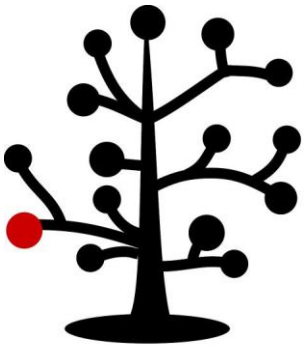


N.i.D.S.

NATIONAL INCOME DYNAMICS STUDY

Household Income: Report on NIDS Wave 1

Technical Paper no. 3

Jonathan Argent
NIDS, University of Cape Town
jonathan.argent@uct.ac.za

July 2009

Contents

1. Introduction

This report gives a brief overview of the income data from the National Income Dynamics Study (NIDS). The derivation of household income measures is addressed first, followed by a discussion of the data quality issues encountered. Non-response emerges as the most significant problem. This is not surprisingly, given that high non-response is a phenomenon consistent across previous household surveys in South Africa that involve income measurement. The treatment of the non-response for the purpose of estimating household income is explained and it's impact is assessed. Discussion of the sample design, household non-response and the weights used to correct for both of these is beyond the scope of this report. A table listing the household and individual level income variables as they appear in the data and the do-files[1] can be found in the appendix.

[1] The relevant do files can be found at http://www.nids.uct.ac.za/home/index.php/Welcome/datasets.html.

2. The construction of household income measurements

There are essentially two sources of income data available in the NIDS data. Firstly there is the 'one-shot' question from the household questionnaire which asks for the total amount of after-tax household income received in the past month. Secondly there are the individual level income questions across all sources. There is a general consensus in the survey literature that incomes calculated by aggregating across questions that individual sources of income are superior to income as measured by a 'one-shot' question.

The total household income measure used to calculate the income quintiles released in the data is calculated using data from *both* sources. As detailed below, where the individual income data is not significantly missing, we aggregate to obtain an estimate of household income. Where individual income data is missing for an entire individual (i.e. unit non-response or mass non-response to the income section) or there is refusal to the questions pertaining to key income sources[2], we substitute the one-shot information from the household questionnaire if this is not missing or refused. Where we still have missing incomes, we use what information we have from the individual income sources. Households that deny all individual income sources and do not give a number in the one-shot are set to zero income. Regardless of the method used in the calculation, implied rental income is added to all households that are not paying rent.

The NIDS questionnaire design focused on the last 30 days prior to the interview in the measurement of income, with a few exceptions in the individual level data. The idea here was to get around the problem of recall bias and obtain a very good snapshot of the welfare of the household over the past month. The trade-off of course is that in such a short period, large once-off incomes (e.g. a retirement gratuity) have an inappropriate effect on the welfare measure. In terms of constructing income from the individual level data some smoothing is thus necessary. The smoothed welfare measure of income does not attempt to estimate national monthly income accurately because (1) some items are excluded due to their distortionary effect, and (2) an implicit ("owner occupied") income is included for housing that is not rented.

[2] This refers to a refusal to acknowledge receipt of this type of income as opposed to a refusal to give the figure. As noted, where a *figure* is refused we can impute.

Certain items that are not included are: inheritance income, retrenchment payments, retirement gratuities, gift income, bridal wealth payments (or lobola) and 'other' income[3]. There were also a few questions that were asked about sources of cashflow that are not technically sources of income (e.g. income from loan repayments), and these are also not included. Some questions asked about the last 12 months. This category includes 13th cheques, payment of profit shares and other bonuses. For the welfare measure these values are converted into monthly amounts by dividing by 12.

The welfare measure defined is intended to be net of income tax. The employment income included was taken from the questions that asked for 'net of tax' employment income. Where individuals gave a gross income and not a net income, their net income was imputed from their gross income using regression. The questions about 13th cheques, profit shares, bonuses and extra piece rate earnings did not ask 'net of tax' questions and so this may result in a slight upward bias in the estimate of income from employment. Self-employment and casual employment questions asked for gross income and so these figures may be overstated too. However in the case of casual wages this is likely to be marginal given that most probably fall beneath the tax threshold. Income from government grants and other government payments (worker compensation, UIF) are not taxed and so this is not relevant. The other sources of income included must also be slightly overstated as they were not asked net of taxes. However these make up a very small portion of total income (and furthermore some such as interest income have a substantial tax exemption), and so this is likely to be a trivial source of bias.

[3] Where 'other' income could be reclassified to a source of income measured by the survey this was done during data cleaning.

3. Non response

The literature on non-response generally acknowledges three main types of non-response, formalised by Rubin (1976). Data is missing completely at random (MCAR) if the probability of response by an individual to a particular question is independent of the answer to that question and independent of all other observed characteristics of the individual. Where the non-response is dependent on at least one observed characteristic of the individual (but is independent of the answer to the question of interest), the data is missing at random (MAR)[4]. Where the non-response is not independent of the answer to the question of interest, the data is said to be 'not missing at random'. These mechanisms of non-response will be referred to in the documentation of the adjustments made for non-response in the data.

The method of mitigating bias from non-response depends on the type of non-response. Where the non-response mechanism is MCAR, there should be no bias in the sample as this is essentially a simple random sample *of the designed sample*. Where the mechanism is MAR (as the response probability depends on some observed variables) the problem can be mitigated by imputations using the variables upon which the probability of response depends. This method may be confounded where too much data is missing or where the variables upon which the probability of non-response depends are unobserved or also have missing data problems. In the case of data being 'not missing at random', it is very difficult to deal with the resulting bias.

In compiling household income figures there are three main problems of non-response. Firstly there is household non-response, which falls outside of the scope of this report. Secondly there are non-respondents within responding households. Thirdly, among individual respondents to the survey there is item non-response. Where an individual professes to earn income from a particular source but does not give the number, we define this as item non-response. This is important because, if left untouched, it implicitly assumes that the person does not earn income from that source even though they have explicitly acknowledged receiving income from particular source[5]. A fourth, subsidiary issue, is that those people who were unavailable for interview, but did not refuse, were interviewed by proxy. Proxy questionnaires makes up 9,4% of the adults from the achieved sample of households. The proxy questionnaire raises some tricky issues because it differs to some extent from the adult questionnaire and it is an open

[4] Clearly MCAR is a sufficient, but not necessary condition to imply MAR.
[5] This includes where the data from the question that asks if the respondent earns income from a particular source is missing; indicates refusal; or indicates that the respondent did not know.

question how the data quality differs given that the questions are not answered by the adult themselves.

3.1 Unit non-response within responding households

Income data for non-responding individuals within responding households is problematic when constructing household income figures. There are three basic options available to the analyst here. Firstly we could assume that the non-response is MCAR within households and we can correct for this by weighting up the aggregate income for the household by the inverse of the non-response percentage. A second option is to assume that those that do not respond have no income and so can simply be ignored when calculating household income. A third option is to assume that the non-response is MAR and, more specifically, MCAR within cells defined by the characteristics of the individual (e.g. race, age, sex, geotype etc).

Table 1: Intra-household adult non-response rate

	Freq.	Percent	Cum.
0%	6,438	88.16%	88.16%
0% - 25%	105	1.44%	89.59%
25% - 49%	323	4.42%	94.02%
50% - 74%	388	5.31%	99.33%
75% - 100%	34	0.47%	99.79%
100%	15	0.21%	100.00%
Total	7,303	100%	

Note: there are only 7303 households in this table (whereas there are 7505 participating NIDS households) because 2 households did not contain any adult members.

Roughly 6,7% of the sample of adults[6] from the achieved sample of households did not respond and for these individuals we have only the information from the household roster. Table 1 above shows the distribution of this unit non-response across responding households. Just over 88% of the 7305 households in the achieved sample had zero unit 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, less than 1% of households had an adult response rate lower than 50%. The 14 households that have 100% non-response to the adult questionnaire are still counted as responding households because household rosters were completed by those households.

[6] This includes adults from proxy questionnaires.

3.2 Item non-response

The table below shows the percentage of missing data for the main income variables. An individual who claimed to have a particular source of income is counted as an observation. The achieved sample includes all those for whom we have a valid figure for this income source.

Table 2: Item non-response for selected income variables (Sections E and F)

Question	Individual level variable	Obs	Achieved	Missing	Imputation
e9-10, e25-26	Main and secondary job	4499	3549	21.1%	Yes (regression)
e42, e43	Casual wages	730	652	10.7%	Yes (regression)
e34	Self employment income*	951	663	30.3%	Yes (regression)
e56	Help friend income	80	71	11.3%	No
e12.1.1	13th Cheque	1207	785	35.0%	Yes (regression)
e12.2.1	Other bonus	552	342	38.0%	Yes (regression)
e12.3.1	Profit share	103	49	52.4%	No
e12.4.1	Extra payment	108	59	45.4%	No
f1.1	Old age pension	2028	1975	2.6%	Yes (rule)
f1.7	Disability grant	871	839	3.7%	Yes (rule)
f1.10	Care dependency grant	47	44	6.4%	Yes (rule)
f1.8	Child Support Grant	2925	2857	2.3%	Yes (rule)
f1.9	Foster care grant	182	172	5.5%	Yes (rule)
f1.5	UIF income	122	81	33.6%	No
f1.6	Workmen's compensation	53	36	32.1%	No
f1.11	Interest/dividend income	136	96	29.4%	Yes (median)
f1.14	Rental income	125	111	11.2%	Yes (regression)
f1.2, f1.3	Private pensions and annuities	290	221	23.8%	Yes (regression)
f1.12	Inheritance	25	19	24.0%	No
f.14, f1.15	Retrenchment payments	62	39	37.1%	No
f2	Inter-household remittances**	1504	1184	24.0%	No

*Contains zeroes - see phrasing of question

**For totalled remittance amounts

There is a general consensus that refusals to income questions are unlikely to be random with respect to income, with those of very high and very low incomes being less likely to respond. There is thus some bias which is inherently difficult to remove. The best that we can do in this situation is to assume that the non-response is in fact MAR and impute missing values accordingly.

3.3 Bracket responses

Historically, questions that probed for sensitive information, particularly income, have experienced a notoriously high rate of non-response (Juster et al, 2007). To mitigate this

problem, surveys now include a backup question for some of these, where the respondent is asked to indicate the earnings *category* into which they fall. This elicits higher response rates from both those that answered 'don't know' and those that refused the original question. There is agreement within the literature that this technique facilitates better estimates of income, although some also argue that the use of so-called unfolding brackets is preferable (see Juster et al, 2007).

Where an individual answers a question with a category rather than a point estimate, we are left with the question of what point estimate to allocate them in the estimation of household and total income. For our purposes we have allocated these individuals to the mid-point of the interval into which they have indicated they belong.

4. Imputation strategy

In the case of item non-response, the problem is essentially that we have individuals who claim a type of income but do not provide the amount of income they receive from this source. Imputation thus takes the form of calculating an appropriate value for these individuals based on similar (in terms of observed characteristics) individuals for whom we have values. This is more complex in the case of unit non-response within households because (1) we do not know if they receive income of the nth type, and (2) we have limited information about them from the household roster. For this reason, the treatment of unit non-response is much more limited.

4.1 Imputation for item non-response

Single imputation inevitably leads to artificially reduced standard errors for estimators (see Rubin, 1976). Multiple imputation techniques provide a means of treatment for this problem (Rubin 1996, Graham and Hofer 2000). At the outset let us note that the imputation pursued for this paper is done with the intention of creating an income variable as a welfare measure for use in papers in which income is not the primary interest. For the purpose of a paper where income is the primary response of interest, multiple imputation techniques would be required.

The purpose of imputation is to reduce the bias created by non-response. Clearly where non-response is too great, the use of this data to impute other data points risks creating more bias than is being mitigated. For income sources where non-response exceeds 40% no imputation is done, following Watson & Wooden (2003). Of course, by not imputing a value we are implicitly imputing the value for these individuals to zero, which would result in understating income where these individuals do in fact have signficicant income from that income source. However, in practice the only cases where this rule actually applied related to income sources with just over 100 observations, which is very small in the context of over 18000 adults. We submit that bias from non-response in these variables is likely to be fairly trivial.

Given the rich content of observed variables available for imputation, regression imputation would appear to be an attractive first option. In terms of a study where the income variables are the main interest, the easy adaptation of this technique for multiple imputation is an added benefit[7]. One problem with this approach is that it requires enough observations to get a

[7] Of course if regression is used in single imputation which is then applied to study the income variable directly this will be problematic as all imputed values will fall on the regression line, strengthening the patterns that exist in the data.

reasonable estimation for the predicting regression. Of course where the observation count is high enough, but the estimation diagnostics suggest the regression is a poor fit, some consideration would have to be given to the underlying assumption that the data are MAR. We propose a minimum observation count for the predicting regression of 100 observations, remembering that a sufficiently good fit is also a requirement even where enough observations are available. Missing data from variables that do not exceed an observation count of 100 are not imputed[8] except in the case of implied rental income which is dealt with separately. The only variables with greater than 100 observations and for which imputations are not done are gift income and remittances; reasonable regression fits could not be found for either of these.

There are some income sources where regression is not an appropriate choice for imputation. For example in the case of the State Old Age pension, where we know that very few individuals receive less than the maximum (Budlender, personal communication, 4 July 2009), we can reasonably impute the maximum amount for all individuals for whom the amount is missing. Similar rules are defined for the other government grants. The Child Support Grant is imputed based on the grant amount that an adult can receive according to the number of biological children that are co-resident with a mother[9].

4.2 Imputation for unit non-response

We take the *a priori* view that unit non-response will be MAR, and so can be dealt with by imputation. However, this type of imputation is difficult because (as previously mentioned) we do not know whether each individual receives income from the nth source, which means that we have to make use of a two step imputation procedure. This type of procedure is demanding in terms of the observation counts and explanatory variables required (especially problematic given our limited information about these individuals from the household roster).

A pragmatic response to this problem is to only impute the most major sources of income and then only when there is sufficient explanatory power available. Together wage income (wages from employment or self-employment, inclusive of 13th cheques and bonuses[10]) and social grant income make up more than 87% of total income; wage income making up about 78% of individual income. Of the social grants only the old age pension and Child Support Grant can be

[8] Of course this is an implicit imputation to zero.
[9] We are aware that this is a simplification of the criteria upon which the Child Support Grant is administered.
[10] The survey asked for 13th cheques or bonues's from the last 12 months. Thus we add 1/12 of these figures to wages to make up a smoothed labour market income figure which we then impute for the non-respondents within responding households.

reliably imputed given observation counts and available explanatory variables, but these are also by far the largest government grants.

The actual method for imputation differs between the two major categories of income for which this is done. For wage income step 1 is to run a probit on all the observations we have with complete wage data. This is then used to predict the probability of a non-responding individual receiving wage income. Step 2 involves a regression of the log of wages on a set of regressors, using the same observations. This is then used to predict a wage for each non-responding individual. The product of the two predicted values (the probability of having income from wages and the amount of wages) gives us the expected value of wages for that individual. Of course this procedure does result in some extremely low numbers, but this should not be problematic in the context of total household income for those households.

For imputation of the income from the Old Age Pension and the Child Support Grant the first step is the same. Probit regressions are used to predict the probability of receiving an Old Age Pension or Child Support Grant (both separately). Using the same procedure that was used for wage income, an expected value of income is generated for both of these grants by multiplying the predicted probability of receiving such a grant by the maximum amount that could be received from that grant.

4.3 Imputation regressions

The regressors used in the imputation regressions include: gender, race, age, trade union membership, province, education, geotype, marital status, home attributes (e.g. number of rooms, type of dwelling etc) and interview month. Some of these regressors have had some prior imputations performed on them. For example where race was missing this was imputed as the mode of household race. The actual specifications across different income categories vary for obvious reasons (e.g. trade union is not included in the regression for self-employment income). Clearly where an individual has missing data (unit or item) and is also missing data for one or more of the regressors that are used in imputation, the regression imputation will fail. Where data is missing for a regressor, the value for that variable is set to zero and a dummy variable is included to control for this.

5. Implied rental income

Implied rental involves problems of non-response and imputation, but is different enough that it warrants individual attention. We have two problems that we face in pursuit of the estimation of implied rental income. The first is the measurement thereof, and the second is the conceptual difficulty involved in the construction of welfare measures. The measurement problem is made up of two parts being (1) the ability of the relevant questions to measure the parameter of interest, and (2) non-response. Part (1) is a consistent problem that cannot be solved, merely mitigated. As for non-response, throughout the rest of this paper we have not imputed where non-response exceeds 40%, but in this case it appears that we have no choice. Table 1 below shows the non-response for each of the questions related to the measurement of implied rentals.

Table 3: Missing data for the use of implied rental income

	Non-owners		Owners	
Item	**Renters**	**Don't rent**	**Mortgage**	**No mortgage**
Amount of bond owing (d7)	n/a	n/a	45%	n/a
Monthly bond payment (d8)	n/a	n/a	33%	n/a
Rent could collect (d9)	n/a	n/a	37%	59%
Rent paid (d11)	8%	n/a	n/a	n/a
Rent would pay (d12)	n/a	64%	n/a	n/a
Market values (d13)	72%	78%	34%	60%
Number in category	1063	845	568	4769

5.1 Renters

This one is the only completely clear cut case. Here we use the rent these people actually pay and this is added to expenditure. This is not an 'implied' rental expenditure, there is actually a flow of payment. Measurement in practice does not really present a problem because NIDS asked for the value of monthly rental payments (if the household claimed to be renting) and non-response was a mere 8%. There is also no obvious reason why measurement in this manner should create a significant bias.

5.2 Don't own, don't rent

These people are living either in an illegal dwelling, or in a house that does not belong to a person from this household (i.e. a person who does not live there, or a firm). The use of the dwelling at no cost constitutes income for these people. The income amount would be the

amount that these people would have to pay in an arm's length rental agreement with the owner, which would be the market rental rate. Clearly this also constitutes an expense since they are making use of the property instead of renting it out, so they are receiving the benefits of the income stream through use. The same figure should thus be added to both income and expenditure. If we did not add the figure to income we would be underestimating their welfare by not counting the monetary value of the free housing. If we do not add the same amount to expenditure then we are ignoring the fact that these people are making use of the housing.

We measure this implied rental income from the question: "how much rent would you pay, if you had to pay to stay here?" The intention here was to measure the market price of rental at this dwelling. However it is possible that people may have interpreted this question to be asking their willingness to pay. There is essentially nothing that can be done to mitigate this possible bias. We do have the 'reasonable market value for the property' as given by the occupants, but unfortunately the non-response in the case of the market value question is 78% for this group; rendering this data almost useless. Non-response of 60%, in the case of the willingness to pay question, is poor too, but somewhat better. The pragmatic thing to do is to impute for the missing 60% off the data we have and simply accept that this method creates some bias. We do not have any other useful source of data for this group.

5.3 Owners

We separate homeowners into two categories; those with mortgage bonds and those without. While both groups receive the same treatment in terms of the flows of implied income from their housing (see below), they are dealt with separately because the mortgage component is an additional complication.

5.3.1 Owners with mortgage bonds still outstanding

This is confusing because there are essentially two linked transactions here that need to be separated out. These are a loan transaction and a purchase transaction. The expense related to the loan is the interest portion of the mortgage repayments. The principal portion of these payments is not an expense but rather an investment (i.e. savings). The purchase of the dwelling is the only part of the transaction that falls into the category of housing. Why should we include the cost of a mortgage in the cost of housing when we do not include the cost of any other means of raising finance to purchase a house?

When a person first purchases the house he then has the right to the economic benefits that flow from ownership: i.e. housing. He may choose to receive the income in the form of rentals,

or he may choose to live in it. If he does choose to live in it, then similarly to the case where a dwelling is owned there is an implied rental income as he is receiving housing from his house. In this case he also has an equal expenditure which is the value of consuming the housing, and which is clearly equal to the implied rental income.

So putting the two transactions together[11]:
Income:

1. Rent* (The value of housing available for consumption)

Expenditure:

1. Rent* (The value of consumed housing)
2. Interest expense (The cost of the use of the bank's financial capital)

Certainly you can measure the rent* portion. This can be measured using the question that asks how much rent the respondent thinks he/she could collect if they were to rent it out.[12] While this question is unlikely to provide an unbiased estimate of the true parameter, it is certainly the best way that this information can be obtained, short of making use of other sources of information (e.g. property agents). The non-response of 37% for this question falls beneath our usual threshold of 40% for imputations.

In contrast, measurement of interest is complicated because while the instalments on a mortgage bond are generally equal amounts, the portion of this that is interest (as opposed to principal) will decline over time (and not in a linear fashion). Thus having no information about where a person is on their repayment timeline and not knowing their interest rate, we cannot guess how much interest they are paying each month on their home loan. In the first month of repayment the interest payment may well be over half the amount of the instalment, but by the final month it will be exactly one month's interest on the final instalment.

However, since NIDS did not collect information on interest expenses (in an effort to avoid questions that may reduce social capital with respondents), the expenditure data does not include any other interest expenses *at all*. It would thus be inconsistent to try and estimate interest expenses from mortgage loans and not from any other type of loan. For this reason, we add rent* to both income and expenditure and acknowledge that there is a general bias in the expenditure data due to the exclusion of interest expenses in general. Households with any debt

[11] Note that this means that the difference in expenditure between the homeowner that is fully paid off and the homeowner with the mortgage bond will tend to zero as the mortgage bond tends towards termination. This is because the interest expense will fall as the time to termination decreases.
[12] Note that this is measuring rent* in a different manner to our measurement by those who don't own and don't rent.

on which interest is charged will thus have understated expenditures in the NIDS estimations and households with mortgage bonds are just a sub sample of this group.

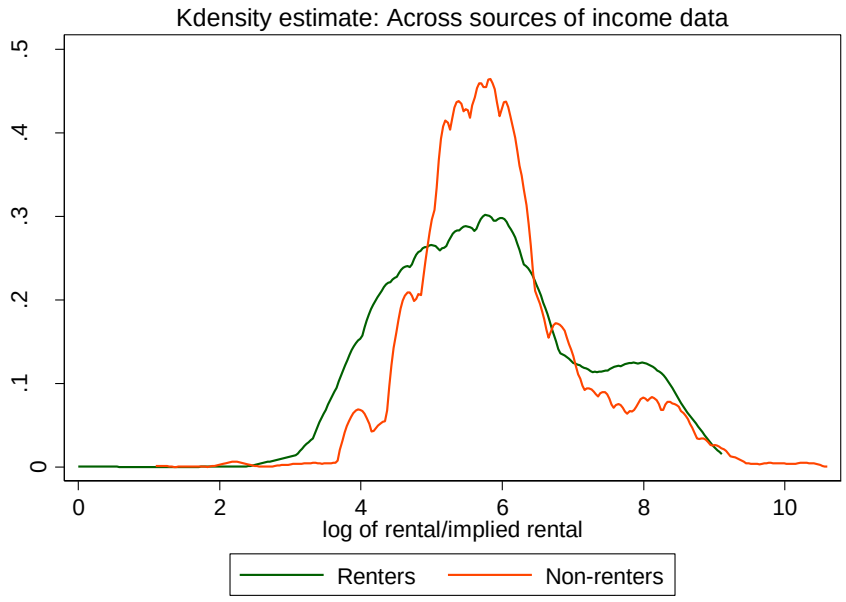
5.3.2 Owners of fully paid off dwellings

These people live in a dwelling that is fully owned (no mortgage) by one of the household members. On the income side, there is clearly an income associated with owning the asset, i.e. the benefits of living in the house. These are valued as the market rental price. Since the household is living in this house, they thus use these benefits themselves and so have an expenditure on housing equal to the income. Clearly we thus add the rent* to both the income and expenditure of the household. Unfortunately we must add the caveat here that there is a measurement problem in terms of non-response as 59% of the respondents in this group did not answer the question on the amount of rent they could collect if they rented the dwelling out.

5.4 The distributions

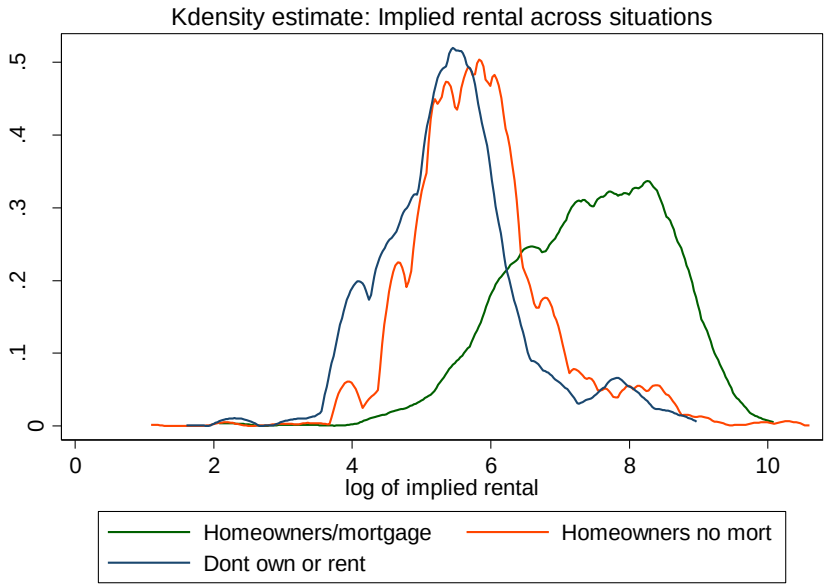
The graph below shows the kernel density estimate of logged implied rentals and logged rental for those that are not renting and are renting respectively. The following graph shows the logged implied rentals across both types of homeowners and those that don't own and don't rent. Those whose housing situation is unknown are not shown (as there are so few of them), but the distribution of their implied rentals is very similar to that of homeowners that do not have a mortgage outstanding.

Figure 1: Kdensity estimate: across sources of income data for renters and non-renters



The distribution of those renting is quite similar to that of those not renting. That the latter distribution is situated slightly to the right is encouraging, given that those renting are likely on average to be less affluent and thus occupying cheaper housing than those who own their properties. In the graph below we can see that the distribution of implied rentals of homeowners that have an outstanding mortgage lies quite far to the right. This is to be expected since those that are able to raise a mortgage are likely to be a more affluent group, occupying more expensive housing.

Figure 2: Kdensity estimate: Implied rental across situations



5.5 Summary

The table below summarises the additions to income and expenditure in the form of rent and implied rent (rent*). The last column summarises the construction of rent/rent*. It does not include interest expenses as these are not housing expenses.

Table 4: Summary of rentals and implied rentals

Category	Expenditure	Income	Notes
Renters	Rent paid	n/a	actual amount paid (D11)
Don't rent / don't own	Rent*	Rent*	amount would pay (D12)
Owners - mortgage	Rent*	Rent*	amount could collect (D9)
Owners - no mortgage	Rent*	Rent*	amount could collect (D9)

It is clear from the table that home owners, with or without mortgages, are treated in the same manner. Those with mortgages would, in a perfect world, also have expenditure equal to the

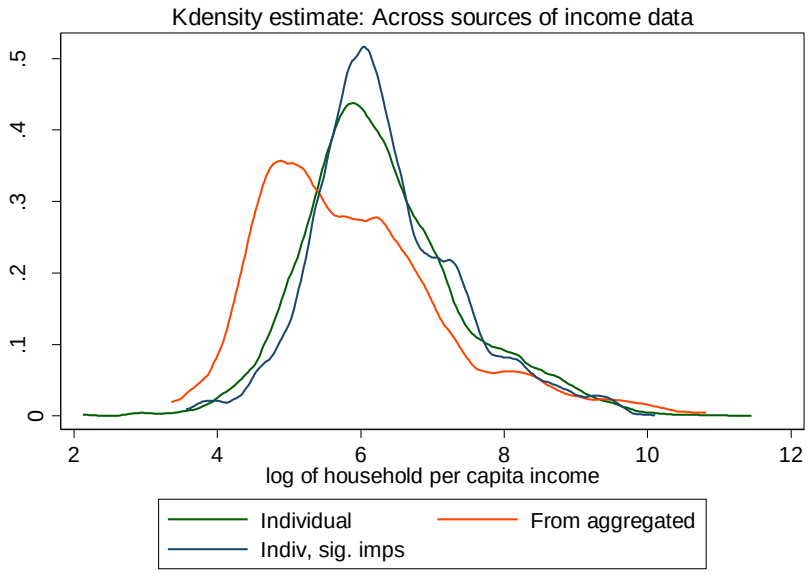
interest payments included under interest expenses. However, as we do not have this data (or any other interest expense data) in NIDS, it is not included and we must simply keep in mind that those with mortgages (or other interest bearing debt) have understated expenses.

Response rates, as previously acknowledged, are particularly poor for those who don't rent and don't own, and those who own and have no mortgage. In these two cases we are forced to depart from our usual rule of not imputing where missing data exceeds 40% of the total observations. Given the size of the values involved, not imputing would very seriously understate monthly income. It is acknowledged that imputation from such a small base will introduce bias in the data. However, we maintain that the gains from not substantially understating income outweigh the bias introduced.

6. Some basic sanity checks on the data

The kernel density estimate below shows the distribution of the household income estimate across the three methods of calculation. Those sourced from the individual questionnaires (green line) represents over 82% of the sample. Those sourced from the one-shot question (red line) represent just under 2% of the sample. The remainder were sourced from the individual with significant and/or unit imputations (blue line). Since these groups are likely to differ it is not clear that the differences in the distributions reflect real differences or differences in data quality.

Figure 3: Kdensity estimate: Across sources of income data



While the final income measure is constructed using both the one-shot source of data and the individual line items from all adults in the household, it is worth comparing the results of the two sources separately for those that have both data. The kernel density estimates below show plots for the log of one-shot per capita household income and the log of the aggregated version[13]. The two are in fact remarkably similar, with the one-shot version being somewhat to the left of the aggregated. The second kernel density plot above shows the same variables as above, but excluding all imputations. The diagrams are very similar, certainly a positive sign for

[13] Note that the two plots use only observations for which we have data for both variables. Where data is missing (and no imputation was possible) observations were not included.

our imputations. This suggests that the differences in the diagram above are more likely explained by real differences rather than measurement differences.

Figure 4: Kdensity estimate: Log of household per capita income – with imputations

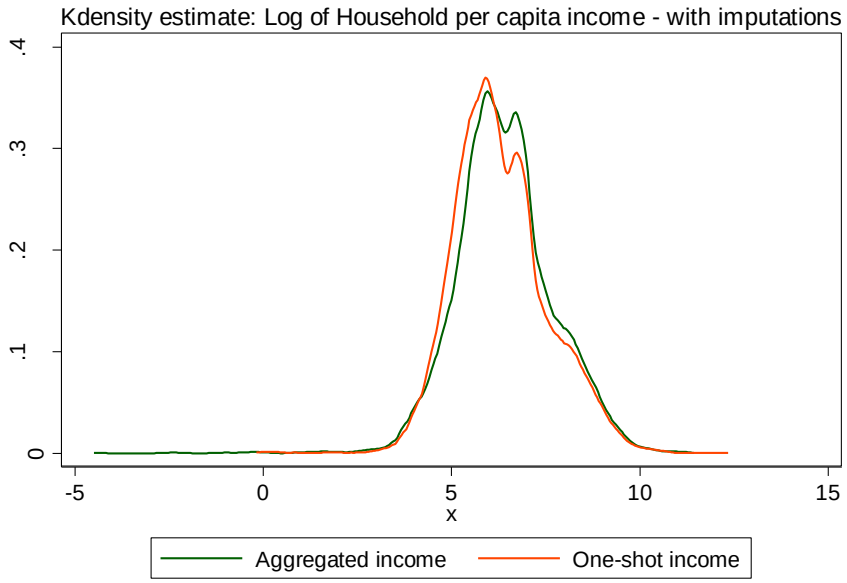
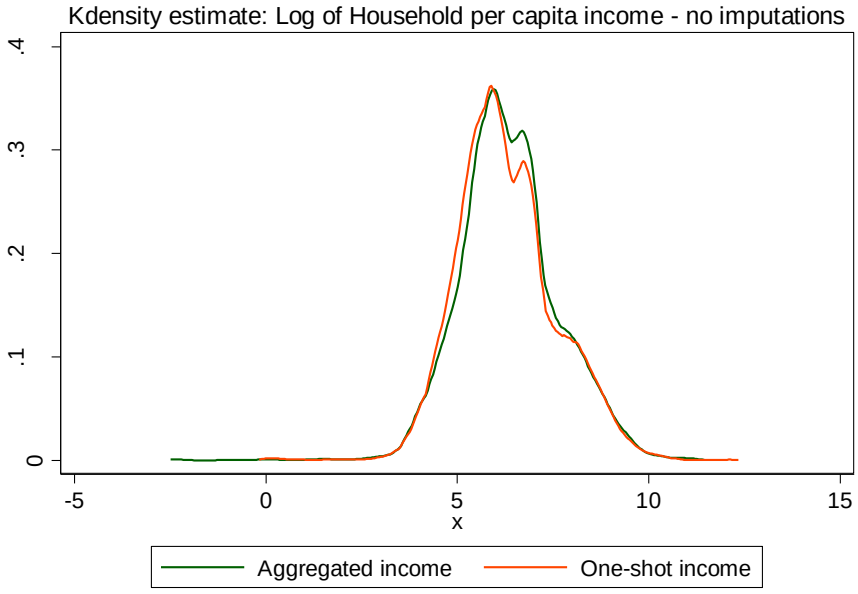


Figure 5: Kdensity estimate: Log of household per capita income – no imputations



The table below (not weighted) compares the final income measure of per capita household income (without implied rental income) with the per capita household income figure obtained from the one-shot measure. In the final row, the one-shot measures being compared are

imputed (since by definition these individuals did not have one-shot data). The first row suggests that the one-shot underestimates per capita household income by just over 10%. The second row when compared with the first row shows that those that don't respond sufficiently to the individual questionnaire have lower median incomes, but higher mean incomes. This could be well explained by some very high income earners refusing the individual questions together with a much larger, poorer group. The result is long tail to the left, but a mean that is slightly higher than the first group. The final row is fairly similar to the first in terms of the differences across measurement techniques. It is not clear which method should be preferred here.

Table 5: Comparing final estimate to one-shot

	Income estimate		**One-shot**	
	Mean	Median	Mean	Median
Individual	991	387	865	300
One-shot	999	200	999	200
Individual sig. Imputation	874	400	749	285

7. Comparison with other data sources in South Africa

The reported IES 2005/2006 data does not give individual or per capita household estimates, which makes comparisons problematic. In addition, the measurement period being annual in IES compared to the previous month for NIDS is likely to result in higher estimates from IES as they will include big once-off incomes that are excluded in NIDS as discussed above. A third problem is that the IES estimates include imputed rent, and given the quality of data NIDS has for this variable, comparisons would be better done without this variable. However if we compare the mean household income as estimated off the weighted NIDS data (R73 176) with that of the IES (R74 589), they are encouragingly close.

Appendix

Table 6: NIDS income variables

Household level variable	*Individual level variable*	*Variable name*
Household (one-shot) (w1_hhq_incb)	n/a	n/a
Labour market income (w1_hhwage)	Main and secondary job	w1_fwag
	Casual wages	w1_cwag
	Self employment income	w1_swag
	13th Cheque	w1_cheq
	Other bonus	w1_bonu
	Profit share	w1_prof
	'Helping friends' income	w1_help
	Extra piece-rate income	w1_extra
Government grant income (w1_hhgovt)	Old age pension	w1_spen
	Disability grant	w1_dis
	Child grant	w1_chld
	Foster care grant	w1_fost
	Care dependency grant	w1_care
Other government income (w1_hhother)	UIF income	w1_uif
	Workmen's compensation	w1_comp
Investment income (w1_hhinvest)	Interest/dividend income	w1_indi
	Rental income	w1_rnt
	Private pensions and annuities	w1_ppen
Income of a capital nature* (w1_hhcapital)	Inheritance	w1_inhe
	Retrenchment payments	w1_retr
	Lobola/bride wealth payments	w1_brid
	Gift income	w1_gift
	Repayment of loans	w1_loan
	Sale of household goods	w1_sale
	Other income	w1_othe
Remittance income (w1_hhremitt)	Inter-household remittances	w1_remt

*Not included in aggregate income calculation

References

Graham, J.W., and Hofer, S.M. 2000. 'Multiple Imputation in Multivariate Research', in Modeling Longitudinal and Multilevel Data: Practical Issues, Applied Approaches, and Specific Examples, edited by Little, T.D., Schnabel, K.U. and Baumert, J., Lawrence Erlbaum Associates, New Jersey.

Juster, F., Cao, H., Couper, M., Hill, D., Hurd, M., Lupton, J., Perry, M., and Smith, J. 2007. 'Enhancing the Quality of Data on the Measurements of Income and Wealth'. Michigan Retirement Research Center, Working Paper WP 2007-151.

Rubin, D.B. 1976. 'Inference and Missing Data', Biometrika, vol. 63, pp. 581-590.

Rubin, D.B. 1996. 'Multiple Imputation After 18+ Years', Journal of the American Statistical Association, vol. 91, pp.473-489.

Watson, N., & Wooden, M. 2003. Towards an imputation strategy for wave 1 of the HILDA survey. HILDA Project.