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

National Income Dynamics Study Panel User Manual

Release 2018

Version 1

Timothy Brophy, Nicola Branson, Reza C. Daniels, Murray Leibbrandt, Cecil Mlatsheni and Ingrid Woolard

SALDRU
Southern Africa Labour and Development Research Unit

University of Cape Town · iYunivesithi yaseKapa · Universiteit van Kaapstad

Forward

The National Income Dynamics Study (NIDS) is the first national household panel study in South Africa. This survey is an initiative of the Department of Planning, Monitoring and Evaluation (DPME) and is part of efforts from the South African government to track and understand the shifting face of poverty. The National Income Dynamics Study is implemented by the Southern Africa Labour and Development Research Unit (SALDRU) based at the University of Cape Town's School of Economics.

NIDS examines the livelihoods of individuals and households over time. It also provides information about how households cope with positive or negative shocks, such as an unemployed relative obtaining a job or a death in the family. Other themes include changes in poverty and well-being; household composition and structure; fertility and mortality; migration; labour market participation and economic activity; human capital formation, health and education; vulnerability and social capital.

The study began in 2008 with a nationally representative sample of over 28,000 individuals in 7,300 households across the country. The survey continues to be repeated with these same household members every two years. These are called our Continuing Sample Members (CSMs). Any other member who becomes part of the household is consequently interviewed but is not tracked in the following waves. These are called Temporary Sample Members (TSMs). Children born to CSM mothers are added to the sample of CSMs and are tracked.

Due to attrition of White, Indian/Asian and high-income respondents, a Top-Up sample was added at Wave 5 (2017) to maintain the representativeness of the sample. In total 2 775 CSMs were added as a result of the Top-Up.

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 data files. 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 5 data and updated versions of Wave 1, 2, 3 and 4 datasets. Highlights in the data are as follows:

Changes across waves:

- ➢ Update to the imputation of Wealth in Wave 4, and the implementation of this update in Wave 5 (See Section 6.12.3.1 2018 Correction of Wave 4 Wealth)
- ➢ Update of the calculation of the *Best school education* variable to include grade 0/R (See Section 6.1 Best Variables)
- ➢ Addition of newly identified CSM babies, who can be identified by the "*w`x'_c_pfr*" variable added to the *Child* data file in waves 2 to 5 (See Section 5.1.1 Addition of Newly Identified CSM Babies to Prior and Current Waves.
- ➢ Introduction of a new method to version datasets (See Section 3.4.1
- ➢ Versioning of the data)

Changes from wave 5:

- ➢ The addition of a Top-Up Sample in Wave 5 and the sample variable (See Section 6.9 Impact of the 2017 Top-Up on Income, Expenditure and Wealth)
- ➢ Changes in the education codes in Wave 5, allowing respondents to accurately define post schooling activities. See Section 5.4 Adjustment To Education Codes
- ➢ The inclusion of interviewer demographics and experience in the *indderived* and *hhderived* data files (See section 6.8 Interviewer Demographics and Experience)
- ➢ The inclusion of financial literacy question and financial literacy derived scores (See Section 6.7 Financial literacy)
- ➢ The inclusion of trust questions capturing respondents' level of trust in others
- ➢ The inclusion of questions regarding smoking behaviour
- ➢ The omission of questions on alcohol consumption

List of Contributors

This document was created by the NIDS team over many years. Contributing authors in alphabetical order include:

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

Readers wishing to cite this User Manual should use the following citation:

Brophy, T., Branson, N., Daniels, R.C., Leibbrandt, M., Mlatsheni, C., & Woolard, I., 2018. *National Income Dynamics Study panel user manual*. Release 2018. Version 1. Cape Town: Southern Africa Labour and Development Research Unit.

Contents

1 Using This Manual

The National Income Dynamics Study (NIDS) is a face-to-face longitudinal survey of individuals living in South Africa and 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 data files.

This document accompanies the release of the Wave 5 data. As There have been updates to the data of previous waves and it is thus necessary to use the latest releases of previous waves when analysing the data. Please refer to the latest documentation for changes in the updated waves if merging to the wave 5 dataset. These are available with the data files on DataFirst's data site https://www.datafirst.uct.ac.za/dataportal/index.php/catalog/NIDS/about and also on the NIDS website www.nids.uct.ac.za

1.1 What All Data Users Must Know

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

- The structure of the data (Section 3).
- The fieldwork schedule (Section 4.3).
- Weights. (See section 6.14).
- Correctly merge NIDS data using Stata (See section 8.1.1).
- Deflate financial data (See section 8.2.4).

1.2 Citation of NIDS Dataset and Documentation

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

Dataset Citation:

Wave 5:

Southern Africa Labour and Development Research Unit. National Income Dynamics Study 2017, Wave 5 [dataset]. Version 1.0.0 Pretoria: Department of Planning, Monitoring, and Evaluation [funding agency]. Cape Town: Southern Africa Labour and Development Research Unit [implementer], 2018. Cape Town: DataFirst [distributor], 2018. https://doi.org/10.25828/fw3h-v708

Wave 4:

Southern Africa Labour and Development Research Unit. National Income Dynamics Study 2014-2015, Wave 4 [dataset]. Version 2.0.0. Pretoria: Department of Planning, Monitoring, and Evaluation [funding agency]. Cape Town: Southern Africa Labour and Development Research Unit [implementer], 2018. Cape Town: DataFirst [distributor], 2018. https://doi.org/10.25828/f4ws-8a78

Wave 3:

Southern Africa Labour and Development Research Unit. National Income Dynamics Study Wave 3, 2012 [dataset]. Version 3.0.0. Pretoria: SA Presidency [funding agency]. Cape Town: Southern Africa Labour and Development Research Unit [implementer], 2018. Cape Town: DataFirst [distributor], 2018. https://doi.org/10.25828/7pgq-q106

Wave 2:

Southern Africa Labour and Development Research Unit. National Income Dynamics Study Wave 2, 2010-2011 [dataset]. Version 4.0.0. Pretoria: SA Presidency [funding agency]. Cape Town: Southern Africa Labour and Development Research Unit [implementer], 2018. Cape Town: DataFirst [distributor], 2018. https://doi.org/10.25828/j1h1-5m16

Wave 1:

Southern Africa Labour and Development Research Unit. National Income Dynamics Study (NIDS) Wave 1, 2008 [dataset]. Version 7.0.0. Pretoria: SA Presidency [funding agency]. Cape Town: Southern Africa Labour and Development Research Unit [implementer], 2018. Cape Town: DataFirst [distributor], 2018. https://doi.org/10.25828/e7w9-m033

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

Documentation Citation:

Brophy, T., Branson, N., Daniels, R.C., Leibbrandt, M., Mlatsheni, C., & Woolard, I., 2018. *National Income Dynamics Study panel user manual*. Release 2018. Version 1. Cape Town: Southern Africa Labour and Development Research Unit.

1.3 Versions of Dataset used to complete this documentation

All figures contained in this document were generated using the following datasets:

Wave	Dataset Version
5	1.0.0
4	2.0.0
3	3.0.0
2	4.0.0
1	7.0.0

2 Number of Observations

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

Table 2.1. shows the total number of observations in each data file for each wave.

Table 2.1.: Summary of n-values across waves

File Name	**Identifiers***	**n**				
		w1	w2	w3	w4	w5
Link File	Pid	-	35167	41396	50369	59677
HHQuestionnaire	w**X**_hhid	7296	9125	10218	11889	13719
HouseholdRoster	w**X**_hhid	7296	9125	10218	11889	13719
	pid	31141	35422	40794	47009	50319
Adult	w**X**_hhid	7289	8841	9965	11605	13464
	pid	16872	21874	22457	26804	30110
Proxy	w**X**_hhid	1375	898	2067	1383	1685
	pid	1750	1124	2714	1597	1952
Child	w**X**_hhid	4327	5062	5638	6307	6878
	pid	9604	11293	12382	13971	14993
hhderived	w**X**_hhid	7296	9125	10218	11889	13719
indderived	w**X**_hhid	7296	9014	10114	11726	13543
	pid	28226	34291	37553	42372	47055

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

2.1 Response Rates

Table 2.2 below presents the numbers of Continuing Sample Members (CSMs[1]) and Temporary Sample Members (TSMs[2]) successfully interviewed in each wave as well as the number of CSMs and TSMs that were added to each wave. 73% of the individuals who were interviewed in Wave 1 were successfully interviewed in Wave 5. 77% of the 1856 CSMs who were either added to the study in Wave 2 or not successfully interviewed in Wave 1were successfully interviewed in Wave 5. 87% of the CSMs who were added in Wave 3 were successfully interviewed in Wave 5, and 92% of the CSMs who were added in Wave 4 were successfully interviewed in Wave 5. 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.

There were low response rates at the Wave 1 baseline sample and subsequent high attrition in waves 2 to 4 of White, Indian/Asian and high-income individuals. A Top-Up sample was therefore added at Wave 5 to increase the number of respondents in these groups to ensure representivity.

[1] Continuing Sample Member: All resident members of the original selected Wave 1 households or the Wave 5 Top-Up Sample (including children) and any children born to female CSMs in subsequent waves.

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

Table 2.2: CSMs and TSMs successfully interviewed by wave

		Interviewed in Wave 1	Interviewed in Wave 2	Interviewed in Wave 3	Interviewed in Wave 4	Interviewed in Wave 5
First Present in Wave 1	CSM	26776	21116	21394	20778	19302
First Present in Wave 2	CSM		1856	1596	1557	1445
	TSM		5565	3144	2281	1845
First Present in Wave 3	CSM			1346	1234	1165
	TSM			5102	2540	1910
First Present in Wave 4	CSM				1723	1584
	TSM				7255	3796
First Present in Wave 5	CSM Total					3278
	CSM Original Sample					1262
	CSM Top-up					2016
	TSM Total					5109
	TSM Original Sample					5109
Total successful individual interviews		**26776**	**26776**	**28537**	**32582**	**37368**
CSMs attempted		28226	26776	29 431	32056	30478
TSMs attempted			5739	5 736	18313	12742

Comparisons on individual outcomes across waves are presented in
Table 2.3, Table 2.4, Table 2.5 and Table 2.6. The most common reason individuals were interviewed in one wave but not the next is because in these households TSMs no longer live with any CSMs in the household. Since TSMs are not tracked, if they no longer live in a household with any CSMs, they will not be re-interviewed.

Table 2.3: Wave 5 and Wave 4 individual outcomes[3]

Wave 5	**Wave 4**						
	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	29730	250	640	0	0	0	480
Refused/ Not Available	419	225	485	0	0	0	12
Household Level Non-Response	2235	130	1376	14	0	0	10
Not Tracked in Wave 4	66	1	32	0	0	0	0
Moved Outside of SA	11	0	1	8	0	0	0
Deceased This Wave	705	7	66	0	0	0	0
Deceased in a Prior Wave	0	0	0	0	882	1583	0
Not Co-resident with any CSMs	4202	112	22	0	0	0	5030

Table 2.4: 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	26555	446	1392	0	0	0	292
Refused/ Not Available	271	59	160	0	0	0	6
Household Level Non-Response	1443	76	1073	0	0	0	20
Not Tracked in Wave 4	90	31	1407	56	0	0	0
Moved Outside of SA	8	1	13	0	0	0	0
Deceased This Wave	768	17	93	0	0	0	4
Deceased in a Prior Wave	0	0	0	0	707	876	0

[3] The Top-Up sample members are not included in these figures

Not Co-resident with any CSMs	3447	89	58	0	1	0	1937

Table 2.5: 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	23619	559	2308	6	0
Refused/Not Available	263	49	81	0	0
Household Level Non-Response	1932	163	2074	3	0
Moved Outside SA	1	0	13	42	0
Deceased this Wave	542	12	152	0	0
Deceased in a Prior Wave	0	0	0	0	876
Not co-resident with any CSMs	2180	79	0	0	0

Table 2.6 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 2.6: Wave 2 and Wave 1 individual outcomes

Wave 2	**Wave 1**	
	Successfully Interviewed	Refused/Not Available
Successfully Interviewed	21116	947
Refused/Not Available	530	94
Household Level Non-Response	4246	365
Moved Outside SA	49	2
Deceased this Wave	834	42

Wave 1 individual household-level non-responses are not presented in Table 2.7. Household non-responses were not specified in Wave 1 and therefore there are no reasons for non-responses available for this wave.

Table 2.7: Reasons for household non-response at the individual level

		Refused / Not Available	Not Located	Not Tracked	Whole HH Dead	Moved Outside SA	Total
Wave 5	Number	3021	1500	1560	205	30	6316
	Percent	47.83	23.75	24.70	3.25	0.48	100
Wave 4	Number	1958	816	1550	189	38	4548
	Percent	43.05	17.94	34.08	4.16	0.84	100
Wave 3	Number	2051	2118	45	176	117	4453
	Percent	46.06	47.56	1.01	3.95	2.63	100
Wave 2	Number	1805	2198	624	158	82	4870
	Percent	37.06	45.13	12.81	3.24	1.68	100

Waves 4 and 5 see 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 and 5 listing that went to fieldwork.

2.2 Attrition

Attrition between waves is defined by comparing the number of successful interviews in a wave to the number in preceding 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 during interviews for Wave 3 but can be deceased at the end of Wave 4.

Table 2.8: Reasons for attrition

	Reason	Refusal	Non-Contact	Deceased	Total
Wave 5	Number	3481	3040	784	7305
	Percent	48	42	11	100
Wave 4	Number	2294	2400	882	5576
	Percent	41	43	16	100
Wave 3	Number	2481	2276	708	5465
	Percent	45	42	13	100
Wave 2	Number	2425	2890	876	6191
	Percent	39	47	14	100

Table 2.8 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. It is important to note that Wave 5 attrition excludes the Top-Up sample.

The racial distribution of attrition is presented below.

Table 2.9: Wave on wave attrition by race

	Pop. Group	Refusal	Non-Contact	Deceased	Total	Attrition Rate
Wave 5	African	2190	2006	635	4831	11.84
	Coloured	673	426	121	1220	18.68
	Asian/Indian	138	95	5	238	44.82
	White	475	512	23	1010	62.69
	Total	3481	3040	784	7305	14.76
Wave 4	African	1410	1489	717	3616	11.17
	Coloured	419	369	120	908	16.75
	Asian/Indian	117	86	10	213	43.74
	White	348	456	35	839	54.41
	Total	2294	2400	882	5576	14.01
Wave 3	African	1366	1748	580	3694	13.37
	Coloured	488	281	98	867	18.3
	Asian/Indian	122	41	5	168	36.44
	White	505	206	25	736	50.07
	Total	2481	2276	708	5465	15.94
Wave 2	African	1201	2185	738	4124	18.57
	Coloured	552	466	102	1120	26.95
	Asian/Indian	134	32	8	174	40.56
	White	538	207	28	773	53.87
	Total	2425	2890	876	6191	21.93

As shown in Table 2.9, non-contacts are the dominant reason for attrition among African respondents in waves 2,3 and 4, while refusals dominate for Asian/Indian and Coloured respondents in all waves. Refusals are the dominant attrition for White respondents in waves 2 and 3, and non-contact is the dominant reason in waves 4 and 5. The population groups with the highest attrition rates are White 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. Attrition rates for the panel will be lower than wave-on-wave attrition rates, for example, a respondent who refused in Wave 2 could be successfully interviewed in Wave 3. This negative attrition is not reflected in Table 2.8. or Table 2.9.

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 which can be viewed by clicking here or follow the steps below.

The steps to follow in order 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 your 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.
b) 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.communications@uct.ac.za 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.
c) Any publications that employ data obtained from DataFirst will cite the source of data in accordance with the Citation Requirement provided with each dataset.
d) A digital copy of all publications based on the requested data, or a link to such publication will be sent to DataFirst.
e) 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 files are only available in Stata format.

3.3 Data Structure

Every resident[4] individual (CSM[5] or TSM[6]) 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 questionnaire maps uniquely to a *household questionnaire* file and *household roster* file using the household identifier*, w**X**_hhid* (where ***X*** denotes the wave[7]). This is the household in which the person is resident at the time they are interviewed. Individual identifiers on their own merge non-uniquely to the household roster file. This lists all the rosters on which they are *household members[8].* 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 rules apply for non-residents:

- If a person left the household more than 12 months before the interview date, and subsequently died, their death and the details of their death are recorded 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, where possible, a proxy questionnaire is 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[9].*

[4] Residency is defined as someone who usually resides at the house for more than four or more nights a week.

[5] Continuing Sample Member: All resident members of the original selected Wave 1 households or the Wave 5 Top-Up Sample (including children) and any children born to female CSMs in subsequent waves.

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

[7] This notation is used throughout this document.

[8] 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.

[9] 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 move outside of South Africa.

- If a respondent moves 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* data files. Out-of-scope households are identified in the Link File with the household and individual outcome identifier variables.
- If the household refuses to participate or there is some other type of non-response (e.g. the household could not be located), the variables from the individual questionnaires will still appear in the data files but will indicate a 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 is the individuals are resident. The Link File also has other pertinent information, such as whether the individual is a CSM or TSM, the *individual questionnaire* data file in which their record can be found for that wave, and the original Wave 1 cluster of the household. The *Link file* also indicates whether respondents originate from the original 2008 sample or the 2017 Top-Up sample. 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. The combination of *w**X**_hhid* and *pid* is unique per person within each wave (Unique identifier for household: *w**X**_hhid*, non-unique identifier for individual: *pid)*.

Adult: One record per entry from the Adult[10] questionnaire. 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 (Unique identifier for individual: *pid, non*-unique identifier for household: *w**X**_hhid*).

Proxy: One record per entry from the Proxy[11] questionnaire (Unique identifier for individual: *pid, non*-unique identifier for household: *w**X**_hhid*).

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

[11] Proxy questionnaires are completed, where possible, for adults who are unavailable or unable to answer their own Adult questionnaire. Proxy questionnaires are also completed for individuals who are out-of-scope at the time of the interview.

Child: One record per entry from the Child questionnaire. Observations with no data beyond Section A 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**_c_outcome* variable (Unique identifier for individual: *pid*, non-unique identifier for household: *w**X**_hhid*).

Derived variables are variables that do not come from questions asked directly of the respondent, but which are calculated or imputed from responses to questions and other information. For example, aggregate income and expenditure variables were constructed from responses to income and expenditure questions. 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. Geographic information of the current location of households and the weights variables are included in this file (Unique identifier for household: *w**X**_hhid*).

indderived: One record per resident person. Deceased and non-resident household members are not included in this file (Unique identifier for individual: *pid,* non-unique identifier for household: *w**X**_hhid)*.

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

3.5 Versioning of the dataset

Wave 5 saw the introduction of a new versioning format. This new format aims to inform users of major, minor and patch fix releases of the data to ensure they use the most up to date data for their research. The versioning format can be broken down as follows:

Version[major releases].[minor releases].[patch releases]

Major release: Refers to the number of a major release. A major release only occurs when a new wave is added to the panel.

e.g. V1.0.0, here 1 indicates the first major release of the dataset.

Minor release: Refers to the number of a minor release. A minor release is the release of a single wave/s that occurs between major releases, i.e. a release of corrected data where no new wave is added to the panel.

e.g. V1.2.0, here the number 2 indicates a second minor release between major releases.

Patch releases: Refers to the number of a patch release. Patches are scripts (Stata do files) which are created to correct specific errors in the data. The scripts are given to users to run on the existing data to fix a specific data issue. Once run, the patch will generate a new minor release version of the data.

e.g. V1.2.3, here the number 3 indicates that 3 patch scripts were released to fix data issues in version 1.2 of the dataset.

3.6 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 to each wave as NIDS is a panel of individuals, and the household identifier is simply a tool to connect each individual to their household 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.

3.7 Merging Datafiles Within and Between Waves

From the release of Wave 2 the longitudinal dimension of NIDS can be explored and, with each subsequent wave, new opportunities to explore this 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 in the Program Library provided with the data. Wave 5 sees the introduction of a Top-Up CSM sample to allow the Wave 5 cross section to maintain national representivity. However, these Top-Up CSM will not form part of between wave merging at this time, as they only exist from Wave 5 onwards and not in any of the proceeding waves.

3.7.1 Identifying CSMs and Residents

The variable *w**X**_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 and in the Wave 5 Top-up were selected to be CSMs. However, all household members in all waves are assigned a *pid*, regardless of their CSM or residency status. To identify if a CSM is from the original 2008 sample or from the Wave 5 Top-Up sample, each data file contains a variable named *w**X**_**Y**_sample.* that indicates from which sample the CSM originates.

The variable *w**X**_r_pres* in each wave's *Household Roster* file can be used to identify residents. The residency criterion is important as a person can appear on multiple rosters but can only be resident (usually sleep 4 or more 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 they are recorded as *resident* only in the household where their individual interview was conducted. They are marked as non-resident in all other households that they are members of. In summary, individuals with multiple household memberships retain the same *pid*

Figure 3.1: CSMs and TSMs across waves

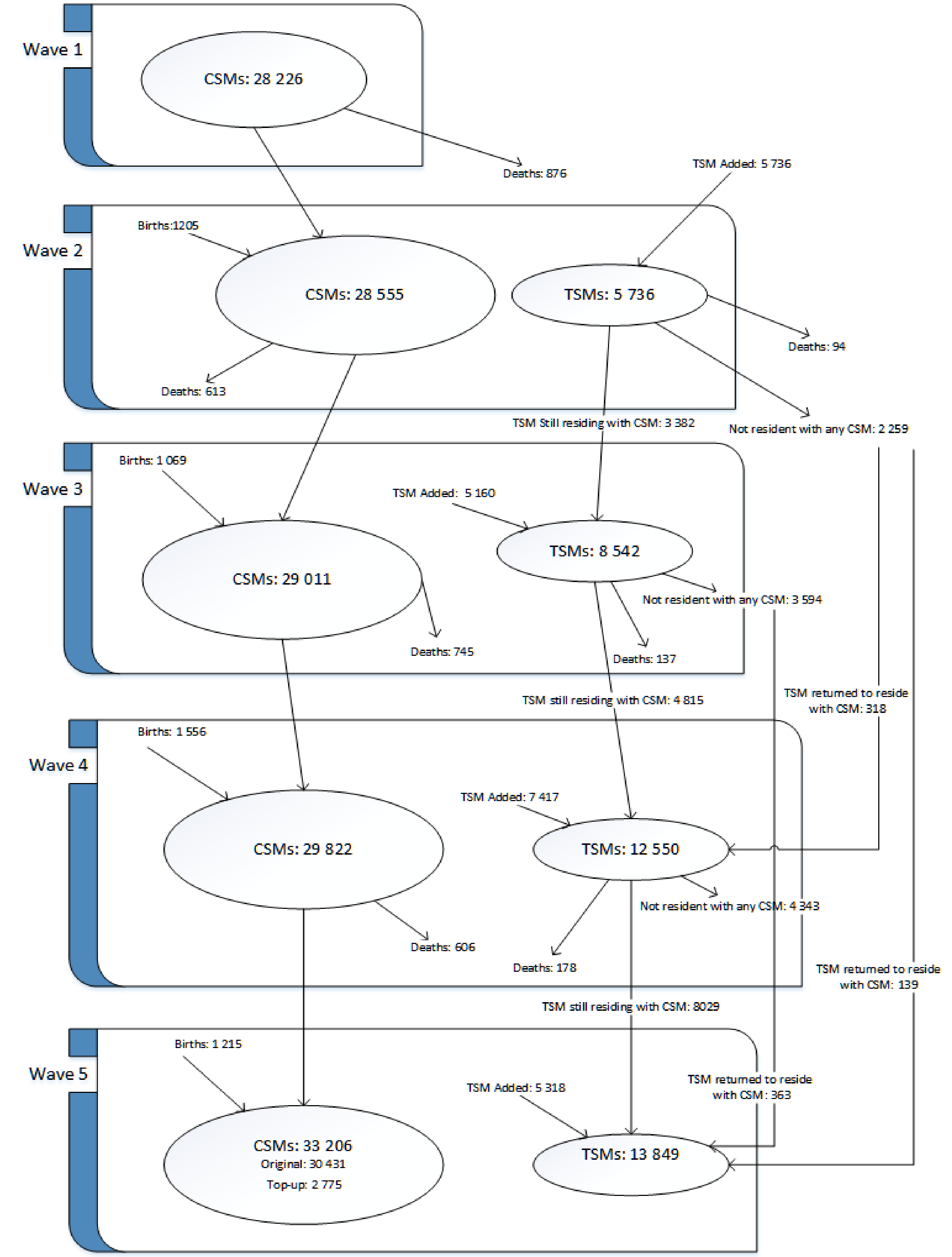


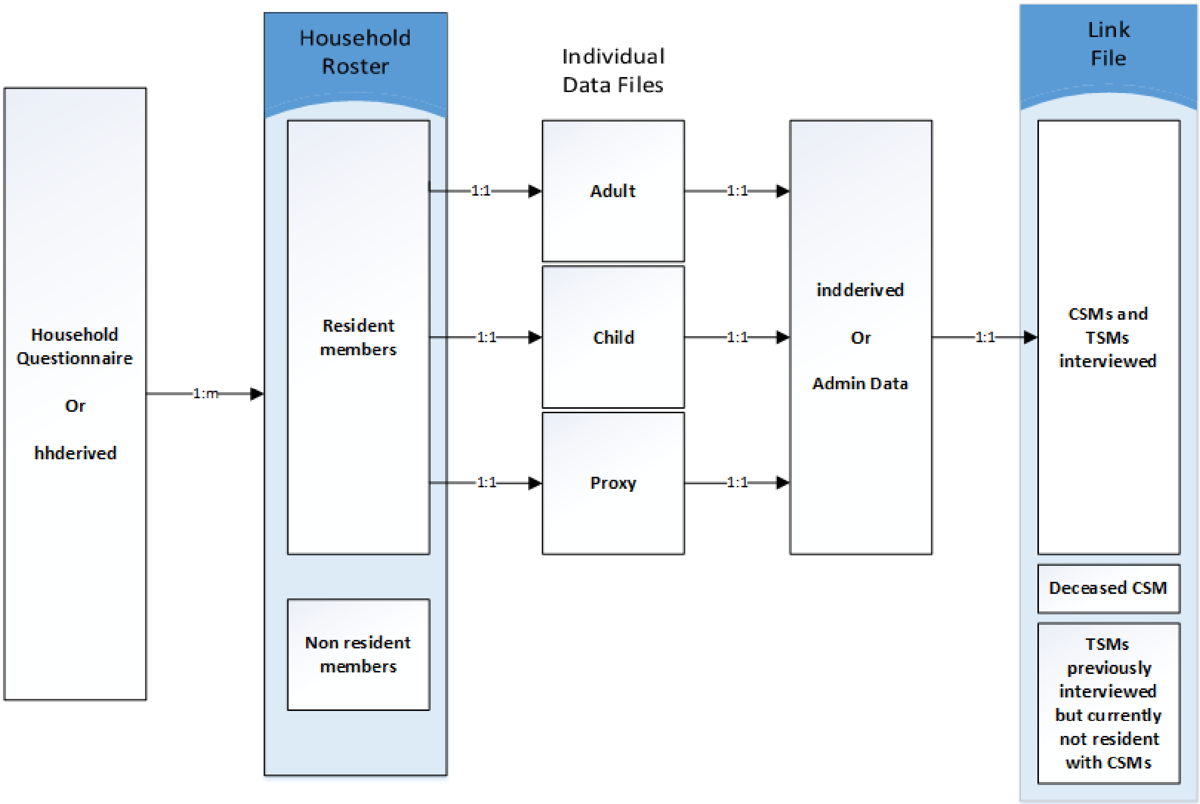
*Diagram adapted from the HILDA User Manual – Release 14

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

3.7.2 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 data files 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 Data File A will match one and only one other observation in Data File 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 files. 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.

3.7.3 Merging Data from different Waves

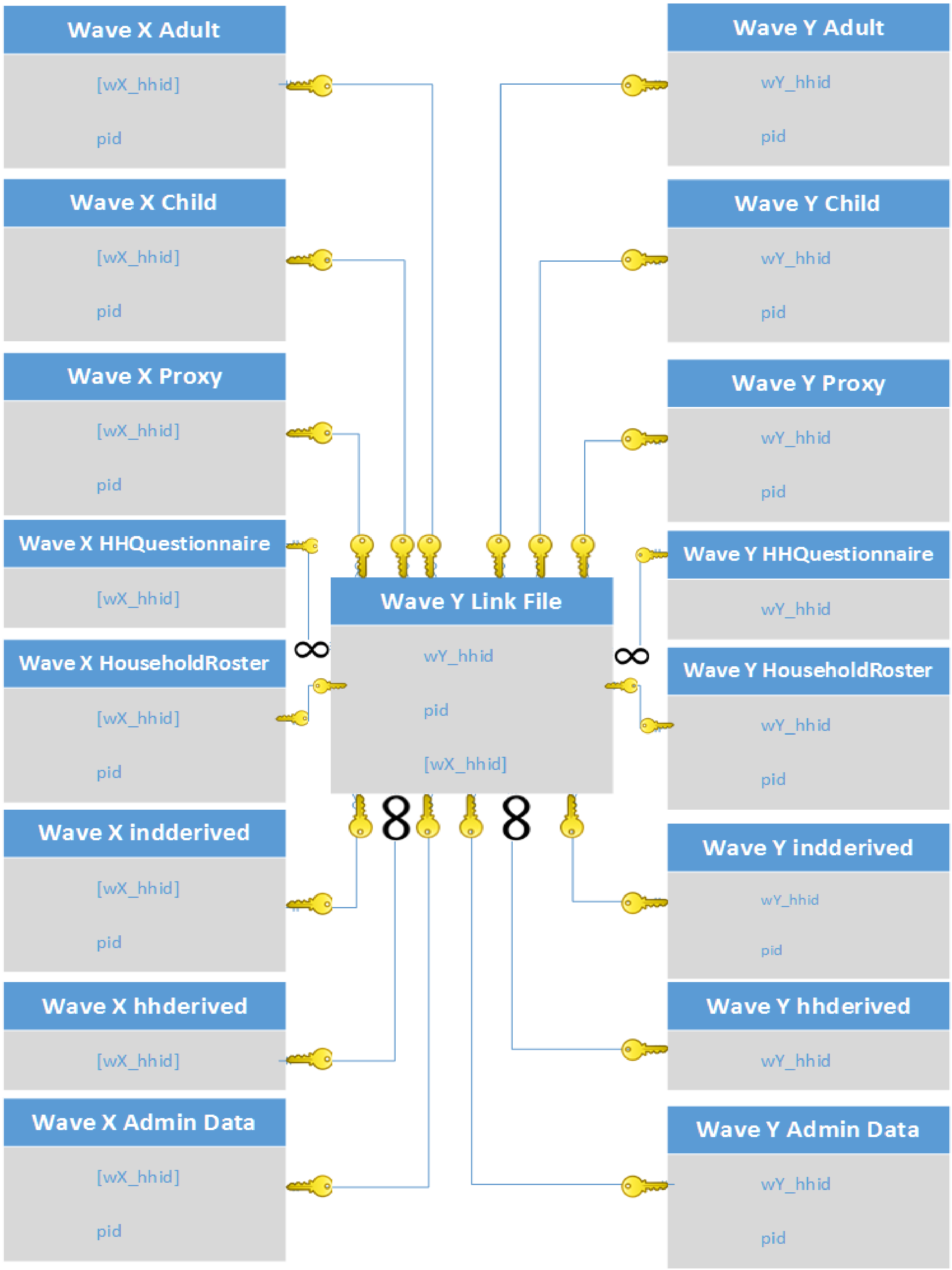
There are two ways to think about merging data from across the NIDS waves:

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 files 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 data files of interest from the remaining wave(s) can be performed.
    - Note that the Link_File contains only resident household members (including deceased members). The Household Roster file contains 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 data files from different 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 data files from different waves as it contains all information of the current and previous waves. In Figure 3.3, 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 *wX_hhid* and *pid* to merge the Household Roster data file in Wave X to the Link File. Once this is done, *wY_hhid* and *pid* can be used to merge the Household Roster in Wave Y to the Link File. Individual data files (Adult, Child, and Proxy) can be merged to the Link File using the *pid* which is a unique identifier in these data files. Merging the Household Questionnaire data file 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.

3.8 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 / subquestion

Details of each component are described below:

3.8.1 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.

3.8.2 Source

The source indicates which data file the variable belongs to, as shown in Table 3.1.

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

3.8.3 Questionnaire 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.

Table 3.2: Examples of significant questionnaire 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

3.8.4 Subsections

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

Employment:

Table 3.3: Example of employment subsections

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

Education:

Table 3.4: Example of education subsections

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		

Health:

Table 3.5: Example of health subsections

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

3.8.5 Descriptors

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

3.8.6 Sub-questions

Note that the sub-question is not a descriptor. Sub-questions only qualify a previous question, with a finite number of qualifying properties, such as location, value or explanation. A sub-question differs from an extension because it qualifies directly from a previous question. For instance, where the question asks whether 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 from a sub-question of the question "Do you sell produce?" and receives the suffix "_v".

3.9 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.

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.

3.10 Anonymisation

In order to protect the identity of our respondents every effort is made to remove personal information which 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).

3.11 Secure (restricted-access) Data

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. In addition to the Public Release Dataset, SALDRU also prepares datasets that include full geo-coding, employment coding and PSU information, as well as text variables as they are captured in the questionnaire. These are referred to as the NIDS Secure datasets. The NIDS Secure data only includes information as collected infield. Special releases are made from time to time of administrative data that has been matched to NIDS data.

The purpose of the Secure datasets is to allow academics 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.

Access to the Secure data is only in DataFirst's Secure Research Data Centre at the School of Economics Building, Middle Campus, University of Cape Town, Researchers go through an accreditation process to be granted access to the Centre. Secure data may not leave the premises, and all research output from the Centre undergoes disclosure control checks before being released. Researchers can apply for access to the Secure NIDS data in the Centre by downloading an accreditation form from http://www.datafirst.uct.ac.za/services/secure-data-services and emailing the completed form to support@data1st.org

3.12 Program Library

NIDS makes several Stata Programs available to users to assist them to use and manipulate the NIDS datasets. The Stata do-files used to create derived variables are also available with the data. See Section 8 Program Library in this User Manual 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
Wave 5	February 2017	December 2017

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 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 of the NIDS survey, 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 to complete in non-agricultural households and 50 minutes to complete in agricultural households.
- **Adult questionnaire**: The Adult questionnaire is applied to all present CSMs and other household members resident in the household 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 available for a direct interview, 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. During Wave 5 fieldwork the ethics committee required NIDS to obtain verbal consent from proxy respondents before conducting proxy interviews.

- **Child questionnaire:** This questionnaire collects information about all CSMs and residents in the 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 sets of 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.

During Wave 5 a youth care consent form is signed for young adults (15-17 years old) by the young adult's caregiver. In addition, assent forms are signed by the young adults themselves.

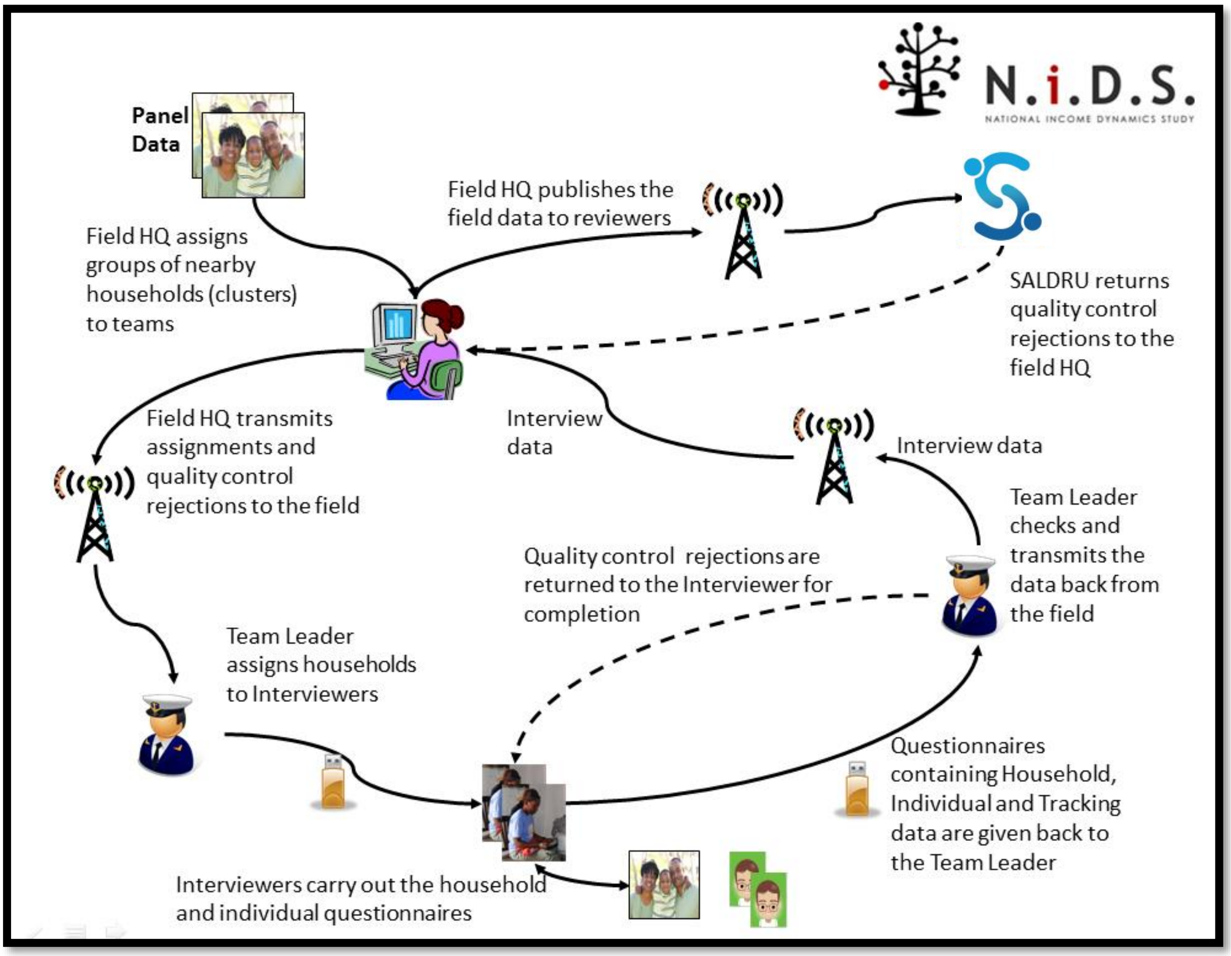
Anthropometric assent forms were required for children 7 to 10 years old. These assent forms were completed based on the child indicating their willingness to be measured as part of the anthropometric module.

During Wave 5 proxy respondents were also contacted directly to get their verbal consent for the interview to be conducted on their behalf, in addition to the person responding for them signing consent.

4.1.1 Overview of CAPI Cycle

The CAPI cycle is illustrated below.

Figure 4.1: The CAPI cycle


N.i.D.S.
NATIONAL INCOME DYNAMICS STUDY
Panel Data
Field HQ assigns groups of nearby households (clusters) to teams
Field HQ publishes the field data to reviewers
SALDRU returns quality control rejections to the field HQ
Field HQ transmits assignments and quality control rejections to the field
Interview data
Interview data
Team Leader checks and transmits the data back from the field
Quality control rejections are returned to the Interviewer for completion
Team Leader assigns households to Interviewers
Questionnaires containing Household, Individual and Tracking data are given back to the Team Leader
Interviewers carry out the household and individual questionnaires

Listing data (PSUs, household addresses, contact details, roster make-up and individual contact details) drawn from the previous wave are 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. 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 were provided in multiple previous waves, an example being head circumference of a child at birth.

The fieldworkers conduct the interviews and validate the questionnaire responses 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 questionnaire type (Household, Adult, Child and Proxy). These flat files are then validated again, and data inconsistencies or unexplained non-responses are returned to the field company directly, or checked via calls to the respondents.

4.1.2 Overview of the Tracking Process

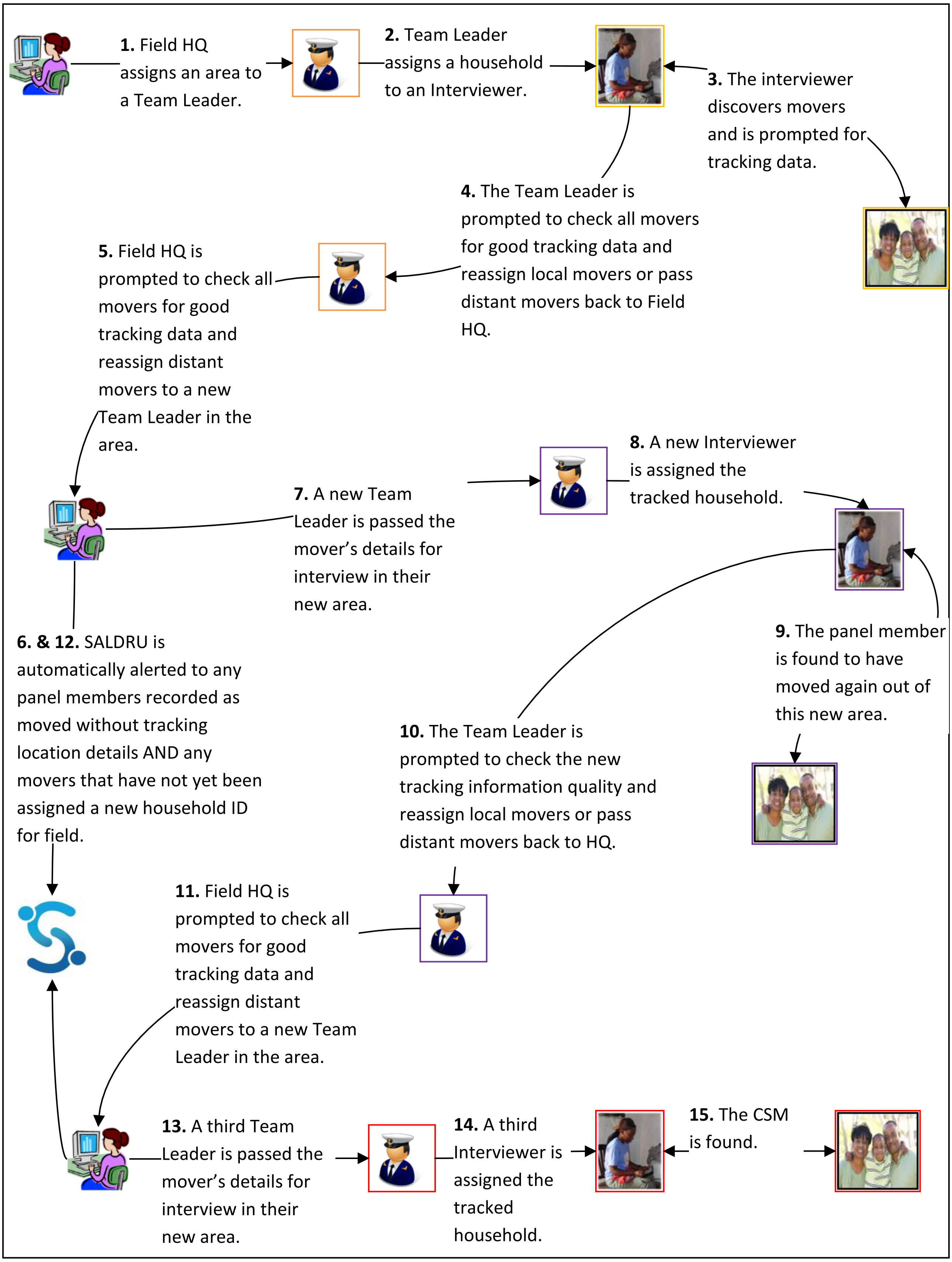
An essential part of the panel aspect of the survey is to track CSMs as they move within the borders of South Africa. CSMs can either be in the same location as they were in the previous wave (or the wave in which they were last located) or they could have moved. Interviewers use the CAPI system to record 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 identifiers (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 cluster signifying those 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 Figure 4.2: Tracking movers

Figure 4.2: Tracking movers



1. Field HQ assigns an area to a Team Leader.

2. Team Leader assigns a household to an Interviewer.

3. The interviewer discovers movers and is prompted for tracking data.

4. The Team Leader is prompted to check all movers for good tracking data and reassign local movers or pass distant movers back to Field HQ.

5. Field HQ is prompted to check all movers for good tracking data and reassign distant movers to a new Team Leader in the area.

6. & 12. SALDRU is automatically alerted to any panel members recorded as moved without tracking location details AND any movers that have not yet been assigned a new household ID for field.

7. A new Team Leader is passed the mover's details for interview in their new area.

8. A new Interviewer is assigned the tracked household.

9. The panel member is found to have moved again out of this new area.

10. The Team Leader is prompted to check the new tracking information quality and reassign local movers or pass distant movers back to HQ.

11. Field HQ is prompted to check all movers for good tracking data and reassign distant movers to a new Team Leader in the area.

13. A third Team Leader is passed the mover's details for interview in their new area.

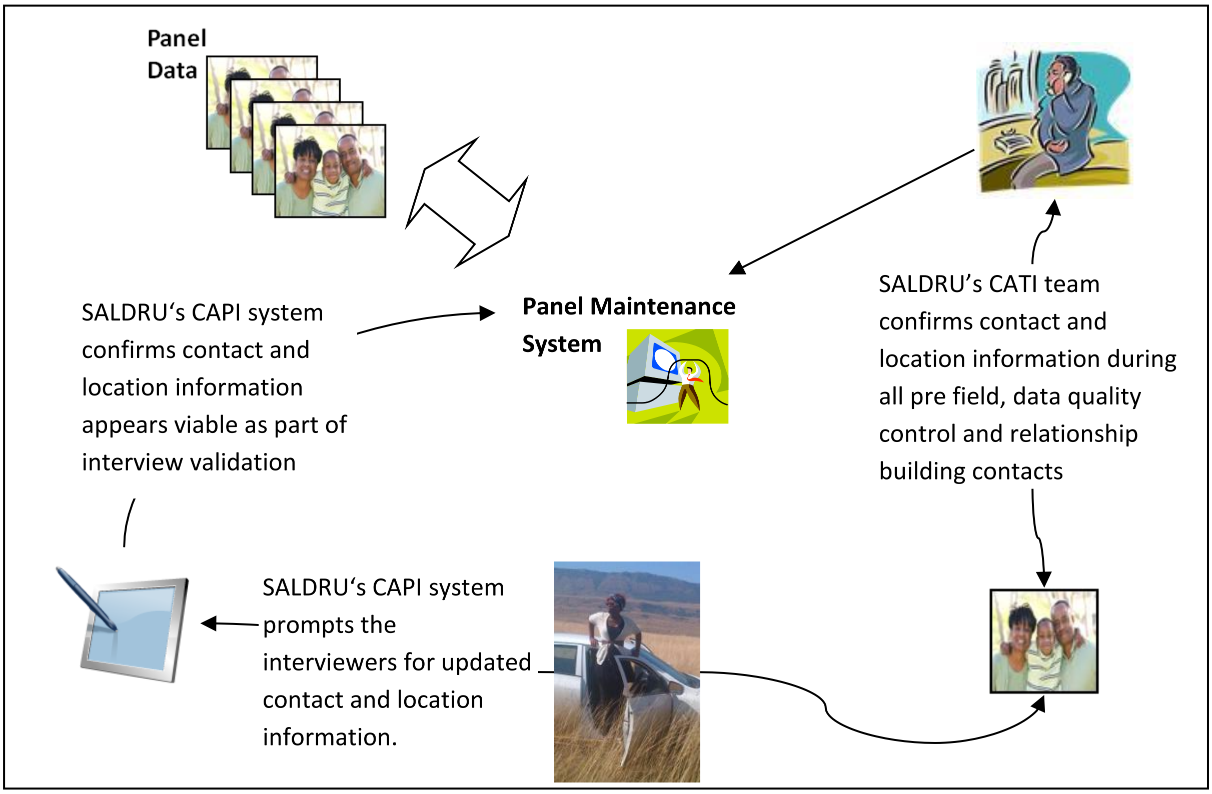
14. A third Interviewer is assigned the tracked household.

15. The CSM is found.

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


Panel Data
SALDRU's CATI team confirms contact and location information during all pre field, data quality control and relationship building contacts
SALDRU's CAPI system confirms contact and location information appears viable as part of interview validation
Panel Maintenance System
SALDRU's CAPI system prompts the interviewers for updated contact and location information.

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.

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 minimised 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 they should 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.

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. **No one at home policy:** Should there be no one at a dwelling, the interviewer is required to visit no less than three times at three different times of day on at least two different days before recording a household non-response.
4. **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.
5. **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.
6. **Individual level non-response call backs:** SALDRU attempts to contact all individuals from 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.
7. **Field organisation rewards:** Field company bonus schemes and targets have been structured to encourage better completion and lower attrition during fieldwork.
8. **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 to ensure that all CSMs are being approached and that the correct roster members are being referred to in their individual questionnaires.

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 has been left blank. There is a strict policy that data is only accepted from field if all sections of the questionnaire have been completed. There is a system for 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 instruments, across instruments, and across waves. Data collection involves several checks and mitigations to ensure data consistency:

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. In addition, cross questionnaire checks are 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 appears to have advanced too far between waves. 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 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 fieldwork. This also reduces interviewer effects. The following are some examples of areas that 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 lie 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 effects.

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 fieldwork to identify interviewer effects and possible mis-capture. 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 in field. If the query is across waves it

could result in a change of data for a previous wave. If the call-back 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 data 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 fieldwork. 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 with 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, data quality checks can commence in real time.

4.2.4.1 *Early Identification and Cleaning of Identifier Mismatches*

As part of cleaning the NIDS dataset, NIDS perform basic cleaning of the data in its raw relational data form, before the data is converted to 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 mismatches typically arise from:

- Erroneous reporting of moving of households, which creates new household identifiers when in fact the household remained intact and at the 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 fieldwork they are returned to the fieldwork company which must 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 interviewed or approached for an interview
- No GPS coordinates were collected for households successfully interviewed or households found but with valid non-response outcomes[12].
- Invalid "No one at home". Field teams have to demonstrate that they have visited these 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").

If these criteria are met, SALDRU then checks for other invalidities, such as:

- Incorrect person interviewed.
- Aberrant field behaviour (for example clear evidence of invention of data, unfeasible numbers of proxies rather than direct interviews).
- 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" are identified when 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 system and 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. 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

[12] Valid unit non-response outcomes are Refused and 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 as the pre-test to ensure consistency. Typically, there are about 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 the 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 year being referred to.

5 Main Data Processes

This section provides an explanation for some of the major sections 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**[13]) to each child in 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 where the birth history is inconsistent across waves, calls are made to respondents 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 is 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 is the one 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.

Wave 5 saw the continuation of the above process, allowing SALDRU to match and confirm children across the panel. This highlighted the fact that many children had been left off mother's birth histories resulting in CSM babies being under-reported in waves where mothers did not report the children.

[13] The asterisk donates a number that indicates the child's position in the mother's birth history, i.e. first born child is 1, second born is 2, and so on.

5.1.1 Addition of Newly Identified CSM Babies to Prior and Current Waves.

During Wave 5 data production a new variable called Post Field Respondent (“*w`x’_c_pfr*”) was added to the Child data files in waves 2, 3, 4 and 5. This variable was included to indicate CSM babies who were added to a given wave after the conclusion of that wave’s fieldwork.

These Post Field Respondent CSMs were identified after CSM mothers confirmed that they had neglected to include these newborn children in their birth histories in prior and current waves. As these children were born to CSMs after the Wave 1 baseline, the children themselves are CSMs and thus form part of the NIDS sample. These children have been added to the prior waves retrospectively with “Not Tracked” interview outcomes. At Wave 5, the total number of CSM babies added across the panel was 354. The number of children added to the data in each wave is represented in Table 5.1: Number of CSM Children added to each wave*:

Table 5.1: Number of CSM Children added to each wave*

Wave	Number of CSM Children Added
2	204 not tracked
3	165 not tracked, 2 deceased
4	58 not tracked, 1 deceased
5	15 not tracked, 2 deceased

**It is important to note that the above totals refer to the numbers of CSM babies added to a particular wave, in most cases the same CSM baby needed to be added to multiple waves. Thus, the above table represents the total number of additions to each wave not the unique number of CSM babies.*

Users may be concerned about the increase in household size after the addition of CSM babies. Table 5.2: Household size, weighted and unweighted illustrates that there is zero difference for weighted average household size and a small difference for unweighted average household size for each of the affected waves:

Table 5.2: Household size, weighted and unweighted

Wave	Weighted		Unweighted	
	Average HH Size BEFORE Addition	Average HH Size AFTER Addition	Average HH Size BEFORE Addition	Average HH Size AFTER Addition
2	3.741	3.741	3.899	3.922
3	3.689	3.689	4.095	4.112
4	3.582	3.582	4.138	4.143
5	3.322	3.322	3.804	3.805

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.

NIDS 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 data was updated for each wave. This process was also applied consistently in Wave 5

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

5.3 Education Progression

In wave 5, all respondents aged between 15 and 30 and respondents who said they were enrolled in an educational institution in Wave 4 were asked about their education progression from 2015 to 2017. For both CSMs and TSMs who were not new in wave 5, we asked questions on education up until and including the year of their last interview. 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 Adjustment To Education Codes

The education codes and categories for education questions were updated in the Wave 5 questionnaires. Table 5.3: Changes in education codes shows a comparison of the education codes and categories used in Wave 1 – 4 compared to Wave 5.

Table 5.3: Changes in education codes

Code	Description	Wave 1 – Wave 4	Wave 5
0	Grade R/0	✓	✓
1	Grade 1 (previously Sub A / Class 1)	✓	✓
2	Grade 2 (previously Sub B / Class 2)	✓	✓
3	Grade 3 (Std 1)	✓	✓
4	Grade 4 (Std 2)	✓	✓
5	Grade 5 (Std 3)	✓	✓
6	Grade 6 (Std 4)	✓	✓
7	Grade 7 (Std 5)	✓	✓
8	Grade 8 (Std 6/Form 1)	✓	✓
9	Grade 9 (Std 7/ Form 2)	✓	✓
10	Grade 10 (Std 8/ Form 3)	✓	✓
11	Grade 11 (Std 9/ Form 4)	✓	✓
12	Grade12 (Std 10/Matric/Senior Certificate/ Form 5)	✓	✓
13	NTC 1	✓	
14	NTC 2	✓	
15	NTC 3	✓	
16	Certificate with less than Grade 12/Std 10	✓	✓
17	Diploma with less than Grade 12/Std 10	✓	✓
18	Certificate with Grade 12/Std 10	✓	✓
19	Diploma with Grade 12/Std 10	✓	✓
20	Bachelors Degree	✓	✓
21	Bachelors Degree and Diploma	✓	✓
22	Honours Degree	✓	✓
23	Higher Degree (Masters, Doctorate)	✓	✓
24	Other (Specify)	✓	✓
25	No Schooling	✓	✓
27	National Certificate Vocational 2 (NCV 2)		✓
28	National Certificate Vocational 3 (NCV 3)		✓
29	National Certificate Vocational 4 (NCV 4)		✓
30	N1 (NATED)/ NTC 1		✓
31	N2 (NATED)/ NTC 2		✓
32	N3 (NATED)/ NTC 3		✓
33	N4 (NATED)		✓
34	N5 (NATED)		✓
35	N6 (NATED)		✓

5.5 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 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 which used *pid* equivalents instead of t*pcodes*. Based on the above reasoning, all the *pcode* variables in Wave 1 have been dropped.

5.6 Surveyed vs. Historical Data

In Waves 4 and 5 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 applied where there were consistent responses to the questions across waves. Where this was the case, the historic data was used in the Wave 4 and 5 data files. In order for users to differentiate between this historical data and the data which was collected in Waves 4 and 5, flag variables have been created. An example of this is *w4_a_brnprov_flg*.

6 Derived Variables

Certain variables are created by the NIDS team. These variables appear in the hhderived and indderived data files. 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 the 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 these variables or variable groups is discussed below.

6.1 Best Variables

Certain information should remain unchanged or at least internally consistent for individuals across waves. Examples include education, gender, population group, date of birth, and age. We might get better data in a subsequent wave or we may get no data if the respondent is not interviewed. 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, the "best" variable 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. Where necessary, additional within-wave calculations are done for variables that will be included in 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.

Wave 5 saw an update to the calculation of best education. In previous releases Grad 0\R was excluded from the calculation. This has been corrected and Grade 0\R is now included in the data for all waves.

6.2 Geography

The Global Positioning System (GPS) information is used to determine 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 each of the next waves. The reason for the shorter distance between the later waves is due to built-in GPS systems being used in these waves which allows 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 Standard Classification of Occupations (SASCO). 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, codes from the 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) at http://www.ilo.org/public/english/bureau/stat/isco/. ISCO coding has been used for all five waves of NIDS for consistency.

A two-stage process is used to classify 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 undertaken 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 stage 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 for inclusion in the public release data. Occupational codes up to the four-digit level are available in the Secure Data.

To highlight the adoption of ISCO in all waves the variables have been renamed to reflect this change as shown in table 6.1.

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

The industry codes used are those found in Statistics South Africa's General Household Survey (2005) industry code list. These codes link the main goods or services provided by the employer to the industry description.

These codes were truncated to the one-digit level and included in the public release data.

6.5 Employment Status

Employment status is coded using the ILO'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, or personal agricultural work, or if they assist others in business activities. Unemployment is differentiated into broad and narrow unemployment according to the standard definitions, by distinguishing those who are actively searching for work and those not actively searching.

6.6 Admin Data

The Admin data file is a data file produced by NIDS in which we match the data we collect in field to external administrative data such as the Master schools list published by the South African Department of Basic Education (DBE).

6.6.1 School's Admin Data

The Admin data files contain school level data for individual records where we are able to match the school name in the NIDS data to school names on the DBE's Ordinary School's Master List, available from the DBE'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 school's education phase.

A scrambled school identifier based on the DBE's unique Education Management Information Systems (EMIS) number for the school is included in the anonymised public release *Admin data* file. 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

Police station districts and location data were published by the South African Police Service (SAPS) in 2015. These have been matched to the NIDS data and included in the data for each wave. The police station data, which is at a household level, was added to the *Admin* data file 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 straight line distance to the district police station as well as the straight line distance to the nearest police station. NIDS assigned each police station in the country a unique identifier which we call the police ID. These police IDs and the banded distances generated by NIDS are included in the public release data. Variables included in the Secure Data are the GPS coordinates, the police station names, and the numerical distance up to 6 decimal points from households to their nearest and district police stations.

6.7 Financial literacy

Wave 5 saw the introduction of financial literacy questions. These questions are based on four topics related to financial literacy and were added to the Adult questionnaire (questions G38 – G42) .

The financial literacy topics on which these questions are based and the names of the associated variables in the *Adult* data file are shown in Table 6.2: Financial literacy topics and questions

Table 6.2: Financial literacy topics and questions

Financial literacy topic on which the question is based	**Variable name**
Numeracy (interest)	*w5_a_flint*
Inflation	*w5_a_flval*
Compound interest	*w5_a_flcomp1*
	w5_a_flcomp2
Risk diversification	*w5_a_flrisk*

Note: The full questions for each of the above variables, and the possible answers categories, can be found in the Wave 5 Adult questionnaire.

Further, following Klapper, Lusardi and van Oudheusden (2015), two new derived variables were created using these financial literacy questions and included in the indderived data file. First, a financial literacy score out of four was created from the five financial literacy questions. The variable is w5_flscore. An explanation of how the score was calculated follows. If the respondent answered one of the questions for a financial topic correctly, they received a point for that topic. Since w5_a_flcomp1 and w5_a_flcomp2 are both based on the compound interest topic, only one of these two questions needed to be answered correctly in order to get a point for that topic. The points for all the topics were then added together to calculate the score. “Refused” and “Don’t know” answers to questions count as incorrect answers. Respondents who had a “Missing” value for any of the five questions, also have a “Missing” answer for the score.

Second, a variable showing whether the score implies that the respondent is financially literate was created. This variable, w5_flyn, equates to “Yes” if the score is at least 3 out of 4 and “No” if the score is between 0 and 2.

6.8 Interviewer Demographics and Experience

The Wave 5 release sees the introduction of interviewer demographics and experience variables that have been added to the Wave 5 *hhderived* and *indderived* datasets. They provide researchers with insight into the interviewer who conducted the interviews, facilitating the analysis of interviewer effects on the data collection process.

6.9 Impact of the 2017 Top-Up on Income, Expenditure and Wealth

NIDS achieved low baseline response rates in predominantly white and Indian areas at baseline. The sample was further reduced between Wave 1 and 4 because of high attrition rates in these groups. In Wave 5 (2017) a sample top-up was undertaken. The aim of the top-up was to increase the number of white, Indian, and high income respondents.

To identify individuals who were added in the 2017 top-up, the variable w5_Y_sample (where Y denotes the relevant data file indicator) was created in all the Wave 5 data files (in the Link File, this variable is called *sample*). This variable identifies which sample households and individual respondents originated from. It takes on the value 1 for “2008 sample” and 2 for “2017 sample”.

The top-up sample has an impact on income, expenditure and wealth variables which is most notable in the processes of imputing missing values. Since the sample top-up was designed to sample higher income respondents, the top-up sample’s inclusion in, or exclusion from the imputation process influences derived values of missing values, particularly at the upper-end of the distribution.

For ease of use and clarity, two sets of income, expenditure, and wealth variables have been created for Wave 5: those including the top-up sample and those excluding the top-up sample. These variables are found in the hhderived and indderived data files. The income, expenditure, and wealth variables that exclude the sample top-up have the suffix “extu”, and those which include the sample top-up do not have a suffix.

If the user wants to use any of these variables for panel analysis, then it is recommended that they use the relevant variable with the “extu” suffix. If the user wants to use any of these variables for cross-sectional analysis, it is recommended that they use the relevant variable without the “extu” suffix.

6.9.1 2017 Top-Up and Imputed Variables

Using the correct version of variables is of special importance for analysing imputed income, expenditure, and wealth variables.

When using income, expenditure, or wealth variables for panel analysis, we recommend using the variables with the “extu” (excluding top-up) suffix. As these values exclude the top up sample in the imputation of item non-response . If the variables including the top-up sample were used instead, an increase in the values for those in the upper-end of the distribution will likely be observed when comparing Waves 1-4 to Wave 5. This may be driven by a calculation change (of derived missing values) rather than an actual change.

Conversely, the use of variables without the “extu” suffix include the top-up sample and are recommended for cross-sectional analysis. This is because the top-up sample was designed to top-up the types of respondents who had a higher probability of having attrited between Waves 1 – 4, their inclusion provides a more representative cross-sectional sample in Wave 5

6.10 Income

Total household income (*w**X**_hhincome)* is derived from variables in the *Adult*, *Proxy* and *Household* data files. 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 are successfully interviewed, *w**X**_hhincome* is the aggregation of all income sources for all individuals in the household. If, however, an adult respondent refuses to be interviewed or is not available, we use the so-called "one-shot" income variable *w**X**_hhq_incb* as the measure of household income. Finally, for households where there is partial unit non-response and one-shot income is 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 6.3 shows how income was aggregated in all waves.

Table 6.3: Sources of aggregation

Wave Number	Source of HH Income	Number of HHs	Percent
W5	Individual aggregation	9457	87.23
	One-shot	1385	12.77
	Total	10842	100
W4	Individual aggregation	8836*	91.90
	One-shot	779*	8.10
	Total	9615	100
W3	Individual aggregation	7134	88.83
	One-shot	897	11.17
	Total	8031	100
W2	Individual aggregation	5508	81.17
	One-shot	1278	18.83
	Total	6786	100
W1	Individual aggregation	7111	97.46
	One-shot	185	2.53
	Total	7296	100

Table 6.44 lists the variables that make up each component of total household income. These variables are located in the indderived data file for each wave.

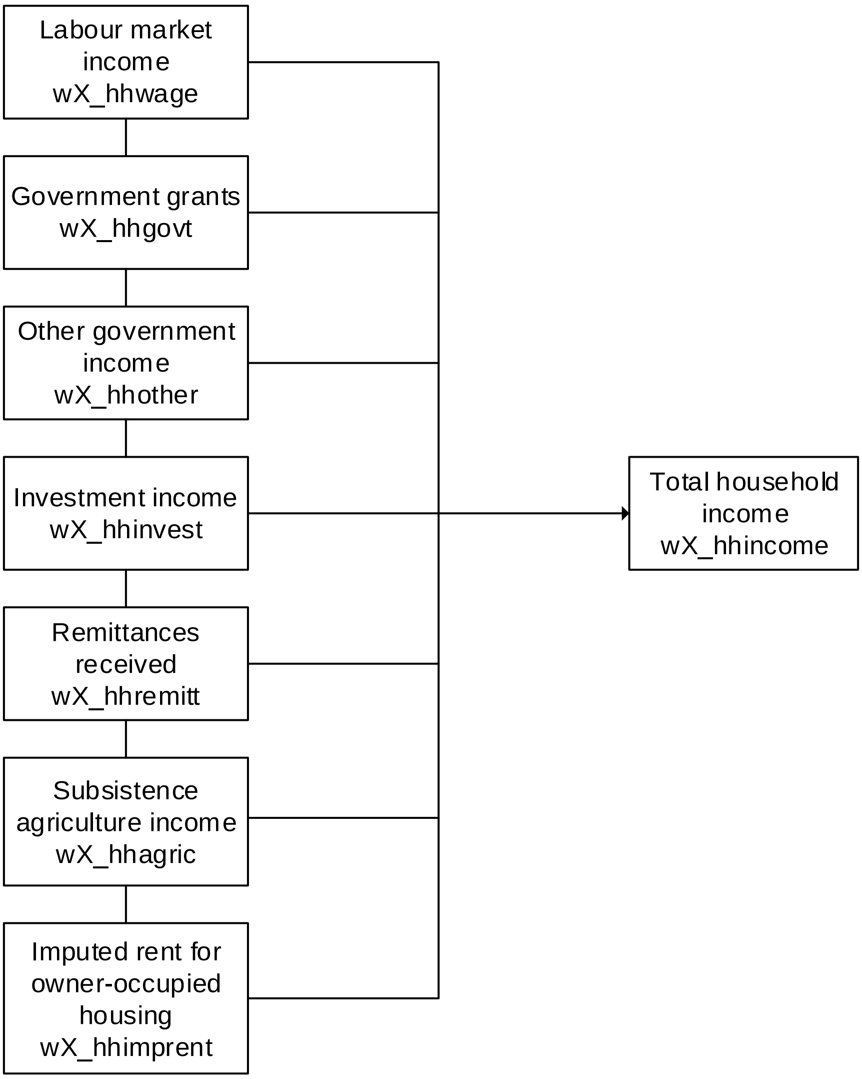
Table 6.4: Components of aggregate household income

Household-level Variable	**Individual-level Variable**	**Variable Name**
Labour Market Income *w**X**_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 *w**X**_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 *w**X**_hhother*	Unemployment Insurance Fund	*w**X**_uif*
	Workmen's compensation	*w**X**_comp*
Investment Income *w**X**_hhinvest*	Interest/dividend income	*w**X**_indi*
	Rental income	*w**X**_rnt*
	Private pensions and annuities	*w**X**_ppen*
Remittance Income *w**X**_hhremitt*	Remittances received	*w**X**_remt*
Subsistence Agricultural Income *w**X**_hhagric*[14]	Income from subsistence agriculture	*w**X**_plot*
	Value of own production consumed	*w**X**_opro*
Imputed Rental Income *w**X**_hhimprent*	N/A	N/A

The seven variables in the first column in Table 6.44 are summed to create aggregate household income. Figure 6.1 shows this aggregation.

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

Figure 6.1: Components of aggregate household income


Labour market income wX_hhwage
Government grants wX_hhgovt
Other government income wX_hhother
Investment income wX_hhinvest
Remittances received wX_hhremitt
Subsistence agriculture income wX_hhagric
Imputed rent for owner-occupied housing wX_hhimprent
Total household income wX_hhincome

6.10.1 Bracket Responses

For certain variables, if respondents are not able to provide a point estimate for the income from a particular source, a response is elicited through a series of unfolding brackets. Where respondents indicate that their income falls inside a bracket, the mid-point of the interval is assigned. Those who indicate that their income is above the value of the highest bracket are assigned twice the value of the upper bound of the top bracket[15].

6.10.2 Item Non-Response and Imputation

Item non-response occurs when the respondent refuses to answer a 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 where a few qualifying conditions are satisfied.

[15] Note that this practice 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.

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 included in the individual derived and household derived data files 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 6.5 summarises the variables imputed, the imputation method used to impute for item non-response, and percentage of missings for Wave 5.

Table 6:5: Wave 5 Income variable imputation[16]

Variable Name	Description	Imputation Method	Wave 5		
			Obs	Achieved	% Missing
w**X**_fwag	Main and secondary wages	Regression	7214	7595	5.01
w**X**_cwag	Casual wages	Regression	968	1006	3.78
w**X**_swag	Self-employment income	Regression	829	1245	33.41
w**X**_cheq	13th cheque	None	109	135	19.26
w**X**_prof	Profit share	None	13	17	23.53
w**X**_extr	Extra payment	None	51	60	15
w**X**_bonu	Bonus income	None	54	60	10
w**X**_othe	Other income	None	60	64	6.25
w**X**_help	Help friend income	None	46	50	8
w**X**_spen	State pension	Rule	2989	2994	0.17
w**X**_ppen	Private pension	Regression	485	531	8.66
w**X**_uif	UIF income	None	45	54	16.67
w**X**_comp	Workmen's compensation	None	24	24	0
w**X**_dis	Disability Grant	Rule	784	789	0.63
w**X**_chld	Child Support Grant	Rule	6059	6065	0.1
w**X**_fost	Foster Care Grant	Rule	319	323	1.24
w**X**_cdep	Care Dependency Grant	Rule	111	111	0
w**X**_indi	Interest/dividend income	None	65	73	10.96
w**X**_rnt	Rental income	Regression	247	255	3.14
w**X**_remt	Remittances	Regression	2541	2896	12.26
w**X**_hhimprent	Imputed rental income	Regression	6,720	8,951	24.92

[16] This table was generated using the full Wave 5 cross sectional sample, including both the original 2008 sample and the Top-up 2017 sample.

Table 6.6 summarizes the variables imputed, the imputation method used to impute for item non-response, and percentage of missings for Wave 4.

Table 6.6: Wave 4 Income variable imputation

Variable Name	**Description**	**Imputation Method**	**Wave 4**		
			Obs	Achieved	% Missing
*w**X**_fwag*	Main and secondary wages	Regression	6659	6914	3.72
*w**X**_cwag*	Casual wages	Regression	1051	1099	4.36
*w**X**_swag*	Self-employment income	Regression	852	1164	26.8
*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	2804	2932	4.37
*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	5634	5631	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	2369	2761	14.2
*w**X**_hhimprent*	Imputed rental income	Regression	6,056	8,108	25.31

Table 6.7 summarizes the variables imputed, the imputation method used to impute for item non-response, and percentage of missings, for Wave 3.

Table 6.7: Wave 3 income variable imputation

Variable Name	**Description**	**Imputation Method**	**Wave 3**		
			Obs	Achieved	% Missing
w**X**_fwag	Main and secondary wages	Regression	5266	5542	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	40	10
w**X**_help	Help friend income	None	47	48	2.08
w**X**_spen	State pension	Rule	2461	2462	0.04
w**X**_ppen	Private pension	Regression	321	341	5.87
w**X**_uif	UIF income	None	48	54	11.11
w**X**_comp	Workmen's compensation	None	14	15	6.67
w**X**_dis	Disability grant	Rule	718	721	0.42
w**X**_chld	Child Support Grant	Rule	4815	4817	0.04
w**X**_fost	Foster Care Grant	Rule	295	302	2.32
w**X**_cdep	Care Dependency Grant	Rule	103	104	0.96
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	1129	1309	13.75
w**X**_hhimprent	Imputed rental income	Regression	4,932	6,914	28.65

Table 6.8 summarizes the variables imputed, the imputation method used to impute for item non-response, and percentage of missings, for Wave 2.

Table 6.8: Wave 2 Income variable imputation

Variable Name	**Description**	**Imputation Method**	**Wave 2**		
			Obs	Achieved	% Missing
w**X**_fwag	Main and secondary wages	Regression	4007	4319	7.2
w**X**_cwag	Casual wages	Regression	528	541	2.4
w**X**_swag	Self-employment income	Regression	505	648	22.07
w**X**_cheq	13th cheque	Regression	154	227	32.16
w**X**_prof	Profit share	None	19	30	36.67
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	334	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	3442	3446	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	5,916	41.99

Table 6.9 summarizes the variables imputed, the imputation method used to impute for item non-response, and percentage of missings, for Wave 1.

Table 6.9: Wave 1 Income variable imputation

Variable Name	Description	Imputation Method	Wave1		
			Obs	Achieved	% Missing
wX_fwag	Main and secondary wages	Regression	3542	4492	21.15
wX_cwag	Casual wages	Regression	650	728	10.71
wX_swag	Self-employment income	Regression	663	951	30.28
wX_cheq	13th cheque	None	783	1204	34.97
wX_prof	Profit share	None	48	102	52.94
wX_extr	Extra payment	None	57	106	46.23
wX_bonu	Bonus income	None	341	550	38
wX_othe	Other income	None	18	18	0
wX_help	Help friend income	None	71	80	11.25
wX_spen	State pension	Rule	1972	2109	6.50
wX_ppen	Private pension	Regression	220	289	23.88
wX_uif	UIF income	None	81	122	33.61
wX_comp	Workmen's compensation	None	36	53	32.08
wX_dis	Disability Grant	Rule	837	869	3.68
wX_chld	Child Support Grant	Rule	2857	3388	15.68
wX_fost	Foster Care Grant	Rule	172	182	5.49
wX_cdep	Care Dependency Grant	Rule	44	47	6.38
wX_indi	Interest/dividend income	None	96	136	29.41
wX_rnt	Rental income	Regression	111	125	11.2
wX_remt	Remittances	Regression	1140	1140	0
wX_hhimprent	Imputed rental income	Regression	2,608	6,237	58.18

6.10.3 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 Wave 2, 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 syntax used to calculate agriculture income.

6.10.4 Bonus Payments

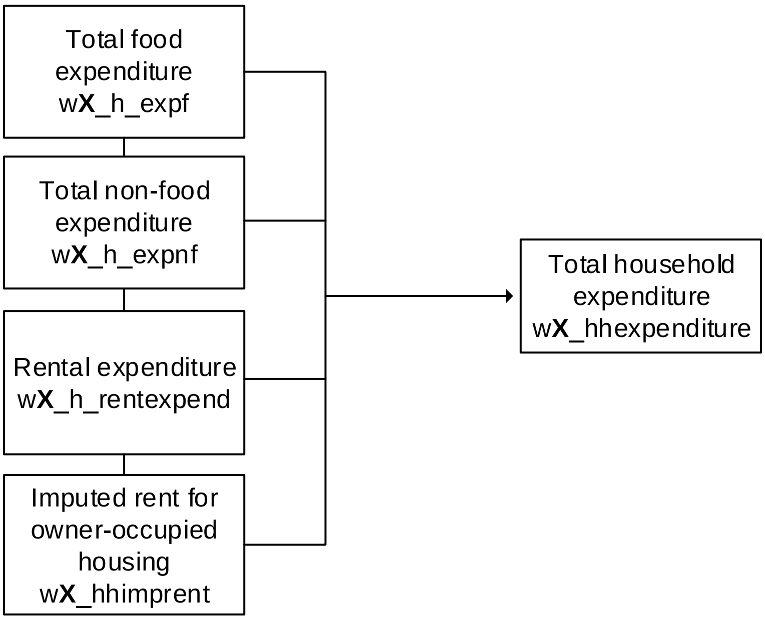
In Wave 1, respondents were asked about the value of 13th cheques, profit shares, and bonus payments received in the past 12 months. This amount was then divided by 12, to reflect an "average" monthly amount. In the Wave 2 Adult questionnaire, respondents were asked about receiving these sources of income 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, Wave 4 and Wave 5 asked for both annual and monthly amounts, and the latter was chosen to be consistent with data from Wave 2.

6.11 Expenditure

All expenditure data comes from the Household questionnaire. The respondent answering the Household questionnaire is asked about total household expenditure in the last 30 days for 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[17] (*w**X**_hhimprent*) to constitute aggregated total household expenditure (*w**X**_h_expenditure*).

[17] 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


Total food expenditure wX_h_expf
Total non-food expenditure wX_h_expnf
Rental expenditure wX_h_rentexpend
Imputed rent for owner-occupied housing wX_hhimprent
Total household expenditure wX_hhexpenditure

6.11.1 Imputations

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

6.11.1.1 *Food Expenditure*

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 expenditure on all food items.

From Wave 3, we asked for detailed food expenditure only if

1. The 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-responses
3. The household received food as payment or ate from own stock or grew their food themselves.

Because of this new rule applied in Wave 3, Wave 4, and Wave 5, we expect the number of missing observations to be the same for each food item in cases where the "one shot" variable is reported.

Table 6.10 shows how food expenditure was aggregated in all waves.

Table 6.10: Sources of aggregation

Wave Number	Source of HH Expenditure	Number of HHs	Percent
W5	One shot	9460	87.25
	Aggregated from food items	1302	12.01
	Imputed (One shot)	80	0.74
	Total	10842	100
W4	One shot	8630	89.76
	Aggregated from food items	955	9.93
	Imputed (One shot)	30	0.31
	Total	9615	100
W3	One shot	6587	82.02
	Aggregated from food items	1255	15.63
	Imputed (One shot)	189	2.35
	Total	8031	100
W2	Survey (One shot or Aggregated from food items)	6345	93.57
	Imputed	62	0.91
	No Data	374	5.52
	Total	6781	100
W1	Survey (One shot or Aggregated from food items)	7250	99.37
	Imputed	46	0.63
	No Data	0	0
	Total	7296	100

6.11.1.2 *Non-food Expenditure*

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.

6.11.1.3 *Rental Expenditure*

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

6.11.1.4 *Imputed Rental Income for Owner-occupied Housing*

This is the same variable that was discussed in the income section of this manual..

6.12 Wealth

The wealth section appears in questionnaires for Wave 2, Wave 4 and Wave 5 only. In this section we describe the derivation of household wealth (in Wave 2, Wave 4, and Wave 5) and individual wealth (in Wave 4 and Wave 5).

We define a household's (individual'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. A broader definition of each of these terms is provided in the following sections:

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/caravans) 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.

6.12.1 Wealth Questions 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.

- **Household questionnaire**
  - Section F2 establishes whether the household would be in debt, break even 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 refuses or do 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 house could be sold.
- A reasonable market value for which all other properties owned by the household could be sold.
- 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, break even 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:

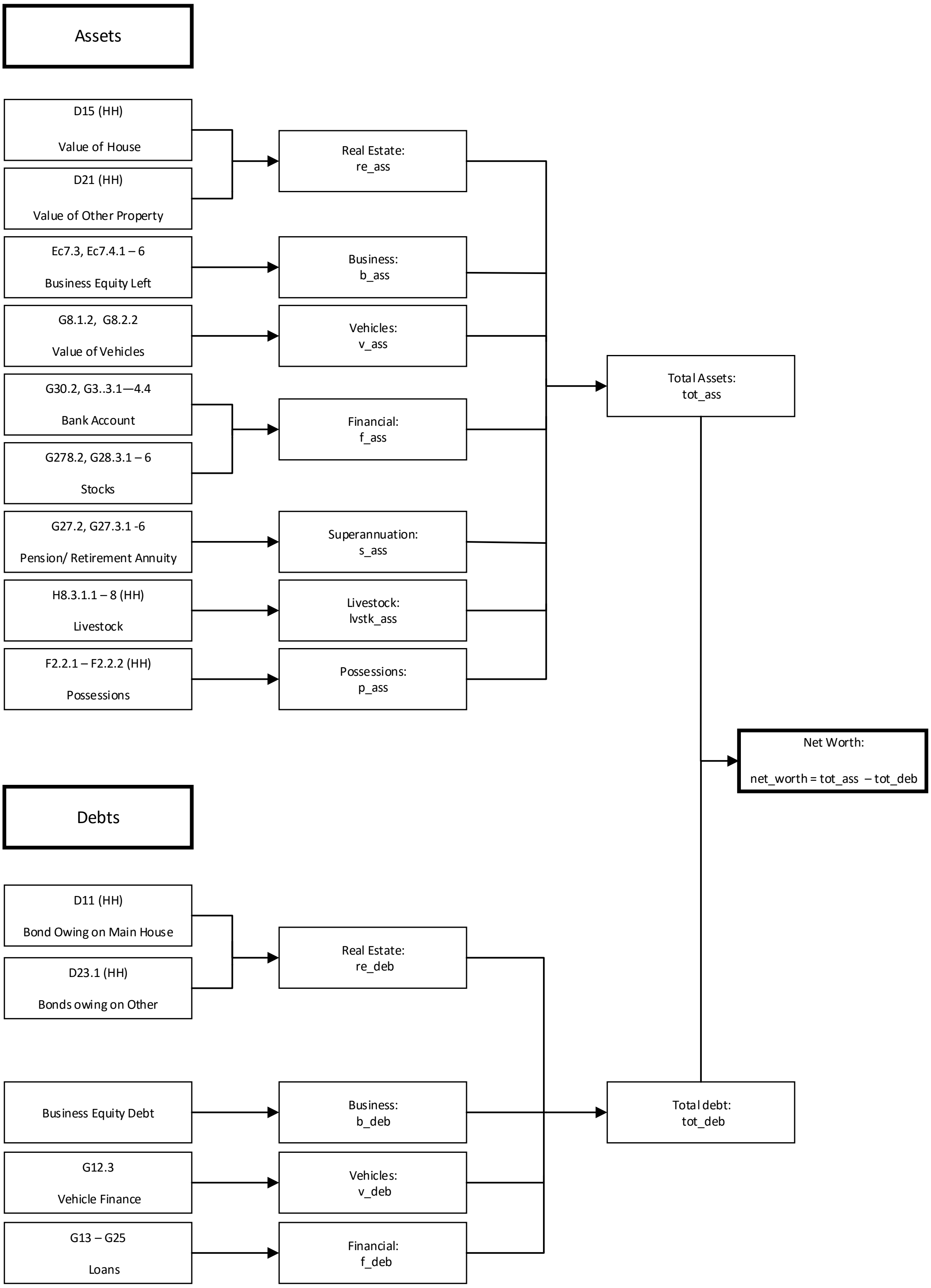
  - The value of all motor vehicles, bakkies/trucks and motorbikes owned by the respondent.
  - Home loans/bonds.
  - Other assets and debts, such as personal bank loans, store cards, and study loans
  - Vehicle finance.
  - Life insurance and unit trusts/stocks/shares.
  - Pensions/retirement annuities.

6.12.2 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 for imputing requires the number of reported observations to be 100 or more, and for the percentage of missing values to be at 40% or below.

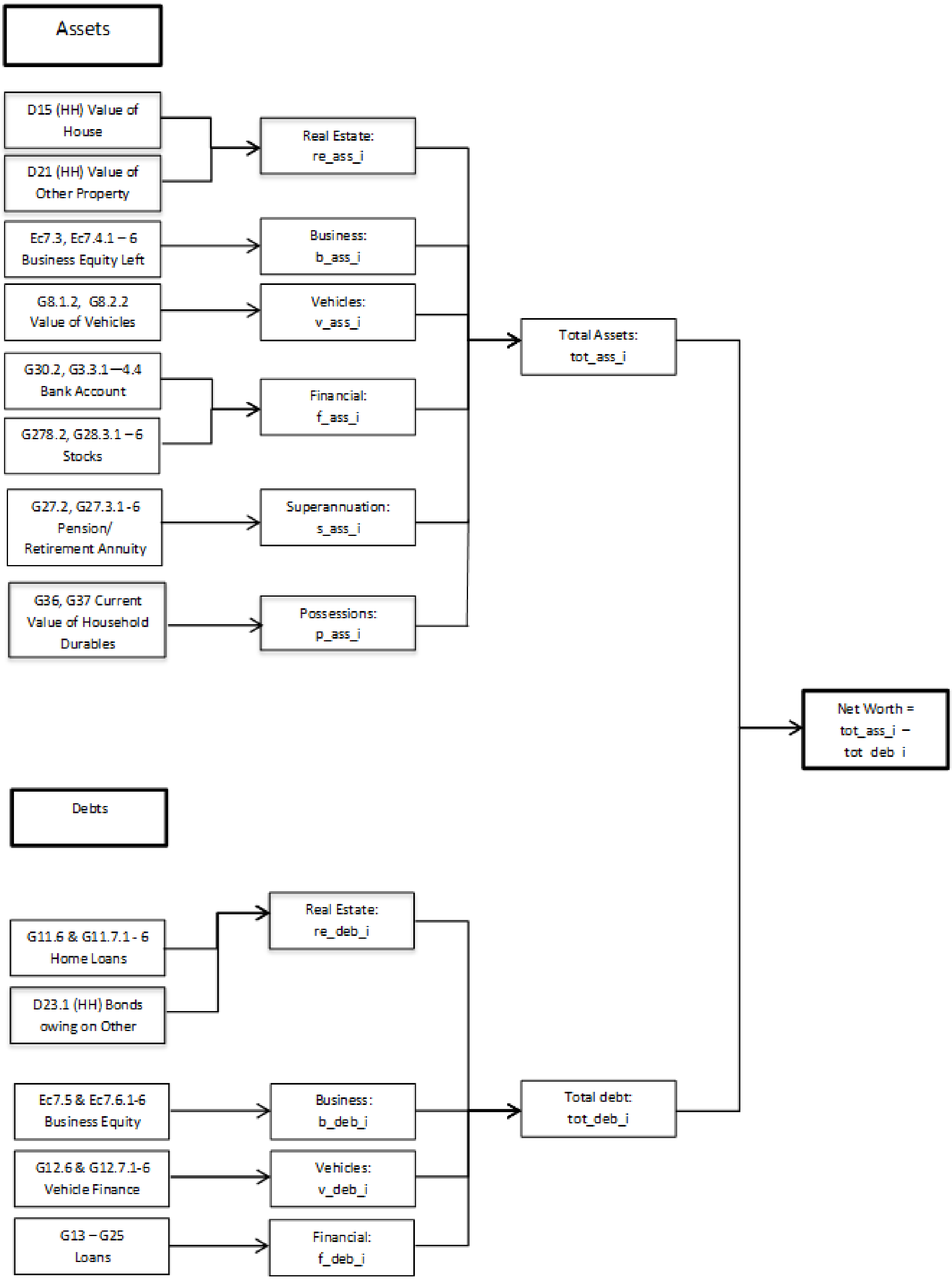
Figure 6.3 and Figure 6.4 outline how the final net worth for each household and individual is calculated.

Figure 6.3: Components of aggregate household wealth



Note: Question numbers in Figure 6.3 (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


Assets
D15 (HH) Value of House
D21 (HH) Value of Other Property
Ec7.3, Ec7.4.1 – 6 Business Equity Left
G8.1.2, G8.2.2 Value of Vehicles
G30.2, G3.3.1—4.4 Bank Account
G278.2, G28.3.1 – 6 Stocks
G27.2, G27.3.1 -6 Pension/ Retirement Annuity
G36, G37 Current Value of Household Durables
Real Estate: re_ass_i
Business: b_ass_i
Vehicles: v_ass_i
Financial: f_ass_i
Superannuation: s_ass_i
Possessions: p_ass_i
Total Assets: tot_ass_i
Net Worth = tot_ass_i – tot deb i
Debts
G11.6 & G11.7.1 - 6 Home Loans
D23.1 (HH) Bonds owing on Other
Ec7.5 & Ec7.6.1-6 Business Equity
G12.6 & G12.7.1-6 Vehicle Finance
G13 – G25 Loans
Real Estate: re_deb_i
Business: b_deb_i
Vehicles: v_deb_i
Financial: f_deb_i
Total debt: tot_deb_i

Note: Question numbers in Figure 6.4 (e.g. G30.2) refer to the Wave 4 question numbers. These may differ for other waves.

Additional things to note in the calculation of individual net worth in Wave 4 and Wave 5 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 are apportioned to members according to percentage ownership of these assets and debts. Questions of percentage ownership were added to the Household questionnaire in Wave 4 and are also included in Wave 5.
- Financial assets in Wave 4 and Wave 5 include bank account balance and value of stocks (unit trusts, stocks and shares) as shown in Figure 6.4. 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 were removed from Wave 4 onwards as these were considered too sensitive. The life insurance value question was also removed from Wave 4 onwards because, firstly, many respondents had no idea of the value of their insurance policy and, secondly, many tended to confuse this life insurance value with the pay-out value.
- In Wave 4 and Wave 5, 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. Negative values were not collected in Wave 2.

6.12.3 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 components of assets and liabilities 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 each question in the section or unit non-response for the individual), then we use the one-shot measure for household net worth. If all adults did not respond to the wealth module and the household one-shot question is also missing, household net worth is set to missing along with all components of household net worth.

6.12.3.1 *2018 Correction of Wave 4 Wealth Calculation*

During 2018 data production an error in the Wave 4 wealth calculation was identified and corrected. The error impacted the wealth component household possessions assets (p_ass) and subsequently the calculation of net worth in Wave 4 V1.0 and V1.1. However, this error has been corrected for Wave 4 V2.0.0 onwards.

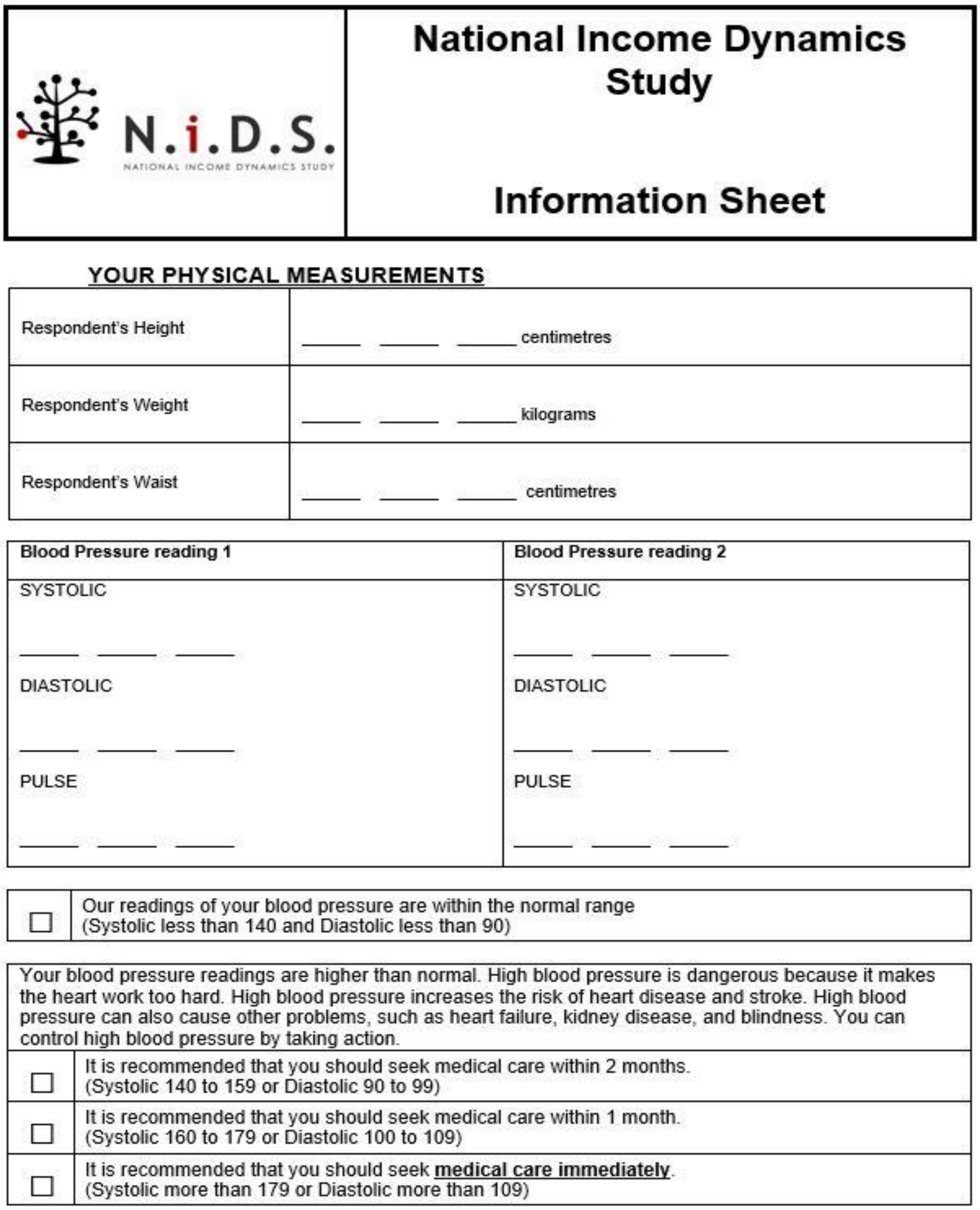
6.12.4 Outliers in Components of Net Worth

The NIDS Operations team has investigated outliers in Wave 5 using the blocked adaptive computationally efficient outlier nominators (BACON) algorithm (see Weber, 2010). Once outliers were identified, households were called to verify with respondents whether the values were indeed correct. If we could not contact the household, the values were left in the data. It is therefore the responsibility of researchers to conduct their own outlier detection checks.

6.13 Anthropometric Z-Scores

Anthropometric measures are collected using the Health Information Sheet as shown in Figure 6.5

Figure 6.5: NIDS health information sheet



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

National Income Dynamics Study

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, weight for age, and BMI for age,. The Stata macros *igrowup* and *who2007* are used to calculate the z-scores. These macros are available 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 therefore 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.

6.13.1 Using the public Release NIDS Data to Create Z-scores

NIDS has received a number of queries from users who have created z-scores using the publically released data 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. It has been estimated that using the WHO macros with age measured in days, 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 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.

6.14 Weights[18]

6.14.1 What is New?

Together with Wave 5 of the National Income Dynamics Study, updates to Waves 1-4 have been released. Since the information on the sample for these waves continues to be improved each wave[19] it has been necessary to recalculate **all** the weights previously released. In addition, this wave saw the inclusion of the top-up sample and led to an across-wave reassessment of the weight calculations. The three substantive changes that resulted are (i) there are two sets of calibrated weights for wave 5, one set inclusive of the top-up sample and one set exclusive of the top-up sample, (ii) the panel weights in this release are adjusted to account for both individual and household characteristics[20] that are predictive of attrition, and (iii) the panel weights are rescaled to sum to the StatsSA population totals in the survey year[21]. NIDS Technical paper 8 provides a comprehensive review of the weights methodology used in NIDS, combining and drawing insight from the documentation released with previous weights (Branson and Wittenberg, 2018). This document should be consulted for further information about the inclusion of the top-up sample.

[18] This section was drafted by Martin Wittenberg and Nicola Branson.

[19] Date of birth, gender, and population group data has been improved. TSM/CSM classification was updated as more relevant information came to light regarding household structures. Some households have also been removed in cases where we find the CSM had been interviewed in another house under a different pid. In these cases, the pid was standardised and the additional household dropped.

[20] In previous versions, only individual characteristics where used to predict attrition.

[21] This step was omitted in previous versions of the weight calculations.

Nevertheless, the **methods** used, i.e. the algorithms underpinning the calculations, have not been changed. This means that, while there will be some differences in the revised weights, they will be similar to the ones released previously.

6.14.2 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 6.6 presents the relationships between weights in diagrammatic form.

There are **three** types of weights:

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

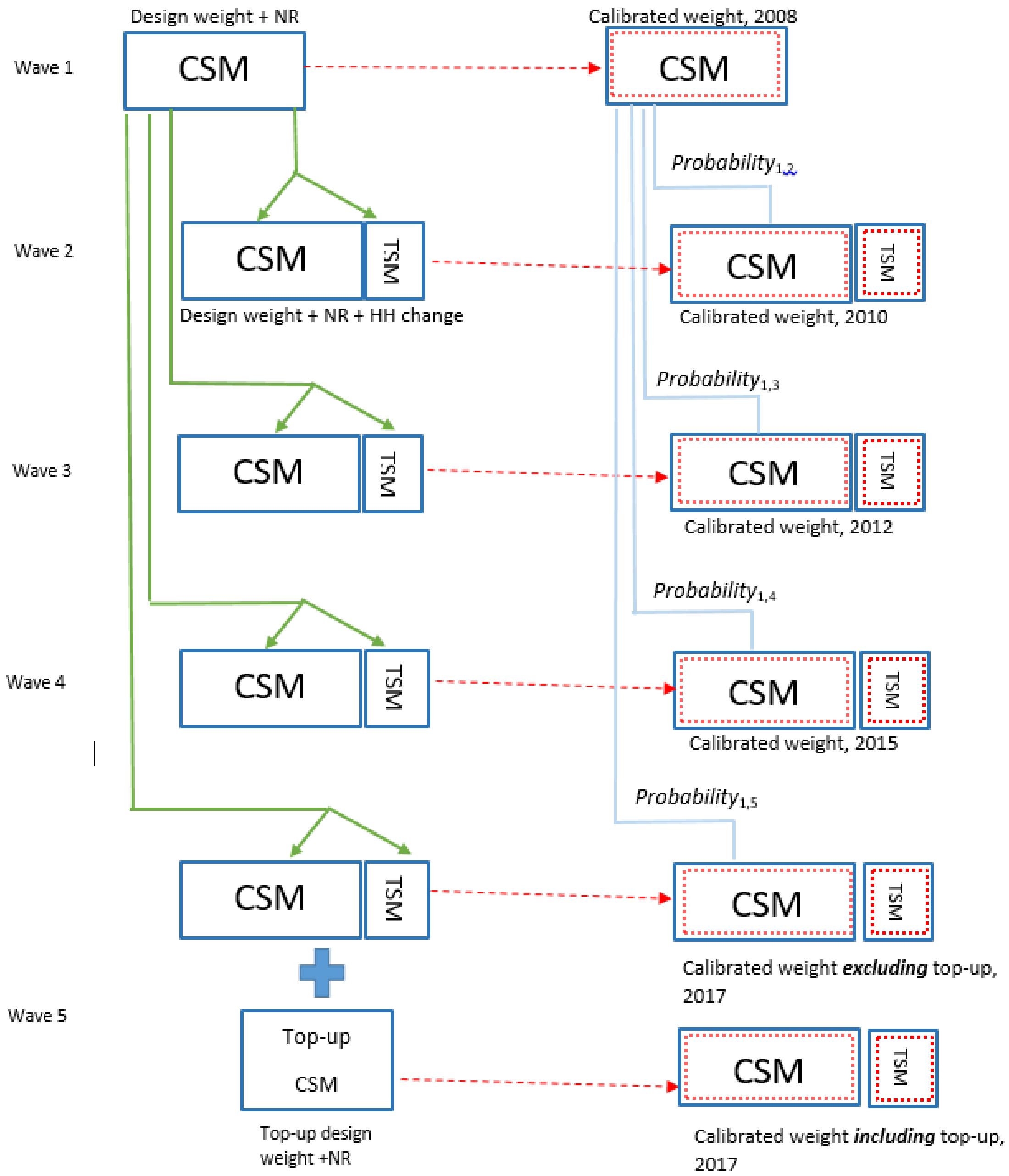
The design weights released with Wave 1 are fundamental to every other weight released with NIDS[22]. They are used to calculate the corresponding design weights for Waves 2-5 (the green arrows in Figure 6.6).

NIDS achieved low baseline response rates in predominantly white and Indian areas in 2008. The sample was further reduced between Wave 1 and 4 because of high attrition rates in these groups, especially between Waves 1 and 2. In Wave 5 (2017) a sample top-up was undertaken. The aim of this resampling exercise was to interview wealthier individuals of all race groups and in doing so increase the number of white and Indian households (Branson, 2018).

To identify individuals who were added in the 2017 top-up, the variable w5_Y_sample (where Y denotes the relevant data file indicator) was created in all the Wave 5 data files (this variable is simply called sample in the Wave 5 Link File and was also included in the Link Files of Waves 2 – 4). This variable identifies which sample households and individual respondents originated from. It takes on the value 1 for “2008 sample” and 2 for “2017 sample”.

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

Figure 6.6: The relationship between the different weights in NIDS



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 (Waves 1-4) and 2017 (Wave 5). This process is indicated in the diagram by the red arrows.

To work with changes over time we need to work with individuals that we observe at least twice. This means that we need to correct for attrition. In order to do this, the probability of observing the individual again is calculated. There are four such probabilities, shown in Figure 6.6:

- **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, 4 or 5

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 blue 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= 2, 3, 4 or 5) are included. In other words, panel weights 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[23] .

We now turn to a more detailed discussion of the different types of weights. Table 6.3: Weights provided in the NIDS data provides a list of the household and individual weights provided in the NIDS Wave 1-5 release, their variable name, which data file they can be found in and which waves they refer to.

Table 6.3: Weights provided in the NIDS data

Weight type	Variable	Data file	Wave/s
Design weight	*w**X**_dwgt*	hhderived	1, 2, 3, 4
Calibration weight	*w**X**_wgt*	hhderived	1, 2, 3, 4
Design weight (incl. top-up sample)	*w5_dwgt*	hhderived	5
Design weight (excl. top-up sample)	*w5_dwgt_extu*	hhderived	5
Calibration weight (incl. top-up sample)	*w5_wgt*	hhderived	5
Calibration weight (excl. top-up sample)	*w5_wgt_extu*	hhderived	5
Panel weight Wave 1 to Wave X	*w**X**_pweight*	indderived	2, 3, 4, 5

Note: In the above table, **X** denotes one of the wave numbers in the right-hand-most column.

[23] 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.

6.14.3 Design Weights

The individuals interviewed in Waves 2, 3, 4 and 5 included 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 (and that used to calculate the NIDS cross-sectional weights) is to average out sample weights within the Wave 2, 3, 4 or 5 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.

Finally, TSM babies are another subpopulation that was not part of the original frame when sampling took place in 2008. To increase the sum of the weights, TSM babies are given the same weight as other members of the household once the above two adjustments are made i.e. they are assigned the household design weight for the specific wave.

The Wave 1 household weights that were used as inputs for the "generalised share method" were the design weights corrected for non-response (i.e. w1_dwgt). The resultant wave specific weight (wX_dwgt) should be thought of as design weights corrected for non-response and for the reshuffling of household membership and births. 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 emigrate and who therefore no longer form part of the South African population.

6.14.3.1 *Wave 5 design weights*

Two sets of wave 5 design weights are included in the release, those including the top-up sample and those excluding the top-up sample.

Identifying the Top-up sample (the sample variable)

To identify individuals who were added in the 2017 top-up, the variable *w5_**Y**_sample* (where **Y** denotes the relevant data file indicator) was created in all the Wave 5 data files (in the Link File, this variable is simply *sample*). This variable identifies from which sample respondents originate. It takes on the value 1 for "2008 sample" and 2 for "2017 sample".

Table 6.3 shows the two categories of cross-sectional weights for Wave 5, **wgt_extu* and **wgt. w5_wgt_extu* weights were constructed on the original 2008 sample only as detailed above. Below we therefore provide details for the construction of the variables that combine the two samples, i.e. *w5_dwgt* and *w5_wgt*.

Design weight including top-up sample *w5_dwgt*

Given the aim of the top-up sample, the sampling frame was restricted to urban residential small areas (SALs) from the 2011 Census where the proportion of white residents was greater than or equal to 50% or the proportion of Indian residents was greater than or equal to 20%.

Similar to the main sample, the top-up sample involved two-stage sampling with stratification at the district council level. 48 households were selected per SAL (Branson, 2018).

Household response in the NIDS top-up was unprecedentedly low (Branson, 2018) . Table 6.4 shows that of the 8202 valid households located, only 1008 households (12%) were interviewed, with the overwhelming majority of households refusing to participate (72%).

Table 6.4: Wave 5 top-up household response

Top-up Households	n	%
Sampled	8752	
Dwelling unit vacant	536	6%
Not located	14	0%
Valid Households	8202	94%
Interviewed	1008	12%
Refused	5902	72%
No one at home	1295	16%
Incomplete	1	0%
Household away	1	0%

It is also worth noting that even once the household agreed to respond, individual response within the household was far lower than NIDS had previously experienced. Only 73% of listed individuals in participating households agreed to respond. As such, while the CSM sample was increased by 2775 individuals, only 2016 successfully completed interviews in Wave 5. No specific adjustment was made for individual level non-response within households.

Table 6.5: Wave 5 top-up individual response

Top-up Individuals (CSMs)	n	%
Existing	2775	
Interviewed	2016	73%
Refused	758	27%
Not Tracked	1	0%

Weights to combine the top-up and original sample in wave 5

Original sample members living in areas in the sampling frame used to select the top-up sample had a non-zero probability of being included in the top-up sample in addition to their original sample interview. To account for this we adjusted these individual weights downwards to ensure this group was not overestimated in our population estimates (Branson and Wittenberg, 2018).

6.14.4 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.

6.14.4.1 *Why is there a need to calibrate the weights?*

The "design weights" have solid theoretical credentials. Nevertheless, there are also good reasons to use 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 no 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.

6.14.4.2 *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.

Information from the calibration exercise shows that the sample has 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 too few prime-age males and too many women (Branson, 2018).

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 is also evident that the pattern seems to have become worse over time. This is probably due, in part, to differential attrition. The inclusion of the top-up sample in Wave 5 has alleviated some of these difficulties, especially for the Indian and white sub-samples (Branson, 2018).

6.14.5 Panel Weights

Individuals who were successfully re-interviewed in waves subsequent to the 2008 baseline are not a random subset of all the individuals surveyed in the first wave. The panel weights provided in the NIDS data are intended to correct for bias resulting from non-random attrition between Wave 1 and a subsequent wave. Table 6.6 provides the response rates of original CSMs by subsequent wave.

Table 6.6: Response rates by wave: CSMs only*

	Wave 1		Wave 2		Wave 3		Wave 4		Wave 5	
Existing	28 226		29 431		29 624		30 567		31 037	
Interviewed	26 776	95%	22 972	78%	24 336	82%	25 292	83%	24 758	80%
Refused	1 450	5%	691	2%	529	2%	433	1%	966	3%
HH level non-response	0	0%	4 628	16%	4 090	14%	2 460	8%	3 088	10%
Moved outside of SA	0	0%	51	0%	56	0%	19	0%	20	0%
Deceased	0	0%	876	3%	613	2%	745	3%	606	2%
Not tracked	0	0%	213	1%	0	0%	1 618	5%	1 599	5%

*Notes: TSMs and Wave 5 top-up members are not included in the sample used in Table 6.6.

The probability of being successfully interviewed in a subsequent wave (blue lines in figure 6.6.) was calculated given the Wave 1 characteristics of the individual and their household using a probit model. Population group, sex interacted with an age quartic, marital status, education level, province, household size, an indicator of whether they live alone or not, whether their household income is missing, geographical type in 2001, questionnaire type, intention to relocate, respondent attention during the interview, respondent attitude during the interview and Wave 1 phase were included as explanatory variables in this estimation.[24]

One of the regrettable features of the pattern of attrition is that particular categories of individuals who had a lower probability of being interviewed in Wave 1 also showed much higher rates of attrition. In the table in Appendix 1 we record the predicted probability of being successfully interviewed in each subsequent wave, according to the probit model. It is evident that Whites and Indians, particularly those in their twenties, had much lower probabilities of being re-interviewed than their African and coloured counterparts. to the NIDS technical paper on weights provides further information (Branson and Wittenberg, 2018).

It seems noteworthy that some of the probabilities are actually higher for a re-interview in Wave 3, 4 and 5 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 in these later waves than in Wave 2.

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 in the subsequent wave, 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

[24] Note that the list of controls was expanded in the construction of the weights for the Wave 12345 release to include household control variables and variables that take account of the respondent questionnaire type and attitude and attention during the interview. The inclusion of household characteristics is to account for the large amount of non-response at the household level.

represented by the calibrated weight, w1_wgt) and an attrition weight, i.e. the inverse of the conditional probability of being re-interviewed.

Some individuals with a high weight in Wave 1 also carried a high attrition weight, and this led to some extreme weights. In order to prevent avoidable errors, we decided to trim the weights to the 1st and 99th percentiles of the weight distribution.

Finally, the panel weights were further rescaled to add up to the StatsSA estimated total population of the survey year.

Given that these are individual level response adjustments, the panel weights are found in the individual derived files.

6.14.6 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 6.6 With the exception of the original Wave 1 and Wave 5 top-up 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.

Table 6.7: Response Probabilities

	Wave 1 - Wave 2							
	Male				Female			
Age group	African	Coloured	Indian	White	African	Coloured	Indian	White
0	0,854	0,799	0,853	0,775	0,847	0,808		0,806
1-4	0,866	0,799	0,780	0,720	0,860	0,798	0,797	0,629
5-9	0,884	0,794	0,663	0,617	0,873	0,782	0,755	0,562
10-14	0,892	0,796	0,640	0,474	0,883	0,800	0,718	0,487
15-19	0,832	0,703	0,487	0,341	0,830	0,718	0,562	0,329
20-24	0,800	0,676	0,489	0,303	0,823	0,727	0,492	0,328
25-29	0,774	0,652	0,588	0,389	0,829	0,724	0,536	0,363
30-34	0,765	0,685	0,600	0,381	0,847	0,757	0,617	0,459
35-39	0,774	0,680	0,632	0,457	0,856	0,772	0,608	0,487
40-44	0,781	0,712	0,682	0,507	0,876	0,793	0,651	0,499
45-49	0,818	0,731	0,744	0,554	0,886	0,807	0,692	0,564
50-54	0,839	0,747	0,644	0,535	0,897	0,832	0,724	0,580
55-59	0,852	0,784	0,710	0,589	0,917	0,828	0,737	0,561
60-64	0,885	0,796	0,723	0,572	0,924	0,843	0,708	0,574
65-69	0,917	0,799	0,728	0,569	0,931	0,835	0,742	0,560
70-74	0,931	0,789	0,615	0,542	0,925	0,829	0,686	0,493
75-79	0,927	0,778	0,505	0,610	0,929	0,829	0,650	0,501
80+	0,919	0,755		0,658	0,927	0,819		0,427

	Wave 1 - Wave 3							
	Male				Female			
Age group	African	Coloured	Indian	White	African	Coloured	Indian	White
0	0,887	0,846	0,888	0,849	0,884	0,870	0,915	0,807
1-4	0,897	0,867	0,828	0,786	0,895	0,863	0,852	0,719
5-9	0,908	0,864	0,734	0,667	0,901	0,855	0,756	0,580
10-14	0,914	0,869	0,707	0,552	0,907	0,868	0,721	0,539
15-19	0,856	0,795	0,547	0,379	0,856	0,799	0,589	0,358
20-24	0,825	0,775	0,545	0,393	0,850	0,806	0,536	0,330
25-29	0,797	0,761	0,562	0,431	0,851	0,796	0,537	0,345
30-34	0,785	0,770	0,658	0,427	0,867	0,816	0,698	0,426
35-39	0,800	0,766	0,701	0,471	0,875	0,818	0,664	0,465
40-44	0,814	0,786	0,728	0,543	0,897	0,838	0,708	0,545
45-49	0,848	0,808	0,742	0,576	0,902	0,862	0,757	0,586
50-54	0,873	0,816	0,669	0,562	0,918	0,884	0,762	0,604
55-59	0,894	0,847	0,719	0,572	0,935	0,891	0,789	0,585
60-64	0,927	0,869	0,738	0,600	0,941	0,903	0,740	0,626
65-69	0,952	0,882	0,742	0,611	0,948	0,911	0,773	0,629
70-74	0,964	0,895	0,653	0,634	0,941	0,920	0,701	0,617
75-79	0,957	0,905	0,575	0,729	0,944	0,908	0,710	0,621
80+	0,951	0,923	0,000	0,856	0,941	0,911	0,000	0,695

	Wave 1- Wave 4							
	Male				Female			
Age group	African	Coloured	Indian	White	African	Coloured	Indian	White
0	0,925	0,908	0,908	0,893	0,928	0,910	0,927	0,872
1-4	0,923	0,897	0,823	0,780	0,922	0,889	0,845	0,736
5-9	0,911	0,864	0,655	0,571	0,904	0,845	0,734	0,521
10-14	0,914	0,863	0,625	0,426	0,908	0,849	0,673	0,436
15-19	0,863	0,797	0,473	0,299	0,866	0,790	0,570	0,249
20-24	0,828	0,781	0,419	0,287	0,854	0,798	0,491	0,193
25-29	0,811	0,758	0,475	0,368	0,858	0,796	0,503	0,255
30-34	0,799	0,775	0,573	0,374	0,875	0,822	0,633	0,335
35-39	0,810	0,764	0,598	0,431	0,882	0,836	0,602	0,404
40-44	0,825	0,791	0,659	0,490	0,902	0,840	0,670	0,483
45-49	0,860	0,802	0,656	0,548	0,909	0,860	0,678	0,546
50-54	0,910	0,825	0,531	0,566	0,925	0,867	0,493	0,579
55-59	0,909	0,833	0,613	0,556	0,932	0,864	0,524	0,536
60-64	0,917	0,883	0,836	0,524	0,952	0,940	0,709	0,666
65-69	0,933	0,879	0,852	0,537	0,955	0,939	0,759	0,644
70-74	0,953	0,922	0,550	0,686	0,941	0,910	0,677	0,519
75-79	0,950	0,917	0,531	0,697	0,947	0,888	0,648	0,499
80+	0,951	0,919		0,729	0,950	0,898		0,593

	Wave 1- Wave 5							
	Male				Female			
Age group	African	Coloured	Indian	White	African	Coloured	Indian	White
0	0,919	0,895	0,915	0,884	0,920	0,897	0,906	0,862
1-4	0,906	0,867	0,816	0,735	0,907	0,871	0,841	0,684
5-9	0,873	0,799	0,623	0,444	0,872	0,810	0,725	0,420
10-14	0,866	0,786	0,554	0,263	0,872	0,812	0,653	0,322
15-19	0,816	0,720	0,450	0,141	0,829	0,751	0,534	0,185
20-24	0,779	0,714	0,368	0,164	0,815	0,757	0,481	0,114
25-29	0,769	0,698	0,443	0,222	0,828	0,759	0,451	0,176
30-34	0,765	0,733	0,455	0,250	0,847	0,787	0,557	0,234
35-39	0,786	0,743	0,488	0,326	0,859	0,798	0,556	0,311
40-44	0,803	0,772	0,597	0,384	0,889	0,807	0,623	0,373
45-49	0,844	0,792	0,587	0,431	0,901	0,827	0,617	0,428
50-54	0,899	0,843	0,540	0,477	0,920	0,843	0,531	0,464
55-59	0,901	0,839	0,536	0,466	0,929	0,838	0,584	0,431
60-64	0,912	0,883	0,841	0,480	0,938	0,915	0,611	0,612
65-69	0,927	0,884	0,855	0,463	0,942	0,914	0,633	0,568
70-74	0,925	0,923	0,524	0,635	0,932	0,876	0,709	0,453
75-79	0,925	0,919	0,535	0,664	0,939	0,858	0,644	0,461
80+	0,919	0,917		0,692	0,944	0,851		0,520

Notes to Table 6.7: Response Probabilities: Predicted probability of being successfully interviewed in a subsequent wave from a probit model including population group, sex interacted with an age quartic, marital status, education level, province, household size, an indicator of whether they live alone or not, whether their household income is missing, geographical type in 2001, questionnaire type, intension to relocate, respondent attention during the interview, respondent attitude during the interview and an indicator of Wave 1 phase. Deceased included as 'responders', those out of scope excluded.

7 Inclusion of Census 2011 Geographic Variables

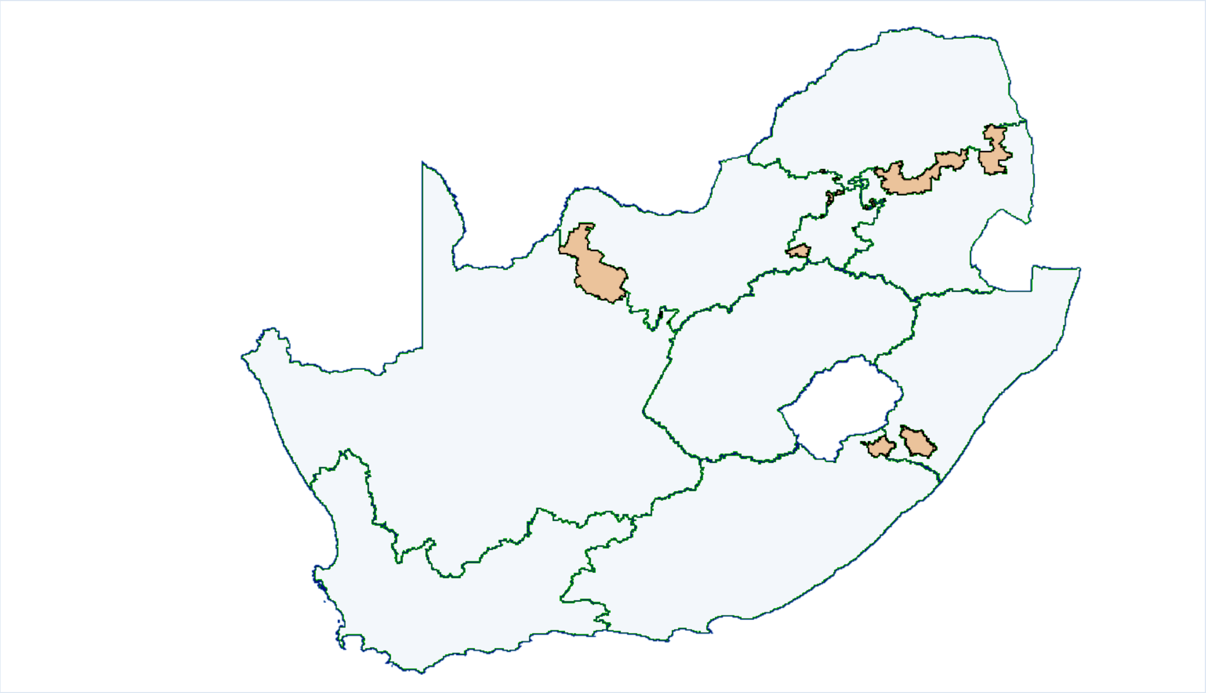
Statistics South Africa (StatsSA) made Census 2011 data publicly available in late 2013. This created the opportunity to update the geographical variables in the NIDS datasets. The household level geographic variables presented in NIDS are Province, District Council and Geo-type. The Secure Data also 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 includes important cautionary notes about the differences between the 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 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

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. The provincial codes have stayed the same.

Figure 7.1: Province changes between 2001 and 2011

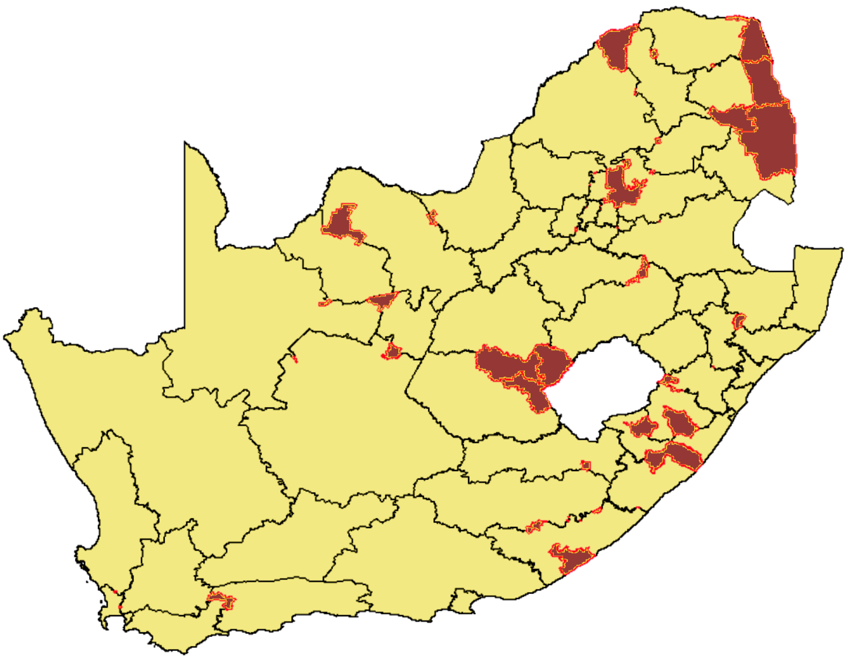


7.2 District Council Changes

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, the City of Cape Town used to be coded 171, but is now coded 199. 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 variable.

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.

Table 7.1 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 version of the data 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, version 1.0 of Wave 4 and version 1.0 of Wave 5, resulting in two sets of geographic variables being available at a provincial, district, and geographical type level in the household derived data file.

7.4.2 Impact of Geographic Variable Changes at an Individual Level

The 2011 variables were included in the migration section of the individual data as of version 6.0 of Wave 1, version 3.0 of Wave 2, version 2.0 of Wave 3, version 1.0 of Wave 4 and version 1.0 of Wave 5. 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* data files of all 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 for NIDS are calculated using the appropriate mid-year population estimates from StatsSA. 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 weight 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. This revision reflects the most accurate data available.

7.5.2 Imputed Income and Expenditure Variables

All derived files for Waves 1, 2, 3,4 and 5 use the 2011 geographic variables. The do-files are available in the Program Library provided 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 with the data on DataFirst's site, as well as on the NIDS website http://www.nids.uct.ac.za/nids-data/program-library

Two kinds of coding files are provided,(i) those that assist with data manipulation of the panel, and (ii) 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 taken 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 data files.

   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 data from the offspring questionnaires (Child and young Adults)

   This program uses the Adult data from any given wave, keeps the mother's 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* data files together and merges this appended data 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 of the mortality history in the *Household* data file. It does this by opening the *Household questionnaire* data file for any given wave, keeping the household identifiers and the mortality data and reshaping the data to create a roster.

8.2 Derived Variables

8.2.1 Income

As explained in section 6.10, 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 are available with the NIDS data on DataFirst's data site, or on the NIDS website here.

8.2.2 Expenditure

As explained in section 6.11, NIDS constructed a derived variable as a measure of total household expenditure in the 30 days preceding the interview. Do-files showing calculation of household expenditure are available with the NIDS data on DataFirst's site, or on the NIDS website here.

8.2.3 Wealth Program Library

As explained in section 6.12, 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 are available with the NIDS data on DataFirst's site, or on the NIDS website 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 with the NIDS data on DataFirst's site, or on the NIDS website here.

8.2.5 Employment Status

NIDS constructed a derived variable using the International Labor Organisation 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 with the NIDS data on DataFirst's site, or on the NIDS website here.

9 References

Branson, N. (2018). Methodology for the NIDS wave 5 top-up sample. Cape Town: SALDRU. (NIDS Technical Paper 9).

Branson, N. and Wittenberg, M. (2018). Longitudinal and cross sectional weights in the NIDS data 1-5. Cape Town: SALDRU. (NIDS Technical Paper 8).

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 A. B. Atkinson & F. Bourguignon (Eds.), *Handbook of income distribution* (Vol. 1, pp. 87 - 166): Elsevier.

de Onis, M., Onyango, A. W., Borghi, E., Siyam, A., Nishida, C., & Siekmann, J. (2007). Development of a WHO growth reference for school-aged children and adolescents. *Bull World Health Organ, 85*(9), 660-667.

Deville, J., & Lavallée, P. (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. *National Income Dynamics Study Technical Paper*(4).

Klapper, L. F., Lusardi, A., & van Oudheusden, P. (2015). *Financial literacy around the world: Insights from the Standard & Poor's Ratings Services Global Financial Literacy Survey*.

Lusardi, A., & Mitchell, O. S. (2014). The economic importance of financial literacy: Theory and evidence. *Journal of Economic Literature, 52*(1), 5-44.

Rendtel, U., & Harms, T. (2009). Weighting and calibration for household panels. *Methodology of Longitudinal surveys*, 265-286.

Weber, S. (2010). BACON: An effective way to detect outliers in multivariate data using Stata (and Mata). *Stata Journal, 10*(3), 331-338.

Wittenberg, M. (2009). Weights: report on NIDS Wave 1. *NIDS Technical Paper, 2*.

Wittenberg, M. (2010). An introduction to maximum entropy and minimum cross-entropy estimation using Stata. *Stata Journal, 10*(3), 315-330.

Wittenberg, M. (2013). *Fat tales of South Africa’s income distribution*. mimeo, School of Economics, 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*: World Health Organization.