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

Longitudinal and Cross-Sectional Weights in the NIDS Data 1-5

Technical Paper no. 9

Nicola Branson
Southern Africa Labour and Development Research Unit

Martin Wittenberg
School of Economics and DataFirst

2019

Longitudinal and Cross-Sectional Weights in the NIDS data 1-5

Nicola Branson and Martin Wittenberg

NIDS Technical Paper 9

1. Introduction

Weights are used to make inferences about the population from a sample by adjusting for unequal probabilities of selection and for non-response. Data users will typically use weights in tabulations, summary statistics and sometimes in regressions.

This paper describes the weighting methodology used in the construction of the National Income Dynamics Study (NIDS) sample weights. NIDS is a longitudinal household-based panel study that follows individuals over time. It began in 2008 with 7296 responding households and, in 2017, the sample was extended through the recruitment of an additional 1008 responding households. Individual interviews are conducted biennially for all members of households containing original CSMs, in person or by proxy for those aged 15 and over and by the caregiver for children under 15. One person also answers questions about the household as a whole. This document draws strongly on NIDS Technical Papers 2 and 6 by Martin Wittenberg and discussion of the weights in the NIDS Wave 1 - 4 user manuals with the aim of producing a single document about the NIDS weights as of Wave 5. Further detail and explanation of the development of the weights can be found in the original technical papers.

A series of longitudinal and cross-sectional weights are provided in the NIDS datasets. The individual panel weights are available in the individual derived files and the cross-sectional weights are available in the household derived files.

2. Cross sectional weights for Wave 1

Before analysis and report-writing on the NIDS data could begin it was necessary to calculate sampling weights. Professor Martin Wittenberg at the University of Cape Town was asked to calculate these weights for NIDS. Technical Paper Number 2 *Calculating the NIDS weights* details the methodologies and assumptions made when calculating the weights.

This is essentially a two-stage procedure. In the first stage, the design weights were calculated as the inverse of the probability of inclusion. In the second stage, the weights were calibrated to the Statistics South Africa (StatsSA) 2008 midyear estimates. Two sets of weights are thus provided, the design weights (and the design weight trimmed) and the calibration weights.

Note that the weights provided in Wave 1 are household level weights that can consistently be used at the individual level. Individual level response within households was high (95%) and therefore no person level adjustments were made.

2.1 Design weights (adjusted for household non-response): *w1_dwgt*

The basis of the calculation of the Wave 1 design weights is the information that StatsSA provided to NIDS about the process of two-stage sampling from their Master sample. Two sets of calculations were necessary in deriving the design weights. First there is a calculation of the probability of sampling each PSU and, second, there is the calculation of the probability of including a household in each PSU in the NIDS sample. The design weights provided in the data also correct for household non-response.

Household design weights

The probability of selecting each household into the sample was determined by the NIDS sample design (see NIDS Technical Paper 1: Methodology).

Households were selected into NIDS using a multi-stage area-based design. The probability of selecting a particular household was dependent on the following:

- The probability of a PSU appearing in the master sample
- The probability of the PSU being selected from the master sample; and
- The probability of the dwelling being selected given that the PSU was selected

The first probability was supplied by StatsSA. The derivation of the probability of the PSU being selected was complicated by the replacement of PSUs as a result of areas not being accessible. Technical Paper 2 (Wittenberg, 2009) provides details of the empirical difficulties involved in adjusting the weights for the replacement PSUs. The key difficulty arises from not knowing *ex ante* what the probability of fieldwork being possible in a PSU is – which therefore requires that an assumption be made – and that no information was provided for the probability of the replacement PSUs being selected.

In the weights provided with the data we therefore take the simplest approach: we assume that the probability that fieldwork is possible is constant within district council and that the replacement PSU's were draw randomly within district council. This allows us to simply replace the original PSUs with the replacement PSUs.

The design weight for a household was calculated as the inverse of the probability of selecting that household.

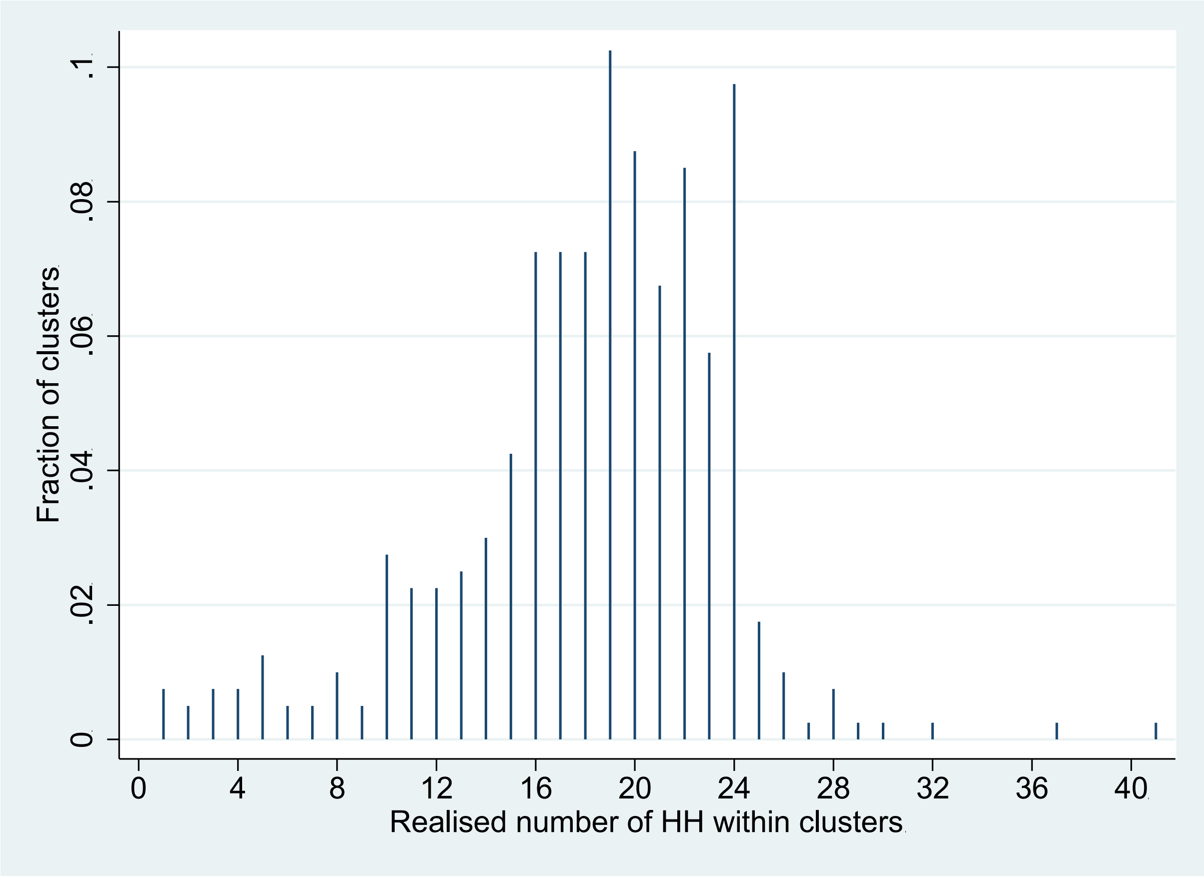
The main source of variability in the 'raw' design weights stems from the difference between the expected number of dwellings in the PSU (based on the 2006 StatsSA listing) and the actual number of dwellings in 2008. Manual listing of PSUs was done in all PSUs selected. Therefore dwellings built since 2006 had zero probability of selection and others had a probability of being vacant. Technical paper 2 (Wittenberg, 2009) provides a discussion on this.

Adjusting the design weights for baseline household non-response

The design weights presented in the NIDS data correct for baseline household non-response. Given the correlation between income, race and area, household non-response adjustments were made at an area level. Within each PSU it was assumed that non-response was random. As such, households that did respond within a PSU were weighted to represent those households that did not respond to the baseline using the inverse probability of the number of households sampled within the PSU.

Figure 1 shows the distribution of number of households attempted in each PSU. While the intension was to interview 24 (48) households per PSU, the realised sample was much lower. Given the very low numbers evident in some PSUs, this adjustment is the source of much of the variability in the 'design' weights presented in the data.

Figure 1: Realised number of households interviewed per PSU (cluster)



Trimming the design weights

In a final step the design weights were "trimmed" to reduce the influence of a few households with very large weights. These arose in PSUs in which only one or two households were interviewed. The weights were trimmed to the 95th percentile of the weights.

2.2 Calibration weights: *w1_wgt*

The second set of weights in Wave 1 are the calibration weights. These weights adjust the design weights such that the age-sex-race marginal totals in the NIDS data match the population estimates produced by StatsSA for the Mid-year Population Estimates for 2008 (see Technical Paper 2 for a discussion of the method). In addition, we imposed the constraint that the population distribution by provinces should correspond to that released in the StatsSA population estimates and that the total weights should add up to the estimated total population of 48,687,000. Finally, a further constraint imposed was that the weights should be constant within households.

Why is there a need to calibrate the weights?

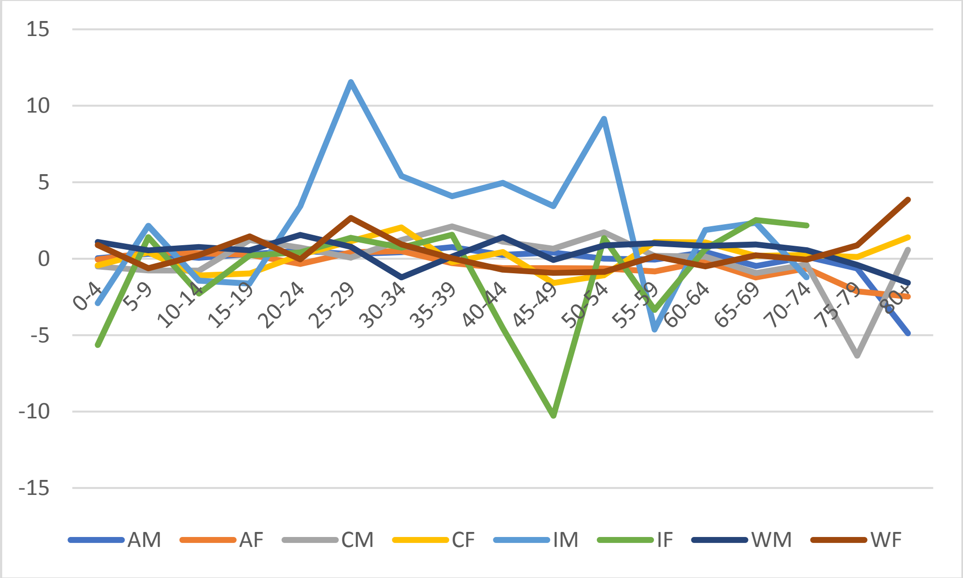
The "design weights" have solid theoretical credentials. Nevertheless, there are also good reasons for using the calibrated weights. Even when we adjust the design weights for household nonresponse, we find that the realised (weighted) sample differs from the national population in systematic ways. For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

Issues to take note of when using the calibrated weights

Nevertheless, getting the sample aligned with the national demography comes at a cost. It is much harder to find weights to align certain "cells" of the age-gender-race cross-tabulation with the national distribution than others. One measure of how far the weights had to be pushed from their baseline is given by the Lagrange multipliers that the ***maxentropy*** command returns. Values close to zero indicate that the constraint did not bind[1].

Figure 2 presents the λ values from the calibration exercise for each race-sex group across age. Large λ value are a sign that the constraint gave problems.

Figure 2: λ values from wave 1 calibration by age group, race and sex



Note to Figure 2: AM=African Male, AF=African female, CM=Coloured Male, CF=Coloured Female, IM=Indian Male, IF=Indian Female, WM=White Male and WF=White Female.

It should be noted that the sign of the multiplier is an indication of whether the weight associated with that group had to be increased (positive multiplier) or decreased (negative multiplier). As noted earlier, the sample shows a clear excess of older Africans and Coloured males age 75-79. It is also evident that the calibration had great difficulty with the Indian subpopulation. The general picture is that there seem to be relatively too few prime-age males and too many women. The fact that we also constrained weights to be common within household would have made this problem much more difficult, hence some of the rather large Lagrange multipliers.

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

The 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.

3. Weights in waves 2-5

3.1 Panel weights Waves 2-5: *wX_pweight*[2]

Individuals who were successfully reinterviewed 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 1 provides the response rates of original CSMs by subsequent wave.

Table 1: Response rates by wave: CSMs only

	Wave 1		Wave 2		Wave 3		Wave 4		Wave 5	
Existing	28226		29225		29453		30502		31021	
Interviewed	26776	95%	22972	79%	24337	83%	25291	83%	24759	80%
Refused	1450	5%	692	2%	375	1%	434	1%	967	3%
HH level non-response	0	0%	4629	16%	4074	14%	2456	8%	3086	10%
Moved outside of SA	0	0%	51	0%	56	0%	19	0%	20	0%
Deceased	0	0%	876	3%	611	2%	743	2%	604	2%
Not tracked	0	0%	5	0%	0	0%	1559	5%	1585	5%

Notes: TSMs and Wave 5 top-up members are not included in the sample used in Table 1. CSM babies are included. The increase in the percentage 'not tracked' in Wave 4 is a result of the decision to no longer attempt to contact respondents who had unsuccessful interviews in both Waves 2 and 3.

The probability of being successfully interviewed in a subsequent wave was calculated given the Wave 1 characteristics of the individual using a probit model. Population group, sex interacted with a quadratic in age, marital status, education level, province, household size, single household status, household income missing, geographical type in 2001, questionnaire type, intension to relocate, respondent attention, respondent attitude and Wave 1 phase were included as explanatory variables in this regression[3]. The reason for using age quartics rather than age dummies is to allow the probability to vary smoothly with age, which given the nature of age-related mortality is more appropriate.

One of the regrettable features of the pattern of attrition is that particular categories of individuals who had a relatively lower probability of being interviewed in Wave 1 also showed much higher rates of attrition. In the table in Appendix A we record the predicted probability of being successfully

[2] X denotes the relevant wave number.

[3] Note that the list of controls was extended in the construction of the weights for the Waves 12345 release to include household control variables and variables that take account of the respondent questionnaire type and attitude and attention. This adjustment was made since a large proportion of non-response is at the household level.

interviewed in each subsequent wave, according to the probit model. It is evident that Whites and Indians, particularly in their twenties had much lower probabilities of being reinterviewed than their African and coloured counterparts.

The panel weights are the inverse of the probability of appearing in the sample. This probability is the product of the probability of being interviewed in Wave 1, times the probability of being successfully reinterviewed in 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 represented by the calibrated weight, *w1_wgt*) and an attrition weight, i.e. the inverse of the conditional probability of being reinterviewed.

Given that some individuals with a high weight in Wave 1 also carried a high attrition weight, this led to some extreme weights. 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.

3.2 Cross-sectional weights for Waves 2 – 4

3.2.1 Household design weights Waves 2-4: *wX_dwgt*

Individuals interviewed in waves subsequent to wave 1 included both household members in the original sample (CSMs) as well as any new individuals 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. Those individuals who were born after Wave 1, CSM or TSM babies, were not part of the original universe and therefore their weight needs to reflect this. Assigning these new household members a weight requires differentiating between new CSM births, TSM babies and TSM adults.

New CSM births

New CSM births are a subpopulation that was not part of the original frame when sampling took place in 2008. If households did not get reshuffled, these respondents 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. NIDS however defines which new borns are classified as CSMs – those born to female CSMs – and they should therefore be thought of as indirectly sampled through their mothers. As such, their mother's weight is assigned to the new born CSM.

TSMs

TSMs born prior to 2008 were part of the original universe covered by the Wave 1 sample and therefore get allocated a share of the sampling weight attached to the individuals with whom they are now co-resident. The most straightforward procedure to share the household sampling weight and the one used to calculate the cross-sectional design weights in NIDS, is to assign TSMs a weight of zero, original household members their initial Wave 1 design weight adjusted for non-response and CSM babies their mother's initial weight and then calculate the average sample weight within the household. This then becomes the household design weight for the specific wave and therefore for the new TSMs.

TSMs born post-2008

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 to get an increase in the total population, TSM babies are given the same weight as other members of the household once the above to adjustment 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 (w**X**_*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 emigrated and who therefore no longer form part of the South African population.

3.2.2 Calibration weights Waves 2-4: *wX_wgt*

The wave specific design weights (*w**X**_dwgt*) were then calibrated to the mid-survey year population estimates released by StatsSA[4]. The data was calibrated to sex-race-age group cell totals (with the oldest three age categories for Indian males and Indian females collapsed) and provincial totals. The calibration was done using the Stata *maxentropy* add-in (Wittenberg 2010). Individuals within the same household were constrained to get the same weight. The resultant weights are contained in the variable *w**X**_wgt*.

As noted in section 2.2, getting the sample aligned with the national demography comes at a cost. It is much harder to find weights to align certain "cells" of the age-gender-race cross-tabulation with the national distribution than others. One measure of how far the weights had to be pushed from their baseline is given by the Lagrange multipliers that the ***maxentropy*** command returns. Appendix B presents the Lagrange multipliers from each of the calibrations by race, sex and age group. Large values indicate that the calibration had difficulty within these cells. Values close to zero indicate that the constraint did not bind[5].

[4] In Wave 2 and 4 the survey fieldwork ran over two years, i.e. 2010/2011 and 2014/2015 respectively. For these waves the 2011 and 2015 mid-year StatsSA estimates were used.

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

The sample shows a clear excess of old Africans and, indeed, Coloured males. It is also evident that the calibration had great difficulty with the Indian subpopulation. The general picture is that there seem to be relatively too few prime-age males and too many women. The fact that we also constrained weights to be common within household would have made this problem much more difficult, hence some of the rather large Lagrange multipliers. It might also be observed that the pattern seems to have become worse over time. This is probably due, in part, to differential attrition.

As noted in section 2.2, 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.

3.3 Cross sectional weights for Wave 5

NIDS achieved low baseline response rates in predominantly white and Indian areas. The sample was further reduced between Wave 1 and 4 due to high attrition rates among these groups, especially between Wave 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, 2019). Given the persistent income inequalities in South Africa, the method used to achieve this aim was to draw sample clusters from the Census 2011 that were predominantly white and Indian. Twenty-three years after democracy neighbourhoods are integrating and, as was the case in Wave 1, not all those sampled in specific areas are likely to be either white or Indian. On the other hand, these areas remain high income, the group with higher attrition rates in general.

3.3.1 Identifying the Top-up sample (the *sample* variable)

To identify individuals that 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”.

The top-up is not designed to be used as a standalone sample therefore the weights provided for Wave 5 are for users to either use the original sample or the original sample including the top-up sample. There are therefore two categories of cross-sectional weights for Wave 5 – those that are for the full Wave 5 sample, including both the original (2008) and the top-up (2017) sample members and those for the original sample only i.e. excluding the new Wave 5 top-up (2017) sample.

The two categories of cross-sectional weights for Wave 5 are as follows:

Table 2: Wave 5 weights

Weight type	Variable including top-up sample	Variable excluding top-up sample
Calibration weight	*w5_wgt*	*w5_wgt_extu*
Design weight	*w5_dwgt*	*w5_dwgt_extu*

The *w5_dwgt_extu* and *w5_wgt_extu* weights were constructed as detailed in section 3.2. Below we therefore provide details for the construction of the variables that combine the two samples, i.e. *w5_dwgt* and *w5_wgt*.

3.3.2 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 were the proportion of white residents was 50% or more or the proportion of Indian residents was 20% or more.

Similar to the main sample, the top-up sample involved two-stage sampling with stratification at the district council level. The number of SALs selected per district council was allocated proportionate to the number of households in the district council relative to the total number in the sampling frame. Geographical photography was used to list households in 84 SALs and the remaining 102 SALs were manually listed in field. 48 households were selected per SAL except for six SALs where there were fewer household were included as a result of low household density (Branson, 2019)[6].

As in Wave 1, the household design weight for the top-up households is therefore constructed as the product of the inverse of:

- The probability of a SAL appearing in the master sample
- The probability of the SAL being selected from the master sample; and
- The probability of the dwelling being selected given that the SAL was selected

Household response in the NIDS top-up was unprecedentedly low (Branson, 2019). Of the 8202 valid households located, only 1008 households (12%) were interviewed, with the overwhelming majority of households refusing to participate (72%). Indeed, there were 37 SALs where there was either no access or not a single household agreed to participate.

Table 3: 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%

[6] SALs with low household density according to the Census data were merged with neighbouring SALs. These six SALs found to have fewer than 48 households reflects change in household density since the census.

As a result of low response within SALs, an approach of using responding households to represent those that did not respond in an area as was done for Wave 1 was not appropriate. Therefore household non-response adjustments were made by SAL type. Whilst in field, the survey company classified each area as residential only, residential and estates mixed, residential and flats mixed, a combination of residential, flats and estates or a combination of flats, estates and compounds. Given that non-response was often driven by area type along these dimensions, e.g. response in estates was close to zero due to limited access to households, this was deemed the most appropriate measure available to adjust for household non-response.

It is also worth noting that even once the household agreed to respond, individual response within the household was far lower than NIDS had experienced in Wave 1. Only 73% of listed individuals in participating households agreed to respond. As such, while the CSM sample was increased by 2775 individuals, only 2016 of these additions completed interviews in Wave 5. No specific adjustment was made for individual level non-response within household.

Table 4: Wave 5 top-up individual response

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

3.3.3 Combining samples

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 needed to adjust these individual weights downwards so as not to allow this group to be overestimated in our population estimates.

We identified households in overlapping areas using their household geographic coordinates in Wave 5 and the sampling frame boundaries. The combined cross sectional design weight for Wave 5 including the top-up sample was constructed as:

w5_dwgt = *w5_dwgt_tu*	for those in the top-up sample
w5_dwgt = *w5_dwgt_extu*	for those in the original sample whose households do not fall within the top-up sampling frame
w5_dwgt = (*w5_dwgt_extu*+*w5_dwgt_tu*)/2	for those in the original sample whose households overlap with the top-up sampling frame

Where *w5_dwgt_extu* and *w5_dwgt_tu* are the Wave 5 design weights adjusted for household non-response for the original sample and top-up sample respectively.

4. Weights in the NIDS data release

4.1 Weights provided

Table 5 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 their refer to.

Table 5: 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.

4.2 Balanced panel weights

As is evident from Table 5, NIDS does not release a balanced panel weight as part of the data release.

4.3 Calculating standard errors

To obtain appropriate standard errors, NIDS users need to that take into account NIDS' complex survey design. While applying the weights in an estimation will correct point estimates, the appropriate standard errors and confidence intervals will only be calculated if the NIDS stratification and clustering are taken into account.

Stata has a suite of commands - 'svy' commands–that deal with complex survey designs. Using the 'svyset' command, users can assign the clustering, stratification and weights. Various statistical procedures are available within the suite of 'svy' commands including means, proportions, tabulations, linear regression, logistic regression, probit models and a number of other commands.

The NIDS original sample was stratified at the district council level (*w1_dc2001*) and clustered at the cluster level (*w1_cluster*).[7] Any new entrants to the household are assigned the same sample design information as the permanent sample member they join. In other words, in subsequent waves new TSMs members were assigned the cluster of the CSMs in the household they joined. New CSM babies were given the cluster of their mother. The overlap between Wave 5 top-up member households and original samples were mapped to ascertain overlap and new clusters added.

[7] StatsSA divided each PSU into clusters that were then allocated to different StatsSA surveys. Two clusters within each PSU were not used and these were provided to NIDS for the initial sample to be drawn.

There are 400 original clusters from Wave 1 and 144 new clusters in Wave 5. Please refer to technical paper 6 "A comment on the use of "cluster" corrections in the context of panel data" (Wittenberg, 2013) for a discussion of the underlying assumptions of using the original cluster variable and recommendations for practise.

To calculate the appropriate standard errors use the *cluster* and *w**X**_dc2001*[8] stratification variables. The name and location of these design variables are listed in Table 6.

```
svyset cluster wX_wgt, strata(wX_dc2001)
```

Table 6: Complex survey design variables

Variable description	**Variable**	**Data file**	**Wave/s**
Original wave 1 sample cluster	*cluster*	Link File	1, 2, 3, 4, 5
District Council (2001 Census)	*w**X**_dc2001*	hhderived	1, 2, 3, 4, 5

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

[8] Note the household derived files also contain district council variables according to the 2011 boundaries as per the Census 2011 and the Municipal Demarcations Board District Councils Census 2011 boundaries. The Wave 5 top-up sample was stratified on the Census 2011 district council boundaries but we use the 2001 to be consistent with the original sample. Use of the Census 2011 information should not have a substantive impact on analysis results.

Appendix A: 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

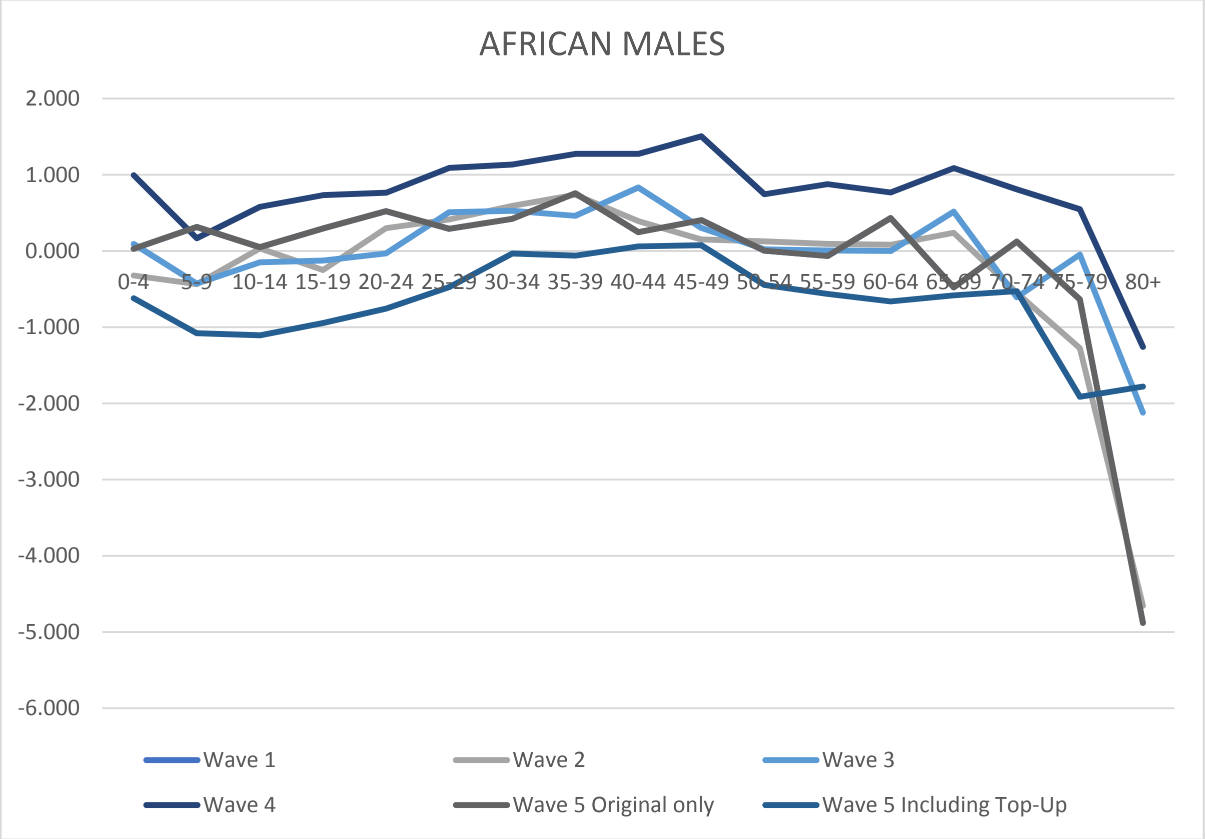
	Wave 1- Wave 3							
	Male				Female			
Age group	African	Coloured	Indian	White	African	Coloured	Indian	White
0	0,887	0,846	0,888	0,849	0,884	0,870	0,915	0,807
1-4	0,897	0,867	0,828	0,786	0,895	0,863	0,852	0,719
5-9	0,908	0,864	0,734	0,667	0,901	0,855	0,756	0,580
10-14	0,914	0,869	0,707	0,552	0,907	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

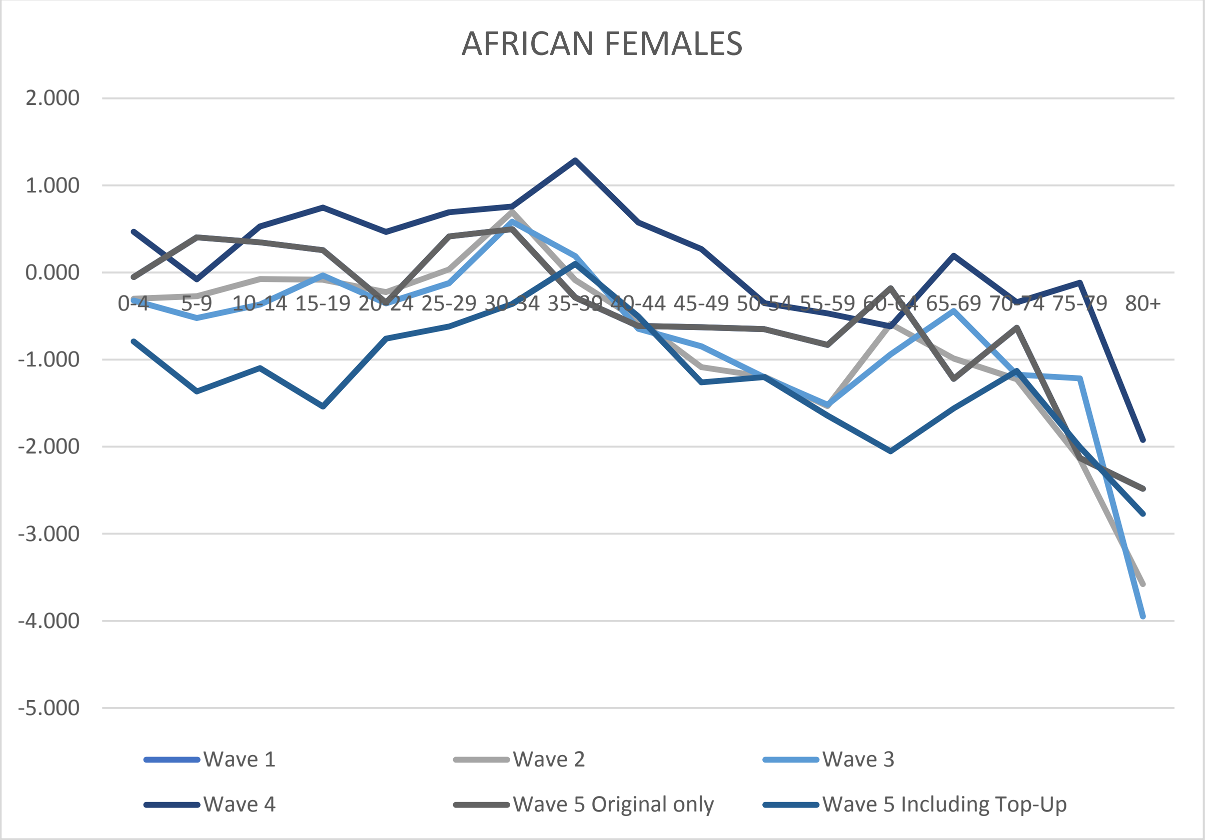
	Wave 1- Wave 4							
	Male				Female			
Age group	African	Coloured	Indian	White	African	Coloured	Indian	White
0	0,925	0,908	0,908	0,893	0,928	0,910	0,927	0,872
1-4	0,923	0,897	0,823	0,780	0,922	0,889	0,845	0,736
5-9	0,911	0,864	0,655	0,571	0,904	0,845	0,734	0,521
10-14	0,914	0,863	0,625	0,426	0,908	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

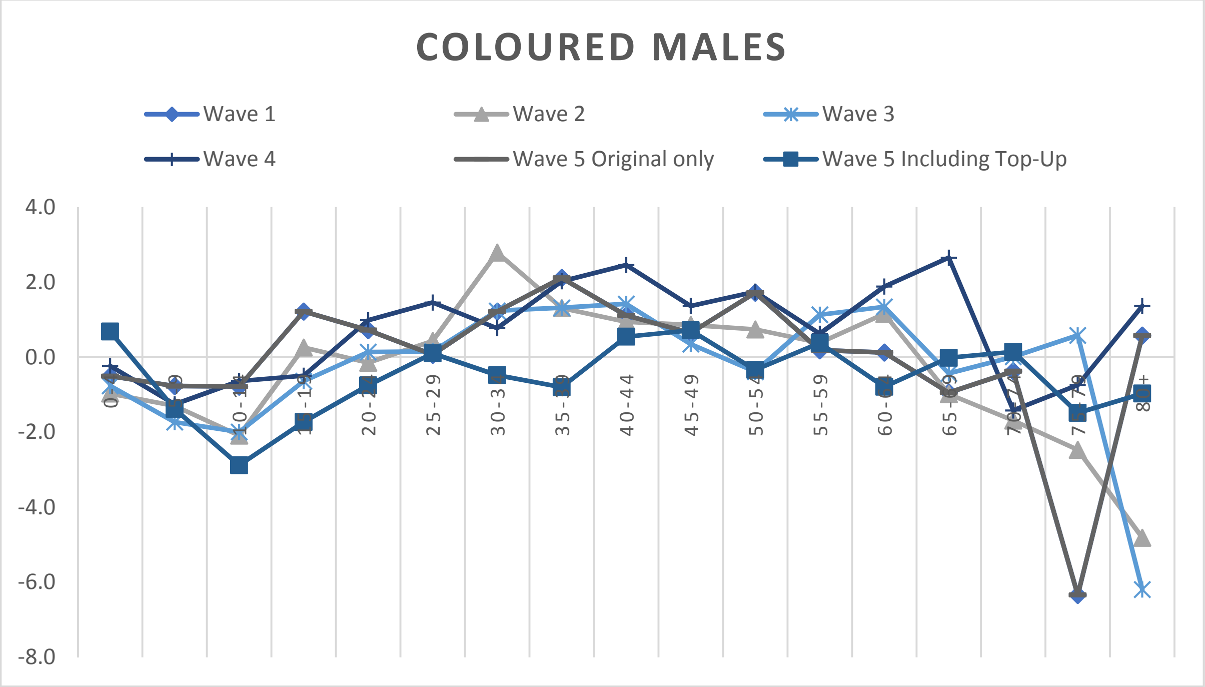
	Wave 1- Wave 5							
	Male				Female			
Age group	African	Coloured	Indian	White	African	Coloured	Indian	White
0	0,919	0,895	0,915	0,884	0,920	0,897	0,906	0,862
1-4	0,906	0,867	0,816	0,735	0,907	0,871	0,841	0,684
5-9	0,873	0,799	0,623	0,444	0,872	0,810	0,725	0,420
10-14	0,866	0,786	0,554	0,263	0,872	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

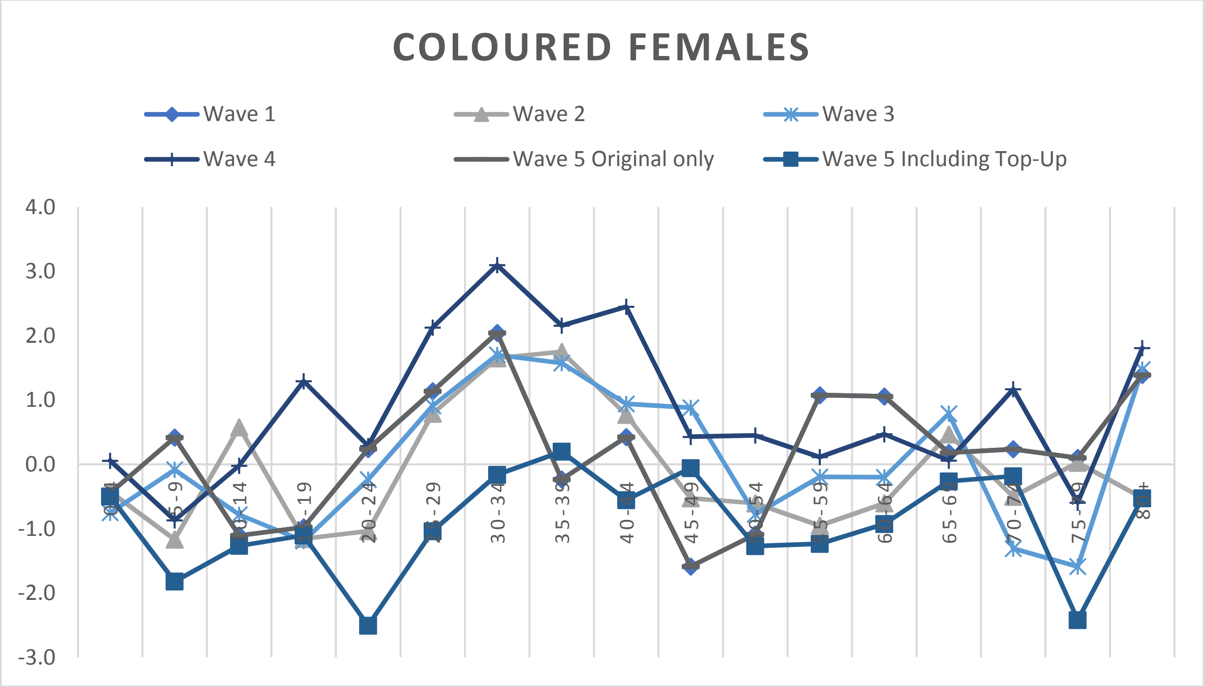
Notes to Appendix A: Predicted probability of being successfully interviewed in a subsequent wave from a probit model including population group, sex interacted with an age quartic, marital status, education level, province, household size, an indicator of whether they live alone or not, whether their 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.

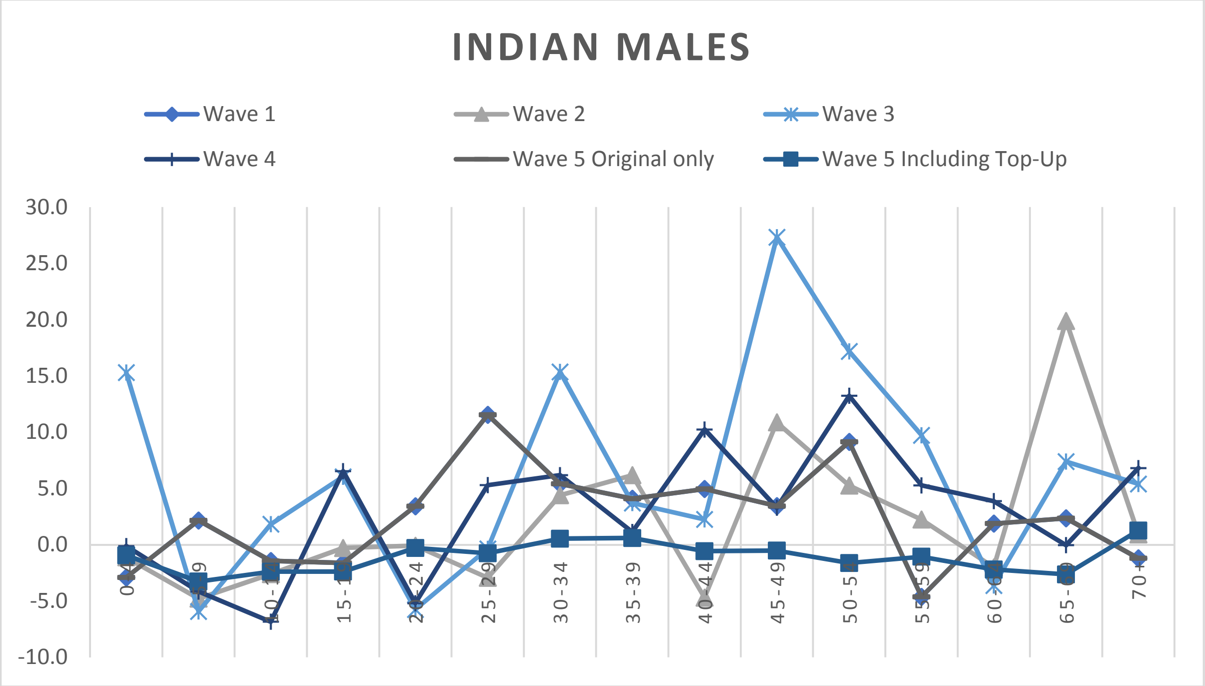
Appendix B: Lambda values from Calibration

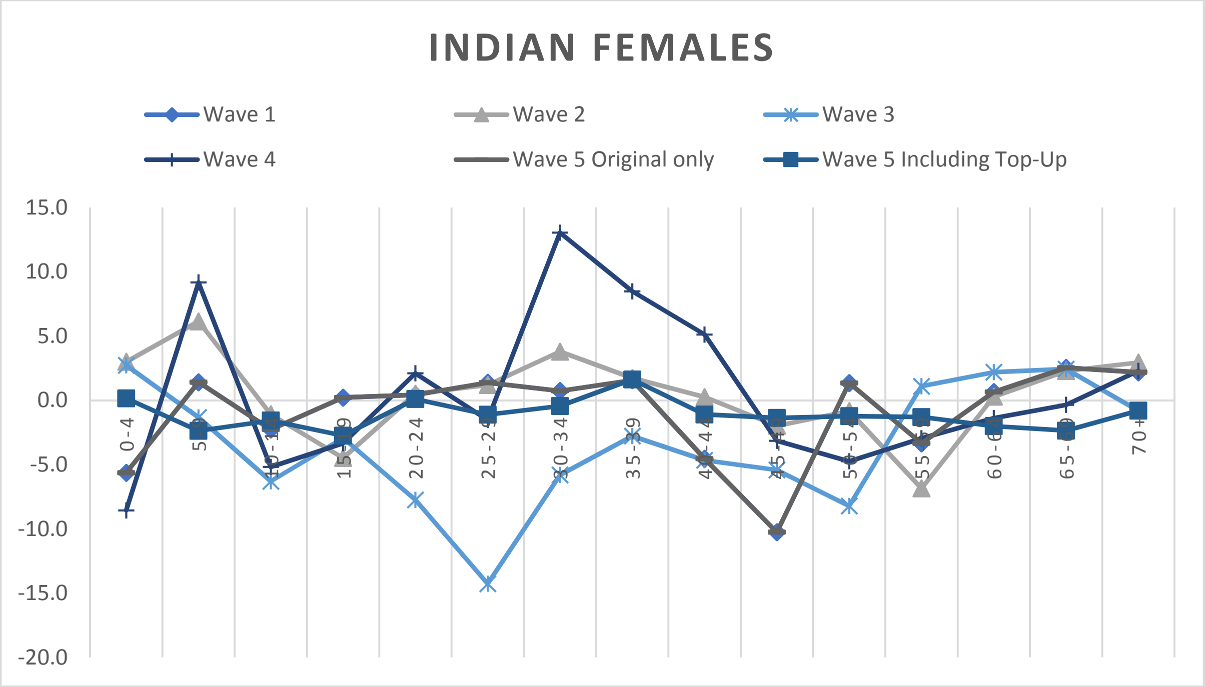


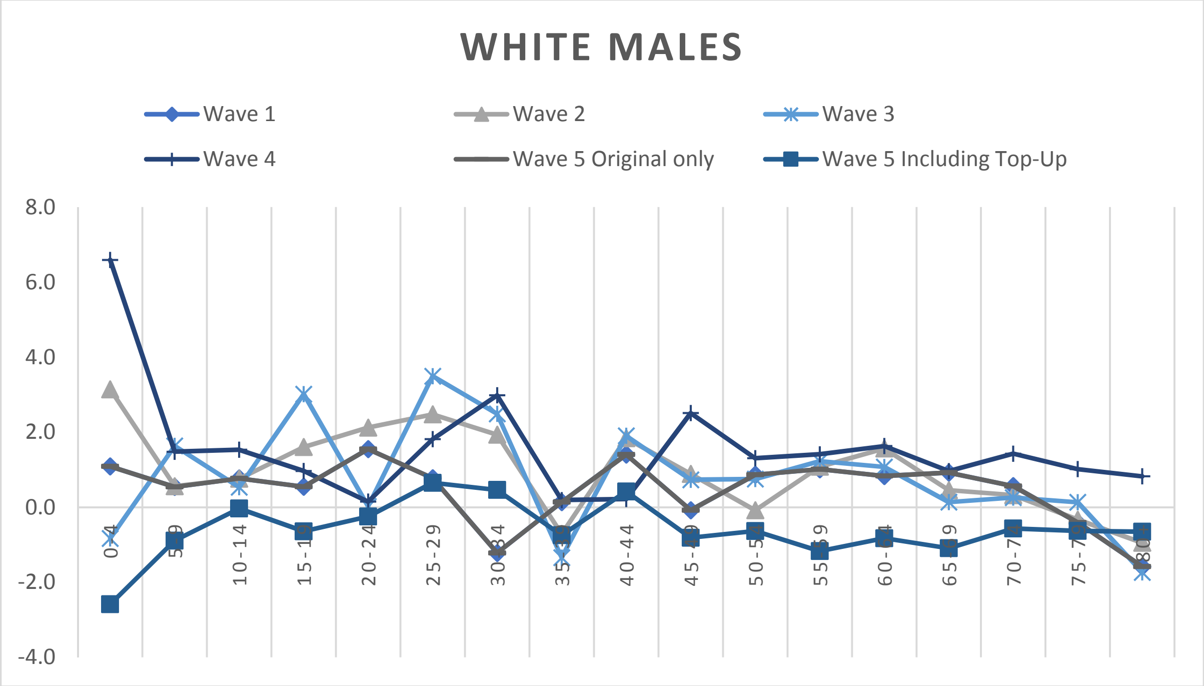


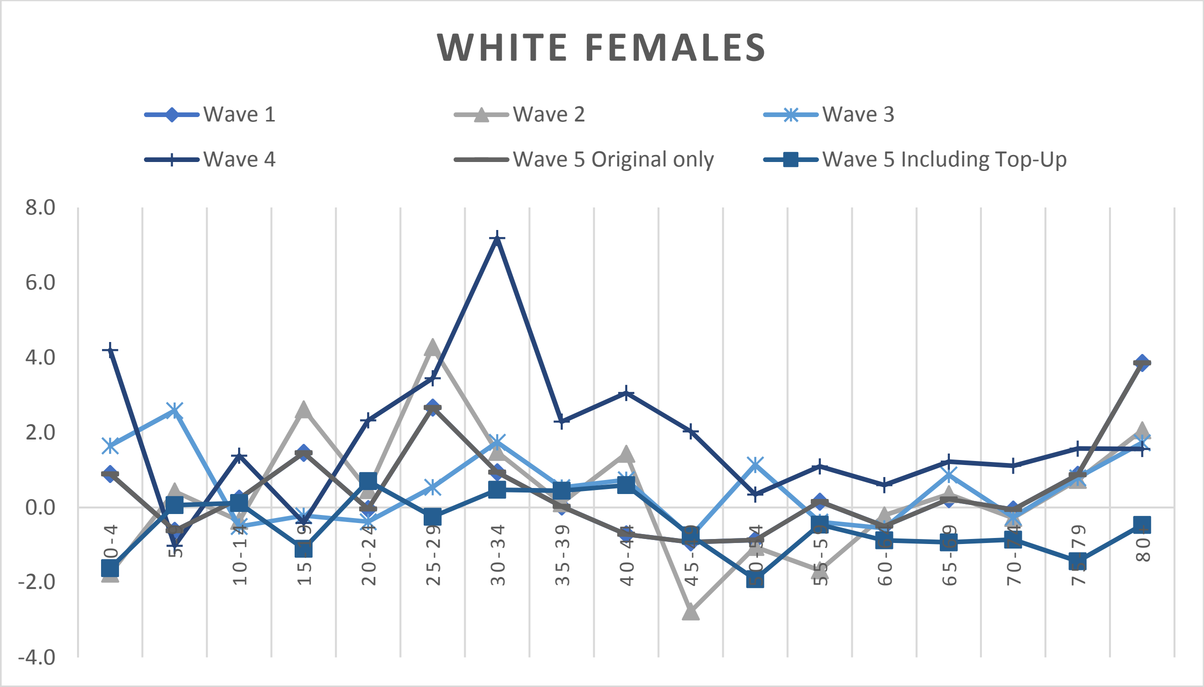

COLOURED MALES
Wave 1
Wave 2
Wave 3
Wave 4
Wave 5 Original only
Wave 5 Including Top-Up
4.0
2.0
0.0
-2.0
-4.0
-6.0
-8.0
0-4
5-9
10-14
15-19
20-24
25-29
30-34
35-39
40-44
45-49
50-54
55-59
60-64
65-69
70-74
75-79
80+


COLOURED FEMALES
Wave 1
Wave 2
Wave 3
Wave 4
Wave 5 Original only
Wave 5 Including Top-Up
4.0
3.0
2.0
1.0
0.0
-1.0
-2.0
-3.0
0-4
5-9
10-14
15-19
20-24
25-29
30-34
35-39
40-44
45-49
50-54
55-59
60-64
65-69
70-74
75-79
80+

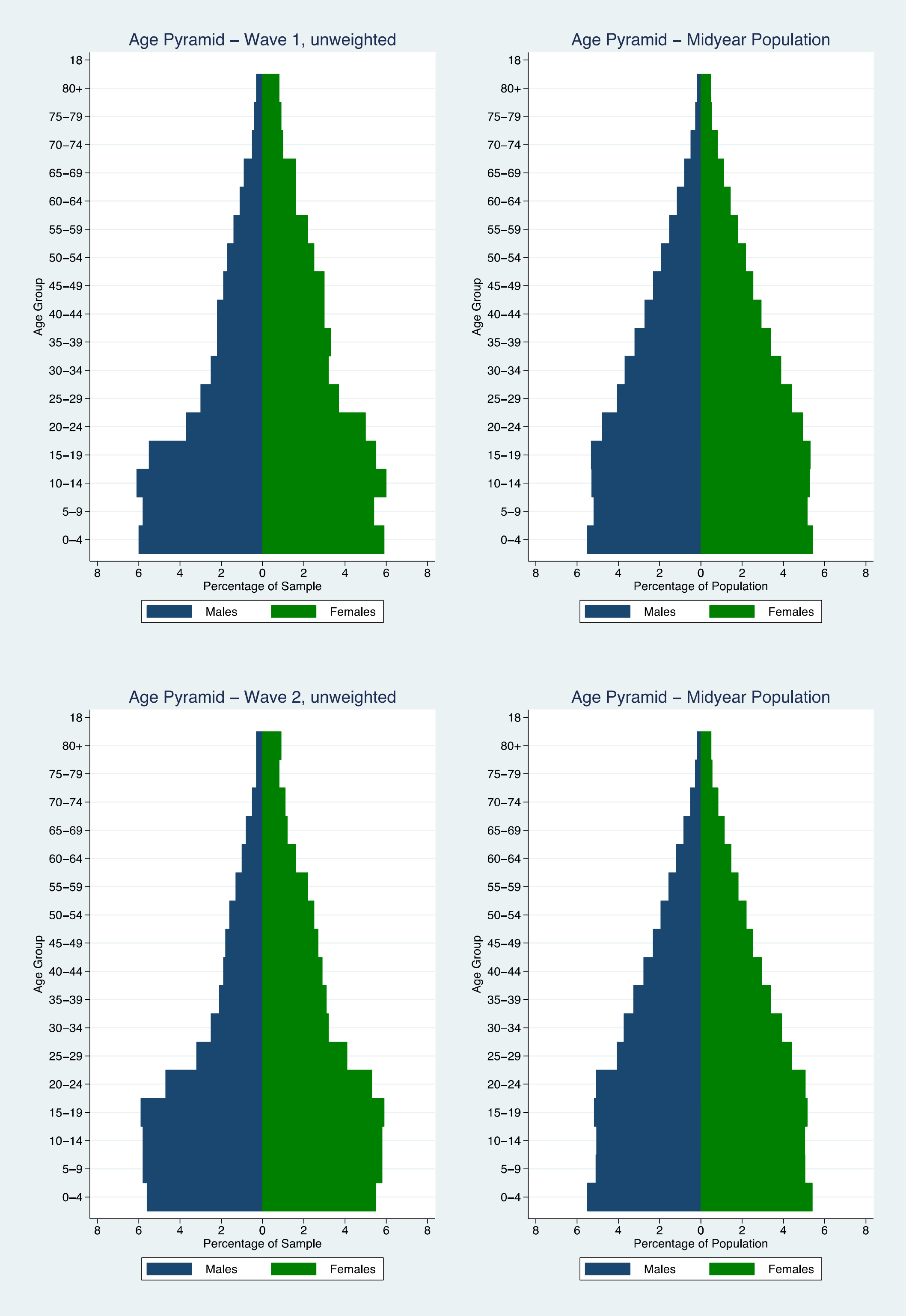

INDIAN MALES
Wave 1
Wave 2
Wave 3
Wave 4
Wave 5 Original only
Wave 5 Including Top-Up
30.0
25.0
20.0
15.0
10.0
5.0
0.0
-5.0
-10.0
0-4
5-9
10-14
15-19
20-24
25-29
30-34
35-39
40-44
45-49
50-54
55-59
60-64
65-69
70+

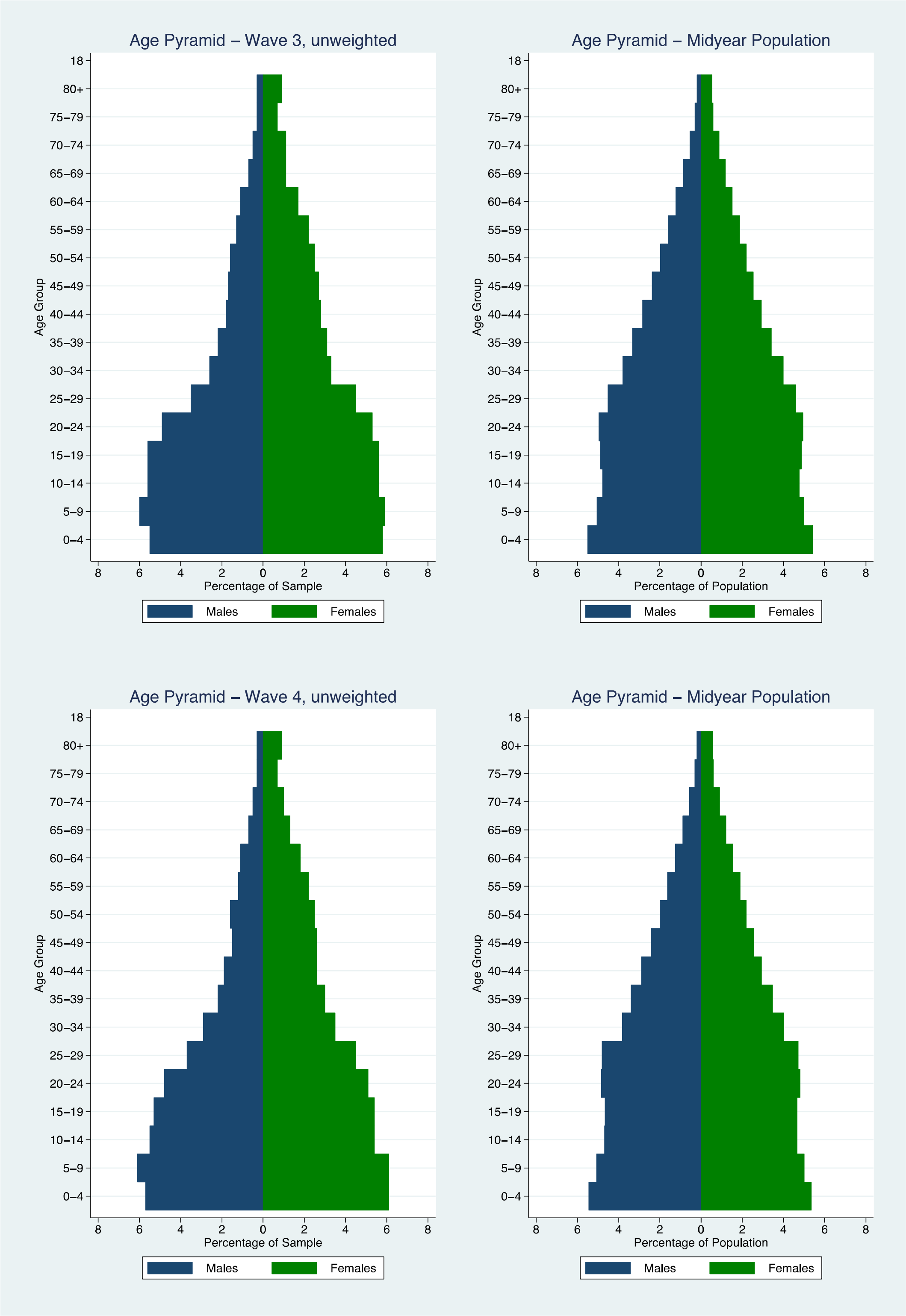

INDIAN FEMALES
Wave 1
Wave 2
Wave 3
Wave 4
Wave 5 Original only
Wave 5 Including Top-Up
15.0
10.0
5.0
0.0
-5.0
-10.0
-15.0
-20.0
0-4
5-9
10-14
15-19
20-24
25-29
30-34
35-39
40-44
45-49
50-54
55-59
60-64
65-69
70+

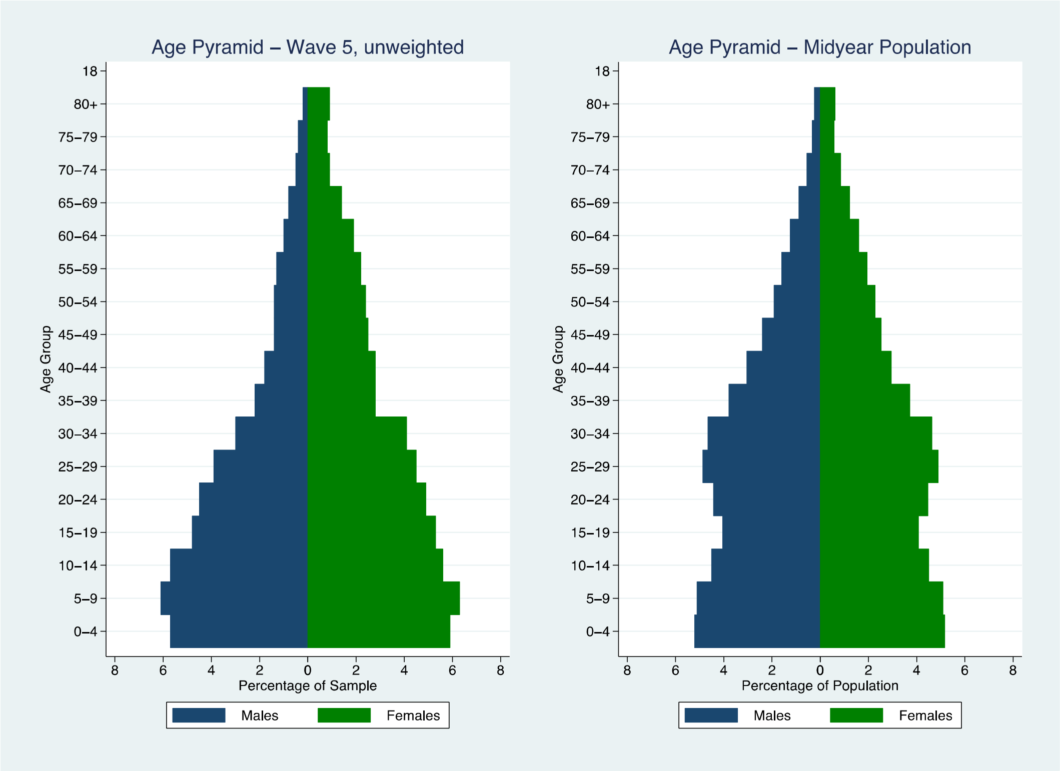

WHITE MALES
Wave 1
Wave 2
Wave 3
Wave 4
Wave 5 Original only
Wave 5 Including Top-Up
8.0
6.0
4.0
2.0
0.0
-2.0
-4.0
0-4
5-9
10-14
15-19
20-24
25-29
30-34
35-39
40-44
45-49
50-54
55-59
60-64
65-69
70-74
75-79
80+


WHITE FEMALES
Wave 1
Wave 2
Wave 3
Wave 4
Wave 5 Original only
Wave 5 Including Top-Up
8.0
6.0
4.0
2.0
0.0
-2.0
-4.0
0-4
5-9
10-14
15-19
20-24
25-29
30-34
35-39
40-44
45-49
50-54
55-59
60-64
65-69
70-74
75-79
80+

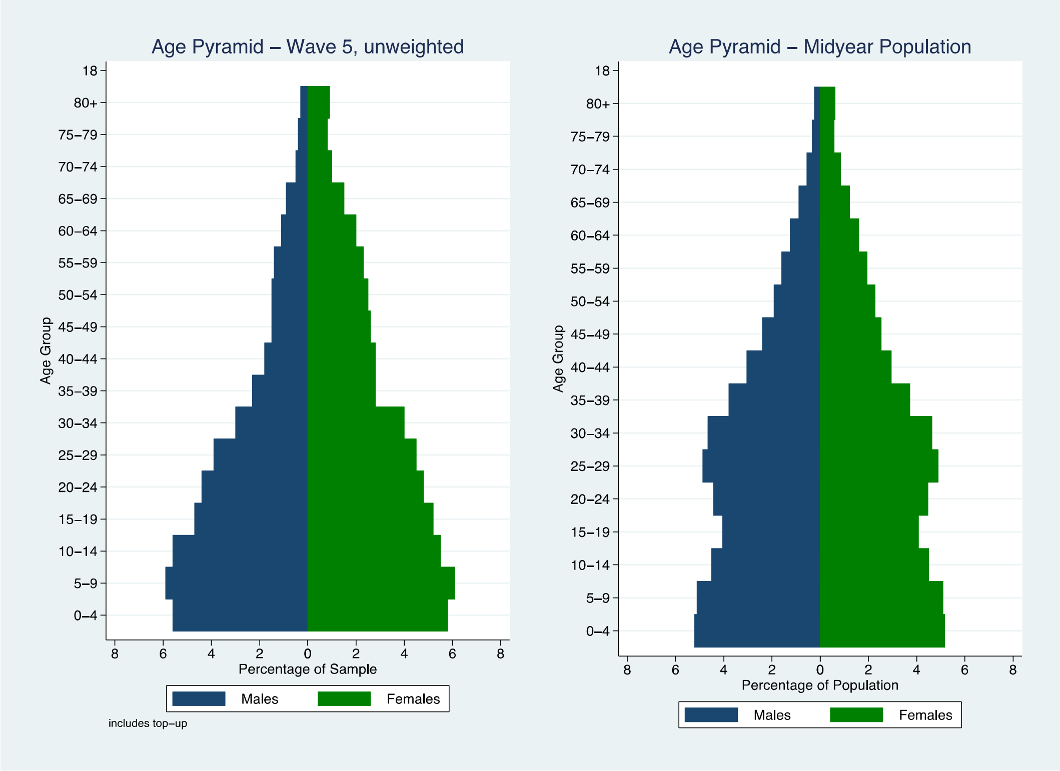
Appendix C: Population Pyramids




Age Pyramid – Wave 3, unweighted
Age Pyramid – Midyear Population
Age Pyramid – Wave 4, unweighted
Age Pyramid – Midyear Population
Age Group
18
80+
75–79
70–74
65–69
60–64
55–59
50–54
45–49
40–44
35–39
30–34
25–29
20–24
15–19
10–14
5–9
0–4
8 6 4 2 0 2 4 6 8
Percentage of Sample
Percentage of Population
Males
Females



Note: Wave 5 sample does not include the Wave 5 top-up respondents



Note: Wave 5 sample includes the Wave 5 top-up respondents

References

Branson (2019). The NIDS Wave 5 Top-up sample. Technical Paper no. 8. Cape Town: SALDRU, UCT.

Deville, J-C., & Lavallée, P. 2006. Indirect Sampling: the Foundations of the Generalised Weight Share Method. *Survey Methodology.* 32(2):165–176.

Leibbrandt, M., Woolard, I., de Villiers, L. 2009. Methodology: Report on NIDS Wave 1. Technical Paper no. 1. Cape Town: SALDRU, UCT.

Rendtel, U. & Harms, T. 2009. Weighting and Calibration for Household Panels In *Methodology of Longitudinal Surveys*. Lynn, P., Ed. Wiley. Chapter 15.

Wittenberg, M. 2009. Weights: Report on NIDS Wave 1. Technical Paper no. 2. Cape Town: SALDRU, UCT.

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

Wittenberg, M. 2013. A comment on the use of “cluster” corrections in the context of panel data. Technical Paper no. 6. Cape Town: SALDRU, UCT.