SOME FACTORS INFLUENCING THE COMPARABILITY AND RELIABILITY OF POVERTY AND INEQUALITY ESTIMATES ACROSS HOUSEHOLD SURVEYS[1]

Derek Yu
Department of Economics, Stellenbosch University

ABSTRACT

In order to evaluate the extent to which a country achieved the objectives of poverty and inequality reduction, up-to-date, reliable and comparable survey data is required. This paper critically reviews the issues which could affect the comparability of the datasets as well as the possible solutions, by investigating whether income or expenditure variable should be used for the analyses, whether the newly adopted diary approach in the Income and Expenditure Survey 2005/2006 is associated with more reliable capture of income and expenditure information compared with the conventional recall method, and if the respondents should be asked to declare the income and expenditure in exact amounts or the relevant intervals. With regard to the exact amount approach, further investigation is conducted to compare the single-estimation approach and the aggregation approach. In contrast, as far as the interval method is concerned, 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 discussed. In addition, it is argued that the survey data should be validated against external sources such as national accounts data in order to improve the reliability of the former data for the subsequent poverty and inequality analyses. Furthermore, since the survey data are, strictly speaking, not time-series data, it is argued that the data should be re-weighted by means of the cross entropy approach in order to be consistent with demographic and geographic numbers presented by the Actuarial Association of South Africa (ASSA) model and Census data in order to improve the reliability of the poverty and inequality estimates and trends. These issues are also discussed.

[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 is 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 and inequality 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 the TBVC (Transkei-Bophuthatswana-Venda-Ciskei) states from the sample.

Since the 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 conduct surveys which in turn provide alternative datasets for poverty and inequality 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 and inequality 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 and inequality estimates and trends, several factors could affect the reliability of the results as well as comparability of the results amongst the surveys. Firstly, whether income or expenditure should be used to measure poverty and inequality. 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 better capture of the income and expenditure information, and the subsequent poverty and inequality estimates.

In some surveys, the respondents were asked to report the exact amount, but only asked to declare the relevant income or expenditure category in others. Looking at the first method (reporting the exact amount) in greater detail, 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 former approach is not precise enough, while the opposing argument is that the latter 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. The other 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.

The last two issues relate to the validation of survey data against external sources, as well as the cross entropy re-weighting approach. First, it is argued that the survey data could be compared

with data from external sources in order to assess the accuracy of the former data, and it has always been argued that household surveys under-estimated income or expenditure, and hence the data should be adjusted (i.e., shifting the distribution) in line with the national accounts data. Secondly, 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 and inequality estimates and trends would be more reliable, if the data is re-weighted to be consistent with demographic and geographic numbers presented by the ASSA and census data by means of cross entropy approach (Branson 2009).

Hence, this paper attempts to discuss the aforementioned issues. Other factors that could also affect the reliability of poverty and inequality 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 and are not discussed.

2. Household surveys for poverty and inequality 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; (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. Household expenditure was derived by adding the respondents' answers on food spending, non-food spending and rent expenditure (i.e., aggregation approach), and also by asking the respondents to declare the 'one-shot' expenditure amount. On the other hand, 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'

[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).

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 was much lower[6], SALDRU eventually decided to use the income and expenditure variables derived by the aggregation approach to conduct poverty and inequality analyses in the official NIDS 2008 reports (e.g., Argent et al. & Finn et al. 2009).

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			

[5] For example, the response rate of the 'one-shot' expenditure question was only 79%.

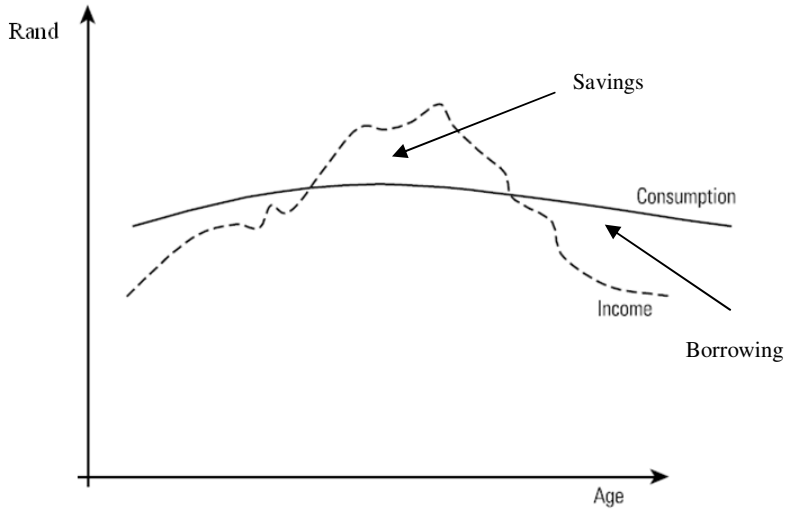
[6] Looking at households that reported both 'one-shot' and aggregated household expenditure, the former figure was only R237 364 million (2000 prices) while the latter figure was R466 683 million (2000 prices). In contrast, with regard to households that reported both 'one-shot' and aggregated household income, the former figure was only R429 590 million (2000 prices) while the latter figure was R627 815 million (2000 prices).

3. Factors affecting the reliability and comparability of poverty and inequality estimates

3.1 Income vs. Expenditure / Consumption

An important question that arises when using the money-metric approach to measure the standard of living, poverty and inequality of the population is whether income or expenditure / consumption should be used. The primary reason for some countries to use the consumption variable is that income typically rises and then falls in the course of a person's lifetime, in addition to fluctuating somewhat from year to year, whereas consumption remains relatively stable, since it could be smoothed to some extent by saving and borrowing (EPRI 2000; McKay 2000: 85-86; Duclos and Araar 2006: 21; Haughton and Khandker 2009: 24-25). This smoothing of short-term fluctuations in income is predicted by the permanent income hypothesis, under which transitory (temporary) income is saved, while long-term (permanent) income is largely consumed (See Figure 1). Hence, 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, it is argued that they are typically smaller than those of income, as consumption is less tied to seasonal and weather-related patterns in agriculture than is income (Deaton and Gross 2000: 93-94).

Figure 1: Lifecycle hypothesis – income and consumption profile over time



Secondly, the concept of consumption – giving money in exchange for a good or service – is clear both to interviewers and interviewees, while the income concept might not be clear (to be discussed later). Consumption is also held to be more readily observed, recalled and measured than income (at least in developing countries, although this is not always the case) (Deaton and Gross 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 deprivation (Haughton and Khandker 2009: 30).

Despite the advantages discussed above, using expenditure / consumption instead of income to measure money-metric poverty and inequality also has its drawbacks. First, there might be a need for collecting data on consumption on goods and services item by item. The number of consumption items could be as many as more than a thousand, while the income source items are

much fewer[7]. Secondly, although the respondents are more likely to remember consumption activities in more detail and to report higher spending 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. 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 Gross 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 (Guenard and Mesple-Somps 2010: 523), but longer recall periods might really be required for durable goods with low purchase frequency.

Households tend to under-declare what they have spent on luxury or illicit items, e.g., alcohol, tobacco, drugs. In addition, with regard to consumption on durable goods, as mentioned above, such expenses are not incurred regularly, so the data could be noisy because recall bias is more likely to happen with longer recall periods. Looking at the 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 (this is the case in all surveys under study), or whether only the change in the value of the asset during the year (i.e., depreciation, plus the cost of locking up one's money in the asset) should be included (Haughton and Khandker 2009: 25).

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, might be properly recorded as both income and consumption, but are often very difficult to value.

Given the pros and cons of using consumption, one might wonder if it is better to use income. However, there are arguments for and against its use. The main argument in favour of using income are that it is easier, cheaper and quicker to collect income data, especially in situations where income comes from one or two sources (e.g., wages and pension) that are easily recalled or for which independent documentation exists (Deaton and Gross 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, as a result of being subject to significant shocks. This is more likely to happen in households engaging predominantly in self-employment, 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 many visits to the household or dependence on the ability of households to remember their income from many months earlier (Deaton and Gross 2000: 93-94; McKay 2000: 84-86). In addition, as a result of the volatile nature of income, the reporting period might not be able to capture the 'average' 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 Gross 2000: 93-94). Respondents might not know how much income they make in these activities, in particular because of either seasonal variations, because income declarations are biased by non-responses and under-declarations, or

[7] 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, with the rest of the questionnaire trying to collect demographic information of the household members.

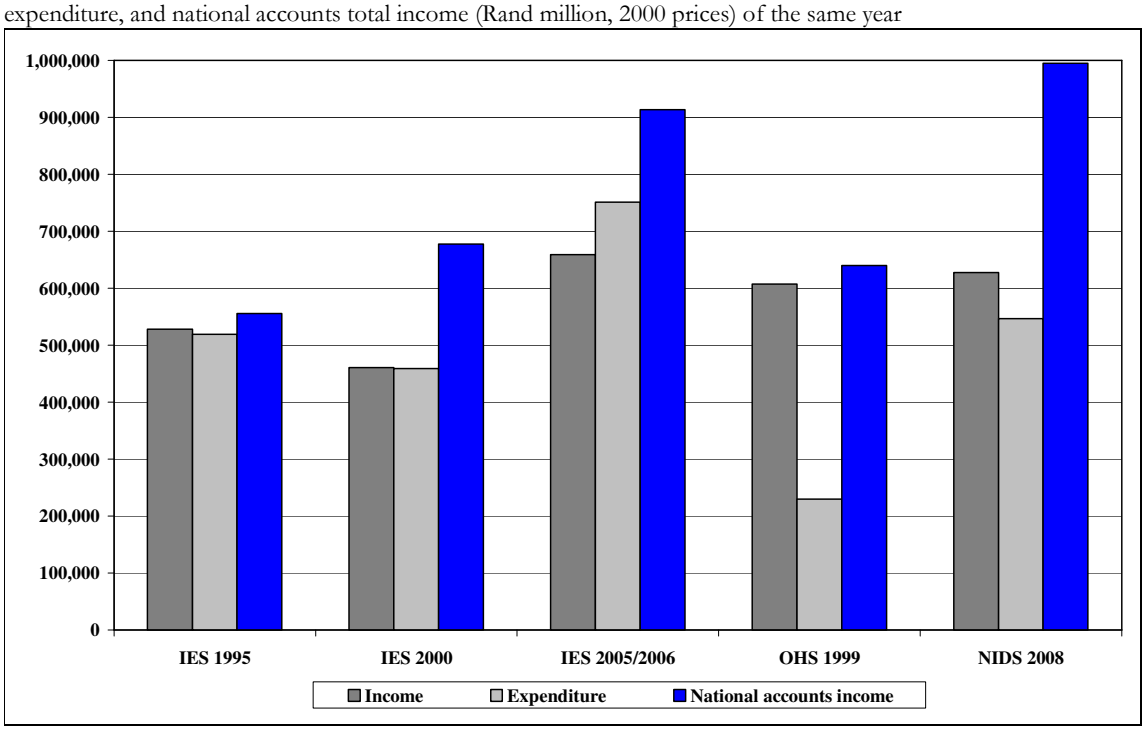
because they might not feel there is a need to report incomes earned infrequently or might not consider them as part of income, e.g., receipt of transfers and remittances (McKay 2000: 95; Haughton and Khandker 2009: 30; Guenard and Mesple-Somps 2010: 527).

It is also argued (Deaton 1997; Deaton and Gross 2000; McKay 2000; Posel and Casale 2005; Haughton and Khandker 2009) that respondents are more likely to lie 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 persuade respondents that the information they give will not be passed on to the tax authorities. As a result, income would be reported inaccurately or understated.

Some respondents might be reluctant to report income earned illegally, such as smuggling, corruption or prostitution, as well as 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 do not want to reveal that they are unsuccessful.

Figure 2 below shows the total income or expenditure of surveys that collected both information, and these amounts are compared with the national accounts total income of the same year, and it can be seen that, in all of these surveys except IES 2005/2006, income was greater than expenditure[8]. To conclude, it is clear that both merits and drawbacks are involved when using either income or expenditure variable to estimate money-metric poverty and inequality.

Figure 2: Total income or expenditure (Rand million, 2000 prices) of surveys that collected both income and expenditure, and national accounts total income (Rand million, 2000 prices) of the same year



[8] However, numerous other factors affect the accuracy of the capture of income and expenditure information, and they will be discussed for the remainder of Section 3.

3.2 Diary vs. Recall method

Regardless of whether income or consumption is chosen to measure poverty and inequality 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.

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 could happen, as the respondents might not remember many purchases long after they have been made. This would lead to either an under-estimation of consumption or to inaccurate guesses (i.e., respondents estimate their consumption over the whole year from their current rate of consumption) (Deaton 1997: 24-25; Deaton and Gross 2000: 109-110). Recall bias becomes more serious as the recall period increases.

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 accurate data is collected on high-frequency non-durable purchases, 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 period is required. In addition, the match between consumption and purchases is more accurate when averaged over a longer recall period (Deaton 2005: 16)[9].

The telescoping phenomenon – respondents tend to include consumption events that took place before the beginning of the recall period (Deaton and Gross 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 Gross 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

[9] For example, if the respondent is asked to declared 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 remembers the monthly food expenditure with good memory), and the resultant food expenditure could be more accurate.

likely to forget the date the consumption events occurred[10].

As a result of these drawbacks, the diary method becomes an alternative approach to collect income and consumption information. Corti (1993) argues that it is a reliable alternative to the traditional interview, recall method for events that are difficult to recall accurately or that are easily forgotten. Moreover, the diary method helps to reduce the problems associated with collecting sensitive information by personal interviews. For example, if 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.

A second advantage of the diary method is that it is more convenient to the respondents, as they could answer the questions at a time and place that are suitable for them (Deaton and Gross 2000: 119-122; Wiseman et al 2005: 395). Thirdly, 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[11]. Finally, the diary method is designed to minimize reliance 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 occur or close to that time (Deaton and Gross 2000: 109; Battinstin 2003: 2; Wiseman et al. 2005: 395).

The diary method is associated with various problems. First, diaries are less appropriate 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[12]. 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 events (Wiseman et al. 2005: 396). Hence, the data collected from the diaries might be biased towards the competent, literate diary keepers (Corti 1993), and a pictorial diary might be required to solve this problem.

Although the diary method reduces the amount of time that the interviewer needs to spend interviewing the households, this method might increase the time that the interviewer must spend travelling, as it requires extra trips to collect the completed diary. Moreover, considerable time might also be spent helping illiterate households fill out the diaries. The interviewers might 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, 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 Gross 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 analyze the sheer volumes of data, especially if the diary is unstructured, since intensive editing and coding will push up the costs

[10] For instance, 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 if asked to declare the expenditure on computer and telecommunication equipment in the past 12 months, this would result in the over-estimation of total expenditure.

[11] 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.

[12] 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

and involve even more time; respondents might be more co-operating and fill in the diaries more accurately, only if offered incentives or gifts.

It is argued by Deaton (2005: 16) and Wiseman et al. (2005: 399-400) 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 person spends. Moreover, 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 might be inaccurate, as some households have unusually low purchase rates in some items (e.g., every two weeks or every month, 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; recall method might work well to record the consumption of the more durable, bulky items with low purchase frequency (Deaton and Gross 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 former focusing on collecting information on income as well as semi-durable and durable goods consumption[13], and the latter 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 Gross 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 and clarify entries with the respondents, but the data would soon become retrospective and once again subject to recall bias (Wiseman et al. 2005: 395).

Finally, Corti (1993), Deaton and Gross (2000: 119-122), Wiseman et al. (2005: 395) and Ahmed et al. (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 shows higher reporting of consumption than the following days/weeks[14]. It could be explained by various factors: the novelty of diary keeping wears off as

[13] This implies that the inaccuracy problem 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 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.

[14] In IES 2005/2006, the respondents were asked to keep the diaries for four weeks. However, Stats SA did not release the weekly expenditure data from the diary approach, so the extent of the first-day effect was not known.

time goes by; the respondents feel exhausted to keep records and eventually become less thorough in their reporting; the diary keepers no longer carry their diaries with them. 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.

Figures 3 and 4 conclude by presenting the information on food and transport expenditure in the three IESs. It can be seen from Figure 2 that food expenditure was clearly lower in IES 2005/2006. 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. 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?

Figure 3: Food expenditure in the IESs (Rand million, 2000 prices)

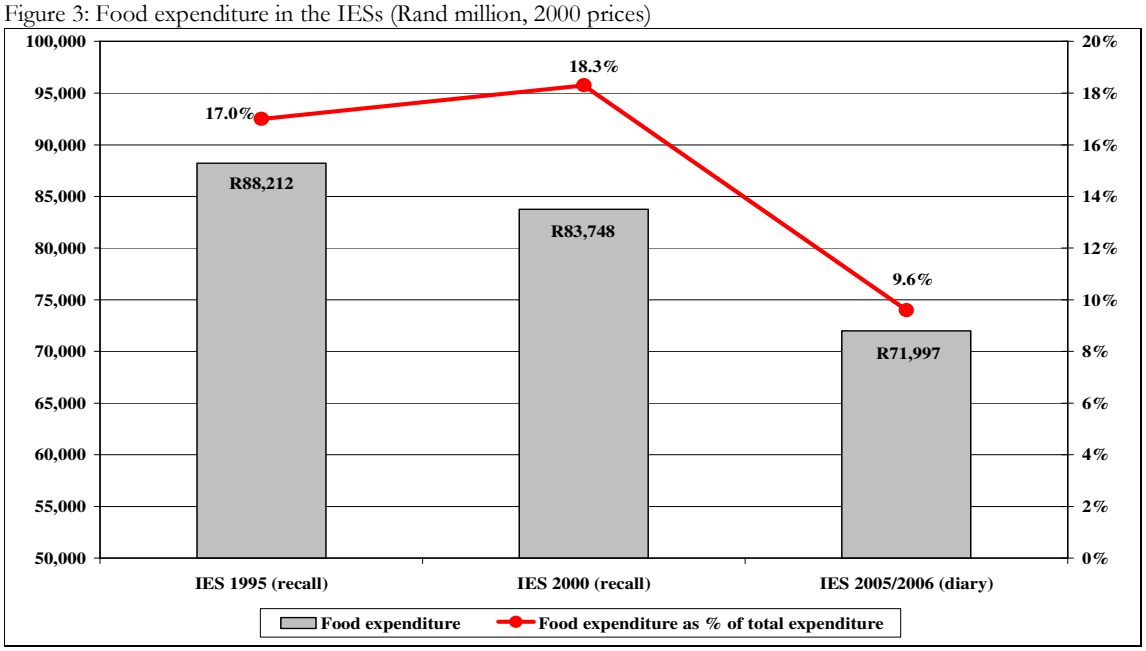
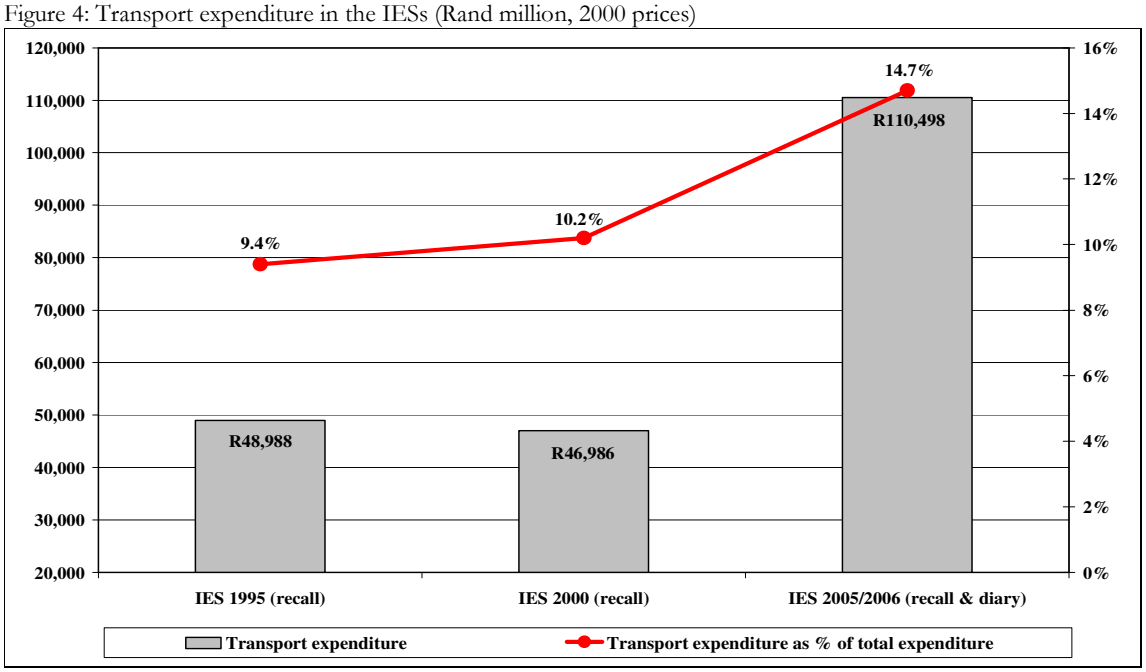


Figure 4: Transport expenditure in the IESs (Rand million, 2000 prices)



3.3 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 / consumption categories, might prefer the anonymity of indicating to what predefined income / consumption 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[15]. 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 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 to option to declare the exact amount, he 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.

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

If income and expenditure information is to be collected by asking 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' overall amount (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).

[15] 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

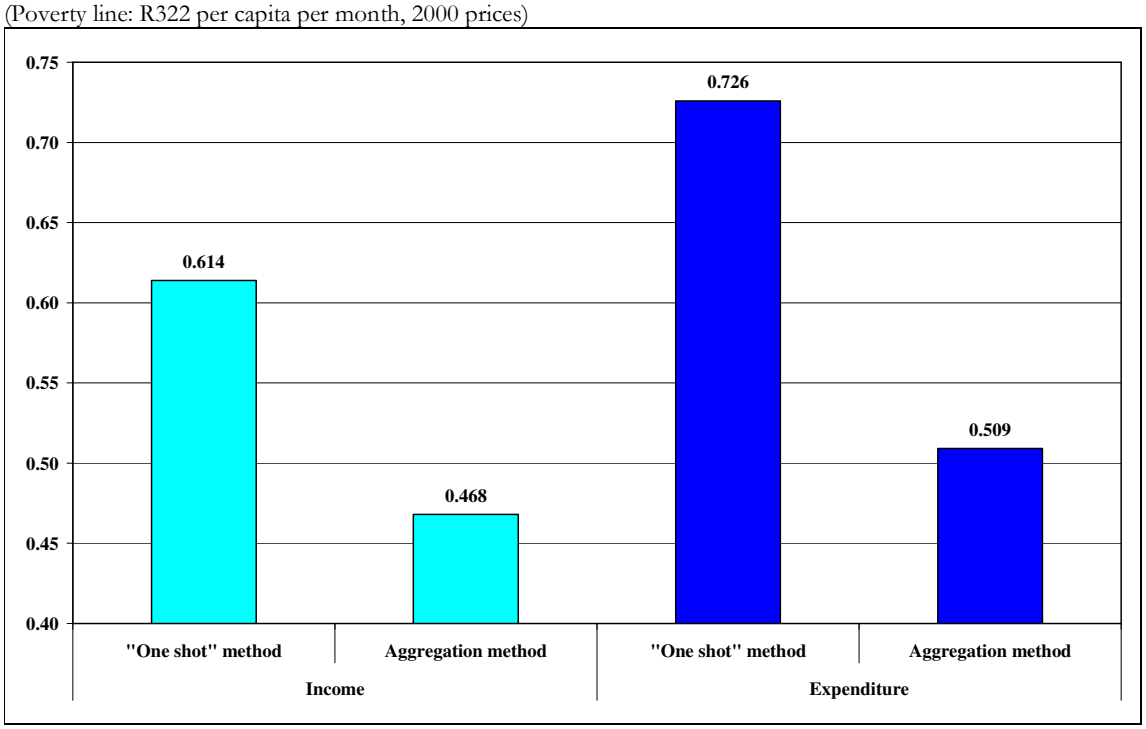
The 'one-shot' amount, single estimate approach, despite being a relatively less time-consuming and costly method to collect the required information, could confuse the respondents, as they be unsure about what items should be included as part of the total income or expenditure. This results in low response rate, and/or under-reporting of total income or expenditure (Deaton 1997: 27; Browning et al. 2002: 7-10). Hence, it is argued that it is necessary 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 providing misleading amounts and even 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 'one-shot' amount method.

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. It 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 find it disturbing to reveal income information even if asked to declare the 'one-shot' amount.

As mentioned in Section 2, 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 5).

Figure 4: Poverty headcount ratios using per capita income and expenditure (2000 prices) variables of NIDS 2008 (Poverty line: R322 per capita per month, 2000 prices)



3.5 Approximation of amount in each band

If the income or expenditure information was collected in bands, the data needs to be made continuous before it could be divided by household size to derive the per capita income or expenditure variable required for poverty and inequality 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.

3.5.1 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 the AMPS 2000 declares its nominal monthly household income falls in the "R5 000 – R5 999" category, the income amount is derived as R5 500. As far as the top category is concerned, since no upper bound exists, it is assumed that the mean exceeds the lower bound 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 equal to R22 000 (R20 000 × 1.1). Although this method lacks theoretical backing (Whiteford and McGrath 1994: 28), it may be attractive because its simplicity.

3.5.2 Midpoint-Pareto method

As the lower income categories are narrow, Whiteford and McGrath (1994: 29) argue that the distribution of income at the bottom end is not markedly influenced by midpoint imputation. However, as greater skewness within groups becomes evident in the higher income categories, a parametric approach is necessary. A Pareto mean is 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 bound 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 the following 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 income bounds of the interval;
- The Pareto coefficient is also used to calculate the mean of the open interval by means of the following equation: $\bar{x} = \left[\frac{\alpha}{\alpha+1}\right] \cdot x_{\infty}$, where x_{∞} represents the lower bound of the open interval.

The midpoint-Pareto method is applied in either of the following ways: (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; (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 the aforementioned category. In the South African studies, (1) is the commonly used approach.

3.5.3 Interval regression

Interval regression tries to predict the income (or expenditure) amount from some well chosen explanatory variables, such as educational attainment, age, gender, race, labour market status of household head, household size, number of employed members in the household, etc. However, instead of the precise income amounts, only the data on the household income range is available. The extreme values of the categories are interval-censored (i.e., each interval is both left- and right-censored), with the exception of the open interval, that is only left-censored. These extreme values must be specified in the interval regression, before the model could predict what income / expenditure each household will earn / spend based on the explanatory variables used.

3.5.4 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 interval. If it is assumed that 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): $Y_{ij} = x_i + sign_{ij} \times U_{ij}(0, x_i - lowerbound_i)$, where Y_{ij} is the new random midpoint income value for income level i and household j, $j = 1,2,...,f_i$, $sign_{ij}$ 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, and U_{ij} is the uniform distribution, with lower bound of 0 and upper bound of $x_i - lowerbound_i$), where $lowerbound_i$ means the lower bound 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 bound of this interval (i.e., $lowerbound_2$) is R400. Assuming $sign_{ij}$ is –1 for this household, and a random draw from the uniform distribution (lower bound and upper bound being 0 and 200 respectively) gives an amount of R50, then the household expenditure amount is derived as: $600 + (-1) \times 50 = \text{R}550$. Similarly, using the same information but if $sign_{ij}$ is +1 for this household, the household expenditure is calculated as: $600 + (+1) \times 50 = \text{R}650$.

Having discussed the various methods to derive the income / expenditure mean of each interval, one might raise questions on 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. In South Africa, Von Fintel (2006), who mainly looked at the LFS 2003 September data, found that the accuracy of the indicators obtained using either a continuous income variable or a categorical income variable were equally accurate, when using a dataset containing both variables. Furthermore, Malherbe (2007) applied the Census 2001 income intervals to the IES 2000 data, and applied the midpoint method, interval regressions method and random midpoint method to derive the amount in each category. At the end, Malherbe found that the poverty and inequality estimates were very similar in the continuous and midpoint datasets, while the interval regressions and random midpoint method obtained different results.

3.6 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 and inequality estimates would be heavily influenced by the number and width of the income / expenditure bands of the survey concerned. From Tables A.1-A.3 in the Appendix, it could be seen the number of bands is as low 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). Therefore, 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) equals to R9 600, R7 500 and R8 500 respectively. It is obvious that these three amounts are different; in this case the AMPS amount (R8 500) is closest to the original amount (R8 200). Is the reliability of the derived amount being influenced by the number and width of bands in each survey? Would the poverty and inequality estimates be over-estimated or under-estimated as a result of these two factors?

Studies that looked at the impact of the aforementioned issues on the poverty and inequality estimates are not found, except Seiver (1979), who 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. His study did not investigate the impact of the number and width of intervals on poverty.

3.7 Households with zero or unspecified income

A serious problem of some surveys is that a high proportion of people reported zero or unspecified personal income, which subsequently resulted in a large proportion of households with zero or unspecified household income. This problem is most serious in the two censuses and CS 2007: the proportion of households with zero income was 13.0% in 1996, 21.0% in 2001 and 8.2% in 2007, while the proportion of households with unspecified income was 11.5%, 16.4% and 11.1% respectively.

Regarding the households with missing household income, Ardington *et al.* (2005) argue that if those with missing data fall disproportionately in the bottom of the income distribution, then levels of poverty will be under-estimated if they are ignored. In contrast, if non-response is higher among the wealthy, measures of inequality are likely to be biased downwards[16]. Furthermore, with regard to the higher proportion of households with zero household income, even allowing for South Africa's high unemployment rates, it is highly unlikely that most of these zero income households had no working-age members earning any income. Hence, if these zero-income households are accepted, this could lead to an over-estimation of measured poverty and inequality.

Hence, when analyzing poverty and inequality, unless the data is missing completely at random (MCAR)[17], ignoring households with unspecified household income would lead to biased results.

---

[16]

[17] With regard to missing data, there are three types of mechanisms, whereby (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

Besides, including households that might incorrectly report zero income might lead to over-estimation of poverty and inequality levels. In general, the four main methods to deal with missing data in general are casewise deletion, available-case deletion, single imputation and multiple imputation. Each method is discussed in greater detail.

3.7.1 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 whose information is complete (Lacerda et al. 2005: 11). Thus, in the case of household income data (or expenditure / consumption), as long as households did not specify the household income amount or category (depending on how the question was asked), the households are immediately excluded from further analyses. However, as mentioned at the beginning of this section, if these households are ignored, it would have serious impact on the reliability of poverty and inequality estimates.

3.7.2 Available-case deletion

Available-case deletion is an extension of casewise deletion, but differs in that it only excludes those cases for which data is missing on the variables necessary to estimate the parameters of interest (Lacerda et al. 2005: 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 and inequality estimates as caused by casewise deletion. Thus, it seems the abovementioned two methods are the best solution to deal with missing data for the purposes of this dissertation.

3.7.3 Single imputation

Imputation aims to provide reasonable estimates of the missing data, instead of simply ignoring 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. 2005: 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 average of the observed values for that variable (Lacerda et al. 2005: 15)[18], while cell mean substitution aims to divide respondents into cells on the basis of some known variables, and the mean values within these cells are used for imputation (Lacerda et al. 2005: 15 & Malherbe 2007: 29)[19]. Moreover, hot deck imputation involves substituting missing values with observed values drawn from similar responding units (Lacerda et al. 2005: 16)[20], while cold deck imputation involves substituting

[18] 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. Hence, the household income of the 10% of households with unspecified income was assumed to be R1 500.
[19] 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 is assumed to earn R1 600.
[20] For instance, using the example as mentioned in footnote 14, households are divided into cells by race and gender of household head. After a random draw on a household headed by a white male, it is found that this household's income is R2 000. Then it is assumed that 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

missing values with a constant value from an external source (Lacerda at el. 2005: 16)[21]. Furthermore, 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. 2005: 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. 2005: 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 head 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 calculated 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.

3.7.4 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. 2005: 17-18):

- X (> 1)[22] plausible versions of the complete data are created by imputing each missing value X times using X independent draws from an appropriate imputation model, conditional on the observed data;
- The X imputed datasets are then each treated as if they are entirely observed and analyzed individually by standard complete-data methods;
- The results from the X analyses are combined in a single and appropriate manner so as to obtain overall estimates and standard errors that reflect both sample variation and uncertainty associated with the imputed values.

A particular multiple imputation technique developed by Raghunathan *et al.* (2001) could be applied when data is missing at random (MAR) but not MCAR, namely sequential regression multiple imputation (SRMI). The SRMI method could be summarized as follows (Ardington *et al.*, 2005: 8-11; Lacerda *et al.* 2008; Vermaark, 2008: 2-3):

- The variables used in the imputation model are arranged from those with the least to those with the most missing values.
- Let the matrix X represent 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 ordered with respect to the extent of missing data they contain.

---

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 is imputed.

[21] For example, if a household 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, each household paid R1 500 personal income tax, then it is assumed that the IES 2000 household as mentioned above spent R1 500 in the last 12 months to pay personal income tax to SARS.

[22] Note that if X = 1, it stands for single imputation.

- Y_1 is regressed on X, and values for Y_1 are imputed using random draws from the appropriate predictive distribution for Y_1. For example, a normal OLS regression model is used when Y_1 is a continuous variable (e.g., earnings amount). However, a Poisson model is used when Y_1 is a count variable (e.g., age), a logistic model is used when Y_1 is binary (e.g., gender), a multinomial logistic model is used when Y_1 is a nominal categorical variable (e.g., province), and an ordered logistic model is used when Y_1 is an ordinal categorical variable (e.g., household income category).
- 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 imputed Y_1, and values are imputed for Y_2.
- This imputation goes on until all Y variables have been imputed using all previously imputed variables as covariates.
- The entire procedure is then repeated m times (i.e., m stands for the number of imputations), to produce m imputed complete datasets.

Yu (2009) is a recent study that adopted the SRMI approach to investigate the poverty and inequality on the two censuses and CS 2007. It was found that (See Table 3 below), after the application of SRMI, poverty headcount ratios and Gini coefficients 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; Gini coefficient increased rapidly between 1996 and 2001, before a moderate decreased took place in 2007.

Table 3: Poverty headcount ratios (Poverty line: R322 per capita per month, 2000 prices) and Gini coefficients, using per capita income (2000 prices) variables of Census 1996, Census 2001 and CS 2007 before and after the application of SRMI

	[1]: Before SRMI	**[2]: After SRMI**	**Difference: [2] – [1]**
Poverty headcount ratio			
Census 1996	0.606	0.576	-0.030
Census 2001	0.670	0.592	-0.078
CS 2007	0.529	0.463	-0.066
Gini coefficient			
Census 1996	0.742	0.694	-0.048
Census 2001	0.825	0.756	-0.069
CS 2007	0.774	0.743	-0.031

3.8 External validation to improve the reliability of survey data

In order to determine the reliability of the survey data, it is argued that the data should be validated against various external sources. These sources are discussed in this section.

3.8.1 Validation against national accounts

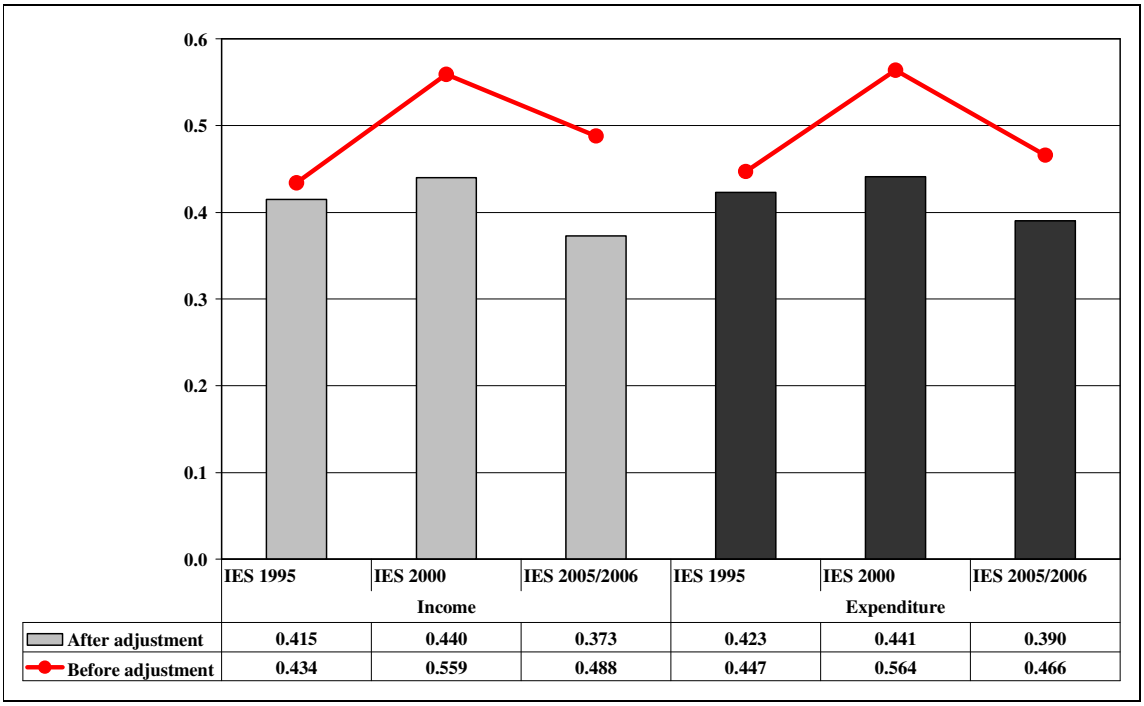
As mentioned before, surveys are more likely to under-estimate income / expenditure / consumption, due to reasons like fatigue, loss of interest, lack of motivation, illiteracy, recall bias, telescoping, refusal to disclose sensitive information, and tendency to declare zero or unspecified income, even if the households contain employed members that are likely to earn income or have income support from non-labour sources. Hence, 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.

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 is, 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. An example is the rapid decline of income and expenditure between IES 1995 and IES 2000. The magnitude of the decline is even greater than the fall in output during the Great Depression. Hence, the poverty and inequality estimates of the two IESs are argued to be totally incomparable, if the data are left unadjusted.

Figure 6 presents the poverty headcount ratios of the three IESs with and without the adjustment of survey means in line with the national accounts mean. Using the original unadjusted data, the poverty headcount ratio increased rapidly between 1995 and 2000 before a decline took place between 2000 and 2005/2006 (the red lines of Figure 6). The former finding was mainly attributed to the serious under-estimation of income and expenditure in 2000. After the adjustment, although the trends remain the same, it could be seen that (the column charts of Figure 6) the increase of poverty between 1995 and 2000 was only moderate.

Figure 6: Poverty headcount ratios (Poverty line: R322 per capita per month, 2000 prices) using per capita income and expenditure (2000 prices) variables of IES 1995, 2000 and 2005/2006 before and after adjustment of survey means in line with national accounts means



	IES 1995	IES 2000	IES 2005/2006	IES 1995	IES 2000	IES 2005/2006
	Income			Expenditure		
After adjustment	0.415	0.440	0.373	0.423	0.441	0.390
Before adjustment	0.434	0.559	0.488	0.447	0.564	0.466

However, it is also possible that adjusting the survey means would create negative effects on the reliability of poverty and inequality estimates. First, it is argued (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 measure of consumption in household survey is based on self-reported expenditures (e.g., cash and from own stock) by the households in the interviews, households are treated as 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.

Secondly, the national account estimates of consumption implicitly include spending by

unincorporated businesses and non-profit organizations, as such religious groups 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 improvements 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). This would, for example, imply that campaign spending by politicians trying to get elected would help reduce poverty, even if none of the spending would go to the poor.
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., unit non-response takes place), 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 the rich households take part in the survey, the included rich people are more likely to understate their income / consumption more than the included poor do, and this implies that inequality is under-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, 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, if 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 (which is greater than the survey mean) might not help improving the survey poverty and inequality estimates, if the unreliable survey distribution is the root of the problem but is not corrected.

It might also be argued 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, and as a result of failing to include these poor households to take part in the survey, the survey income / consumption estimates would be biased upwards. Hence, 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 and inequality 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 the replacement of survey means by national accounts means does not improve the poverty and inequality estimates at all, and might even worsen them, if the issues relating to the survey weights are not sorted out right at the beginning.

The 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, it is argued that items like imputed rent and in-kind income are included in the national accounts income and private consumption estimates, but might not be recorded in household surveys, and this eventually results in the differences between the two series[23].

3.8.2 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 4 below shows that, in general, the IES 2000 and IES 2005/2006 did a decent job to capture social grants income, despite the fact that disability grant income was under-captured.

Table 4: Social grants income of IES 2000 and 2005/2006 compared with social grants expenditure of National Treasury (Rand million, nominal terms)

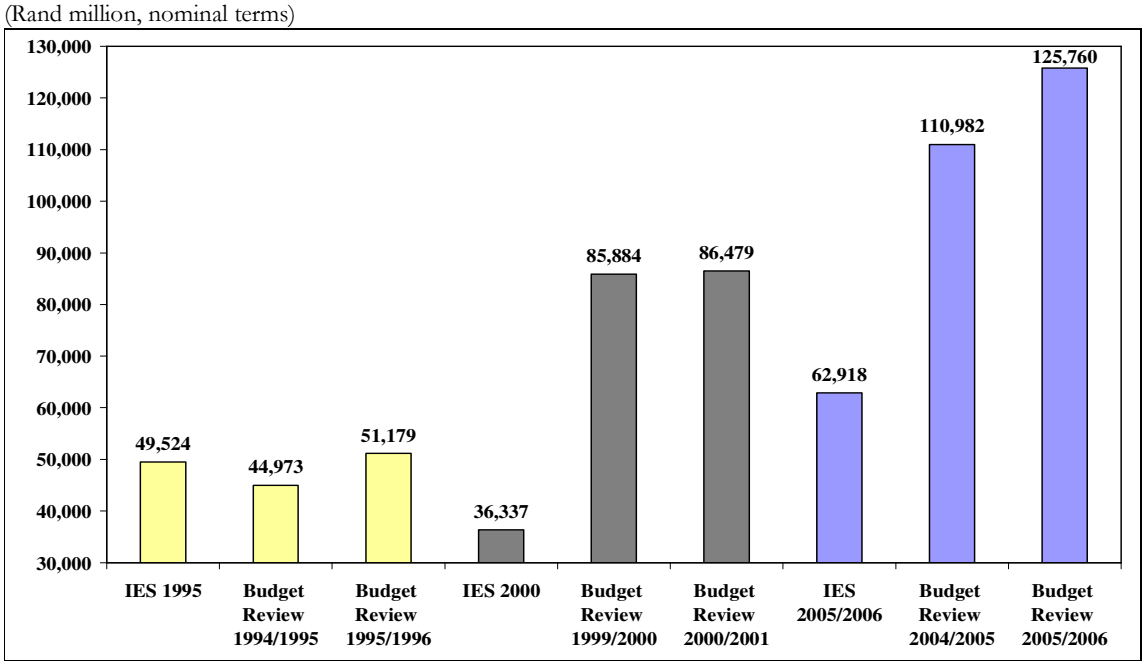
	Old-age/War pension	**Disability grant**	**Child/Family/Other grants**
[A]: IES 2000	R15 402	R3 058	R1 533
[B]: Treasury - 1999/2000	R11 660	R3 823	R944
[C]: Treasury - 2000/2001	R12 208	R4 066	R1 770
[A] / [C]	132.1%	80.0%	162.4%
[A] / [C]	126.2%	75.2%	86.6%
[D]: IES 2005/2006	R25 301	R10 375	R19 981
[E]: Treasury - 2004/2005	R18 540	R12 570	R13 774
[F]: Treasury - 2005/2006	R20 025	R14 438	R17 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 could be compared with the net personal income tax revenue received by SARS, and from Figure 7, it can be seen that IES 1995 did an outstanding job to capture this tax expenditure accurately. 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. In fact, the under-estimation of the tax expenditure in IES 2000 could be one of the reasons to account for 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 less serious (about 57% of the income tax revenue of SARS as reported in the 2004/2005 and 2005/2006 budget).

[23] IES 2005/2006 and NIDS 2008 are the two questionnaires 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.

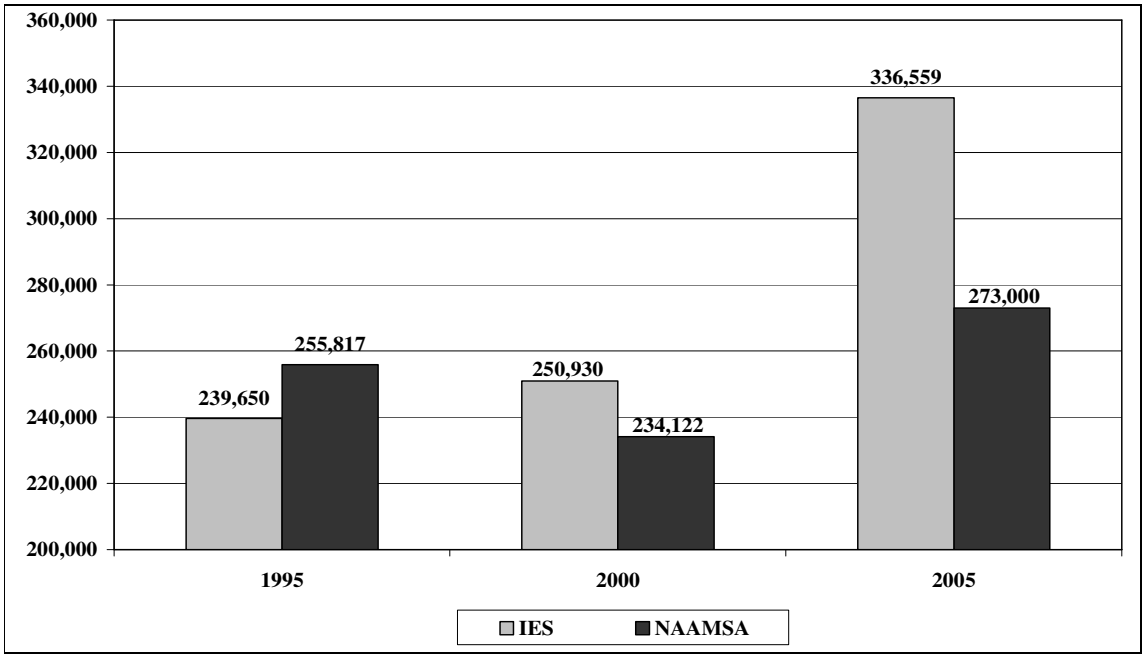
Figure 7: Net personal income tax expenditure of IESs compared with net personal income tax revenue of SARS (Rand million, nominal terms)



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

In the three IESs, the 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 8 show that the IES 2005/2006 over-estimated the number of new motor vehicle purchases.

Figure 8: 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.

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 5 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, same as the personal income tax expenditure, was seriously under-estimated, when compared with external sources.

Table 5: 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 is compared with the aggregate of SAPIA's 1994Q4, 1995Q1, 1995Q2 and 1995Q3 data, the IES 2000 data is compared with the aggregate of SAPIA's 1999Q4, 2000Q1, 2000Q2 and 2000Q3 data, and the IES 2005/2006 data is compare with the sum of SAPIA's 2005Q4, 2006Q1, 2006Q2 and 2006Q3 data.

3.9 Post-stratification weighting

With the exception of Census 1996 and Census 2001, the remaining data used for poverty and inequality analyses in this study is 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, but 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 weigh 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 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 is out-of-date 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, temporary consistency is not considered. This creates problems when the data is 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 adopted to re-weigh the person weights of all the data under study, with the person weights being adjusted to conform to the race, gender and age distribution of the population estimates as calculated by the Actuarial Association of South Africa 2003 (ASSA 2003) model. Branson argues (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, mortality, as well as HIV/AIDS prevalence rate. There are 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. For the remainder of the study, 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 $\sum_{k=1}^{K} p_k x_t = y_t \qquad t \in [1,...,T]$ 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 which can be written as:

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

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 the labour market, poverty and inequality 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 after the entropy approach was conducted (the study did not look analyze the poverty and inequality trends). She re-weighted the person weights of OHS 1995-1999 and the March LFSs in 2000-2004[24]. 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. Her study did not investigate the impact of the CE approach on poverty and inequality estimates and trends.

[24] 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.

4. Conclusion

This paper first discussed the pros and cons of using income and expenditure (consumption) for poverty and inequality analyses. Although the general conclusion is that expenditure is the preferred variable to be used in developing countries, further investigation shows 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 discussed, 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, and 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 and inequality analyses; the impact of the number of bands and width of each band on the poverty and inequality estimates; and how to deal with households with zero or unspecified income or expenditure. At the end, it was found that the midpoint-Pareto method was most appropriate to make the interval data continuous, there is insufficient research done both domestically and internationally that investigate how the number and width of bands affect the poverty and inequality estimates, and the sequential regression multiple imputation (SRMI) approach is 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 and inequality 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.

References

Ahmed, N., Brzozowski, M. and Crossley, T.F. (2005). *Measurement errors in recall food expenditure data*. SEDAP Research Paper No. 133. Hamilton: The Program for Research on Social and Economic Dimensions of an Aging Population (SEDAP).

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.

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

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.

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

Economic Policy Research Institute (ERPI) (2001). *Impact of social security system on poverty in South Africa*. EPRI Research Paper No. 19.

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

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). 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.

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). 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.

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.

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

Appendix

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+		