The South African National Income Dynamics Study: Design and methodological issues

Article *in* Journal for Studies in Economics and Econometrics · January 2010

CITATIONS
16

READS
947

3 authors, including:

Ingrid Woolard
Stellenbosch University
84 PUBLICATIONS **2,687** CITATIONS
SEE PROFILE

Murray V Leibbrandt
University of Cape Town
118 PUBLICATIONS **2,619** CITATIONS
SEE PROFILE

THE SOUTH AFRICAN NATIONAL INCOME DYNAMICS STUDY: DESIGN AND METHODOLOGICAL ISSUES

I Woolard[*], M Leibbrandt and L de Villiers

Abstract

The National Income Dynamics Study (NIDS) is a nationally representative panel survey of 28 255 individuals that were resident in 7 305 households in South Africa at the time of the base wave in 2008. Attempts will be made to interview each of these individuals and all of their current household members at two-year intervals in the future. NIDS is the first national panel study of individuals of all ages in South Africa. As the panel unfolds, it will reveal the dynamic structure of households in South Africa and changes in the living conditions and well-being of household members. This article presents the core methodological decisions in the design of the first wave of the NIDS panel survey. It describes the data production process, the sampling methodology, the response rates, the derivation of weights, data processing issues and how researchers can download the data. The article concludes with a discussion of some key panel issues for NIDS going forward.

1. Introduction

Since 1993 there has been an explosion of household survey data in South Africa, almost all from cross-sectional surveys which provide a snapshot of conditions or attitudes at one point in time. Such surveys have generated a wealth of findings on life in South Africa. But they are much better in answering "what?"-type questions than "why?" or "how?"-type questions. For example, a cross-sectional survey can tell us what proportion of women are employed at any given point in time, but it cannot tell us whether the same women are moving into and out of employment and what life events preface, accompany or follow on from these changes in labour market status. Similarly, repeated Income and Expenditure Surveys can tell us

[*]Respectively Associate Professor and SALDRU Research Associate, Professor and Director of SALDRU and Survey Manager, National Income Dynamics Study SALDRU, School of Economics, University of Cape Town, Private Bag, Rondebosch 7701, Republic of South Africa. The authors acknowledge financial support for this paper from the Programme to Support Pro-poor Policy Development in the South African Presidency. Professor Leibbrandt acknowledges the Research Chairs Initiative of the Department of Science and Technology and National Research Foundation for funding the Research Chair in Poverty and Inequality.
Email: Ingrid.Woolard@uct.ac.za

whether poverty rates are decreasing, increasing or holding level but cannot tell us about the fate of individual households over time. Suppose that two Income and Expenditure Surveys reveal that the poverty rate is the same in each period. This could be the result of the same households having been in poverty in both time periods. Alternatively, it may be that some households exited poverty over the period, while an equal number entered. Such distinctions, missed by cross-sectional surveys, might be very important in determining an effective policy response which may differ for chronic versus transitory poverty (Chaudhuri and Ravallion, 1994).

Most developed countries and several developing countries now engage in household panel surveys. Woolard and Leibbrandt (2006) provide a detailed review of these studies. Such surveys study the same group of households or individuals over time in order to better understand social change, income mobility and poverty dynamics. Baulch and Hoddinott (2000) provide a review of this literature in a developing country context. Panel studies are particularly important for monitoring and evaluation purposes as these surveys observe households and individuals both before and after a sudden change in their circumstances or their participation in a government programme. This allows for a richer and more precise assessment of the impact of the programme or the shock. Finally, panel data provides some scope for dealing with unobserved heterogeneity (Wooldridge, 2002). Social behaviour is complex and any cross-sectional quantitative model omits many sources of variation across individuals, households and communities. These unobserved factors may bias the coefficient estimates in which the analyst is interested. To the extent that this unobserved heterogeneity remains constant over time and one is using panel data to investigate changes over time, the panel controls for this unobserved heterogeneity by differencing it out of the model.

As reviewed by Woolard and Leibbrandt (2006) there are a number of regional panel surveys in South Africa. However, NIDS will be the first national panel study to document the dynamic structure of a sample of household members in South Africa and changes in their incomes, expenditures, assets, access to services, education, health and other dimensions of well-being. A key feature of the panel study is its ability to follow people as they move out of their original households. In doing this, the movement of household members as they leave and/or return to the household or set up their own households will be adequately captured in subsequent waves.

The first (baseline) wave of NIDS was conducted by the Southern Africa Labour and Development Research Unit (SALDRU) based at the University of Cape Town's School of Economics. The fieldwork for this first wave commenced in February 2008, with the public release of the data in July 2009. This first wave of the survey provides the baseline information on the well-being of 28 255 sample members in 7 305 households against which to measure all future changes. The design of NIDS envisages data collection every two years and the second wave is currently in the field, with data release planned for late 2011.

As of now though, only one wave of NIDS data is available and the articles in this special edition make use of this wave as a nationally representative picture of South

Africa in 2008. Even as a cross-section, in order to use these data appropriately and effectively users need the details of how the data were produced. This article sets out to present these core methodological decisions in the design of Wave 1 of the NIDS panel. It follows the data production process, starting with questionnaire design, then moving to sampling, fieldwork, response rates, the derivation of weights, data processing issues and how researchers can download the data. We conclude by returning to some key panel issues as NIDS goes forward to Wave 2 and beyond.

2. The scope of the survey and questionnaire design

NIDS collects a broad range of information on a large number of topics. Despite the name of the survey, the emphasis is not on income but on a wide range of measures of well-being. Some of the topics on which information is gathered include:

- Expenditures of the household;
- The wages, social grants and other incomes of individuals in the household;
- The assets owned by the household and the services to which the household has access;
- The level of education and health status of household members;
- Whether household members are still in school or working or looking for work or helping at home or retired; and
- The community groups to which members of the household belong, whether household members would like to remain within their current communities and how well-off they are relative to others in their community.

These topics must be viewed against the backdrop of the key concerns highlighted by the Presidency for NIDS to investigate. These included the need to measure whether all South Africans are benefiting from economic growth and social stability, and the concern that numbers of South Africans might end up being 'socially excluded', left behind or trapped in a 'second economy' where they are unable to benefit from economic opportunity. NIDS is expected to shed light on the circumstances in which South Africans find themselves, how these conditions impact on their ability to improve their well-being and how government policy can play a positive role in these livelihood strategies.

The 2008 NIDS questionnaires were designed through many iterations with the overriding goal of ensuring that they gathered baseline information on all of these factors. To do this, information is gathered on all members of the household; including those that were resident and those that were non-resident at the time of the interview. Those that were resident provide the base sample of individuals who will remain in the NIDS sample over time. Information about non-resident members is essential in understanding the household and family support systems that individuals had around them at the time of the interview. A household questionnaire as well as individual questionnaires for each adult and each child in

the household were administered at each dwelling unit. Respondents aged 12-59 were also asked to take a short numeracy test[1].

3. Sample design

The target population for NIDS is private households and residents in workers' hostels, convents and monasteries. The frame excludes other collective living quarters such as students' hostels, old age homes, hospitals, prisons and military barracks.

A stratified, two-stage cluster sample design was employed in sampling the dwelling units to be included in the base wave. In the first stage, a sample of 400 Primary Sampling Units (PSUs)[2] was drawn (by statisticians at Stats SA) from Stats SA's 2003 Master Sample of 3000 PSUs. At the time that the 2003 Master Sample was compiled, eight non-overlapping samples of ten or twelve dwelling units were systematically drawn within each PSU. Each of these samples is termed a "cluster" by Stats SA. These clusters were then allocated to the various household surveys that were conducted by Stats SA between 2004 and 2007 (such as the Labour Force Surveys, General Household Surveys and the 2005/06 Income and Expenditure Survey). However, two clusters in each PSU were never used by Stats SA and these were allocated to NIDS.

In the first stage, a sample of 400 PSUs had to be drawn from the 3000 PSUs in the Master Sample. The explicit strata in the Master Sample are the 53 district councils (DCs). The sample was proportionally allocated to these 53 strata and PSUs were selected within strata with probability proportional to size. It should be noted that the sample was not designed to be representative at provincial level, implying that analysis of the results at the province level is not recommended.

Tables 1 and 2 compare the distribution of the PSUs that are in the NIDS sample against those that are in the Master Sample. It can be seen from both tables that the sample spread per province and per geography type is quite similar between the two samples. Thus the selected sample was deemed satisfactory in this regard.

[1]Copies of these questionnaires are available on the NIDS website at http://www.nids.uct.ac.za/home/index.php?/Nids-Questionnaires/nids-questionnaires.html

[2]A PSU is defined as a geographical area that consists of at least one Enumeration Area (EA) or several EAs from the 2001 Census, when the originally selected EA was found to have fewer than 74 households. In some cases it has been necessary to add EAs to the original EA to meet the requirement of a minimum of 74 households per PSU. The EA or EAs added to the original EA has to be of the same settlement type as the original EA. An EA is the smallest portion of land that the country was demarcated into for the purpose of Census enumeration.

Table 1: The distribution of the PSUs in the NIDS sample and in the full master sample, by province

	NIDS Sample		Master Sample	
Province	**Frequency**	**Percent**	**Frequency**	**Percent**
Western Cape	52	13,0	385	12,8
Eastern cape	53	13,3	396	13,2
Northern Cape	27	6,8	207	6,9
Free State	31	7,8	245	8,2
Kwa-Zulu Natal	86	21,5	640	21,3
North West	35	8,8	259	8,6
Gauteng	48	12,0	353	11,8
Mpumalanga	30	7,5	233	7,8
Limpopo	38	9,5	282	9,4
South Africa	**400**	**100,0**	**3 000**	**100,0**

Table 2: The distribution of the PSUs in the NIDS sample and in the full master sample, by type of geographical area

	NIDS Sample		Master Sample	
Geography Type	**Frequency**	**Percent**	**Frequency**	**Percent**
Rural	49	12,3	310	10,3
Traditional Authority Areas	131	32,8	957	31,9
Urban Formal	194	48,5	1535	51,2
Urban Informal	26	6,5	198	6,6
South Africa	**400**	**100,0**	**3000**	**100,0**

Within each PSU, Stats SA provided two clusters with a total of 24 dwelling units. Stats SA provided maps for all PSUs and detailed listings with these 24 dwelling units marked. These listings had been updated several times since originally compiled in 2003 in order to increase the ease with which fieldworkers could find the specific dwelling units. (The sample of dwelling units itself had not, of course, been changed.). In spite of this, it was sometimes necessary to re-list a PSU if dramatic changes had occurred since the listing had last been updated. For example, if an informal settlement had been cleared to make way for formal houses, the listing was unusable. In these cases, the PSU was re-listed and a new systematic sample of dwelling units was selected. The drawback of re-listing a PSU is that the chance of sample overlap with dwelling units that had already been selected for other surveys is substantially increased. The extent of this overlap cannot be quantified as the lists are no longer comparable.

In summary, the first stage of the sampling resulted in a sample of 400 PSUs. Within each of these PSUs there were two unused clusters (drawn by means of systematic sampling at the time that the Master Sample was created in 2003). This gave us a sample of 24 dwelling units in each of 400 PSUs, making a total of 9600

dwelling units. Based on the results of the pre-test, it was expected that this sample would yield 8000 participating households.

However, during the initial round of fieldwork for Wave 1 we did not achieve the target number of households. Therefore we went back to the field to attempt to overturn refusals in 48 PSUs and to visit 24 new dwelling units in each of 32 of these areas. Stats SA drew a random sample of an additional 24 dwelling units in each of these 32 PSUs. These were PSUs in which the predominant population groups were White and Asian households. This was done in order to improve representation of these race groups in the data. This exercise became known as Phase 2 and is discussed further below.

3.1 The sample of households

Fieldworkers were instructed to interview all households living at these 9600 selected dwelling units. If they found that the dwelling unit was vacant or the dwelling no longer existed they were not permitted to substitute the dwelling unit. Where more than one household resided at the selected dwelling unit, a separate household questionnaire was completed for every household and they are treated in the data as separate units. In order to qualify as separate households, the household members should not share resources or food. For example, lodgers and live-in domestic workers were considered to be separate households.

3.2 The sample of individuals

All resident household members at selected dwelling units were included in the NIDS panel, providing that at least one person in the household agreed to participate in the study. The household roster in the household questionnaire was used to identify potential participants in the study. Firstly, respondents were asked to list all individuals that had lived under the same roof (or within the same compound/homestead) for at least 15 days during the last 12 months or who had arrived in the last 15 days and for whom this was now their usual residence. In addition the persons listed should share food from a common 'pot' and share resources from a common resource pool. All those listed on the household roster are considered household members. However, only persons who "usually stay here four nights a week" were classified as resident household members. Household members who did not fulfil this criterion are termed "non-resident members".

All *resident* household members became NIDS sample members. Most non-resident members had a non-zero probability of being sampled at the place where they usually reside and were thus excluded from the sample. The exception to this was non-resident members that were "out of scope" at the time of the survey. Out-of-scope household members were those living in institutions (such as boarding school hostels, halls of residence, prisons or hospitals) which were not part of the sampling frame. These individuals thus had a zero probability of selection at their usual place of residence and were therefore included in the NIDS sample as part of the household that had listed them as non-resident members. Thus out-of-scope

non-resident members also became NIDS sample members and proxy questionnaires were completed for this group.

Every resident and non-resident "out of scope" individual in a participating household became a Continuing Sample Member (CSM). Each CSM should have had an individual questionnaire (adult, child or proxy) completed for them. Importantly, these individuals are CSMs even if they refused to be interviewed in the base wave.

4. Wave 1 data collection

4.1 Pre-test and publicity

The fieldwork for Wave 1 was put out for tender; with the tender being awarded to Development Research Africa (DRA) in consortium with Take Note Trading. A pre-test was conducted in eight areas (that were not part of the NIDS sample) in September 2007. The pre-test was enormously useful in terms of highlighting problematic parts of the questionnaire and honing fieldwork, quality control and data capture procedures.

One of the issues that was highlighted in the pre-test was the need for a publicity campaign ahead of the main study. In discussions with Stats SA during the latter part of 2007, it was proposed that Stats SA should play an active role in this publicity campaign. Since Stats SA's Labour Force Survey, General Household Survey and Income and Expenditure Survey had all been conducted in these same PSUs it was agreed that the nine Stats SA provincial co-ordinators would identify freelance fieldworkers that had experience in conducting publicity campaigns for use in the NIDS PSUs. These publicity officers would be employed directly by the fieldwork organisation for the duration of the publicity campaign. Stats SA provided a trainer who trained the 40 freelance publicity officers in Johannesburg in early January 2008.

Publicity was then conducted over a ten-day period with each publicity officer doing an average of one PSU per day. These publicity officers were responsible for updating the listings, making contact with police, community leaders and "gatekeepers" in the area and dropping off NIDS brochures at every selected dwelling unit.

The quality of the publicity campaign seems to have been variable, with some publicity officers doing an excellent job while others seem to have done very little. Monitoring the quality of the work was very difficult since the publicity officers worked alone.

4.2 Fieldwork

Wave 1 fieldwork commenced with training in Durban in the last week of January 2008. Fieldwork began in one region at a time in order for any teething problems to be contained and remedied within a specific region before moving on. Training of

about 150 fieldworkers took place at DRA's four regional offices in Durban, Johannesburg, Port Elizabeth and Cape Town. The model allowed for a large number of fieldworkers to be in the field simultaneously. With the benefit of hindsight, fewer, top-quality fieldworkers staying in field for a longer period of time would probably have been a better strategy. Managing fewer teams at any one time reduces the logistical burden on the fieldwork organisation and allows for tighter supervision of teams.

The first week in each regional office consisted of a full week of training conducted jointly by NIDS and DRA staff. Each fieldworker was given a package containing a training manual, copies of all letters and publicity information and two complete sets of questionnaires to be used for training and annotated for the field. Daily tests were held and marks were recorded in order to allow the fieldwork organisation to assess the quality of all fieldworkers. A qualified nurse assisted with the anthropometric training and a special training session was organised at which practice measurements were taken of babies and young children.

Once fieldwork in each PSU was complete, the bundles of questionnaires for each household in the PSU were sent to DRA's regional office for internal quality control. Once the quality was deemed satisfactory, the completed household bundles were couriered to the NIDS offices at SALDRU for another round of quality control, coding and data capture.

Fieldwork teams consisted of a team leader and between three and four fieldworkers. Team leaders were responsible for identifying the correct dwelling unit to be approached based on the listing, introducing fieldworkers to households and taking GPS co-ordinates for all selected dwelling units. Team leaders also co-ordinated the use of anthropometric equipment within a PSU. Each field team spent 5 days in a PSU. In that time they were meant to identify each of the selected dwelling units, interview all respondents and overturn refusals. In the case of non-response, the fieldworker had to visit the dwelling unit at three different times on three different days before it was accepted as a valid non-response. Attempting to overturn soft refusals was also the responsibility of the team leader.

Wherever possible, households and fieldworkers were matched based on language and race. Due to the insufficient number of fieldworkers trained it was not always possible to match fieldworkers to households in this way. There is some evidence that potential respondents did not respond well to fieldworkers who could not speak their language or who were from a different population group and therefore the response rates were negatively affected.

Fieldwork began in early February 2008. It was initially scheduled to be complete in May, although NIDS had always built in June as a potential spill-over month. However, there were even more delays than envisaged and fieldwork for the first phase was only completed in July 2008. The completion of an initial round of fieldwork in all PSUs resulted in about 6500 successful households from the 9600 dwelling units that were sampled.

Given the fact that the target number of households (8 000) was not realised, we started investigating the feasibility of expanding the fieldwork period and going back into the field. There were a number of options available at the time. Our Steering Committee asked us to prepare a detailed report on non-response and refusals as the basis for a decision. This was done and, based on extensive input from our Steering Committee, we decided on the following strategy: Firstly, we asked Stats SA to draw replacement PSUs for 9 areas in which no interviews were conducted during the first phase of fieldwork. Secondly, it was decided that we would re-visit (with the intention of overturning refusals) all predominantly white PSUs in Gauteng, Mpumalanga, Limpopo and the Western Cape, all Indian/Asian PSUs in Gauteng and KwaZulu-Natal and all Coloured PSUs in Gauteng.

The additional responses generated by this strategy were precious to the panel and made the re-visitation campaign worth pursuing. However, it was felt that this alone would not yield sufficient numbers of White and Asian/Indian panel members. Stats SA were asked to draw a new sample of 24 dwelling units in each of the predominantly white and Asian/Indian PSUs. These new households were visited at the same time as the campaign to overturn refusals.

From a sampling point of view this amounted to an *ex ante* decision to oversample by predominant racial group in specific provinces. This *ex ante* simplicity was deemed to be good as it implied that the derivation of sampling weights would be easier. Phase 2 was implemented in the same way as Phase 1 and the same protocols were followed in field. Flags have been inserted in the data to differentiate Phase 1 and Phase 2 households and participants from each other.

The result of the phase 2 fieldwork was that 1 856 households were (re)visited during field work. In the majority of cases these were “difficult” PSUs that had already proven themselves to be tricky to access and the majority of the households that were being visited had previously refused to participate. Low response rates were, therefore, expected. The re-visit exercise resulted in an additional 807 successful households. After the additional fieldwork phase NIDS achieved a total of 7 305 participating households with 28 255 individuals.

5. Response rates

As mentioned before, response rates in phase 1 were disappointing and phase 2 was embarked upon to realise a more acceptable base wave sample. A detailed analysis of household level response rates follows. Response rates were calculated using the number of visited dwelling units as the denominator and the number of participating households as the numerator. In the instances where response rates are given by race the predominant race group of the PSU is assigned to all households in that PSU. This is done because we do not have any demographic information about non participating households.

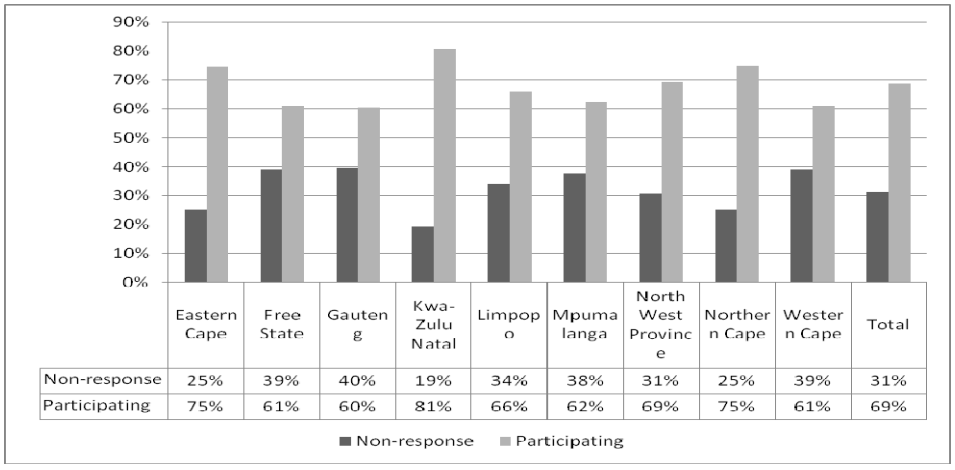
Every effort was made to correctly identify all resident household members at the time of the interview. For different reasons not all resident household members were interviewed. Proxy questionnaires were completed for 1 754 adults who were

unavailable. For a further 1 250 adults no questionnaires were completed. For these individuals we only have the information supplied in the household roster which is part of the household questionnaire. By virtue of being resident members in households in which at least one other person agreed to participate, these individuals are panel members and we will attempt to make contact with them in the next wave.

Over the combined fieldwork periods NIDS fieldworkers knocked on 10 642 household doors. Of these, 7 305 households agreed to participate. This equates to a 69% response rate. The total sample for NIDS consists of 409 PSUs. Of those, 9 were replaced in phase 2 because the whole PSU was inaccessible in phase 1. They are therefore excluded from the rest of the calculations.

Figure 1 below presents the breakdown of household response rates by province. It can be seen that KwaZulu-Natal had the best response rate at 81%, while Gauteng, Free State and Western Cape had the worst response rates at 60%. Thus, there was fairly wide variation in response rates at the provincial level. Provinces vary by predominant geo-type and race and, in understanding what lies behind this provincial variance, it is useful to look at the breakdown of response rates according to these demographic markers. Figures 2 and 3 present the pictures for geo-type and race respectively.

It can be seen that the response rate in urban formal areas was 60%. In all geo-types other than this, response rates of 70% or more were achieved with particularly high response rates in rural informal and traditional authority areas. Despite the phase 2 strategy, Figure 3 makes it clear that white response rates were low at 36%. In sharp contrast, the response rates for all other racial groups were much higher with the figures for Indian, Coloured and African being 66%, 73% and 76% respectively.



	Eastern Cape	Free State	Gauteng	Kwa-Zulu Natal	Limpopo	Mpumalanga	North West Province	Northern Cape	Western Cape	Total
Non-response	25%	39%	40%	19%	34%	38%	31%	25%	39%	31%
Participating	75%	61%	60%	81%	66%	62%	69%	75%	61%	69%

Figure 1: Response rates by province

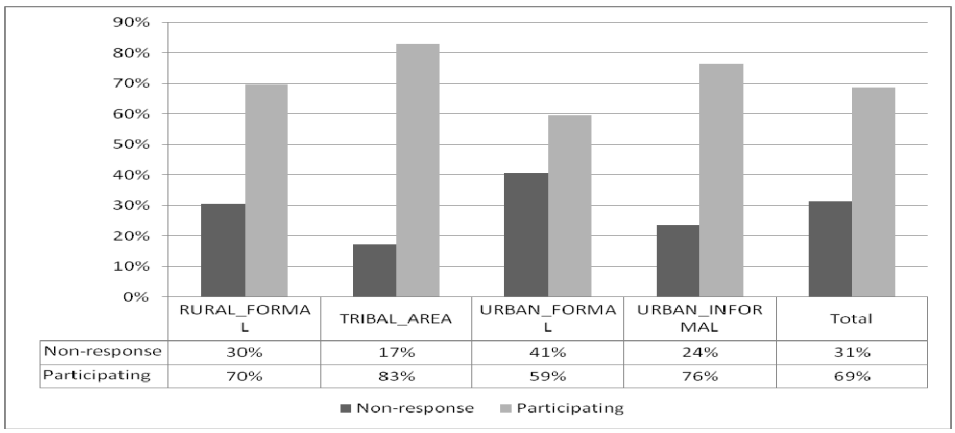


Figure 2: Response rates by geo-type

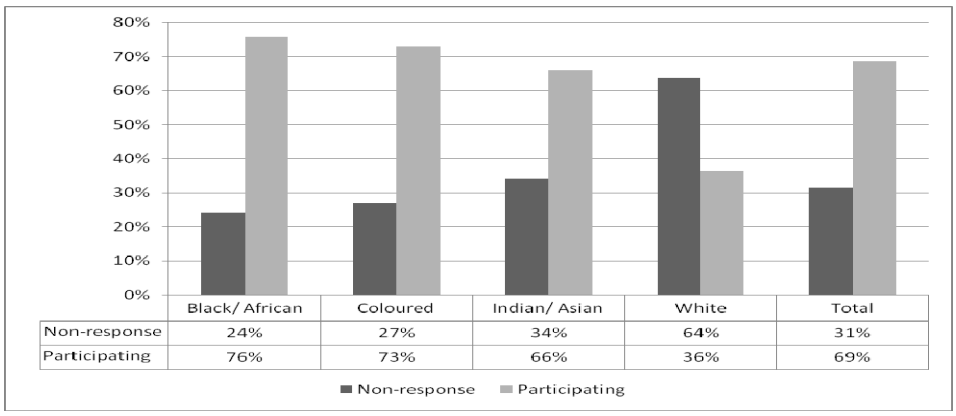


Figure 3: Response rates by race

Table 3a below reports the number of households per race group that were achieved in phase 1 and phase 2 of the fieldwork. This is an important table as it makes it clear that the phase 2 strategy more than doubled the number of white households in the sample. Thus, although the response rate of white households was disappointing across both phases of fieldwork, the phase 2 fieldwork was very successful in bolstering the number of white households and individuals that constitute the base sample of NIDS going forward. The same is true of Indian households and individuals.

As mentioned earlier, we have no information on the race of the members of non-responding households and we had no choice but to look at household non-response using our sampling data base to tell us the predominant race group per PSU. Table 3a is calculated in this way and in this sense it is consistent with the preceding analysis in this section. However, given that it reports respondents rather than non-respondents, we can compare these figures to those in the NIDS data. As such, it offers an interesting point of comparison between the PSU-based figures and the actual data that we gathered in the two phases of fieldwork. Table 3b presents the

same racial breakdown as Table 3a based on the NIDS data. There are fewer households in Table 3b than 3a because there is no race data for 20 households.

Table 3b shows that, relative to predictions using the predominant race of the PSU, we realised notably fewer whites and Indians, notably more Africans and the expected number of Coloureds in the NIDS sample. This is unsurprising. We are fifteen years into our new democracy and, even though racial desegregation of residential areas has been slow, there has been some movement of Coloured and African households into wealthier residential areas but very little movement the other way.

6. Weights

Before analysis and report-writing on the NIDS data could begin it was necessary to calculate sampling weights. Martin Wittenberg at the University of Cape Town took on the responsibility of calculating the weights for the base wave of NIDS. For a detailed explanation of the weighting procedure see Wittenberg (2009).

Table 3: Number of participating households by race group from Phase 1 and Phase 2

a. By predominant race group in the sampled PSUs

	Phase 1	**Phase 2**	**Total**
Black/ African	5,225	272	5,497
Coloured	951	76	1,027
Indian/ Asian	32	98	130
White	290	360	650
Total	6,498	806	7,304

b. *By actual race of household respondent in questionnaire*

	Phase 1	**Phase 2**	**Total**
Black/African	5,202	396	5,598
Coloured	938	77	1,015
Indian/Asian	48	62	110
White	283	270	553
Other	8	0	8
Total	6,479	805	7,284

The basis of the calculation of the design weights is the information that Stats SA provided to NIDS about the process of two-stage sampling from the 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 a calculation about the probability of including each specific household in each PSU in the NIDS sample.

The second set of weights are the post-stratification 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 Stats SA for the Mid Year Population Estimates. In addition, we imposed the constraint that the population distribution by province should correspond to that released in those 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, i.e. each household member needed to have the same weight. This is based on the assumption that NIDS disproportionately missed certain types of households, rather than that we disproportionately under-enumerated particular groups within participating households. The post-stratified weights were created using the "cross-entropy" estimation procedure (Golan, Judge and Miller, 1996). The program used to calculate the weights is available (Wittenberg, 2009a).

7. Data cleaning and data preparation

7.1 Preserving anonymity in the data

It is the responsibility of the NIDS team to ensure that respondent identities are protected. During the interviews information was collected that would enable tracking and re-contact of respondents for subsequent waves of data collection. However, some of this information is excluded from the public release data set to preserve the anonymity of the respondents. The types of data collected but excluded from public release fall into three broad categories.

Firstly, the names and addresses of the respondents will never be released. This information is kept separately as part of the panel maintenance system. Secondly, we will not release the detailed geographical information about the respondent's current or past location. The names of suburbs and towns are not released. The lowest level of geography which is released is the district council level. Thirdly, detailed information regarding schools, clinics and occupations were collected. These "text"-fields will not be released as they could potentially be used to identify respondents.

NIDS has three versions of data: public, internal and secure. The public data does not have any personalised information or the names of schools attended or clinics/hospitals utilised by the respondent. The internal data has all the components of the public data as well as coding for schools and hospitals and more detailed information on the geographic location of households (such as the sub-place and main-place names within which households fall) and the occupations and sectors of employment of panel members. This dataset, or parts thereof, can be accessed only from within NIDS and is only made available for analytical purposes to researchers on special request and subject to them signing an agreement. Researchers can only work on this data inside the NIDS office and may only take processed data away with them. The secure data includes all the information on respondents and is used only for operational (fieldwork and tracking) purposes. The secure data is thus never made available to researchers.

7.2 Integration of community level data

The questionnaires were designed in such a way as to facilitate the integration of community and administrative data. Respecting the anonymity of respondents, GPS information and clinic or school data could be used to calculate distances to nearest school or clinic. There is great research and policy potential for such links with administrative data. A pilot project which links the school data to the National Educational Infrastructure Management System (NEIMS) data from the Department of Basic Education has been embarked on, with funding provided by the Programme for the Support of Pro-Poor Policy Development (PSPPD).

7.3 Derived variables and imputed values

Some of the preparation of the NIDS data for public release required the calculation of derived variables. In deriving these variables one is moving beyond cleaning and preparation of primary data. Two important examples are the calculation of total household income or total household expenditure. Both require the aggregation of all income sources or all expenditure categories for each household and, in addition, require some assumptions about the treatment of missing incomes or expenditures.

The key principle for NIDS is that it is the anonymised primary data that forms the basis of the public release NIDS data. Therefore, any analyst will be able to start from the primary data and aggregate data for themselves and make their own assumptions about how to treat missing data. However, some derived variables are provided in the public release data. Work that was done within NIDS to impute missing data for individual variables and/or to derive new variables has been placed in derived data files that are distinguishable from the primary NIDS data files. In addition the programs used to derive these variables and clear documentation about the decisions are available to users.

Most of the contestable assumptions in working on NIDS data arise from three kinds of non-response. Firstly there is household non-response. We have discussed such non-response in detail above. Our discussion of sampling weights made it clear that the key decisions about dealing with such non-response are imbedded in the derivation of these weights. Analysts are likely to assume that the recommended weights take care of such non-response.

Secondly there are non-respondents within responding households. Table 4 below shows the distribution of this unit non-response across responding households. Just over 88% of the 7 305 households in the achieved sample had zero non-response. This is an encouraging sign in terms of the extent of bias from unit non-response as only about 12% of households are affected at all. In addition, only 6% of households had an individual response rate lower than 50% within the household. The 14 households that have 100% non-response to the adult questionnaire are still counted as responding households because household rosters were completed for these households.

Roughly 6,7% of the sample of adults from the achieved sample of households did not respond and, for these individuals, we have information only from the household roster. Relatedly, there are adults who were unavailable for an interview and for whom proxy questionnaires were completed. Proxy questionnaires make up 9,4% of the adults from the achieved sample of households. For such people we have more information than that contained in the household roster but not complete information. Also, it is an open question as to how the data quality differs given that the questions are not answered by the adult themselves.

Table 4: Intra-household adult non-response rate

	Frequency	Percent	Cumulative Percent
0%	6,438	88,2%	88,2%
0% - 25%	105	1,4%	89,6%
25% - 49%	321	4,4%	94,0%
50% - 74%	391	5,4%	99,3%
75% - 100%	34	0,5%	99,8%
100%	14	0,2%	100%
Total	7,303	100%	

Each analyst needs to make assumptions about how to deal with non-responding individuals varying from assuming that those that do not respond have no income to assuming that such non-response can be imputed based on the characteristics of the individual (e.g. race, age, sex, geotype, etc) that are known from the household roster.

Finally, among individuals or households who do respond to the survey there is item non-response. For example, where an individual professes to earn income from a particular source but does not give the amount, we define this as item non-response. There is very little that can be said about such item non-response in general. The other papers in this special edition touch on issues of item response as they specifically relate to their research topic. Many derived variables, such as aggregate incomes and expenditures, include imputed values for such non-response. In these cases it is essential that the imputation decisions are made explicit.

8. Conclusion: Looking forward to Wave 2 and beyond

Sections 2 through 7 above provided details on the production of the Wave 1 data. This is information that users need in order to use these data responsibly. These data are publicly available and instructions to users on how to download them are given in the introduction to this special edition. This Wave 1 dataset provides the baseline information for the National Income Dynamics Study going forward and we end this paper by returning to the discussion of the National Income Dynamics Study as a panel survey.

In the design of panel surveys, rules are needed to determine how the samples for the second and subsequent waves of surveys are to be generated and to specify

which units of the sample remain and which drop out from one wave to the next. For the original panel members, there are problems in determining whether to follow all sample members as their status changes. Some surveys attempt to follow all sample members into institutions (such as hospitals and old age homes) where it is practicable to do so. Other surveys do not follow all persons into institutions though their whereabouts are still tracked. If, however, a sample member is determined to have entered the institutionalized population on a permanent basis, then that person is usually removed from the sample entirely (Woolard and Leibbrandt, 2006). If necessary, rules are adopted/implemented for adding new sample members to the panel so that valid and desired samples are maintained for representing the population. Decisions concerning the rules to be followed need to be considered and balanced by a number of factors, including the purpose of the survey, and analytic and cost concerns.

The tracking rules must ensure that the sample replenishes itself in the same way as the population. It was decided at the outset that NIDS would employ an "indefinite life" design and would involve interviews with all adult members and proxy interviews for all children. In line with leading panel studies conducted in other countries (Wooden, 2001), the sampling unit is the household, and members of those households will be traced over an indefinite life. The Wave 1 sample is then automatically extended over time by following rules that add to the sample. In the first wave, a representative sample of SA households was drawn. All the resident and out-of-scope members of households in which at least one person agreed to participate in the base wave are termed "continuing sample members". They will be tracked in subsequent waves of NIDS. Children born to female continuing sample members (after the baseline survey) will themselves become continuing sample members. All other persons who join the sample in subsequent years (by virtue of being co-resident with a CSM) are temporary sample members (TSMs). If they cease living with a CSM they cease being sample members.

These following rules, in combination with the initial sample that is intended to be representative of all South African households, provide a mechanism for ensuring that the panel retains its cross-sectional representativeness over time. It is expected that the number of households will grow as CSMs move out of their initial households and attach themselves to other households or establish new households. At the limit, each of the 28255 CSMs from Wave 1 could, in theory, move into their own household. We will need to wait for the results of Wave 2 to see how many households needed to be visited in order to locate all of our CSMs.

8.1 Refreshing the sample

Refreshing the sample refers to adding units to the sample over time. It is done in order to represent new members of the population (such as immigrants that moved into South Africa after the original sample was selected) and/or to compensate for losses from attrition.

The longer the panel study is continued, the less representative the panel will become of the current population. Serious thought needs to go into the possibility of refreshing the NIDS sample at a later stage.

8.2 Maintaining high response rates across waves

Panel respondents may change residence between waves of data collection, and time and money are needed to locate such respondents. It is important to trace respondents who have moved or changed telephone numbers so that the panel study can maintain the required sample size and reduce attrition. Adopting a comprehensive locating procedure is essential to minimize non-response bias. NIDS has a team of full-time survey assistants who are responsible for maintaining contact with respondents by phone and mail. Customized software has been created in-house to manage this system.

To minimize the effects of attrition, it is important not to write off sample members who become non-respondents after the initial wave. Many of these "wave non-respondents" may be willing to participate in later rounds. If wave non-respondents are kept in the sample and some are "converted" in later waves, the effects of attrition may not be cumulative. Respondents that could not be located in one wave may be found later on; respondents that were too busy to take part in one round may have more time in the next. In any panel sample, there will be respondents that insist on being dropped from the panel; it may make sense to simply write off such respondents since the chances of converting them are very low. Nonetheless, a substantial portion of wave non-response is due to temporary circumstances (Wooden, 2001) and wave non-respondents will not be automatically dropped from the panel.

8.3 Modifying the questionnaires across waves

A defining feature of a panel design is the administration of the same items to a sample of respondents on several occasions over time. It is this feature of panel surveys that permits the direct measurement of change in individual units. It would, therefore, seem logical that questionnaires and data collection instruments should be kept the same across each wave of a panel study. Any changes in appearance, content, or wording of the instruments, or in the data recording or coding procedures, could compromise the comparability of the data in the different waves.

Two considerations may, however, make it necessary to change the data collection instruments used in a panel survey. In the first place, new issues may arise and the panel sample may be the best means for collecting information about them. One of the virtues of a panel study is its ability to provide timely information about emerging issues. When new issues arise, it may make sense to add a module or supplement to the existing instruments. In effect, the panel sample can be used to collect cross-sectional data on the new topic. Although this strategy may not capitalize on all the strengths of a panel design, it can save time and money as compared to selecting and interviewing a new cross-sectional sample. In addition, the data collected about the panel members in previous waves may enrich the

analysis of the data collected in the new module. However, since adding questions to the instrument will increase the burden placed on the panel respondents, the number of new items should be kept to a minimum. In some cases, it may be better to conduct a separate survey than to jeopardize the success of an ongoing panel.

A second circumstance that can argue for change in a panel questionnaire involves problems with an item. When a question yields unreliable data in each wave, the estimates of change become doubly unreliable. For this reason, it is important even in panel surveys to revise poorly worded questions or questions that appear to yield suspect data for other reasons. Although replacing faulty questions or instruments interrupts the sequence of comparable measurements, it may be necessary if the measurements are to be interpretable at all. The likelihood of finding faulty items can be substantially reduced through pilot testing of the instruments in advance of the main survey. However, sometimes the problem with an item is not that it was poorly conceived in the first place, but that it becomes less and less meaningful over time.

References

Baulch, B and Hoddinott, J (2000): "Economic Mobility and Poverty in Developing Countries", *Journal of Development Studies*, 36(6), 1-24.

Chaudhuri, S and Ravallion, M (1994): "How Well do Static Indicators Identify the Chronically Poor?", *Journal of Public Economics*, 53 (3), 367-394.

Golan, A, George, J, and Douglas M (1996): *Maximum Entropy Econometrics: Robust Estimation with Limited Data*, Wiley, Chichester.

Wittenberg, M (2009): "Weights: Report on NIDS Wave 1", *NIDS Technical Paper Number 2*, Southern Africa Labour and Development Research Unit, University of Cape Town.

Wittenberg, M (2009a): "Sample Survey Calibration: An Information-theoretic Perspective", Unpublished Mimeograph, Southern Africa Labour and Development Research Unit, University of Cape Town.

Wooden, M (2001): "Design and Management of a Household Panel Survey: Lessons from the International Experience", HILDA *Project Discussion Paper Series No. 2/01*, The Commonwealth Department of Family and Community Services, Australia.

Woolard, I and Leibbrandt, M (2006): "Planning for the South African National Income Dynamics Study (NIDS): Lessons from the international experience", Unpublished Mimeograph, Southern Africa Labour and Development Research Unit, Cape Town.

Wooldridge J.M (2002): *Econometric Analysis of Cross Section and Panel Data*, MIT Press, Cambridge, MA.