Centre for Longitudinal Studies

CLS Cohort Studies

Working Paper 2008/10

Missing Income Data in the Millennium Cohort Study: Evidence from the First Two Sweeps

Denise Hawkes
Ian Plewis

December 2008

Missing Income Data in the Millennium Cohort Study: Evidence from the First Two Sweeps

Denise Hawkes and Ian Plewis
Institute of Education

December 2008

First published in December 2008 by the
Centre for Longitudinal Studies
Institute of Education, University of London
20 Bedford Way
London WC1H 0AL
www.cls.ioe.ac.uk

ISBN: 978-1-906929-04-6

The Centre for Longitudinal Studies (CLS) is an ESRC Resource Centre based at the Institution of Education. It provides support and facilities for those using the three internationally-renowned birth cohort studies: the National Child Development Study (1958), the 1970 British Cohort Study and the Millennium Cohort Study (2000). CLS conducts research using the birth cohort study data, with a special interest in family life and parenting, family economics, youth life course transitions and basic skills.

This document is available in alternative formats.
Please contact the Centre for Longitudinal Studies.
tel: +44 (0)20 7612 6875
email: info@cls.ioe.ac.uk

CONTENTS

Acknowledgements

We would like to thank Kelly Ward for access to a revised household income variable for MCS2. We would also like to thank the participants of the 29th General Conference of The International Association for Research in Income and Wealth in Joensuu, Finland and the Bedford Group Internal Seminar Series for their comments. All remaining errors are of course our own.

1. Introduction

The Millennium Cohort Study (MCS) is the fourth in a series of internationally renowned cohort studies in the UK. It includes 18818 babies in 18552 families born over a 12 month period and living in selected UK wards at age 9 months. Areas with high proportions of Black and Asian families, disadvantaged areas and the three smaller UK countries are all over-represented in the sample which is disproportionately stratified and clustered. The first and second sweeps took place when the cohort members were 9 months and 3 years old. In addition, at sweep two, families who were living in sampled areas with a child of the appropriate age but who were not located at the first sweep were introduced. These "new families" tend to be more mobile than those already part of the MCS. Partners were interviewed whenever possible and detailed questions about individual and household income were included in both sweeps.

There are four ways in which income data can be missing. There was unit non-response at sweep one such that the response rate then was 72%. There was further partner non-response; the partner response rate at sweep one, among respondent families with partners, was 88%. In addition there was item non-response for income: about 6% of main respondents and 6% of partners did not provide income data at sweep one. Moreover, there was attrition between sweeps one and two: 79% of eligible cases responded at sweep two. The correlates of unit and partner non-response at sweep one are set out in Plewis (2004); the evolution of the sample from sweep one to sweep two is described in Plewis and Ketende (2006).

The paper will address the following questions:

(i) Are there (a) within household and (b) within individual correlations for missing income data?

(ii) Is a female interviewer more successful than a male interviewer in getting responses to income questions from main respondents and their partners?

(iii) Is there a systematic tendency for income data to be missing at sweeps one and two over and above what we know about unit and partner non-response?

(iv) Is attrition at sweep two related to (a) household income at sweep one; (b) the failure to provide income data at sweep one?

The paper will conclude by considering the implications for statistical modelling and future data collection of our findings on the patterns and correlates of income non-response at both sweeps.

1.1 Background Literature

Unlike the previous UK cohort studies, MCS has been designed with a focus on social and economic data rather than health data. As a consequence the quality of the MCS dataset could be reduced by the failure of some participants to report their income. If income non-response were truly random then it would merely result in a loss of precision in any statistical analysis based on complete cases. However,

income non-respondents are often different from those who do provide income information. Therefore, any analysis of income undertaken without considering the type of people who do not provide income data could produce biased estimates.

Much has been written about the quality of income data obtained in surveys. A selection of these include Miller (1953) who compared the income data on the 1950 US census with that of the 1950 US Current Population Survey (CPS). He found that income is usually under-reported in surveys as respondents often forget minor or irregular sources of income. Miller (1953) also found that those who were self-employed were more likely to misreport their earnings. The self-employed were asked about their earnings separately to the employed at both sweeps one and two of the MCS. In addition MCS respondents were asked to report the income from their main job, which we shall consider in this paper, as well as being asked about earnings from second/occasional jobs.

Weinberg et al (1999) compare the CPS benchmarks from the National Income and Product Accounts supplemented with data from the Internal Revenue Service tax returns and the Social Security Administration. They consider the income data from the CPS from 1947 to 1997. They claim that the tendency to under-report income is largely from sources other than wages or salaries, for example asset income, and interest and dividend payments.

Siminski et al. (2003) compare the Australian Bureau of Statistics (ABS) Household Income Data with the Australian System of National Accounts, ABS population data and the Department of Family and Community Services expenditure data. They note that the under-reporting of income is not restricted to the bottom end of the income distribution. They give various reasons why survey data sets and their external data sources provide different reports. These include (i) a problem with the external data source, (ii) different scopes of the surveys, (iii) different definitions used to define income groups, (iv) the appropriateness of the weights used and (v) the misreporting of income.

Rodgers et al. (1993) consider measurement error in income data for the Panel Survey of Income Dynamics (PSID) validation study by comparing employee and employer reports of earnings. This study is limited to the male employees of a single large manufacturing firm but it is found that hourly wages are the most likely to suffer from measurement error. Jäckle et al (2004) used a sample of low income respondents from the European Community Household Panel Survey (ECHP) and undertook a validation study. They compared the income data obtained from employers' records and government benefit data from the Department for Work and Pensions (DWP). They found that obtaining consent from respondents to contact their employer was more difficult than obtaining consent to contact the DWP about benefit records.

This paper will, however, focus less on the quality of the data actually obtained in the MCS and more on item non-response associated with the income data. Rodgers et al. (1993) cite earlier work on the PSID data by Duncan et al. (1985) who found from company records those who were unit non-responders had earnings 5.5% higher than earnings respondents, and those who were item non-responders had earnings 11.3% lower than earnings respondents.

Nicoletti et al. (2001) consider the effects of non-sampling errors on the quality of various income measures in the ECHP. They consider both the effect of unit non-response, distinguishing between attrition and new/re-entry participants, as well as item non-response, distinguishing between those who provide partial information and those who do not respond to any income questions. They find that for those households with both unit and partial item non-responses final imputed household income is higher than for responding households. In addition, for those with complete income non-response final imputed household income is lower than for responders. They also find that full income non-response is common for those who are self-employed.

Essig et al. (2003) use a controlled field experiment to consider interviewer and mode (personal interview and drop off questionnaire) effects on item non-response to income and financial questions. They find that whilst the mode of the survey affected non-response, that is the interview had a higher response rate than the drop off question, there was no additional effect on the respondents' propensity to respond to financial questions. They also found that respondent, household and interviewer characteristics do not have a strong and consistent effect on item non-response to income questions.

Lynn et al. (2004) used a sample of low income respondents from the ECHP to consider the effects of interview style on income response. They designed an experiment to compare dependent interviewing (both proactive and reactive) with traditional independent interviewing. Dependent interviewing was found to have less non-response for income than independent interviewing, especially for income sources which are relatively common or easy to forget.

Schräpler (2003) uses the British Household Panel Study (BHPS) to consider income non-response in panel studies. He finds that refusals and don't knows to gross income relate to different characteristics of respondents. Those who refuse are mainly male without dependent children whilst those who don't know are mainly females, in low or middle occupational groups and work irregularly. Interviewer/area effects are also found.

Riphahn et al. (2002) use the German Socioeconomic Panel (GSOEP) to consider item non-response on income and wealth questions. They find that if an interviewer is female, especially if the respondent is female, there is a higher non-response rate on income. In addition if the respondent is younger than the interviewer this improves the response to income questions. They also find that don't knows are different in their characteristics to other non-responders in the sample.

4

2. Income Non-Response in the Millennium Cohort Study (MCS)

2.1 Sweeps 1 and 2 separately

The MCS sample is made up of two groups of respondents: the original sample of 18552 families and an additional 692 families who were missed in the first sweep. For each family a main respondent was identified, who was usually the mother of the cohort child, and if possible a partner respondent was identified, who was usually the father of the cohort child. In both cases we shall consider only those who report being currently employed (either in post at present or on leave including maternity leave) and include both the employed and self-employed. For a small proportion of partners a proxy interview was completed by the main respondent. The proxy response to income is considered separately from the partner response. In both sweeps of the MCS the main respondent and their partner individually report their income. They were asked to report both their net and gross income if they were employed and their take home income if they were self employed.[1] For this analysis of income non-response we shall count someone as responding to the income question if they provide a response to their gross and/or net income if they are employed, and provide their net pay if self-employed.

For the 18552 original respondents in the first sweep, the distribution of the main respondent income response is given in Table 1(a). Of those who did not respond 1.8% were 'don't knows' and 0.9% refused to respond. Nearly all of those who are not eligible were not employed at present (51.4%).

Correspondingly for the partner respondent (Table 1(a)): those who did not respond can be divided into 'don't knows' (2.1%) and refusers (2.1%). Those who were ineligible can be divided into those who were not employed at present (8.2%), those who completed the proxy questionnaire (1.0%), those where no partner lives in the household, that is lone parents, (13.9%) and other not applicable (8.0%).

At sweep two 14898 of the 18552 original respondents were interviewed and 4.4% either refused or did not know their income, a higher proportion than for MCS1 (see Table 1(a)). Moreover, for the partner respondents, 8.7% refused or did not know their income compared with 4.3% in MCS1. Excluding the 'not applicable' group, we find that item non-response for income goes up from 5.6% to 8.0% for the main respondents and from 6.2% to 12% for the partners.

[1] Those employed are asked the following two questions:

1. Last time you were paid (in your main job) what was your total take home pay – that is after all deductions for tax, National Insurance, union dues, pension and so on, but including overtime, bonuses, commission and tips? Range 1..999997 [refuse, don't know, missing]
2. And the last time you were paid what was your gross pay – that is before any deductions? Range 1..999997 [refuse, don't know, missing]

Those who were self employed were asked:

1. I know that it is sometimes difficult for self employed people to give an exact figure for their income, but could you please think about your take home income in the last 12 months. That is, the amount you personally took out of the business after all taxes and costs. About how much is this? Range 1..999997 [refuse, don't know, missing]

Table 1 summarises the income response for sweeps one and two by type of respondent: original sample, new families and proxy respondents. Both the 'new family' respondents and the proxy respondents have higher rates of non-response among the eligible sample than the original sample (Tables 1(b) and 1(c)). The lower response with the proxy is to be expected as the main respondent is less likely to know the partner's actual earnings than the partner themselves. We shall focus on the responses from the original sample for the rest of the paper.

Table 2 shows the relationship between main and partner respondents' responses to the income questions at each sweep for the original sample. The first panel contains the within family income response for MCS1. We can see that if the main respondent responds to the income question then the partner is also most likely to respond to their income question (78.5%). If the main respondent does not respond to the income question then 26.6% of partners also do not respond. A similar pattern of results is found for MCS2.

Table 1: Pattern of Income Non-Response, MCS sweeps 1 and 2

(a) Original Sample

	Sweep One		**Sweep Two**	
	Main	**Partner**	**Main**	**Partner**
income response	45.9%	64.7%	50.6%	62.9%
don't know/refusal	2.7%	4.3%	4.4%	8.7%
not applicable	51.5%	31.0%	45.1%	28.4%
sample	18552		14898	

(b) Sweep Two including New Families

	New Families Only		**All Families (New & Original)**	
	Main	**Partner**	**Main**	**Partner**
income response	27.9%	42.4%	49.5%	61.9%
don't know/refusal	5.7%	11.5%	4.4%	8.9%
not applicable	66.4%	46.2%	46.1%	29.3%
Sample	692		15590	

(c) Proxy

	Sweep One	**Sweep Two**
income response	32.9%	59.3%
don't know/refusal	22.0%	39.6%
not applicable	45.1%	1.1%
sample	338	226

NOTES:
1. weighted percentages, unweighted observations

The second panel contains the within family income response for sweep two of the MCS. Similar patterns to panel one can be seen. However, we can see that there are generally larger proportions of respondents in each cell who don't know or refuse to respond to the income questions.

6

Table 2: Within Family Income Response by MCS Sweep

SWEEP ONE		Partner respondent			
		don't know/refusal	not applicable	income response	Total
Main Respondent	don't know/refusal	26.6% 16.6% 128	27.4% 2.4% 147	45.9% 1.9% 189	100% 2.8% 464
	not applicable	3.9% 46.2% 418	42.7% 71.0% 5135	53.4% 42.5% 4711	100% 51.5% 10264
	income response	3.5% 37.2% 278	18.0% 26.6% 1685	78.5% 55.6% 5861	100% 45.9% 7824
	total	4.3% 100% 824	31.0% 100% 6967	64.7% 100% 10761	100% 100% 18552

SWEEP TWO		Partner respondent			
		don't know/refusal	not applicable	income response	Total
Main Respondent	don't know/refusal	26.7% 13.3% 163	29.0% 4.4% 200	44.3% 3.1% 251	100% 4.3% 614
	not applicable	9.6% 49.3% 728	36.5% 58.1% 2982	54.0% 38.6% 3480	100% 45.1% 7190
	income response	6.5% 37.4% 473	21.0% 37.5% 1697	72.5% 58.3% 4924	100% 50.6% 7094
	total	8.7% 100% 1364	28.4% 100% 4879	62.9% 100% 8655	100% 100% 14898

NOTES:
1. weighted percentages, unweighted observations
2. each cell contains: row %, column % and observations

2.2 Across Sweeps 1 and 2

Table 3 shows the relationship between each respondents' response to income questions in sweeps one and two. This has been restricted to those who are the same respondent across the two sweeps. The first panel contains the within individual income response across the sweeps of the MCS for the main respondent. If the main respondent provided income data in sweep one they are most likely to provide income data at sweep two (79.9%). If the main respondent was not applicable in sweep one they are largely not applicable in sweep two (74.4%). This group is mostly those who have not been in the labour market at each of the two sweeps. The main respondent was more likely not to report their income in sweep two (4.4%) than in sweep one (2.6%).

The second panel contains the within individual income respondent across sweeps of the MCS for the partner respondent. If the partner responded to income at sweep

one they are highly likely to respond to income in the second sweep (91.2%). There are only 82 cases in sweep two that are not applicable. Finally if the partner respondent did not respond to the income questions at sweep one, 35.2% also refused at sweep two, a higher proportion than for the main respondent (17.9%).

Tables 2 and 3 tell us that there are substantial within household correlations in response behaviour: item non-response by the main respondent predicts item non-response by the partner. There are also important within individual correlations across sweeps: a don't know or refusal at sweep two is more likely if there was a don't know or refusal at sweep one. On the other hand, both Tables 2 and 3 show considerable movement across response categories: a don't know or refusal by the main respondent is more likely to be accompanied by a response rather than a non-response from the partner and those who are item non-respondents at sweep one are more likely than not to be respondents at sweep two.

Table 3: Within Individual Income Response across MCS Sweeps

MAIN		**Sweep Two**			
		don't know/refusal	**not applicable**	**income response**	**Total**
Sweep One	don't know/refusal	17.9% 10.4% 64	26.7% 1.5% 87	55.4% 2.8% 206	100% 2.6% 357
	not applicable	2.9% 32.0% 198	74.4% 82.5% 5920	22.8% 22.0% 1615	100% 49.3% 7733
	income response	5.3% 57.6% 347	14.8% 16.0% 953	79.9% 75.2% 5204	100% 48.1% 6504
	total	4.4% 100% 609	44.5% 100% 6960	51.1% 100% 7025	100% 100% 14594

PARTNER		**Sweep Two**			
		don't know/refusal	**not applicable**	**income response**	**Total**
Sweep One	don't know/refusal	35.2% 13.9% 174	0.4% 4.3% 4	64.4% 3.5% 323	100% 4.7% 501
	not applicable	22.9% 27.1% 421	2.4% 73.4% 59	74.7% 12.0% 1298	100% 14.1% 1778
	income response	8.7% 59.0% 707	0.1% 22.4% 19	91.2% 84.5% 6707	100% 81.2% 7433
	total	11.9 100% 1302	0.5% 100% 82	87.6% 100% 8328	100% 100% 9712

NOTES:
1. weighted percentages, unweighted observations
2. each cell contains: row %, column % and observations
3. only including providing an interview at both sweeps one and two, therefore excluding unit non responders at sweeps one and two
4. restricted to those who are the same main and partner respondents at both sweeps

3. Modelling Income Non-Response

Table 4 presents estimates for models which predict income non- response for the main respondent at sweeps one and two. The dependent variable is 1 if the main respondent refused to respond or didn't know their income and 0 if they provided income data. Those who were not eligible for the income question have been excluded from this analysis. As the dependent variable is a binary variable these models have been estimated using a logistic regression (allowing for the survey design).

The sex of the interviewer is only known for sweep one. The first row of Table 4 shows that being a male interviewer (18% of all interviewers) is not a statistically significant predictor of income non-response. The strongest and most consistent predictor is self-employment status. This matches with much of the literature discussed above. Social class and country have an important effect on non-response for sweep one only. Northern Ireland has higher odds of non-response than the reference category, England in sweep one.

For sweep two only, an important predictor is if the main respondent has a partner. Lone parents who are employed are less likely to respond to the income questions. In addition, at sweep two, an Indian ethnic background increases the chances of a non-response at sweep two relative to the reference category, white respondents. Also the Northern Ireland effect found in column 4 is removed in column 5 once we condition on response at sweep one. If the main respondent was a non-responder to the income question at sweep one they are more likely to be a non-responder to the income question at sweep two. Finally, if the main respondent is the same main respondent as at sweep one this increases the chance of a non-response at sweep two.

Table 5 presents the corresponding estimates to Table 4 for the partner respondent at sweeps one and two. Once again the sex of the interviewer is not a significant predictor of income non response of the partner. Also, as with the main respondent, the strongest and most consistent predictor is self employment status across the two sweeps. Northern Ireland has higher odds of non-response than the reference category, England consistently across the two sweeps for the partner respondent. Ethnicity has a positive and significant effect on income non-response across the sweeps.

For sweep one the older the partner is at interview predicts income non-response. Also being in the social classes 'lower supervisors and technical' and 'semi routine and routine' improve the chances of response to the income questions. The more educated partners are more likely to respond to income questions, as measured by the NVQ levels. At sweep one Wald tests on social class, NVQ levels, ethnicity and country find all four sets of variables have an important effect on model fit.

For sweep two those living in owner occupied housing are predicted to be more likely to respond to the partner income questions. Those who did not respond at sweep one are more likely not to respond at sweep two. As with the main respondents, this knocks out the Northern Ireland effect. Finally, if the partner respondent is the same

9

respondent across the two sweeps this increases the likelihood of an income response at sweep two. At sweep two, Wald tests on social class, NVQ levels, ethnicity and country show that only ethnicity and country have an important effect on model fit.

Table 6 presents multinomial logistic regressions for non response at sweep one to consider the relative impacts of the explanatory variables on refusals and don't knows separately. The dependent variable is 2 if the main respondent refused to respond, 1 if they didn't know their income and 0 if they provided income data (this is the reference category). For the main respondent (columns one and two) self employed status remains important only for those who report that they do not know their income. In addition to self employment, social class is an important predictor of a don't know response for the main respondent. For those main respondents who refuse to report their income only the Northern Ireland category is a significant predictor.

Columns three and four of Table 6 present estimates for the same analysis for the partner respondent. A don't know response to the income question is predicted by self employment status, working for a small employer, not being educated to first degree level or equivalent (NVQ Level 4) and living in Northern Ireland. A refusal is predicted by self employment status, being an older respondent, not having an occupation considered to be semi routine and routine, not being educated beyond GCSE (NVQ Level 2), having a larger family, belonging to an ethnic minority group and living in Northern Ireland.

Given the importance of self-employment status in predicting income non-response, Table 7 considers whether a change in self-employment status leads to a change in income response behaviour. The table presents separate panels for the main and partner respondents. We find an increase in self-employment across the two sweeps, especially for the partners. For the main respondent, and as predicted by Table 4, moving out of self-employment leads to a greater likelihood of an income response at sweep two whereas moving into self-employment reduces that likelihood. The picture for the partners is somewhat different in that both movement out of and into self-employment increase the chance of non-response at sweep two. The numbers of respondents who change employment status are, however, rather small.

10

Table 4: Main Respondent Income Non-Response in Sweeps One and Two

	Sweep One			Sweep Two		
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
Interviewer was male		1.2	1.3			
		(0.73 - 1.9)	(0.80 - 2.2)			
Main self employed	6.3		6.4	6.8	6.6	6.7
	(3.7 – 11)**		(3.8 – 11)**	(4.6 – 10)**	(4.4 - 9.8)**	(4.5 - 9.9)**
Main's age at interview	1.0		1.0	1.0	1.0	1.0
	(0.98 - 1.0)		(0.98 - 1.0)	(1.0 - 1.0)	(0.98 - 1.0)	(0.98 - 1.0)
Main has a partner	1.0		1.0	0.58	0.57	0.56
	(0.66 - 1.5)		(0.67 - 1.6)	(0.43-0.77)**	(0.42 - 0.76)**	(0.42 - 0.76)**
Social Class of Main Respondent: reference category Managerial and Professional						
Intermediate	1.5		1.6	1.1	1.1	1.1
	(1.0 - 2.3)*		(1.1 - 2.3)*	(0.80 - 1.5)	(0.78 - 1.5)	(0.79 - 1.5)
Small employers and self employment	1.8		1.8	1.3	1.0	1.0
	(1.1 - 3.1)*		(1.1 - 3.0)*	(0.86 – 2.0)	(0.68 - 1.6)	(0.66 - 1.5)
Lower supervisors and technical	0.85		0.86	1.1	1.1	1.2
	(0.44 - 1.7)		(0.44 - 1.7)	(0.55 - 2.2)	(0.57 - 2.2)	(0.59 - 2.3)
Semi routine and routine	1.4		1.4	0.93	0.95	0.96
	(0.90 - 2.0)		(0.90 - 2.0)	(0.61 - 1.4)	(0.62 - 1.5)	(0.62 - 1.5)
NVQ Levels: reference category none						
NVQ Level 1	1.2		1.3	1.5	1.4	1.4
	(0.61 - 2.5)		(0.61 - 2.6)	(0.56 - 3.8)	(0.54 - 3.8)	(0.54 - 3.7)
NVQ Level 2	0.85		0.85	0.76	0.77	0.76
	(0.50 - 1.4)		(0.50 - 1.4)	(0.34 - 1.7)	(0.34 - 1.8)	(0.34 - 1.7)
NVQ Level 3	0.74		0.74	1.1	1.1	1.1
	(0.40 - 1.3)		(0.40 - 1.3)	(0.44 - 2.7)	(0.45 - 2.7)	(0.44 - 2.6)
NVQ Level 4	0.89		0.89	0.84	0.87	0.85
	(0.51 - 1.5)		(0.51 - 1.5)	(0.35 - 2.0)	(0.36 - 2.1)	(0.35 - 2.0)
NVQ Level 5	1.2		1.2	1.1	1.1	1.1
	(0.64 - 2.3)		(0.64 - 2.3)	(0.41 - 2.9)	(0.40 - 2.9)	(0.39 - 2.9)

11

	Sweep One			Sweep Two		
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
Other/overseas quals only	1.3		1.3	1.2	1.3	1.3
	(0.57 - 2.8)		(0.57 - 2.8)	(0.31 - 4.5)	(0.35 - 4.8)	(0.34 - 4.8)
Cohort child is the first born	1.0		1.0	0.85	0.85	0.85
	(0.80 - 1.3)		(0.79 - 1.3)	(0.65 - 1.1)	(0.65 - 1.1)	(0.65 - 1.1)
Ethnicity: reference white						
Mixed	0.48		0.46	0.85	1.0	1.0
	(0.11 - 2.1)		(0.10 - 2.1)	(0.077 - 9.5)	(0.096 – 11)	(0.096 – 11)
Indian	1.5		1.4	2.4	2.3	2.3
	(0.75 - 2.9)		(0.72 - 2.7)	(1.4 - 4.2)**	(1.4 – 4.0)**	(1.4 – 4.0)**
Pakistani and Bangladeshi	1.6		1.6	0.58	0.46	0.45
	(0.62 - 4.1)		(0.61 - 4.1)	(0.19 - 1.8)	(0.12 - 1.7)	(0.12 - 1.7)
Black or Black British	1.6		1.6	0.88	0.80	0.81
	(1.0 - 2.5)		(1.0 - 2.5)*	(0.36 - 2.1)	(0.33 – 2.0)	(0.33 – 2.0)
Other ethnic group	0.98		1.0	2.3	2.2	2.1
	(0.34 - 2.8)		(0.35 - 2.9)	(1.0 - 5.0)*	(0.99 - 4.7)	(0.97 - 4.6)
Owner occupier	0.81		0.81	0.97	1.0	0.99
	(0.59 - 1.1)		(0.58 - 1.1)	(0.70 - 1.3)	(0.72 - 1.4)	(0.71 - 1.4)
Country: reference England						
Wales	0.82		0.79	1.2	1.2	1.2
	(0.55 - 1.2)		(0.52 - 1.2)	(0.88 - 1.6)	(0.88 - 1.6)	(0.88 - 1.6)
Scotland	1.4		1.4	1.1	1.0	1.0
	(0.68 - 2.8)		(0.69 - 2.7)	(0.71 - 1.5)	(0.68 - 1.5)	(0.67 - 1.5)
Northern Ireland	1.7		1.7	1.5	1.44	1.4
	(1.1 - 2.4)**		(1.2 - 2.5)**	(1.0 - 2.3)*	(0.94 - 2.2)	(0.94 - 2.2)
If main respondent not responded to income questions in sweep 1					3.0	3.0
					(2.0 - 4.6)**	(2.0 - 4.6)**
Main respondent the same respondent as sweep one						5.3
						(1.0 – 28)*
Observations	8190	8190	8190	5800	5800	5800

Note: 95% confidence intervals in parentheses * significant at 5%; ** significant at 1%

12

Table 5: Partner Respondent Income Non-Response in Sweeps One and Two

	Sweep One			Sweep Two		
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
Interviewer was male		1.2	1.3			
		(0.75 - 1.9)	(0.79 - 2.1)			
Partner self employed	1.7		1.7	3.6	3.6	3.6
	(1.3 - 2.3)**		(1.3 - 2.3)**	(2.7 - 4.8)**	(2.7 - 4.9)**	(2.7 - 4.9)**
Partner's age at interview	1.0		1.0	1.0	1.0	1.0
	(1.0 - 1.0)**		(1.0 - 1.0)**	(1.0 - 1.0)*	(1.0 - 1.0)	(1.0 - 1.0)
Social Class of Partner Respondent: reference category Managerial and Professional						
Intermediate	0.84		0.83	0.89	0.92	0.92
	(0.44 - 1.6)		(0.43 - 1.6)	(0.55 - 1.4)	(0.58 - 1.4)	(0.59 - 1.5)
Small employers and self employment	3.0		3.0	1.3	1.1	1.1
	(2.3 - 3.9)**		(2.3 - 3.9)**	(0.90 - 1.8)	(0.77 - 1.6)	(0.78 - 1.6)
Lower supervisors and technical	0.68		0.68	1.1	1.1	1.1
	(0.47 - 0.98)*		(0.47 - 0.97)*	(0.81 - 1.4)	(0.84 - 1.4)	(0.83 - 1.4)
Semi routine and routine	0.67		0.66	0.89	0.92	0.91
	(0.49 - 0.91)*		(0.48 - 0.89)**	(0.69 - 1.1)	(0.71 - 1.2)	(0.71 - 1.2)
NVQ Levels: reference category none						
NVQ Level 1	0.77		0.77	1.1	1.0	1.0
	(0.47 - 1.2)		(0.47 - 1.2)	(0.69 - 1.7)	(0.68 - 1.6)	(0.66 - 1.6)
NVQ Level 2	0.63		0.63	0.77	0.83	0.83
	(0.44 - 0.90)*		(0.44 - 0.91)*	(0.54 - 1.1)	(0.58 - 1.2)	(0.58 - 1.2)
NVQ Level 3	0.59		0.59	0.78	0.82	0.82
	(0.39 - 0.88)*		(0.39 - 0.88)*	(0.52 - 1.2)	(0.55 - 1.2)	(0.55 - 1.2)
NVQ Level 4	0.47		0.47	0.69	0.75	0.76
	(0.32 - 0.68)**		(0.32 - 0.68)**	(0.46 - 1.0)	(0.51 - 1.1)	(0.51 - 1.1)
NVQ Level 5	0.34		0.34	0.73	0.81	0.82
	(0.19 - 0.60)**		(0.19 - 0.59)**	(0.45 - 1.2)	(0.50 - 1.3)	(0.51 - 1.3)
Other/overseas quals only	0.57		0.57	0.89	0.96	0.92
	(0.32 - 1.0)		(0.32 - 1.0)	(0.50 - 1.6)	(0.53 - 1.7)	(0.51 - 1.7)

13

	Sweep One			Sweep Two		
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
Cohort child is the first born	1.1		1.1	0.91	0.88	0.88
	(0.94 - 1.4)		(0.93 - 1.4)	(0.77 - 1.1)	(0.74 - 1.0)	(0.74 - 1.0)
Ethnicity: reference white						
Mixed	1.1		1.1	2.3	2.4	2.5
	(0.44 - 2.6)		(0.44 - 2.6)	(1.0 - 5.3)*	(1.0 - 5.7)*	(1.1 - 5.8)*
Indian	1.9		1.8	2.5	2.3	2.3
	(1.1 - 3.4)*		(1.0 - 3.2)*	(1.5 - 4.1)**	(1.4 - 3.9)**	(1.4 - 3.9)**
Pakistani and Bangladeshi	2.2		2.2	2.4	2.2	2.2
	(1.3 - 3.8)**		(1.3 - 3.7)**	(1.6 - 3.7)**	(1.4 - 3.4)**	(1.5 - 3.4)**
Black or Black British	1.7		1.7	1.4	1.3	1.3
	(0.90 - 3.3)		(0.89 - 3.3)	(0.78 - 2.6)	(0.7 - 2.6)	(0.69 - 2.6)
Other ethnic group	2.0		2.0	1.0	0.97	0.99
	(1.1 - 3.5)*		(1.1 - 3.5)*	(0.53 – 2.0)	(0.49 - 1.9)	(0.49 – 2.0)
Owner occupier	0.99		0.98	0.76	0.76	0.77
	(0.76 - 1.3)		(0.75 - 1.3)	(0.63 - 0.91)**	(0.64 - 0.92)**	(0.65 - 0.93)**
Country: reference England						
Wales	0.81		0.78	1.2	1.3	1.3
	(0.54 - 1.2)		(0.52 - 1.2)	(0.96 - 1.6)	(0.98 - 1.7)	(0.98 - 1.7)
Scotland	1.5		1.5	0.84	0.79	0.79
	(0.71 - 3.1)		(0.72 - 3.1)	(0.60 - 1.2)	(0.57 - 1.1)	(0.58 - 1.1)
Northern Ireland	1.8		1.9	1.7	1.6	1.6
	(1.3 - 2.6)**		(1.3 - 2.6)**	(1.3 - 2.3)**	(1.1 - 2.1)**	(1.2 - 2.2)**
If partner respondent not responded to income questions in sweep 1					4.6	4.5
					(3.5 - 6.0)**	(3.4 – 6.0)**
Partner respondent the same respondent as sweep one						0.39
						(0.21 - 0.71)**
Observations	10754	10754	10754	7893	7893	7893

Note: 95% confidence intervals in parentheses * significant at 5%; ** significant at 1%

14

Table 6: Main and Partner Respondent Income Non-Response in Sweep One

	Main Respondent		Partner Respondent	
	Don't Know	**Refusal**	**Don't Know**	**Refusal**
	(1)	**(2)**	**(3)**	**(4)**
Respondent self employed	12	1.6	1.8	1.7
	(6.9 – 21)**	(0.50 – 5.5)	(1.2 – 2.6)**	(1.1 – 2.5)**
Respondent's age at interview	0.99	1.0	1.0	1.0
	(0.96 – 1.0)	(0.98 – 1.1)	(0.99 – 1.0)	(1.0 – 1.1)**
Main has a partner	1.4	0.61		
	(0.79 – 2.6)	(0.29 – 1.2)		
Social Class of Respondent: reference category Managerial and Professional				
Intermediate	1.9	1.3	1.2	0.66
	(1.1 – 3.3)*	(0.79 – 2.0)	(0.46 – 3.2)	(0.37 – 1.2)
Small employers and self employment	1.8	1.4	6.6	1.1
	(1.0 – 3.2)*	(0.37 – 5.6)	(4.5 – 9.8)**	(0.78 – 1.7)
Lower supervisors and technical	0.50	1.2	0.66	0.71
	(0.16 – 1.5)	(0.51 – 2.9)	(0.38 – 1.1)	(0.45 – 1.1)
Semi routine and routine	2.2	0.59	0.81	0.59
	(1.3 – 3.7)**	(0.30 – 1.2)	(0.47 – 1.4)	(0.41 – 0.85)**
NVQ Levels: reference category none				
NVQ Level 1	1.2	1.5	1.1	0.54
	(0.55 – 2.5)	(0.39 – 5.7)	(0.57 – 1.9)	(0.29 – 1.0)
NVQ Level 2	0.79	1.0	0.73	0.55
	(0.42 – 1.5)	(0.33 – 3.2)	(0.46 – 1.2)	(0.35 – 0.86)**
NVQ Level 3	0.87	0.51	1.0	0.29
	(0.43 – 1.7)	(0.15 – 1.7)	(0.61 – 1.7)	(0.16 – 0.53)**
NVQ Level 4	0.91	0.89	0.48	0.42
	(0.49 – 1.7)	(0.29 – 2.7)	(0.29 – 0.81)**	(0.26 – 0.68)**
NVQ Level 5	1.0	1.5	0.51	0.23
	(0.46 – 2.4)	(0.46 – 4.7)	(0.24 – 1.11)	(0.11 – 0.50)**
Other/overseas quals only	1.4	0.95	0.76	0.43
	(0.59 – 3.3)	(0.20 – 4.4)	(0.32 – 1.8)	(0.21 – 0.86)*

15

	Main Respondent		Partner Respondent	
	Don't Know	**Refusal**	**Don't Know**	**Refusal**
	(1)	**(2)**	**(3)**	**(4)**
Cohort child is the first born	0.94	1.2	1.1	1.2
	(0.67 – 1.3)	(0.87 – 1.8)	(0.87 – 1.4)	(0.92 – 1.5)
Ethnicity: reference white				
Mixed/Other	0.57	1.4	2.0	1.4
	(0.20 – 1.6)	(0.38 – 5.2)	(0.93 – 4.1)	(0.75 – 2.8)
Indian	1.7	1.0	1.4	2.4
	(0.72 – 3.8)	(0.42 – 2.6)	(0.65 – 3.0)	(1.2 – 4.7)*
Pakistani and Bangladeshi	1.5	1.8	1.7	2.8
	(0.47 – 5.0)	(0.26 – 12)	(0.76 – 3.7)	(1.6 – 5.1)**
Black or Black British	1.1	2.5	1.0	2.3
	(0.57 – 2.0)	(1.2 – 5.4)*	(0.45 – 2.4)	(1.1 – 4.7)*
Owner occupier	0.73	1.0	0.84	1.2
	(0.51 – 1.0)	(0.52 – 2.1)	(0.61 – 1.2)	(0.84 – 1.7)
Country: reference England				
Wales	0.84	0.78	0.76	0.85
	(0.49 – 1.5)	(0.44 – 1.4)	(0.38 – 1.5)	(0.54 – 1.3)
Scotland	1.2	1.8	1.5	1.5
	(0.61 – 2.2)	(0.77 – 4.3)	(0.70 – 3.4)	(0.64 – 3.3)
Northern Ireland	1.1	2.8	1.8	1.9
	(0.73 – 1.6)	(1.6 – 5.0)**	(1.1 – 2.8)*	(1.2 – 3.1)*
Observations	8190	8190	10754	10754

95% confidence intervals in parentheses * significant at 5%; ** significant at 1%

16

4. Modelling Attrition at MCS2 using Household Income and Income Response in MCS1 as Predictors

Here we consider the impact of household income and income non-response at sweep one on attrition at sweep two. These results are summarised in Table 8. The dependent variable here is 0 if a response is obtained and 1 if a non response is obtained. Those families where the cohort member died or the family has emigrated are not included in this model. Column one considers household income as measured at sweep one. Household income at sweep one does have an important impact on drop out at sweep two. Larger household income predicts less unit non response at sweep two.

Columns two and three consider unit non response at sweep two using item non response with regard to income at sweep one. For both the main and the partner respondent income non response at sweep one predicts unit non response at sweep two.

Table 8: Predicting Attrition at Sweep Two from Household Income at Sweep One.

	Unit Non-Response at Sweep Two		
	(1)	**(2)**	**(3)**
Household Income	0.66		
	(0.63 - 0.70)**		
Main Income Non-Response		1.7	
		(1.3 - 2.2)**	
Partner Income Non-Response			1.7
			(1.4 - 2.1)**
Observations	16790	8205	11464

Note: 95% confidence intervals in parentheses * significant at 5%; ** significant at 1%, all explanatory variables measured at sweep one

17

5. Conclusions

The paper has found that there do appear to be within household and within individual correlations for missing income data. Also, unlike other papers, we do not find interviewer effects. Female interviewers are no more successful than male interviewers in getting responses to income questions from main respondents and their partners. We must bear in mind, however, that this lack of association could be affected by the processes that lead to male and female interviewers being assigned to particular respondents.

We also find that there is a systematic tendency for income data to be missing at sweeps one and two over and above what we know about unit and partner non-response. For both the main and partner respondents income non-response is consistently related to self-employment. This suggests that how income questions are asked to the self- employed may need to be considered in more detail. At sweep one, both main and partner respondents in Northern Ireland are less likely to respond to the income questions. At sweep two, previous income non-response at sweep one and maintaining the same respondents are important predictors of non-response.

Finally, attrition at sweep two does appear to be related to both household income at sweep one, in that low income respondents are more likely to drop out at sweep two, as well as previous item non response to income questions at sweep one.

18

6. Implications for Analysis

Longitudinal data are collected in order to measure and to model change and so the results from this paper need to be considered in this light. We should also bear in mind that both unit and partner non-response in sweep one was associated with income in that poorer families and, in addition, poorer partners were more likely to be missed (Plewis, 2004). This process continued as we moved from sweep one to sweep two in that, as we have seen, poorer families were more likely to drop out (although we do not, of course, know whether the process of becoming either poorer or richer between sweeps was related to attrition). The under-representation of poorer families, although partially compensated for by the strategy of over-sampling more disadvantaged areas, could have implications for model estimates if, for example, the relation between an outcome of interest and income was non-linear. In a similar vein, the lack of information about the income of the self-employed might have an impact on model estimates. There are, of course, techniques for adjusting for non-response – weighting, multiple imputation and selection modelling for example - but discussion of these goes well beyond the purposes of this paper. What we have shown here is that members of households containing a young child do not always report their income and their reluctance or inability to do so is related to how they earn a living, who they live with (if anyone), where they live and what their ethnic background is. These are all factors to bear in mind when using income either as a response or, perhaps more commonly, as an explanatory variable in models of change.

19

References

Duncan, G. J. and N. A. Mathiowetz (1985) *A Validation Study of Economic Survey Data*, Ann Arbor, MI: Institute for Social Research, The University of Michigan

Essig, L. and J. Winter (2003) "Item Nonresponse to Financial Questions in Household Surveys: An Experimental Study of Interviewer and Mode effects" *Working Paper No*. 05-18. University of Mannheim: Germany

Jackle, A., E Sala, S. P. Jenkins and P. Lynn (2004) "Validation of Survey Data on Income and Employment: the ISMIE Experience" *Working Papers of the Institute for Social and Economic Research,* paper 2004-14. Colchester: University of Essex.

Lynn, P., A. Jäckle, S. Jenkins and E. Sala (2004) "The Effects of Dependent Interviewing on Responses to Questions on Income Sources" *Working Papers of the Institute for Social and Economic Research*, paper 2004-16. Colchester: University of Essex

Miller, H. R. (1953) "An Appraisal of the 1950 Census Income Data" *Journal of the American Statistical Association,* Vol. 48, No. 261 (March 1953), pp28-43

Nicoletti, C. and F. Peracchi (2001) *Sample Participation and Income Nonresponse in the ECHP,* Unpublished manuscript, Tor Vergeta University, Rome

Plewis, I. (ed.) (2004) *Millennium Cohort Study First Survey: Technical Report on Sampling (3rd Edition).* Centre for Longitudinal Studies: London

Plewis, I. and S. Ketende (eds.) (2006) *Millennium Cohort Study: Technical Report on Response (1st Edition).* Centre for Longitudinal Studies: London

Riphahn, R. T., O. Sefling (2002) "Item Non-Response on Income and Wealth Questions" *IZA Discussion Paper Series*, IZA DP No. 573, September 2002, Institute for the Study of Labor (IZA), Bonn: Germany

Rodgers, W. L., C. Brown, G. J. Duncan (1993) "Errors in Survey Reports of Earnings, Hours Worked, and Hourly Wages" *Journal of the American Statistical Association,* Vol. 88, No. 424 (Sec 1993) 1208-1218

Scräpler, J.-P. (2003) "Respondent Behaviour in Panel Studies – A Case Study for Income Non-Response by Means of the British Household Panel Study (BHPS)" *Working Papers of the Institute for Social and Economic Research,* paper 2003-08. Colchester: University of Essex

Siminski, P., P. Saunders and B. Bruce (2003) "Reviewing the Intertemporal Consistency of ABS Household Income Data through Comparisons with External Aggregates" *The Australian Economic Review,* vol. 36, no. 3, pp333-349

Weinberg, D. H., C. T. Nelson, M. I. Roemer, E. J. Welniak, Jr (1999) "Fifty Years of U.S. Income Data from the Current Population Survey: Alternatives, Trends and Quality" *The American Economic Review,* Vol. 89, No. 2, Papers and Proceedings of the One Hundred Eleventh Annual Meeting of the American Economic Association (May 1999), pp18-22

Centre for Longitudinal Studies
Institute of Education
20 Bedford Way
London WC1H 0AL
Tel: 020 7612 6860
Fax: 020 7612 6880
Email cls@ioe.ac.uk
Web http://www.cls.ioe.ac.uk

Centre for Longitudinal Studies

CLS Cohort Studies

Working Paper 2008/10

Missing Income Data in the Millennium Cohort Study: Evidence from the First Two Sweeps

Denise Hawkes
Ian Plewis

December 2008