception to the rule would be when specifically looking for people who are resident in more than one household, in which case *pid* alone may be used. The roster is the only file where merging with *pid* only will yield different results to merging on *pid* and *w**X**_hhid*. The relationship of the datasets in each wave is shown in Figure 3.2 below.

Figure 3.2: Link of data files within wave



Only one household questionnaire is administered for each household. Each household questionnaire or hhderived file merges to many records on the household roster, as the household roster exists on an individual level. Using the *pid* and *w**X**_hhid*, a one-to-one merge exists when merging the Household Roster to the individual questionnaires (one-to-one relationship is when a single observation in Dataset A will match one and only one other observation in Dataset B). Non-resident members on the Household Roster will not merge to any individual data file. Only residents in a given wave will have records in the indderived or the Admin Data datasets. A one-to-one merge exists when the individual data files are merged to the Link File. When merging the individual datasets to the Link File, CSMs who died and TSMs who were part of the sample in previous waves but not interviewed in the current wave will not merge to any individual file.

3.6.3 Merging Between Waves

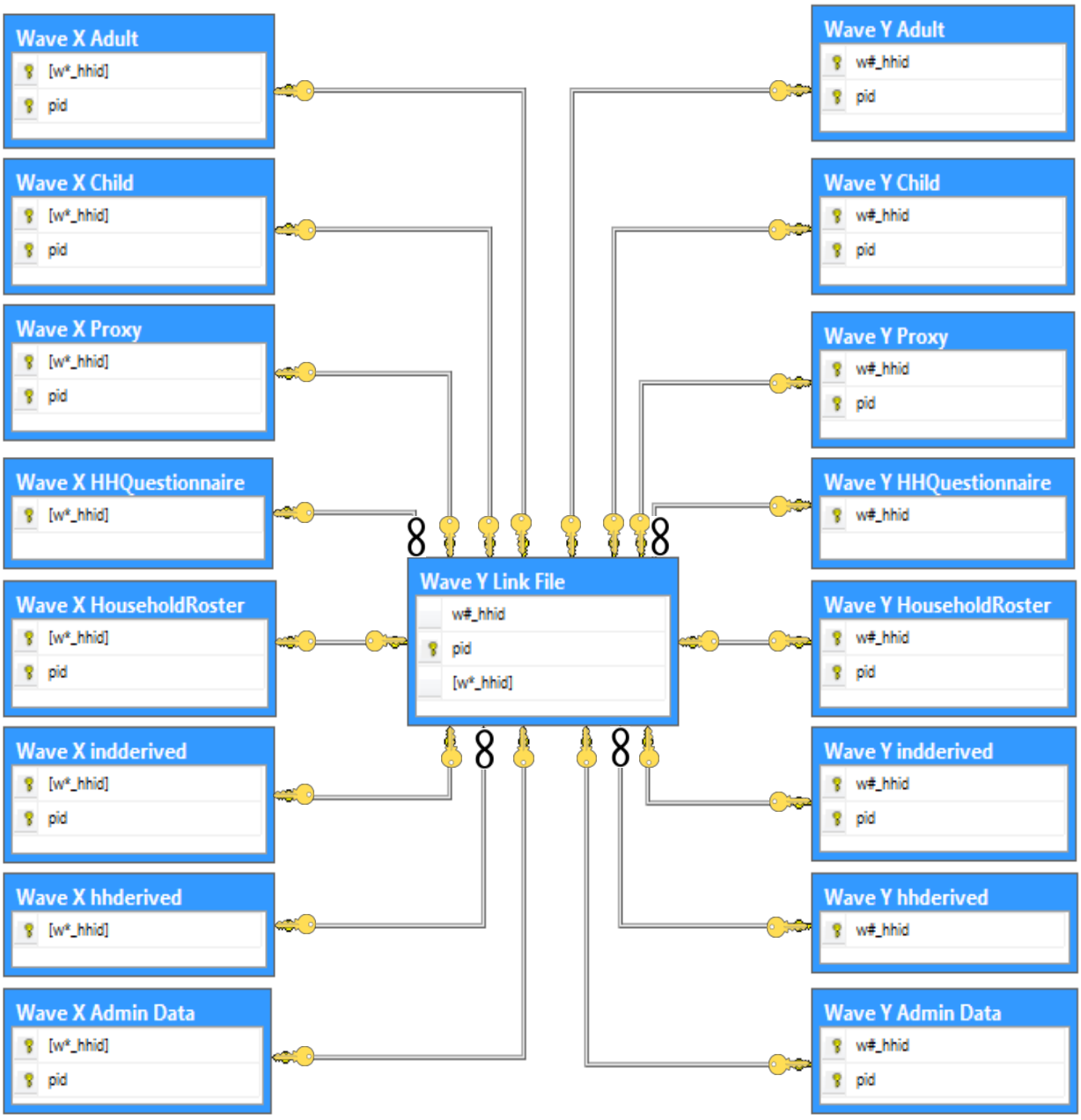
There are two ways to think about merging between waves:

1. NIDS is a panel of individuals, therefore the person identifier (*pid*) is central to merging across waves. Within a given wave, a particular *pid* will not be unique in the roster file if the same individual is a member of more than one household. This prevents a simple one-to-one 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 should 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 one-to-one 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 (*w**X**_hhid*) for all waves. 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. Each wave's data can be merged to the Link File using *pid* and the wave-specific household identifier (*w**X**_hhid*). 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(s) 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.

   Figure 3.3 shows how the Link File may be used to merge datasets between waves.

Figure 3.3: Linking data files between waves



Note: In the above diagram the symbol of the key at one end of the line and a key on the other end represents a one-to-one relationship whereas a key at one end and the infinity symbol at the other end represents a one-to-many relationship.

The latter wave Link File must be used when merging datasets between waves as it contains all information of the current and previous waves. In Figure 3.3 above, given that Wave Y was conducted after Wave X, the Wave Y Link File will be used to merge the datasets. Since NIDS is a panel that follows individuals, the household identifier for the same *pid* will be different across waves. The *pid* and the wave specific *hhid* for each wave should be used to merge to the Link File. As an illustration: Figure 3.3 shows that we can use *w#_hhid* and *pid* to merge the Household Roster dataset in Wave X to the

Link File. Once this is done, *w*_hhid* and *pid* can be used to merge the Household Roster in Wave Y to the Link File. Individual datasets (Adult, Child and Proxy) can be merged to the Link File using the *pid* which is a unique identifier in these data sets. Merging the Household Questionnaire to the Link File results in a one-to-many relationship (each *hhid* will be related to many rows in the Link File) since the Link File is on a *pid* level.

3.7 Variable Naming Convention

Variables are named consistently across waves for ease of reference. Where questions are identical across waves the core of the variable name will be the same.

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 / sub question

Details of each component are described below:

3.7.1 Wave

The wave prefix indicates in which wave the data was collected, e.g. *w1_* indicates Wave 1, *w2_* indicates Wave 2, and so forth.

3.7.2 Source

The source indicates which dataset the variable belongs to. See Table 3.1 below.

Table 3.1: The source indicators

Source Indicator	Meaning
A	Adult file
C	Child file
P	Proxy file
H	Household file
R	Household Roster file

3.7.3 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 are shown in the Table 3.2 below.

Table 3.2: Examples of significant section leaders

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

3.7.4 Subsections

The subsections are used for grouping similar questions. There are a number of subsections to many of the main sections. Examples include:

Within Employment:

Table 3.3: Example of employment variable names

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

Within Education:

Table 3.4: Example of education variable names

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:

Table 3.5: Example of health variable names

Ailments in last 30 days	hl30	Lifestyle	hllf
Recent consultations	hlcon	Smoker	hllfsmk
Vision	hlvis	Difficulty of activities	hldif

3.7.5 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.

3.7.6 Subquestions

Note that the subquestion is not a descriptor. Subquestions only qualify a previous question, with a finite number of qualifying properties, such as location, value or explanation. A subquestion 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. *w**X**_a_empsll_v*). This variable is classified as a sub question of the question "Do you sell produce?", and receives the suffix "_v".

3.8 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. The codes are detailed in Table 3.6 below.

Table 3.6: Non-response codes

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.

3.9 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).

3.10 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.

3.11 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 the Program Library section of this User Guide for a detailed list of these files.

4. Data Collection

Data collection periods for all waves are as follows:

Table 4.1: Interview dates

	Start	End
Wave 1	February 2008	December 2008
Wave 2	May 2010	September 2011
Wave 3	May 2012	December 2012
Wave 4	September 2014	August 2015

Every effort has been made to be consistent in the data collection methodology applied across waves, while also paying attention to being more efficient in field operations. From Wave 2 onwards, all data have been collected using Computer Assisted Personal Interviewing (CAPI) software, which has been extended and improved upon over time. Use of paradata to monitor interviewer performance has also been developed in order 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 monitoring of fieldworker behaviour during field operations and other quality control measures taken.

4.1 Data Collection Process

In each wave, four types of questionnaires are administered:

- **Household questionnaire**: One Household questionnaire is completed per household by the oldest woman in the household or another person knowledgeable about household affairs and particularly household spending. Household questionnaires take approximately 39 minutes in non-agricultural households and 50 minutes in agricultural households to complete.
- **Adult questionnaire**: The Adult questionnaire is applied to all present CSMs and other household members resident in their households that are aged 15 years or over. This questionnaire takes 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) is taken on their behalf with a present resident adult. On average, a Proxy questionnaire takes 12 minutes to complete. Proxy questionnaires are also asked for CSMs who have moved out of scope (out of South Africa or to a non-accessible institution such as prison), except if the whole household has moved out of scope, and can therefore not be tracked or interviewed directly.
- **Child questionnaire:** This questionnaire collects information about all CSMs and residents in their household younger than 15. Information about the child is gathered from the care-giver of the child. The questionnaire focuses on the child's educational history, education, anthropometrics and access to grants. This questionnaire takes an average of 16 minutes per child to complete.

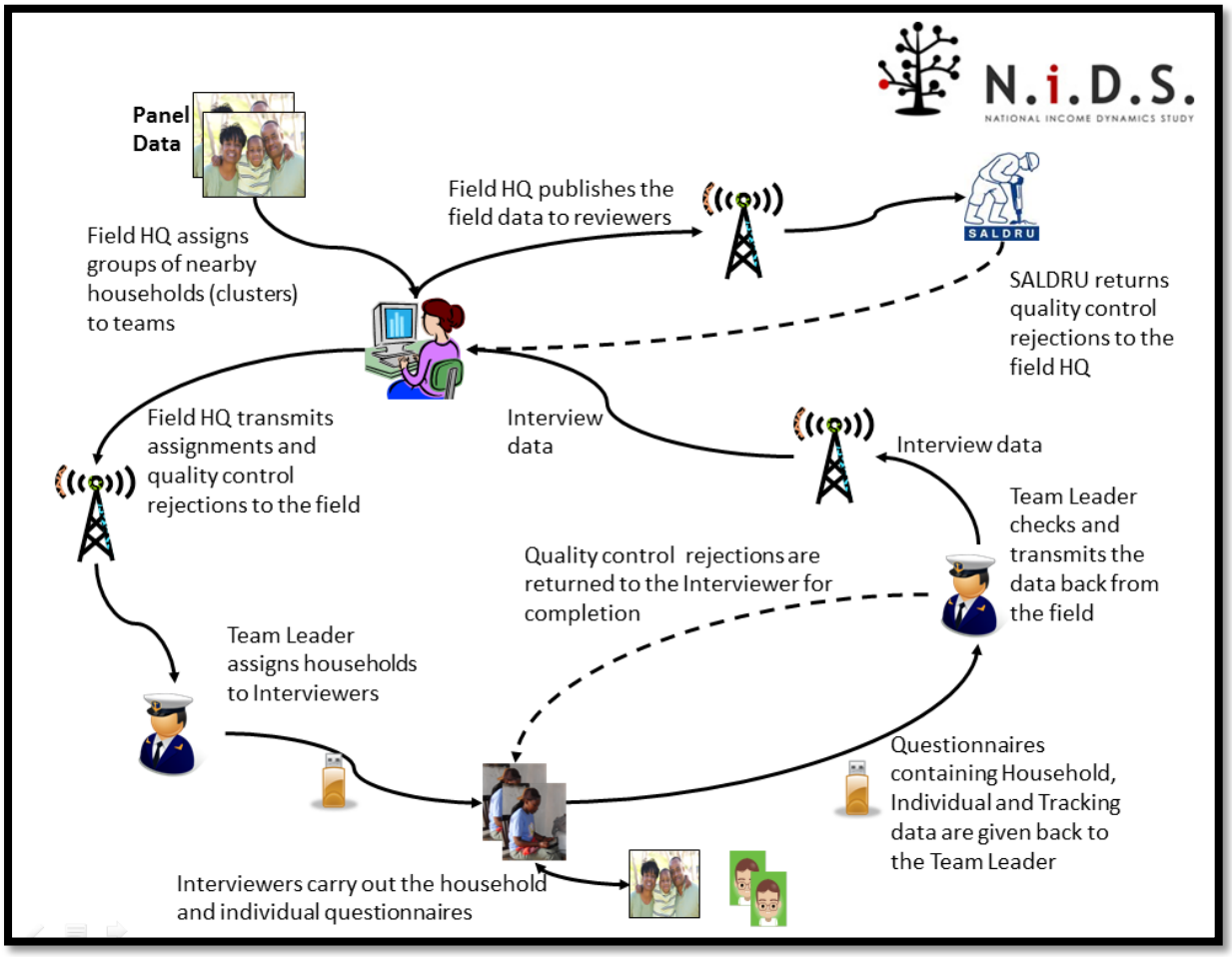
Paper consent forms are issued in all languages and the informed consent process is conducted in the respondent's language of choice. For each questionnaire, two consent forms are signed. One signed

copy remains with respondents and the other is returned to SALDRU. These forms carry unique bar-coded numbers that are entered into the CAPI system. Similarly, the household and person level IDs are displayed on the CAPI system and written onto the consent forms so that cross-referencing is possible. Data coming in from the field are accepted as valid only if SALDRU has a signed consent form for each interview that produced the data. If signed consent forms are not located, the associated interviews are deleted from the dataset.

4.1.1 Overview of CAPI Cycle

The CAPI cycle is illustrated below.

Figure 4.1: The CAPI cycle



Listing data (PSUs, household addresses, contact details, roster make up and individual contact details) drawn from the previous wave is pre-loaded into the CAPI system. Respondents who were not located in the previous wave are listed with the area and household information from the wave in which they were last observed, in order to allow fieldworkers to reattempt to gather information about them. This process allows CSMs to re-enter the sample when they would otherwise have been lost due to insufficient information collected during the previous wave. Listing data is centrally distributed via modems to field teams on a cluster by cluster basis prior to their arrival.

Also included are 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 are 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 is sometimes not displayed, but is used by the CAPI system to challenge inconsistent answers (e.g. attendance at school during the previous wave). Where answers are inconsistent with data previously collected, the interviewer is challenged to confirm the answer and enter substantiating notes for the change.

Certain pre-populated data are used to skip questions if valid and consistent answers had been discovered in multiple previous waves, an example being head circumference of a child at birth.

The fieldworkers conduct the surveys and validate the content using tablet computers. Field Team Leaders then re-validate the fieldworker data prior to transmission back to NIDS (SALDRU in the diagram above).

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

4.1.2 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 can either be in the same location as they were in the previous wave (or the wave in which they were last located) or they could have moved. Interviewers use the CAPI system to load address and contact details for movers (either "Whole Household Moved" or "Household Splitters"). The field team leader then assesses these details to:

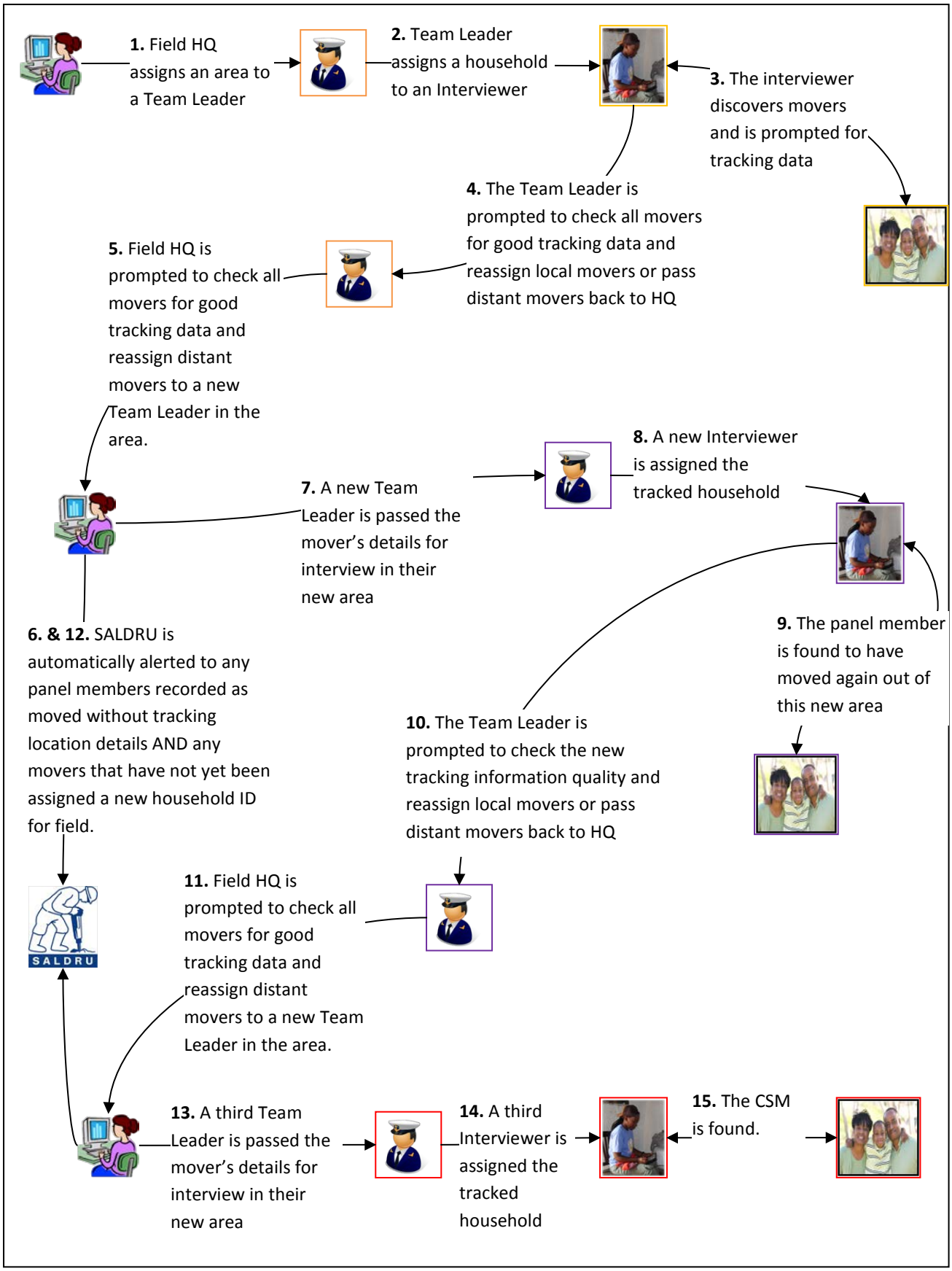
1. Generate new household IDs locally containing the movers to be dealt with by that team; or
2. Transmit the location details back to field control to generate household identifiers for movers and assign them to the relevant team on a geographical level.

Households are created around these location details which are indexed and linked to respondents. A household ID is 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 are finalised only after the location of the CSM is confirmed.

Where no useable data is available for movers, household and person records are moved to a dummy PSU signifying lost in tracking. In these cases SALDRU examines the location information available and the contact details of the originating household in an attempt to improve or verify the mover details. Where this is successful, these households are sent "back to field" for completion. By making use of the extensive family networks represented in the Panel Maintenance System, the SALDRU office team is 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:

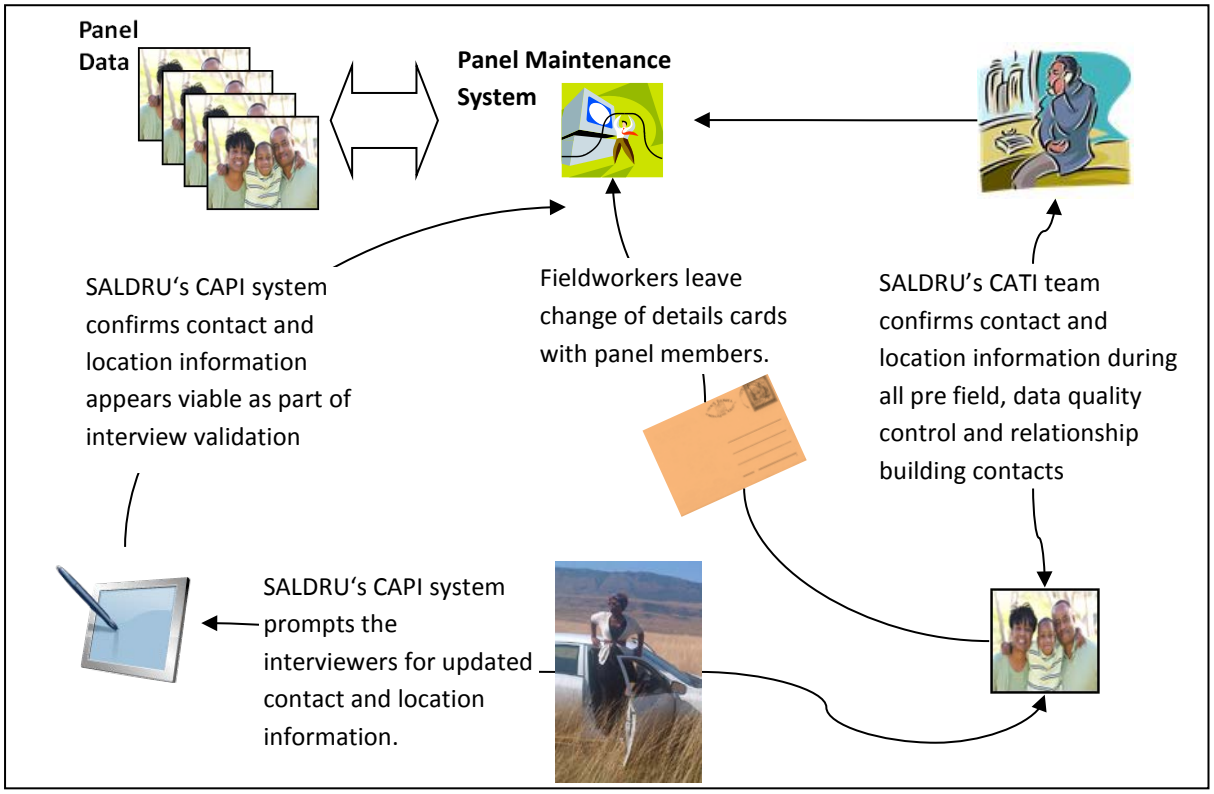
Figure 4.2: Tracking movers



4.1.3 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 4.3: 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 prepaid change of address card is left with panel members).

4.2 Data Quality Issues and Data Collection

Data quality issues that arise and are mitigated in the data collection process include the following:

4.2.1 Unit Non-Response

Unit non-response is 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-explain what the survey is about and the value of the respondent's contribution. Similarly, written records are left with respondents about their anthropometric data including whether to seek medical advice over their blood pressure readings; anecdotal evidence is that this information is highly prized by respondents. SALDRU also carries 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 is encouraged as much as possible.

2. **Tracking systems:** The CAPI software carries 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 is also done in an effort to minimise non-contact.
3. **Field status for temporarily away respondents:** since Wave 3, a "temporarily away" status for households has been included in the system. This catches instances where no one is at a dwelling but it is discovered that they will return within the fieldwork period (but not while the team is currently in the relevant cluster). These dwellings are then 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 are interviewed and the number of "no one at home" respondents contains a smaller proportion of these respondents than is the case for Wave 2.
4. **Household level non-response call backs:** Households may come back from field as a refusal, dwelling-unit vacant or un-locatable/un-traceable. Households that came back from field as refused are contacted by SALDRU to confirm the refusal and attempt to overturn it; where a refusal is overturned these are returned to the field company for re-interview. Where the field organisation fails to track individuals, SALDRU investigates further using the history of co-residents and alternative contacts for movers. Operationally, this is done through the NIDS call-centre with the Panel Maintenance System.
5. **Individual level non-response call backs:** SALDRU attempts to contact all individual level refusals to confirm the refusal and attempt to overturn it; where a refusal is overturned these are returned to the field company for re-interview.
6. **Field organisations rewards:** Field company bonus schemes and targets have been structured so as to encourage better completion and lower attrition during fieldwork.
7. **CAPI pre-population:** Pre-populating the CAPI roster along with the automatic insertion of the relevant names into the individual's questions ensures easy monitoring that all CSMs are being approached and that the correct roster members are being referred to in their individual questionnaires.
8. **No one at home policy:** Should there be no one at a dwelling, the interviewer is 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-response.

4.2.2 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. A strict policy is in place such that data is only accepted from field if all sections have been completed. There is a system for accepting exceptions, but each

exception has to be approved by SALDRU staff. Any questionnaires submitted that are not completed correctly and which do not have an exception raised are returned to field for completion.

4.2.3 Data Consistency

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

1. **Translation, respondent understanding and measurement error:** The CAPI system holds 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. To reduce interviewer effects, SALDRU makes some use of the context sensitive help afforded by the use of CAPI.
2. **CAPI consistency checks:** The CAPI system has 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 are 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 is prompting the interviewer if schooling appeared to have advanced too far between waves. All of these checks are 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 are closely monitored during field. This also reduces the interviewer effects. The following areas are 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 are usually taken periodically from about 6 weeks into fieldwork (or when there is enough data to estimate meaningful averages). Where interviewers' performance measures lay outside of ±50% of the mean they are investigated, retrained, moved to different teams for closer supervision or removed; in some cases the households are re-interviewed to include hitherto missed respondents. The nature of the measures used and their commencement date therefore need to be considered when addressing issues of interviewer effect.
4. **Within wave and across wave consistency checks in office:** SALDRU carries out a range of pattern searches and consistency checks on the data during field to identify interviewer effects and possible miscapture. When areas of concern are found, the respondents/households are contacted to ensure that the data are correct. If a call-back is successful the data collected during the call-back are used to correct the information collected infield. If the query is across waves it

could result in a change of data for a previous wave. If the call is unsuccessful, the conflicting information is left 'as is' in the data. A number of key variables (gender, 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.

5. **Live behavioural correction:** The use of CAPI allows live checking of data quality from the commencement of field. Through returning data "back to field" for recollection in a timely fashion, NIDS is able to mitigate and normalise the most obvious interviewer effects.

4.2.4 The Mechanics of Data Quality Checks

In this section we discuss three main data quality checks that are 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 allows the interviews to be downloaded by SALDRU in real time, the data quality process can commence in real time.

4.2.4.1 *Early identification and cleaning of identifier mismatches*

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

The cleaning at this level consists of ensuring identifiers for these files are correct and consistent. Identifier mismatch typically arise from:

- Erroneous moving of households, which creates new household identifiers when in fact the household remained intact and at their original physical address. In these cases the household identifiers are returned to their original household ID.
- Mover CSMs splitting from differing households but moving in together, which creates the situation of one CSM being recorded as a TSM (the new household having been created around the other splitter). This happens very infrequently.
- CSMs who split from their household in one wave and then return to that household in a later wave. In the CAPI system a new record gets created for the returned CSMs. Through careful identification of likeness within household dynasties, such cases can be identified. Sometimes the identification takes place before the fieldwork company attempts to track the original CSM and they can be informed that it is no longer necessary to track that respondent.
- Conversely, there is the need to identify people who are incorrectly identified as a CSM when in fact the wrong person has been interviewed. Where these cases are identified during field they are returned to the fieldwork company to attempt to interview the right person.

Identification of these problems occurs through:

- Automatic checks built into the flat file creation process that highlight 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 is 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 are built into the flat file production code for the next daily CAPI data upload.

4.2.4.2 *Returning incorrect data "Back To Field"*

A "status" control, visible on the CAPI systems, is used by interviewers and through all management layers. This status system allows more quality control checks to be included in the CAPI system itself, which means more sophisticated checks can be carried out by the SALDRU quality control office.

The CAPI status system automatically rejects 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[9].
- Invalid "No one at home". Field teams have to demonstrate that they have 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 then checks 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.

"Invalid non-response" is where the SALDRU team attempts to call all non-response households to ensure that the field teams have tried enough times to get hold of the respondents, refusals are genuine or that households could really not be contacted or physically located. If the SALDRU team gets in contact with the respondents and they are willing to participate in the survey then these are returned as "back to fields" to the field company in the form of an exception report.

If a questionnaire is deemed invalid by SALDRU's data quality checks , it is marked as rejected in the CAPI systems and therefore sent "back to field" and a further in-person interview is required (i.e. telephonic interviews are also not permitted in resolving "back to field" issues).

4.3 Fieldwork Schedule

4.3.1 Pre-Test

As part of the preparations for fieldwork a full system pre-test is conducted that acts 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-tests in previous waves, all aspects of the panel and pre-population can be tested. The pre-test tracking initially included 586 individuals from 160 households. These households originated in 8 clusters (4 in KwaZulu-Natal, 3 in

[9] Valid unit non-response outcomes – Refused, No one at home.

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.

4.3.2 Main Data Collection

Fieldworker training is generally conducted at the same time to ensure the highest amount of consistency. Typically, there are in excess of 100 fieldworkers who operate in teams of 4, comprised of 1 team leader and 3 interviewers. Occasionally team sizes vary depending on the region and/or typical household characteristics for that area.

Typically fieldwork is completed within one calendar year. For waves conducted across two years, all questions refer to the actual year in order to avoid confusion. In the case of multi-year data collection, it is advised to pay attention to the date of interview variables (*w**X**_intrv_y*) to understand the specific year being referred to.

5. Main Data Processes

This section provides an explanation for some of the major section that have been adjusted or improved over time in the NIDS data cleaning process.

5.1. Birth History

To enhance the usability of the NIDS data, Wave 4 saw the allocation of unique identifiers (*bhchild_id**) to each child on the birth history. This is to assist with the process of identifying children across waves. Previously, only children who were members in the household had identifiers assigned to them.

The process of allocating each child with an identifier is performed by algorithmically matching children across waves. Fuzzy string matching is used for string variables along with direct comparison of numeric variables, such as dates of birth and gender. In cases were the birth history is inconsistent across waves, calls are made by the NIDS Call Centre to determine the children the respondent has given birth to. Where the Call Centre is unable to make contact with respondents, information on some birth histories will remain inconsistent across waves. Once the children are determined to be the same child across waves, identifiers are allocated using a two stage process:

1. The same algorithm for identifying wave matches was repeated to match the children using the birth history to the household roster. If a perfect match is established the child is allocated the same identifier as that which is on the roster.
2. The children who do not match any record on the household roster are then randomly assigned identifiers in the second step.

5.2. Parental Data

Wave 4 saw new processes to reduce inconsistencies in the parental information in the data (Adult questionnaire section d, Child questionnaire section e, and Household Roster questionnaire section b) which have made the use of parental variables problematic.

We identified cases where inconsistencies existed by comparing parental related variables across waves. Examples of variables which were examined include birth year of parent, death year of parent, and cases where a parent “came back to life” in a successive wave. Where respondents had at least three parental data issues, a call was placed to confirm all the parental data for both parents in each wave across the panel. Once the data was confirmed with the respondents via calls, the respective data was then updated in each wave.

Data of respondents that we could not contact via calls was left unchanged.

5.3. Education Progression

In wave 4, a subsample of individuals who had education progression inconsistencies between any of the four waves were asked an additional set of education-related questions relating to previous waves (i.e. impossible progressions between grades over time). The additional questions include educational activities from 2008 to 2011 (Adult questionnaire: h14.1_ed11att – h17.6_ed08wdex; Child questionnaire: c12.1_ed11att – c15.7_ed08wdex). These additional variables have not been ‘pushed back’ into previous waves corresponding to their respective years but left for the user to decide whether to use them to ‘clean’ previous wave data.

5.4. Pcode Variables in Wave 1 Data

Both the *pcode* and respective *pid* have been released in Wave 1 data since V4.0 in February 2012. From V5.0, released in Sep 2013, non-resident individuals were assigned a *pid* for the first time. Since non-resident individuals now have a *pid*, the *pcode* variable became an unnecessary duplicable identifier. In addition to this, the cleaning process of these identifiers (*pcode* and *pid* variable) became more time consuming due to every *pid* adjustment requiring a *pcode* adjustment. Furthermore, the *pcode* variables were inconsistent with the rest of the panel where *pid* equivalents instead of the *pcodes* were used. Based on the above reasoning, all the *pcode* variables in Wave 1 have consequently been dropped.

5.5. Surveyed vs. Historical Data

In Wave 4 selected variables in the demographics, parental data, and education sections were not re-asked of respondents. This was done to avoid re-asking respondents time-invariant data that we have collected previously. This was only done in instances where we had consistent responses to the questions across waves. In order for users to differentiate between this historical data and the data which was surveyed in Wave 4, flag variables have been created. An example of this is *w4_a_brnprov_flg*.

6. 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:

- 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.

6.1. 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 the respondent is not interviewed for any particular reason. 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, 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 *w**X**_best_age* is calculated within each wave using the best date of birth and the date of interview for that wave.

6.2. Geography

The Global Positioning System (GPS) information is 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.

From Wave 2 onwards, a variable has been defined *(w**X**_stayer)* at the individual level for respondents that remained within 100 metres between Wave 1 and 2 and within 40 metres between Wave 2 and 3 and between Wave 3 and Wave 4. The reason for the shorter distance between the later waves is due to built-in GPS systems being used in these waves which allowed for more accurate GPS coordinates. 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.

6.3. Occupation

The classification of occupations in Wave 1 was initially done using the South African Classification of Occupations (SASCO). In order to provide data on occupations that are comparable across waves, the SASCO codes have been dropped from Wave 1. In place of the SASCO codes, International Standard Classification of Occupations (ISCO) have been adopted to classify occupations according to the job

title and main tasks or duties stated by the respondent. ISCO codes belong to the international family of economic and social classifications which is maintained by the United Nations and are published by the International Labour Organization (ILO) - see http://www.ilo.org/public/english/bureau/stat/isco/. ISCO coding has been used for all four waves for consistency.

A two stage process is used to occupations. Firstly, occupations are automatically grouped together based on the descriptions given to us by respondents into a list of occupational codes found in the ISCO code list. This grouping process is initially done and quality controlled electronically using a fuzzy string matching algorithm, which groups similar words together and matches words incorrectly spelled by the interviewer into likely alternatives. The second part involves hand-coding the descriptions that the algorithm cannot identify by manually reviewing the occupation descriptions and ISCO codes, as well as the work description data given to us by respondents. The codes are then truncated down to the one-digit level and included in the Public Release Data. Disaggregated occupational codes are available in the Secure Dataset.

To highlight the adoption of ISCO in all waves the variables have been renamed to reflect this change as follows:

Table 6.1: Variable naming convention for employment codes

Variable description	Old Variable Name	New Variable Name
One digit level ISCO code	*_c	*_isco_c
Full ISCO code (Available only in Secure Data)	*_fc	*_isco_fc

6.4. Industry

Industry coding is done in two parts, similar to occupational coding. Part one also involves 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 involves hand coding the descriptions that the algorithm could not identify.

These codes are 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.

6.5. Employment Status

Employment Status is 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 is determined to be employed if they are economically active and reported having any form of employment at the time of the interview, including a primary job, secondary job, self-employment, paid casual work, personal agricultural work, or if they assist others in business activities. Unemployment is differentiated into broad and narrow unemployment as per the standard

definitions, by distinguishing those who were actively searching for work and those not actively searching.

6.6. Admin Data

The Admin dataset is a dataset produced by NIDS whereby we match the data we collect in field to external administrative data such as the Master schools list published by the Department of basic education.

6.6.1. School's Admin Data

The Admin datasets contain school level data for individual records where we were able to match the school name collected by NIDS to school names on the Ordinary School's Master List, as available from the Department of Basic Education's website. The matching process is performed by implementing approximate or fuzzy string algorithms, taking the geographic distance between the school and the household into account as well as the schools education phase.

A scrambled school identifier - based on the schools unique EMIS number published by the Department of basic education (DBE) – is included in the anonymised Admin dataset. Descriptive data for the matched schools is also included - such as the quintile, province, no fees school status, phase and the department of education responsible for the governance of the school. The Secure Data contains additional variables describing the number of learners, number of teachers and the learner teacher ratio for each school.

6.6.2. Police Station Data

The 2015 dataset made available by the South African Police Service (SAPS) is included in the data for each wave. The police station data, which is at a household level, was added to the *Admin* data on an individual level. The suffix "15" was added to all the police station variables to indicate that it pertains to the 2015 police station data. Police station IDs (*w**X**_poldistr_id_15*) were generated, as these were not available in the data provided by the SAPS.

Variables include data on the distance to the district police station as well as the straight line distance to the nearest police station. The police IDs and the banded distances generated by NIDS are included in the public version of the data release. Variables included in the Secure Data are the GPS coordinates, the police names, and the numerical distance up to 6 dp from households to their nearest and district police stations.

6.7. Income

Total household income (*w**X**_hhincome)* is 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 is derived in one of three ways. If all adult household resident members were successfully interviewed, *w**X**_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 use the so-called "one-shot" income variable *w**X**_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 aggregate any income data we have from the remaining responding household resident members. Imputed rental income from owner-occupied housing, *w**X**_hhimprent,* is added to all households, irrespective of the method of aggregation, where appropriate. Table 6.2 shows how income was aggregated in all waves.

Table 6.2: Sources of aggregation

Wave Number	Source of HH Income	Number of HHs	Percent
W4	Individual aggregation	8291	86.19
	One-shot	1329	13.81
	Total	9620	100
W3	Individual aggregation	7272	90.53
	One-shot	761	9.47
	Total	8033	100
W2	Individual aggregation	5659	83.38
	One-shot	1128	16.62
	Total	6787	100
W1	Individual aggregation	7097	97.27
	One-shot	199	2.73
	Total	7296	100

Table 6.3 below lists the variables that make up each component of total household income. These variables are located in the indderived data file for each wave.

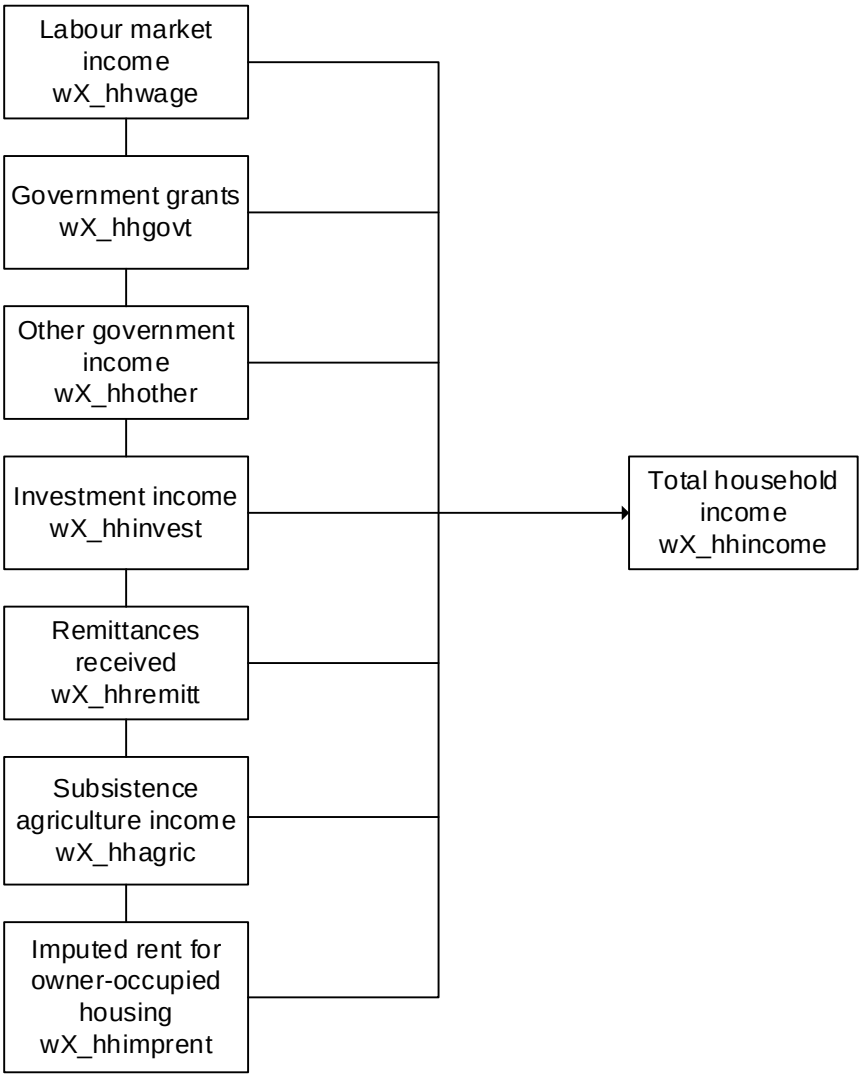
Table 6.3: Components of aggregate household income

Household-level Variable	Individual-level Variable	Variable Name
Labour Market Income wX_hhwage	Main and second job	w**X**_fwag
	Casual wages	w**X**_cwag
	Self-employment income	w**X**_swag
	13th cheque	w**X**_cheq
	Bonus payment	w**X**_bonu
	Profit share	w**X**_prof
	"Help friends" income	w**X**_help
	Extra piece-rate income	w**X**_extra
Government Grant Income wX_hhgovt	State Old Age Pension	w**X**_spen
	Disability Grant	w**X**_dis
	Child Support Grant	w**X**_chld
	Foster Care Grant	w**X**_fost
	Care Dependency Grant	w**X**_cdep
Other Income from Government wX_hhother	Unemployment Insurance Fund	w**X**_uif
	Workmen's compensation	w**X**_comp
Investment Income wX_hhinvest	Interest/dividend income	w**X**_indi
	Rental income	w**X**_rnt
	Private pensions and annuities	w**X**_ppen
Remittance Income wX_hhremitt	Remittances received	w**X**_remt
Subsistence Agricultural Income wX_hhagric[10]	Income from subsistence agriculture	w**X**_plot
	Value of own production consumed	w**X**_opro
Imputed Rental Income wX_hhimprent	N/A	N/A

The seven variables in the first column in Table 6.3 are summed to create aggregate household income. Figure 6.1 shows this aggregation.

[10] Agricultural Income was not used in calculating aggregate household income in Wave 2.

Figure 6.1: Components of aggregate household income



6.7.1. 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 is assigned. Those who indicated that they received income above the value of the highest bracket are assigned twice the value of the upper bound of the top bracket[11].

6.7.2. 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 is conducted only once a few qualifying conditions are

[11] 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.

satisfied. Single imputation regressions are run only when there are: a) 100 or more "valid" responses for a variable and b) the percent of missings does not exceed 40%. Pre-imputation, post-imputation and imputation flags are available in the individual derived and household derived datasets for each variable that has been imputed.

A rule-based imputation process is 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, are 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.

Table 6.4 summarizes the variables imputed, the imputation method used to impute for item non-response and percentage of missings, for Wave 4.

Table 6.4: Wave 4 Income variable imputation

Variable Name	Description	Imputation Method	Wave 4		
			Obs	Achieved	% Missing
w**X**_fwag	Main and secondary wages	Regression	6663	6921	3.73
w**X**_cwag	Casual wages	Regression	1052	1100	4.36
w**X**_swag	Self-employment income	Regression	853	1165	26.78
w**X**_cheq	13th cheque	Regression	130	143	9.09
w**X**_prof	Profit share	None	9	9	0
w**X**_extr	Extra payment	None	26	31	16.13
w**X**_bonu	Bonus income	None	47	55	14.55
w**X**_othe	Other income	None	26	26	0
w**X**_help	Help friend income	None	76	78	2.56
w**X**_spen	State pension	Rule	2809	2937	4.36
w**X**_ppen	Private pension	Regression	259	278	6.83
w**X**_uif	UIF income	None	56	61	8.2
w**X**_comp	Workmen's compensation	None	10	10	0
w**X**_dis	Disability grant	Rule	855	857	0.23
w**X**_chld	Child support grant	Rule	5636	5639	0.05
w**X**_fost	Foster care grant	Rule	353	360	1.94
w**X**_cdep	Care dependency grant	Rule	86	86	0
w**X**_indi	Interest/dividend income	None	32	34	5.88
w**X**_rnt	Rental income	Regression	238	239	0.42
w**X**_remt	Remittances	Regression	2371	2763	14.19
w**X**_hhimprent	Imputed rental income	Regression	6,059	8,111	25.3

Table 6.5 summarizes the variables imputed, the imputation method used to impute for item non-response and percentage of missings, for Wave 3.

Table 6.5: Wave 3 income variable imputation

Variable Name	Description	Imputation Method	Wave 3		
			Obs	Achieved	% Missing
w**X**_fwag	Main and secondary wages	Regression	5270	5546	4.98
w**X**_cwag	Casual wages	Regression	663	681	2.64
w**X**_swag	Self-employment income	Regression	664	830	20
w**X**_cheq	13th cheque	None	69	82	15.85
w**X**_prof	Profit share	None	9	9	0
w**X**_extr	Extra payment	None	6	6	0
w**X**_bonu	Bonus income	None	31	33	6.06
w**X**_othe	Other income	None	36	36	0
w**X**_help	Help friend income	None	47	48	2.08
w**X**_spen	State pension	Rule	2462	2463	0.04
w**X**_ppen	Private pension	Regression	321	341	5.87
w**X**_uif	UIF income	None	46	54	11.11
w**X**_comp	Workmen's compensation	None	14	15	6.67
w**X**_dis	Disability grant	Rule	717	720	0.42
w**X**_chld	Child Support Grant	Rule	4816	4818	0.04
w**X**_fost	Foster Care Grant	Rule	296	303	2.31
w**X**_cdep	Care Dependency Grant	Rule	103	104	0.096
w**X**_indi	Interest/dividend income	None	38	43	11.63
w**X**_rnt	Rental income	Regression	132	134	1.49
w**X**_remt	Remittances	Regression	1128	1308	13.76
w**X**_hhimprent	Imputed rental income	Regression	4,933	6,915	28.66

Table 6.6 summarizes the variables imputed, the imputation method used to impute for item non-response and percentage of missings, for Wave 2.

Table 6.6: Wave 2 Income variable imputation

Variable Name	Description	Imputation Method	Wave 2		
			Obs	Achieved	% Missing
w**X**_fwag	Main and secondary wages	Regression	4008	4320	7.2
w**X**_cwag	Casual wages	Regression	528	541	2.4
w**X**_swag	Self-employment income	Regression	478	648	26.08
w**X**_cheq	13th cheque	Regression	154	227	32.16
w**X**_prof	Profit share	None	19	31	38.71
w**X**_extr	Extra payment	None	63	73	13.7
w**X**_bonu	Bonus income	None	62	82	24.39
w**X**_othe	Other income	Regression	118	120	1.67
w**X**_help	Help friend income	None	51	57	10.53
w**X**_spen	State pension	Rule	2138	2147	0.42
w**X**_ppen	Private pension	Regression	333	361	7.48
w**X**_uif	UIF income	None	47	61	22.95
w**X**_comp	Workmen's compensation	None	5	5	0
w**X**_dis	Disability Grant	Rule	589	598	1.51
w**X**_chld	Child Support Grant	Rule	3440	3447	0.12
w**X**_fost	Foster Care Grant	Rule	230	238	3.36
w**X**_cdep	Care Dependency Grant	Rule	58	59	1.69
w**X**_indi	Interest/dividend income	None	23	26	11.54
w**X**_rnt	Rental income	Regression	82	84	2.38
w**X**_remt	Remittances	Regression	534	679	21.21
w**X**_hhimprent	Imputed rental income	Regression	3,432	5,917	42

Table 6.7 summarizes the variables imputed, the imputation method used to impute for item non-response and percentage of missings, for Wave 1.

Table 6.7: Wave 1 Income variable imputation

Variable Name	Description	Imputation Method	Wave1		
			Obs	Achieved	% Missing
w**X**_fwag	Main and secondary wages	Regression	3542	4492	21.15
w**X**_cwag	Casual wages	Regression	650	728	10.71
w**X**_swag	Self-employment income	Regression	663	951	30.28
w**X**_cheq	13th cheque	None	785	1206	34.91
w**X**_prof	Profit share	None	48	102	52.94
w**X**_extr	Extra payment	None	58	107	45.79
w**X**_bonu	Bonus income	None	341	550	38
w**X**_othe	Other income	None	18	18	0
w**X**_help	Help friend income	None	71	80	11.25
w**X**_spen	State pension	Rule	1972	2108	6.45
w**X**_ppen	Private pension	Regression	220	289	23.88
w**X**_uif	UIF income	None	81	122	33.61
w**X**_comp	Workmen's compensation	None	36	53	32.08
w**X**_dis	Disability Grant	Rule	837	869	3.68
w**X**_chld	Child Support Grant	Rule	2857	3388	15.68
w**X**_fost	Foster Care Grant	Rule	172	182	5.49
w**X**_cdep	Care Dependency Grant	Rule	44	47	6.38
w**X**_indi	Interest/dividend income	None	96	136	29.41
w**X**_rnt	Rental income	Regression	111	125	11.2
w**X**_remt	Remittances	Regression	1140	1140	0
w**X**_hhimprent	Imputed rental income	Regression	2,608	6,237	58.18

6.7.3. 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.

From Wave 3 onwards, the Household questionnaires differ 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 is calculated from the Household questionnaire. The aggregated value of all crops and/or animals harvested or consumed by the household forms the measure of this

income source. The process used is similar to that applied in Wave 1. This is deemed as the best estimate 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 has been calculated.

6.7.4. Bonus Payments

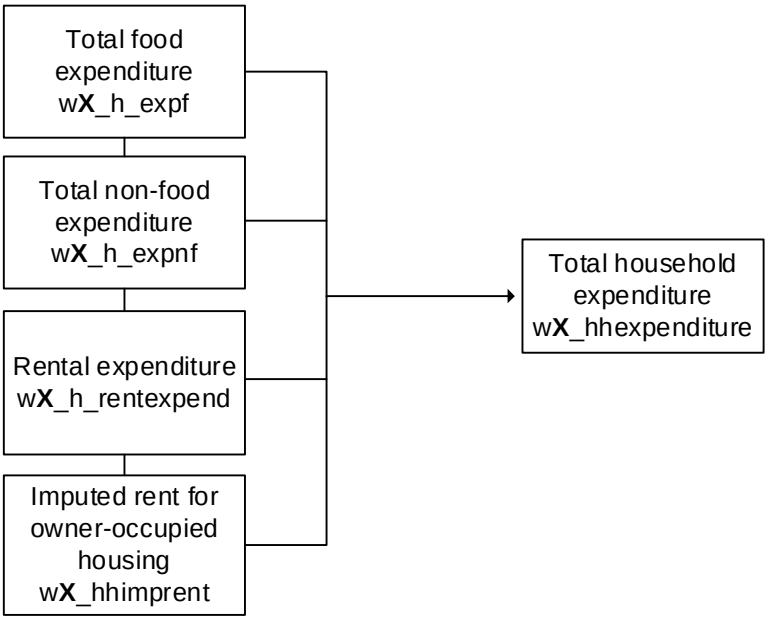
In the first wave, respondents were asked about the value of 13th cheques, profit shares and bonus payments received in the past 12 months. This amount was then divided by 12, to reflect an "average" monthly amount. In the Wave 2 Adult questionnaire, respondents were asked about receiving these sources 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 and Wave 4 asked for both annual and monthly amounts, and the latter was chosen so as to be consistent with Wave 2.

6.8. Expenditure

All expenditure data come from the Household questionnaire. The respondent answering the Household questionnaire is asked about total household expenditure in the last 30 days for each of food and non-food items. These are summed to provide total food expenditure (*w**X**_h_expf*) and total non-food expenditure (*w**X**_h_expnf*), respectively. These two components are added to total rental expenditure (*w**X**_h_rentexpend*) and imputed income from owner occupied housing[12] (*w**X**_hhimprent*) to constitute aggregated total household expenditure (*w**X**_h_expenditure*).

[12] Imputed rental income from owner-occupied housing is 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).

Figure 6.2: Components of aggregate household expenditure



6.8.1. Imputations

There are 4 categories for imputation of expenditure. These are discussed below.

6.8.1.1. Food

If a respondent indicates that the household purchased one of the food items in the last 30 days, but cannot give an expenditure amount, this value is imputed using a single regression imputation approach. If a household is unable to provide a value for any of the food items, the "one-shot" food expenditure is used, rather than an aggregation over all the food line items. We maintain the rule-of-thumb that imputation only takes place when there are at least 100 recorded observations and missings do not exceed 40%.

In Wave 1 and Wave 2, we asked for both the 'one shot' food expenditure amount and the details on all food expenditure items.

Since Wave 3, we asked for detailed food expenditure only if the

1. Household didn't answer the 'one shot' food question or the 'one shot was suspicious in that it was less than 5% or more than 80% of total household income**.**
2. Both the 'one shot' and the bracketed questions were non-response.
3. Household received food as payment or ate from own stock or grew it themselves.

Because of this new rule in Wave 3 and Wave 4, one would expect the number of missing observations to be the same for each food item in cases where the 'one shot variable is reported.

Table 6.8 shows how food expenditure was aggregated in all waves.

Table 6.88: Sources of aggregation

Wave Number	Source of HH Expenditure	Number of HHs	Percent
W4	One-shot	8634	89.75
	Aggregated from food items	955	9.93
	Imputed (One shot)	31	0.32
	Total	9620	100
W3	One-shot	6589	82.02
	Aggregated from food items	1255	15.62
	Imputed (One shot)	189	2.35
	Total	8033	100
W2	Survey (One shot or Aggregated from food items)	6345	93.56
	Imputed	62	0.91
	No Data	375	5.53
	Total	6782	100
W1	Survey (One shot or Aggregated from food items)	7250	99.37
	Imputed	46	0.63
	No Data	0	0
	Total	7296	100

6.8.1.2. Non-food

If a respondent indicates that the household purchased one of the non-food items in the last 30 days, but cannot give an expenditure amount, this value is imputed using the same single regression imputation approach.

6.8.1.3. Rental expenditure

Missing values for households that rent the dwelling unit that they live in are imputed using a single imputation approach.

6.8.1.4. 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.

6.9. Wealth

The wealth section appears in Wave 2 and Wave 4 only. In this section we describe the derivation of household wealth (in both Wave 2 and Wave 4) and individual wealth (in Wave 4 only).

We define household's net worth as household (individual) assets less household (individual) debts. This concept of household net worth is spread over six different asset types, namely: net financial

wealth, net business equity, net real estate equity, value of vehicles, total value of pension/retirement annuities and livestock wealth. Individual net worth is spread over the first 5 asset types and excludes livestock wealth. The reason for this is that livestock net worth cannot be apportioned among household members as we do not know who in the household owns the livestock. A broader definition of each of these terms is provided below:

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

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

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

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

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

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

6.9.1. Wealth in the Household and Adult Questionnaires

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

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

- **<u>Household questionnaire</u>**
  - Section F2 establishes whether the household would be in debt, breakeven or have something left over if the home and all major possessions were sold, all investments were turned into cash and all debts were paid off.
  - If something would be left over, then we ask for the Rand value. If respondents refuse or don't know, then a series of unfolding brackets kicks in.
  - If the household would be in debt, then we ask for the Rand value of that debt. Once again, if the respondent refused or did not know, a series of unfolding brackets kicks in.
  - Section H8 asks about the value of livestock in the household's possession, over seven categories of animals.

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

Section D asks for:

- The amount of bond still owing on the property if it is owned by a member of the household.
- A reasonable value for which the household could be sold.
- A reasonable market value for which all other properties owned by the household could be sold for.
- The total value of bonds that are still owing on all other properties owned by resident household members.

- **Adult questionnaire**

  Section E establishes:

  - Whether the respondent would be in debt, breakeven or have something left over if all business assets and investments were turned into cash and all debts were paid off; and
  - How much money would be left over; or
  - How much debt would be left over?

  Section G asks about:

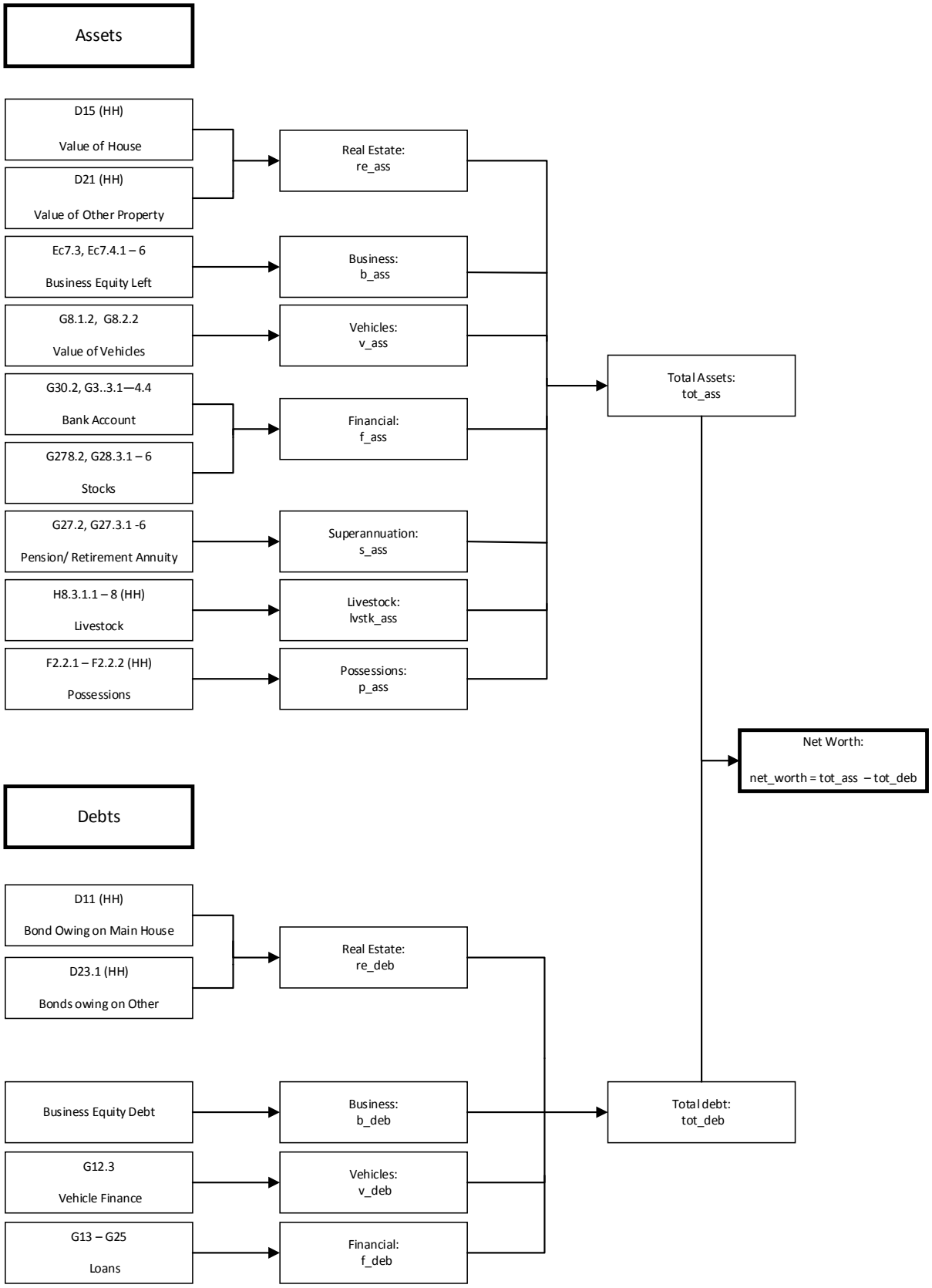
  - The value of all motor vehicles, bakkies/trucks and motorbikes owned by the respondent.
  - Home loans/bonds.
  - A variety of other assets and debts, such as personal bank loans, store cards and study loans, etc.
  - Vehicle finance.
  - Life insurance and unit trusts/stocks/shares.
  - Pensions/retirement annuities.

6.9.2. Imputation

Where a household acknowledges an asset or a debt, but is unable to provide a value, we impute using a single equation imputation regression approach. Our rule-of-thumb requires the number of reported observations to be 100 or more, and for the percentage of missing values to be at 40% or below.

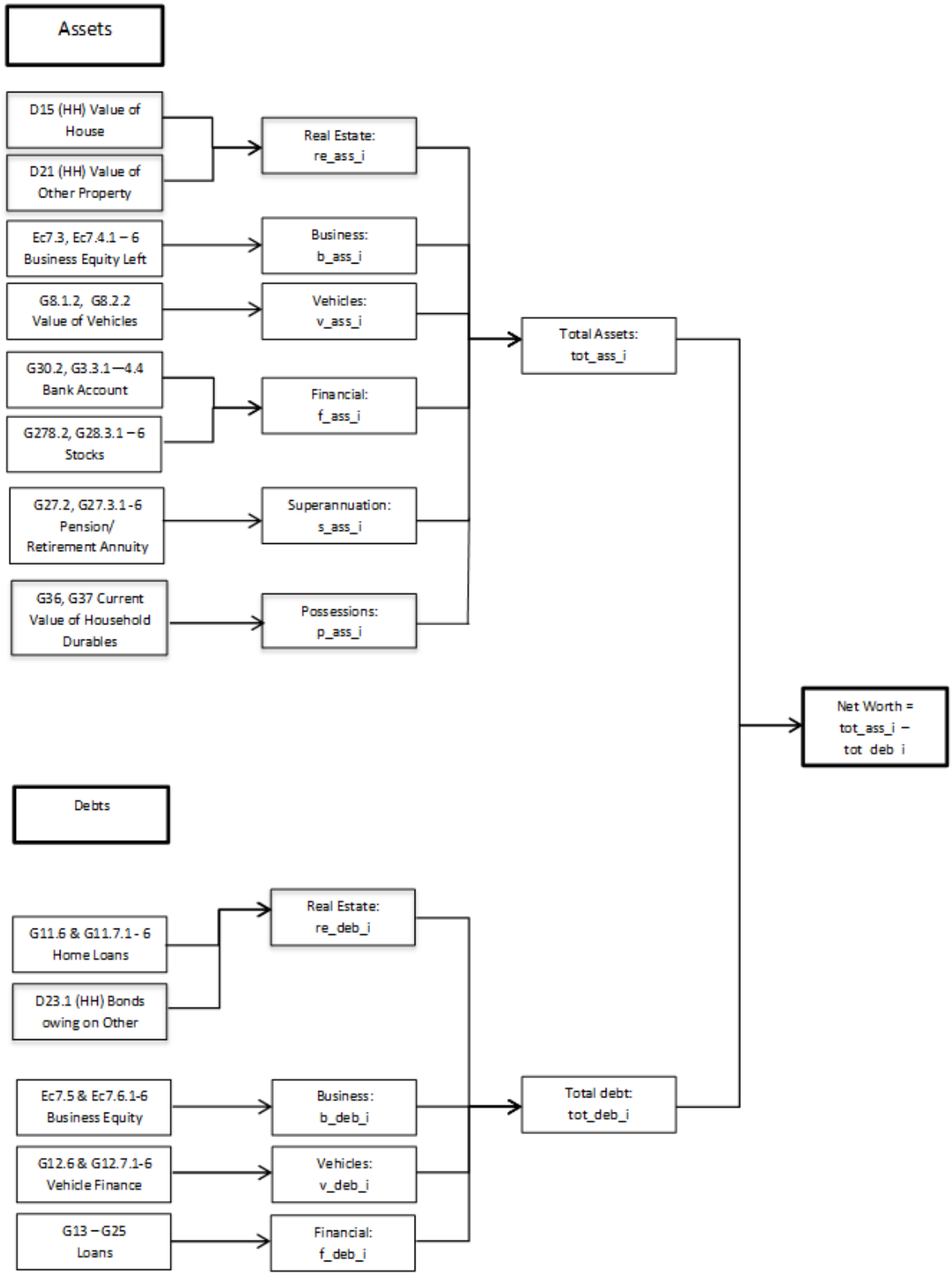
Figure 6.3 and Figure 6.4 below outline how the final net worth for each household and individual is calculated, respectively.

Figure 6.3: Components of aggregate household wealth



Note: Question numbers in the diagram above (e.g. D15 (HH)) refer to the Wave 4 question numbers. These may differ for other waves.

Figure 6.4: Components of aggregate individual wealth



Note: Question numbers in the diagram above (e.g. G30.2) refer to the Wave 4 question numbers. These may differ for other waves.

A number of additional things to note in the calculation of individual net worth in Wave 4 are as follows:

- Livestock wealth is not included in the calculation for individual wealth as we do not have information on who in the household owns the livestock or part of the livestock. However, livestock wealth forms part of household wealth.
- Real estate assets and debts were apportioned to members according to percentage ownership of assets and debts respectively. Questions of percentage ownership where first introduced into the household questionnaire in Wave 4.
- Financial assets in Wave 4 included bank account balance and value of stocks (unit trusts, stocks and shares) as shown in the diagram above. However, in Wave 2 financial assets included cash balance and life insurance policy value in addition to bank account balance and stocks (unit trusts, stocks and shares) value. The cash balance questions was removed in wave 4 as it was sensitive questions for respondents. The life insurance value question was also removed in Wave 4 for the following two reasons: many respondents had no idea of the value of their insurance policy and many tended to confuse this life insurance value with the pay-out value.
- In Wave 4, the bank account value question had some negative values which translated into an overdraft. These negative values formed part of financial debt. In addition, individuals who did not know their bank account balance had the option of answering the unfolding brackets which included negative ranges. No negative values were collected in Wave 2.

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

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

6.10. Anthropometric Z-Scores

Anthropometric measures are collected using the Health information sheet in Figure 6.5 below.

Figure 6.5: NIDS health information sheet

N.i.D.S.
NATIONAL INCOME DYNAMICS STUDY

National Income Dynamics Study

Wave 4 (2014)

Information Sheet

YOUR PHYSICAL MEASUREMENTS

Respondent's Height	_____ _____ _____ centimetres
Respondent's Weight	_____ _____ _____ kilograms
Respondent's Waist	_____ _____ _____ centimetres

Blood Pressure reading 1	**Blood Pressure reading 2**
SYSTOLIC _____ _____ _____	SYSTOLIC _____ _____ _____
DIASTOLIC _____ _____ _____	DIASTOLIC _____ _____ _____
PULSE _____ _____ _____	PULSE _____ _____ _____

☐	Our readings of your blood pressure are within the normal range. (Systolic less than 140 and Diastolic less than 90)	
☐	Your blood pressure readings are higher than normal. High blood pressure is dangerous because it makes the heart work too hard. High blood pressure increases the risk of heart disease and stroke. High blood pressure can also cause other problems, such as heart failure, kidney disease, and blindness. You can control high blood pressure by taking action.	
	☐	It is recommended that you should seek medical care within 2 months. (Systolic 140 to 159 or Diastolic 90 to 99)
	☐	It is recommended that you should seek medical care within 1 month. (Systolic 160 to 179 or Diastolic 100 to 109)
	☐	It is recommended that you should seek **medical care immediately**. (Systolic more than 179 or Diastolic more than 109)

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 are 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) are 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* are used to calculate the z-scores and are available for downloaded from www.who.int/childgrowth/software/en/.

The following variables were created:

*w**X**_zhfa* - height or age for individuals up to 19 years of age
*w**X**_zwfa* - weight for age for individuals up to 10 years of age
*w**X**_zwfh* - weight for height for individuals up to 5 years of age
*w**X**_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 z-scores, we assume 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 assume that the measured height is standing height. Age in days is used to calculate the z-scores.

NIDS fieldworkers are instructed to take two height measures and then a third if the first two measures are more than one centimetre apart. Similarly, a third weight measure is required if the first two weight measures are more than one kilogram apart. In practice, the third measures are very seldom taken. For calculating z-scores, we use the average of the first two measures. In instances were these first two measures differ by more than one centimetre in the case of height and one kilogram in the case of weight, we use the third measure if it is available.

6.10.1. 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 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 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. Using the WHO macros with age measured in days, it has been estimated that the prevalence of stunting among children aged 2 to 10 years is approximately 17%, while the corresponding estimates using the zanthro macro with age measured in years is around 8%. The underestimation from using zanthro is most pronounced at the youngest ages.

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 (and including) -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.

6.11. Weights[13]

6.11.1. What is New?

Together with Wave 4 of the National Income Dynamics Study, updates to Waves 3, 2 and 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. Furthermore, the way deceased respondents are handled has been adjusted. While the initial calculation of weights included a correction for people who died, this is conceptually wrong, therefore the new set of weights only correct for non-response.

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 also be consulted 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 (StatsSA) in 2015. 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, 2010 and 2012 will now gross up to slightly larger totals than before.

6.11.2. The relationship between the different weights

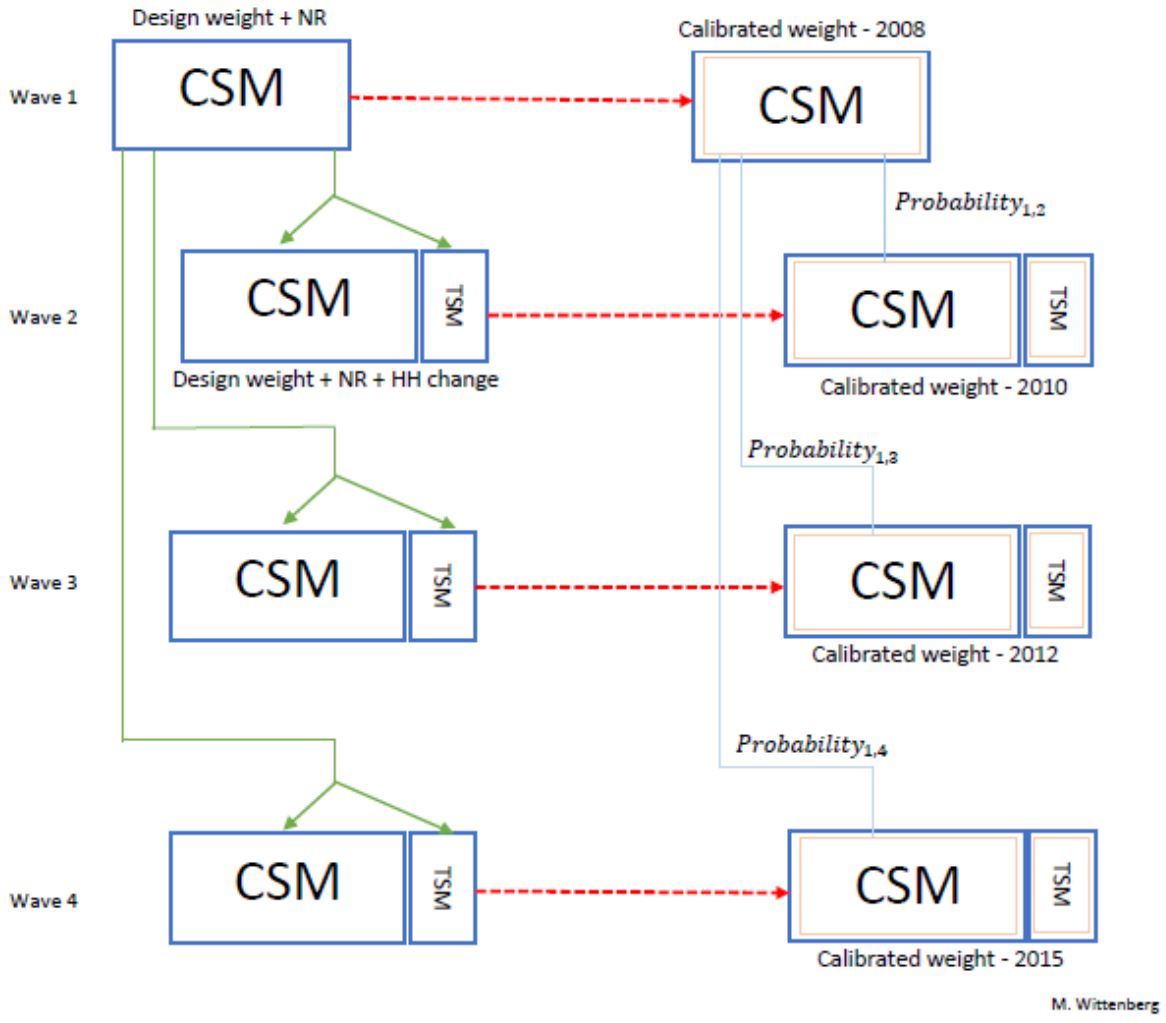
It can be rather difficult to keep track of all the different types of weights that there are in the National Income Dynamics Study. Figure 1 presents the relationships in diagrammatic form.

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

[13] This section was drafted by Martin Wittenberg.

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

Figure 6.6: The relationship between the different weights in NIDS



The design weights released with Wave 1 are fundamental to every other weight released with NIDS[14]. They are used to calculate the corresponding design weights for waves 2 and 3 (the green arrows in Figure 6.6).

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 2015. This process is indicated in the diagram by the red arrows.

[14] 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.

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$_{1,X}$ –** This is the probability of observing an individual from Wave 1 (i.e. one of the CSMs) again in Wave X where X is 2, 3 or 4

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.

It should be noted that only panel weights that correct for attrition between Wave 1 and Wave X (X=1, 2, 3 or 4) are included. In other words panel weight such as the one between Wave 2 and 3 are excluded. This is done to keep the number of weights manageable going forward. Users are welcome to create panel weights that correct for attrition between intermediate waves. When these weights are calculated it should be noted that attrition of TSMs between waves (e.g. Wave 2 and Wave 4) 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 waves are therefore conceptually much messier than the corresponding weights for CSMs[15] .

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

6.11.3. Design Weights

The individuals interviewed in waves 2, 3 and 4 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, 3 or 4 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.

[15] Note that if one wanted to restrict the analysis of changes between Wave 2 and Wave 3 (for example) only to CSMs then the "wave 1 to wave 3" panel weights would still be appropriate.

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 (*w**X**_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.

6.11.4. The Calibrated Weights

All waves were calibrated to provincial totals and to gender-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 previous 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.

6.11.4.1. 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 visited) than households in which there were neither younger children or pensioners. Any statistics which are correlated with the age-gender-race or provincial breakdowns are likely to be measured more accurately with the calibrated weights.

6.11.4.2. 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-gender-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[16].

[16] 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.

The following cells in Table 6.9 gave difficulties (taking a λ value in excess of four as a sign that the constraint gave problems):

Table 6.9: λ values greater than of 4 or less than -4

Wave	Constraint	λ	Constraint	λ	Constraint	λ
1	African Male 80+	-4.9	Col Male 75-79	-6.1	Indian Male 25-29	11.3
	Indian Male 30-34	4.0	Indian Male 50-54	8.7	Indian Male 55-59	-5.0
	Indian Female 0-4	-5.0	Indian Female 40-44	-4.5	Indian Female 45-49	-9.7
	White Female 80+	4.1				
2	African Male 80+	-5.0	Col Female 80+	-5.1	Indian Male 40-44	-5.2
	Indian Male 45-49	10.2	Indian Male 50-54	5.4	Indian Male 65-69	17.5
	Indian Female 5-9	6.2	Indian Female 15-19	-4.7	Indian Female 55-59	-4.5
3	African Female 80+	-4.1	Col Male 80+	-5.8	Indian Male 0-4	16.3
	Indian Male 5-9	-6.7	Indian Male 30-34	14.0	Indian Male 45-49	22.7
	Indian Male 50-54	14.6	Indian Male 55-59	9.5	Indian Male 65-69	7.8
	Indian Male 70-74	6.0	Indian Female 20-24	-8.1	Indian Female 25-29	-15.5
	Indian Female 30-34	-4.5	Indian Female 45-49	-5.4	Indian Female 50-54	-8.7
4	Indian Male 10-14	-4.7	Indian Male 15-19	4.7	Indian Male 30-34	5.6
	Indian Male 50-54	10	Indian Male 70+	6.0	Indian Female 0-4	-9.1
	Indian Female 5-9	11.1	Indian Female 30-34	12	Indian Female 35-39	9
	Indian Female 40-44	4.6	White Male 0-4	4.8	White Female 0-4	4.0
	White Female 30-34	6.0				

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 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.

6.11.5. Panel Weights

The individuals who were successfully re-interviewed in waves 2, 3 and 4 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.

6.11.5.1. Wave 1 to Wave 2 attrition

As shown in Table 6.10, the pattern of probabilities varies quite strongly with race and age[17]. Unfortunately, this pattern of attrition is correlated quite strongly with the initial pattern of nonresponse.

Table 6.10: Average probabilities of successful re-interview - Wave 1 to Wave 2 – by age, gender and race

Age Category	Male				Female			
	African	Coloured	Asian/ Indian	White	African	Coloured	Asian/ Indian	White
-9	0.0661	0.0616				0.0981		
0-1	0.8304	0.7687	0.8493	0.7412	0.8326	0.7812		0.8151
1-4	0.8610	0.7884	0.7924	0.7268	0.8534	0.7908	0.8259	0.6726
5-9	0.8820	0.8028	0.6645	0.6355	0.8701	0.7902	0.7342	0.5947
10-14	0.8652	0.7666	0.5228	0.4432	0.8529	0.7738	0.6395	0.4141
15-19	0.8214	0.7112	0.4918	0.3360	0.8208	0.7231	0.5155	0.3098
20-24	0.7736	0.6785	0.4556	0.2876	0.7961	0.7141	0.5634	0.3164
25-29	0.7462	0.6410	0.4930	0.3705	0.8040	0.7147	0.6047	0.3990
30-34	0.7285	0.6591	0.6085	0.3654	0.8123	0.7350	0.6006	0.4531
35-39	0.7197	0.6669	0.6296	0.4245	0.8216	0.7639	0.6284	0.4832
40-44	0.7202	0.6746	0.6926	0.4597	0.8375	0.7720	0.6532	0.4845
45-49	0.7398	0.6872	0.7401	0.5092	0.8497	0.7849	0.6965	0.5217
50-54	0.7428	0.7166	0.6667	0.4985	0.8523	0.7972	0.6852	0.5437
55-59	0.7602	0.7177	0.6786	0.5610	0.8697	0.7728	0.7202	0.5475
60-64	0.7873	0.7229	0.6614	0.5443	0.8671	0.7732	0.6670	0.5357
65-69	0.8130	0.6860	0.4008	0.5242	0.8691	0.7769	0.5663	0.5165
70-74	0.8124	0.6506		0.5154	0.8551	0.7433	0.5264	0.4238
75-79	0.7777	0.5522	0.2856	0.4657	0.8457	0.6967	0.3276	0.4478
80-84	0.7306	0.5487		0.4008	0.8059	0.5519		0.3591
85+	0.5170	0.2463		0.2757	0.7082	0.5232		0.4359

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 re-interviewed, 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 re-interviewed. 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,

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

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.

6.11.5.2. Wave 1 to Wave 3

The table of average probabilities by age-gender-race cells is given in Table 6.11.

Table 6.91: Average probabilities of successful re-interview - Wave 1 to Wave 3 – by age, gender and race

Age Category	Male				Female			
	African	Coloured	Asian/ Indian	White	African	Coloured	Asian/ Indian	White
-9		0.0381			0 .0843	0.0653		
0-1	0.8620	0.8358	0.8816	0.8370	0.8663	0.8361		0.8165
1-4	0.8854	0.8530	0.8121	0.7793	0.8796	0.8497	0.8345	0.7204
5-9	0.9022	0.8710	0.7005	0.6760	0.8932	0.8558	0.7458	0.6187
10-14	0.8853	0.8473	0.5696	0.4996	0.8724	0.8367	0.6580	0.4612
15-19	0.8394	0.7999	0.5477	0.3736	0.8404	0.7941	0.5406	0.3420
20-24	0.7862	0.7643	0.5617	0.3263	0.8133	0.7769	0.5875	0.3423
25-29	0.7530	0.7425	0.6051	0.4042	0.8085	0.7704	0.5892	0.4041
30-34	0.7135	0.7428	0.6506	0.4309	0.8129	0.7805	0.6676	0.4505
35-39	0.7116	0.7278	0.6917	0.4502	0.8156	0.7928	0.6630	0.4806
40-44	0.7021	0.7244	0.7412	0.5114	0.8280	0.7929	0.6913	0.5147
45-49	0.7117	0.7209	0.7319	0.5184	0.8319	0.8071	0.7379	0.5227
50-54	0.7091	0.7243	0.6493	0.5311	0.8341	0.8159	0.7180	0.5503
55-59	0.7197	0.7119	0.6250	0.5280	0.8483	0.7987	0.7333	0.5601
60-64	0.7470	0.6942	0.5667	0.5165	0.8411	0.7997	0.6773	0.5338
65-69	0.7544	0.6233	0.3781	0.4606	0.8310	0.7808	0.5635	0.5451
70-74	0.7284	0.5742		0.4654	0.8071	0.7396	0.4218	0.5054
75-79	0.6659	0.4075	0.2723	0.4715	0.7760	0.6632	0.3396	0.5353
80-84	0.5809	0.2768		0.4409	0.6972	0.5970		0.5578
85+	0.3399			0.3560	0.5420	0.3667		0.5550

It seems noteworthy that some of the probabilities are actually higher for a re-interview 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.

6.1.1.1 *Wave 1 to Wave 4*

The table of average probabilities by age-gender-race cells is given in Table 6.12. The age profile is as expected, with fewer survivors at high baseline ages (relative to wave 1 to Wave 3).

Table 6.102: Average probabilities of successful re-interview - Wave 1 to Wave 4 – by age, gender and race

Age Category	Male				Female			
	African	Coloured	Asian/ Indian	White	African	Coloured	Asian/ Indian	White
-9	0.1514	0.0939			0.1109	0.1225		
0-1	0.8672	0.8463	0.8678	0.8531	0.8730	0.8536		0.8343
1-4	0.8868	0.8577	0.7960	0.7792	0.8803	0.8511	0.8094	0.7287
5-9	0.8967	0.8625	0.6265	0.6198	0.8858	0.8409	0.7053	0.5559
10-14	0.8802	0.8377	0.4864	0.4135	0.8649	0.8166	0.5974	0.3520
15-19	0.8340	0.7872	0.4504	0.2917	0.8329	0.7719	0.4975	0.2381
20-24	0.7798	0.7599	0.4376	0.2676	0.8050	0.7575	0.5284	0.2728
25-29	0.7486	0.7245	0.5573	0.3153	0.8001	0.7565	0.5294	0.3202
30-34	0.7079	0.7314	0.5683	0.3432	0.8036	0.7736	0.5879	0.3637
35-39	0.6949	0.7150	0.6211	0.4188	0.8010	0.7859	0.5935	0.4234
40-44	0.6800	0.7014	0.6342	0.4691	0.8063	0.7798	0.6297	0.4444
45-49	0.6748	0.6844	0.5856	0.4812	0.8040	0.7827	0.6590	0.4680
50-54	0.6594	0.6677	0.5206	0.4856	0.7981	0.7728	0.6409	0.5081
55-59	0.6468	0.6365	0.4565	0.4651	0.7955	0.7297	0.6519	0.5126
60-64	0.6517	0.5857	0.3911	0.4407	0.7726	0.7154	0.5669	0.4964
65-69	0.6331	0.5068	0.2345	0.3452	0.7445	0.6555	0.4053	0.4595
70-74	0.5724	0.4308		0.3396	0.7007	0.5909	0.3931	0.3874
75-79	0.4777	0.3043	0.1592	0.2189	0.6406	0.4693	0.2507	0.3453
80-84	0.3772	0.2583		0.2005	0.5170	0.4216		0.3617
85+	0.0911				0.3808	0.3149		0.1895

6.11.6. 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 subsample) analysing any subsample that is too small is probably inviting trouble.

7. Inclusion of Census 2011 Geographic Variables

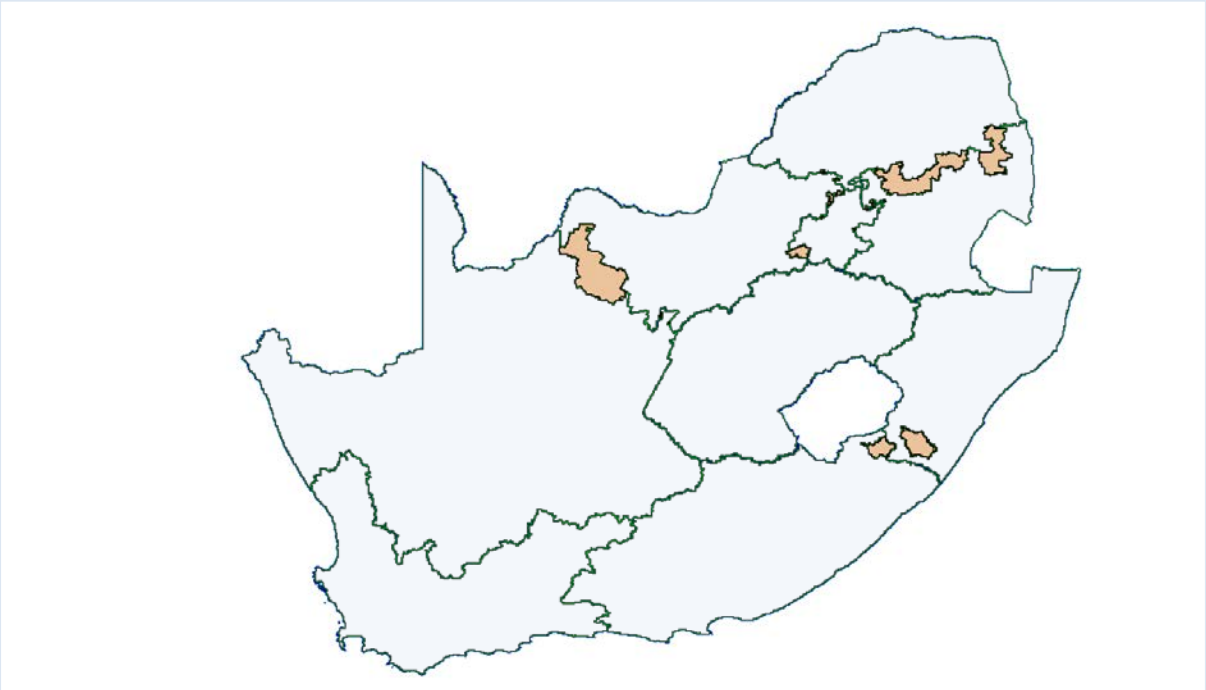
Statistics South Africa (StatsSA) made Census 2011 data publicly available late 2013. This created the opportunity to update the geographical variables in the NIDS Datasets. The household level geographic variables presented in NIDS are Province, District Council and Geo-type. The Secure Data also includes the Main Place and EA number. Prior to the public release of Census 2011, NIDS had calculated these variables based on the 2001 Census boundaries. This section outlines the differences and has important cautionary notes about the differences between 2001 and 2011 geography.

To assist users all **previously released geography variables** are still included in all waves, they **have just been renamed to include the suffix *'2001'***. The new geography variables have the suffix *'2011'*. See the detail of the changes in the respective Change documents for each wave. Care should be taken when comparing to household level variables. If using the 2011 household variables, then users must also use the 2011 migration equivalents e.g. *w**X**_a_brndc2011* for the district council in which the individual was born. The same applies to the 2001 variables.

7.1 Provincial Boundary Changes

The Provincial Boundaries changed between the 2001 Census and the 2011 Census. Figure 7.1 is a map showing the provincial boundary changes. The light shaded areas are the areas that have changed boundaries. All the provincial codes have stayed the same.

Figure 7.1: Province changes between 2001 and 2011

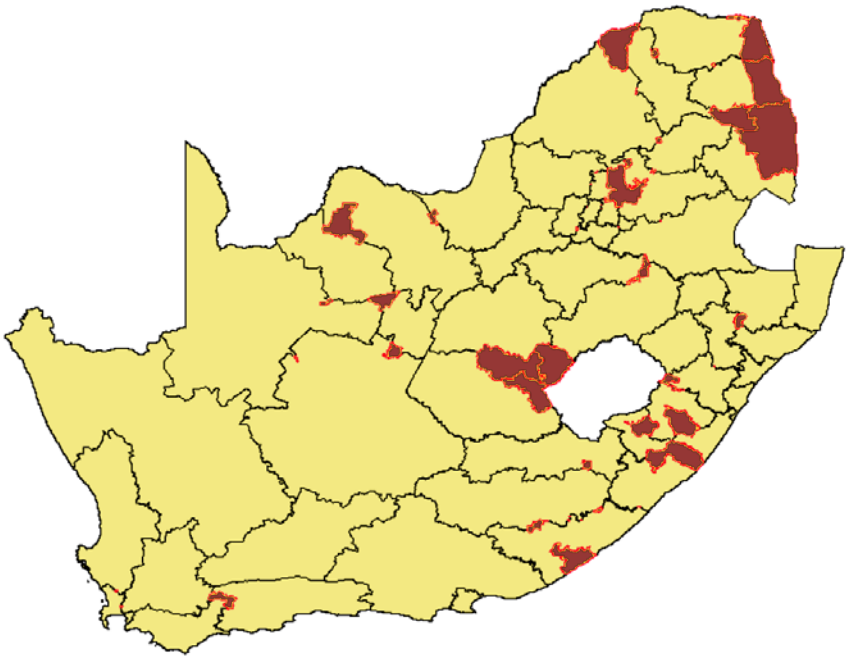


7.2 District Council Changes

There were significant changes to the District Municipal Boundaries between 2001 and 2011. The number of metropolitan municipalities increased from 6 to 8, while the number of district municipalities remained 52 (44 plus 8 metros). District Council codes were also changed. For example, while City of Cape Town used to be code 171, it is now code 199. In order to assist users we release the original District Council variable renamed to *w**X**_dc2001* as well as the new District Council codes (*w**X**_dc2011).* We also include the Municipal Demarcation Board code (*w**X**_mdbdc2011)*. The Municipal Demarcation Board variable is only available for 2011. Also note that this is a string/text variable, not a numeric.

It is very important to note that the 2001 and 2011 District Council codes are not comparable at all. Given the change in numbering and the change in boundaries, comparisons cannot be made. Figure 7.2 is a map outlining where the District Councils changed shape.

Figure 7.2: District Council changes between 2001 and 2011



7.3 Geographical Type Variables

In the 2011 Census data there are two geographical type variables, *EA_GTYPE* and *EA_TYPE*. Previously there was only one variable that classified EAs as Rural Formal, Traditional Authority Area, Urban Informal or Urban Formal. The new geographical type classifications are Urban, Traditional and Farms.

Below is Table 7.1 that gives a description of each new category.

Table 7.1: *EA_GTYPE* source, StatsSA

EA_GTYPE_C	EA_GTYPE	Description
1	Traditional	Communally-owned land under the jurisdiction of traditional leaders. Settlements within these areas are villages.
2	Urban	A continuously built-up area that is established through township establishment such as cities, towns, 'townships', small towns, and hamlets. The areas are identified by "erf/erven/cadastre" from the Surveyor General or Municipal planning units.
3	Farms	Land allocated for and used for commercial farming including the structures and infrastructure on it. The areas are identified by farm and farm portion cadastre from the Surveyor General.

EA_TYPE classifies the EA by land use and human settlement in the area. Table 7.2 gives the descriptions of the land use and settlement types.

Table 7.2: *EA_TYPE* source, StatsSA

EA_TYPE_Code	EA_TYPE	Example
1	Formal residential	Single houses, town houses, high rise flats, scheme housing, estates
2	Informal residential	Illegal informal structures
3	Traditional residential	Villages in tribal areas
4	Farms	Farms
5	Parks and recreation	State forests, military training ground, holiday resorts, nature reserves, national parks
6	Collective living quarters	School hostels, tertiary education hostels, workers' hostels, military barracks, prisons, hospitals, hotels, old age homes, orphanages, monasteries
7	Industrial	Factories, large warehouses, mining areas, saw mill, railway stations and shunting areas, airports
8	Small holdings	Small holdings, agricultural holdings
9	Vacant	Open areas within urban and traditional areas
10	Commercial	Mixed CBD, office parks, shopping malls

The *EA_TYPE* variable with categories given in Table 7.2 is only available in the Secure datasets and is named *w**X**_eatype2011*

Very important: Do not merge across census periods, i.e. 2001 to 2011 variables. Matches might be false and not represent the same space or values.

7.4 Impact of Geographic Variable Changes on Data

7.4.1. Impact of Geographic Variable Changes at a Household Level

The inclusion of the 2011 variables was effective as of version 5.2 of Wave 1, version 2.2 of Wave 2, version 1.2 of Wave 3 and version 1.0 of Wave 4, resulting in two sets of geographic variables being available at a provincial, district, and geographical type level in the household derived dataset.

7.4.2. Impact of Geographic Variable Changes at an Individual Level

The 2011 variables were included into the migration section of the individual datasets as of version 6.0 of Wave 1, version 3.0 of Wave 2, version 2.0 of Wave 3 and version 1.0 of Wave 4. The District Council of birth (*brndc*) as well as the District Council prior to current location (*lvbfdc*) were affected in the Adult, Child and Proxy datasets of all four waves. In addition to this, District Council in 1994 and in 2008, asked in Wave 2, were changed to include 2011 variables. Similarly, District Council in 1994 and in 2006, asked in Wave 1, were changed to include 2011 variables.

7.5. Impact of Geography Variable Changes on Other Variables

7.5.1. Weights

Weights are calculated using the appropriate mid-year population estimates from Stats SA. The mid-year population estimates have used the latest provincial boundaries since 2007. However, as described above, NIDS initially reported provincial boundaries as they appeared in the sample originally provided by StatsSA, which reflected the 2001 boundaries. All the weights calculations in all waves were updated to use the 2011 Census boundaries. The result is that almost all weights changed slightly. Although individual cases might have shifted by seemingly significant proportions, the overall changes are insignificant. We also believe that this revision reflects the most accurate data available.

7.5.2. Imputed Income and Expenditure Variables

All derived files for Waves 1, 2 and 3 use the 2011 geographic variables. The do-files are available from our Program Library for users who want to recreate these variables or understand how they are created.

8. 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.

8.1 Data Manipulation

8.1.1. Merging Datasets

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

1. Program 1 - Merging all the data into a panel

   This program creates a panel dataset by merging all of the NIDS datasets together.

   It must be noted that this does not create a balanced panel dataset and as such the interview outcomes need to be take into account when performing analysis.

2. Program 2 - Merging files for a given wave into a cross-section

   This program creates a cross-sectional dataset for a given wave by merging together the Individual and, Household questionnaires, Household Roster and Derived datasets.

   It must be noted that both non-resident and deceased respondents will be included in this dataset.

8.1.2. Reshaping data

3. Program 3 - Reshaping the Birth History section and merging in the offspring questionnaires

   This program uses the Adult dataset form any given wave, keeps the mothers identifiers, along with her birth history and reshapes the data into a roster form. Then in a separate process it appends the Adult, Child and Proxy datasets together and merges this appended dataset to the reshaped birth histories against the offspring's identifiers that appear on the birth history.

4. Program 4 - Reshaping of the mortality section to create a roster

   The function of this program is to create a roster form of the mortality history located in the Household dataset. It does this by opening the Household questionnaire for any given wave, keeping the household identifiers and the mortality data and reshaping the dataset to create a roster.

8.2.Derived Variables

8.2.1. 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. Do-files showing calculation of household income can be found here.

8.2.2. Expenditure

As explained above in section 6.8, NIDS constructed a derived variable as a measure of total household expenditure in the 30 days preceding the interview taking place. Do-files showing calculation of household expenditure can be found here.

8.2.3. Wealth Program Library

As explained above in section 6.9, NIDS constructed derived variables as a measure of both total household wealth and total individual wealth in the 30 days preceding the interview taking place. Do-files showing calculation of both household and individual wealth can be found here.

8.2.4. Deflators

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

8.2.5. Employment Status

NIDS constructed a derived variable using the International Labor Organization definitions to assign respondents to one of the following categories - Employed, Unemployed (strict definition), Unemployed (broad definition) and Not Economically Active. Do-files can be found here.

9. References

Chinhema, M., Brophy, T., Brown, M., Leibbrandt, M., Mlatsheni, C., & Woolard, I., eds. 2016. "National Income Dynamics Study Panel 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, M, 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, M, 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