ANNUAL FORUM 2005

Trade and Uneven Development: Oppo rtunities and Challenges

Who replies in brackets and what are the implications for earnings estimates? An analysis of earnings data from South Africa

Dorrit Posel & Daniela Casale

"Who replies in brackets and what are the implications for earnings estimates? An analysis of earnings data from South Africa"

Dorrit Posel & Daniela Casale

University of KwaZulu-Natal, Durban, South Africa

(posel@ukzn.ac.za and casaled@ukzn.ac.za)

Abstract

In household surveys, earnings data typically can be reported as point values, in brackets or as 'missing'. In this paper we consider South African household survey data that contain these three sets of responses. In particular, we examine whether there are systematic differences between the sample of the employed with earnings reported as point values and those with earnings responses in brackets; we compare five different methods of reconciling bracket and point responses so as to generate descriptive measures of earnings; and we investigate empirically how earnings measures differ by approach.

1. Introduction

This paper interrogates information collected in household surveys on earned income. Household surveys provide a key source of information about individuals, the work that they do, and the households in which they live. A primary concern with survey data, however, is that respondents may not be willing or able to provide information about themselves or others in the household (cf. Hawkins, 1975; DeMaio, 1980; Moore, 1988; Duncan & Petersen, 2004).

Item non-response in surveys has been observed particularly for questions about earnings and wealth (Atkinson & Mickelwright, 1983; Lillard, Smith & Welch, 1986; Juster & Smith, 1994 and 1997; Moore & Loomis, 2001; Riphahn & Serfling, 2004). Individuals may be reluctant to disclose information about how much income is earned partly because of confidentiality or privacy concerns. Another reason for not providing earnings information is that respondents do not know 'exactly' how much they (or others in the household) earn.

The introduction of earnings brackets in household surveys helps overcome these problems of disclosure (Juster & Smith, 1997). Respondents may be more willing to report income in brackets, and brackets 'permit' respondents to report with a margin of error. Prompting reticent respondents with earnings brackets therefore typically reduces the number of missing values for earned income. But it also means that two (non-overlapping) sets of earnings information are collected – point values and bracket responses.

Our paper investigates these two groups of earnings data using information gathered in a nationally representative South African household survey (the September 2002 Labour Force Survey). The paper has three main objectives. First, it explores whether the sample of the employed whose earnings are reported in brackets is systematically different from the full sample of the employed, and more specifically from those whose earnings are reported as point values. Second, it examines different methods of reconciling bracket and point data on earnings, particularly when the two sets of responses are not randomly distributed across the employed. Third, the paper considers

1

empirically the robustness of national estimates of earnings, of the working poor and of earnings inequality, to the treatment of earnings in brackets.

The survey data that we analyse clearly show bracket responses to be a non-random sample of all the earnings responses. One set of differences between bracket responses and point values reflects the respondent's knowledge about income earned. Individuals are more likely to report earnings in brackets when they are proxy- rather than self-reporting, and when there are more people in the household about whom information must be provided. Another key variable affecting the probability of reporting a bracket is income: individuals are significantly more likely to provide a bracket response when earnings fall in the lower, and particularly the upper, ends of the distribution. This finding mirrors the more general result reported in the literature, that non-response to questions on income and wealth is concentrated in the tails of the income distribution (Lillard et al, 1986; Juster & Smith, 1994; Riphahn & Serfling, 2004).

The distribution of bracket responses has important implications for imputing point values for these earnings. We would expect to derive 'better' imputed earnings values for bracket responses when more information is taken into account. We find particularly that summary measures for imputed earnings when selection information is *not* used (in an ordinary least squares regression on all reported point earnings values) are significantly different to estimates derived using a Heckman selection model. In addition, we find that controlling for selection produces estimates for bracket responses that are highly consistent with estimates derived using bracket information (from the simplest of the bracket midpoint to estimating multivariate earnings equations for each bracket based on the point values in that bracket).

Our analysis also shows that, out of the sample of initial non-respondents, that is, all those who did not provide a point value at first, the bracket responses are a distinct sub-group. The characteristics of those whose earnings are reported in brackets on average are significantly different from those with no earnings information. We therefore also briefly consider the implications of our findings for imputing point earnings values for the employed with missing earnings information.

 The paper is structured as follows. In the next section, we briefly review the survey literature on item non-response, particularly on questions about income. In section

3 we present the data that are used for the study and in section 4 we compare the samples of the employed, focusing on those whose earnings are reported as point values and those with earnings reported in brackets. We investigate different methods of imputing point earnings values for bracket responses in section 5. In section 6 we explore the implications of our findings for deriving summary measures of earnings estimates.

2. Income non-response: Background

There are a number of problems that can undermine the quality of data collected in surveys. Individuals may refuse to be interviewed and samples drawn therefore may not be representative. If interviewed, respondents may not provide accurate information and data collected may be subject to measurement error and recall bias. A further problem concerns item non-response, where individuals are not willing or able to answer certain questions, and information collected is therefore incomplete.

Item non-response has been documented in the economics literature principally for questions on earnings and wealth. Key areas of investigation in this literature concern the characteristics of non-respondents and possible determinants of response probabilities (cf. De Maio, 1980; Bell, 1984; Juster & Smith, 1994; Riphahn & Sefling, 2004), as well as how to reduce the number of missing data through survey design and implementation (cf. Lillard et al, 1986; Heeringa, Hill & Howell, 1995; Juster & Smith, 1994 & 1997; Moore & Loomis, 2001).

A general finding in the empirical literature is that non-response to income questions is not random. Rather, there are clear correlates of income non-response, the most well-documented being income itself. Non-response rates have been found to be significantly larger for high and low income earners, giving rise to a U-shaped relation between income and the probability of income non-response (Lillard et al, 1986; Juster & Smith, 1994).

Non-response can be attributed to two broad reasons: either a respondent knows the information but is not willing to disclose it; or a respondent does not know and therefore cannot say. There are a number of factors that may explain why the proportion

of both 'not willing' and 'not knowing' respondents would be higher in the tails of the income distribution.

Low-income earners may not want to disclose income information because they do not want to reveal to the interviewer that they are not successful. High income-earners may not want the interviewer to know just how successful they are (Riphahn & Serfling, 2004). High-income earners may particularly fear "governmental or other uses of the data" (Lillard et al, 1986:492).

Among willing respondents, difficulties in providing exact values for income earned may also be more pronounced among very low and very high income earners. Where income sources are irregular or sporadic (as may be expected among the survivalist self-employed, for example) or diverse (as among the professionally self-employed), a greater "cognitive requirement" (Riphahn & Serfling, 2004: 4) of providing information will reduce response rates.

Another correlate of income non-response that has been identified, particularly in the earlier empirical literature on surveying, concerns whether or not respondents are self-reporting. Where household surveys rely on information provided by principal (or single) respondents, we would expect respondents to be more knowledgeable about their own income than about the income of others in the household. This may explain why item non-response has been reported to be significantly higher for proxy-reporting than for self-reporting (cf. Coder, 1980; and Moore, 1988 for a review of early studies).

A simple survey extension, which has been found to reduce substantially the extent of income non-response, is the introduction of "follow-up brackets" (Juster & Smith, 1997: 1286). Respondents may be more willing to disclose income in brackets, particularly among the highest income-earners where the top income bracket is open-ended. Brackets also signal to respondents that even if they do not know the exact amount of income, "an approximation constitutes a legitimate response" (Duncan & Petersen, 2004:4).

Much of the literature identified above focuses on income non-response in general. In this study we examine particularly the sample of employed whose earnings are reported in brackets. Using household survey data for South Africa, we show that the sample of bracket responses displays characteristics typically associated in the literature

with income non-response. But we also show that bracket responses constitute a distinct sub-sample of all initial non-responses.

The focus of our analysis lies in comparing the sample of bracket responses and the sample that reported point values for earnings. A primary motivation of the study is to consider how best to reconcile point and bracket information on earnings in order to derive reliable summary measures of earnings. Most studies of earnings and labour market outcomes in South Africa convert earnings in brackets to point values by assigning to bracket responses the midpoint of their respective bracket (cf. Hofmeyr, 2002; Casale, Muller & Posel, 2004; Kingdon & Knight, 2004; Meth & Dias, 2004; Vermaak, 2005). We consider this, as well as other methods of imputing point values for the earnings brackets, particularly in light of our finding that bracket responses are a non-random sample of the employed with earnings information.

3. Data

The first nationally representative, comprehensive household survey in South Africa was undertaken in 1993. In this survey, the Project for Statistics on Living Standards and Development, respondents were not given the opportunity to report earnings information in brackets. Of all those in the sample who were reported as being employed, approximately 14 percent had missing data for earnings.

In all subsequent national household surveys conducted in the country, bracket responses have been introduced. Individuals first have been asked how much income they (or another household member) earned, and if they did not know or refused to answer, they were then prompted by being shown a set of earnings brackets. As expected, the non-response rate on income earned has fallen. In the September Labour Force Survey of 2002 (LFS 2002:2)[1], the survey used in this study, only seven percent of the employed are without any earnings data.

---

[1] The Labour Force Survey, which was introduced in 2000, is a biannual survey conducted by the national statistical agency (Statistics South Africa). Approximately 30 000 households in South Africa are interviewed - in the September 2002 survey, this amounted to just over 100 000 individuals.

In the LFS 2002:2 about 25 000 individuals in the sample were reported as being employed. Where respondents would not, or could not, provide a point value for the earnings of the employed, they were shown 14 categories of earnings, the lowest bracket representing zero earnings and the highest, open-ended, bracket representing earnings of more than 30 000 rands a month. Respondents were also given the opportunity to report that they did not know the earnings bracket, or to refuse to provide this information.

A particular problem that arises when analysing employment and earnings data collected through household surveys in South Africa concerns the significant number of individuals who are reported as being employed (if for only one hour of the previous week), and whose earnings are reported not as missing, but as zero. This earnings information has been recorded in the lowest earnings bracket representing zero income. In the LFS 2002:2, for example, over 1 000 of the approximately 7 000 employed with bracket responses for earnings are reported as being in the lowest bracket. Furthermore, there are *no* individuals for whom zero income has been recorded as a point value. This would suggest that these data are not a true sample of bracket responses for zero income, but rather that zero income has been captured or post-coded *only* as a bracket response.[2] We have therefore excluded all the employed with zero income from our overall sample.

This leaves us with a complete set of household- and individual-level data for 22 094 employed individuals for whom either a point or a bracket earnings response has been reported in the 2002 sample. Of these, approximately 26 percent (5 728 individuals) have earnings values reported in brackets. For a further 1 637 individuals who are reported as being employed, information on earnings is missing altogether.

4. Whose income is reported in brackets?

In this section, we describe the samples of the employed according to whether, and which, earnings information is provided. We then investigate econometrically what

---

[2] This practice of recording zero income as a bracket response only, rather than as a point value, is used in other rounds of the LFS as well.

influences the probability that bracket responses rather than point values for earnings will be reported.

Table 1 summarises the individual and household characteristics of the three sub-samples of the employed: the sample with point values for earnings (sample A); the sample whose earnings are reported in brackets (sample B); and the sample for whom no earnings information has been provided (sample C). With few exceptions noted in Table 1, the mean characteristics of these samples are significantly different from each other.

Table 1. Mean characteristics of the employed (standard deviations in parentheses).

	All the employed	Employed with point earnings *value* (A^1)	Employed with earnings *bracket* (B)	Employed with no earnings information (C^2)
Self-reporting	0.5419	0.5796	0.4958	0.3268
	(0.4983)	(0.4936)	(0.5000)	(0.4692)
Female	0.4349	0.4379	0.4412	0.3836
	(0.4958)	(0.4961)	(0.4965)	(0.4864)
White	0.1493	0.0710	0.2853	0.4569
	(0.3564)	(0.2568)	(0.4516)	(0.4983)
Age	38.82	38.56	39.25	39.84
	(11.38)	(11.38)	(11.21)	(11.91)
Years of schooling	8.860	7.970	10.718	11.257
	(4.143)	(4.136)	(3.518)	(2.964)
Rural dweller	0.3507	0.4077	0.2428	0.1588
	(0.4772)	(0.4914)	(0.4288)	(0.3656)
Self-employed	0.1550	0.1304	0.2018	0.2376
	(0.3619)	(0.3367)	(0.4014)	(0.4258)
Informal sector employment	0.1733	0.1850	0.1580	0.1106
	(0.3785)	(0.3883)	(0.3648)	(0.3137)
Household size	4.1670	4.138	4.204	4.3280
	(2.596)	(2.682)	(2.375)	(2.464)
N	23 731	16 366	5 728	1637

Source: LFS 2002:2

Notes: 1. Except for 'Female' and 'Household size', the differences in the means between those reporting a point value and those reporting a bracket are significant at the five percent level at least.
2. Except for 'Household size' and 'Age', the differences in the means between those reporting a bracket and those with missing information are significant at the five percent level at least.

A clear difference across samples is the extent of self-reporting on income. Close to sixty percent of the employed with point earnings values reported on their own income; among bracket respondents, the proportion of self-reporters was significantly lower at fifty percent; among those with missing information, it was significantly lower still at 33 percent. These statistics suggest that part of the reason for initial income non-response is that respondents do not know exactly how much income is earned because they are proxy-reporting for others in the household. Many, although not all, can then be persuaded to "estimate" earnings when being shown earnings brackets.

The type of employment individuals are involved in also differs significantly across the three samples. Individuals in sample A were the least likely to be self-employed. A significantly larger proportion of sample B was self-employed with the largest proportion of self-employment among those in sample C. Self-employment may capture how regular, and regulated, earnings are likely to be, and may therefore reflect the difficulty in computing "exact" or even approximate earnings.

The descriptive statistics presented in Table 1 would suggest, however, that reporting on earnings is not simply influenced by the cognitive requirement of providing that information. Average years of education are highest among those for whom no earnings information is reported; they are lowest among the sample with point values for earnings. Higher education usually reflects higher earnings, as does being older, white and living in an urban area. Compared to those who reported a point value, all these characteristics are more pronounced in the sample of bracket respondents and most pronounced in the sample of complete non-respondents. The descriptive statistics therefore are also consistent with there being differences in how earnings information is reported according to income.

We explore this further in Figures 1 and 2. Figure 1 compares the income distribution of those with bracket responses to those with point values (which we have assigned to their respective brackets). The earnings distribution for sample B clearly lies to the right of the distribution for sample A. Approximately 46 percent of all the employed with bracket responses were earning more than 3 500 rands a month compared to less than 15 percent of those with point values for earnings.

Figure 1. Income distribution of the employed (% in each bracket)

Earned income bracket (Rands)
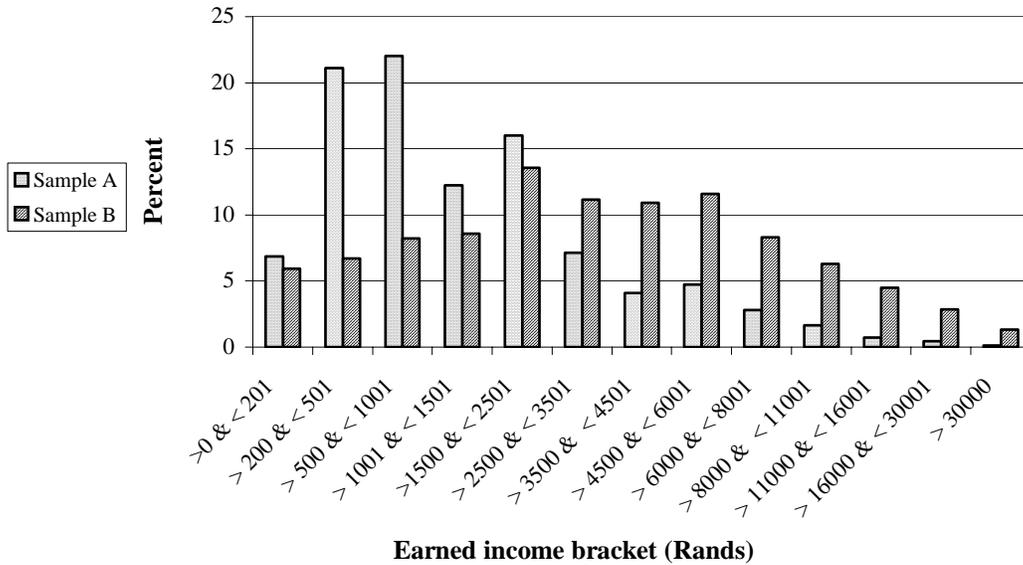
Figure 2. Percentage of the employed with earnings information who reported in brackets

Earned income bracket (Rands)
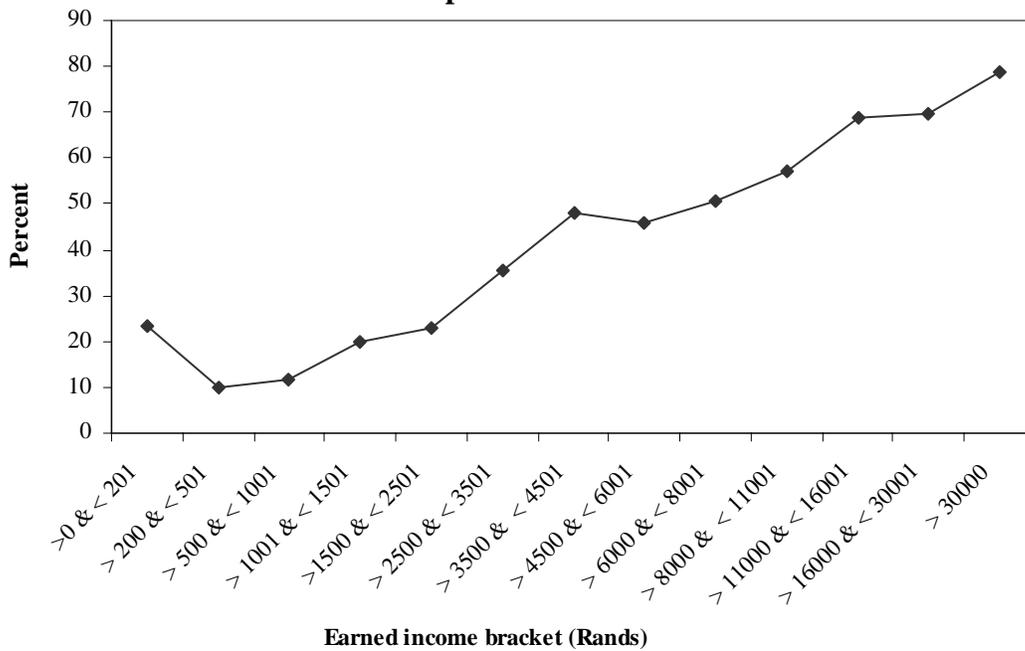
Source: LFS 2002:2

However, Figure 2 illustrates that the relationship between income and bracket responses is not simply linear. Rather, the proportion of bracket responses initially declines as income increases, before rising to almost eighty percent of all those in the top, open-ended bracket. Our findings are consistent with the U-shaped relationship between income and income non-response identified in the literature, although for our sample, we find a very early turning point and a relationship that would better be described as J-shaped (at least partly explained by our sample excluding zero income-earners).

We estimated the probability of providing point values rather than bracket information on earnings in a multivariate context, using a probit model. We chose to specify the equation with the dependent variable equal to one if earnings were reported as point values (and therefore as zero if earnings were in brackets) because this captures the first stage of a selection model which we consider in the next section.[3]

The results of the estimation, which are reported in Table 2, show that in a multivariate analysis, income continues to be a large and significant predictor of reporting propensities. The non-linear relationship between income and reporting is also clearly evident from the marginal effects of the earnings brackets. These initially increase, reaching an early maximum and then decline as the bracket value increases.

All the other individual variables that were significant in the descriptive analysis in differentiating the samples of point and bracket responses are also significant in the regression analysis. Point values are less likely (and bracket values therefore more likely) to be reported among those employed who are older, more educated, white, living in urban areas and not reporting on their own income. Even after controlling for income, therefore, the correlates of earnings continue to have a significant independent effect on the probability of reporting point values. The type of employment further affects how income is reported. Individuals who are self-employed and who are employed in the informal sector are significantly less likely to have point values for earnings.

Household size and gender now also show up as significant – living in large households and being female reduces the probability that point values will be reported. Household size may proxy for how much there is to know in the household and therefore for the likelihood of respondents knowing 'exactly' how much is earned. It is not clear

---

[3] Our findings are robust to estimating the probability using a linear probability regression model.

why bracket responses are more likely for women than for men – one possibility is that women may not want to disclose their earnings to others in the household as a way of retaining control over this income.

Table 2. Reporting point values or brackets for earnings, probit estimation

Dependent variable =1 if point value is reported	Coefficients[1]		Marginal effects[2]
> 0 & < 201 (Rands)	0.91223	(0.15503)*	0.18894
> 200 & < 501	1.41220	(0.15240)*	0.27688
> 500 & < 1001	1.36206	(0.15103)*	0.27524
> 1001 & < 1501	1.05835	(0.15101)*	0.21667
>1500 & < 2501	0.95604	(0.14963)*	0.21203
> 2500 & < 3501	0.70499	(0.15022)*	0.16186
> 3500 & < 4501	0.46710	(0.15078)*	0.11697
> 4500 & < 6001	0.58933	(0.15009)*	0.14059
> 6000 & < 8001	0.50464	(0.15187)*	0.12340
> 8000 & < 11001	0.42934	(0.15422)*	0.10791
> 11000 & < 16001	0.18428	(0.16085)	0.05133
> 16000 & < 30001	0.19717	(0.16953)	0.05456
Self-reporting	0.21798	(0.02105)*	0.06580
Female	-0.09275	(0.02080)*	-0.02787
White	-0.40491	(0.03066)*	-0.13378
Age	-0.00368	(0.00097)*	-0.00110
Years of schooling	-0.05586	(0.00348)*	-0.01672
Rural dweller	0.05718	(0.02340)**	0.01701
Self-employed	-0.32237	(0.03549)*	-0.10419
Informal sector employment	-0.12239	(0.03651)*	-0.03770
Household size	-0.01720	(0.00415)*	-0.00515
Constant	0.49076	(0.16518) *	
Percent correctly predicted	76.80		
N	22 094		

Source: LFS 2002:2
Notes: [1]Standard errors in parentheses.
[2]Marginal effects on the probability that y=1 calculated for infinitesimal changes in the continuous explanatory variables and for discrete changes in the dummy variables from 0 to 1.
* Significant at the one percent level; ****** Significant at the five percent level.

In sum, in the South African data that we use, bracket responses represent a large portion (over three quarters) of the cases where earnings information was not initially provided as a point value. Those employed for whom bracket responses are reported exhibit the same characteristics identified in the literature as the non-response cases more generally.

However, our study also shows that bracket responses are, not surprisingly perhaps, a distinct sub-sample of all non-responses. A significantly larger proportion of those for whom *no* earnings information are provided are white, urban and self-employed compared to those whose earnings are provided in brackets. The complete non-response cases also have significantly more years of education.

Our primary interest in this study lies in a comparison between those employed whose earnings are reported in brackets and those with point values for earnings. We have shown that there are a number of significant differences between these two samples, including that earnings in brackets are concentrated in the tails, and particularly the upper tail, of the earnings distribution. The question which we now explore empirically is whether we need to take this information into account when reconciling bracket responses with point values for earnings. Should we use information on the sample of point earnings respondents to help us predict point values for those with earnings in brackets?

5. Estimating point values for earnings in brackets

Nationally representative household surveys in South Africa provide an important source of information about labour force participation in the country. Using these data to describe labour market outcomes – such as average earnings, earnings poverty and inequality – requires that earnings be aggregated or ranked across the employed. In turn, this entails assigning point values to earnings reported in brackets. In this section, we consider empirically five different ways of estimating point values for bracket responses.

The most common method adopted in the analysis of South African survey data has been "the midpoint method" (cf. Hofmeyr, 2002; Casale, Muller & Posel, 2004; Kingdon & Knight, 2004; Meth & Dias, 2004; Vermaak, 2005). All bracket responses are assigned a point value equal to the midpoint of their corresponding earnings bracket. The approach is desirable for its simplicity. Another simple alternative would be an "actual average method", in which all point values for earnings are allocated to their respective

brackets. The mean of the point values by bracket would then be assigned to the bracket responses.

Table 3. Earnings (Rands) Across Brackets

Earnings Bracket (Rands)	Brackets Midpoint	Point Values Reported by Bracket Mean (Standard Deviation)	Midpoint/ Mean
>0 & < 201	100	151.29 (48.05)	0.6610
> 200 & < 501	350	370.75 (87.06)	0.9440
> 500 & < 1001	750	754.91 (148.24)	0.9932
> 1001 & < 1501	1250	1278.48 (152.67)	0.9777
>1500 & < 2501	2000	1992.40 (291.40)	1.0404
> 2500 & < 3501	3000	3042.53 (288.55)	0.9860
> 3500 & < 4501	4000	4092.22 (265.10)	0.9775
> 4500 & < 6001	5250	5385.78 (466.21)	0.9748
> 6000 & < 8001	7000	7156.89 (610.39)	0.9781
> 8000 & < 11001	9500	9593.00 (799.70)	0.9903
> 11000 & < 16001	13500	13522.58 (1501.37)	0.9983
> 16000 & < 30001	23000	21484.49 (3836.88)	1.0705
> 30000	35000	74752.05 (95528.15)	0.4682

Source: LFS 2002:2

Table 3 compares earnings derived from these two methods. Overall, the results from the midpoint and the mean of actual values are very similar – reported mean earnings vary from the bracket midpoint typically by between one and three percent. The exception is for the lowest and highest earnings brackets, where reported earnings are significantly larger than the bracket midpoint. For the highest income earners, mean earnings for

sample A are more than double the value we assigned to the highest bracket (the midpoint between 30 000 and 40 000 rands), and given our findings in section 3, it seems likely that using this "midpoint" will artificially truncate the upper tail of the earnings distribution.

The overall similarity between the two sets of measures may seem to be reassuring, but it is not immediately clear why a similarity would be appropriate. We have shown that the sample of the employed whose earnings are reported in brackets is systematically different to the sample whose earnings have been reported as point values. We would therefore only expect average earnings in sample A to provide a good reflection of the true average earnings of sample B, if selection into B operates at the level of which bracket, rather than where in the bracket, the individual falls.

Both methods could also be considered 'crude' because they give the same earnings value, per bracket, to all the employed with bracket responses. Neither method therefore assigns a distribution to earnings within the brackets for sample B.

This latter problem could be addressed by estimating an earnings function to impute point earnings values for bracket responses, thereby generating a continuous distribution of earnings for those in sample B. The simplest of these estimations, which we identify as a third method, is to estimate an ordinary least squares (OLS) regression. Using actual earnings values for sample A and information on the characteristics of A, a standard log earnings regression can be run on a variety of explanatory variables, representing personal, job and regional characteristics (commonly used in South African earnings equations). We would then use the estimated coefficients from A's earnings function to predict earnings for sample B based on B's characteristics. (The results from this regression are reported in the Appendix. We find that almost all the estimated coefficients are highly significant.)

However, the simple OLS estimation assumes that it is appropriate to use information about earnings for sample A to predict earnings for sample B. The structural relationship between earnings and the explanatory variables is assumed to be the same for both A and B. But this may not be the case: group B may be different not only because of observable differences, but also because of unobservable differences that affect the nature of the earnings function.

Furthermore, estimating a simple earnings equation for sample A to impute values for sample B introduces a new concern – we are not using all the earnings information that we have available to predict earnings for bracket responses. Specifically, we are ignoring information about the end points of the earnings brackets reported for sample B. The implications of disregarding this information are clear when we examine how 'well' the OLS regression predicts earnings for those in brackets.

In Table 4, we detail the percentage of predicted earnings that correspond to the actual earnings bracket reported for the employed in sample B, and in sample A. The match overall is not high – imputed earnings fell within the correct bracket for only about a third of all the employed with earnings information. The match for sample B is also substantially lower than that for sample A. Predicted earnings match reported brackets for only approximately 23 percent of those in sample B compared to just over 36 percent of the employed in sample A (indicating again that bracket responses do not represent a random sample of the employed with earnings information).[4]

Table 4. Imputed Earnings that Match Reported Brackets – Not Using Bracket Information

Earnings Bracket (Rands)	Percent imputed earnings correctly assigned		
	All	Sample A	Sample B
>0 & < 201	5.81	6.60	3.23
> 200 & < 501	56.61	58.57	38.90
> 500 & < 1001	43.10	44.02	36.09
> 1001 & < 1501	30.93	31.99	26.63
>1500 & < 2501	41.43	42.30	38.48
> 2500 & < 3501	21.02	18.58	25.47
> 3500 & < 4501	16.58	16.22	16.96
> 4500 & < 6001	16.84	16.32	17.44
> 6000 & < 8001	13.45	10.17	16.63
> 8000 & < 11001	15.21	11.11	18.28
> 11000 & < 16001	9.31	5.93	10.85
> 16000 & < 30001	1.71	0	2.45
> 30000	0	0	0
Total	33.01	36.48	23.08

Source: LFS 2002:2

---

[4] Similar results are found in Juster & Smith (1997) in a study on imputing point estimates for missing wealth data in U.S. household surveys. Using a simple OLS regression model, they find that only 35 percent of predicted bracket responses fell within the relevant reported range, with the misallocation being even higher at the ends of the distribution.

The final two methods that we consider estimate earning values for sample B by explicitly recognising selection effects and bracket information respectively. We estimate a maximum likelihood (ML) Heckman selection model to control for possible selection effects into sample A. The selection equation is captured by the probit regression reported in Table 2; the earnings equation is estimated using the same set of explanatory variables as in the simple OLS regression.

Not surprisingly, we find that there is a significant selection effect: a likelihood ratio test that there is zero correlation between the error terms of the two equations is rejected at better than the one percent level. Table 5 shows that controlling for sample selection produces both a higher overall match between actual and predicted earnings brackets, and a closer match between samples A and B. Nonetheless, the selection model still only correctly assigns predicted earnings to their actual brackets for well below half of the employed.

Table 5. Imputed Earnings that match reported brackets – ML Heckman Selection

	Percent imputed earnings correctly assigned		
Earnings Bracket (Rands)	All	Group A	Group B
>0 & < 201	4.31	4.46	3.81
> 200 & < 501	48.63	49.94	36.81
> 500 & < 1001	50.64	51.35	45.22
> 1001 & < 1501	24.48	25.35	20.93
>1500 & < 2501	52.84	54.15	48.39
> 2500 & < 3501	43.42	46.40	37.97
> 3500 & < 4501	33.08	32.14	34.08
> 4500 & < 6001	45.39	48.07	42.26
> 6000 & < 8001	49.63	52.38	46.95
> 8000 & < 11001	51.51	50.00	52.63
> 11000 & < 16001	62.50	59.32	63.95
> 16000 & < 30001	63.68	52.11	68.71
> 30000	46.94	14.29	55.84
Total	42.95	43.83	40.43

Source: LFS 2002:2

As an alternative to the selection model, we incorporate bracket information into our estimations by running OLS earnings regressions using 'bracket restrictions'. That is, we

estimated separate log earnings equations for each bracket based on the point values reported in that bracket. Some of the upper brackets had to be combined because of small sample sizes, and in the end we ran ten separate earnings equations. By estimating with bracket restrictions, we use a different set of coefficients to predict sample B's earnings per bracket, which also ensures that predicted earnings values are more likely to match their actual brackets.[5]

In a final step, we tried accounting for selection into sample A *within* each bracket, but found that there was no significant selection bias in the separate earnings equations by bracket. Possible reasons for this could be that our sample sizes are too small for some of the estimations, and that there is not sufficient information to adequately distinguish the selection equation from the wage equation (Deaton, 1997). But another reason may be that the selection effect operates principally on which bracket the individual falls into and not where in the bracket the individual falls. Estimating OLS earnings equations with bracket restrictions therefore may be effectively functioning as if we were controlling for selection bias in the estimated coefficients for sample A. Hence, within each bracket, actual values may indeed provide a good reflection of the earnings distribution of bracket responses[6]. In the section that follows, we consider how these five different methods of estimating point values for bracket responses – midpoint, actual average, unrestricted OLS regression, ML Heckman selection model and restricted OLS regression – affect a range of descriptive measures of earned income.[7]

---

[5] The match between predicted earnings and reported brackets for the full regression sample (i.e. A and B) rises to 99.6 percent when separate earnings equations are estimated per bracket. There is not a perfect match because the earnings model does not fully explain earnings. That the 100 observations that are incorrectly assigned are found mostly at the top and bottom ends of the bracket distribution indicates that there are unexplained factors that determine earnings operating mostly in the tails of the earnings distribution.

[6] Within each bracket, therefore, the functional relationship between earnings and the explanatory variables for sample A may provide a good reflection of this relationship for the bracket responses; but the characteristics between these two samples may be different within brackets.

[7] Another method of imputation that we do not consider in this paper is 'hot-decking', a procedure that involves matching each non-respondent to a respondent according to a set of observed characteristics, and then assigning the respondent's point value to the non-respondent. While this method is commonly used in the U.S. (for example, the Census Bureau uses hot-decking to impute missing values for income and wealth, see also Juster & Smith, 1997), to our knowledge it is not used much in South Africa. See Lillard *et al* (1986) on the disadvantages of using the hot-deck procedure for imputation compared to regression models.

6. Descriptive measures of earnings by method of estimation

An objective of assigning point values to earnings bracket responses is to derive descriptive measures of earnings that can be used to evaluate labour market performance. For example, what do people earn on average, what is the extent of poverty among the employed and how unequal are earnings? In this section we investigate empirically whether, and how, earnings estimates differ depending on how point values for bracket responses are imputed.

We first examine the range of estimates of average earnings for those with bracket responses. Table 6 reports these average earnings by bracket and by method. The data show that at this disaggregated level, there is a wide spread in average earnings by bracket across the different approaches. The midpoint, mean and OLS with bracket restrictions methods produce consistent average earnings for sample B that obviously correspond to the bracket value reported. The simple OLS and ML Heckman selection models, in contrast, do not. Rather, earnings particularly at the lower and the upper ends of the earnings distribution, and particularly for the OLS estimation, fall significantly outside the reported brackets. The result is that the earnings distributions generated by these two methods are compressed.

However, there are also clear differences between average earnings predicted by the simple OLS estimation and by the selection model, differences that would be expected given our findings in section 4. By taking into account possible selection bias of being in sample A (a selection which corresponds to a low presence in the tails of the earnings distribution), the selection model generates earnings for sample B that are significantly lower at the bottom end of the distribution and significantly higher at the top end. The difference in predicted earnings is greatest among the highest income-earners – the selection model estimates average earnings that are more than 250 percent larger than those generated through the simple OLS regression.

Table 6. Imputed Average Earnings by Bracket for Sample B

Earnings bracket (Rands)	Bracket midpoint	Mean of bracket based on A's values	OLS – no bracket restrictions	ML Heckman selection model	OLS – bracket restrictions
>0 & < 201	**100**	**151.29**	**696.87**	**530.81**	**143.49**
		(48.05)	(800.81)	(446.61)	(35.12)
> 200 & < 501	**350**	**370.75**	**863.75**	**746.59**	**366.80**
		(87.06)	(982.02)	(507.75)	(37.43)
> 500 & < 1001	**750**	**754.91**	**1176.29**	**1120.01**	**762.65**
		(148.24)	(1159.19)	(604.94)	(63.31)
> 1001 & < 1501	**1250**	**1278.48**	**1660.96**	**1828.22**	**1284.35**
		(152.67)	(1085.36)	(625.22)	(48.66)
>1500 & < 2501	**2000**	**1992.40**	**2032.19**	**2414.08**	**1986.99**
		(291.40)	(1279.47)	(733.93)	(97.64)
> 2500 & < 3501	**3000**	**3042.53**	**2783.69**	**3590.56**	**3053.85**
		(288.55)	(1560.82)	(905.93)	(85.54)
> 3500 & < 6001	**4750**	**4786.28**	**3938.99**	**5357.62**	**4772.69**
		(751.98)	(2150.01)	(1367.66)	(205.54)
> 6000 & < 8001	**7000**	**7156.89**	**4962.30**	**7350.30**	**7195.08**
		(610.39)	(2520.31)	(1650.06)	(297.12)
> 8000 & < 16001	**12000**	**10788.08**	**6573.32**	**10853.45**	**10799.94**
		(2098.19)	(3239.66)	(2957.94)	(1036.43)
> 16001	**28 000***	**33643.39**	**8119.63**	**21537.61**	**34430.80**
		(50222.36)	(3934.40)	(7830.91)	(37610.0)

Source: LFS 2002:2

Notes: Some income categories (3501 to 4500 and 4501 to 6000; 8 001 to 11 000 and 11 001 to 16 000; 16 001 to 30 000 and in excess of 30 000) have been merged because of the small number of observations for the earnings equations with bracket restrictions. Standard deviations are in parentheses. * Calculated as the midpoint between 16 000 and 40 000 rands.

These comparisons help account for differences in the aggregated earnings estimates reported in Table 7. The table provides estimates of average earnings for those in sample B, for all the employed with earnings information (using the actual point values for sample A and the imputed values for sample B), as well as measures of poverty and inequality among the employed. Considerable variation remains in aggregate earnings by method for the employed with bracket responses. In particular, estimated earnings using the simple OLS method are significantly lower than all other predicted values – the OLS 'underestimates' earnings among high-income earners by considerably more than it 'overestimates' earnings among the low-income earners. As expected, the selection model predicts significantly higher average earnings than the simple OLS because it generates an earnings distribution that is more extended at the upper tail. Average

earnings for sample B are highest using actual average earnings and OLS estimations with bracket restrictions, because these methods further extend the distribution among high income-earners.

Table 7. Earnings Estimates – Average, Working Poor, Inequality – by Method (Standard Errors in Parentheses)

			Predicted earnings		
	Bracket Midpoint	Reported earnings, average by bracket	OLS - no bracket restrictions	ML Heckman selection model	OLS - bracket restrictions
Average earnings for sample B (Rands) [1]	4763.12 (76.46)	5311.10 (122.04)	3274.48 (37.96)	4950.71 (66.97)	5098.10 (136.46)
Average earnings for all (Rands) [2]	2730.36 (35.72)	2872.43 (43.74)	2344.42 (30.47)	2779.00 (34.55)	2817.21 (46.39)
Headcount index of earnings poverty (<467 Rands/month) [3][4]	0.2118 (0.0027)	0.2118 (0.0027)	0.1992 (0.0027)	0.1951 (0.0027)	0.2117 (0.0027)
Gini coefficient [5]	0.5976 (0.0047)	0.6152 (0.0051)	0.5523 (0.0047)	0.5832 (0.0041)	0.5997 (0.0049)
Variance of Logs	1.4818 (0.0139)	1.4713 (0.0135)	1.2410 (0.0101)	1.4085 (0.0125)	1.4573 (0.0128)

Source: LFS 2002:2

Notes: 1. Average earnings estimates (whether for sample B or for all the employed with earnings information) from the 'OLS – no bracket restrictions' method are significantly lower than the estimates from all other methods at the five percent level at least.
2. Earnings values will differ from national averages estimated for South Africa because i) data are not weighted; ii) all reported zero income earners are not included in the sample.
3. The national poverty line used for this table equals 467 rands per adult equivalent per month in 2002 prices. The poverty line represents the per adult equivalent household subsistence level (HSL) set by The Institute for Development Planning Research at the University of Port Elizabeth, South Africa (see Woolard & Leibbrandt, 2001:49).
4. The poverty estimates derived from the 'OLS – no bracket restrictions' and the 'Heckman selection' methods are significantly lower than the poverty estimates derived from the other estimates (at the five percent level).
5. The Gini coefficient and variance of logs derived from the 'OLS – no bracket restrictions' method are significantly lower than all the other estimates. The Gini and variance of logs estimates from the 'Heckman selection' method are both significantly lower than the estimates from the 'average by bracket' method, while the variance of logs is also significantly lower than that from the 'midpoint' method (all at the five percent level).

All the methods predict earnings for sample B that are significantly higher than actual earnings reported for sample A. The inclusion of earnings for bracket responses therefore raises average earnings for all the employed with earnings information. But because those with bracket responses constitute just over a quarter of the aggregate sample, the

difference in average earnings across the methods is less pronounced when the whole sample is considered. Although average earnings predicted by the simple OLS estimation remain considerably lower, the midpoint, actual average, Heckman selection and OLS with bracket restrictions models produce largely comparable measures.

The simple OLS estimation also consistently produces the lowest measures of poverty and inequality – again not unexpectedly given that this method raises average earnings at the bottom end and truncates average earnings at the top end of the distribution. The selection model also generates lower estimates of poverty and inequality compared to the other methods. However, because the earnings distribution is less compressed than that derived from the simple OLS method, the selection model still generates significantly higher measures of inequality, reflected in a higher Gini coefficient and variance of logs, compared to the simple OLS method.

Estimates of average earnings therefore seem consistent when we take either bracket information (as in the midpoint, actual average or the restricted OLS methods) or selection information (as with the ML Heckman selection model) into account. In fact, taking bracket information into account may effectively control for sample selection. An OLS estimation that recognises neither selection nor brackets produces average earnings measures that differ most from the other estimates. However, when calculating distributional measures, such as indicators of poverty and inequality, methods that use bracket information produce more consistent estimates than both the simple OLS and selection models.

There is no evidence to suggest therefore that, although crude, the midpoint method commonly used in South African studies, generates biased earnings estimates (or at least estimates that are more biased than the other estimations). The value of actual earnings among those in the uppermost bracket, however, would suggest that, with our data, we need to extend the 'midpoint' value assigned to the top bracket.

Our findings also have implications for imputing earnings for the sample of the employed with missing earnings data. Most studies in South Africa, including our own, have simply dropped these missing values. In the LFS 2002:2, the sample of the employed with no earnings information is relatively small, representing only seven percent of the employed in the aggregate sample. But we know also that this sample

(sample C) is not randomly distributed – as shown earlier the average characteristics of those in sample C are significantly different from those in samples A and B.

The findings presented above suggest that if we were to also include those with no earnings information in our analysis, we should not use a simple OLS earnings equation based on A's actual values to predict point earnings estimates for C. As was the case with predicting earnings for bracket responses, we would expect this model to considerably underestimate average earnings for the sample of complete non-responses. Without any information on bracket intervals, the most appropriate method to use therefore would be a Heckman selection model, controlling for selection into sample C. For the earnings equation of the Heckman model, we recommend using the estimates of sample B's earnings that we consider most reliable, that is, the imputed values derived from the OLS with bracket restrictions method[8].

As expected, the results indicate a significant selection effect, and average earnings for the full sample of employed increase following the inclusion of the imputed earnings for the complete non-response cases (to 2946.95 rands). However, the imputed average earnings estimate obtained for sample C using this method (4698.01 rands) is lower than that obtained for sample B using the OLS with bracket restrictions method (5098.10 rands, from Table 7). The Heckman selection model used here underestimates C's average earnings by compressing the upper tail of the earnings distribution for sample C. This result again highlights the usefulness of having bracket information for the employed in producing an earnings distribution that is closer to the 'true' distribution.

Conclusion

Item non-response in household surveys is a problem particularly for questions on income and wealth, both key variables in socio-economic research. For example, not

---

[8] This follows Juster & Smith (1997) who use the bracket responses as the 'donors' for imputing earnings estimates for the complete non-response cases, rather than the more conventional method of using the point value responses. They propose that bracket respondents are a more representative sample of complete non-respondents than those who provided point values, because of their shared tendency for initial non-response. Our descriptive findings in section 4 provide further justification, as they show that sample C's mean characteristics are more similar to sample B's mean characteristics than to sample A's.

having any earnings information on a significant proportion of the employed sample will bias earnings estimates if we expect that the missing observations are a non-random sample of all the employed. Complete income non-response, however, is substantially reduced by allowing initial non-respondents to provide bracket information on earnings.

In this study, we use data from a 2002 South African household survey to show that the employed can be clearly distinguished according to how their earnings information is reported. Consistent with international findings, the sample with point values for earnings has characteristics that differ significantly on average from those employed without point values.

Our analysis focuses on the particular features of the employed for whom bracket information has subsequently been provided. We find that those with bracket responses for earnings are a distinct sample of the employed in general, as well as a distinct sub-sample of all initial non-responses. The probability of earnings being reported in brackets is affected particularly by the level of earnings – those reporting in brackets are found to be concentrated in the tails, and especially the upper tail, of the income distribution. Ignoring bracket information would therefore lead to an underestimation of earnings.

In this paper we consider five methods of reconciling earnings reported as point values with bracket responses, with a view to calculating summary measures of earnings, poverty and inequality. Our results show that a simple OLS regression, that predicts earnings for bracket responses based on point value responses, and that does not take systematic differences between these two samples into account, performs the least favourably. Estimates of average earnings, poverty and inequality were all substantially lower than estimates derived from the other techniques explored: the "midpoint", "average", "Heckman selection" and "OLS with bracket restrictions" methods.

We also find that those methods that use the information on the bracket intervals (the "midpoint", "average", and "OLS with bracket restrictions"), produce highly consistent summary measures. The advantage of the OLS with bracket restrictions model is that it produces a continuous distribution within brackets, at the same time ensuring a good match between predicted earnings and the reported earnings range.

We also briefly considered the implications of our findings for imputing earnings values for the complete non-response cases. Using the sample of point responses to

predict values for those with no earnings information would lead to a considerable underestimation of this sample's earnings. However, even taking into account the possible non-randomness of the sample of complete non-responses in a Heckman selection model, it is very likely that imputed earnings for this group are underestimated. Without bracket interval information, the imputed earnings distribution is compressed, particularly at the upper end.

Our findings underline the importance of having, and using, bracket information to impute point earnings values. Statistical agencies therefore should not only continue to allow survey participants this option for sensitive questions such as income, but should also explore surveying techniques that would encourage more initial non-respondents to provide bracket information.

Appendix

Earnings Function (Ordinary Least Squares), on sample A's point values

Dependent variable = log of monthly earnings if point value reported	Coefficients
Age	0.08453 (0.00906)*
Age^2	-0.00125 (0.00021)*
Age^3	5.49e-06 (1.55e-06)*
Years of schooling	0.05564 (0.00173)*
Female	-0.25996 (0.01290)*
Married	0.08014 (0.01173)*
Informal sector worker	-0.51993 (0.02017)*
Self-employed	-0.06002 (0.02291)*
White	0.65197 (0.02277)*
Indian	0.31209 (0.03173)*
Coloured	0.16772 (0.01985)*
Log of hours worked per month	0.24470 (0.01269)*
Agriculture	0.10809 (0.05109)**
Mining	1.05009 (0.05689)*
Manufacturing	0.615808 (0.05460)*
Electricity	0.90034 (0.08244)*
Construction	0.59790 (0.05727)*
Trade	0.43576 (0.05416)*
Transport	0.70654 (0.05809)*
Finance	0.63287 (0.05672)*
Community services	0.88594 (0.05371)*
Exterior organisations	1.39136 (0.33462)*
Legislative/managerial	1.04546 (0.06189)*
Professional	0.98637 (0.06683)*
Technical/associate professional	0.78292 (0.05857)*
Clerks	0.51343 (0.05839)*
Service, shop, sales workers	0.18157 (0.05777)*
Skilled agriculture	0.34280 (0.04253)*
Craft and related trade	0.30441 (0.05833)*
Plant and machine operators	0.27695 (0.05796)*
Elementary occupations	0.09590 (0.05528)***
Rural	-0.16953 (0.01316)*
Eastern Cape	-0.34006 (0.02411)*
Northern Cape	-0.20506 (0.02537)*
Free State	-0.47928 (0.02525)*
KwaZulu-Natal	-0.11257 (0.02330)*
Northwest	-0.16868 (0.02513)**
Gauteng	-0.05317 (0.02348)*
Mpumalanga	-0.13875 (0.02590)*
Northern Province	-0.25405 (0.02648)*
Constant	3.16051 (0.14040)*
F(40, 16325)	743.15*
R^2	0.6455
N	16366

Notes: Standard errors in parentheses. * significant at the one percent level; ** significant at the five percent level; *** significant at the ten percent level

Omitted categories: male, not married, formal sector worker, employee, African, domestic work, private households, urban and western cape.

References

Atkinson A.J and Micklewright, J. (1983) 'On the reliability of income data in the Family Expenditure Survey 1970-1977'. *Journal of the Royal Statistical Society*. Series A (General), Volume 146, Issue I: 33-61.

Bell, R. (1984) 'Item nonresponse in telephone surveys: An analysis of who fails to report income'. *Social Science Quarterly*, 65: 207-215.

Bhat, C.R. (1994) 'Imputing a continuous income variable from grouped and missing Income Observations'. *Economic Letters*, 46: 311-319.

Casale, D.; Muller, C. and Posel, D. (2004) ' "Two million net new jobs": A reconsideration of the rise in employment in South Africa, 1995 –2003'. *South African Journal of Economics*, 72(5): 978-1002.

Daniels, R.C. and Rospabé, S. (2005) 'Estimating an earnings function from coarsened data by an interval censored regression procedure'. Development Policy Research Unit, Working Paper 05/91.

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

DeMaio T.J. (1980) 'Refusals: Who, where and why'. *The Public Opinion Quarterly,* 44(2): 223-233.

Duncan, G.J. and Hill, D.H. (1985) 'An investigation of the extent and consequences of measurement error in labor- economic survey data'. *Journal of Labor Economics*, 3(4): 508-532.

Duncan, G.J. and Petersen E. (2004) 'The long and short of asking questions about income, wealth and labour supply'. Mimeo, Northwestern University.

Flinn, C.J., Kulka, R., Moffit, R. and Wolpin, K.I. (2001) 'Introduction to the Journal of Human Resources special issue on data quality'. *The Journal of Human Resources*, 36(3): 413 – 415.

Hawkins, D.F. (1975) 'Estimation of nonresponse bias'. *Sociological Methods and Research*, 3: 462-485.

Hofmeyr, J. (2002) 'The importance of segmentation in the South African labour market'. Mimeo. University of Natal, Durban.

Juster, F. T. and Smith, J.P. (1994) 'Improving the quality of economic data: Lessons from the HRS'. Health and Retirement Study Working Paper Series, Paper No.94-027.

Juster, F. T. and Smith, J.P. (1997) 'Improving the quality of economic data: Lessons from the HRS and AHEAD'. *Journal of the American Statistical Association*, 92(440): 1268 – 1278.

Kingdon, G. and Knight, J. (2004) 'Unemployment in South Africa, 1995-2003: Causes, problems and policies'. Report prepared for National Institute for Economic Policy (NIEP), Pretoria.

Lillard, L., Smith, J.P. and Welch F. (1986) 'What do we really know about wages? The importance of nonreporting and census imputation'. *Journal of Political Economy*, 94(3): 489-506.

Meth, C. and Dias, R. (2004) 'Increases in poverty in South Africa, 1999-2002'. *Development Southern Africa*, 21(1): 59-85.

Moore, J.C. (1988) 'Self/proxy response status and survey response quality'. *Journal of Official Statistics*, 4(2): 155-172.

Moore, J. C. and Loomis, L.S. (2001) 'Using alternative question strategies to reduce income nonresponse'. Research Report Series (Survey Methodology #2001-03). Statistical Research Division, U.S. Bureau of the Census, Washington D.C.

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

Rodgers, W.L, Brown C. and Duncan, G.J. (1993) 'Errors in survey reports of earnings, hours worked, and hourly wages'. *Journal of the American Statistical Association*, 88(424): 1208 – 1218.

Stern, S. (1991) 'Imputing a continuous income variable from a bracketed income variable with special attention to missing observations'. *Economic Letters*, 37: 287-291.

Woolard, I. and Leibbrandt, M. (2001) 'Measuring poverty in South Africa', in H. Bhorat, M. Leibbrandt, M. Maziya, S. van der Berg, and I. Woolard. *Fighting poverty. Labour markets and inequality in South Africa*. Cape Town: UCT Press.

Zweimuller, J. (1992) 'Survey non-response and biases in wage regressions'. *Economic Letters*, 39: 105-109.