Institute of Education

UCL

# Millennium Cohort Study

## Sixth Survey 2015-2016

## Technical report on response (Age 14)

By Tarek Mostafa and George Ploubidis

February 2017

CENTRE FOR LONGITUDINAL STUDIES

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

# Contents

# Introduction

The Millennium Cohort Study (MCS) is a multi-disciplinary, multi-purpose longitudinal survey following the lives of more than 19,000 children born in the UK in 2000-01. It is the most recent of the UK's world-renowned national longitudinal birth cohort studies. The study has been tracking the Millennium children through their childhood years and plans to follow them through adulthood.

As with any longitudinal survey, the MCS is subject to attrition. Attrition takes place when respondents drop out of the survey over time. This leads to two problems: (i) a reduction in sample size, and (ii) bias in sample composition. Sample bias arises when the likelihood of dropping out from the survey is correlated with the socio-demographic characteristics of the respondents. In this case, the survey will lose a particular type of respondents (e.g. disadvantaged families, ethnic minorities, etc) and the sample will no longer be representative of the population it was drawn from. However, there are statistical methods to deal with this, so as to ensure the remaining sample recovers (under reasonable assumptions) population parameters, which are the topic of this report.

This report examines attrition in sweep 6 (age 14) of MCS and presents the procedures used in the construction of MCS6 unit non-response weights. For a full description of attrition in previous sweeps, refer to the MCS Technical Report on Response (3rd edition, 2010) and Technical Report on Response in sweep 5 (2014). For a description of how to use the weights in Stata and SPSS, refer to the respective guides (Stata, SPSS). For a description of the MCS sample refer to the Technical Report on Sampling (4th edition, 2007).

# Response in MCS

In table 1, the proportion of productive and unproductive cases are presented by category. The table shows that the proportion of productive cases decreased over time from 96.4 per cent in MCS1 to 60.9 per cent in MCS6. The two categories of non-response which have seen a marked increase over time are 'Refusal' and 'Not issued'. 'Refusals' consist of respondents who refused to take part in a particular sweep of data collection, and 'Not issued' are respondents who have not participated in the survey on two consecutive occasions, and therefore were no longer issued for fieldwork (i.e. the survey agency no longer tries to contact them).

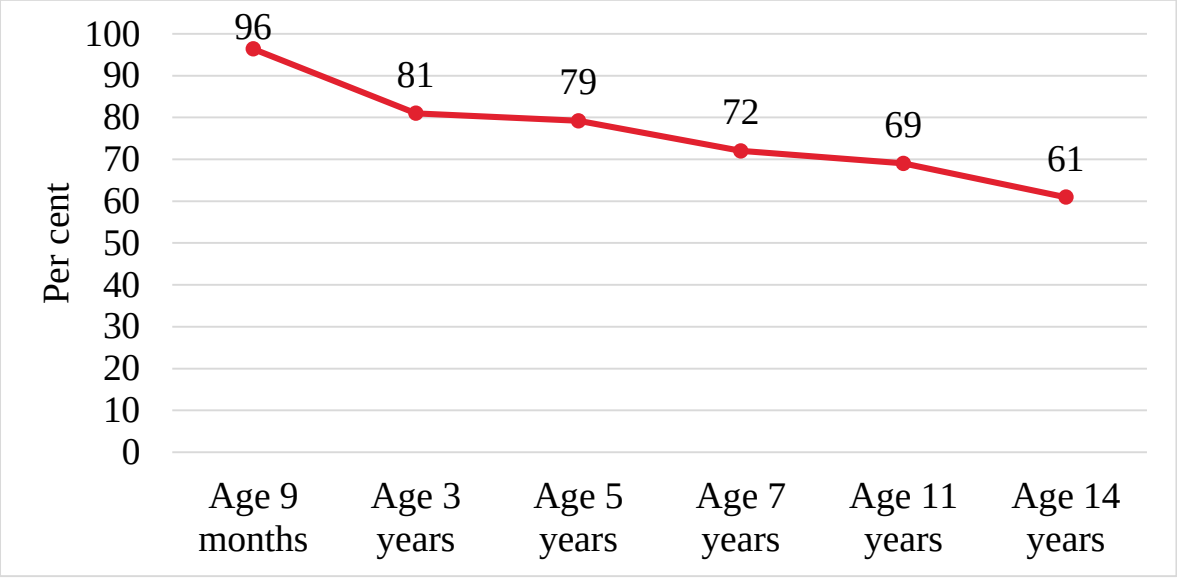
Non-contact has declined over time because respondents in this category have either been located and contacted again, or have moved to the not issued category. All other types of non-response are relatively stable over time. Note that 'Ineligible' includes child deaths, sensitive cases and temporary and permanent emigrants. The category 'Untraced movers' refers to respondents who have changed address and were not located, including possible emigrants. Respondents who were not issued in MCS1 are labelled as 'New Families'. These were eligible families who were not contacted in MCS1 because their addresses were not know in time for them to be included in the first wave of data collection.

**Table 1: Productive and unproductive cases in all MCS sweeps**

| | MCS1 Age 9 months | | MCS2 Age 3 years | | MCS3 Age 5 years | | MCS4 Age 7 years | | MCS5 Age 11 years | | MCS6 Age 14 years | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Freq. | % | Freq. | % | Freq. | % | Freq. | % | Freq. | % | Freq. | % |
| Productive | 18,551 | 96.4 | 15,590 | 81.0 | 15,246 | 79.2 | 13,857 | 72.0 | 13,287 | 69.0 | 11,726 | 60.9 |
| Refusal | | | 1,739 | 9.0 | 2,315 | 12.0 | 1,811 | 9.4 | 2,195 | 11.4 | 3,029 | 15.7 |
| Ineligible | | | 167 | 0.9 | 300 | 1.6 | 126 | 0.7 | 78 | 0.4 | 45 | 0.2 |
| Untraced movers | | | 686 | 3.6 | 546 | 2.8 | 706 | 3.7 | 388 | 2.0 | 428 | 2.2 |
| Non-contact | | | 930 | 4.8 | 546 | 2.8 | 123 | 0.6 | 438 | 2.3 | 75 | 0.4 |
| Not issued | 692 | 3.6 | | | | | 2,212 | 11.5 | 2,851 | 14.8 | 3,828 | 19.9 |
| Other unproductive | | | 131 | 0.7 | 290 | 1.5 | 408 | 2.1 | 6 | 0.0 | 112 | 0.6 |
| Total | 19,243 | 100 | 19,243 | 100 | 19,243 | 100 | 19,243 | 100 | 19,243 | 100 | 19,243 | 100 |

Figure 1 presents the proportion of productive cases in MCS in all sweeps. The figure shows that the sample decreased by 40% by the time of the age 14 survey.

**Figure 1: Proportion of cases productive in all MCS sweeps**




Note: The total number of MCS respondents ever interviewed is 19,243.

We now show how the proportion of productive cases at MCS6 vary along key dimensions. First, Table 2 shows how the MCS6 proportion of productive cases vary by country of sampling. The proportion productive is higher than the UK average in England while it is lower than the average in Scotland and Northern Ireland.

**Table 2: Productive and unproductive cases by country of sampling in MCS6**

| | England | | Wales | | Scotland | | Northern Ireland | |
|---|---|---|---|---|---|---|---|---|
| | Freq. | % | Freq. | % | Freq. | % | Freq. | % |
| Productive | 7,678 | 62.8 | 1,669 | 60.5 | 1,263 | 54.1 | 1,116 | 58.0 |
| Refusal | 1,853 | 15.2 | 455 | 16.5 | 401 | 17.2 | 320 | 16.6 |
| Ineligible | 35 | 0.3 | 2 | 0.1 | 6 | 0.3 | 2 | 0.1 |
| Untraced | 199 | 1.6 | 92 | 3.3 | 101 | 4.3 | 36 | 1.9 |
| Non-contact | 60 | 0.5 | 6 | 0.2 | 6 | 0.3 | 3 | 0.2 |
| Not issued | 2,335 | 19.1 | 503 | 18.2 | 548 | 23.5 | 442 | 23.0 |
| Other unproductive | 64 | 0.5 | 33 | 1.2 | 11 | 0.5 | 4 | 0.2 |
| Total | 12,224 | 100 | 2,760 | 100 | 2,336 | 100 | 1,923 | 100 |

Sample size=19,243

Table 3 shows that the proportion of productive cases vary across sampling strata in each country. Respondents sampled from the socially advantaged stratum are more likely to be productive in all four countries, compared to those sampled from the disadvantaged stratum. Respondents sampled from the ethnic minority stratum are less likely to be productive than those in the advantaged stratum in England.

**Table 3: Proportion of productive cases by stratum in MCS6**

| | England | | | Wales | | Scotland | | Northern Ireland | |
|---|---|---|---|---|---|---|---|---|---|
| | Adv. | Dis. | Ethn. | Adv. | Dis. | Adv. | Dis. | Adv. | Dis. |
| Productive | 67.1 | 59.8 | 60.3 | 65.1 | 58.5 | 60.1 | 48.3 | 63.4 | 54.8 |
| Unproductive | 32.9 | 40.2 | 39.7 | 34.9 | 41.6 | 39.9 | 51.7 | 36.7 | 45.2 |

Note: Adv stands for advantaged stratum. Dis stands for disadvantaged stratum and Ethn stands for ethnic minority stratum. Sample size=19,243.

In table 4 we look at different response patterns. Table 4 shows that 47.2 per cent of all respondents participated in all six sweeps of MCS. In contrast, 22.1 per cent have interrupted response patterns (i.e. non-monotone response). In other words, they participated in a number of sweeps, and then dropped out before participating again in subsequent sweeps. 30.7 per cent of all respondents have monotone response patterns. That is, they participated in a number of sweeps before dropping out for all subsequent sweeps.

**Table 4: Monotone vs. non-monotone response in MCS**

| Patterns | Freq. | Percent |
|---|---|---|
| Monotone | 5,908 | 30.7 |
| Non-monotone | 4,247 | 22.1 |
| All waves | 9,088 | 47.2 |
| Total | 19,243 | 100 |

Table 5 shows the percentages of respondents participating in n sweeps (n=1…6). We see that 64.1 per cent of respondents participated in at least five out of six sweeps of MCS indicating that more than half of the sample have almost complete records.

**Table 5: Number of times productive up to MCS6**

| Times productive | Freq. | Percent |
|---|---|---|
| One | 1,932 | 10.0 |
| Two | 1,391 | 7.2 |
| Three | 1,559 | 8.1 |
| Four | 2,027 | 10.5 |
| Five | 3,246 | 16.9 |
| Six | 9,088 | 47.2 |
| Total | 19,243 | 100 |

## Predicting response at MCS6

The procedure used for predicting response at sweep 5 was used also at sweep 6 [Technical Report on Response in sweep 5 (Mostafa 2014)].We estimate a logit model in which the dependent variable is binary (=1 for response and 0 otherwise) and the predictors are:

1. The cohort member's gender.
2. Mother's age at first live birth.
3. The cohort member's ethnic group.
4. Housing tenure in MCS5.
5. Accommodation type in MCS5.
6. The main respondent's highest educational qualification in all sweeps.
7. Whether the cohort member was breastfed.
8. Number of parents living in the household in MCS5.
9. The main respondent's highest social and economic status in all sweeps.
10. Ratio of number of times not answering the income question divided by the number of sweeps productive.
11. Ratio of number of times reporting having a job divided by the number of times productive.

12. Whether the household is a 'new' family. 701 children joined the survey in sweep 2 because their addresses were not known in sweep 1 and therefor did not take part in the first sweep. These children and their families were labelled as 'new families'.

Missing data for predictor variables due to non-monotone non-response or item missingness were imputed using simple and multiple imputations, as described below. Multiple imputations were carried out using the MI command in Stata 13. As a result of the use of simple and multiple imputations, the sample used in the logit response model consisted of 15,415 observations (i.e. the issued sample in MCS6). Weights were constructed for all respondents in MCS6. Imputations were carried out in the following way:

Replacement of missing values:

Since ethnicity is a fixed attribute over time and the main respondent's highest educational qualification is unlikely to change from one sweep to the other, we replaced the missing values on these two variables in MCS6 using the most recent available information from previous sweeps. Mother's age at first live birth was missing only for 60 cases. These were replaced by the average of non-missing cases.

## Multiple imputations:

Four variables were imputed using multiple imputation with chained equations in Stata 13. The imputation was carried out for item missingness and missingness caused by non-monotone response patterns. Accommodation type had the largest proportion missing with 22%. We generated 25 multiple imputations. The number of parents in the household and housing tenure were missing for 2,177 respondents in the analytical sample (i.e. issued sample in MCS6). Whether the cohort member was breastfed was missing for 612 respondents - mostly new families who joined the survey in sweep two - and accommodation type was missing for 3,435 respondents.

Breastfeeding and type of accommodation were imputed using the following variables as predictors of item specific missingess: highest educational qualification, whether family is a new family, cohort member's gender, cohort member's ethnic group, the main respondent's highest social and economic status, and sampling stratum.

The number of parents living in the household was imputed using highest educational qualification, whether family is a new family, cohort member's gender, cohort member's ethnic group, the main respondent's highest social and economic status, sampling stratum, in addition to the number of parents in the household in MCS1 and MCS4.

Housing tenure was imputed using highest educational qualification, whether family is a new family, cohort member's gender, cohort member's ethnic group, the main respondent's highest social and economic status, and sampling stratum, in addition to housing tenure in all previous sweeps.

We note that multiple imputation returns valid estimates assuming the data are Missing at Random (MAR) (Enders, 2010, Seaman et al., 2013, Sterne et al., 2009). This implies that any differences between the missing values and the observed values can be explained by the variables that were included in the imputation models. Put differently, conditional on the variables in the imputation model, missingness in not due to unobserved or observed variables not included in the model.

## No imputation required:

The ratio of number of times not answering the income question divided by the number of sweeps productive did not have any missing data since it was constructed using non-missing observations from all sweeps. The same applies for the ratio of number of times reporting having a job divided by the number of times productive. The main respondent's highest social and economic status was constructed as the maximum of social and economic status reported in each sweep.

Finally, some variables such as cohort member's gender and whether the household is a new family did not have any missing values and therefore did not require any imputation.

Table 6 shows the odds ratios of the response logit model estimated using the 25 imputed datasets. The linear predicted values were generated from this model; then the predicted values were converted into predicted probabilities using an inverse logit transformation. The non-response weights for sweep 6 were constructed as the inverse of the predicted probabilities [Wooldridge 2007]. Two overall weights were constructed by multiplying the sampling weights in sweep 1 by the attrition weights in each sweep of MCS (i.e. 2 to 6). The weights were scaled to make their total equal to the productive sample size.

FOVWT1: sweep 6 overall weight for single country analysis

FOVWT2: sweep 6 overall weight for whole of UK analysis.

**Table 6: Logit response model**

| Explanatory variables | Odds Ratio | SE | t | P>t | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| New family | 1.03 | 0.11 | 0.28 | 0.78 | 0.84 | 1.26 |
| Cohort member is a boy | 0.87 | 0.03 | -3.49 | 0.00 | 0.81 | 0.94 |
| *Cohort member's ethnic group Reference: White)* | | | | | | |
| Mixed | 0.85 | 0.10 | -1.42 | 0.15 | 0.68 | 1.06 |
| Indian | 1.35 | 0.19 | 2.17 | 0.03 | 1.03 | 1.78 |
| Pakistani, Bangladeshi | 2.01 | 0.19 | 7.31 | 0.00 | 1.67 | 2.43 |
| Black | 0.79 | 0.09 | -2.19 | 0.03 | 0.64 | 0.97 |
| Other, NA, not known, refusal | 1.22 | 0.17 | 1.43 | 0.15 | 0.93 | 1.61 |
| *Main respondent's highest social and economic status (Reference: Managerial and Professional)* | | | | | | |
| I ntermediate | 0.83 | 0.05 | -2.87 | 0.00 | 0.74 | 0.94 |
| Small employers and self employed | 0.77 | 0.07 | -2.65 | 0.01 | 0.64 | 0.94 |
| Lower supervisory and technical | 0.66 | 0.06 | -4.47 | 0.00 | 0.54 | 0.79 |
| Semi-routine and routine | 0.52 | 0.03 | -9.81 | 0.00 | 0.46 | 0.59 |
| NA | 0.45 | 0.05 | -7.96 | 0.00 | 0.37 | 0.55 |
| *Highest educational qualification (Reference: NVQ level 1)* | | | | | | |
| NVQ level 2 | 1.00 | 0.08 | -0.05 | 0.96 | 0.85 | 1.16 |
| NVQ level 3 | 1.11 | 0.10 | 1.21 | 0.23 | 0.94 | 1.32 |

| | | | | | | |
|---|---|---|---|---|---|---|
| NVQ level 4 | 1.35 | 0.12 | 3.46 | 0.00 | 1.14 | 1.60 |
| NVQ level 5 | 1.84 | 0.22 | 5.16 | 0.00 | 1.46 | 2.33 |
| Overseas qualifications only | 1.22 | 0.16 | 1.47 | 0.14 | 0.94 | 1.58 |
| None of these | 0.80 | 0.07 | -2.60 | 0.01 | 0.67 | 0.95 |
| *Number of parents in household (reference: One parent)* | | | | | | |
| Two parents/carers | 1.20 | 0.06 | 3.70 | 0.00 | 1.09 | 1.32 |
| Child was breastfed at least once | 1.23 | 0.06 | 4.49 | 0.00 | 1.12 | 1.35 |
| *Accommodation type (Reference: Other)* | | | | | | |
| House or bungalow | 1.14 | 0.11 | 1.39 | 0.17 | 0.95 | 1.37 |
| *Housing tenure (Reference: Own outright)* | | | | | | |
| Own - mortgage/loan | 1.09 | 0.10 | 0.94 | 0.35 | 0.91 | 1.32 |
| Rent from local authority | 0.73 | 0.08 | -2.96 | 0.00 | 0.59 | 0.90 |
| Rent from housing association | 0.72 | 0.08 | -2.87 | 0.00 | 0.57 | 0.90 |
| Rent privately | 0.74 | 0.08 | -2.79 | 0.01 | 0.60 | 0.91 |
| Other (rent free, living with parents) | 0.67 | 0.09 | -3.02 | 0.00 | 0.51 | 0.87 |
| Ratio income item non-response | 0.26 | 0.03 | -11.43 | 0.00 | 0.20 | 0.32 |
| Ratio times having a job | 0.33 | 0.03 | -14.08 | 0.00 | 0.29 | 0.39 |
| Mother's age at first birth | 1.05 | 0.00 | 10.68 | 0.00 | 1.04 | 1.06 |
| Constant | 1.81 | 0.36 | 2.96 | 0.00 | 1.22 | 2.69 |
| N | | | 15,415 | | | |

Note: The analytical sample in table 6 includes all issued cases in MCS6 with 15,415 observations.

In tables 7 and 8, the means, minima and maxima of the two weights are presented by stratum.

**Table 7: FOVWT1, sweep 6 overall weight for single country analysis**

| Sampling stratum | N | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|---|
| England - Advantaged | 3,240 | 1.24 | 0.66 | 0.68 | 8.96 |
| England - Disadvantaged | 2,876 | 1.00 | 0.68 | 0.38 | 10.90 |
| England - Ethnic | 1,562 | 0.42 | 0.30 | 0.14 | 4.28 |
| Wales - Advantaged | 542 | 1.54 | 0.87 | 0.79 | 10.80 |
| Wales - Disadvantaged | 1,127 | 0.79 | 0.49 | 0.32 | 6.93 |
| Scotland - Advantaged | 688 | 1.06 | 0.79 | 0.32 | 7.95 |
| Scotland - Disadvantaged | 575 | 0.95 | 0.81 | 0.20 | 9.23 |
| Northern Ireland - Advantaged | 458 | 1.25 | 0.86 | 0.34 | 6.07 |
| Northern Ireland - Disadvantaged | 658 | 1.03 | 0.88 | 0.20 | 8.45 |
| All strata | 11,726 | 1.01 | 0.73 | 0.14 | 10.90 |

**Table 8: FOVWT2, sweep 6 overall weight for whole of the UK analysis**

| Sampling stratum | N | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|---|
| England - Advantaged | 3,240 | 1.60 | 0.88 | 0.86 | 12.47 |
| England - Disadvantaged | 2,876 | 1.31 | 0.91 | 0.49 | 14.03 |
| England - Ethnic | 1,562 | 0.56 | 0.40 | 0.18 | 5.53 |
| Wales - Advantaged | 542 | 0.52 | 0.29 | 0.27 | 3.48 |
| Wales - Disadvantaged | 1,127 | 0.27 | 0.16 | 0.11 | 2.37 |
| Scotland - Advantaged | 688 | 0.82 | 0.59 | 0.25 | 6.26 |
| Scotland - Disadvantaged | 575 | 0.74 | 0.63 | 0.16 | 7.30 |
| Northern Ireland - Advantaged | 458 | 0.49 | 0.32 | 0.14 | 2.43 |
| Northern Ireland - Disadvantaged | 658 | 0.40 | 0.33 | 0.08 | 3.33 |
| All strata | 11,726 | 1.01 | 0.86 | 0.08 | 14.03 |

For a description of how to use the weights in Stata and SPSS refer to the respective guide: Stata, SPSS.

We note that the effectiveness of the response weights to correct for bias depends on the inclusion of all important predictors of unit non response in the logit response model (Table 6) (Seaman and White, 2013).

## Links to supporting documents

MCS Technical Report on Sampling (4th edition, 2007)

User Guide to Analysing MCS Data Using Stata (1st edition, 2011)

User Guide to Analysing MCS Data Using SPSS (1st edition, 2010)

MCS Technical Report on Response (3rd edition, 2010)

Technical Report on Response in sweep 5 (2014).

## References

ENDERS, C. K. 2010. Applied missing data analysis, Guilford Press.

SEAMAN, S., GALATI, J., JACKSON, D. & CARLIN, J. 2013. What is meant by “missing at random”? Statistical Science, 28, 257-268.

SEAMAN, S. R. & WHITE, I. R. 2013. Review of inverse probability weighting for dealing with missing data. Stat Methods Med Res, 22, 278-95.

STERNE, J. A. C., WHITE, I. R., CARLIN, J. B., SPRATT, M., ROYSTON, P., KENWARD, M. G., WOOD, A. M. & CARPENTER, J. R. 2009. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ, 338.

WOOLDRIDGE, J. M. 2007. Inverse probability weighted estimation for general missing data problems. Journal of Econometrics, 141, 1281-1301.