Some factors influencing the comparability and reliability of poverty estimates across household surveys

DEREK YU

Stellenbosch Economic Working Papers: 03/13
March 2013

KEYWORDS: POVERTY, INCOME, EXPENDITURE, RECALL METHOD, DIARY METHOD, IMPUTATIONS, ASSA MODEL, SOUTH AFRICA
JEL: I32

DEREK YU
DEPARTMENT OF ECONOMICS
UNIVERSITY OF STELLENBOSCH
PRIVATE BAG X1, 7602
MATIELAND, SOUTH AFRICA
E-MAIL: DEREKY@SUN.AC.ZA

UNIVERSITEIT
STELLENBOSCH
UNIVERSITY

BER
BUREAU FOR ECONOMIC RESEARCH

A WORKING PAPER OF THE DEPARTMENT OF ECONOMICS AND THE
BUREAU FOR ECONOMIC RESEARCH AT THE UNIVERSITY OF STELLENBOSCH

Some factors influencing the comparability and reliability of poverty estimates across household surveys

DEREK YU[1]

ABSTRACT

In order to evaluate the extent to which a country achieves the objectives of poverty and inequality reduction, up-to-date, reliable and comparable survey data is required. This paper critically reviews the factors which could affect the comparability and reliability of poverty estimates and trends across various household surveys. First, whether income or expenditure variable should be used for the analyses and whether the diary approach is associated with more reliable capture of income and expenditure information compared with the conventional recall method are looked at. If the respondents are asked to declare the income and expenditure in exact amounts, whether they are asked to report these as 'one-shot' amounts (single estimation approach) or aggregate amount derived from the sum of the amounts for sub-categories (aggregation approach) could affect the poverty estimates. If the respondents are asked to report income and expenditure in intervals, issues that could affect the reliability of this approach, such as the number and width of the intervals, the appropriate method used to approximate the income (expenditure) amount in each interval, as well as the possible methods to deal with households reporting zero or unspecified income (expenditure) are investigated. In addition, survey data is validated against external sources such as national accounts data to investigate if it would lead to improved reliability of the former data for the subsequent poverty analyses. Furthermore, since the survey data are, strictly speaking, not time-series data, the data are re-weighted by means of the cross entropy approach in order to be consistent with demographic and geographic numbers presented by the Actuarial Society of South Africa (ASSA) model and Census data so as to find out if the comparability and reliability of the poverty estimates and trends are improved.

Keywords: poverty, income, expenditure, recall method, diary method, imputations, ASSA model, South Africa
JEL codes: I32

[1] The author gratefully acknowledges the valuable comments by Servaas van der Berg.

1. Introduction

To evaluate the extent to which a country achieved the objectives of poverty and inequality reduction, up-to-date, reliable and comparable data are required. Before the transition, the census conducted by Statistics South Africa (Stats SA) was seemingly the only data source available to analyze money-metric poverty trends. Although the Income and Expenditure Survey (IES) was also a usable dataset, the sample only covered a limited sub-set of households in metropolitan areas of the country. In addition, the 1993 October Household Survey (OHS) excluded the people residing in homelands (Transkei-Bophuthatswana-Venda-Ciskei) from the sample.

Since the political transition in 1994, a major advance by Stats SA was the improvement of the IES and OHS, as the sample was extended to all areas. In addition, new surveys were conducted, such as the General Household Survey (GHS) introduced in 2002, the Labour Force Survey (LFS) which replaced the OHS since 2000, and the Quarterly Labour Force Survey (QLFS) which replaced the LFS since 2007. The sampling design and questionnaire structure of the aforementioned surveys have also been improved throughout the years.

Institutions other than Stats SA also conduct surveys, thereby providing alternative datasets for poverty analyses, such as the Project for Statistics on Living Standards and Development (PSLSD) as well as the National Income Dynamic Study (NIDS) conducted by Southern Africa Labour and Development Research Unit (SALDRU). Moreover, although the All Media Products Survey (AMPS) has been conducted by the South African Advertising Research Foundation (SAARF) since 1975, it has only been used as an alternative data source for poverty analyses in recent years.

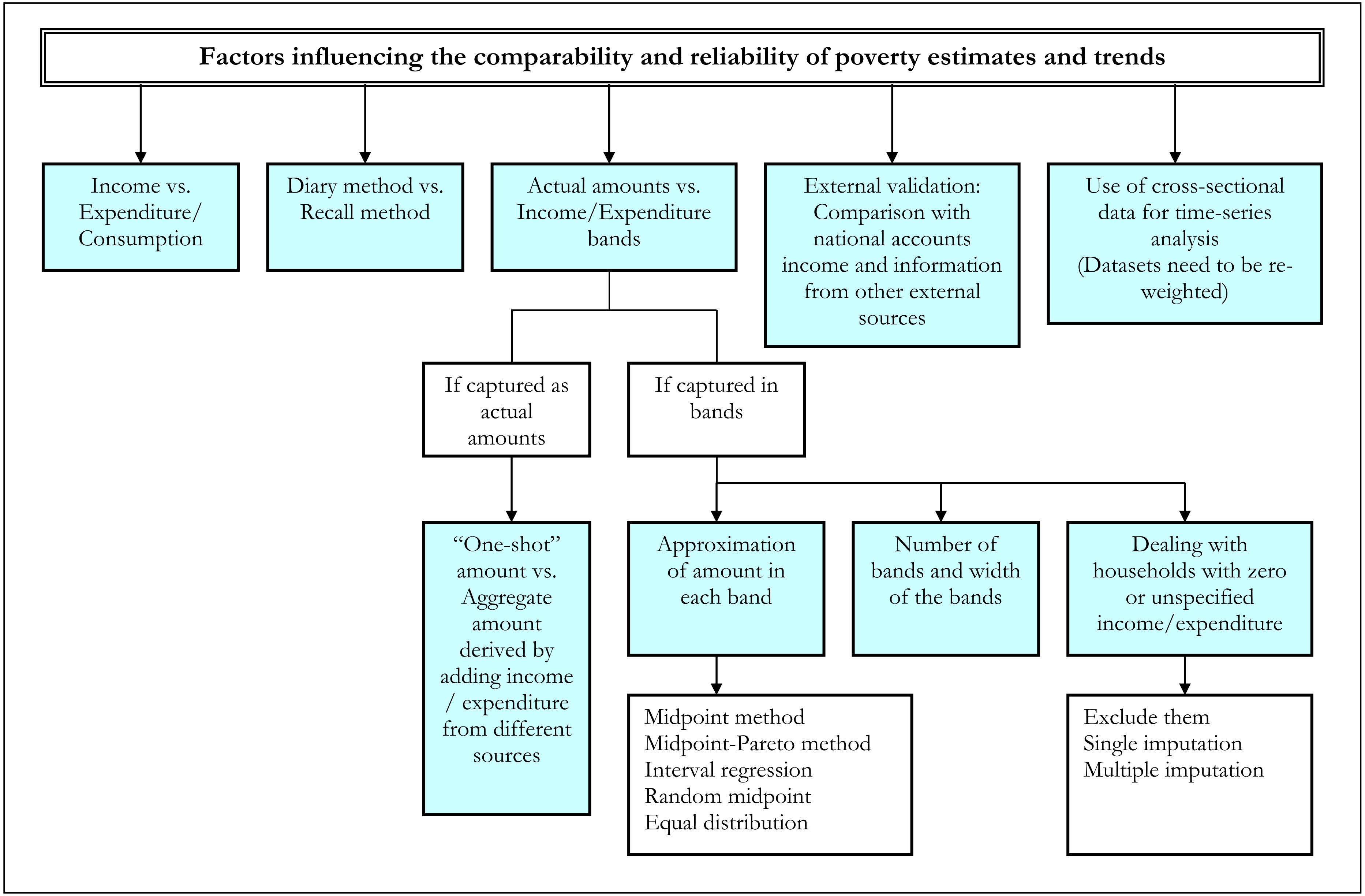
With regard to the use of money-metric variables (e.g., per capita income and per capita expenditure) to derive poverty estimates and trends, several factors could affect the reliability and comparability of the results amongst the surveys (See Figure 1). Firstly, the variable used (income or expenditure) to measure poverty. Secondly, the commonly used method in the South African surveys to collect the income and expenditure information is the recall method, except that IES 2005/2006 adopted both diary and recall methods. It is not certain if the diary method result in better capture of the income and expenditure information as well as poverty estimates.

In some surveys, respondents were asked to report the exact amount, but they were only asked to declare the relevant income or expenditure category in other surveys. Looking at the first method (reporting the exact amount), it could be derived as a 'one-shot', single estimate or derived as the sum of the amounts from different sources. Some argue that the 'one-shot' amount is not precise enough, while the opposing argument is that the aggregation approach is too costly and time-consuming, resulting in inaccuracy of the data obtained due to reasons like interviewee fatigue.

The accuracy of the second method (declaring the relevant category) could be influenced by the number of bands and the width of bands. Another issue is the appropriate method to approximate the income or expenditure amount in each band. Furthermore, almost all surveys included households with zero or unspecified income or expenditure, and this proportion was very high in some surveys (e.g., the two censuses and Community Survey 2007). Rather than simply excluding these households from the analyses, various methods could be applied to impute the income or expenditure of these households.

Survey data could be compared with data from external sources in order to assess the accuracy of the former data, and it has been found that household surveys under-estimated income or expenditure, and hence the data should be adjusted (i.e., shifting the distribution rightwards) in line with the national accounts data.

Figure 1: Factors influencing the comparability and reliability of poverty estimates and trends, using survey income and expenditure data


Factors influencing the comparability and reliability of poverty estimates and trends
Income vs. Expenditure/ Consumption
Diary method vs. Recall method
Actual amounts vs. Income/Expenditure bands
External validation: Comparison with national accounts income and information from other external sources
Use of cross-sectional data for time-series analysis (Datasets need to be re-weighted)
If captured as actual amounts
If captured in bands
"One-shot" amount vs. Aggregate amount derived by adding income / expenditure from different sources
Approximation of amount in each band
Number of bands and width of the bands
Dealing with households with zero or unspecified income/expenditure
Midpoint method
Midpoint-Pareto method
Interval regression
Random midpoint
Equal distribution
Exclude them
Single imputation
Multiple imputation

Furthermore, the survey data were not designed for time-series comparison, as the sampling frame and methodology were not consistent amongst different surveys (e.g., IES vs. CS 2007) and even in a particular survey from different years (e.g., IES 1995 vs. IES 2000 vs. IES 2005/2006 adopted different sampling methodologies). Hence, it is argued that poverty estimates and trends would be more reliable, if the data is re-weighted to be consistent with demographic and geographic numbers presented by the Actuarial Society of South Africa (ASSA) and census data by means of cross entropy approach (Branson 2009).

This paper attempts to discuss these issues by using the aforementioned datasets between 1993 and 2009. Other factors that could also affect the reliability of poverty estimates such as the length of the questionnaire, quality of training received by the interviewers prior the start of the interviews, their experience and efforts devoted to capture information during the interviews fall beyond of the scope of this paper.

2. Household surveys for poverty analyses in South Africa

Table 1 summarizes the collection of income and expenditure information in the seven commonly used household surveys in South Africa. The income information was collected in some surveys but expenditure was collected in other surveys. Some surveys (e.g., IESs) collected both income and expenditure information. In addition, respondents were asked to declare the actual amounts in some surveys (e.g., IESs), but the relevant category in other surveys[2] (e.g., censuses). Looking at the former approach in detail, respondents were asked to declare a one-shot, single-estimate total household income or expenditure amount in some surveys (e.g., AMPSs), but had to report the amounts on each source of income or expenditure, before these amounts were added to derive the total household income or expenditure amount in others (e.g., IESs). Furthermore, IES 2005/2006 was the only survey that adopted the diary approach[3].

Two further issues need to be taken into consideration. First, the Standard Trade Classification (STC) approach was adopted to categorize the income and expenditure items in IES 1995 and IES 2000, but the Classification of Individual Consumption According to Purpose (COICOP) approach was used in IES 2005/2006[4]. Since the COCIOP approach is very different from the STC, in order to have consistent income and expenditure variables across all three IESs for meaningful comparative analyses to be conducted, there are two options: (1) Re-categorize the income and expenditure items in the 1995 and 2000 surveys, using the 2005 COICOP structure; or (2) Re-categorize the income and expenditure items in the 2005/2006 survey using STC.

Secondly, NIDS 2008 was the only survey that asked the respondents to declare the income and expenditure amounts by using both the single-estimate approach and aggregation approach. In that survey, Household expenditure was derived by adding the respondents' answers on food spending, non-food spending and rent expenditure (i.e., aggregation approach), and by asking the respondents to declare the 'one-shot' expenditure amount. Household income was derived by adding the respondents' answers on seven broad components (i.e., aggregation approach), namely wage income, government grant income, other government income, investment income, remittances income, implied rent income and agricultural income. Income information was also collected alternatively by asking the respondents to declare the 'one-shot' income amount. Since SALDRU was worried about the low response rate to the one-shot amount questions[5] and that poverty would be seriously over-estimated as the amounts derived from the one-shot approach

[2] Tables A.1-A.3 in the Appendix present the nominal monthly household income or expenditure categories of surveys that collect the income or expenditure information using the interval method.

[3] Although the diary approach was adopted in IES 2005/2006, it was used in conjunction with the recall approach. The former approach was used mainly to collect non-durable expenditure. For detailed discussion on how the two approaches were adopted in IES 2005/2006, refer to Yu (2008).

[4] For detailed discussion on the difference between STC and COICOP approaches, refer to Yu (2008).

[5] The response rates of the 'one-shot' income and expenditure questions were only 81.1% and 79.4% respectively.

was much lower[6], they decided to use the income and expenditure variables derived by the aggregation approach to conduct poverty analyses in the official NIDS 2008 reports (e.g., Argent, Franklin, Keswell, Leibbrandt and Levinsohn 2009; & Finn, Leibbrandt and Woolard 2009). That is, the 'one-shot' amount variables were not used by SALDRU to derive poverty estimates.

Table 1: Availability of income and expenditure information in South African household surveys: a summary

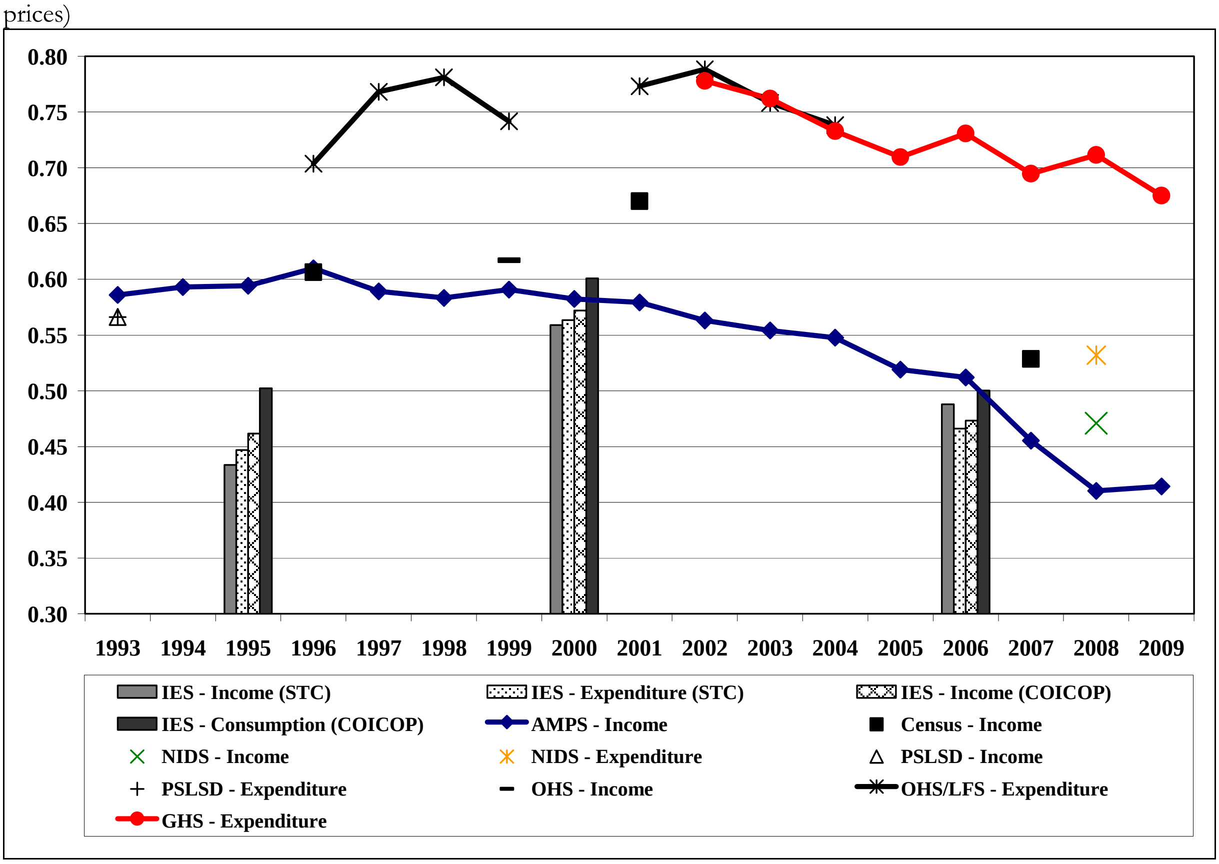
Survey	**Year**	**Question asked?**	**Recall or diary method?**	**Data captured in bands or actual amounts?**	**Overall amount or aggregation of amounts from different sources?**	**Number of bands, if the data is captured in bands**
Income						
Census	1996 2001 2007	Yes	Recall	Bands	Overall	Between 12 and 14
IES	1995 2000 2005/2006	Yes	Recall	Actual amounts	Aggregation	N/A
OHS	1995 – 1999	Yes (1999 only)	Recall	Bands	Overall	8
LFS	2000 – 2007	No	N/A			
QLFS	2008 –	No				
GHS	2002 – 2009	No				
PSLSD	1993	Yes	Recall	Actual amounts	Aggregation	N/A
NIDS	2008	Yes	Recall	Actual amounts	Aggregation Overall	15
AMPS	1993 – 2009	Yes	Recall	Bands	Overall	Between 29 and 32
Expenditure						
Census	1996 2001 2007	No	N/A			
IES	1995 2000 2005/2006	Yes	Recall in 1995 and 2000; recall and diary methods in 2005/2006	Actual amounts	Aggregation	N/A
OHS	1995 – 1999	Yes (In 4 surveys)	Recall	1996 – 1998: Actual amounts 1999: Bands	Overall	8 (1999)
LFS	2000 – 2007	Yes (In 4 surveys)	Recall	Bands	Overall	8
QLFS	2008 –	No	N/A			
GHS	2002 – 2009	Yes	Recall	Bands	Overall	Between 8 and 10
NIDS	2008	Yes	Recall	Actual amounts	Aggregation	N/A
AMPS	1993 – 2009	No	N/A			

[6] Looking at households that reported both 'one-shot' and aggregated household expenditure, the former figure was only R237 364 million (2000 prices) while the latter figure was R466 683 million (2000 prices). In contrast, with regard to households that reported both 'one-shot' and aggregated household income, the former figure was only R429 561 million (2000 prices) while the latter figure was R531 525 million (2000 prices).

3. Poverty trends since the transition

This section uses a lower bound poverty line (R322, per capita per month, 2000 prices)[7] proposed by Woolard and Leibbrandt (2006) to examine the poverty estimates and trends across the surveys between 1993 and 2009[8]. Figure 2 presents the results in terms of poverty headcount ratios between 1993 and 2009[9]. With regard to the poverty trends using the two censuses and CS 2007, the poverty headcount ratio increased between 1996 and 2001, before a sharp decline took place between 2001 and 2007. The 2007 poverty headcount ratio was lower than the 1996 ratio.

Figure 2: Poverty headcount ratios in each survey, 1993-2009 (Poverty line: R3 864, per capita per annum, 2000 prices)



The poverty headcount ratio increased rapidly between IES 1995 and IES 2000, before a downward trend was observed between IES 2000 and IES 2005/2006. This trend took place regardless of whether the STC or COCIOP approach was adopted. However, the IES 2005/2006 poverty headcount ratio was still slightly above the IES 1995 ratio. It was argued by Van der Berg, Louw and Du Toit (2008) that the extent of increase of poverty could be over-estimated, since there was a large drop of recorded income (or expenditure) between IES 1995 and IES 2000[10]. Such a great drop in income between the two surveys was unlikely, as it was larger than the decrease experienced by the South African economy during the Great Depression of the 1930s. In addition, this decrease was also larger than the decline experienced by some of the affected countries during the 1998 Asian economic crisis. Thus, it seems certain issues (e.g.,

[7] This poverty line was calculated by observing the essential non-food expenditure of households that spent approximately R211 on food (i.e., the food poverty line), and it was found that the former amount was R111. Hence, the lower bound poverty line was equal to R322 (= R211 + R111).

[8] Since the household income variable was not derived correctly by Stats SA in Census 1996, the income variable derived by Yu (2009) will be used for the forthcoming analyses. In addition, the Census 2001 income variable before hotdeck imputation was conducted by Stats SA will be used for the remainder of the paper.

[9] Figures A.1 and A.2 in the Appendix show the poverty gap ratios and squared poverty gap ratios in each survey using the same poverty line. In this paper, the focus of discussion is on the poverty headcount ratios.

[10] The national accounts income data showed that national income increased between the two years.

differences in sampling methodology) made the comparability of IES 1995 and IES 2000 difficult, and the poverty results between the two surveys should be interpreted with caution. Furthermore, since income (or expenditure) was very poorly captured in IES 2000, while IES 2005-2006 was the survey that captured income best, the extent of the decline of poverty between these two surveys could be over-estimated.

Using the OHS and LFS per capita expenditure variable, the poverty headcount ratio increased since 1996, before a downward trend was observed from 2002. In addition, the 2004 poverty headcount ratio was slightly higher than the 1996 ratio. In the GHSs, a continuous downward trend in poverty was observed between 2002 and 2005, before an unstable downward trend was observed between 2005 and 2009. The LFS 2002-2004 poverty headcount ratios were extremely close to the GHS 2002-2004 results. Furthermore, the poverty headcount ratios in OHSs, LFSs and GHSs were much higher (always above 0.70) than the results using censuses and IESs (and also AMPS, NIDS and PSLSD, to be discussed below).

In AMPSs, there was not too much change in the poverty headcount ratio before 2000, as it stabilised at approximately 0.59 between 1993 and 1999, before a continuous downward trend took place between 2000 and 2008. This trend is very different than what was found when looking at the censuses, IESs and OHSs, as these surveys indicated that poverty increased since the transition, before a downward trend took place since 2000. Furthermore, the AMPS poverty headcount ratios have always been lower than the ratios derived in OHSs, LFSs and GHSs.

The 1993 PSLSD poverty headcount ratios, regardless of whether the income or expenditure variable was used, were slightly below the 1993 AMPS ratio. However, the poverty headcount ratio was slightly higher for the income variable (0.598, compared with 0.566 when using the expenditure variable). In contrast, in NIDS, the poverty headcount ratio was higher if expenditure was used (0.532) while the ratio using income (0.471) was closer to the ratio in AMPS 2008 (0.410).

To conclude, despite the fact that the levels of poverty differed across the surveys, it was found that poverty increased since the advent of democracy until about 2000 in all surveys under study (except AMPSs, which showed that poverty stagnated in the 1990s), before a downward trend took place in the 2000s.

4. Factors affecting the reliability and comparability of poverty estimates and trends

4.1 Income vs. Expenditure / Consumption

An important question that arises when using the money-metric approach to measure the poverty of the population is whether income or expenditure / consumption should be used. The general argument (Haughton and Khandker 2009: 30) is that most rich countries use the income variable, as most of the income comes from salaries and wages and hence it is comparatively easy to measure, while it is difficult to quantify both the volumes and amounts of purchase when it comes to capturing expenditure. In contrast, in poorer countries, income is harder to measure as much of it comes from self-employment in informal activities, but consumption / expenditure is more straightforward and easier to estimate. Thus, consumption is the preferred variable.

Figure 3 shows the total income or expenditure of surveys that collected information on both income and expenditure, and it can be seen that, in all of these surveys except IES 2005/2006, income was greater than expenditure, contrasting with the general argument as discussed above that expenditure is captured better in poor, developing countries (with South Africa being one of them). In addition, Figure 4 shows that the poverty headcount ratios were higher using the per capita expenditure variables in all of these surveys, except PSLSD 1993 and IES 2005/2006.

Figure 3: Total income and expenditure (Rand million, 2000 prices) of surveys that collected both income and expenditure

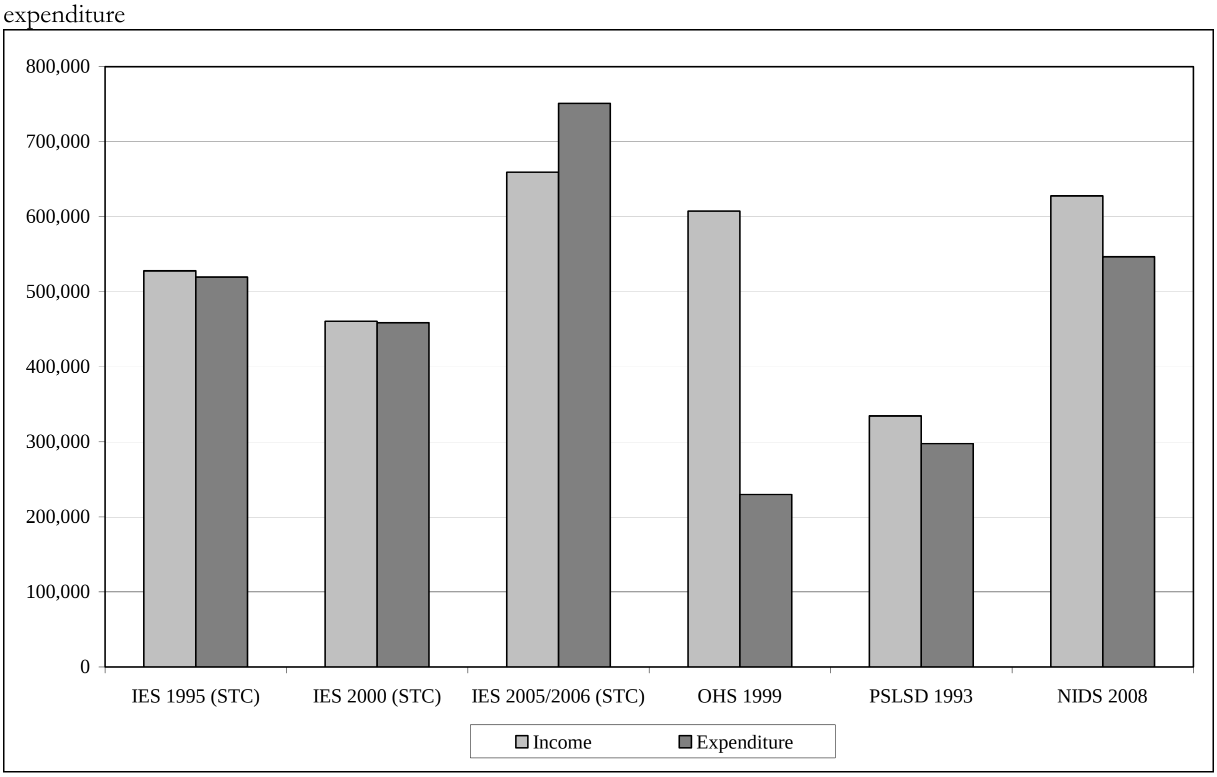
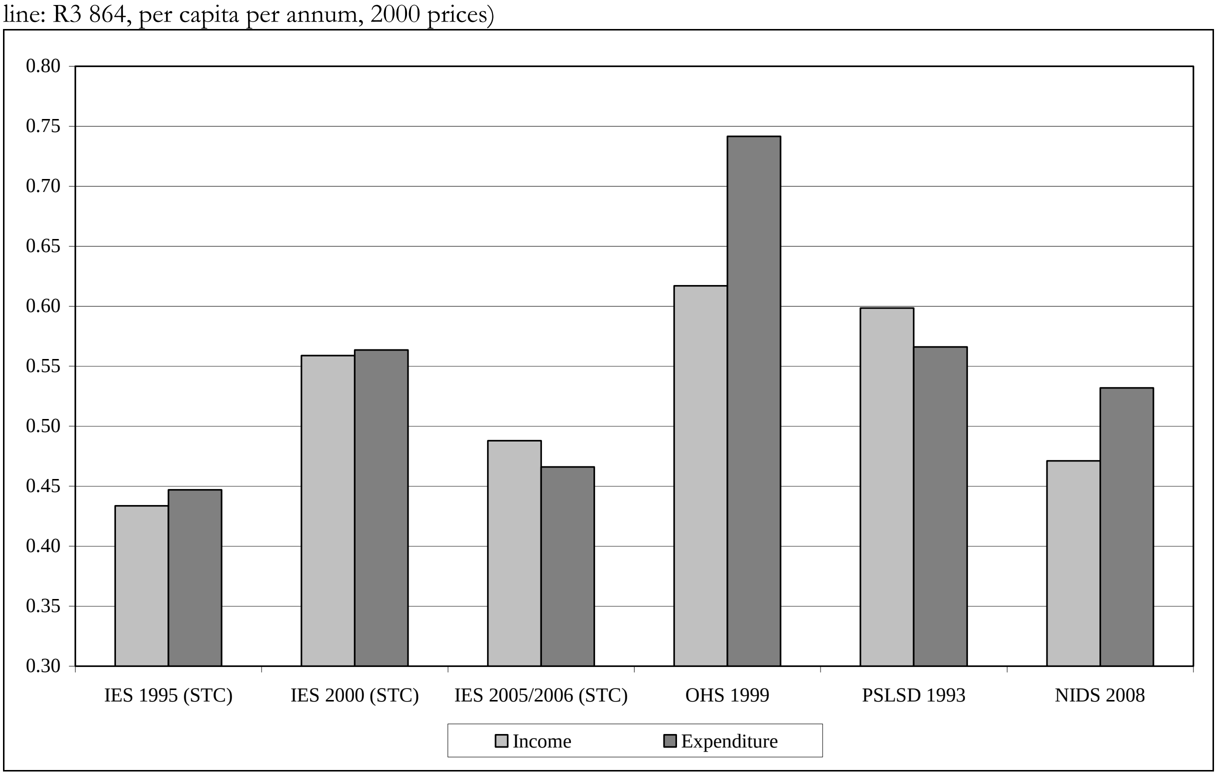
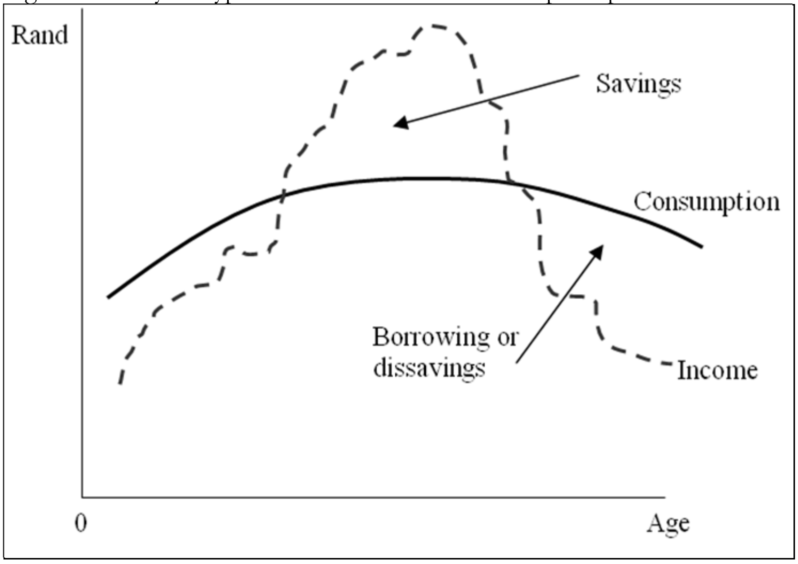


Figure 4: Poverty headcount ratios in each survey that collected both income and expenditure information (Poverty line: R3 864, per capita per annum, 2000 prices)



The primary reason to use the expenditure / consumption variable is that, in addition to fluctuating somewhat from year to year, income normally increases and then decreases in the course of a person's lifetime. In contrast, consumption remains relatively stable, since it could be smoothed to some extent by saving and borrowing (Blundell and Preston 1998: 603; McKay 2000: 85-86; Duclos and Araar 2006: 21; Haughton and Khandker 2009: 24-25). The permanent income hypothesis predicts this smoothing of short-term income fluctuations. Under this hypothesis, transitory (temporary) income is saved, while long-term (permanent) income is largely consumed (See Figure 5). Information on consumption over a relatively short period is more likely to represent a household's general level of welfare than the equivalent information on the more volatile income (Haughton and Khandker 2009: 25). Although random irregularities and seasonal patterns are present in consumption, they are normally smaller than those of income, as consumption is less tied to seasonal and weather-related patterns in agriculture than is income (Deaton and Grosh 2000: 93-94)

Figure 5: Lifecycle hypothesis – income and consumption profile over time



Source: Haughton and Khandker (2009: 24).

Secondly, the concept of consumption – giving money in exchange for a good or service – is clear to both interviewers and interviewees, while the income concept might not be clear (to be discussed later). Consumption is also more readily observed, recalled and measured than income (at least in developing countries, although this is not always the case) (Deaton and Grosh 2000: 93-94; Duclos and Araar: 2006: 21). Thus, it is easier to recall information on consumption. Finally, consumption is preferred over income as the former shows the current actual material standard of living by reflecting more directly the degree of commodity possession (Haughton and Khandker 2009: 30).

Using expenditure / consumption instead of income to measure money-metric poverty also has its drawbacks. First, there is a need for collecting data on consumption on goods and services item by item, in the case of an aggregation approach. The number of consumption items could be as many as more than a thousand, while the income source items are much fewer[11]. Secondly, although respondents are more likely to remember consumption activities in greater detail and to report higher expenditure if the questions are more detailed (Haughton and Khandker 2009: 25), such a longer questionnaire (e.g., if the aggregation approach is adopted) devoted to collecting consumption information is very costly and time-consuming. However, if a shorter questionnaire is used in order to save money and time (e.g., if the single estimate approach is adopted or if the respondents are only asked to report the overall food and non-food spending, instead of being asked to declare spending on each food and non-food item), this could result in inaccurate estimates of total consumption (Guenard and Mesple-Somps 2010: 523).

Thirdly, the respondents might not provide answers to all consumption items or might not remember the amounts spent on all items, and so imputations have to be made (Deaton and Grosh 2000: 93-94). Fourthly, overly long recall periods (e.g., one year) could lead to under-estimation of consumption as memories fade as time goes by, i.e., recall bias arises (Guenard and

[11] For example, in the 57-page IES 2000 questionnaire, only 6 pages were devoted to collecting information on income, while about 45 pages of the questionnaire were aimed at collecting consumption / expenditure information.

Mesple-Somps 2010: 523), but longer recall periods might really be required for durable goods with low purchase frequency. Furthermore, households are likely to under-report what they have spent on luxury or illicit items, e.g., alcohol, tobacco, or drugs. With regard to consumption on durable goods, such expenses are not regularly incurred, so the data could be noisy because recall bias is more likely to happen with longer recall periods. Looking at durable goods consumption in greater detail, it is difficult to measure it, as it is not sure whether the full consumption amount on a durable good should be included, or whether only the change in the asset's value during the year (i.e., depreciation, plus the cost of locking up one's money in the asset) should be included[12] (Haughton and Khandker 2009: 25).

Deaton (1997: 32) argues that the presence of substantial inflation tends to overstate consumption relative to income, given that surveys usually have different reporting periods for the two variables. The reference period for consumption varies from item to item (e.g., the reference period could be one week, one month or one year for different items in IES 2000), while the importance of seasonality of income means that reference periods for income items are usually a year. Consumption is then denominated in more recent, higher prices than is income, resulting in an upward bias to measures of consumption[13]. Another disadvantage of using consumption relates to the difficulty of disentangling production and consumption (Deaton 1997: 28). As most agricultural households are both producers and consumers, they might find it difficult to distinguish consumption from production. In addition, home-produced items, typically food grown or raised on the farm or in kitchen gardens, should be properly recorded as both income and consumption, but are often very difficult to value.

Having discussed the pros and cons of using the consumption variable, there are also arguments for and against using the income variable. The main argument in favour of using income is that it is easier, cheaper and quicker to collect income data, especially in circumstances where income comes from only one or two sources (e.g., wages and pension) that are easily remembered or for which independent documentation exists (Deaton and Grosh 2000: 93-94). This is more likely to happen in richer, developed countries. Even if the household's income might come from many sources, it is still relatively easier to measure income than consumption, given the limited number of income sources (e.g., salaries and wages, pensions, remittances, interest received, income from businesses, etc.).

As far as the problems of using income are concerned, as mentioned previously (the lifecycle hypothesis), income of many households could be very volatile seasonally during the year, or from one year to another year, as a result of being subject to significant shocks. This is more likely to happen in households engaging predominantly in self-employment, or agricultural activities or households that are heavily reliant on transfers from either public or private sources. As a result, measuring the household annual income might require a lot of visits to the household or dependence on the ability of households to remember their income from many months earlier (Deaton and Grosh 2000: 93-94; McKay 2000: 84-86). In addition, as a result of the volatile

[12] In all surveys under study, the full consumption amount on a durable good was included, providing it took place during the reference period as specified in the questionnaire. For instance, assuming two respondents taking part in IES 2000 was asked the expenditure on vehicles in the past 12 months. The first respondent purchased a vehicle during 2000, and hence he reported the full expenditure on this vehicle. The second respondent bought a vehicle in 1999, so he would not report this expenditure in IES 2000, despite the fact that he was using the vehicle at the time of the survey.

[13] The following example could simplify the explanation: if a respondent takes part in the survey in December and is asked to declare the food consumption for the past month, and his answer is R100, then the annual food consumption is derived as R1 200 (R100 × 12 months). However, the same basket of food might be cheaper in the earlier months of the year, and if the respondent is asked to declare the annual food consumption (i.e., longer reference period) instead, the actual amount could be lower than R1 200 (providing the respondent remembers the food expenditure month by month clearly – keep in mind recall bias is more likely to happen with longer recall periods), i.e., the R1 200 amount derived using a shorter reference period might be biased upwards due to the impact of inflation.

nature of income (the income could be temporarily high or low), the reporting period might not be able to capture the 'mean' income of the household accurately.

The concept of income, especially income from self-employment or own-account agricultural and informal activities, is often unclear (Deaton and Grosh 2000: 93-94). Respondents might not genuinely know how much income they make in these activities, in particular due to reasons like seasonal variations, income declarations are biased by under-declarations and non-responses, or respondents might not feel there is a need to report incomes earned infrequently or might not consider them as part of income, e.g., receipt of transfers and remittances and other non-labour income (McKay 2000: 95; Haughton and Khandker 2009: 30; Guenard and Mesple-Somps 2010: 527). In addition, although income sources / items are fewer compared with consumption sources / items as mentioned above, Riphahn and Serfling (2004) argue that income sources could be diverse, especially among the professionally self-employed in rich countries. A greater cognitive requirement on the respondent to provide information could result in lower response rates or the reporting of unreliable income information.

Posel and Casale (2005: 4) argue that each member of the household is more knowledgeable about his/her own income than about the income of the other members. Hence, item non-response – not specifying the household income – is significantly higher for proxy-reporting (i.e., only the household head or one member of the household is asked to declare the total household income earned by all members of the household, as in the OHSs, LFSs, GHSs and AMPSs) than for self-reporting (i.e., each member of the household is asked to declare his/her personal income, before the personal incomes of all members are added to derive household income, as in the two censuses and CS 2007)[14].

It is also argued by various researchers (Deaton 1997; Deaton and Grosh 2000; McKay 2000; Posel and Casale 2005; Haughton and Khandker 2009) that respondents are more likely to report inaccurate information about their income or refuse to declare the full extent of their income, as income is a more sensitive topic to ask about than consumption. This could be due to the fact that, as income is taxable in almost all countries, it is difficult for interviewers to convince respondents that the information they provide will not be passed on to the tax authorities. As a result, income would be reported inaccurately or understated.

Some respondents might be hesitant to report income earned illegally, such as smuggling, corruption or prostitution, and income earned from informal activities not reported to the tax authorities, such as street vending. Another reason the respondents might feel sensitive to disclose income information is that, income from assets is harder to capture, with the wealthy being typically thought to be less likely to co-operate as they might fear governmental or other uses of the data. In contrast, low-income earners might overstate their income, as they might think that by reporting low earned income they are considered being unsuccessful.

4.2 Recall method vs. Diary method

Regardless of whether income or consumption is chosen to measure poverty estimates, an important issue is how to collect the information. In all South African surveys under study, the recall method was adopted. The only exception is IES 2005/2006, which adopted the diary method for the first time to complement the recall method. Table 2 presents how the total income or expenditure was derived in this survey.

[14] However, this is not the case in the South African surveys. A fairly high proportion of people did not specify their personal income in the two censuses and CS 2007. This consequently resulted in a higher proportion of households with unspecified household income (because household income was derived from personal incomes). In OHSs, LFSs and GHSs, only the household heads were asked to report household income or expenditure, and the proportion of households with unspecified answers was lower than in the censuses and CS 2007. This will be discussed in greater detail in Section 4.7.

Table 2: Derivation of the annual income and expenditure, IES 2005/2006

Type of data item	**Reference period**		**Annualized figure**
	[A]: Diary (Survey month)	**[B]: Main questionnaire**	
Non-durable items	1 month	–	[A] × 12
Semi-durable items	1 month	11 months	[A] + [B]
Durable items	1 month	11 months	[A] + [B]
Services	–	1 or 12 months	[B] (if reference period is 1 month) [B] × 12 (if reference period is 12 months)
Regular income	–	1 and 11 months#	Monthly figure + 11-month figure#
Irregular income	–	12 months	[B]

In IES 2005/2006, respondents were asked to declare income for the previous month and income for the 11 months prior to the survey month for all regular income items. These two figures were then added before the annualized figure was derived.

Note: When Stats SA released the IES 2005/2006 data, only the aggregate income and expenditure amount of each item was given (e.g., assuming expenditure on food was R1 000, it was not known if, for example, R600 of this amount was derived from the diary method and the remaining R400 from the recall method).

The recall method is problematic for various reasons. First, recall bias is very likely to happen, as the respondents could not remember many purchases long after they have been made. This is likely to result in either an under-estimation of consumption (as it is not easy for people to remember their consumption from long ago) or inaccurate guesses (i.e., respondents estimate their consumption over the whole year from their current rate of consumption) (Deaton 1997: 24-25; Deaton and Grosh 2000: 109-110). This recall bias becomes more serious as the recall period increases.

The telescoping phenomenon – respondents tend to include consumption events that took place before the beginning of the recall period (Deaton and Grosh 2000: 110) – is also likely to happen under the recall method. As a result, consumption could be over-estimated. For instance, when asked about expenditures during the previous year, respondents might include items they bought 13 months ago. Deaton and Grosh argued further that telescoping is more likely to happen in durable goods purchases and/or if the recall period becomes longer, since respondents are more likely to forget the date the consumption events occurred. Hence, for example, if a household taking part in the survey in October 2009 purchased a personal computer worth R5 000 in September 2008 (i.e., more than a year ago), but wrongly thought that it was bought in October 2008 and included it as part of expenditure for the recall period, this would result in the over-estimation of total expenditure.

Deaton (2005: 16) suggests a shorter recall period for accuracy of memory. Moreover, if the respondents' memories of their consumption fade quickly, many visits might be required throughout the year to ensure that data on high-frequency non-durable purchases are collected accurately, but the resultant increase of the frequency of the survey could be costly. In contrast, as the consumption of some items might only take place occasionally during a year, a longer recall period is required, e.g., consumption of durable items like motor vehicles might not have taken place in the last month but rather in the last year.

Looking at the issue of recall period further, the match between consumption and purchases is more accurate when averaged over a longer recall period (Deaton 2005: 16). For example, if the respondent is asked to declare consumption on food in the past month and the respondent takes part in the survey in December, it is likely that his/her food expenditure is higher than usual due to the festive season, and the resultant annual food expenditure derived from this monthly expenditure could be over-estimated. However, if the respondent is asked to declare the total food expenditure in the past 12 months, the seasonal fluctuations (i.e., food expenditure is lower at the start of the year but then higher in certain months) might be considered by the respondents (providing he/she remembers the monthly food expenditure with good memory), and the resultant food expenditure could be more accurate.

As a result of the drawbacks of using the recall method as discussed above, the diary method becomes an alternative approach to collect income and consumption information. Corti (1993) argues that it is a reliable alternative to the conventional interview method (which adopts the recall approach) for events that are easily forgotten or difficult to recall correctly; the diary method is designed to minimize dependence on respondents' memories and consequently reduces the likelihood of recall bias, especially on frequently purchased (non-durable) items which are normally more difficult to recall, since consumption events are recorded as they take place or close to that time (Deaton and Grosh 2000: 109; Battinstin 2003: 2; Wiseman, Conteh and Matovu 2005: 395). The diary method is also more convenient to the respondents, as they could answer the questions at a time and place that are suitable for them (Deaton and Grosh 2000: 119-122; Wiseman et al 2005: 395).

The diary method also helps to reduce the problems associated with gathering sensitive information by personal interviews. For example, the respondent might feel uncomfortable if he/she is asked by the interviewer to recall total consumption on items like alcohol and tobacco, but will feel more comfortable to report the consumption on these items on a diary without the presence of the interviewer. Finally, diaries allow for the analysis of events over time (Wiseman et al. 2005: 395). For instance, it is possible to look at the effect seasonality has on expenditure, particularly in poor rural communities, if the diary method is adopted[15].

Despite its merits, the diary method is associated with various problems. First, diaries are less suitable where literacy levels are low, because the diary keepers might not be able to write down the purchase items correctly if given an unstructured diary so as to enter consumption activities on a blank page (this is the case in the IES 2005/2006 diary approach, as the respondents were asked to describe the items, place of purchase and the consumption value on the weekly diary). Even if the diary is structured like a questionnaire in which the participants are only required to tick the printed boxes containing the consumption events and fill in the consumption amounts, some of them might not be literate enough to understand the meaning of these consumption items (Wiseman et al. 2005: 396). Hence, the data collected from the diaries might be biased towards the competent, literate diary keepers (Corti 1993). A pictorial diary might be required to improve the accuracy of the responses of the people with lower literacy levels.

Deaton (2005: 16) and Wiseman et al. (2005: 399-400) argue that the diary method might not suit the more diverse, well-off households with bigger household size; if the responsibility for spending lies with more than one person in the household, individuals have insufficient knowledge of what each household member spends. Moreover, as some family members are outside home most of the time, multiple diaries per household should be considered, but it would become much more costly and time-consuming to collect and edit the information. Consequently, overlap in entries made by different family members could happen.

If the households are asked to keep the diaries for a very short period of time (e.g., one week, or four weeks in the case of IES 2005/2006), the resultant consumption estimate might be inaccurate, as some households have unusually low purchase rates in some items (e.g., every month or every few months, especially the semi-durable and durable goods). Hence, the diary method might work better for non-durable items as the purchases of these items take place more frequently; the recall method might work well to record the consumption of the more durable,

[15] It is also possible to observe this seasonality effect in the recall period, providing the respondents are, for example, asked to declare expenditure on the items in each of the last 12 months. However, this approach was not adopted in all surveys under study.

bulky items with low purchase frequency (Deaton and Grosh 2000: 119-122; Battinstin 2003: 2), despite the fact that recall bias is more likely to happen in the latter approach due to the longer reference period required. This argument might explain why the recall method (questionnaire) was still used in IES 2005/2006 to complement the diary method, with the recall method focusing on collecting information on income as well as semi-durable and durable goods consumption[16], and the diary method primarily concentrating on the collection of non-durable consumption information.

Telescoping and recall bias, as discussed previously, could still happen even if the diary method is adopted, despite the fact that the likelihood of it happening becomes lower, as the diaries still rely on the respondents' memory and might not be filled out every day (Deaton and Grosh 2000: 119-122). The chance that these two problems would occur increases if entries are not made as close as possible to the time of actual expenditure, since the respondents are left to their own devices to complete the diary and there is no guarantee that the respondents would report events immediately after they took place (Deaton 1997: 24-25 & Wiseman et al. 2005: 398). For example, if the respondent purchased various goods at a supermarket one day but the entries were only made on the dairy a few days later, consumption amounts might not be recalled correctly and the consumption of some goods might be forgotten and eventually not entered at all on the diary. Hence, the researchers might need to visit the households frequently to actively encourage them to regularly update the diaries. If it is found that there are missing data (e.g., consumption items are entered on the diary but the amounts spent are not reported), then the researchers have to go back to the respondents to clarify entries, but the data would eventually become retrospective and once again subject to recall bias (Wiseman et al. 2005: 395).

Corti (1993), Deaton and Grosh (2000: 119-122), Wiseman et al. (2005: 395) and Ahmed, Brzozowski and Crossley (2006: 9-10) argue that the 'first-day effect' is likely to happen in the diary approach: the first day and first week of diary keeping show higher reporting of consumption than the following days/weeks. It could be explained by various factors: the novelty of diary keeping wears off as time goes by; the respondents feel exhausted to keep records and eventually become less detailed in their reporting; the diary keepers no longer carry their diaries with them[17]. This is why, as mentioned above, intermediate visits from the interviewers or even incentives are required to preserve good diary keeping until the end of the period.

The recording of the use of illicit drugs or income earned that is not declared to tax authorities might remain inaccurate under the diary method, even though it does not involve face-to-face communication as it happens in the recall (interview) method, as the respondents could still feel sensitive to enter such information on the diaries, and eventually decide not to fill in the above information at all on the diaries (Wiseman et al. 2005: 395).

Although the diary method reduces the duration that the interviewer spends interviewing the households, this method might increase the time that the interviewer must spend travelling, as it requires additional trips to collect the completed diary. Moreover, considerable time might also be spent assisting illiterate households to fill out the diaries. Furthermore, the interviewers might

[16] This implies that inaccuracy in the durable goods consumption data is inevitable to a certain extent, regardless of which method is adopted: if the recall method is adopted, a longer reference period is required to collect reliable information since such consumption happens only occasionally, but a longer reference period is associated with a greater likelihood of recall bias and telescoping. If the diary method is adopted, durable goods consumption might be reported as low as zero. It is because the participants are only asked to keep the diaries for a few weeks and durable goods consumption might not have taken place at all during the diary-keeping period. However, when comparing the two approaches, it seems the recall method is the relatively better approach to collect information on durable goods consumption.

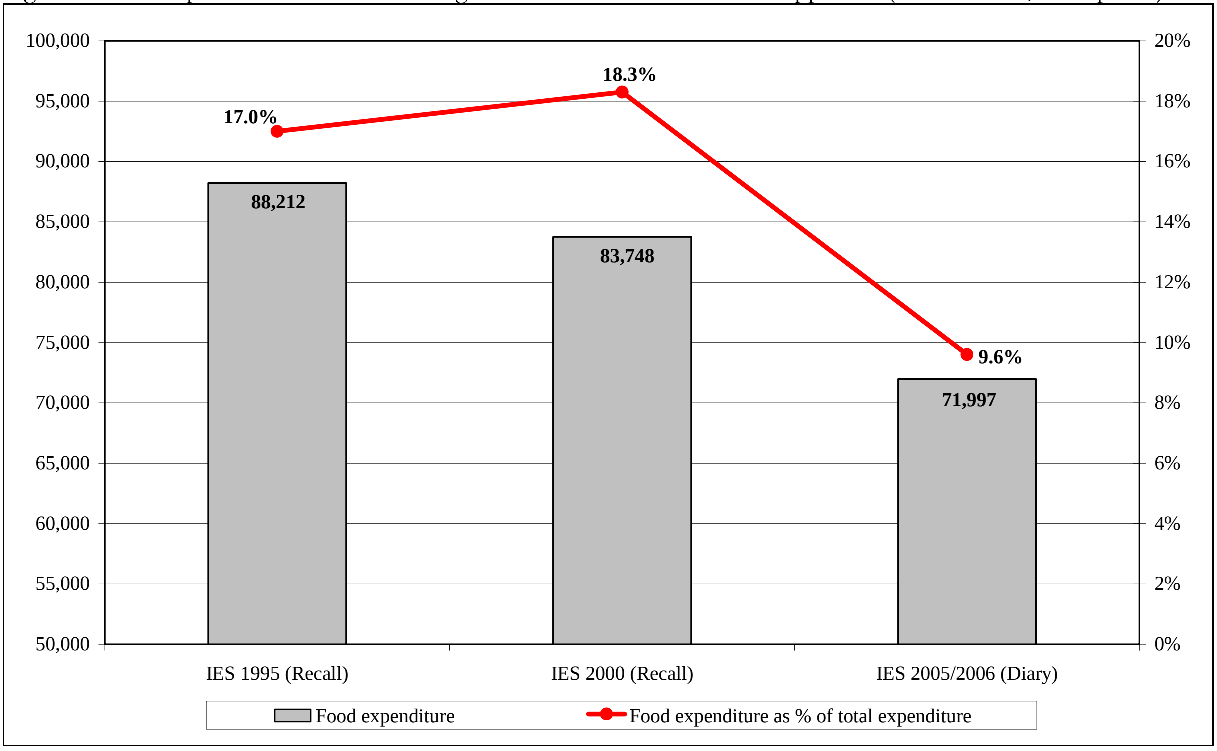
[17] In IES 2005/2006, Stats SA decided that only households that completed the main questionnaire and at least two weekly diaries were accepted. Missing acquisitions for households with two or three diaries were imputed, before household income and consumption were derived.

also need to visit the households frequently to examine the diary briefly, or to prompt the respondents to fill it out more completely if the diary appears to be incomplete. Consequently, the diary method could become more time-consuming to the interviewers compared with the recall method, might transform the situation back into an interview, and could even affect the motivation and competence of the interviewers due to reasons like fatigue (Corti 1993; Deaton and Grosh 2000: 119-122).

The diary method could be time-consuming and expensive (Sudman and Ferber 1971: 726; Corti 1993; Wiseman et al. 2005: 395): time is required to train the diary keepers and to maintain their support; intensive labour work is required to collect, edit and analyse the sheer volumes of data, especially if the diary is unstructured, since intensive editing and coding will push up the costs and involve even more time; respondents might be more co-operating and fill in the diaries more accurately only if offered incentives or gifts.

Figures 6 and 7 conclude the discussion on the diary and recall methods by presenting the information on food and transport expenditure in the three IESs[18]. It can be seen from Figure 6 that food expenditure was clearly lower in IES 2005/2006. This reported decline in food expenditure took place over a period when child hunger has been declining, according to GHS data.

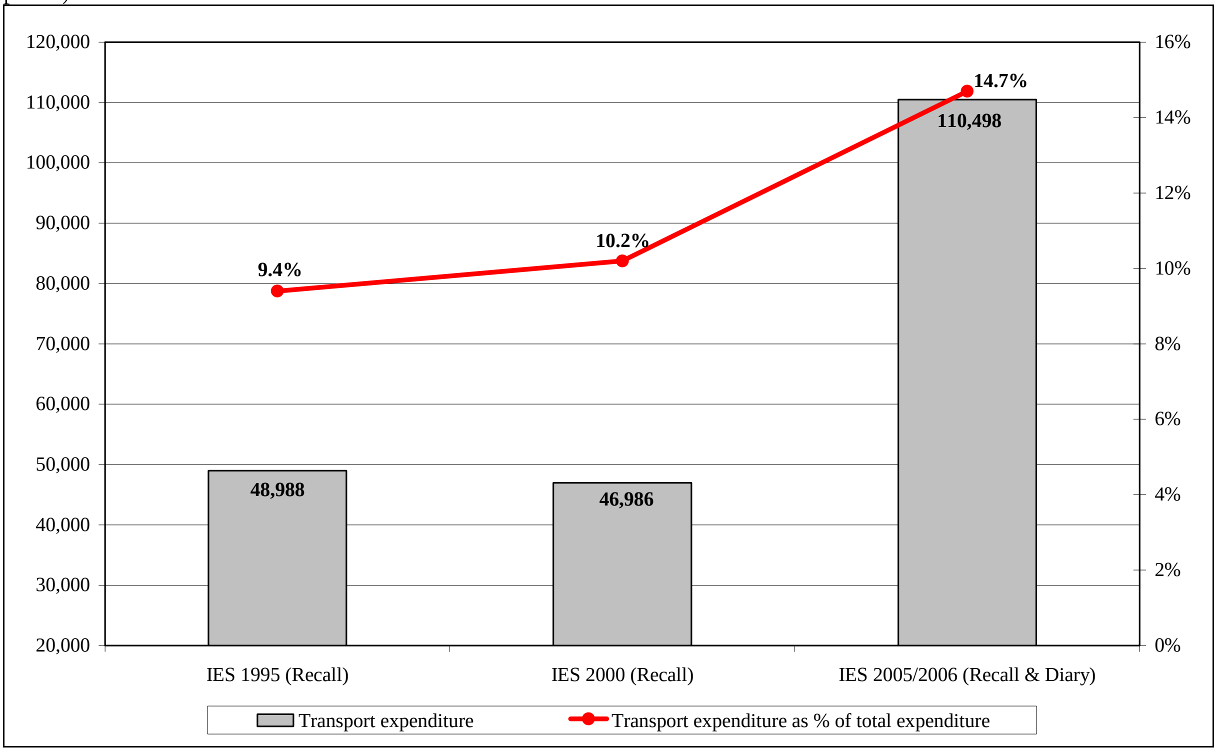
Figure 6: Food expenditure in the IESs using Standard Trade Classification approach (Rand million, 2000 prices)



Is it possible that the diary method resulted in the under-estimation of food expenditure in this survey (e.g., due to factors like first-day effect, illiteracy of respondents), or is it rather due to the fact that the recall method resulted in over-estimation of food expenditure in 1995 and 2000 (e.g., due to reasons like telescoping)? In contrast, the transport expenditure was much higher in IES 2005/2006, as shown in Figure 7. Is it possible that the use of the diary to complement the recall method resulted in a better capture of transport expenditure in this survey?

[18] Table A.4 in the Appendix provides more information by showing the total expenditure in all 20 expenditure categories, as well as total expenditure in each category as proportion of total expenditure, in the three IESs.

Figure 7: Transport expenditure in the IESs using Standard Trade Classification approach (Rand million, 2000 prices)



4.3 Actual amount vs. Bands

Participants in the surveys were asked to declare the exact income and expenditure (consumption) amounts in some surveys, or the relevant income and expenditure category in other surveys. An important question that arises is which method is more appropriate to collect the information better. Posel and Casale (2005: 10), Von Fintel (2006: 1) and Malherbe (2007: 25) argue that two major reasons the respondents did not declare the exact income amounts in the surveys are that they are reluctant to disclose such information due to confidentiality or privacy concerns, and that they really do not know exactly how much they or other members in the households earn and/or spend. As a result, this leads to a high proportion of households with unspecified income or consumption information and also possible bias in the data collected.

Hence, respondents, especially those in the higher income categories, might prefer the anonymity of indicating to what predefined interval (band) they belong. In fact, Posel and Casale (2005) found with regard to the information on income from the main job in the 2002 September LFS that bracket values instead of the actual amounts were more likely to be reported among those employed who are older, more educated, white, residing in urban areas, self-employed, informally employed and staying in larger households[19]. Von Fintel (2006) also found that people with higher earnings from the main job in the 2003 September LFS were more likely to report the relevant income category. Hence, the 'income bracket option' question should also be asked along with the 'exact income amount' question in the questionnaire in order to boost the response rate and obtain more reliable income or expenditure information (this is not the case in all surveys under study, except the income information in NIDS 2008).

Furthermore, this income band approach also permits respondents to report with a margin of error, especially if they really do not know the exact amounts earned. For example, if someone aged 35 years taking part in Census 1996 did not quite remember clearly that his/her nominal personal income was R4 450.75, but he/she still remembered that his/her income was

[19] Note that with regard to the question on income from the main job in the LFSs, the respondents were given two options to declare the income – either the exact amount or the relevant income category

somewhere between R4 400 and R4 500, then he/she would report his income to be under the "8: R3 501 – R4 500" interval. If he/she was only allowed the option to declare the exact amount, he/she might end up refusing to answer this question, which would eventually cause his/her household income to be unspecified. As a result, a significant greater response for income variables could be achieved and a better dataset with possibly more correct results created, if the interval approach is adopted.

A final problem of using the interval approach is that, as survey years progress, income brackets will invariably change with inflation. Alternatively, if the brackets are left unadjusted, a higher and higher proportion of households would fall in the higher categories due to the impact of inflation[20].

In Figure 8, the total income and expenditure amounts derived from these two approaches are compared, and there is no indication that a particular method resulted in a higher amount being captured. For instance, the 1995 IES total income (respondents were asked to declare the information in exact amounts) was greater than the 1995 AMPS total income (the interval method), while in 2008, the AMPS income (interval method) was greater than the NIDS income and expenditure (exact amount method), but the GHS expenditure (interval method) was the lowest.

Figure 8: Total income or expenditure (Rand million, 2000 prices) of selected surveys

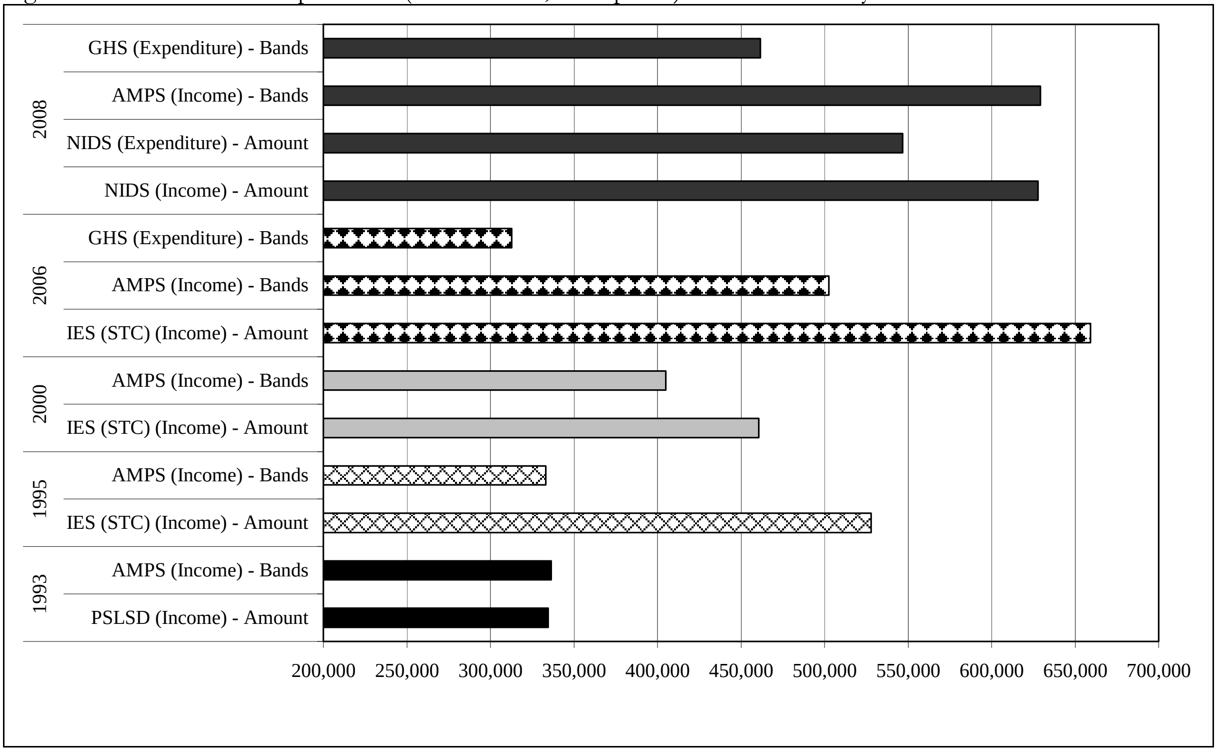
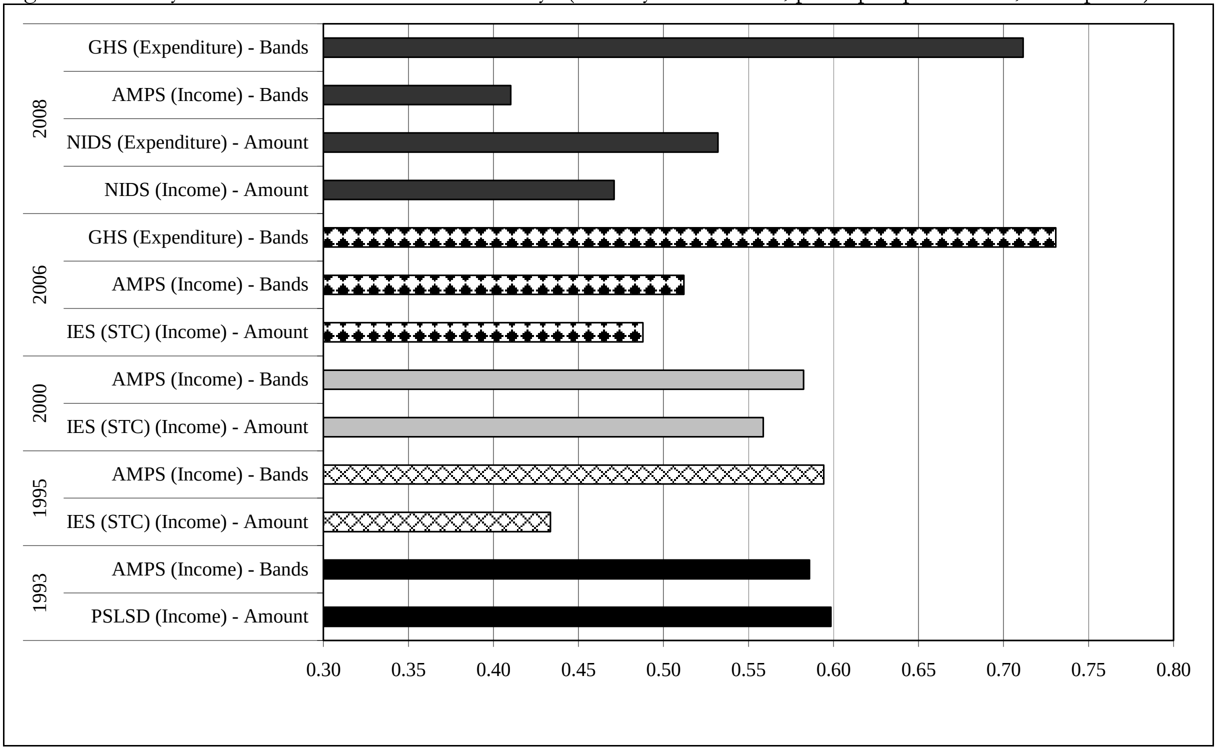


Figure 9 shows the poverty headcount ratios using the surveys as mentioned in Figure 8, and the results do not suggest that either method would lead to a lower poverty estimate. For example, in 1993, the PSLSD (exact amount) and AMPS (intervals) poverty headcount ratios were quite close (but the former ratio was slightly higher), while in 2006, the GHS (intervals) headcount ratio was much higher, followed by the AMPS (intervals) ratio, and finally the IES (exact amount) ratio.

[20] This explains why the nominal income brackets are adjusted between Census 1996 and Census 2001, as shown in Table A.1.

Figure 9: Poverty headcount ratios of selected surveys (Poverty line: R3 864, per capita per annum, 2000 prices)



4.4 Actual amount: One-shot overall amount vs. Aggregation of amounts from sub-items

If income and expenditure information is to be collected by asking the respondents to declare the exact amounts earned or spent, the next issue to decide is whether to ask the respondents to declare the 'one-shot', single estimate (by asking questions like "What is the total income you earned from all sources in the past 12 months?" and "How much do you spent on all items in the past month?") or to aggregate the amounts from sub-items (i.e., by asking questions like "How much do you earn from income source X?", "How much do you earn from income source Y?", and so forth, and then the total income is derived by adding the amounts from the answers of these questions).

The 'one-shot' amount, single estimate approach, despite being a relatively less time-consuming and costly method to collect the required information, could confuse the respondents, as they are unsure about what items should be included as part of the total income or expenditure they declare. This may result in low response rate, and/or under-reporting of total income or expenditure (Deaton 1997: 27; Browning et al. 2002: 7-10). Hence, there is a need to disaggregate to some extent so as to obtain more satisfactory estimates.

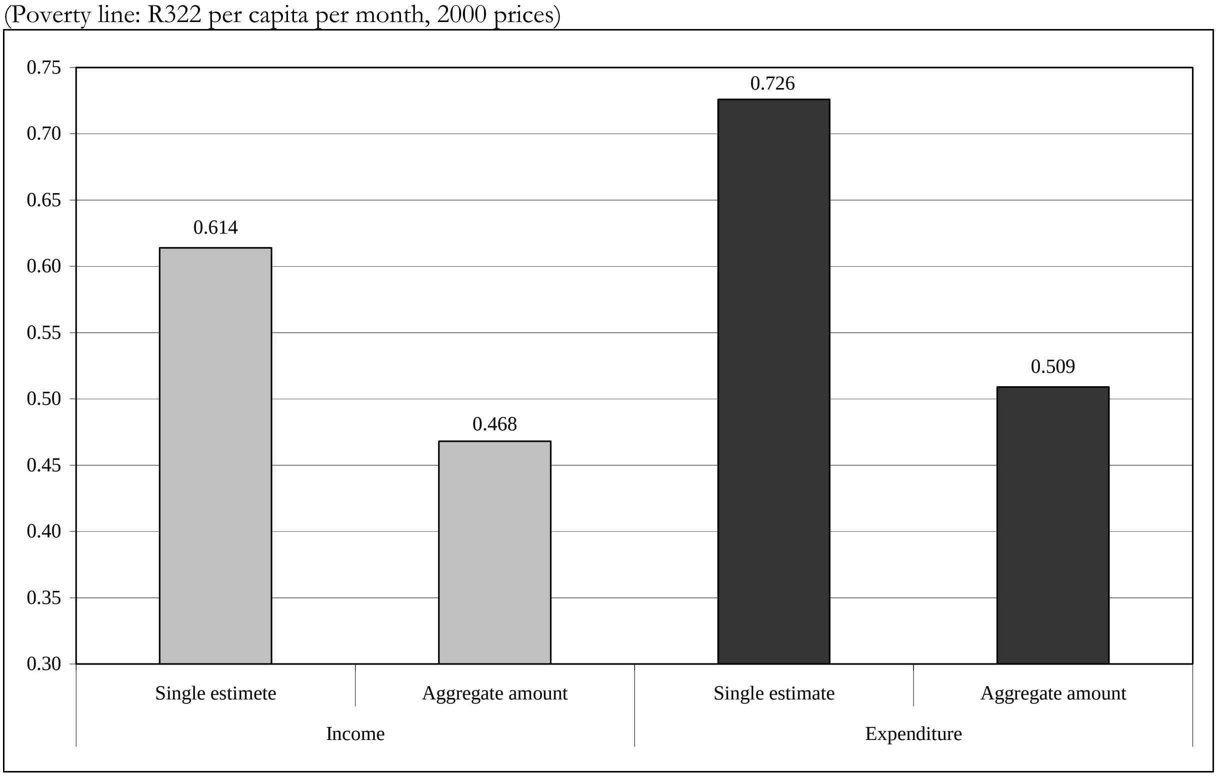
If a series of questions are asked on all of the sub-items in order to derive the overall income or expenditure amount, an issue to consider is the appropriate level of disaggregation. Deaton (2005: 16) claims that the greater the degree of disaggregation of the number of items that are separately distinguished, the more accurate is the measured consumption (expenditure) in total. However, Deaton (2005: 16) as well as Browning et al. (2002: 12-18) suggest that, if the level of disaggregation is too high, it could be very demanding, time-consuming and exhausting to both the interviewers and interviewees, and the latter might end up deliberately reporting inaccurate amounts or not answering some questions (i.e., item non-response). This eventually results in the derivation of an even more inaccurate aggregate consumption amount, compared with the single estimate method.

Browning et al. (2002: 19) also argue that, for non-durable items, a non-exhaustive list method should be more than enough to obtain reliable information on consumption (expenditure), e.g., the two questions "expenditure on food at home" and "expenditure on food outside home" should result in a pretty good predictor of total food expenditure. In contrast, for durable items, they suggest that the exhaustive method works better.

With regard to the derivation of the aggregate income, Davern et al. (2005: 1535) claim that the 'one-shot' amount approach might work better, as asking respondents to declare exact amounts earned from each income source could prove quite burdensome and intrusive for the respondents. This is because people generally do not like to divulge how much money they earn in too great detail, as a result of the questions' sensitive nature. In fact, some respondents already find it disturbing to reveal income or even consumption information even if asked to declare the 'one-shot' amount.

As mentioned before, NIDS 2008 is the only survey that collected the actual income and expenditure amounts by using both the single estimation and aggregation methods. As the former method seriously under-captured income and expenditure (see footnote 6), the poverty headcount ratios were subsequently higher (see Figure 10 below).

Figure 10: Poverty headcount ratios using per capita income and expenditure (2000 prices) variables of NIDS 2008 (Poverty line: R322 per capita per month, 2000 prices)



Note: Only households with specified income (expenditure) amounts under both single estimate and aggregation approaches were included.

4.5 Approximation of amount in each band

If the income or expenditure information was collected in bands, the data needs to be made continuous before dividing it by household size to derive the per capita income or expenditure variable required for poverty analyses. Hence, the income or expenditure amount of each band needs to be determined. This section discusses the commonly used approaches to deal with this problem.

4.5.1 Midpoint method

The midpoint method is simple and widely used. In this method, each household who supplies its income / expenditure bracket is assumed to earn / spend the category mean – its midpoint. For example, if a household taking part in AMPS 2009 declares its nominal monthly household income falls in the "R5 000 – R5 999" category, the income amount is derived as R5 500. Similarly, if a household participating in GHS 2009 claims its nominal monthly household expenditure falls in the "R5 000 – R9 999" category, then the expenditure amount is approximated as R7 500. As far as the top category is concerned, since no upper limit exists, it is often assumed that the mean exceeds the lower limit by 10% (Fields 1989). For instance, if the nominal monthly household income category of a household from the AMPS 2000 sample is "R20 000+", the income amount is assumed to R22 000 (R20 000 × 1.1).

Although this method lacks theoretical backing (Whiteford and McGrath 1994: 28), it may be attractive because of its simplicity. However, Seiver (1979: 230) is concerned that the true mean of any interval will always be below its midpoint, regardless of the number and width of the intervals, given intervals starting with "0"[21], as reported earnings or incomes tend to heap at levels ending in "0", or to a lesser extent, "5". For example, if the Census 1996 income categories were given as "R1 000 – R1 499", "R1 500 – R2 499" and so forth, then the people earning R1 500 would fall in the latter category, while the former category would be dominated by people earning R1 000. As a result, the true mean of the "R1 000 – R1 499" category would be smaller than its midpoint (R1 250). However, if the categories were given as "R1 001 – R1 500", "R1 501 – R2 500", etc., (i.e., ending in "0") like they were asked in Census 1996, the former category would probably be dominated by people stating they earned R1 500, and the true mean of this interval would exceed the midpoint (R1 250).

4.5.2 Midpoint-Pareto method

As the lower income / expenditure categories are narrow (as is the case in the surveys under study – see Tables A.1-A.3), Whiteford and McGrath (1994: 29) argue that the distribution of income at the bottom end is not noticeably influenced by midpoint imputation. However, as greater skewness within groups becomes evident in the higher income categories, a parametric approach is necessary there. A Pareto mean can be estimated for the open interval. This value could deviate from the midpoint, according to the heaviness of the tail (Von Fintel 2006: 15).

The Pareto mean (in the case of household income) is calculated as follows (Cloutier, 1988: 417; Gustavsson 2004: 20; Whiteford and McGrath 1994: 83):

- A Pareto function is fitted to the data by regressing $\log N$ against $\log Y$, i.e., $\log N = c + \alpha \log Y$, where Y stands for the lower limit of a household income interval and N represents the number of households with the household income above Y;
- Successive regressions are conducted each time eliminating the lowest income interval, until the highest coefficient of determination (R^2) is found, subject to the constraint that no less than the last three intervals before the open interval are used;
- The Pareto coefficient (α) from the chosen Pareto function is used in this equation to calculate the means of each of the bounded income intervals: $\bar{x} = \left[\frac{\alpha}{\alpha+1}\right] \cdot \left[\frac{x_1^{\alpha+1} - x_2^{\alpha+1}}{x_1^{\alpha} - x_2^{\alpha}}\right]$, where x_1 and x_2 are the upper and lower bounds of the interval;
- The Pareto coefficient is also used to calculate the mean of the open interval. That is, $\bar{x} = \left[\frac{\alpha}{\alpha+1}\right] \cdot x_\infty$, where x_∞ represents the lower limit of the open interval.

The midpoint-Pareto method is applied in the categorical data in either of the following ways: (1) the midpoint is used for all categories except the open category, while the Pareto method is applied to derive the Pareto mean for the latter category; (2) the midpoint is used for categories

[21] This is the case in the OHSs/LFSs, GHSs and AMPSs.

up to and including the category containing the population median income, and the Pareto mean is used for categories above the aforementioned category.

In the South African studies, method (1) discussed above is the commonly used approach. Whiteford and McGrath applied method (2) on the Census 1991 household income data. The Pareto equation from the regression of 10 observations was: $\log N = 14.04 - 1.938 \log Y$, where -1.938[22] was the Pareto coefficient. Table 3 presents their results and the means of each interval had method (1) and the simple midpoint method been applied[23].

Table 3: Applications of midpoint and midpoint-Pareto methods on Census 1991

	Mean of each interval		
Nominal monthly household income	**Midpoint method**	**Midpoint-Pareto method (1)** #	**Midpoint-Pareto method (2)** ##
1: No income	R0	R0	R0
2: R1 – R499	R250	R250	R250
3: R500 – R699	R600	R600	R600
4: R700 – R999	R850	R850	R850
5: R1 000 – R1 499	R1 250	R1 250	R1 250
6: R1 500 – R1 999	R1 750	R1 750	R1 750
7: R2 000 – R2 999	R2 500	R2 500	R2 500
8: R3 000 – R4 999	R4 000	R4 000	R4 000
9: R5 000 – R6 999	R6 000	R6 000	R6 000
10: R7 000 – R9 999	R8 500	R8 500	R8 500
11: R10 000 – R14 999	R12 500	R12 500	R12 500
12: R15 000 – R19 999	R17 500	R17 500	R17 106
13: R20 000 – R29 999	R25 000	R25 000	R23 899
14: R30 000 – R49 999	R40 000	R40 000	R37 253
15: R50 000 – R69 999	R60 000	R60 000	R58 163
16: R70 000 – R99 999	R85 000	R85 000	R82 083
17: R100 000 – R149 999	R125 000	R125 000	R119 495
18: R150 000 – R199 999	R175 000	R175 000	R171 061
19: R200 000 – R299 999	R250 000	R250 000	R238 990###
20: R300 000 – R499 999	R400 000	R400 000	R372 531###
21: R500 000 or above	R550 000	R880 193	R880 193###

Source: Whiteford and McGrath (1994: 84).

Method 1: Midpoint is used for all categories except the open category, while the Pareto method is applied to derive the Pareto mean for the latter category.

## Method 2: Midpoint is used for categories up to and including the category containing the population median income, and the Pareto mean is used for categories above this category.

### The mean of the R200 000 – R299 999 interval = $\left[\frac{-2.3151}{-1.3151}\right]\cdot\left[\frac{299999^{-1.3151}-200000^{-1.3151}}{299999^{-2.3151}-200000^{-2.3151}}\right] =$ R238 990,

while the mean of the R300 000 – R499 999 interval = $\left[\frac{-2.3151}{-1.3151}\right]\cdot\left[\frac{499999^{-1.3151}-300000^{-1.3151}}{499999^{-2.3151}-300000^{-2.3151}}\right] =$ R372 531.

The Pareto mean of the open interval is derived as: $\left[\frac{-2.3151}{-1.3151}\right]\cdot 500000 =$ R880 193.

[22] However, it is suspected that the Whiteford and McGrath (1994) did not use this coefficient to derive the Pareto mean. In fact, the coefficient should be -2.3151 (instead of -1.938).

[23] From the last column of Table 3, the Pareto mean was derived from the category '11: R10 000 – R14 999' onwards. However, it is unlikely that the 1991 median monthly household income fell in this range. For instance, the median monthly household income in Census 1996, Census 2001 and CS 2007 were about R16 000, R13 000 and R19 000 respectively (in 2000 prices). Using these three amounts, the median income in 2000 prices in Census 1991 ranged between R6 364 and R9 300, i.e., falling in either the '9: R5 000 – R6 999' or '10: R7 000 – R9 999' categories. Thus, it is not sure if Whiteford and McGrath (1994) derived the Pareto mean from the interval *containing the median income* or rather *from the median income interval onwards.*

4.5.3 Interval regression

Interval regression predicts the income/expenditure amount from some well chosen explanatory variables, such as educational attainment, age, gender, race, labour market status of household head, household size, and number of employed members in the household. The lower and upper limits of each income or expenditure category (with the exception of the open interval – there is no upper limit) must be specified in the interval regression, before the model could predict what income/expenditure each household earns/spends based on the explanatory variables used.

4.5.4 Random midpoint method

This method uses the midpoint of an income / expenditure interval and then distributes the households falling within the income / expenditure level randomly across the interval. If f_i stands for the frequency of households falling within income level i and x_i represents the midpoint of income level i, the following model is applied to obtain the random midpoint dataset (Malherbe 2007: 37): $Y_{ij} = x_i + \text{sign}_{ij} + U_{ij}(0, x_i - \text{lower limit}_i)$, where Y_{ij} is the new random midpoint income value for income level i and household j, j = 1, 2, …, f_i, while sign_{ij} is the sign for income level i and household j, where sign_{ij} has a 50% chance of being +1 and 50% chance of being –1. U_{ij} is the uniform distribution, with lower limits of 0 and upper limit of (x_i – lower limit_i), with lower limit_i representing the lower limit of income level i.

For example, if a household fell in the "R400 – R799" monthly household expenditure category in GHS 2008, the midpoint (i.e., x_2) is R600, while the lower limit of this interval (i.e., limit limit_2) is R400. Assuming sign_{ij} is –1 for this household, and a random draw from the uniform distribution (lower limit and upper limit being 0 and 200 respectively) gives an amount of R50, then the estimated household expenditure amount is derived as: 600 + (-1)×50 = R550. Similarly, using the same information but if sign_{ij} is +1 for this household, the household expenditure is calculated as: 600 + (+1)×50 = R650.

4.5.5 Equal distribution method

This method assumes that income recipients are equally distributed within each category. For example, if 400 households fell in the "2: R400 – R799" monthly household expenditure category in GHS 2008, the first randomly chosen household from this interval is assumed to have monthly expenditure of R400, the second and third randomly chosen households are assumed to have monthly expenditure of R401 and R402, and so forth, and the 400-th and the last randomly chosen household is supposed to spend R799. However, the method is cumbersome since it generates a huge number of records (Whiteford and McGrath 1994: 30), as the width of the interval and the number of households falling in the interval increase.

4.5.6 Conclusion

Having discussed the various methods to derive the income / expenditure mean of each interval, the comparability of the results of these methods, as well as the quality of the data captured in the actual amount method and the interval method are considered. In South Africa, Von Fintel (2007), who looked at the 2003 September LFS data on earnings from the main job and applied various methods (midpoint method, mid-point Pareto method, interval regression and lognormal distribution) to make the categorical earnings data continuous, found that coefficients of the Mincerian earnings regressions were LARGELY invariant to the methods used. His study did not investigate the impact of each method on poverty estimates. In contrast, Malherbe (2007) applied the Census 2001 income intervals to the IES 2000 data, and separately applied the midpoint method, interval regressions method and random midpoint method to derive the amount in each category. Malherbe found that the poverty estimates were very similar for the continuous and midpoint data, while the interval regressions and random midpoint method provided different results. The interval regression data under-estimated poverty, while the results obtained from the random midpoint data were not usable and the data were eventually rejected by Malherbe.

4.6 Number of bands and width of each band

If the respondents in a survey report their income or expenditure by declaring the relevant category, one might be concerned that the results of the poverty estimates would be heavily influenced by the number and width of the income / expenditure bands of the survey concerned. From Table 4, it could be seen the number of bands is as few as eight in the OHSs/LFSs/GHSs but as many as 32 in the AMPSs. The width of the bands ranges from R100 (e.g., in AMPS 2009) to R102 400 (e.g., in Census 2001 and CS 2007).

Table 4: Number and width of income and expenditure bands in selected surveys

Census 1996 – Income	**Width**	**AMPS 2009 – Income**	**Width**
R1 – R200	200	R1 – R499	500
R201 – R500	300	R500 – R599	100
R501 – R1 000	500	R600 – R699	100
R1 001 – R1 500	500	R700 – R799	100
R1 501 – R2 500	1 000	R800 – R899	100
R2 501 – R3 500	1 000	R900 – R999	100
R3 501 – R4 500	1 000	R1 000 – R1 099	100
R4 501 – R6 000	1 500	R1 100 – R1 199	100
R6 001 – R8 000	2 000	R1 200 – R1 399	200
R8 001 – R11 000	3 000	R1 400 – R1 599	200
R11 001 – R16 000	5 000	R1 600 – R1 999	400
R16 001 – R30 000	14 000	R2 000 – R2 499	500
Census 2001 & CS 2007 – Income	**Width**	R2 500 – R2 999	500
R1 – R400	400	R3 000 – R3 999	1 000
R401 – R800	400	R4 000 – R4 999	1 000
R801 – R1 600	800	R5 000 – R5 999	1 000
R1 601 – R3 200	1 600	R6 000 – R6 999	1 000
R3 201 – R6 400	3 200	R7 000 – R7 999	1 000
R6 401 – R12 800	6 400	R8 000 – R8 999	1 000
R12 801 – R25 600	12 800	R9 000 – R9 999	1 000
R25 601 – R51 200	25 600	R10 000 – R10 999	1 000
R51 201 – R102 400	51 200	R11 000 – R11 999	1 000
R102 401 – R204 800	102 400	R12 000 – R13 999	2 000
OHSs/LFSs/GHSs – Expenditure	**Width**	R14 000 – R15 999	2 000
R0 – R399	400	R16 000 – R19 999	4 000
R400 – R799	400	R20 000 – R24 999	5 000
R800 – R1 199	400	R25 000 – R29 999	5 000
R1 200 – R1 799	600	R30 000 – R39 999	10 000
R1 800 – R2 499	700	R40 000 – R49 999	10 000
R2 500 – R4 999	2 500		
R5 000 – R9 999	5 000		

For instance, if a household's exact monthly income and expenditure are both R8 200 in nominal terms, this household would fall in the 'R6 401 – R12 800' in CS 2007 (12 categories), 'R5 000 – R9 999' in GHS 2009 (10 categories) and 'R8 000 – R8 999' in AMPS 2009 (30 categories), and the derived income or expenditure amount (assuming the Pareto method is applied to the open interval and the midpoint method is applied to the other categories) would be estimated as R9 600, R7 500 and R8 500 respectively. In this case the AMPS amount (R8 500) is closest to the original amount (R8 200). The following questions arise: is the reliability of the derived amount being influenced by the number and width of bands in each survey? Would the poverty estimates be over-estimated or under-estimated as a result of these two factors?

There are no South African studies done to investigate the impact of the aforementioned issues on poverty estimates. However, Figure 11 shows that the total expenditure in OHS/LFSs and GHSs (with fewer intervals) was clearly lower. On the other hand, the total income in Census

1996 and 2001 (with very wide intervals in the higher-income categories) was lower than total income in AMPS 1996 and 2001 (AMPS is associated with more income categories and narrow width in each category), but the opposite happened when comparing CS 2007 with AMPS 2007. Furthermore, Figure 12 presents the poverty headcount ratios using these surveys that adopted the interval approach, and it can be seen that the ratios were always higher in OHS/LFSs and GHSs. These results suggest that fewer intervals could be associated with under-capturing of income / expenditure and over-estimation of poverty. Research needs to be done in South Africa to examine the impact of the number of bands and width of each band on poverty estimates.

Figure 11: Total income or expenditure (Rand million, 2000 prices) of surveys that adopted the interval method

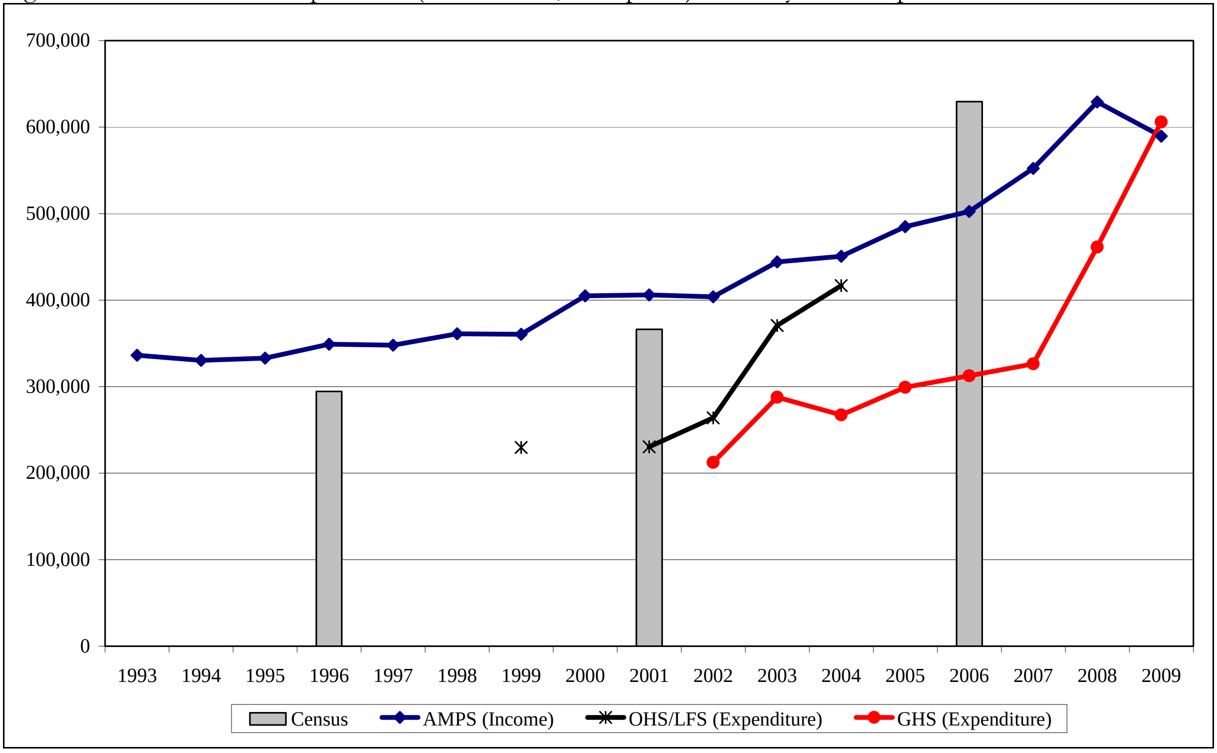
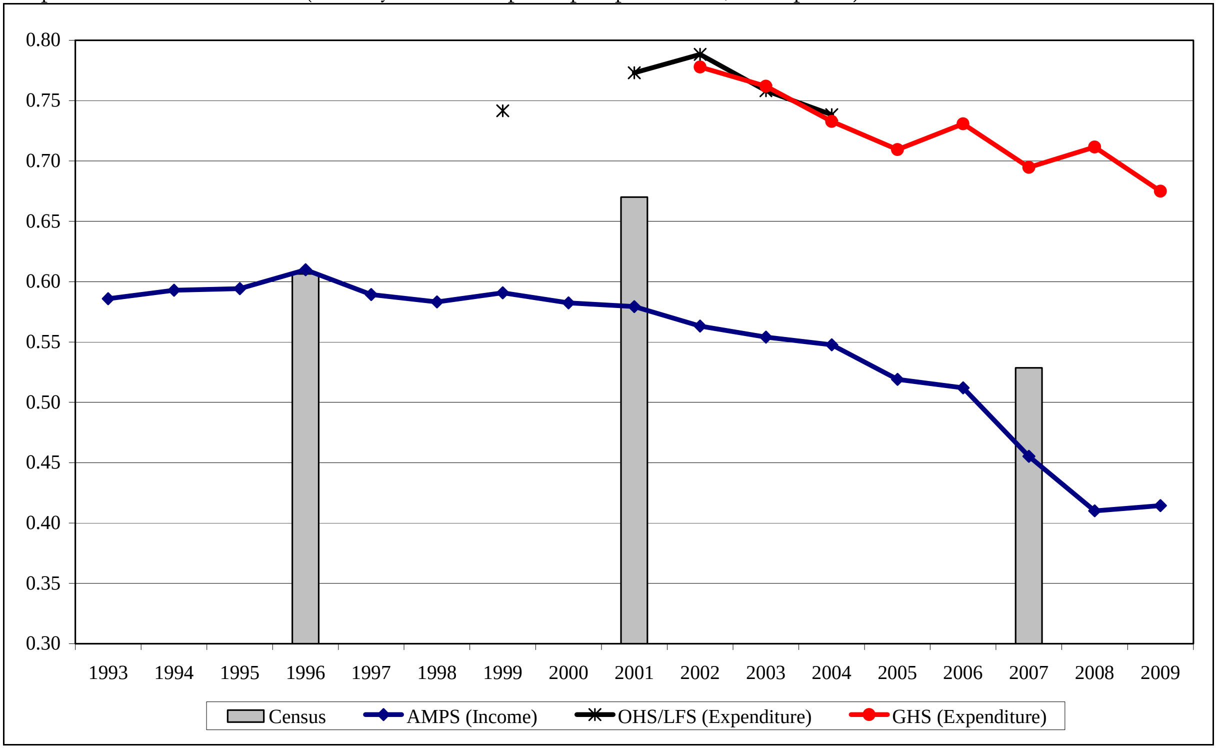


Figure 12: Poverty headcount ratios using per capita income and expenditure (2000 prices) variables of surveys that adopted the interval method (Poverty line: R322 per capita per month, 2000 prices)



Looking at international studies, Seiver (1979) found that income distribution results are influenced by the number and width of intervals chosen to span the range: fewer, wider brackets result in over-estimation of inequality measures. He did not investigate their impact on poverty.

Due to the lack of local and international studies investigating the impact of the number of bands and the width of each band on poverty estimates, the aim of the remainder of Section 4.6 is to apply the intervals from various surveys on the IES income data. First, some AMPS 2000 intervals are collapsed in order to investigate if these estimates would change significantly. Next, the Census 1996, Census 2001, AMPS 2000 and GHS 2009 intervals are applied in all three IESs to study the poverty and inequality trends across the three surveys. Note that if the income brackets are left unadjusted, a greater proportion of households would fall in the high-income categories in the more recent surveys due to the impact of inflation. Hence, the intervals above will be adjusted to 2000 prices before they are applied in all three IESs. Also, the midpoint-Pareto method (2) as discussed in Section 4.5.2 is used to derive the income amount in each interval.

First, the AMPS intervals are collapsed as follows (Table 5):

- Some of the low-income intervals are collapsed together so that these intervals become wider. There are 25 intervals in total after collapsing.
- Some of the high-income intervals are collapsed together so that these intervals become wider. There are 21 intervals in total after collapsing.
- Both the low-income and high-income intervals are collapsed. There are 11 intervals in total after collapsing, and this method results in the widest intervals.

Table 5: Collapsing selected AMPS 2000 intervals

Original intervals	**Collapsed intervals (1)**	**Collapsed intervals (2)**	**Collapsed intervals (3)**
R1-R199	R1-R199	R1-R199	R1-R499
R200-R299	R200-R399	R200-R299	
R300-R399		R300-R399	
R400-R499	R400-R599	R400-R499	
R500-R599		R500-R599	R500-R999
R600-R699	R600-R799	R600-R699	
R700-R799		R700-R799	
R800-R899	R800-R999	R800-R899	
R900-R999		R900-R999	
R1 000-R1 099	R1 000-R1 199	R1 000-R1 199	R1 000-R1 999
R1 100-R1 199			
R1 200-R1 399	R1 200-R1 399	R1 200-R1 399	
R1 400-R1 599	R1 400-R1 599	R1 400-R1 599	
R1 600-R1 999	R1 600-R1 999	R1 600-R1 999	
R2 000-R2 499	R2 000-R2 499	R2 000-R3 999	R2 000-R3 999
R2 500-R2 999	R2 500-R2 999		
R3 000-R3 999	R3 000-R3 999		
R4 000-R4 999	R4 000-R4 999	R4 000-R5 999	R4 000-R5 999
R5 000-R5 999	R5 000-R5 999		
R6 000-R6 999	R6 000-R6 999	R6 000-R7 999	R6 000-R7 999
R7 000-R7 999	R7 000-R7 999		
R8 000-R8 999	R8 000-R8 999	R8 000-R9 999	R8 000-R9 999
R9 000-R9 999	R9 000-R9 999		
R10 000-R10 999	R10 000-R10 999	R10 000-R11 999	R10 000-R11 999
R11 000-R11 999	R11 000-R11 999		
R12 000-R13 999	R12 000-R13 999	R12 000-R15 999	R12 000-R15 999
R14 000-R15 999	R14 000-R15 999		
R16 000-R17 999	R16 000-R17 999	R16 000-R19 999	R16 000-R19 999
R18 000-R19 999	R18 000-R19 999		
R20 000+	R20 000+	R20 000+	R20 000+

The Foster-Greer-Thorbecke (FGT) poverty indices are reported in Table 6. First, the poverty indices at all three poverty lines by collapsing the low-income intervals (method 1) are almost the same as the results obtained by using the original AMPS intervals. However, the indices only decreased after collapsing the high-income intervals (method 2) and declined slightly further after collapsing both low-income and high-income intervals (method 3). Thus, it seems poverty estimates only showed negligible changes after the application of fewer and wider intervals.

Table 6: FGT poverty estimates and Gini coefficients, after applying the AMPS income intervals on the IES 2000 income (STC approach) data, and collapsing some intervals

	FGT poverty index		
	P_0	**P_1**	**P_2**
Poverty line: R211 per month per annum (2000 prices)			
The actual continuous income variable	0.429	0.206	0.127
AMPS intervals	0.422	0.202	0.123
AMPS collapsed intervals (1)	0.411	0.204	0.124
AMPS collapsed intervals (2)	0.414	0.201	0.123
AMPS collapsed intervals (3)	0.409	0.186	0.111
Poverty line: R322 per month per annum (2000 prices)			
The actual continuous income variable	0.559	0.307	0.204
AMPS intervals	0.562	0.303	0.200
AMPS collapsed intervals (1)	0.562	0.305	0.202
AMPS collapsed intervals (2)	0.547	0.300	0.199
AMPS collapsed intervals (3)	0.534	0.286	0.185
Poverty line: R593 per month per annum (2000 prices)			
The actual continuous income variable	0.710	0.462	0.342
AMPS intervals	0.713	0.458	0.339
AMPS collapsed intervals (1)	0.713	0.459	0.340
AMPS collapsed intervals (2)	0.713	0.452	0.334
AMPS collapsed intervals (3)	0.703	0.443	0.323

AMPS collapsed intervals (1): The low-income intervals are collapsed together
AMPS collapsed intervals (2): The high-income intervals are collapsed together
AMPS collapsed intervals (3): Both the low-income and high-income intervals are collapsed together
Note: P_0: Poverty headcount ratio
P_1: Poverty gap ratio
P_2: Squared poverty gap ratio

Next, in addition to the AMPS 2000 intervals, the following intervals are also applied on the IES 2000 income data, before investigating the poverty and inequality estimates: Census 1996 intervals, Census 2001 intervals, GHS 2009 intervals, equal R500 intervals, equal R1 000 intervals and equal R2 000 intervals. For the latter three approaches, the open interval is "R20 000+". However, due to the impact of inflation, the nominal intervals of the two censuses and GHS 2009 are converted to intervals in 2000 prices (See Table 7), before deriving the poverty estimates obtained from the application of these intervals. Furthermore, once again the midpoint method was applied on all intervals, except that Pareto method was used for the open interval, in order to make each dataset continuous again.

Table 7: Adjusting the Census 1996, Census 2001 and GHS 2009 nominal intervals into 2000 prices intervals

Nominal intervals	Real intervals
Census 1996	
1: None	1: None
2: R1 – R200	2: R1 – R251
3: R201 – R500	3: R251 – R627
4: R501 – R1 000	4: R627 – R1 254
5: R1 001 – R1 500	5: R1 254 – R1 880
6: R1 501 – R2 500	6: R1 880 – R3 134
7: R2 501 – R3 500	7: R3 134 – R4 387
8: R3 501 – R4 500	8: R4 387 – R5 641
9: R4 501 – R6 000	9: R5 641 – R7 521
10: R6 001 – R8 000	10: R7 521 – R10 028
11: R8 001 – R11 000	11: R10 028 – R13 788
12: R11 001 – R16 000	12: R13 788 – R20 055
13: R16 001 – R30 000	13: R20 055 – R37 603
14: R30 001 or more	14: R37 603+
Census 2001	
1: None	1: None
2: R1 – R400	2: R1 – R377
3: R401 – R800	3: R377 – R754
4: R801 – R1 600	4: R754 – R1 509
5: R1 601 – R3 200	5: R1 509 – R3 017
6: R3 201 – R6 400	6: R3 017 – R6 035
7: R6 401 – R12 800	7: R6 035 – R12 070
8: R12 801 – R25 600	8: R12 070 – R24 140
9: R25 601 – R51 200	9: R24 140 – R48 279
10: R51 201 – R102 400	10: R48 279 – R96 558
11: R102 401 – R204 800	11: R96 588 – R193 117
12: R204 801 or more	12: R193 117+
GHS 2009	
1: R0	1: R0
2: R1 – R199	2: R1 – R115
3: R200 – R399	3: R115 – R231
4: R400 – R799	4: R231 – R461
5: R800 – R1 199	5: R461 – R692
6: R1 200 – R1 799	6: R692 – R1 038
7: R1 800 – R2 499	7: R1 038 – R1 442
8: R2 500 – R4 999	8: R1 442 – R2 884
9: R5 000 – R9 999	9: R2 884 – R5 767
10: R10 000 or more	10: R5 767+

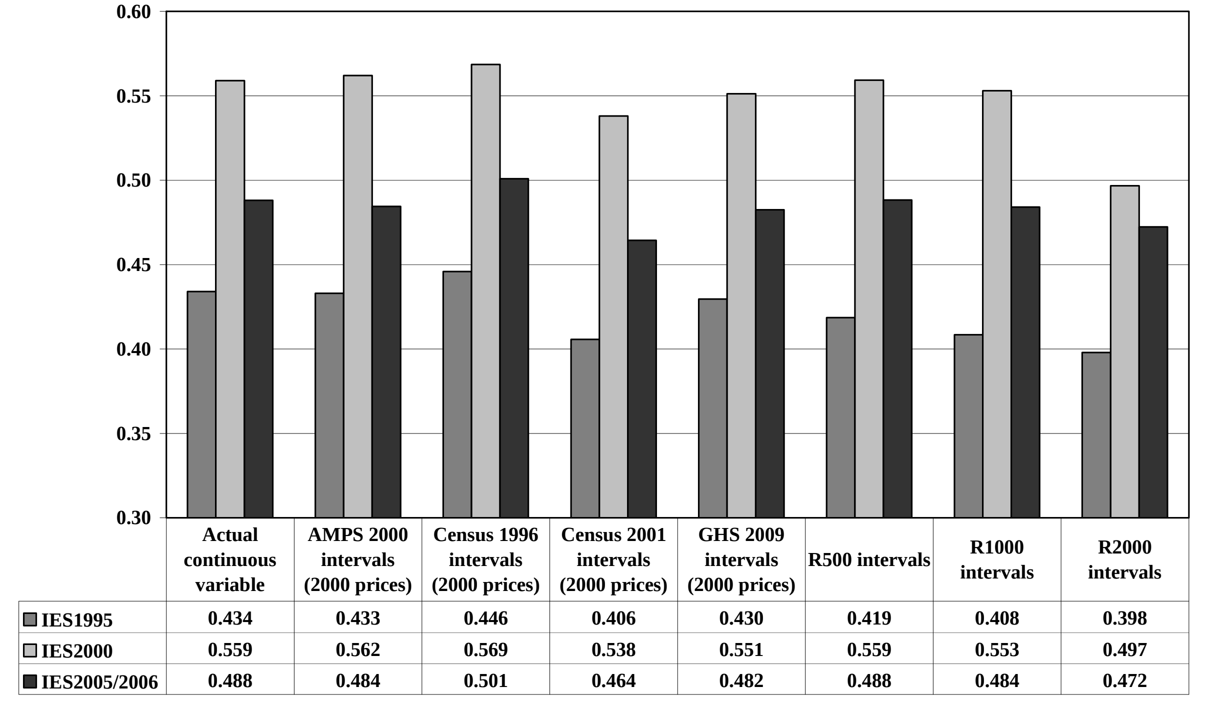
Table 8 reports the poverty headcount ratios at the three poverty lines by applying the different intervals to the STC income variable of the three IESs. It can be seen that the FGT poverty indices at all three poverty lines when applying the AMPS 2000 and the R1 000 intervals are closest to those obtained by using the IES actual continuous income variable. Furthermore, poverty is clearly lower if the R2 000 intervals are applied. This could be explained by the fact that these R2 000 intervals are much wider at the lower end of the distribution (e.g., “R0 – R1 999”, “R2 000 – R3 999”, etc.). Hence, the income of the poor households could be over-estimated, which eventually causes the under-estimation of poverty.

Table 8: FGT poverty indices, after applying various intervals on the IES 2000 income (STC approach) data

		FGT poverty index		
		P_0	P_1	P_2
Poverty line: R211 per month per annum (2000 prices)				
The actual continuous income variable		0.429	0.206	0.127
Applying the intervals on the income data	AMPS 2000 intervals (2000 prices)	0.422	0.202	0.123
	Census 1996 intervals (2000 prices)	0.417	0.207	0.129
	Census 2001 intervals (2000 prices)	0.412	0.198	0.123
	GHS 2009 intervals (2000 prices)	0.411	0.198	0.123
	R500 intervals	0.416	0.192	0.114
	R1 000 intervals	0.426	0.199	0.116
	R2 000 intervals	0.391	0.127	0.055
Poverty line: R322 per month per annum (2000 prices)				
The actual continuous income variable		0.559	0.307	0.204
Applying the intervals on the income data	AMPS 2000 intervals (2000 prices)	0.562	0.303	0.200
	Census 1996 intervals (2000 prices)	0.569	0.306	0.205
	Census 2001 intervals (2000 prices)	0.538	0.297	0.197
	GHS 2009 intervals (2000 prices)	0.551	0.295	0.197
	R500 intervals	0.559	0.296	0.192
	R1 000 intervals	0.553	0.300	0.195
	R2 000 intervals	0.497	0.241	0.133
Poverty line: R593 per month per annum (2000 prices)				
The actual continuous income variable		0.710	0.462	0.342
Applying the intervals on the income data	AMPS 2000 intervals (2000 prices)	0.713	0.458	0.339
	Census 1996 intervals (2000 prices)	0.705	0.459	0.340
	Census 2001 intervals (2000 prices)	0.717	0.448	0.331
	GHS 2009 intervals (2000 prices)	0.695	0.448	0.331
	R500 intervals	0.701	0.454	0.333
	R1 000 intervals	0.709	0.455	0.334
	R2 000 intervals	0.706	0.417	0.284

Note: P_0: Poverty headcount ratio
P_1: Poverty gap ratio
P_2: Squared poverty gap ratio

Figure 13: Poverty headcount ratios, after applying various intervals on the three IESs (Poverty line: R322 per month in 2000 prices)



	Actual continuous variable	AMPS 2000 intervals (2000 prices)	Census 1996 intervals (2000 prices)	Census 2001 intervals (2000 prices)	GHS 2009 intervals (2000 prices)	R500 intervals	R1000 intervals	R2000 intervals
▨ IES1995	0.434	0.433	0.446	0.406	0.430	0.419	0.408	0.398
□ IES2000	0.559	0.562	0.569	0.538	0.551	0.559	0.553	0.497
■ IES2005/2006	0.488	0.484	0.501	0.464	0.482	0.488	0.484	0.472

The intervals mentioned above are now applied on the STC income variable of the other two IESs to investigate the poverty levels and trends across the three IESs. Figure 13 and Table A.5 in the Appendix present the FGT poverty indices across the three IESs. Focusing on the poverty headcount ratios at the R322 poverty line, it can be seen that the same poverty trends (i.e., rapid increase between IES 1995 and IES 2000, before a decline took place in IES 2005/2006, but the IES 2005/2006 poverty headcount ratio was still above the IES 1995 ratio) are observed, regardless of which intervals were applied. In addition, the poverty headcount ratios obtained by using the AMPS 2000, GHS 2009 and R500 intervals are closest to the results obtained by using the original continuous variable in all three surveys, but poverty was seriously under-estimated with the application of the Census 2001 and R2 000 intervals (fewer and wider intervals).

The results discussed above should be interpreted with caution, as it had to be assumed that the respondents, who declared their income or expenditure by aggregation approach in the IESs, would report similar income or expenditure if asked to report the 'one-shot' amount or in intervals. For instance, if the aggregated income of a respondent in IES is equal to R940, then it is assumed that he/she would report that his/her income falls in the 'R900 – R999' interval in AMPS, and then the midpoint method is applied by converting his/her categorical answer into an amount of R950 (which is quite close to the original actual amount of R940, and hence this would not have a significant impact on poverty and inequality estimates). However, the discussions in Sections 4.3 and 4.4 have shown that this might not always be the case. For example, it is possible that someone reporting his/her income falls in the 'R800 – R899' interval in AMPS would declare his aggregate income as R750 in the IES.

4.7 Households with zero or unspecified income

A serious problem in some surveys (especially the two censuses and CS 2007) is the high proportion of people reporting zero or unspecified personal income, which subsequently resulted in a large proportion of households with zero or unspecified household income (See Table 9).

Table 9: Proportion of households with zero or unspecified income or expenditure in each survey

		Zero	**Unspecified**
Census / CS (Income)	1996	13.0%	11.5%
	2001	21.0%	16.4%
	2007	8.2%	11.1%
OHS/LFS (Expenditure)	1996	0.0%	7.6%
	1997	0.0%	5.0%
	1998	0.0%	4.5%
	1999	0.0%	7.6%
	2001	0.0%	3.7%
	2002	0.0%	3.0%
	2003	0.0%	2.3%
	2004	0.0%	2.7%
GHS (Expenditure)	2002	0.0%	3.5%
	2003	0.0%	3.7%
	2004	0.0%	3.1%
	2005	0.0%	2.1%
	2006	0.0%	1.4%
	2007	0.0%	1.5%
	2008	0.0%	2.6%
	2009	0.5%	4.4%

Note: In the IESs, all households had specified income/consumption/expenditure in all three surveys. No households reported zero income/consumption/expenditure amounts in IES 1995, while only a very negligible proportion of households (less than 1% in each survey) had zero amounts in the other two IESs.

Note: In PSLSD 1993 and all AMPSs, all households had non-zero, specified income and expenditure.

Note: In NIDS 2008, all households had specified income and expenditure. No households reported zero expenditure, while only a negligible proportion of households (less than 1%) had zero income.

Regarding the people/households with missing personal/household income, Ardington et al. (2005) argue that if those with missing data fall excessively in the bottom of the income distribution, then poverty levels will be under-estimated if they are ignored. In contrast, if non-response is higher among the affluent, inequality measures are likely to be biased downwards[24]. Furthermore, with regard to the higher proportion of households with zero household income, even taking South Africa's high unemployment rates into consideration, it is highly unlikely that most of these zero-income households distinguished had no working-age members earning any income[25]. If these zero-income households are included for analyses, this could lead to an over-estimation of measured poverty.

There are three types of missing data (Lacerda et al., 2008: 6-9):

- Missing completely at random (MCAR): The distribution of missingness is independent of both the observed and missing data.
- Missing at random (MAR): The distribution of missingness is independent of missing data, but is dependent on some or all of the observed variables for each observational unit.
- Missing not at random (MNAR): The distribution of missingness is dependent on both the observed and missing data.

When examining poverty, unless the data are MCAR, ignoring households with unspecified household income would lead to biased results. Including households that might incorrectly report zero income might lead to over-estimation of poverty levels. In general, the four main methods to deal with missing data are casewise deletion, available-case deletion, single imputation and multiple imputation. Each method is discussed in greater detail.

4.7.1 Casewise deletion

Casewise deletion, also commonly known as listwise deletion or complete-case analysis, is the simplest method to deal with missing data. It discards any observational unit with incomplete information (Lacerda et al. 2008: 11). Thus, in the case of household income data (or expenditure / consumption), those households that did not specify the household income amount or category (depending on how the question was asked) are immediately excluded from further analyses. However, as mentioned at the beginning of this section, if these households are ignored, it would have a serious impact on the reliability of poverty estimates.

4.7.2 Available-case deletion

Available-case deletion is an extension of casewise deletion, but differs in that it only excludes those cases for which data are missing on the variables required to estimate the parameters of interest (Lacerda et al. 2008: 11). For example, if all households taking part in a survey reported dwelling type while 10% of households did not specify household income, but the latter variable is not used at all by a researcher in his/her analysis, then there is no need to worry about the missing income data, and all observations are kept in the dataset. However, if household income is an important variable for analysis (as in the case of this study), these 10% observations are immediately eliminated. However, excluding these households would have the same negative impact on poverty estimates as caused by casewise deletion. Thus, it seems the abovementioned two methods are not the best solution to deal with missing data for the purposes of this study.

4.7.3 Single imputation

Imputation aims to provide reasonable estimates of the missing data, instead of simply ignoring

[24] Yu (2009: 61) found that, among households with unspecified household income, 35.%, 29.4% and 47.6% contained at least one employed member, in Census 1996, Census 2001 and CS 2007 respectively. In addition, 27.9%, 22.2% and 37.1% of the heads of these households were employed at the time of each survey respectively. This implies that ignoring them would result in the over-estimation of poverty and narrowing of inequality.

[25] When looking at the households with zero income in Census 1996, Census 2001 and CS 2007, 1.8%, 1.5% and 5.5% of them were headed by an employed member in each survey, while 2.2%, 2.0% and 6.2% of these households had at least one employed member.

observations with missing data. If it is applied to impute one value for each missing item of a variable, this is known as single imputation (Lacerda et al. 2008: 13). The commonly used single imputation methods are unconditional mean substitution, cell mean substitution, hot deck imputation, cold deck imputation and stochastic regression imputation.

Unconditional mean substitution means that the missing values are replaced by the mean of the observed values for that variable (Lacerda et al. 2008: 15). For example, assuming household income information from a survey was collected as exact amounts, 90% of households declared their household income and the mean household income for these households was R1 500. The household income of the 10% of households with unspecified income would then be assumed to be R1 500.

Cell mean substitution aims to divide respondents into cells on the basis of some known variables, and the average values within these cells are used for imputation (Malherbe 2007: 29 & Lacerda et al. 2008: 15). For example, the mean household income for a household headed by each race and gender could be derived. To apply this mean, a household headed by a black male has a mean household income of R1 600, then a household with exactly the same race and gender characteristics but with unspecified household income would also be assumed to earn R1 600.

Hot deck imputation involves "substituting missing values with observed values drawn from similar responding units" (Lacerda et al. 2008: 16). For example, using the example above, households are divided into cells by race and gender of household head. After a random draw on a household headed by a white male, this household's income is R2 000. Then household A with unspecified household income but exactly the same race and gender characteristics has its household income imputed as R2 000. Similarly, after the second random draw on households from the same cell, a household with income level of R2 500 is chosen, and then household B with unspecified household income but the same race and gender characteristics has its household income imputed as R2 500. This process would carry on in each cell, until all missing household income data are imputed.

Cold deck imputation involves substituting missing values with a constant value from an external source (Lacerda at el. 2008: 16). For example, if a household headed by a black male taking part in IES 2000 did not answer the question "How much personal income tax did you pay the South African Revenue Service (SARS) in the last 12 months?", and from the National Treasury Budget Review 2000 document, it was found that, on average, a black male-headed household paid R1 500 personal income tax, then it would be assumed that the IES 2000 household as mentioned above spent R1 500 in the last 12 months to pay personal income tax to SARS.

Stochastic mean substitution is employed when "imputed values are randomly generated from a specified theoretical distribution with mean equivalent to the cell mean and variance equal to the cell variance" (Lacerda et al. 2008: 16). An extension to the above methods is known as stochastic regression imputation, in which "missing values are replaced by a value predicted by regression imputation plus a residual drawn to represent the uncertainty in the predicted value" (Lacerda et al. 2008: 17). For example, in the household income example above, in addition to race and gender of household head, other demographic characteristics such as the province of residence, age of household head, marital status of household head, as well as the number of children and elderly in the household should also be considered as explanatory variables to predict household income.

Finally, there are some less commonly used methods to deal with missing data. For example, the logical imputation method: A consistent value is estimated or deduced from other information relating to the individual or household, e.g., if two members from a household both declared they received old-age pension income in the last 12 months, but one of them stated he earned R1 500 from it while the other member did not specify his/her answer, then it is assumed that he/she also earned R1 500 from old-age pension during the same period. As another example, if both income and expenditure questions were asked in a household survey, but the respondent only declared the monthly household income as R10 000 but did not specify household expenditure, then one could impute the household expenditure as R10 000.

4.7.4 Multiple imputation

The multiple imputation method involves imputing several values for each missing item to allow for the inherent uncertainty in the imputation procedure. It consists of the following three steps (Lacerda et al. 2008: 17-18):

- *m* (which is greater than one – if X equals one, it stands for single imputation) plausible versions of the complete data are created by "imputing each missing value *m* times using *m* independent draws from an appropriate imputation model, conditional on the observed data" (Lacerda et al. 2008: 17);
- The *m* imputed datasets are then treated as if they are fully observed and analysed individually by standard complete-data methods;
- The results from the *m* analyses are combined in a single and proper manner so as to obtain overall estimates and standard errors that reflect both sample variation and uncertainty in association with the imputed values.

In this study, values for the households with unspecified personal or household income are imputed using a particular multiple imputation technique developed by Raghunathan, Lepkowski, Van Howeyk and Solenberger (2001), which is applied when data are missing at random (MAR), namely sequential regression multiple imputation (SRMI). The SRMI method could be summarized as follows (Raghunathan et al. 2001: 86-87; Ardington et al. 2005: 8-11; Ardington, Lam, Leibbrandt and Welch 2006: 826-827; Lacerda et al. 2008; Vermaak 2008: 2-3):

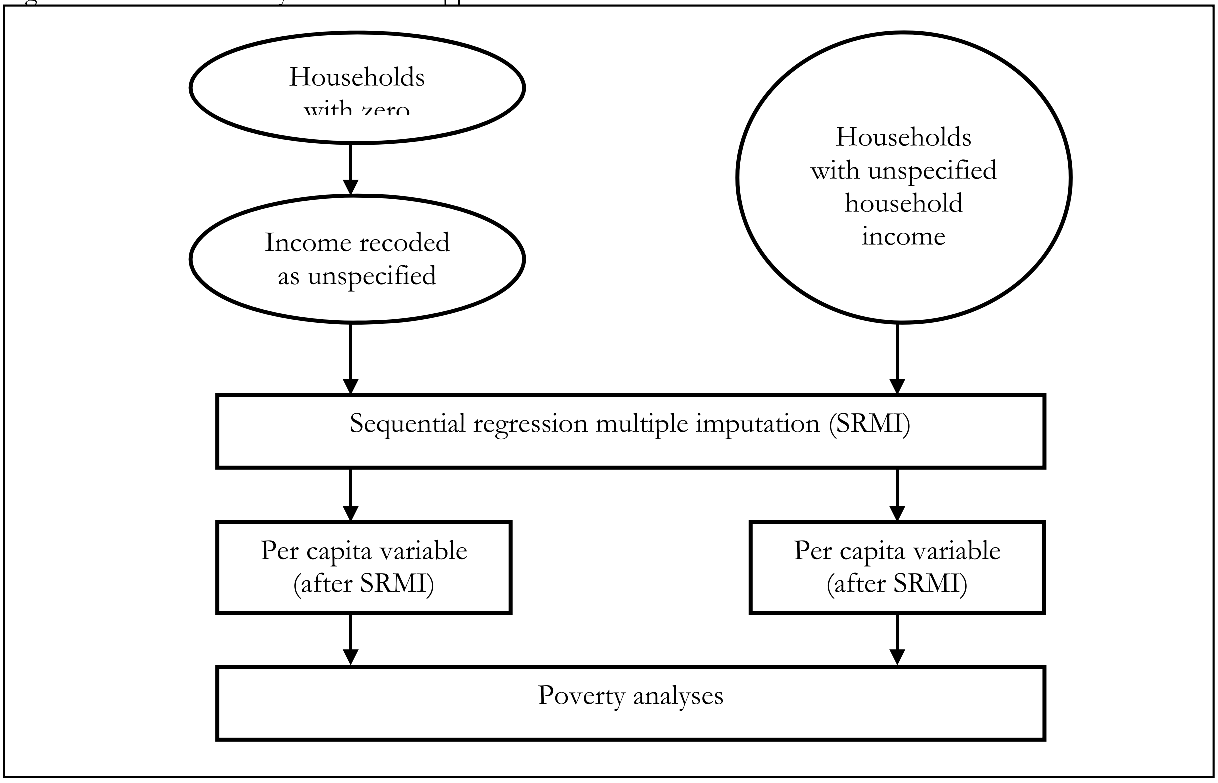
- The variables used in the imputation model are arranged from those with the least to those with the most missing values. The variables could be continuous (e.g., earnings amount), binary (e.g., gender), count (e.g., age), nominal categorical (e.g., province) or ordinal categorical (e.g., household income category).
- The matrix X represents all variables that are fully observed (i.e., there are no unspecified responses), while $Y_1, Y_2, ..., Y_k$ stand for the ordered variables that contain missing values. The variables are arranged according to the extent of missing data they contain, i.e., Y_1 and Y_k have the least and most missing values respectively.
- All missing values are imputed as part of a process to estimate the joint conditional density of $Y_1, Y_2, ..., Y_k$ given X. In other words, $f(Y_1, Y_2, ..., Y_k | X, \beta_1, \beta_2, ..., \beta_k)$ $= f_1(Y_1 | X, \beta_1) f_2(Y_2 | X, Y_1, \beta_2) ... f_k(Y_k | X, Y_1, Y_2, ..., Y_k, \beta_k)$, where f_i and β_i stands for the conditional density functions and a vector of parameters in the conditional distribution respectively. In all cases, the β_i vectors are the estimated coefficients and estimates of the disturbance term.
- The first round of the imputation starts with Y_1 regressed on X in order to obtain an estimate of the β_1 vector. The missing values in Y_1 are then imputed by random draws from the predictive distribution. In other words, by first drawing a vector β_1^* from the posterior distribution of β_1 and then using β_1^* to generate a set of predicted values to replace the missing Y_1 values. A normal OLS regression model is used when Y_1 is a continuous variable. However, a Poisson model is used when Y_1 is a count variable, a logistic model is used when Y_1 is binary, a multinomial logistic model is used when Y_1 is a nominal categorical variable, and an ordered logistic model is used when Y_1 is an ordinal categorical variable.
- Since its missing values have now been imputed, Y_1 is appended to the set of predictor variables. Next, Y_2 is regressed on X and the newly derived Y_1 that includes the imputed

values. The values are then imputed for Y_2. This imputation goes on until all Y variables have been imputed using non-missing variables (X) and all previously imputed variables of Y as covariates, before the first round is completed. At the end, the first complete set of data with no missing values is available.

- The imputation process is then repeated in the second round, updating the regression parameters β_i with parameters drawn from the now-complete distribution. That is, regress Y_1 on X and Y_2, Y_3, …, Y_k; regress Y_2 on X and Y_1, Y_3, …, Y_k; and so on. This cycle is repeated for a pre-specified number of rounds, or until the imputed values and parameters converge to a stable distribution.
- Assuming m stands for the number of imputations, m imputed complete datasets are produced at the end.

This SRMI approach could be applied at both person and household levels to impute the household income or expenditure of missing data. Households with zero income or expenditure are recoded as missing, before SRMI is also applied on them (See Figure 14). For the remainder of the study, SRMI at person level and SRMI at household level will be referred to as SRMI1 and SRMI2 respectively[26].

Figure 14: A brief summary of the SRMI approach


Households with zero
Income recoded as unspecified
Households with unspecified household income
Sequential regression multiple imputation (SRMI)
Per capita variable (after SRMI)
Per capita variable (after SRMI)
Poverty analyses

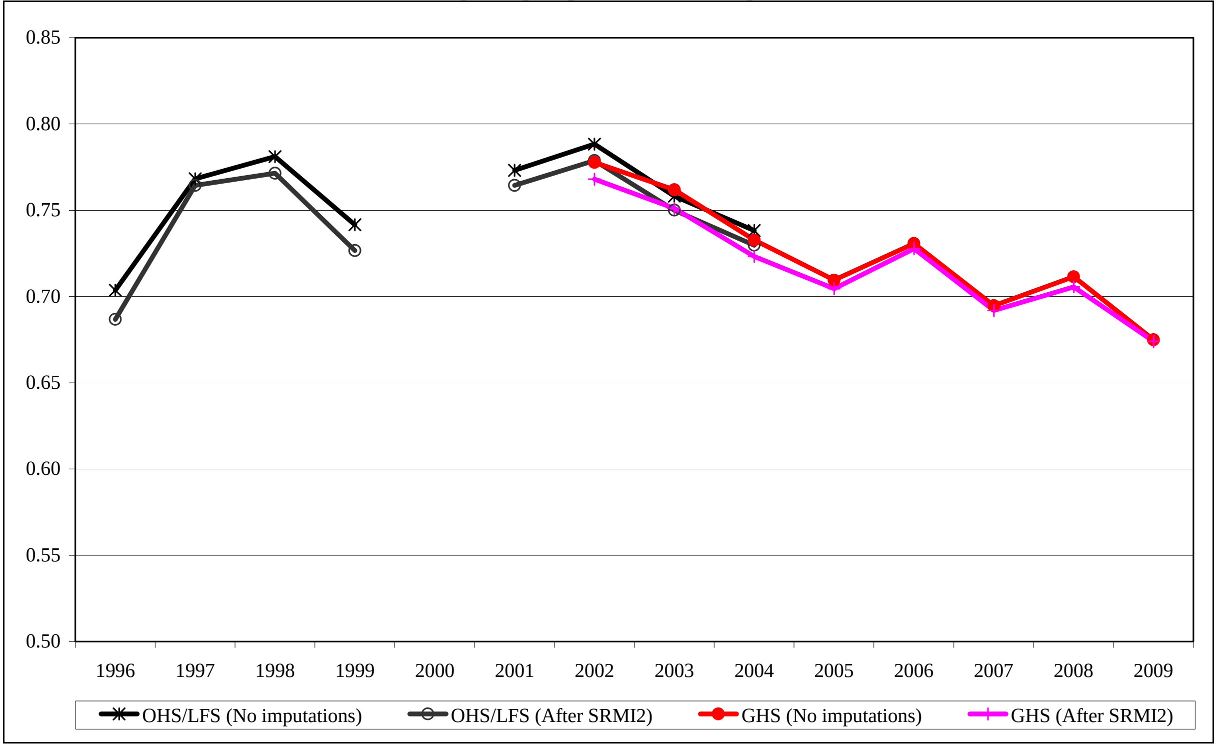
Yu (2009) is a recent study that adopted the SRMI approach to investigate the poverty on the two censuses and CS 2007. It was found that, after the application of SRMI, poverty headcount ratios decreased in all three surveys, but the trends remained the same, i.e., poverty increased between 1996 and 2001, before a rapid decline took place between 2001 and 2007 (See Table 10). In contrast, after SRMI2 was conducted on the OHSs, LFSs and GHSs, there was only a negligible decrease of poverty headcount ratios in these surveys, as shown in Figure 15, and the poverty trends as discussed in Section 3 remain the same.

[26] For detailed explanation of the SRMI1 and SRMI2 methods, refer to Yu (2009).

Table 10: Poverty headcount ratios using per capita income (2000 prices) variables of Census 1996, Census 2001 and CS 2007, after SRMI1 and SRMI2 (Poverty line: R322 per capita per month, 2000 prices)

	Without SRMI	After SRMI1	After SRMI2
Census 1996	0.606	0.601	0.576
Census 2001	0.670	0.647	0.592
CS 2007	0.529	0.478	0.463

Figure 15: Poverty headcount ratios using per capita expenditure (2000 prices) variables of OHSs, LFSs and GHSs, before and after SRMI2 (Poverty line: R322 per capita per month, 2000 prices)



4.8 External validation to improve the reliability of survey data

Survey data should be validated against various external sources in order to determine the reliability of it. These sources are discussed in this section.

4.8.1 Validation against national accounts

Surveys are more likely to under-estimate income / expenditure / consumption than to over-estimate it, due to reasons like fatigue, loss of interest, lack of motivation, illiteracy, recall bias, telescoping, and the tendency to declare zero or unspecified income, even if the households contain members who are employed or have income support from non-labour sources[27]. As a result, poverty could be over-estimated. Therefore, one view is that the distributional estimates of the survey data should be adjusted rightwards to be consistent with the national accounts series for aggregate household income / consumption (Van der Berg, Burger, Burger, Burger, Louw and Yu 2005 & 2009). That is, household survey means are replaced by national accounts means, but the distribution of the household survey is retained.

Table 11 shows the poverty headcount ratios after adjusting the survey means in line with the national accounts income mean, and naturally, these ratios declined after the adjustment (except for the OHS 1999 post-SRMI2 income variable). However, the poverty trends in each survey remain the same as discussed in Section 3.

[27] Table A.4 in the Appendix shows the total income, expenditure or consumption in each survey as percentage of the national accounts total income in the same year, and it can be seen that this percentage is below 100% in all surveys, except the post-SRMI2 OHS 1999 income variable.

Table 11: Poverty headcount ratios with and without adjustment of survey means in line with national accounts means, using the per capita variables (Poverty line: R322 per capita per month, 2000 prices)

Survey	Per capita variable	Year	Poverty headcount ratio	
			Without adjustment	With adjustment
Census/ CS	Income – No imputations	1996	0.606	0.445
		2001	0.670	0.518
		2007	0.529	0.409
	Income – After SRMI1	1996	0.601	0.326
		2001	0.647	0.432
		2007	0.478	0.335
	Income – After SRMI2	1996	0.576	0.334
		2001	0.592	0.440
		2007	0.463	0.335
IES	Income – STC	1995	0.434	0.415
		2000	0.559	0.440
		2005/2006	0.488	0.373
	Expenditure – STC	1995	0.447	0.423
		2000	0.564	0.441
		2005/2006	0.466	0.390
	Income - COICOP	1995	0.462	0.427
		2000	0.572	0.440
		2005/2006	0.473	0.379
	Consumption - COICOP	1995	0.502	0.341
		2000	0.601	0.343
		2005/2006	0.500	0.270
OHS	Expenditure – No imputations	1996	0.704	0.343
		1997	0.768	0.345
		1998	0.781	0.326
		1999	0.742	0.408
	Income – No imputations	1999	0.617	0.584
	Expenditure – After SRMI2	1996	0.687	0.337
		1997	0.764	0.374
		1998	0.771	0.345
		1999	0.727	0.442
	Income – After SRMI2	1999	0.596	0.622
LFS	Expenditure – No imputations	2001	0.773	0.476
		2002	0.788	0.515
		2003	0.758	0.555
		2004	0.738	0.599
	Expenditure – After SRMI2	2001	0.764	0.466
		2002	0.779	0.520
		2003	0.750	0.635
		2004	0.730	0.599
GHS	Expenditure – No imputations	2002	0.778	0.452
		2003	0.762	0.523
		2004	0.733	0.397
		2005	0.710	0.400
		2006	0.731	0.384
		2007	0.695	0.369
		2008	0.712	0.490
		2009	0.675	0.539

Table 11: Continued

Survey	Per capita variable	Year	Poverty headcount ratio	
			Without adjustment	With adjustment
GHS	Expenditure – After SRMI2	2002	0.768	0.449
		2003	0.751	0.510
		2004	0.723	0.410
		2005	0.705	0.440
		2006	0.728	0.381
		2007	0.692	0.366
		2008	0.706	0.528
		2009	0.674	0.536
PSLSD	Income	1993	0.598	0.474
	Expenditure	1993	0.566	0.346
NIDS	Income	2008	0.471	0.288
	Expenditure	2008	0.532	0.318
AMPS	Income	1993	0.586	0.438
		1994	0.593	0.420
		1995	0.594	0.434
		1996	0.610	0.437
		1997	0.589	0.407
		1998	0.583	0.415
		1999	0.591	0.415
		2000	0.582	0.428
		2001	0.579	0.425
		2002	0.563	0.387
		2003	0.554	0.388
		2004	0.548	0.362
		2005	0.519	0.345
		2006	0.512	0.328
		2007	0.455	0.298
		2008	0.410	0.283
		2009	0.414	0.282

However, adjusting survey means in line with national accounts mean implies the following must be true (Deaton 2001: 135): (1) the national accounts estimates are correct; (2) survey estimates of the mean are incorrect; (3) in spite of (2), the income / consumption levels of each household in the survey are correct up to a multiplicative factor. Proponents of the adjustment procedure generally believe that national accounts data are, in general, superior to survey data, and argue that not adjusting the survey means is more likely to introduce a larger error into the trends than adjusting the means[28]. However, if the sources of data disagree and there is no reason to favour one over the other, a more modest version of adjustment is suggested, that is, the survey data are scaled up by some weighted average of the national accounts mean and the survey mean, at least after correcting for conceptual differences and coverage (Deaton 2001: 136). The possible problems of national accounts data as well as reasons why adjusting the survey means might even create more negative effects on the reliability of poverty estimates are the focus of this section.

First, it is argued by some (Ravallion 2000; Deaton 2001: 133-134; Karshenas 2003: 694; Ravallion 2003: 646) that the national accounts estimates of consumption might not be the ideal variable to be treated as the gold standard to which the survey estimates should correspond. While the consumption measure in household survey is derived from self-reported expenditures (e.g., cash and from own stock) by the households in the interviews, households are treated as

[28] An example is the rapid decline of income and expenditure between IES 1995 and IES 2000. The magnitude of the measured decline is even greater than the fall in output during the Great Depression, as mentioned earlier. Hence, the poverty rate would show a rapid decline between the two surveys, and such decrease would be smaller had the distribution of the 2000 data been adjusted in line with national accounts mean.

residual claimants in the national accounts, as aggregate consumption is simply the residual obtained by subtracting other measured forms of domestic absorption from aggregate output. Hence, the errors and omissions in the estimation of the other components of the gross domestic product (GDP) all impinge on aggregate consumption.

The second problem with the national account estimates of consumption is that they implicitly include spending by unincorporated businesses and non-profit organizations, for example, religious groups, trade unions, clubs, and political parties. However, these estimates are not captured in surveys, as the aforementioned institutions are not households and hence did not take part in the surveys. Hence, the growth measured in the national accounts consumption might not really show up in progress in the living standards of the poor, and if the survey income / consumption distribution is adjusted (rightwards) in line with the national accounts consumption mean, this would result in an under-estimation of poverty (Ravallion 2000; Deaton 2001: 133-134; Karshenas 2003: 694; Ravallion 2003: 646-647).

Thirdly, Ravallion (2000 & 2003: 646-647) and Deaton (2001: 133-134 & 2005: 10) argue that rich households are missed more than the poor by surveys (i.e., there is unit non-response), as the well-off households are more likely to refuse to participate in the survey, or it is relatively more difficult to penetrate the gated communities (e.g., getting past the guard dogs) in which many rich people live. Hence, such households could be replaced by the more compliant but perhaps less well-off ones. Furthermore, even if rich households take part in the survey, the included rich people are likely to understate their income / consumption more than the included poor do, and this implies that poverty could be over-estimated.

If the survey mean is simply replaced by the national accounts mean, it assumes that the survey under-estimates income / consumption by a constant proportion across all levels. Thus, if this were untrue, after the adjustment, the income / consumption of the poor households could be seriously over-estimated, and poverty would in turn be under-estimated. As an example, the bottom 20% and top 20% of the population under-stated their expenditures by 25% and 50% respectively, while the average household under-stated its expenditure by 35% (when comparing with national accounts mean), if there is a uniform rightward adjustment of the survey mean in line with the national accounts mean by 35%, this clearly results in the over-estimation of expenditure of the poor households, and a subsequent under-estimation of poverty. This implies that the simple adjustment of the survey distribution upwards in line with the national accounts mean might not help improving the survey poverty estimates, if the unreliable survey distribution is the root of the problem but still not corrected.

It might also be true that surveys have missed the poor rural households (as it is expensive or dangerous to visit these places) as well as the very poor without fixed abode (i.e., homeless), and as a result of failing to include these poor households in the survey, the survey income / consumption estimates would be biased upwards. Once again, the main problem has to do with the incorrect distribution of survey data as a result of failing to capture these poor households as part of the sample, and simply adjusting the survey mean in line with national accounts by assuming the extent of adjustment is uniform across the whole population might not improve the reliability of poverty estimates, but rather complicate matters.

Based on the above arguments, different kinds of households have different likelihoods of being included in household surveys. As a result, survey results need to be weighted correctly to give an accurate representation of the population as a whole, with the calculation of suitable weights depending on the availability of accurate, up-to-date information about the population (Deaton 2001: 133-134). This implies that the replacement of survey means by national accounts means does not improve the poverty estimates at all, and might even worsen them, if the issues relating to the survey weights are not sorted out right at the beginning.

Other problems affecting the comparability between national accounts and household survey estimates are related to the capture of informal economic activities and certain income items. First, Deaton (2005) and Ravallion (2003: 646-647) argue that the value of informal activities is notoriously difficult to measure in the national accounts. Hence, as an economy grows and its structures change, many production activities shift from the informal sector to the formal sector. Consequently, economic activity is increasingly accurately captured in the national accounts data. This implies that the level of national accounts income is understated but growth is overstated as the economy develops and grows. This could partly explain the diverging gap between national accounts and household survey estimates of income in countries like India (Deaton and Kozel 2005). Secondly, in the national accounts income and private consumption estimates, items like imputed rent and in-kind income are taken account of, but they might not be recorded in household surveys, and this could result in differences between the two series[29].

4.8.2 Validation against other external sources

In addition to the national accounts, the survey data could also be validated against other external sources. Some of the commonly chosen external sources are discussed here. The focus is on the validation of IES data against these sources. First, the survey data on social grants income could be compared with the social grants expenditure by the National Treasury. For example, Table 12 below shows that, in general, the IES 2000 and IES 2005/2006 did a decent job of capturing social grants income, despite the fact that disability grant income was under-captured.

Table 12: Social grants income of IES 2000 and 2005/2006 compared with social grants expenditure of National Treasury (Rand million, nominal terms)

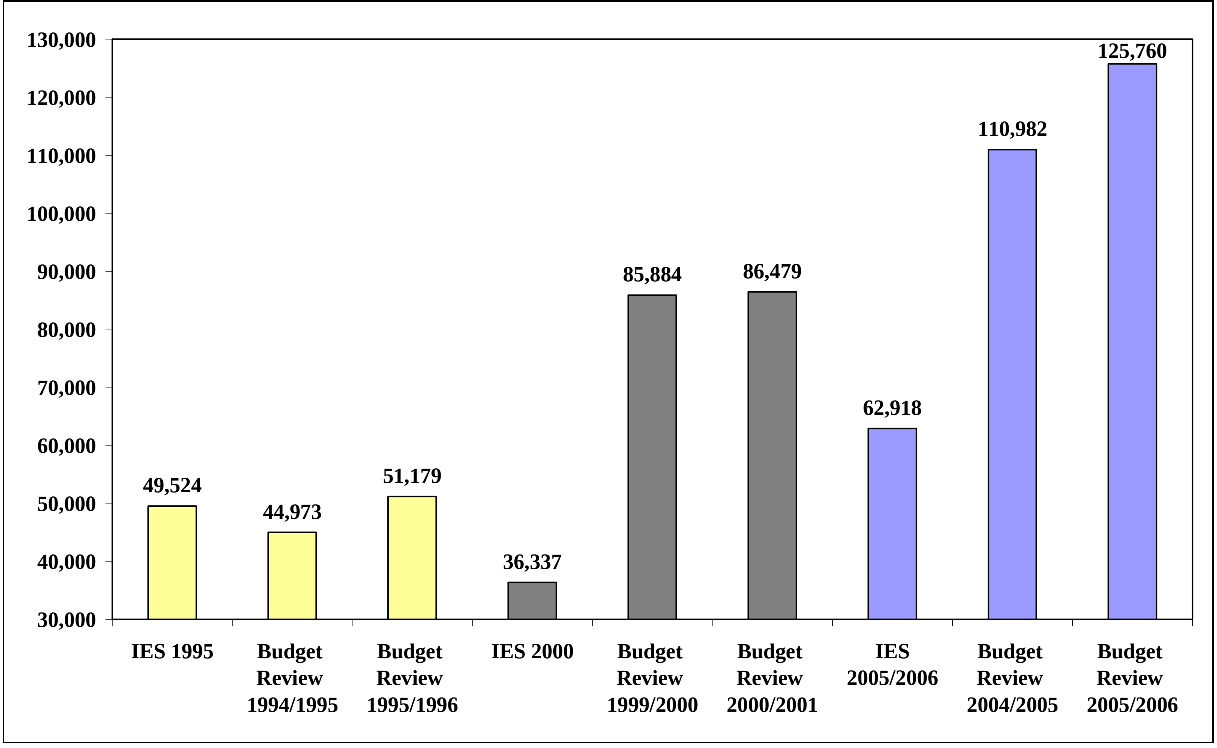
	Old-age/War pension	**Disability grant**	**Child/Family/Other grants**
[A]: IES 2000	15 402	3 058	1 533
[B]: Treasury - 1999/2000	11 660	3 823	944
[C]: Treasury - 2000/2001	12 208	4 066	1 770
[A] / [B]	132.1%	80.0%	162.4%
[A] / [C]	126.2%	75.2%	86.6%
[D]: IES 2005/2006	25 301	10 375	19 981
[E]: Treasury - 2004/2005	18 540	12 570	13 774
[F]: Treasury - 2005/2006	20 025	14 438	17 465
[D] / [E]	136.5%	82.5%	145.1%
[D] / [F]	126.3%	71.9%	114.4%

Data sources: Own calculations using IES data and National Treasury Budget Review (various issues).

Secondly, net personal income tax expenditure data of the survey is compared with net personal income tax revenue received by SARS. Figure 16 shows that IES 1995 did an outstanding job of capturing this tax expenditure accurately. However, the income tax expenditure captured in IES 2000 is only equivalent to slightly above 40% of the income tax revenue of SARS in both the 1999/2000 and 2000/2001 budget. This under-estimation of tax expenditure in IES 2000 could be associated with the very low total income captured in the survey (compared with the national accounts total income in the same year). The under-capture of income tax expenditure also took place in IES 2005/2006, despite the extent of it being a little (about 57% of the income tax revenue of SARS as reported in the 2004/2005 and 2005/2006 budget was captured).

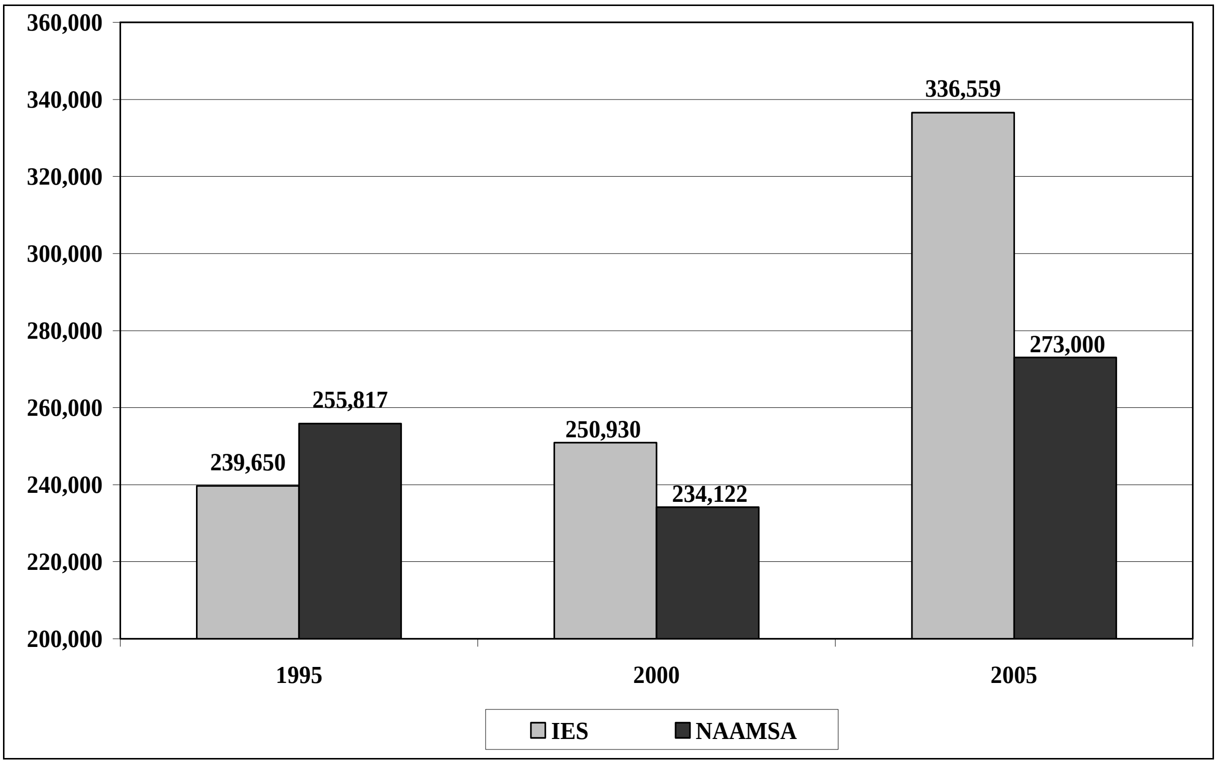
[29] IES 2005/2006 and NIDS are two surveys containing questions that clearly asked the respondents to declare imputed rent and in-kind income, and these items were taken into consideration when household income and consumption were derived. This is not the case in other surveys under study, as respondents were simply asked to declare income or expenditure from all sources, but some respondents might not be aware that imputed rent and in-kind income are income or expenditure items. The poverty headcount ratio in IES 2005/2006 was 0.473, using the income variable that included the imputed rent, but increased to 0.504 after excluding the imputed rent. Similarly, the NIDS income poverty increased from 0.471 to 0.534 as a result of excluding the imputed rent.

Figure 16: Net personal income tax expenditure of IESs compared with net personal income tax revenue of SARS (Rand million, nominal terms)



Data sources: Own calculations using IES data and National Treasury Budget Review (various issues).

Figure 17: Number of households with non-zero expenditure on new vehicle purchase in IESs compared with number of new vehicles sold from NAAMSA data



Data sources: Own calculations using IES and NAAMSA data.

In the three IESs, household heads were asked to declare expenditure on new and used vehicles. Thus, the statistics on the number of new cars sold from the National Association of Automobile Manufacturers of South Africa (NAAMSA) could be compared with the number of households with non-zero expenditure on new and used vehicles in the IESs. A drawback of the latter data is that it is impossible to know the number of new vehicles purchased in each household, and hence, the IES and NAAMSA data could only be compared based on the assumption that each household reporting non-zero new vehicle spending in the IESs only purchased one new vehicle. The results from Figure 17 above show that the IES 2005/2006 over-estimated the number of new motor vehicle purchases.

Finally, the survey data on petrol expenditure could be compared with the estimated petrol cost released by the South African Petroleum Industry Association (SAPIA). For instance, Table 13 compares the estimated total cost of petrol as reported by SAPIA and the total petrol expenditure from the IESs, and the results show that petrol expenditure in IES 2000 and 2005/2006, as for the personal income tax expenditure, was seriously under-estimated, when compared with external sources.

Table 13: Petrol expenditure in the IESs compared with estimated petrol cost from SAPIA

IES	[A]: IES petrol expenditure (Rand million)	[B]: SAPIA (million litre)	[C]: SAPIA: Fuel price per litre (97, Coast)	[D] = [B] × [C] Estimated total cost (Rand million)	[A]/ [D]
1995	R7 277	10 020	0.5708	R5 720	127%
2000	R12 852	10 556	1.9511	R20 593	63%
2005/2006	R23 533	11 158	4.9527	R55 263	43%

Data sources: Own calculations using IES and SAPIA data.

Note: The IES 1995 data are compared with the aggregate of SAPIA's 1994Q4, 1995Q1, 1995Q2 and 1995Q3 data, the IES 2000 data are compared with the aggregate of SAPIA's 1999Q4, 2000Q1, 2000Q2 and 2000Q3 data, and the IES 2005/2006 data are compared with the sum of SAPIA's 2005Q4, 2006Q1, 2006Q2 and 2006Q3 data.

4.9 Post-stratification weighting

With the exception of Census 1996 and Census 2001, all the data sources for poverty analyses in this study are survey data, as only a sample of people from the population took part in the survey. Design weights are created to make the sample represent the population. Different households have different inclusion probabilities as a result of both designed and unplanned factors. Hence, some households are over-represented relative to the others, and vice versa. In order for the sample estimates to accurately reflect the population, there is a need to weight each household according to its true inclusion probability.

In addition, due to the presence of non-coverage and unit non-response, post-stratification adjustment to the design weights is necessary by benchmarking the survey data to external aggregate population data so as to impose consistency between survey results and those from external sources. In the Stats SA survey data under study (IESs, OHSs/LFSs/QLFSs and GHSs), the person weights were post-stratified to the external population totals, i.e., the mid-year population estimates at the time of the survey derived by using the Census 1991, 1996 and 2001 information, with the pre- and post-census year population information being calculated using exponential interpolation and extrapolation.

Nonetheless, some concerns were raised regarding the reliability of the post-stratification design weights (Branson 2009):

- The auxiliary data (i.e., the mid-year population estimates) used as a benchmark in the post-stratification adjustment could be unreliable, inconsistent over time and of poor quality, thereby resulting in temporal inconsistencies even at the aggregate level. Branson (2009: 14) argues that this is likely the case in the population data derived by the Census, as the data are outdated to be used to project population estimates over a long period. Hence, the increased precision of the post-stratification weights could be offset by the potential bias introduced by using the questionable auxiliary data;
- Since the survey data are cross sectional, the purpose of the post-stratification adjustment is to produce the best estimates of the population, given the information available at the time of the survey. However, temporal consistency is not considered. This creates

problems when the data are used for time-series analyses;

- As the post-stratification adjustment of the Stats SA data was conducted at the person level (i.e., the person weight), this could result in inconsistency between person-level and household-level data, and the resultant analyses done at person and household levels would not necessarily agree.

Hence, the entropy post-stratification approach is conducted to re-weight the person weights of all the data under study to conform to the race, gender and age distribution of the population estimates as calculated by the Actuarial Society of South Africa 2003 (ASSA 2003) model. Branson (2009: 17) argues that the population data derived from the ASSA model is more time consistent.

The ASSA 2003 model aims to project the South African mid-year population from 1985, on the basis of various demographic, epidemiological and behavioural assumptions. The model could also be used to project trends in fertility and mortality as well as HIV/AIDS prevalence rate. There were two ASSA 2003 models at the time of this study: the full model projects the population of the four race groups by gender and age category (18 categories in total: 0-4 years, 5-9 years, and so forth, with the last category being "85 years or above) as well as the provincial population, while the lite model does not divide the population by race.

The entropy approach could be explained as follows: let x be a random variable with possible outcomes $x_k, k = 1,2,..., K$ and probabilities, $p = (p_1, p_2,..., p_k)'$, then the entropy measure is: $H(p) = -\sum_k p_k \ln p_k$, where $0 \cdot \ln(0)$ is defined to be 0. H(p) = 0 presents the degenerate solution, one possible outcome with certainty. H(p) reaches a maximum when the probability distribution is uniform. This is referred to as the maximum entropy (ME) approach.

The maximum entropy approach can be generalized to include prior information about the probability distribution with the aim to improve the accuracy of the estimates. This is known as the cross entropy (CE) approach and could be explained as follows: consider a survey sample of K individuals prior to adjustment probabilities q_k, i.e., the initial Stats SA person weights converted into proportions to the sum of one. Each individual has a vector of x_k characteristics (e.g., race, gender, age group). The CE estimate of p is the estimate which minimizes the difference from q, given the constraints to the problem. Alternatively, this implies the person weights are adjusted to meet aggregate trends (as derived by the ASSA model) which appear realistic over time, while simultaneously diverging as little as possible from the original Stats SA person weights.

In equation terms, the CE approach could be explained as follows (Golan, Judge and Miller 1996; Branson 2009: 34-36):
$\underset{p_k}{Min}\, I(p,q) = \underset{p_k}{Min}\left(\sum_{k=1}^{K} p_k \ln\left(\frac{p_k}{q_k}\right)\right) = \underset{p_k}{Min}\left(\sum_{k=1}^{K} p_k \ln p_k - \sum_{k=1}^{K} p_k \ln q_k\right)$, subject to the moment consistency constraints $\sum_{k=1}^{K} p_k x_t = y_t \quad t \in [1,...,T]$ and adding-up normalization constraint $\sum_{k=1}^{K} p_k = 1$.

Each x_t stands for a person-level indicator, indicating which demographic group the individual is in (e.g., the individual's gender, age category and race). T represents the number of restrictions. For example, if race (4 categories), gender (2 categories) and age groups (18 categories) are used, altogether there are 144 race-gender-age constraints (4 × 2 × 18), nine provincial constraints, plus the category "missing" (i.e., those with unspecified race, gender or age), i.e., 154 (144 + 9 + 1) constraints in total.

The new probability person weights are estimated as follows:

$$\underset{p_k}{Min}\, L = \underset{p_k}{Min}\left(\sum_{k=1}^{K} p_k \ln\left(\frac{p_k}{q_k}\right) + \sum_{t=1}^{T} \lambda_t \left(y_t - \sum_{k=1}^{K} p_k x_k \right) + \mu\left(1 - \sum_{k=1}^{K} p_k\right)\right)$$

The first-order conditions are:

$$\frac{\partial L}{\partial p_k} = \ln p_k - \ln q_k + 1 - \sum_{t=1}^{T} \lambda_t x_t - \mu = 0 \qquad k \in [1,...,K]$$

$$\frac{\partial L}{\partial \lambda_t} = y_t - \sum_{k=1}^{K} p_k x_k = 0 \qquad t \in [1,...,T]$$

$$\frac{\partial L}{\partial \mu} = 1 - \sum_{k=1}^{K} p_k = 0$$

The solution to this can be written as:

$$p_k = \frac{q_k}{\Omega\left(\tilde{\lambda}_1, \tilde{\lambda}_2, ..., \tilde{\lambda}_T\right)} \overset{\sim}{\exp}\left[\sum_{t=1}^{T} \tilde{\lambda}_t x_k\right] \qquad k \in [1,...,K], \text{ where } \Omega\left(\tilde{\lambda}\right) = \sum_{k=1}^{K} q_k \exp\left[\sum_{t=1}^{T} \tilde{\lambda}_t x_k\right]$$

Once the entropy person weights are derived, the household entropy weight variable is created and is equal to the mean entropy person weight within the household. The CE weights will be later used to investigate poverty estimates and trends, with their results compared to those obtained by using the original person and household weights.

The most efficient way to adjust the person weights would be to use the original design person weights (i.e., before the post-stratification adjustment against the Census mid-year population estimates). However, these weights are not publicly available and hence the adjusted design person weights (i.e., after the adjustment against the Census estimates) are used.

The approach discussed above was adopted by Branson (2009), the only South African study that investigated the labour market trends using the entropy approach (her study did not analyse the poverty trends). The person weights of OHS 1995-1999 and the March LFSs in 2000-2004[30] were re-weighted. After that, Branson looked at the trends in the share of single-person households, population shares by gender and area type of residence respectively, economically population and the number of employed, by using the Stats SA person weights as they were, the adjusted person weights after ME approach and the adjusted person weights after the CE approach.

In particular, she investigated whether the abrupt changes during certain years (especially in the OHSs and the changeover from OHS to LFS) were attributable to the inappropriate post-stratification technique by Stats SA or rather due to other reasons like changes in the questionnaire design, etc. After the entropy approach was adopted, it was found that "although there are small changes, the entropy weights have no significant effect in creating a more consistent trend in the labour market variables between 1995 and 2004. In other words, the large

[30] When imposing the ASSA 2003 model's population estimates constraints on the entropy model, Branson (2009) combined the "80-84 years" and "85 years or above" categories together as "80 years or above". In other words, there were 17 age categories in total. Altogether there are 136 race-gender-age constraints (4 × 2 × 17), 9 provincial constraints, plus the category "missing" (i.e., those with unspecified race, gender or age), i.e., 146 (136 + 9 + 1) constraints in total.

inconsistencies in the labour market variables are not a result of shifts in the weights" (Branson 2009: 53). The same findings were observed regardless of whether the ME or CE approach was conducted. Furthermore, Branson (2009: 53) found that the relatively higher employment levels in OHS 1995 (compared with OHS 1996-1997) and LFS 2000a (a rapid 1.5 million increase from the OHS 1999 employment level) were "unlikely to be a function of incorrect weights caused by post-stratification errors", but these abrupt changes were rather "either real or the result of measurement error".

For the remainder of this section, the poverty estimates and trends in the two censuses and CS 2007, the three IESs, OHSs, LFSs and GHSs are re-visited after the application of the minimum cross entropy (CE) approach to re-weight these datasets[31]. Table 14 reports the findings on the poverty headcount ratios. Looking at the poverty trends in the two censuses and CS 2007, after re-weighting the latter survey by the CE approach, the poverty headcount ratios showed a negligible increase at all three poverty lines compared with the ratios using the original Stats SA weights. The poverty trends remain the same, i.e., a moderate increase of poverty between the censuses, before a rapid decrease took place between Census 2001 and CS 2007.

Table 14: Poverty headcount ratios at different poverty lines before and after the cross entropy approach was conducted, using the per capita variables

Survey	Per capita variable	Year	Poverty headcount ratio	
			Stats SA weights	Cross entropy weights
Census/ CS	Income – No imputations	1996	0.606	0.606
		2001	0.670	0.670
		2007	0.529	0.534
	Income – After SRMI1	1996	0.601	0.601
		2001	0.647	0.647
		2007	0.478	0.484
	Income – After SRMI2	1996	0.576	0.576
		2001	0.592	0.592
		2007	0.463	0.469
IES	Income – STC	1995	0.434	0.445
		2000	0.559	0.557
		2005/2006	0.488	0.479
	Expenditure – STC	1995	0.447	0.457
		2000	0.564	0.561
		2005/2006	0.466	0.457
	Income - COICOP	1995	0.462	0.472
		2000	0.572	0.570
		2005/2006	0.473	0.464
	Consumption - COICOP	1995	0.502	0.514
		2000	0.601	0.599
		2005/2006	0.500	0.493

[31] As the two censuses were not surveys, they were not re-weighted. In addition, since the QLFS took place during a 3-month period, the February population figure derived by the ASSA model was used to derive the CE weights in the Q1 survey. Similarly, the May, August and November ASSA model's population figures were used to derive the CE weights in the Q2, Q3 and Q4 surveys respectively. Since IES 2005/2006 was conducted between September 2005 and August 2006, the March 2006 population figure derived by the ASSA model was used to derive the CE weights for this survey. As NIDS took place between January and December 2008, the mid-year population figure derived by the ASSA model was used to derive the CE weights.

Table 14: Continued

Survey	Per capita variable	Year	Poverty headcount ratio	
			Stats SA weights	Cross entropy weights
OHS/ LFS	Expenditure – No imputations	1996	0.704	0.722
		1997	0.768	0.755
		1998	0.781	0.774
		1999	0.742	0.736
		2001	0.773	0.765
		2002	0.788	0.780
		2003	0.758	0.750
		2004	0.738	0.731
	Expenditure – After SRMI2	1996	0.687	0.701
		1997	0.764	0.751
		1998	0.771	0.765
		1999	0.727	0.721
		2001	0.764	0.756
		2002	0.779	0.770
		2003	0.750	0.742
		2004	0.730	0.722
GHS	Expenditure – No imputations	2002	0.778	0.772
		2003	0.762	0.758
		2004	0.733	0.722
		2005	0.710	0.701
		2006	0.731	0.723
		2007	0.695	0.687
		2008	0.712	0.708
		2009	0.675	0.683
	Expenditure – After SRMI2	2002	0.768	0.762
		2003	0.751	0.747
		2004	0.723	0.712
		2005	0.705	0.696
		2006	0.728	0.719
		2007	0.692	0.684
		2008	0.706	0.702
		2009	0.674	0.682

With regard to the poverty trends in the IESs, after using the CE weights, the poverty headcount ratio increased slightly in IES 1995, but the opposite took place in IES 2000 and IES 2005/2006. However, the same poverty trends were still observed, i.e., a rapid increase between 1995 and 2000, before it decreased between the 2000 and 2005/2006 IESs, but the IES 2005/2006 poverty headcount ratios were higher than the IES 1995 ratios.

Next, looking at OHS 1996-1999 and the 2001-2004 September LFSs, the use of the CE weights resulted in slightly lower poverty headcount ratios in all surveys, except in OHS 1996. On the other hand, the poverty headcount ratios in GHS 2002-2009 also experienced a slight decrease in all surveys after using the CE weights, except in GHS 2009. Finally, the use of the CE weights did not cause any changes in the poverty trends in the OHSs, LFSs and GHSs in general.

5. Conclusion

This paper examined various factors affecting the comparability and reliability of poverty estimates and trends across household surveys. First, the pros and cons of using income and expenditure (consumption) for poverty analyses were discussed. Although the general consensus was that expenditure is the preferred variable to be used in developing countries, further investigation found that this might not be the case. Secondly, the possible merits and drawbacks of using the traditional recall approach and the diary approach to capture the income and expenditure were looked at, and it seems durable expenditure would always be captured with some flaws, regardless of which approach is adopted.

The issue of whether the income and expenditure should be captured in actual amounts or in bands / intervals / categories was investigated, as each method involves advantages and disadvantages. If the information is collected in actual amounts, the next question that arises is whether the amounts should be captured as a 'one-shot' single estimate or rather the aggregation of amounts from different sources. The pros and cons of each approach were discussed. If the information is collected in intervals instead, three issues come up: the appropriate method to convert the interval data into continuous data for the subsequent poverty analyses; the impact of the number of bands and width of each band on the poverty estimates (an issue that needs further investigation, as there is lack of South African studies on it); how to deal with households with zero or unspecified income or expenditure. It was found that the midpoint-Pareto method was most appropriate to make the interval data continuous, but there is insufficient research both domestically and internationally that investigates how the number and width of bands affect the poverty estimates. The sequential regression multiple imputation (SRMI) approach was used to impute the income (or expenditure) of households reporting zero or unspecified income (or expenditure).

The possible merits and drawbacks of adjusting the survey income (or expenditure) distribution in line with the national accounts income mean, as well as the validation of the survey data against external sources (e.g., income tax revenue data by the National Treasury) to evaluate the reliability of the former data were discussed. Finally, since the post-stratification adjustment of the survey weights in the Stats SA survey datasets did not take account of temporal consistency issue, concerns were raised with regard to using these cross-sectional datasets to investigate the change of poverty estimates over time. It was found that the cross entropy approach would address the temporal inconsistency problems and the minimum cross entropy (CE) would be adopted to re-weigh the datasets for further analyses on the aforementioned estimates over time. However, after the datasets were re-weighted, there were only negligible changes to the poverty estimates.

To conclude, as the income and expenditure information were collected so differently in each survey, the levels of poverty and inequality could differ a lot across the surveys. Yet, there is still a need to undertake the sort of analyses as done in this study in order to make valid comparisons of both the poverty and inequality levels and trends across the surveys.

6. References

Ahmed, N., Brzozowski, M. and Crossley, T.F. (2006). *Measurement errors in recall food consumption data*. IFS Working Paper W06/21. London: Institute for Fiscal Studies.

Ardington, C., Lam, D., Leibbrandt, M. and Welch, M. (2005). *The sensitivity of estimates of post-apartheid changes in South African poverty and inequality to key data imputations*. CSSR working paper no. 106. Cape Town: Centre for Social Science Research.

Argent, J., Franklin, S., Keswell, M., Leibbrandt, M. & Levinsohn, J. (2009). *Expenditure: Report on NIDS Wave 1*. Technical Paper No. 4. Cape Town: Southern African Labour and Development Research Unit, University of Cape Town.

Battinstin, E. (2003). *Errors in survey reports of consumption expenditures.* IFS Working Papers W03/07. London: Institute for Fiscal Studies.

Blundell, R. and Preston, I. (1998). Consumption inequality and income uncertainty. *Quarterly Journal of Economics*. 113(2): 604 – 640.

Branson, N. (2009). *Re-weighting the OHS and LFS national household survey data to create a consistent series over time: A cross entropy estimation approach.* SALDRU Working Paper Number 38. Cape Town: Southern Africa Labour and Development Research Unit, University of Cape Town.

Browning, M., Crossley, T.F. and Weber, G. (2002). *Asking consumption questions in general purpose surveys*. SEDAP Research Paper No. 77. Hamilton: The Program for Research on Social and Economic Dimensions of an Aging Population (SEDAP).

Cloutier, N.R. (1988). Pareto extrapolation using grouped income data. *Journal of Regional Science.* 28(3): 415-419.

Corti, L. (1993). Using diaries in social research. *Social Research Update*. March 1993. Guildford: University of Surrey.

Davern, M., Rodin, H., Beebe, T.J. and Thiede Call, K. (2005). The effect of income question design in health surveys on family income, poverty and eligibility estimates. *Health Services Research*. 40(5): 1534-1552.

Deaton, A. (1997). *The analysis of household surveys: A microeconomic approach to development policy*. Baltimore: The John Hopkins University Press.

Deaton, A. (2001). Counting the world's poor: Problems and possible solutions. *World Bank Research Observer*. 16(2): 125 – 147.

Deaton, A. (2005). Measuring poverty in a growing world (or measuring growth in a poor world). *The Review of Economics and Statistics*. 87(1): 1-19.

Deaton, A. and Grosh, M. (2000). Consumption. In Grosh, M. and Glewwe, P. (ed.), *Designing household survey questionnaire for developing countries: Lessons from 15 years of the living standards measurement study – Volume One*. Washington: The World Bank: 91 – 133.

Deaton, A. and Kozel, V. (2005). Data and dogma: The great Indian poverty debate. *The World Bank Research Observer*. 20(2): 177-199.

Duclos, J. and Araar, A. (2006). *Poverty and equity: Measurement, policy and estimation with DAD.* 1st edition. Ottawa: Springer.

Fields, G.S. (1989). *A compendium of data on inequality and poverty for the developing world.* Unpublished report. New York: Cornell University.

Finn, A., Leibbrandt, M. and Woolard, I. (2009). *Income and expenditure inequality: Analysis of the NIDS wave 1 dataset.* Discussion Paper No. 5. Cape Town: Southern Africa Labour and Development Research Unit, University of Cape Town.

Guenard, C. and Mesple-Somps, S. (2010). Measuring inequalities: Do household surveys paint a realistic picture? *Review of Income and Wealth*. 56(3): 519-538.

Haughton, J. and Khandker, S.R. (2009). *Handbook on poverty and inequality*. Washington: The World Bank.

Karshenas, M. (2003). Global poverty: National accounts based versus survey based estimates. *Development and Change*. 34(4): 683 – 712.

Lacerda, M., Ardington, C. and Leibbrandt, M. (2008). *Sequential regression multiple imputation for incomplete multivariate data using Markov chain Monte Carlo*. SALDRU paper series number 13. Cape Town: Southern African Labour and Development Research Unit.

Malherbe, J.E. (2007). *An analysis of income and poverty in South Africa.* Unpublished Master thesis. Stellenbosch: Stellenbosch University.

McKay, A. (2000). Should the survey measure total household income? In Grosh, M. and Glewwe, P. (ed.), *Designing household survey questionnaire for developing countries: Lessons from 15 years of the living standards measurement study – Volume Two*. Washington: The World Bank: 83 – 104.

National Association of Automobile Manufacturers of South Africa (NAAMSA). [Online] Available: http://www.naamsa.co.za/

National Treasury. *Budget Review* (various issues). Pretoria: Government Printers.

Posel, D. and Casale D. (2005). *Who replies in brackets and what are the implications for earnings estimates? An analysis of earnings data from South Africa.* Paper presented at the ESSA Conference, Durban.

Raghunathan, T.E., Lepkowski, J.M., Van Hoewyk, J. and Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*. 27(1): 85-95.

Ravallion, M. (2000). Should poverty measures be anchored to the national accounts? *Economic and Political Weekly*: August 26 – September 2: 2345-2352.

Ravallion, M. (2003). Measuring aggregate welfare in developing countries: How well do national accounts and survey agree? *Review of Economics and Statistics*. 85(3): 645 – 652.

Riphahn, R.R. and Serfling, O. (2004). *Item non-response on income and wealth questions*. Unpublished manuscript. Basel: University of Basel, Switzerland.

Seiver, D.A. (1979). A note of the measurement of income inequality with income data. *The Review of Income and Wealth*. 25(2): 229-234.

South African Petroleum Industry Association (SAPIA). [Online] Available: www.sapia.co.za/

Sudman, S. and Ferber, R. (1971). Experiments in obtaining consumer expenditures by diary methods. *Journal of American Statistical Association*, 66(336): 725-735.

Van der Berg, Burger, R., Burger, R.P., Louw, M. and Yu, D. (2005). *Trends in poverty and inequality since political transition*. Stellenbosch Economic Working Papers: 01/05. Stellenbosch: Stellenbosch University.

Van der Berg, S., Burger, R., Burger, R.P., Louw, M. and Yu, D. (2009). *A series of national account-consistent estimates of poverty and inequality in South Africa*. Stellenbosch Economic Working Papers: 09/07. Stellenbosch: Stellenbosch University.

Van der Berg, S., Louw, M. and Du Toit, L. (2008). *Poverty trends since the transition: What we know*. Stellenbosch: Stellenbosch University.

Vermaak, C. (2005). *Trends in income distribution, inequality and poverty in South Africa, 1995 to 2003*. Paper presented at the ESSA Conference, Durban.

Von Fintel, D. (2006). *Earnings bracket obstacles in household surveys – how sharp are the tools in the shed?* Stellenbosch Economic Working Papers: 08/06. Stellenbosch: Stellenbosch University.

Von Fintel, D. (2007). Dealing with earnings bracket responses in household surveys – How sharp are midpoint imputations. *South African Journal of Economics*. 75(2): 293-312.

Whiteford, A. and McGrath, M. (1994). *The distribution of income in South Africa*. 1st edition. Pretoria: Human Sciences Research Council.

Wiseman, V., Conteh, L. and Matovu, F. (2005). Using diaries to collect data in resource-poor settings: questions on design and implementation. *Health Policy and Planning*. 20(6): 393 – 404.

Woolard, I. and Leibbrandt, M. (2006). *Towards a poverty Line for South Africa: Background note*. Cape Town: Southern Africa Labour and Development Research Unit, University of Cape Town.

Yu, D. (2008). *The comparability of Income and Expenditure Surveys 1995, 2000 and 2005/2006*. Stellenbosch Economic Working Papers: 11/2008. Stellenbosch: Stellenbosch University.

Yu, D. (2009). *The comparability of Census 1996, Census 2001 and Community Survey 2007*. Stellenbosch Economic Working Papers: 21/09. Stellenbosch: Stellenbosch University.

Appendix

Figure A.1: Poverty gap indices in each survey (Poverty line: R3 864 per capita per annum, 2000 prices)

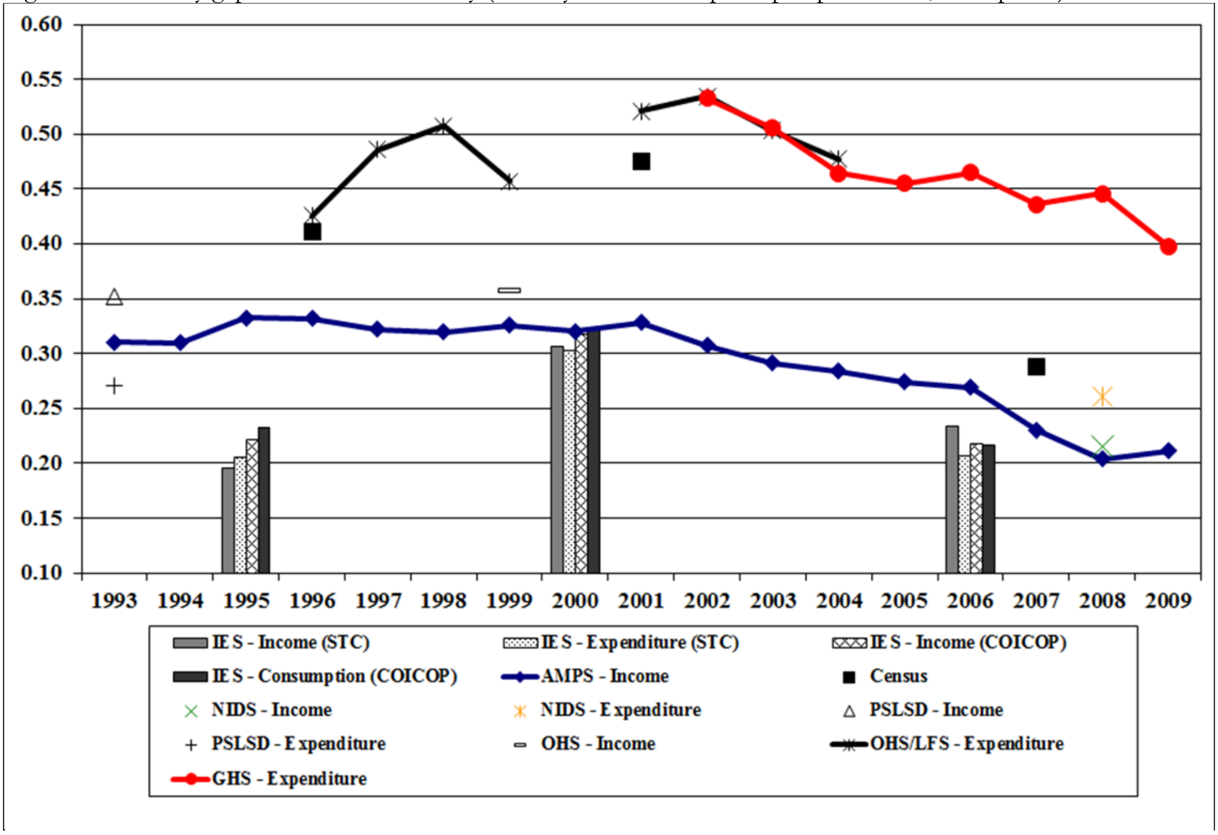


Figure A.2: Squared poverty gap indices in each survey (Poverty line: R3 864 per capita per annum, 2000 prices)

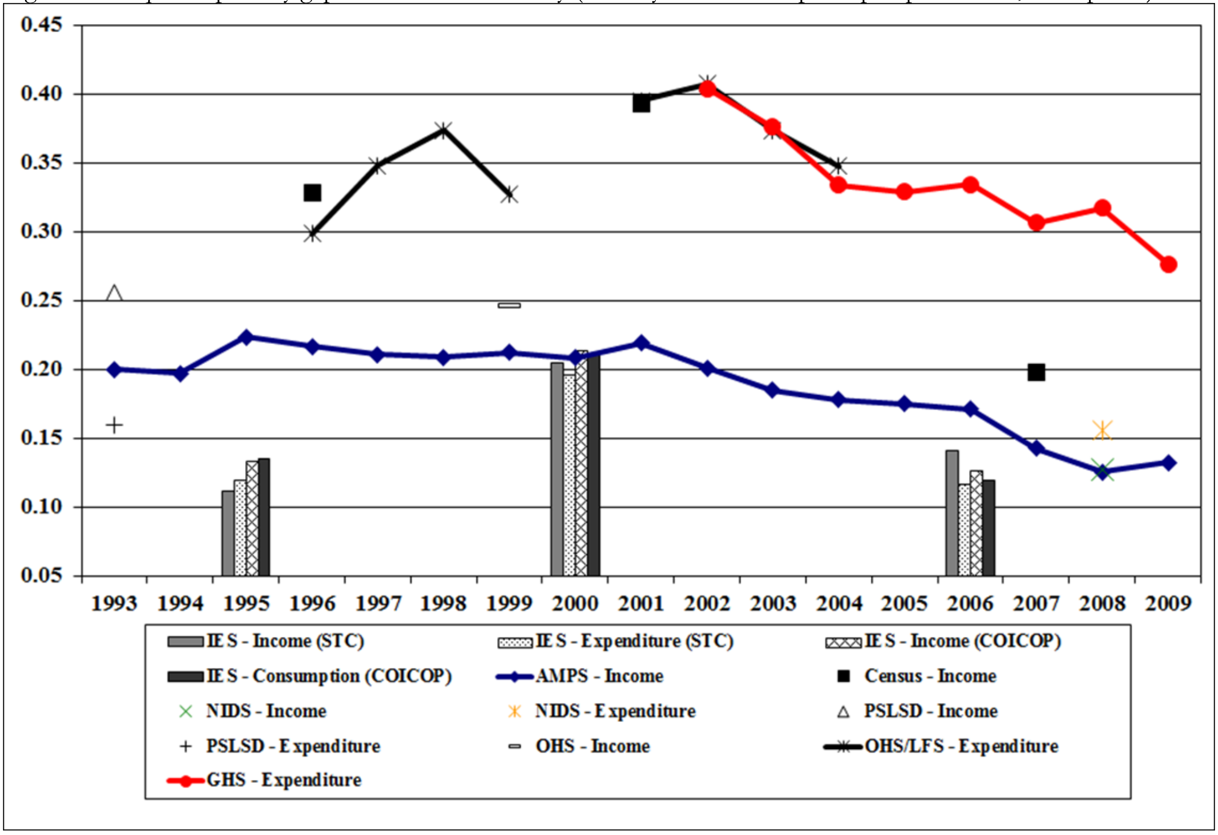


Table A.1: Nominal monthly household income categories in Census 1996, Census 2001 and CS 2007

Census 1996	Census 2001 & CS 2007
1: None	1: None
2: R1 – R200	2: R1 – R400
3: R201 – R500	3: R401 – R800
4: R501 – R1 000	4: R801 – R1 600
5: R1 001 – R1 500	5: R1 601 – R3 200
6: R1 501 – R2 500	6: R3 201 – R6 400
7: R2 501 – R3 500	7: R6 401 – R12 800
8: R3 501 – R4 500	8: R12 801 – R25 600
9: R4 501 – R6 000	9: R25 601 – R51 200
10: R6 001 – R8 000	10: R51 201 – R102 400
11: R8 001 – R11 000	11: R102 401 – R204 800
12: R11 001 – R16 000	12: R204 801 or more
13: R16 001 – R30 000	13: Unspecified
14: R30 001 or more	
99: Unspecified	

Table A.2: Nominal monthly household income or expenditure categories in OHSs, LFSs and GHSs

OHS 1999 (Income), OHS 1999 (Expenditure), LFS 2001-2004 September (Expenditure), and GHS 2002-2008 (Expenditure)	GHS 2009 (Expenditure)
1: R0 – R399	1: R0
2: R400 – R799	2: R1 – R199
3: R800 – R1 199	3: R200 – R399
4: R1 200 – R1 799	4: R400 – R799
5: R1 800 – R2 499	5: R800 – R1 199
6: R2 500 – R4 999	6: R1 200 – R1 799
7: R5 000 – R9 999	7: R1 800 – R2 499
8: R10 000 or more	8: R2 500 – R4 999
9: Don’t know	9: R5 000 – R9 999
10: Refuse	10: R10 000 or more
	11: Don’t know
	12: Refuse

Table A.3: Nominal monthly household income or expenditure categories in AMPSs

	1993	1994-1996	1997-1999	2000-2001	2002-2006	2007-2008	2009
1	R1-R99	R1-R99	R1-R99	R1-R199	R1-R199	R1-R299	R1-R499
2	R100-R199	R100-R199	R100-R199	R200-R299	R200-R299	R300-R399	R500-R599
3	R200-R299	R200-R299	R200-R299	R300-R399	R300-R399	R400-R499	R600-R699
4	R300-R399	R300-R399	R300-R399	R400-R499	R400-R499	R500-R599	R700-R799
5	R400-R499	R400-R499	R400-R499	R500-R599	R500-R599	R600-R699	R800-R899
6	R500-R599	R500-R599	R500-R599	R600-R699	R600-R699	R700-R799	R900-R999
7	R600-R699	R600-R699	R600-R699	R700-R799	R700-R799	R800-R899	R1 000-R1 099
8	R700-R799	R700-R799	R700-R799	R800-R899	R800-R899	R900-R999	R1 100-R1 199
9	R800-R899	R800-R899	R800-R899	R900-R999	R900-R999	R1 000-R1 099	R1 200-R1 399
10	R900-R999	R900-R999	R900-R999	R1 000-R1 099	R1 000-R1 099	R1 100-R1 199	R1 400-R1 599
11	R1 000-R1 099	R1 000-R1 099	R1 000-R1 099	R1 100-R1 199	R1 100-R1 199	R1 200-R1 399	R1 600-R1 999
12	R1 100-R1 199	R1 100-R1 199	R1 100-R1 199	R1 200-R1 399	R1 200-R1 399	R1 400-R1 599	R2 000-R2 499
13	R1 200-R1 399	R1 200-R1 399	R1 200-R1 399	R1 400-R1 599	R1 400-R1 599	R1 600-R1 999	R2 500-R2 999
14	R1 400-R1 599	R1 400-R1 599	R1 400-R1 599	R1 600-R1 999	R1 600-R1 999	R2 000-R2 499	R3 000-R3 999
15	R1 600-R1 999	R1 600-R1 999	R1 600-R1 999	R2 000-R2 499	R2 000-R2 499	R2 500-R2 999	R4 000-R4 999
16	R2 000-R2 499	R2 000-R2 499	R2 000-R2 499	R2 500-R2 999	R2 500-R2 999	R3 000-R3 999	R5 000-R5 999
17	R2 500-R2 999	R2 500-R2 999	R2 500-R2 999	R3 000-R3 999	R3 000-R3 999	R4 000-R4 999	R6 000-R6 999
18	R3 000-R3 999	R3 000-R3 999	R3 000-R3 999	R4 000-R4 999	R4 000-R4 999	R5 000-R5 999	R7 000-R7 999
19	R4 000-R4 999	R4 000-R4 999	R4 000-R4 999	R5 000-R5 999	R5 000-R5 999	R6 000-R6 999	R8 000-R8 999
20	R5 000-R5 999	R5 000-R5 999	R5 000-R5 999	R6 000-R6 999	R6 000-R6 999	R7 000-R7 999	R9 000-R9 999
21	R6 000-R6 999	R6 000-R6 999	R6 000-R6 999	R7 000-R7 999	R7 000-R7 999	R8 000-R8 999	R10 000-R10 999
22	R7 000-R7 999	R7 000-R7 999	R7 000-R7 999	R8 000-R8 999	R8 000-R8 999	R9 000-R9 999	R11 000-R11 999
23	R8 000-R8 999	R8 000-R8 999	R8 000-R8 999	R9 000-R9 999	R9 000-R9 999	R10 000-R10 999	R12 000-R13 999
24	R9 000-R9 999	R9 000-R9 999	R9 000-R9 999	R10 000-R10 999	R10 000-R10 999	R11 000-R11 999	R14 000-R15 999
25	R10 000-R10 999	R10 000-R10 999	R10 000-R10 999	R11 000-R11 999	R11 000-R11 999	R12 000-R13 999	R16 000-R19 999
26	R11 000-R11 999	R11 000-R11 999	R11 000-R11 999	R12 000-R13 999	R12 000-R13 999	R14 000-R15 999	R20 000-R24 999
27	R12 000-R12 999	R12 000-R13 999	R12 000-R13 999	R14 000-R15 999	R14 000-R15 999	R16 000-R19 999	R25 000-R29 999
28	R13 000-R13 999	R14 000-R15 999	R14 000-R15 999	R16 000-R17 999	R16 000-R19 999	R20 000-R24 999	R30 000-R39 999
29	R14 000+	R16 000+	R16 000-R17 999	R18 000-R19 999	R20 000-R24 999	R25 000-R29 999	R40 000-R49 999
30			R18 000+	R20 000+	R25 000-R29 999	R30 000-R39 999	R50 000+
31					R30 000-R39 999	R40 000+	
32					R40 000+		

Table A.4: Total household annual income and expenditure in the IESs using Standard Trade Classification approach (2000 prices, Rand million)

	IES1995		**IES2000**		**IES2005/2006**	
Total expenditure						
Housing	76 084	14.6%	78 656	17.1%	118 512	15.8%
Domestic workers	7 251	1.4%	11 703	2.6%	10 615	1.4%
Food	88 212	17.0%	83 748	18.3%	71 997	9.6%
Beverages	8 433	1.6%	9 781	2.1%	7 616	1.0%
Cigarettes and smokers' requisites	4 343	0.8%	4 530	1.0%	3 680	0.5%
Personal care	11 354	2.2%	14 242	3.1%	6 603	0.9%
Other household consumer goods	6 534	1.3%	4 821	1.1%	4 229	0.6%
Household services	1 612	0.3%	446	0.1%	323	0.0%
Household fuel	2 726	0.5%	4 087	0.9%	3 386	0.5%
Clothing and footwear	23 440	4.5%	16 981	3.7%	26 304	3.5%
Furniture/Equipment	18 923	3.6%	10 602	2.3%	21 234	2.8%
Health services	18 678	3.6%	16 937	3.7%	29 978	4.0%
Transport	48 988	9.4%	46 986	10.2%	110 498	14.7%
Computer and telecommunication equipment	1 502	0.3%	3 071	0.7%	4 655	0.6%
Communication for household purposes	10 907	2.1%	9 613	2.1%	16 414	2.2%
Education	8 822	1.7%	13 160	2.9%	18 558	2.5%
Reading matter and stationery	2 298	0.4%	3 109	0.7%	2 678	0.4%
Recreation, entertainment and sports	6 457	1.2%	7 147	1.6%	15 258	2.0%
Miscellaneous expenditure	166 270	32.0%	110 123	24.0%	274 949	36.6%
Expenditure on own harvest/livestock	6 714	1.3%	9 123	2.0%	3 667	0.5%
Total household annual expenditure	519 549	100.0%	458 867	100.0%	751 153	100.0%

Table A.5: FGT poverty estimates, after applying different intervals on the three IESs

		FGT poverty index		
		$\mathbf{P_0}$	$\mathbf{P_1}$	$\mathbf{P_2}$
Poverty line: R211 per month (2000 prices)				
IES 1995				
The actual continuous income variable		0.286	0.106	0.053
Applying the intervals on the income data	AMPS 2000 intervals (2000 prices)	0.275	0.106	0.053
	Census 1996 intervals (2000 prices)	0.271	0.109	0.056
	Census 2001 intervals (2000 prices)	0.253	0.102	0.053
	GHS 2009 intervals (2000 prices)	0.252	0.100	0.052
	R500 intervals	0.292	0.119	0.067
	R1 000 intervals	0.305	0.123	0.063
	R2 000 intervals	0.227	0.050	0.014
IES 2000				
The actual continuous income variable		0.429	0.206	0.127
Applying the intervals on the income data	AMPS 2000 intervals (2000 prices)	0.422	0.202	0.123
	Census 1996 intervals (2000 prices)	0.417	0.207	0.129
	Census 2001 intervals (2000 prices)	0.412	0.198	0.123
	GHS 2009 intervals (2000 prices)	0.411	0.198	0.123
	R500 intervals	0.416	0.192	0.114
	R1 000 intervals	0.426	0.199	0.116
	R2 000 intervals	0.391	0.127	0.055
IES 2005/2006				
The actual continuous income variable		0.338	0.137	0.075
Applying the intervals on the income data	AMPS 2000 intervals (2000 prices)	0.326	0.137	0.075
	Census 1996 intervals (2000 prices)	0.324	0.133	0.073
	Census 2001 intervals (2000 prices)	0.317	0.132	0.073
	GHS 2009 intervals (2000 prices)	0.319	0.133	0.074
	R500 intervals	0.341	0.138	0.076
	R1 000 intervals	0.332	0.140	0.077
	R2 000 intervals	0.354	0.128	0.060
Poverty line: R322 per month (2000 prices)				
IES 1995				
The actual continuous income variable		0.434	0.195	0.111
Applying the intervals on the income data	AMPS 2000 intervals (2000 prices)	0.433	0.195	0.111
	Census 1996 intervals (2000 prices)	0.446	0.194	0.112
	Census 2001 intervals (2000 prices)	0.406	0.187	0.107
	GHS 2009 intervals (2000 prices)	0.430	0.187	0.106
	R500 intervals	0.419	0.203	0.123
	R1 000 intervals	0.408	0.208	0.125
	R2 000 intervals	0.398	0.149	0.065
IES 2000				
The actual continuous income variable		0.559	0.307	0.204
Applying the intervals on the income data	AMPS 2000 intervals (2000 prices)	0.562	0.303	0.200
	Census 1996 intervals (2000 prices)	0.569	0.306	0.205
	Census 2001 intervals (2000 prices)	0.538	0.297	0.197
	GHS 2009 intervals (2000 prices)	0.551	0.295	0.197
	R500 intervals	0.559	0.296	0.192
	R1 000 intervals	0.553	0.300	0.195
	R2 000 intervals	0.497	0.241	0.133

Note: P_0: Poverty headcount ratio
P_1: Poverty gap ratio
P_2: Squared poverty gap ratio

Table A.5: Continued

		FGT poverty index		
		P_0	**P_1**	**P_2**
IES 2005/2006				
The actual continuous income variable		0.488	0.234	0.141
Applying the intervals on the income data	AMPS 2000 intervals (2000 prices)	0.484	0.234	0.141
	Census 1996 intervals (2000 prices)	0.501	0.227	0.137
	Census 2001 intervals (2000 prices)	0.464	0.225	0.135
	GHS 2009 intervals (2000 prices)	0.482	0.225	0.136
	R500 intervals	0.488	0.235	0.142
	R1 000 intervals	0.484	0.235	0.143
	R2 000 intervals	0.472	0.228	0.130
Poverty line: R593 per month (2000 prices)				
IES 1995				
The actual continuous income variable		0.622	0.352	0.236
Applying the intervals on the income data	AMPS 2000 intervals (2000 prices)	0.625	0.351	0.235
	Census 1996 intervals (2000 prices)	0.612	0.350	0.234
	Census 2001 intervals (2000 prices)	0.623	0.339	0.226
	GHS 2009 intervals (2000 prices)	0.605	0.340	0.226
	R500 intervals	0.618	0.356	0.243
	R1 000 intervals	0.612	0.358	0.245
	R2 000 intervals	0.606	0.322	0.197
IES 2000				
The actual continuous income variable		0.710	0.462	0.342
Applying the intervals on the income data	AMPS 2000 intervals (2000 prices)	0.713	0.458	0.339
	Census 1996 intervals (2000 prices)	0.705	0.459	0.340
	Census 2001 intervals (2000 prices)	0.717	0.448	0.331
	GHS 2009 intervals (2000 prices)	0.695	0.448	0.331
	R500 intervals	0.701	0.454	0.333
	R1 000 intervals	0.709	0.455	0.334
	R2 000 intervals	0.706	0.417	0.284
IES 2005/2006				
The actual continuous income variable		0.657	0.395	0.275
Applying the intervals on the income data	AMPS 2000 intervals (2000 prices)	0.659	0.394	0.274
	Census 1996 intervals (2000 prices)	0.651	0.389	0.268
	Census 2001 intervals (2000 prices)	0.665	0.381	0.264
	GHS 2009 intervals (2000 prices)	0.639	0.381	0.264
	R500 intervals	0.663	0.396	0.275
	R1 000 intervals	0.665	0.395	0.275
	R2 000 intervals	0.656	0.388	0.266

Note: P_0: Poverty headcount ratio
P_1: Poverty gap ratio
P_2: Squared poverty gap ratio

Table A.6: Comparison of annual total income/expenditure/consumption in various surveys with annual total income in the national accounts in the same year

Survey	Variable	Year	Amount (R million) (2000 prices)	As % of total income in the national accounts
Census/CS	Total income – without any imputations involved	1996	294 475	50.5%
		2001	366 341	52.5%
		2007	629 421	68.9%
	Total income – After SRMI1	1996	339 993	58.3%
		2001	470 360	67.4%
		2007	776 476	85.0%
	Total income – After SRMI2	1996	350 345	60.1%
		2001	506 896	72.7%
		2007	782 283	85.6%
IES	Total income – STC	1995	527 850	95.0%
		2000	460 572	71.9%
		2005/2006	659 229	72.2%
	Total expenditure – STC	1995	519 549	93.5%
		2000	458 867	71.7%
		2005/2006	751 153	82.2%
	Total income - COICOP	1995	495 411	89.2%
		2000	441 795	69.0%
		2005/2006	705 713	77.3%
	Total consumption - COICOP	1995	365 935	65.9%
		2000	324 026	47.8%
		2005/2006	531 386	58.2%
OHS	Total expenditure – No imputations	1996	190 111	32.6%
		1997	172 608	28.6%
		1998	151 399	24.6%
		1999	229 693	35.9%
	Total income – No imputations	1999	607 350	94.9%
	Total expenditure – After SRMI2	1996	195 845	33.6%
		1997	183 153	30.4%
		1998	161 717	26.3%
		1999	252 422	39.4%
	Total income – After SRMI2	1999	746 173	116.5%
LFS	Total expenditure – No imputations	2001	230 514	33.1%
		2002	264 065	36.9%
		2003	370 790	50.4%
		2004	417 062	52.4%
	Total expenditure – After SRMI2	2001	241 690	34.7%
		2002	280 567	39.2%
		2003	414 435	56.3%
		2004	443 144	55.6%
GHS	Total expenditure – No imputations	2002	212 412	29.7%
		2003	287 893	39.1%
		2004	267 470	33.6%
		2005	299 400	34.9%
		2006	312 736	34.2%
		2007	326 385	33.9%
		2008	461 528	46.7%
		2009	606 047	61.1%

Table A.6: Continued

Survey	Variable	Year	Amount (R million) (2000 prices)	As % of total income in the national accounts
GHS	Total expenditure – After SRMI2	2002	229 177	32.0%
		2003	308 977	42.0%
		2004	289 165	36.3%
		2005	312 468	36.5%
		2006	314 442	34.4%
		2007	334 237	34.7%
		2008	486 045	49.2%
		2009	612 482	61.7%
PSLSD	Total income	1993	334 531	65.3%
	Total expenditure	1993	297 679	58.1%
NIDS	Total income	2008	627 815	63.1%
	Total expenditure	2008	546 682	54.9%
AMPS	Total income	1993	336 394	65.6%
		1994	330 381	62.5%
		1995	333 057	59.9%
		1996	349 167	59.9%
		1997	347 982	57.7%
		1998	361 044	58.7%
		1999	360 573	56.3%
		2000	404 993	59.8%
		2001	406 077	58.2%
		2002	403 762	56.4%
		2003	444 193	60.4%
		2004	450 696	56.6%
		2005	485 001	56.6%
		2006	502 572	55.0%
		2007	552 266	57.3%
		2008	629 142	63.2%
		2009	589 559	59.4%