University of Cape Town

CENTRE FOR
SOCIAL SCIENCE RESEARCH

COMPARING ALTERNATIVE MEASURES OF HOUSEHOLD INCOME: EVIDENCE FROM THE KHAYELITSHA/MITCHELL'S PLAIN SURVEY

Jolene Skordis
Matthew Welch

CSSR Working Paper No. 25

Published by the Centre for Social Science Research
University of Cape Town
2002

Copies of this publication may be obtained from:

The Administrative Officer
Centre for Social Science Research
University of Cape Town
Private Bag
Rondebosch, 7701
Tel: (021) 650 4656
Fax: (021) 650 4657
Email: kforbes@cssr.uct.ac.za

Price in Southern Africa (incl. VAT and postage): R 15.00

or it can be downloaded from our website
http://www.uct.ac.za/depts/cssr/pubs.html

ISBN: 0-7992-2156-2

CENTRE FOR SOCIAL SCIENCE RESEARCH

Social Surveys Unit

COMPARING ALTERNATIVE MEASURES OF HOUSEHOLD INCOME: EVIDENCE FROM THE KHAYELITSHA/MITCHELL'S PLAIN SURVEY

Jolene Skordis
Matthew Welch

CSSR Working Paper No. 25

December 2002

Jolene Skordis is a Junior Researcher in the Social Surveys Unit, Centre for Social Science Research at the University of Cape Town.

Matthew Welch is Deputy-Director of the Data First Resource Unit, Centre for Social Science Research at the University of Cape Town.

Comparing Alternative Measures of Household Income: Evidence from the Khayelitsha/Mitchell's Plain Survey

Abstract

Household income is a variable that is used widely for economic and sociological analysis. Little has been written about the optimal way to generate the information necessary to calculate household income. Most South African analyses use a household income variable generated by a single household respondent reporting on the household income. The Khayelitsha/Mitchell's Plain Survey provides a unique opportunity to explore alternative ways of generating this variable. We compare the estimates of household income obtained from the household module to estimates of household income obtained by aggregating the detailed income data from the adult module of the survey. We show that household income estimates for the KMP survey tend to be higher and to have greater variation when estimated by aggregating individual income data compared to the estimates obtained in the household module. The difference between income estimates has a material impact on the secondary analysis of data. This is illustrated through the use of Gini coefficients, a simple measure of income-inequality. Household income measured at the household level appears to underestimate household income-inequality in a sample.

Introduction

Measures of household income are used to estimate a variety of socio-economic phenomena, from absolute poverty (Podder and Chatterjoo, 2002) to relative inequality (Quisumbing, Haddad, and Peña 2001; Mandel, 2002). Household income itself has been compared across health outcomes (Wang, Patterson and Hills, 2002) educational outcomes (Lanot, 2002) and labour force outcomes (Kingdon and Knight, 2000). Given this widespread application as both an independent and a dependent variable, the accurate measurement of this variable is vital to data analysis. Despite this importance, little has been written about the optimal way to collect the information necessary to generate a measure of household income. Should one ask a household representative for a total household estimate or, alternatively, ask all the income-earning members of the household for their individual incomes and calculate the household income post-

hoc? Do these methods yield similar estimates and, if not, how do the estimates differ?

The Khayelitsha/Mitchell's Plain Survey (KMPS) (SALDRU/Centre for Social Science Research, 2000) was conducted in the magisterial district of Mitchell's Plain in Cape Town, which is an historically black/African and coloured district. The survey asked questions at both the household and the individual level. All household residents over the age of 18 years were asked to participate in the study. A household representative was asked to report household income at the household level and personal income was asked of all age-eligible participants at the individual level. This dataset therefore provides the opportunity to explore the two ways in which one could estimate household income i.e. from a single household representative or by "estimating up" using individual responses from all adults in the household.

Measuring household income at the household level

Household income as measured at the household level by the KMP survey is R1680.19 per month on average (see *Table One* below). This estimate is taken from a single question in the KMPS household module.[1]

Table One: Total monthly household income (all sources)

Obs	1086
Median	R1000
Mean	R1680.19
Std. Dev.	R2122.924
Variance	4506808
Skewness	4.958602
Kurtosis	47.60871

The median income (exactly R1000) and the "neat" round values at each percentile tell their own story (illustrated further in *Figure One*). Household income collected at the household level seems to be particularly vulnerable to heaping. The phenomenon of rounding up income responses is not unusual. But later comparison with household income derived from the individual level

[1] Question 16 of the household module is worded as follows: "How much money comes to the household from all sources in a typical month?"

data shows that the intervals between heaping are larger in the household level data, i.e. that household level data exhibits less variation or larger intervals between discrete points.

Figure One: Plotting household income from the household module

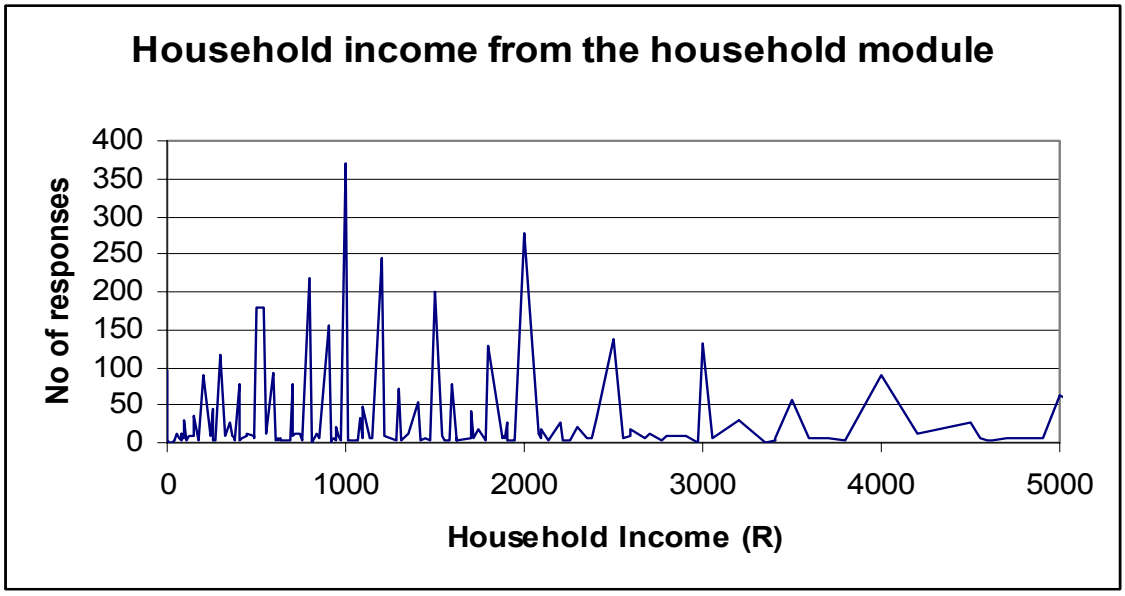
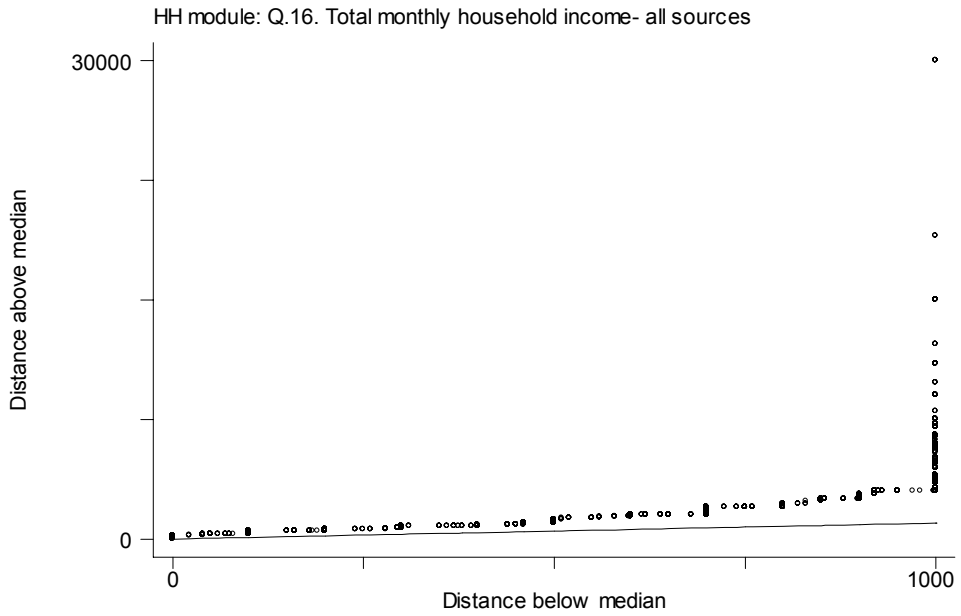


Figure Two presents the household data in the form of a symmetry plot. The interpretation of a symmetry plot is described by Hamilton (1992: 10-11):

> 'The median divides a distribution in half. If the distribution is exactly symmetrical [as is a normal distribution], then for each value above the median there is another value the same distance below the median. A symmetry plot graphs the distance from the median of the ith value above the median against the distance from the median of the ith value below the median. Each pair of values defines one point, so a symmetry plot based on n cases will contain about n/2 points.'

This symmetry plot suggests that there are outliers in the data. Looking at the characteristics of those households which might be considered outliers does not, however, provide any firm grounds for dropping the observations from further analysis. If we were to estimate household income using regression techniques with household income from the household module as the dependant variable, it might be wise to apply a log transformation, bootstrapping techniques or some other more robust technique less sensitive to outliers. (Statacorp, 2001).

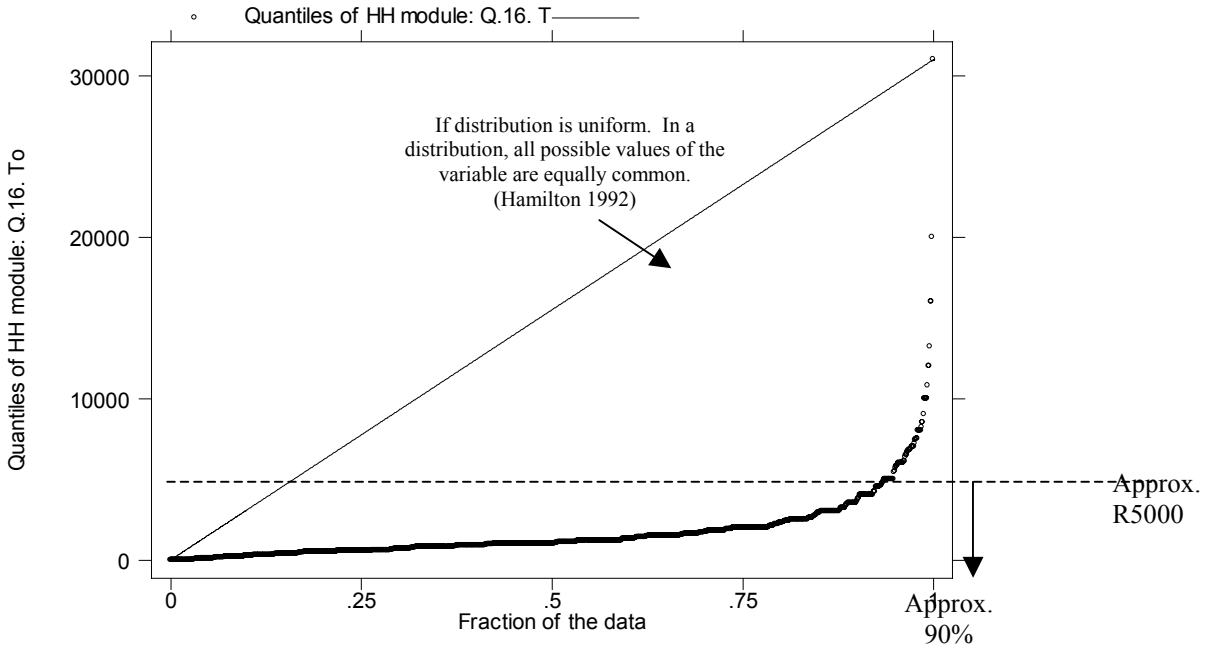
Figure Two: A symmetry plot of household income from the household module



Aside from showing the outliers in the data, the symmetry plot shows that the data is symmetrical i.e. the points are close to the line, for much of the distribution. At higher levels of income however, the data becomes asymmetrical and the distance above the median is far greater for these observations than the corresponding distance below the median.

Figure Three shows a quantile plot that reinforces this asymmetry in the data and shows us the fraction of the cases that lie below a given value. It is apparent that more than 90% of the sample have a household income of less than R5000 per month (measured at the household level). This is not surprising in a relatively impoverished urban area such as Khayelitsha/Mitchell's Plain.

Figure Three: A quantile plot of household income from the household module



Due to a design fault in the household roster, data was not collected on the person who provided the household level estimate of household income. Therefore it is not possible to segment the findings based on any particular characteristics of that individual. However, this figure (or something like it) would normally be sufficient for inter-household analysis (Podder and Chatterjoo, 2002). For the purposes of the following discussion, this estimate serves as a baseline figure against which subsequent estimates of household income (derived from individual level data) will be compared.

The table below gives us the confidence intervals around the household level estimate of household income (henceforth referred to as the "baseline" estimate). Three levels of confidence are tabulated: the 90%, 95% and 99% confidence intervals. As this measure will serve as the baseline figure for comparison in the remainder of this paper we need to understand whether the subsequent derived estimates fall within the bounds of the confidence intervals for the baseline measure.

Table Two: Confidence intervals for baseline household income

Variable	*Obs*	*Mean*	*Std. Err.*	*[90% Conf.*	*Interval]*
Household Income	1086	R1680.19	64.41983	1574.14	1786.243
				[95% Conf.	*Interval]*
				1553.79	1806.593
				[99% Conf.	*Interval]*
				1513.965	1846.419

Measuring household income at the individual level

Measuring individual incomes

In order to derive a measure of household income from individual responses (a process referred to in this paper as "estimating-up"), one must first have an estimate of individual income. With the KMP dataset, the challenge is to construct a reliable income figure that consists of a number of different components, namely:

1. Wage income from first and (possibly) second jobs; which in turn consists of:
    - Pure wage payments
    - Overtime pay,
    - Bonus payments,
    - Piece rate payments,
    - Share of profits,
    - Bonuses when the business is doing well, and
    - Productivity pay.
2. Casual income,
3. Income from self employment,
4. Income from grants and investments,
5. Income from other people,
6. Other forms of income not captured by the categories above.

Each of these components could be presented in a number of ways and two decisions need to be made. Firstly, which manner of calculation is the most "correct"? And, secondly, which manner of calculation will readily enable us to add together all components to arrive at a reliable income figure? Individual income is calculated below as gross income rather than net income. This is necessitated by the questionnaire design, which allows pure wage income to be given in either net or gross terms but explicitly asks for overtime payments in gross terms. The estimate of gross monthly income (including these overtime and other payments) can then be adjusted downwards to estimate net monthly income using an appropriate tax schedule (SARS, 2000).[2]

1) Wage Earnings

Income from wage earning activity is asked in two separate places in the KMP questionnaire. In the first instance (question E9), respondents are simply asked for their basic wage and then a follow-up question is used to ascertain whether the amount given is net or gross. In the second instance (question E19), respondents are asked to specify their gross earnings, to list all deductions and finally to specify their net earnings. In order to reduce the effect of item non-response on these wage income questions, data from both questions has been used to measure wage income in this paper. The second question (which asks for gross earnings and full details of wage deductions) had the higher number of positive responses for gross income. However, where responses were missing, they were supplemented in two ways:

- Firstly, by using a value from the first question (E9[3]) if the wage specified in that question was a gross wage; and

[2] Question E19 is worded as follows: "Please fill in the following detailed about deductions...". The question asks for gross earnings and subsidies (before deductions), details of all wage deductions and finally for net earnings after all deductions. This net earnings figure is not used for a number of reasons; firstly, the "sub-sections" of this question (i.e. the details of deductions from gross earnings) have a high incidence of non-response. Secondly, using this estimate for the few people that did provide details of salary deductions means that we would have to treat observations in a varied (and inconsistent) way, based purely on whether they had given adequate responses to question E19. Thirdly, the deductions that are provided in E19 seldom sum up to the gross income provided when added to the net earnings figure.

[3] Question E9 of the Adult Module is worded as follows: "What is your basic wage (i.e. excluding overtime payments)? To determine the frequency of that wage payment it needs to be combined with additional information from question E8, which asks; "Do you get paid every day, every week, every fortnight, or every month?" .

- Secondly, by adding the wage deductions and net income specified, to arrive at a gross estimate of wage income.

Gross wage income was calculated on a monthly basis and as such, some of the daily, weekly and fortnightly estimates needed to be adjusted accordingly. In a few cases, respondents had specified their wage income but had failed to specify the regularity with which that income was paid. In these instances the wage provided was compared against the mean of the daily, weekly, fortnightly and monthly wage earners to asses the likely regularity of payment. This is not a fool-proof process as the response being categorised may be an outlier, however there were only four such cases and this will not significantly affect the final income estimates for the sample.

In short, the final construction of total wage income uses data from the detailed wage question (E19) as the primary variable and, where it is missing, replaces it with data from a simpler wage question (E9) if that question is a gross measure. Respondents who were not in wage income were not required to fill in wage income (either in question E9 or question E19); these respondents have been allocated a zero wage income. This still leaves approximately 529 missing estimates of gross wage income but the variable is as complete as possible i.e. approximately 20% of the sample did not give positive income estimates and could not be classified as zero income earners. As the following table shows, many of these missing individual observations were captured in later "emergency[4]" and "proxy[5]" questionnaires.

For the purposes of this discussion, we have chosen not to use emergency responses as these are asked in net income terms. Proxy responses are not used because the respondent may be the same person who is giving the household level estimate in the household module – thus convoluting our attempts to compare these supposedly different estimates. Consideration was given to the inclusion of proxy responses if the payslip had been shown, however none of the proxy respondents showed pay slips to the interviewers. The proxies and emergency modules have also not been used because, while both of the questions in these questionnaires ask for income, one has a higher response rate

[4] Emergency questionnaires are shorter versions of the original questionnaire. Some emergency questionnaires were answered by proxy and others by the adult identified in the household roster.

[5] Proxy questionnaires are the same questionnaires that were administered to the bulk of the sample however, the adult identified in the household roster did not answer his or her own questionnaire. Instead, a suitable proxy provided the information.

than the other; 856 positive responses in E9[6] and 601 in question E19[7]. This raises concerns about getting different responses if questions are ordered differently and asked differently.

The following table shows how the wage income information is distributed between employment categories (as defined by Nattrass (2002)), proxy respondents and emergency respondents.

Table Three: Breakdown of wage income responses

Labour Force Categories (Nattrass, 2002)	*Total number of respondents in each category*	*Number of respondents who gave wage income*	*Number of respondents who did not give wage income*	*Number of respondents categorised by proxy information*[8]	*Number of respondents categorised by emergency module information*[9]
	(a)	*(b)*	*(c)*	*(d)*	*(e)*
Wage-employed	882	767	115	85	2
Self-employed	210	3	207	201	0
Casual-employed	66	5	61	56	0
Active job-seekers	448	448	0	0	0
Exclusive network job-seekers	173	173	0	0	0
Marginalised unemployed	390	390	0	0	0
Non-labour force participants	329	329	0	0	0
Unclassified	146	0	146	187	76
Total	**2644**	**2115**	**529**	**431**	**78**

All adult respondents in the KMPS should have an allocated wage income regardless of whether or not they are in wage employment. For those who are definitely not in wage employment – i.e. the job-seekers (active and exclusive), the marginalised unemployed, and the non-participants in the labour force – their wage income would simply be zero. This said then, columns "b" and "c" add to the same total as column "a" i.e. 2644. Columns "d" and "e" add to 509, which imply that only 20 of the 529 missing wage income observations cannot be accounted for by proxy and emergency information (which is approximately

[6] Of these 856 responses, 257 were gross income estimates.

[7] This in itself is an interesting phenomenon – why should one income question elicit a higher response rate than another? Is it due to question ordering or perhaps the wording of the two questions? This is a subject for a different paper.

[8] No pay slip was provided by the proxy to verify any of these responses.

[9] The reader is reminded that the emergency information was asked in net terms and as such is unsuitable for the current analysis.

0.8% of the total sample). The decision not to use the proxy information was taken with a particularly conservative view in mind, i.e. that the findings of this paper should not be obscured by debate around the validity or suitability of proxy respondents who did not provide payslips to verify the income information provided.

The numbers detailed in the table above raise the question of the five casual workers and the six self-employed respondents who have not given wage information and were not classified by proxy or emergency module information. There is a strong argument for including these respondents as zero wage income earners. However, there is an equally strong argument for omitting these respondents as the Nattrass categories classify respondents on the basis of their main occupation and not on the basis of their sole occupation. These respondents may be in more than one form of employment and as such, they may also be involved in wage employment. This is illustrated by the fact that other casual and self-employed workers have given positive wage incomes. The eleven respondents who did not give any wage income information are unlikely to have a significant impact on the final analysis. As before then, the decision was taken to follow a conservative path and leave these respondents out of the wage income variable. Their income from casual work and self-employment however will be captured subsequently (if they provided it) and treating them as "missing" wage income values does not mean that they are necessarily treated as "missing" in any subsequent analysis.

2) Overtime Pay

Monthly overtime pay was calculated by multiplying the wage that the respondent is paid for an hour of overtime before tax, by the number of hours of overtime they worked in the past month. In this instance it was not possible to accommodate for item non-response by using information from other sources.

Table Four: Summary of the overtime variable

Variable	*Obs*	*Mean*	*Std. Err.*	*[95% Conf.*	*Interval]*
Overtime pay for first job[10]	228	R1411.559	580.3182	268.0602	2555.059
Hours spent working overtime in first job last month	228	R16.46491	1.423546	13.65986	19.26997

[10] Only 8 people claimed to receive overtime pay for their second job and as such this information is not tabulated above.

3) Bonus payments, piece-rate payments etc

These bonus payments refer to thirteenth cheques or regular annual payments converted to monthly amounts. The treatment of bonus payments is more complex as there is a real risk that respondents stipulated their annual payments rather than the monthly equivalent, despite the instruction in the question. This is clearly the case in instances where the "monthly" bonus exceeds or is equal to the monthly wage income. Unfortunately it is less clear in instances where the bonus comprises a significant portion of the monthly income but does not exceed it. This problem thus requires the application of a decision rule and it was decided that any bonus payment that exceeded 25% of the monthly salary was likely to be an annual payment. These payments were divided by twelve to arrive at a monthly bonus value.

Table Five: Summary of the bonus pay variable

Variable	*Obs*	*Mean*	*Std. Err.*	*[95% Conf.*	*Interval]*
Bonus pay for first job[11]	293	R425.78	44.96482	337.2785	514.271

The question of how best to treat other piece rate payments collected by the questionnaire remains. Piece rate payments and other forms of productivity related pay might be paid quarterly or on an ad hoc basis. Unfortunately it is difficult to make a reliable assumption about the frequency of these payments and the nature of the recorded figure (i.e. did respondents remember to multiply quarterly payments by four before dividing by twelve to get a monthly figure?) As such, we have treated all aspects of this data consistently, i.e. by dividing by twelve if the amount exceeds 25% of the gross monthly wage. Given the small number of respondents who claimed to receive piece rate payments (25 in total), these assumptions are unlikely to significantly affect the final income estimates.

Table Six: Summary of the piece rate pay variable

Variable	*Obs*	*Mean*	*Std. Err.*	*[95% Conf.*	*Interval]*
Piece rate payments from first job	21	R202.91	57.8941	82.13971	323.6698
Piece rate payments from second job	4	R 63.17	22.1679	-7.381547	133.7149

[11] The questionnaire asked about similar bonus payments for the second job and bonus payments received when the business was doing well. These values are not shown here as less than 30 respondents claimed to receive these payments.

4) Arriving at total monthly wage income

One can now derive total monthly wage income by adding together gross monthly wages, monthly overtime pay and monthly bonus payments (including piece rate payments) using the measures discussed above. As mentioned previously, individuals classified as non-labour force participants or unemployed[12] were not classified as missing respondents, their incomes were assumed to be zero if they had given no alternative income estimates. This means that all respondents in the adult dataset should have a wage income amount, even if that amount is zero. The reader is reminded that many of the missing responses for total monthly wage income are in fact captured in proxy and emergency questionnaires; however these questionnaires have not been used for the reasons detailed earlier. This process was then repeated for the second wage job and income from wage earning activities for both jobs was added together under the broad banner of "gross monthly wage income". The resulting value looks as follows:

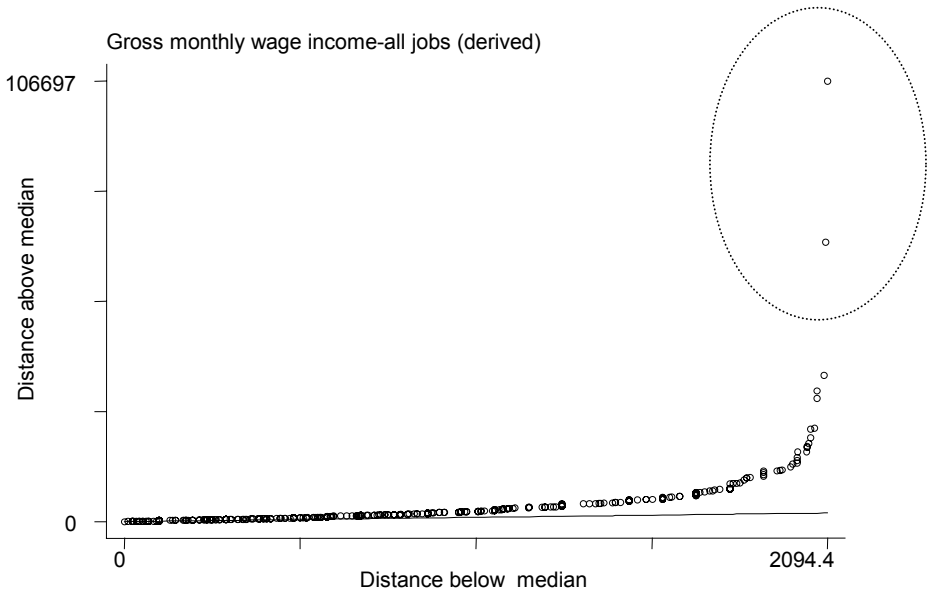
Table Seven: Derived gross monthly wage income from all jobs[13]

Obs	2115
Median	0
Mean	R1286.659
Std. Dev.	R3952.106
Variance	1.56e+07
Skewness	13.85656
Kurtosis	313.4484

[12] This is in line with the classifications used by Nattrass (2002)

[13] This table includes wage incomes of zero for the unemployed and for non-participants in the labour force.

Figure Four: Plotting gross monthly wage income from all jobs



The symmetry plot and table above clearly show the effect of outliers. The question now posed is how best to deal with these observations? The highest gross monthly wage income in the dataset earns R108 800 per month. This wage is reported by a cleaner for a rugby club and is distorted by the fact that she claims to receive R1200 per hour of overtime worked and to have worked 90 hours of overtime in the past month. This is likely to be a recording error and as such, we have decided to drop this observation from the following analysis. A similar problem occurs with the second obvious outlier, a respondent who claims to earn R1800 per hour of overtime and to have worked 17 hours of overtime in the past month. We treat this observation in the same way as the previous one and drop it from the following analysis.

Dropping these outliers affects the standard deviation of gross monthly wage income and reduces the mean gross monthly wage income by R83.83 or approximately 7%. *Table Eight* shows the values of gross individual income without outliers.

Table Eight: Derived gross monthly wage income - all jobs (controlling for outliers)

Obs	2113
Median	0
Mean	R1203.329
Std. Dev.	R2815.688
Variance	7928099
Skewness	5.237873
Kurtosis	44.8186

Dropping the two obvious outliers also affects the previous estimates of overtime pay and the new means are as follows:

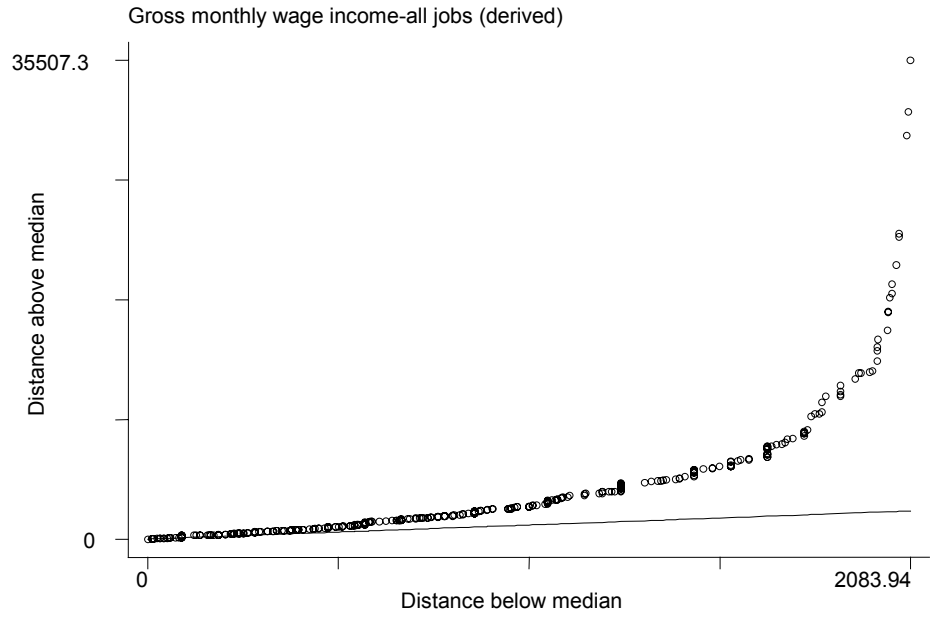
Table Nine: Summarising the overtime variable without the outliers

Variable	*Obs*	*Mean*	*Std. Err.*	*[95% Conf.*	*Interval]*
Overtime pay for first job	224	R665.34	196.1512	278.7902	1051.884
Overtime pay for second job	8	R818.25	719.8776	-883.9901	2520.49
Hours spent working overtime in first job last month	224	R16.11	1.406246	13.33591	18.87837

Graphing the individual wage income variable without the effect of the two obvious outliers (see *Figure Five*) still shows some observations which are significantly different from the mean but one cannot necessarily conclude that these are erroneous.[14]

[14] All subsequent analysis excludes the two outliers.

Figure Five: Plotting gross monthly wage income from all jobs without the outliers



Bonus payments (including piece rate payments) and overtime payments each comprise a significant portion of total monthly wage income amongst positive wage income earners. Overtime payments comprise approximately 14% of total wage income for those who receive overtime pay. Similarly, bonus payments comprise approximately 10% of total monthly wage income. We have not yet begun to compare household income provided by a single household respondent[15] with household income derived from the income of all working household members, but this comparison is our primary objective. Given that bonus and overtime payments are likely to be less regular than standard wage payments, subsequent regression analysis probes the possibility that the single household respondent is not reporting these payments received by other household members and that this accounts for some proportion of the difference between household income estimates.

[15] In response to question 16 on the Household Module

Table Ten: Bonus and overtime payments as a proportion of total wage income

Variable	*Obs*	*Mean*	*Std. Err.*	*[95% Conf.*	*Interval]*
Total bonus payments as a proportion of total wage income (including piece rate payments)	773	.1000165	.0086649	.083007	.1170261
Total overtime payments as a proportion of total wage income	224	.1416544	.0172129	.1077337	.1755751

5) Adding Other Forms of Income

To this wage income, we add (where applicable) income from respondents' own businesses, income from casual work, income from grants and investments, transfers from other people not living in the household, and any other forms of income not falling into these categories. The individual estimates are added by household to arrive at a total household income figure.[16]

The household income estimate generated by adding up the incomes of adult household members is henceforth referred to as "derived household income". Household income from the household module is referred to as "baseline household income". This terminology should not be taken to imply that one measure is necessarily more accurate than another.

Derived gross monthly household income is summarised below in *Table Ten*. This income measure has a mean of R2465.35 which falls well outside the 90%, 95% and 99% confidence intervals of the baseline household income.

Table Eleven: Derived gross monthly household income (all sources)

Obs	1174
Sum of Wgt.	1174
Median	R833
Mean	R2465.352
Std. Dev.	R4635.597
Variance	2.15e+07
Skewness	6.07501
Kurtosis	66.99254

[16] Claiming that this is in fact total household income, assumes that no-one under 18 years of age works (or generates income) in the household.

While the derived household income is still prone to heaping, the following graph shows that the intervals between "heaps" are narrower than the intervals for the baseline household income.

Figure Six: Gross household income from the individual modules

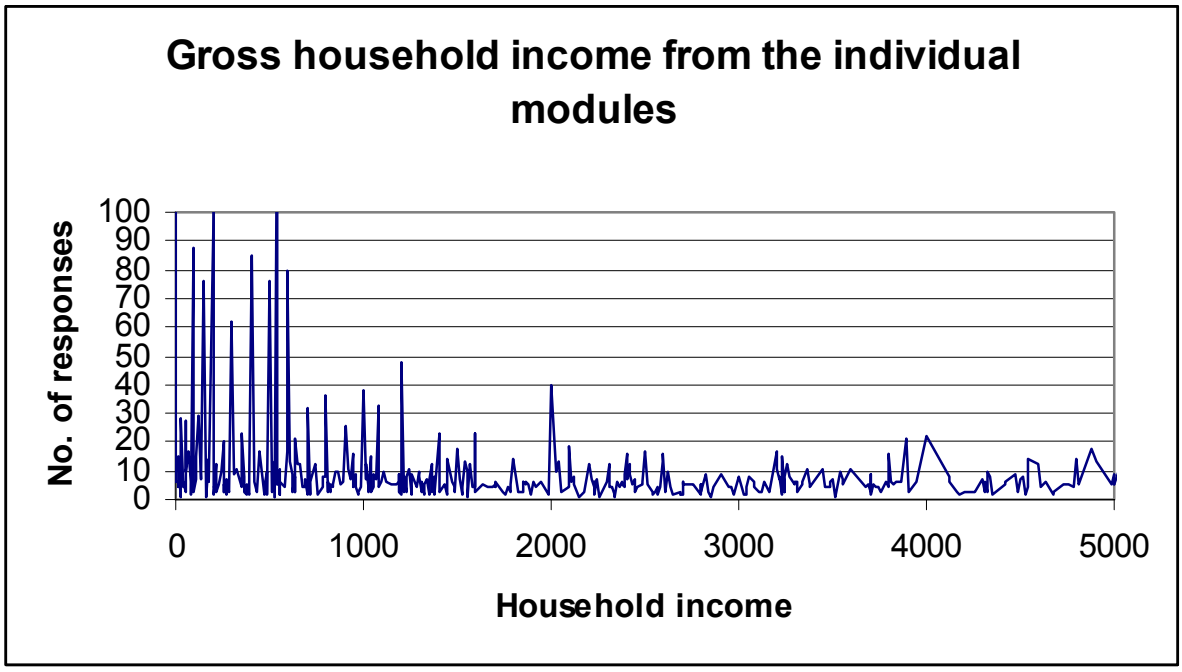


Table Twelve shows that, on average, wage income comprises more than half of gross monthly income from all sources.

Table Twelve: Wage income as a percentage of gross monthly income from all sources

Variable	*Obs*	*Mean*	*Std. Err.*	*[95% Conf.*	*Interval]*
Wage income as a percentage of gross monthly income from all sources	1220	.5463884	.0138515	.519213	.5735638

The remaining 45% of gross monthly individual income is comprised of a large number of disparate sources of income. These include income from self employment, from casual work, previously undisclosed casual income, state old age pensions, disability pensions, veterans pensions, employers pensions, workers compensation, UIF, State child support, private child maintenance grants, foster care grants, alimony payments, transfers from other people (not

household members), rental income, income from other financial investments and other forms of income. Small (i.e. statistically insignificant) base sizes in the majority of these categories prevents the detailed individual analysis of these components. However, their impact on the difference between household income measures will be explored on aggregate in subsequent regression analysis.

6) Arriving at a net estimate of income[17]

As we cannot know whether the household-level estimate of income is gross or net, it is necessary to derive both gross and net income estimates from the individual observations. Calculating net household income necessitates annualising the gross monthly income estimates and applying the tax schedule of the South African Revenue Services (SARS, 2000) (see *Table Thirteen*). This schedule is the simplest means of adjusting for tax payments and it does not include consideration for tax rebates, age-related tax thresholds or other deductions such as interest and dividends exemptions. The constraints of the dataset prohibit the extensive application of such tax adjustments and as such, a straightforward tax schedule is the only obvious means of adjustment.[18] Of course, any form of tax adjustment assumes that all respondents pay their taxes.

Table Thirteen: Tax Rates for Natural Persons (SARS, 2000)

Taxable income (R)		*Rates of Tax (R)*
0 - 35 000	18%	of each R1
35 001 – 45 000	R6 300 + 26%	of the amount above R35 000
45 001 – 60 000	R8 900 + 32%	of the amount above R45 000
60 001 – 70 000	R13 700 + 37%	of the amount above R60 000
70 001 – 200 000	R17 400 + 40%	of the amount above R70 000
200 001 and above	R69 400 + 42%	of the amount above R200 000

The net annual income is then converted into a monthly estimate to enable comparison and the resultant variable has the characteristics detailed in *Table Fourteen.* It is interesting to note that the mean of derived net household income falls just outside the 99% confidence interval of the baseline household income estimate. This suggests that household level respondents are giving net household income rather than gross.

[17] This net estimate is net of tax only and not net of other deductions.

[18] Analysts wishing to apply a more rigorous tax adjustment are referred to the South African Revenue Services (SARS) documentation.(SARS, 2000)

Table Fourteen: Derived net monthly household income (all sources)

Obs	1174
Sum of Wgt.	1174
Median	R683.06
Mean	R1854.033
Std. Dev.	R3088.922
Variance	9541438
Skewness	4.750367
Kurtosis	44.35492

Comparing income estimates

The KMP dataset has thus enabled the generation of three measures of household income: a baseline estimate from a household level question, a derived gross estimate from the individual level questionnaires and a derived net estimate from the individual level questionnaire. The correlation figure presented in the table below is a spearman rho correlation, a non-parametric measure that enables the rigorous testing of the relationship between variables. It shows relative rank and so does not assess income levels.

The spearman rho correlation was selected because it is less sensitive to outliers in the data. The spearman rho correlation is also well suited to instances such as this where the data is not normally distributed (Statacorp, 2001).

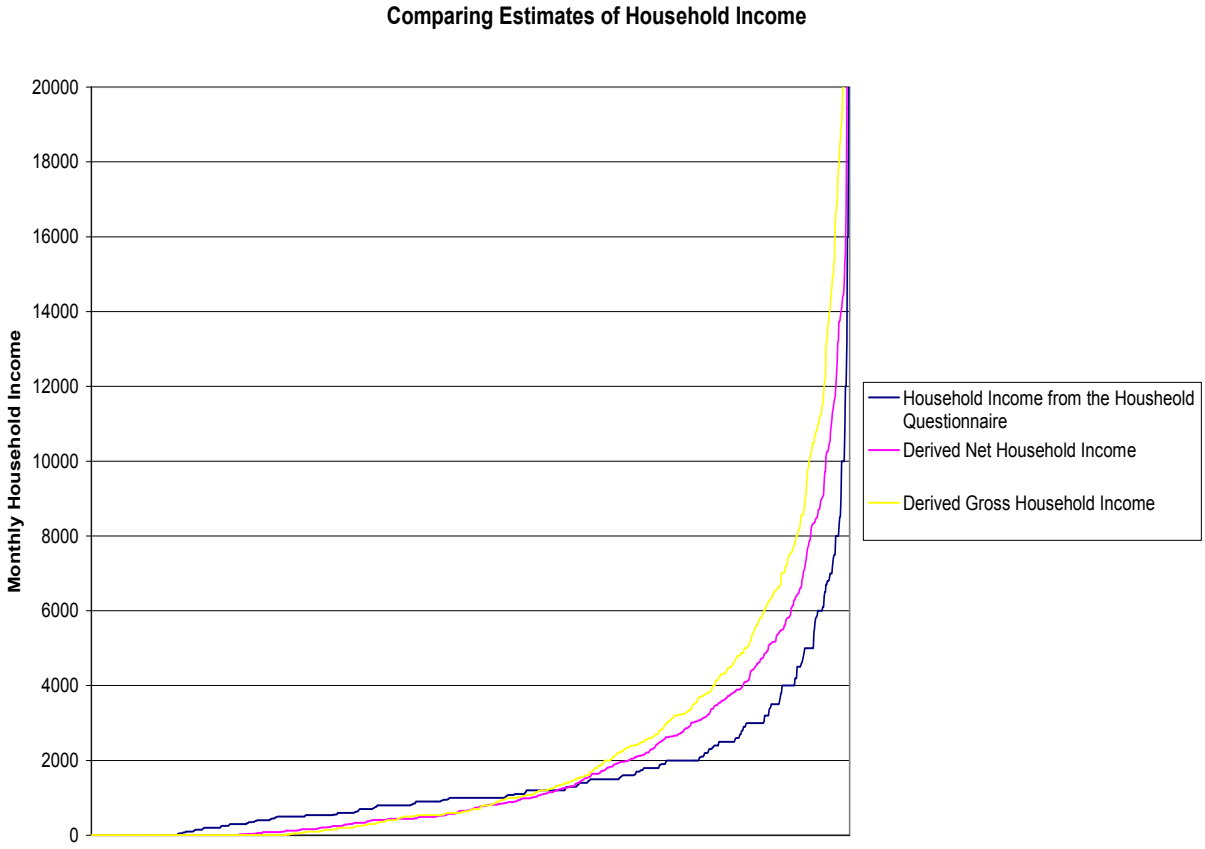
Comparing these estimates with one another, we see that there is less than 50% correlation between the derived estimates and the baseline estimate. If the manner in which household income was collected (i.e. from a single household respondent or from all income earning individuals in the household) then one would anticipate a correlation coefficient closer to 1. This lower result suggests that the way in which household income estimates are collected has a material impact on the responses received.

Table Fifteen: Correlation with the baseline estimate

	Mean	*Median*	*Correlation with baseline estimate*
Total Sample:			
Household Level Estimate (n=1086)	R1680.19	R 1000	1.0000
Derived Gross Household Income (n=1174)	R2465.35	R 833	0.4908
Derived Net Household Income (n=1174)	R1854.03	R683.06	0.4908

Figure Seven shows how the derived net and gross estimates of household income first lie below baseline household income and then above at high levels. Although estimates of household income go above R20,000 per month, the R20,000 "cut-off" point was chosen as this is the point at which the three estimates converge. This chart was generated by mapping the ordered distribution of one variable against the ordered distribution of another.

Figure Seven: Comparing estimates of household income

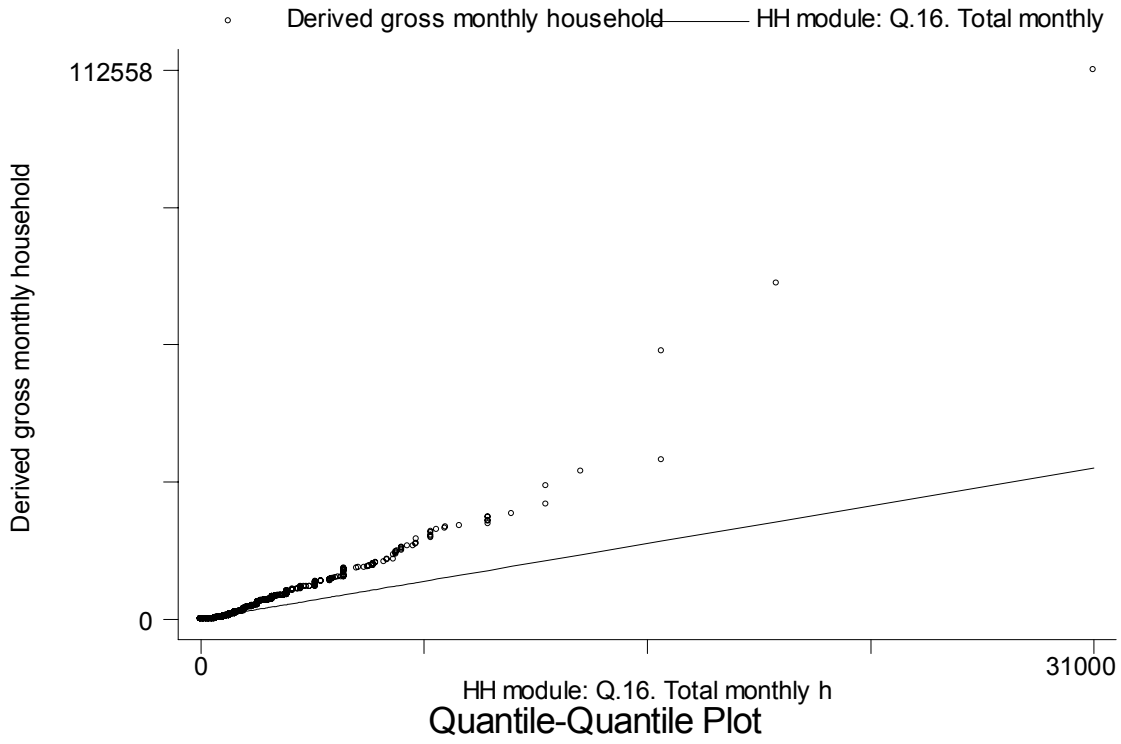


An alternate way of showing this information is to summarise the data in a quantile-quantile plot. Since sample quantiles are the data values in ascending order, a quantile-quantile plot is the equivalent of graphing the sorted values of one variable against the sorted values of another (Hamilton, 1992). If the two distributions were identical, they would lie on the diagonal line.

Figure Eight shows that the distribution of the two variables is similar at very low levels of income but diverge at higher levels of income. The quantiles of the derived variables are systematically higher than those of the income

estimates from the household modules. This applies to both the net and gross derived estimates.[19]

Figure Eight: Quantile-quantile plot of derived gross monthly household income



Understanding the Differences between Income Estimates

In order to probe the difference between the baseline household income and derived household income it is necessary to create a variable that deducts the baseline measure from the derived measure for each household. Where the baseline estimate exceeds the derived estimate, the difference would obviously

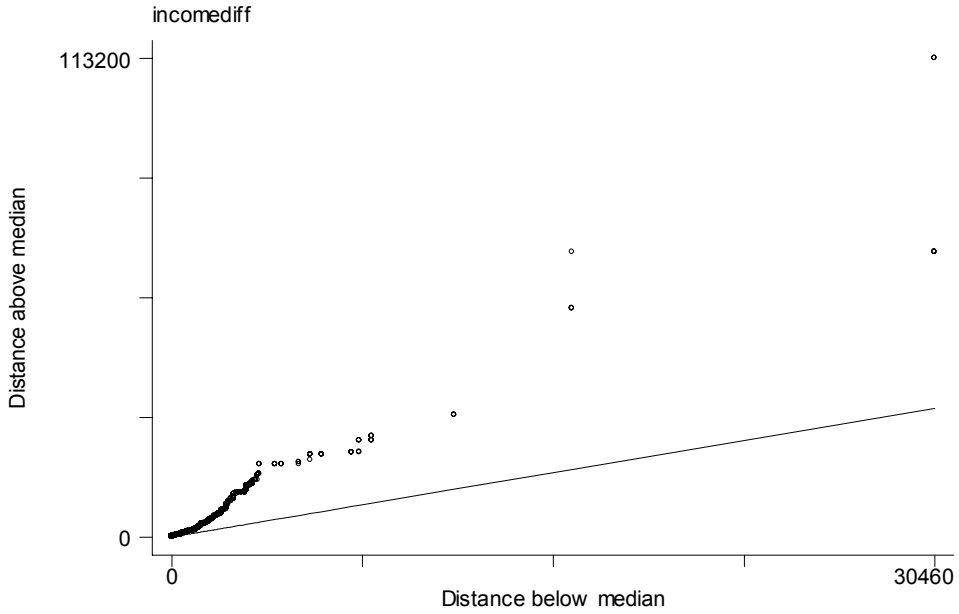
[19] Because derived net household income was calculated from derived gross household income (which includes fewer analytical assumptions), these two income estimates are perfectly correlated. They also share an almost identical relationship with the baseline household income estimate. For these reasons and for the purpose of parsimony, further comparison will be conducted using the derived gross income and the baseline figures only.

be negative. The mean difference however, is a positive R826,07 which, considering the median value of zero, suggests that the distribution of the data is skewed i.e. that the derived estimate exceeds the baseline estimate more often (or by a greater magnitude) than does the baseline estimate exceed the derived estimate. This is reinforced by the symmetry plot shown in *Figure Nine.*

Table Sixteen: Difference between derived gross household income and household income from the household module

Obs	1084
Sum of Wgt.	1084
Median	R0
Mean	R826.066
Std. Dev.	R4248.794
Variance	1.81e+07
Skewness	6.313676
Kurtosis	87.58781

Figure Nine: A symmetry plot of the difference between baseline and derived household income



The variable capturing the difference between the estimates is then used as the dependent variable in an OLS regression that attempts to identify the statistically significant determinants of the difference. The independent variables used in the model need to be household level variables rather than individual level variables and this limits the range of possible variables for inclusion.

At the risk of introducing multicollinearity into the model, both the baseline household income and the derived household income are included as independent variables in order to confirm (or refute) the finding that the difference between the estimates is greatest at higher levels of household income. A variable is also created to capture the number of age eligible adults identified in the household questionnaire who did not complete an adult questionnaire.

In order to investigate the possibility that single household respondent might be neglecting to include ad hoc payments such as payments for casual work, overtime and bonus payments, three new variables were created:

1. One capturing the proportion of total individual income comprised of non-wage income (including casual income, grants etc)
2. Another capturing the proportion of total income comprised of bonus payments; and
3. A third capturing the proportion of income comprised of overtime payments.

Only overtime as a proportion of total income was found to be statistically significant and to have a positive co-efficient. This means that as overtime increases as a proportion of total income, so too does the difference between the baseline and derived estimates. The following regression results do not contain these new variables as their inclusion dramatically reduces the base size of the regression from 1080 to 65.[20] Instead, a new regression model was specified without the "proportion variables". The results of this regression are set out in *Table Seventeen*.

[20] This occurs because the data processing programme used to generate the results only includes those observations for which there are values for every independent variable.

Table Seventeen: OLS regression on the difference in income estimates

Summary Statistics		
Number of obs	=	1080
F (13, 1066)	=	46.27
Prob > F	=	0.0000
R-squared	=	0.3607
Adj R-squared	=	0.3529
Root MSE	=	3415.8

Independent Variable	*Coef.*	*Std. Err.*	*t*	*P>t*	*[95% Conf.*	*Interval]*
Household size	20.51834	50.65277	0.41	0.686	-78.87211	119.9088
Coloured	-784.7417	272.9573	-2.87	0.004	-1320.336	-249.1472
Second quintile of baseline household income	-598.119	306.5856	-1.95	0.051	-1199.699	3.460703
Third quintile of baseline household income	-1038.516	344.7717	-3.01	0.003	-1715.024	-362.0075
Fourth quintile of baseline household income	-1540.215	364.7534	-4.22	0.000	-2255.931	-824.4987
Fifth quintile of baseline household income	-3405.39	423.265	-8.05	0.000	-4235.917	-2574.863
Second quintile of derived household income	162.1518	327.2145	0.50	0.620	-479.9058	804.2095
Third quintile of derived household income	1004.207	330.8535	3.04	0.002	355.0088	1653.405
Fourth quintile of derived household income	2629.681	355.8462	7.39	0.000	1931.442	3327.919
Fifth quintile of derived household income	8292.174	384.6123	21.56	0.000	7537.491	9046.857
No. of age eligible respondents missing from the adult module	-193.4438	152.7811	-1.27	0.206	-493.2296	106.342
Constant	-172.0929	324.6684	-0.53	0.596	-809.1546	464.9687

The adjusted R-squared of the regression above suggests that the independent variables in the model explain approximately 35% of the difference between the derived estimate of household income and the baseline estimate.

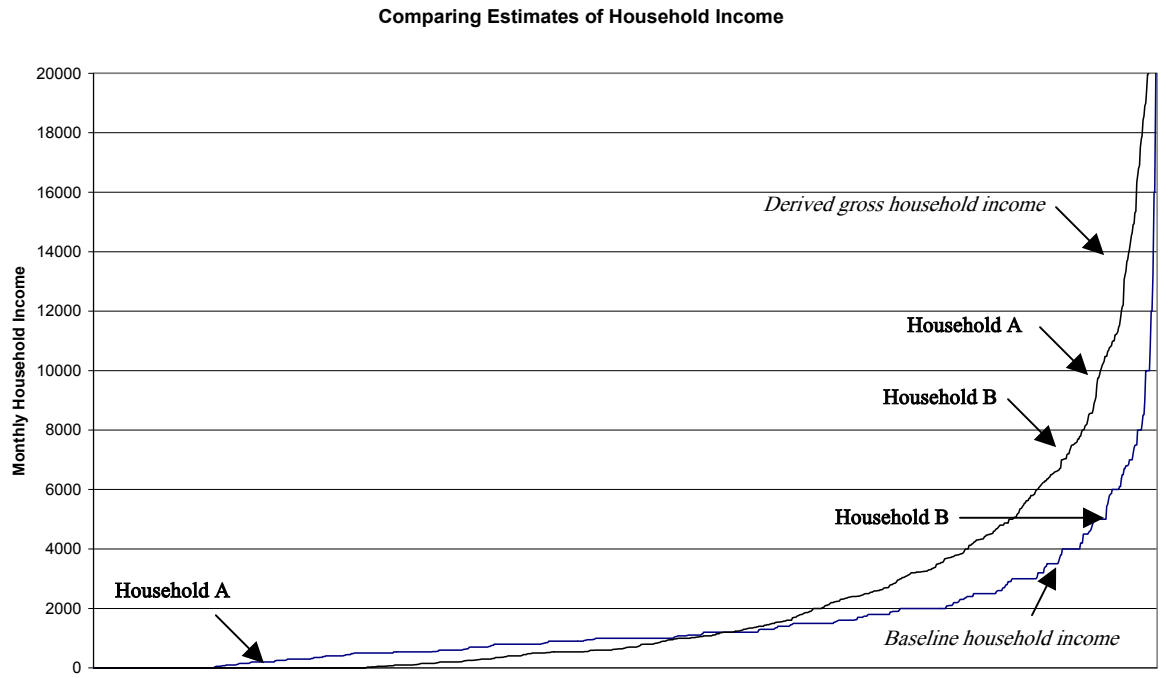
While household size is not a statistically significant predictor of difference (at the 95% confidence level), the population group of the household, the upper quintiles of the baseline household income, the upper quintiles of the derived estimate of household income and the number of age eligible household members who did not answer the adult questionnaire, are all statistically significant indicators at the 95% confidence level.

The third, fourth and fifth quintiles of the baseline household estimate are statically significant and they illustrate a definite pattern. Differences between the household income measures are lower for households with higher baseline incomes (i.e. households that fall into higher baseline income quintiles). Conversely, households with lower baseline incomes exhibit larger differences between the baseline and derived measures. This is consistent with the finding for the derived income measure where the difference between income measures is greater at the higher quintiles of derived income.

To illustrate this graphically, we refer to *Figure Ten,* showing the ordered distribution of the baseline and derived household income estimates. A single household could be positioned at different points on the two distributions and the regression tells us that this is in fact the case. According to the regression results, the difference between the two positions that one household may have is greater for households positioned in the lower quintiles of the baseline distribution.

Two hypothetical households are plotted on the following graph to show the nature and direction of this relationship. The specific points chosen for these households are somewhat arbitrary and are selected purely to impart a qualitative understanding of the relationship between the variables. Notice then that the difference in household income estimates is greater for Household A than it is for Household B and that each household has two points on the graph; one point on the derived distribution and one point on the baseline distribution.

Figure Ten: Plotting the positions of two households on the ordered distributions of the two household income variables (i.e. baseline and derived gross income)



Being a coloured household as opposed to a black/African household also has a significant effect on the dependent variable, decreasing the difference between the derived and baseline estimates by approximately R784.74.

The number of age eligible respondents missing from the adult sample does not appear to be a statistically significant indicator of the difference in income estimates at the 95% confidence level.

What Difference Does This Make in Practice?

In short then, household income generated from information provided by a single household respondent generates lower estimates of household income than does household income derived from the individual responses of all age eligible household members.[21] While the KMP dataset gives the option of using

[21] This is true even when the derived estimated is expressed in net terms.

either variable (i.e. baseline or derived) this is a luxury afforded by few other datasets. For this reason it is important to understand the practical implications of the difference in income estimates. If the differences have no effect on the ultimate analysis for which they are used, then the choice is arbitrary. However, if the choice does have an effect on final outcomes, the analyst needs to be aware of possible bias.

The following table shows the Gini coefficients calculated from each measure of household income.[22]

Table Eighteen: Gini Coefficients for the different income variables

Variable	*Gini*	*Bias*	*Std. Err.*	*[95% Conf. Interval]*	
Baseline Household Income	.4980588	-.0036703	.0207521	.4568823	.5392354
Derived Gross Household Income	.6263095	-.0024316	.0209909	.584659	.66796
Derived Net Household Income	.6005475	-.0006094	.0174101	.566002	.635093

These Gini coefficients illustrate that the different measures of income do in fact result in different measures of household inequality.[23] The baseline estimates show the lowest levels on inequality and the derived gross estimates the highest. If we accept the hypothesis that capturing the "missing" adults from the individual questionnaire would increase the difference between the derived gross estimates and the baseline, then we might see even greater differences between the measures of inequality.

It is concerning that in a country such as South Africa, which is undergoing significant restructuring, we may not be accurately measuring the extent of existing income inequality in areas such as Khayelitsha and Mitchell's Plain.

[22] Note that the KMPS sample was drawn from the Khayelitsha/ Mitchell's Plain area of Cape Town and that the Gini coefficients presented reflect the relative inequality within those areas only.

[23] The reader is reminded that this is a household level Gini and that the Gini coefficient varies between zero and one. A perfect equal society will have a Gini of zero and a perfectly unequal society will have a Gini of one (Deaton, 1998).

Conclusions

Estimates of household income are dependent on the manner in which income data is collected. The experience of the KMP survey suggests that a household level question may result in a systematic downward bias in the estimates, particularly at higher income levels. We know the direction of the bias from the preceding comparison with the derived estimates but we do not know the exact extent of the bias as the comparator itself is incomplete due to non-response.

The differences between income estimates have a material impact on the secondary analysis. The Gini coefficient was used to illustrate how different estimates of household income can yield different estimates of household income inequality for the same sample.

The KMPS experience also suggests that asking household income at the individual level is particularly vulnerable to non-response. However, techniques exist for the treatment of such non-response (see Skordis and Welch, 2002). By comparison, few techniques exist to account for the systematic downward bias of a response when the extent of that bias is uncertain and inconstant (i.e. when the extent of the bias changes over the range of the variable) as is the case when household income is asked at the household level. As such, while household income may be less costly to collect at the household level and may be less prone to the problems of non-response, these estimates do pose their own problems, which may be more difficult to account for in the final analysis.

If the analyst has no choice but to use household income estimates collected at the household level, then care should be taken to identify the likely downward bias of the estimates and the impact of this bias (as well as the lack of variation in the estimates) on the secondary analysis.

Bibliography

Deaton, A. 1998. *The Analysis of Household Surveys: A Microeconometric Approach to Development Policy*. The Johns Hopkins University Press for the World Bank, Baltimore.

Hamilton, L. C. 1992. *Regression with Graphics: A Second Course in Applied Statistics* , Duxbury Press, Belmont, California.

Kingdon, G. & J. Knight. 2000. Are Searching and Non-searching Unemployment Distinct States when Unemployment is High? The Case of South Africa. Unpublished paper, Centre for the Study of African Economies, Oxford University.

Lanot, G. 2002. The Relative Effect of Family Characteristics and Financial Situation on Educational Achievement. *Education Economics*, 10, 2, 165.

Mandel, M. J. 2002. The Rich Get Richer, and That's OK. *Business Week*, 3796, 88.

Nattrass, N. 2002. Unemployment, Employment and Labour Force Participation in Khayelitsha/ Mitchell's Plain. University of Cape Town, Centre for Social Science Research *Working Paper* No. 12 , October.

Podder, N. and S. Chatterjoo. 2002. Sharing the national cake in post reform New Zealand: income inequality trends in terms of income sources. *Journal of Public Economics*, 86, 1-27.

Quisumbing, A. R., L. Haddad & C. Peña. 2001. Are women overrepresented among the poor? *Journal of Development Economics*, 66, 225-269.

SALDRU/Centre for Social Science Research. 2000. Khayelitsha/ Mitchell's Plain Survey, University of Cape Town.

SARS. 2000. *Budget Review 2000*. Pretoria, South African Revenue Services.

Statacorp. 2001. *Stata Statistical Software: Release 7.0*, College Station, TX:Stata Corporation, Reference Manuals and On-Line Assistance.

Skordis, J. & Welch, M. 2002. Measuring the Impact of Non-Response on Household Income in the Khayelitsha/Mitchell's Plain Survey. University of Cape Town, Centre for Social Science Research *Working Paper* (forthcoming).

Wang, Z., C. M. Patterson & A. P. Hills. 2002. Association between overweight or obesity and household income and parental body mass index in Australian youth: analysis of the Australian National Nutrition Survey. *Asia Pacific Journal of Clinical Nutrition*, 11, 3, 200.

THE CENTRE FOR SOCIAL SCIENCE RESEARCH

Working Paper Series

RECENT TITLES

12/02	*Unemployment, Employment and Labour-Force participation in Khayelitsha/Mitchell's Plain*	By N. Nattrass
13/02	*Devising Social Security Interventions for Maximum Poverty Impact*	By S. van der Berg & C. Bredenkamp
14/02	*Perceptions of and Attitudes to HIV/AIDS among young adults at the University of Cape Town*	By S. Levine & F. Ross
15/02	*A Matter of Timing: Migration and Housing Access in Metropolitan Johannessburg*	By J. Beall, O. Crankshaw & S. Parnell
16/02	*Popular Attitudes towards the South African Electoral System*	By R. Southall & R. Mattes
17/02	*Are Urban Black Families Nuclear? A Comparataive Study of Black and White South African Family Norms*	By M. Russell
18/02	*AIDS and Human Security in Southern Africa*	By N. Nattrass
19/02	*Public Works as a Response to Labour Market Failure in South Africa*	By A. McCord
20/02	*Race, Inequality and Urbanisation in the Johannesburg Region, 1946-1996*	By O. Crankshaw & S. Parnell
21/02	*The "Status" of Giving in South Africa: An Empirical Investigation into the Behaviour and Attitudes of South Africans towards Redistribution*	By C. Pengelly
22/02	*How Important is Education for Getting Ahead in South Africa?*	By M. Keswell & L. Poswell
23/02	*Missing Links? An Examination of the Contributions made by Social Surveys to our Understanding of Child Well-Being in South Africa*	By R. Bray
24/02	*Unemployment and Distributive Justice in South Africa: Some Inconclusive Evidence From Cape Town*	By J. Seekings

The Centre for Social Science Research

The CSSR is an umbrella organisation comprising five units:

The Aids and Society Research Unit (ASRU) supports quantitative and qualitative research into the social and economic impact of the HIV pandemic in Southern Africa. Focus areas include: the economics of reducing mother to child transmission of HIV, the impact of HIV on firms and households; and psychological aspects of HIV infection and prevention. ASRU operates an outreach programme in Khayelitsha (the Memory Box Project) which provides training and counselling for HIV positive people

The Data First Resource Unit ('Data First') provides training and resources for research. Its main functions are: 1) to provide access to digital data resources and specialised published material; 2) to facilitate the collection, exchange and use of data sets on a collaborative basis; 3) to provide basic and advanced training in data analysis; 4) the ongoing development of a web site to disseminate data and research output.

The Democracy In Africa Research Unit (DARU) supports students and scholars who conduct systematic research in the following three areas: 1) public opinion and political culture in Africa and its role in democratisation and consolidation; 2) elections and voting in Africa; and 3) the impact of the HIV/AIDS pandemic on democratisation in Southern Africa. DARU has developed close working relationships with projects such as the Afrobarometer (a cross national survey of public opinion in fifteen African countries), the Comparative National Elections Project, and the Health Economics and AIDS Research Unit at the University of Natal.

The Social Surveys Unit (SSU) promotes critical analysis of the methodology, ethics and results of South African social science research. One core activity is the Cape Area Panel Study of young adults in Cape Town. This study follows 4800 young people as they move from school into the labour market and adulthood. The SSU is also planning a survey for 2004 on aspects of social capital, crime, and attitudes toward inequality.

The Southern Africa Labour and Development Research Unit (SALDRU) was established in 1975 as part of the School of Economics and joined the CSSR in 2002. SALDRU conducted the first national household survey in 1993 (the Project for Statistics on Living Standards and Development). More recently, SALDRU ran the Langeberg Integrated Family survey (1999) and the Khayelitsha/Mitchell's Plain Survey (2000). Current projects include research on public works programmes, poverty and inequality.