The impact of attrition and non-response in birth cohort studies: a need to incorporate missingness strategies

Tarek Mostafa   Institute of Education, University College London, UK
T.Mostafa@ioe.ac.uk
Richard D. Wiggins   Institute of Education, University College London, UK

Abstract

This paper reveals the extent of attrition in the British Cohort Study begun in 1970 (BCS70) and how it affects sample composition over time. We examine the determinants of response and then construct inverse probability weights (IPWs) to adjust for sample loss. Secondly, we create a hypothetical substantive data set from BCS70 across data collection waves 3 and 4 to illustrate the effectiveness of the use of weights and multiple imputations (MI) in handling the impact of unit non-response and item missingness respectively. Our findings show that when the predictive power of the response models is weak, the efficacy of non-response weights is undermined. Further, multiple imputations are effective in reducing the bias resulting from item missingness when the magnitude of the bias is high and the imputation models are well specified.

Introduction

This paper has two objectives. Firstly, to provide a historical account of the extent of unit non-response in the British Cohort Study begun in 1970 (BCS70) across the nine waves of data collection between 1970 and 2012 together with an illustration of the construction of inverse probability weights (IPWs) to adjust for unit non-response. Secondly, we illustrate the impact of using IPWs and the application of multiple imputation (MI) for an artificially constructed set of patterns of missingness under a substantive question for analysis. The construction of IPWs under the first objective are based upon a set of birth characteristics for cohort members (CMs) because these values are available for almost the entire sample of CMs originally recruited into the study. This second objective is to assess the impact of non-response weights and imputation techniques on the bias resulting from unit non-response and item missingness respectively by conditioning on a fully observed subset of data from wave 4. The next section presents a brief overview of the literature

on approaches to handle missing data and more background to response patterns in BCS70. We follow this section with a description of wave-by-wave non-response in BCS70 and examine the extent to which the representativeness of the study may be revealed over time in terms of the cumulative loss of CMs according to their birth characteristics. There then follows a section to illustrate the impact of the use of IPWs and MI. The final sections provide a brief discussion and conclusion. Our underlying motivation for this paper is to help raise users' appreciation of the need to incorporate strategies to handle missingness in any longitudinal analysis of BCS70 and, indeed birth cohort studies more generally.

Background

Statistical description and analysis are persistently challenged by the problem of missing data (Little & Rubin 2002). Survey samples are threatened by both unit non-response and individual item missingness where a respondent fails to provide all of the information requested

(also known as partial non-response). In longitudinal surveys, the problem of maintaining co-operation with CMs over time adds another dimension to the problem of non-response. Attrition refers to situations where CMs drop out of a study and never return, and situations where individual CMs have an interrupted response pattern over time. For instance, a respondent may not co-operate during a particular wave of a study and then return subsequently creating an instance of 'wave non-response'. These patterns are distinguished as monotone and non-monotone response, respectively.

Missing data constitutes a problem for two reasons. First, missingness leads to the loss of observations and to the reduction of sample size. For instance, in BCS70 if only CMs who have responded in all nine waves (since 1970) are considered available for any substantive analysis, the resulting sample would represent only 20 per cent of the original sample of 17,284 CMs. Secondly, missingness may lead to selection bias and inaccurate inference. In order to tackle these problems and make best use of all of the available data in any analysis it becomes necessary to make assumptions about how a proportion of our data came to be missing at all. Many authors (for example Carpenter & Plewis, 2011) term this process to be 'the missingness mechanism'. Following Rubin (1976), Little and Rubin (2002), Carpenter and Plewis (2011) and others we adopt a typology of missingness mechanisms described as 'missing completely at random' (MCAR), 'missing at random' (MAR) and 'missing not at random' (MNAR). Briefly, MCAR implies that the probability of not answering a particular question is uncorrelated with the characteristics of the respondent, and in any longitudinal survey it means that the probability of dropping-out from any wave is uncorrelated with the characteristics of the CM. MCAR is a very strong assumption to make since missingness is more likely to be at random (MAR) or not at random (MNAR). Under MAR the probability of non-response to a question or the probability of dropping-out from a particular sweep are related to some of the observable characteristics of the respondent such as gender, social class, or educational level. Under MNAR, the probability of item non-response or the probability of dropping-out from a particular wave is related to characteristics or traits yet to be observed (or

simply described as unobserved variables). If the pattern or nature of missingness is related to any observable or unobservable variables then ignoring it would lead to the loss of a particular type of respondents (e.g. men, the less well educated) and hence the sample will no longer be random or representative of the parent population.

Historically, two broad approaches have been typically adopted to tackle the problem of missing data involving the application of weighting adjustments (IPWs) to compensate for unit non-response and/or multiple imputation (MI) to handle item non-response (Kalton (1986) and Lepkowski (1989)). Both adjustment strategies assume MAR. Under MI, there are several different approaches appropriate to the type of data and/or data structure (Carpenter & Plewis, 2011). In this paper we demonstrate that analyses which draw upon the longitudinal history of the cohort require strategies to use all of the available data, including any partial information (incomplete wave responses).

The application of weights or IPWs in longitudinal surveys is typically used to adjust or re-balance the distributions of the responders so that the relative importance of each CM's characteristic in any particular wave is reweighted according to the importance of the characteristics of those who dropped out. In other words, if the survey is losing men over time, then men will be given a relatively higher weight than women (see Hawkes & Plewis (2006), Plewis (2007a) for illustrations using the Millennium Cohort Study (MCS), Plewis (2007b) and Seaman and Wright (2011) for a review). The probability of response at each sweep is estimated using logistic regression models for a binary outcome (response or non-response) or multinomial-logistic regression models (allowing for more than two outcome categories). These models draw upon CM characteristics as covariates to improve the goodness of fit of the underlying model. Additionally, analysts may use external metadata as auxiliary covariates to enhance efficiency. For useful illustrations, see Plewis (2011), Schouten and de Nooij (2005) and Micklewright, Skinner and Schnepf (2012).

In survey research, there is a long tradition of applying weights to adjust for differential probabilities of selection (the sample design) and/or response. It is fairly straightforward to construct weights to adjust for differential non-response, and relatively easy to apply and make them routinely

available as part of an archived dataset. For example the British Household Panel Survey (BHPS) (Taylor, Brice, Buck & Prentice-Lane, 2010) and MCS both provide longitudinal weights to adjust for cumulative drop-out over time as well as wave-specific or cross-sectional weights. However, they have a number of disadvantages:

1.   Weights (both longitudinal and cross-sectional or wave specific) adjust for unit non-response at the level of the individual CM. Implicitly the same weights adjust for all item response appertaining to a lost individual. There may be more effective solutions available to handle item missingness, namely MI as described below.

2.   If variables x, y and z are used in predicting unit non-response, and thus in the construction of weights, the results of analyses using x, y and z as dependent and independent variables will yield unbiased results. However, if there is a fourth omitted variable, w, that happens to be strongly related to x, y and z, then the inclusion of w could well improve the efficiency of the construction of weights.

3.   Under conditional regression applications, if we are regressing an outcome variable from sweep t+1 on a number of independent variables collected during an earlier sweep t where attrition has possibly occurred, the weighted analysis will be constrained to using only the non-missing cases in both waves (Goldstein, 2009). This further undermines the efficacy of non-response weights because they will only adjust for non-response in one wave (usually the wave in which the dependent variable was observed). As an alternative, there is the possibility of combining IPW and MI as a remedy (Wiggins, Schofield, Bartley, Sacker & Head, (2004) McDonald & Ketende (2009) and Seaman, White, Copas & Li (2012)).

Under MI, (Little & Rubin (2002), Schafer & Olsen (1998) and Rubin (1987, 2004)) missing values under MAR assumptions are replaced several times to create filled-in replicates of our data. These replicates of multiply-imputed data are analysed separately and ultimately combined under Rubin's Rules (Rubin, 1987) to provide parameter estimates which take account of the uncertainty introduced into any analysis by filling-in under the MAR assumption. There are a number of approaches to MI which vary according to assumptions made about the type of data to be imputed (e.g.) multivariate Normal together with the application

of transformation before and after imputation, for categorical data and/or where data is missing at varying levels of aggregation or clustering, (i.e. for multilevel or hierarchical data), and various software packages offer these solutions. The interested reader is encouraged to consult Carpenter and Kenward (2013) for a valuable overview of MI and its application.

MI approaches appear more complex to use than weights but their application presents two main advantages:

1.   MI can handle both the treatment of item and unit non-response. Indeed wave non-response can be considered a special case of item missingness where all variables are missing for the same respondent within a longitudinal record.

2.   MIs can be custom-made according to the needs of the researcher. When properly specified, they are robust and generate valid inference. MI can be implemented according to the structure of the data (e.g. for handling a multilevel structure see Goldstein, Carpenter & Browne (2014) and the type of variables (e.g. continuous, ordinal or multinomial variables) see Nathan (1983), Nathan & Holt (1980), Pfeffermann (2001) and Carpenter & Plewis (2011)).

According to Carpenter and Kenward (2013) in their chapter entitled 'Sensitivity analysis: MI unleashed', in order to explore how robust inferences are to the assumption of MAR, analysts have an obligation to impute data under MNAR or at least 'approximate the results of doing so'. One approach that may appeal to the user is known as 'joint modelling' where the substantive model of interest is modelled jointly with a model for missingness (also referred to as 'Heckman modelling', Heckman (1979)). In this way, it is proposed that the unobserved variables that simultaneously influence both the outcome and the missingness are captured by the residuals in the two models, which are allowed to correlate. Technically, the challenge for the analyst is to identify variables (or instruments) for the missingness model which predict the probability of missingness but do not correlate with the substantive outcome (see Carpenter & Plewis (2011) for an illustration using NCDS data).

Clearly, in this brief overview of various approaches to handling missingness we have placed the use of IPW and MI in the foreground of our coverage simply because good illustrations are available in the literature for users to consult as

well as software to match (e.g. STATA). For more recent methodologies which specifically draw upon the longitudinal nature of information databases, users are referred to the application of MI under a 'two-fold fully conditional specification (FCS)' which fills-in data at time t conditional upon data in time t and adjacent time points t-1 and t+1 (Nevalainen, Kenward & Virtanen (2009) and Welch, Bartlett & Petersen (2014)). Additionally, users may prefer to find solutions to maximising their use of available data in the context of the application of a particular analysis, such as structural equation modelling where the emphasis is upon the temporal structure of relationships. Here 'full information maximum likelihood (FIML)' has been developed under well-known SEM algorithms (e.g. AMOS, Arbuckle (1996)) where item missingness is not handled directly but the likelihood function is adjusted so that incomplete data is used in the estimation under MCAR or MAR (Davey & Savla, 2010). The application of FIML also incorporates the use of auxiliary variables (Enders, 2008).

Non-response in the British Cohort Study (BCS70)

All nine available waves of BCS70 are used to examine attrition and model response in terms of several birth characteristics. These covariates include: gender, father's social class, father's and mother's age at completion of education, mother's age at delivery, whether mother lived in London in 1970, whether or not the CM's mother attempted breast-feeding, her marital status and the number of older siblings at the time of the CM's birth.

In table 1 we summarise the pattern of missing data for BCS70 over the nine waves of data collection from 1970 to 2012. Just under 1 in 5 (19.8 per cent, labelled as non-missing) of the CMs participated in all nine waves, whereas over half (52 per cent, labelled as non-monotone) dropped out from at least one wave but returned to the study in a subsequent wave, and nearly a third (27.2 per cent, labelled as monotone) dropped out from the survey after participating in a number of waves without ever returning, to date. The base sample of 17,284 CMs consists of the original birth sample (i.e. excluding immigrants who joined the study later on).

Table 1. Patterns of missing data in BCS70 (1970 to 2012)

Pattern	Frequency	Percentage
Monotone	4,716	27.2
Non monotone	9,153	53.0
Non missing	3,423	19.8
Total	17,284	100

Table 2 below shows that over 42 years, from birth in 1970 up to and including the ninth wave in 2012, 7,930 CMs dropped out of the study for various reasons. Some have died, others have left Great Britain, while some have refused to participate or disappeared from the study for one or more waves, only to reappear again. The category labelled 'dead'

describes the total number of deaths over the nine sweeps, while the category 'unproductive' describes all other possibilities for dropout: e.g. permanent and temporary immigrants, refusals, non-contact. One should note that dropout is not always permanent since some respondents return in later waves.

Table 2. Detailed response and non-response categories for BCS70 from 1970 to 2012

Response categories	Wave 1	Wave 2	Wave 3	Wave 4	Wave 5	Wave 6	Wave 7	Wave 8	Wave 9
Age	Birth	**5**	10	16	26	30	34	38	42
Full or partial response	16,569	12,939	14,349	11,206	8,654	10,833	9,316	8,545	9,354
Dead	0	565	585	597	697	748	795	824	853
Unproductive	715	3,780	2,350	5,481	7,933	5,703	7,173	7,915	7,077
Total	17,284	17,284	17,284	17,284	17,284	17,284	17,284	17,284	17,284

Note. The mode of data collection changed between wave four and wave five. In wave five, the data was collected through postal services.

Figure 1 shows that there was a substantial drop in the achieved sample size between age 16 and age 26 years. There are several possible reasons for this drop including the length of the period of 10 years separating the two waves, a teacher's strike at age 16, the use of a self-completion postal survey at age 26, and the fact that the responsibility for CMs to provide consent to participate shifted from their parents to themselves as young adults. Furthermore, the drop in the achieved sample size at age 38 years can also be partially attributed to the use of a telephone survey as a mode of data collection rather than face-to-face interviews. Interestingly, the achieved sample size increased by 4.7 per cent for wave 9 (CM age 42) as some CMs were successfully traced and followed-up. For further information on the data collection, please see the BCS70 Technical Report (Centre for Longitudinal Studies, 2014).

Figure 1. Sample size in the different waves of BCS70
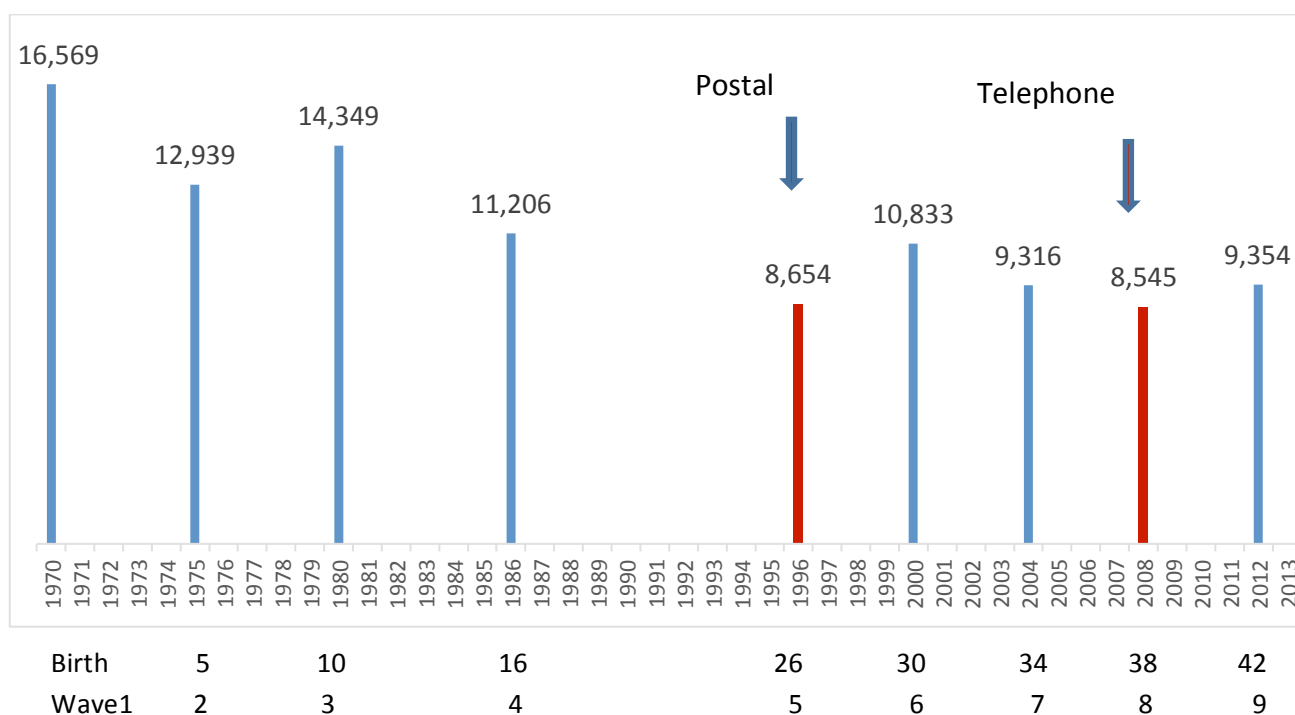
Figures 2 and 3 show the evolution of the sample composition according to CM characteristics recorded at birth. We find that the proportions of male CMs, CMs with fathers who have manual occupations, and CMs with mothers with low levels of education remaining in the study are steadily falling over time. Likewise, the proportions of CMs whose parents were single in 1970, whose mothers were living in London in 1970, and those who have at least three older siblings, have also fallen. It is also worth noting that the rise in sample size in wave 9 (age 42), and the switch from use of the telephone in wave 8 to face-to-face data collection in wave 9, resulted in a change in the point estimates of these proportions. Indeed, they have slightly converged towards their original values at birth. This suggests that the non-response bias in wave 9 will be lower than that in wave 8. Moreover, the differential in the ratio of males to females attained was at a maximum by age 26 (wave 5) and has declined ever since.

Figures 2 and 3. Change in the BCS70 sample composition over time (1970 to 2012). Change is computed in percentages
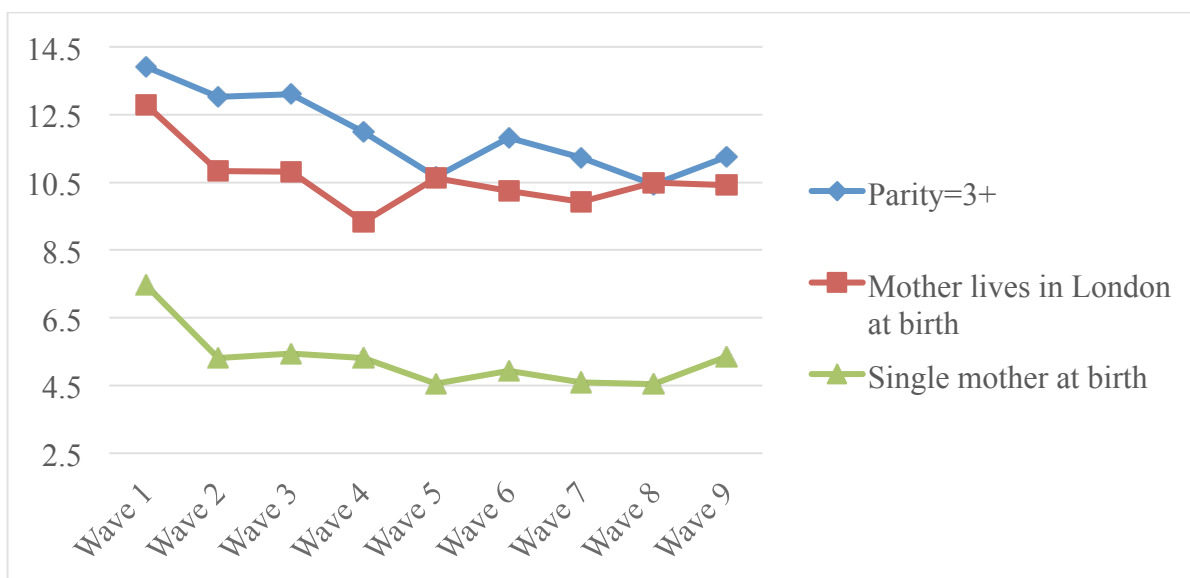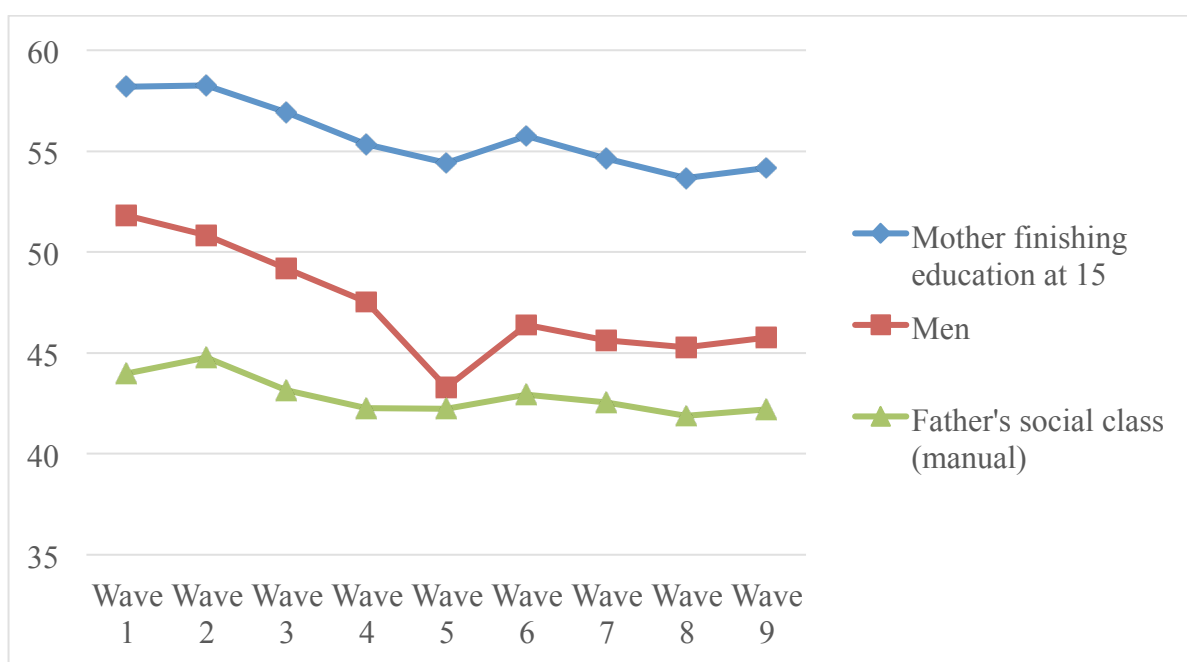
In general, we can say that men from lower social backgrounds whose parents were single in 1970 are more likely to drop out from the survey. The drop out within these groups could have also been exacerbated by the lack of cohort maintenance. Based on these findings, it is clear that any dropout or lack of co-operation is not a random phenomenon. Differential subject loss according to key birth characteristics will weaken the representativeness of the study, and unless the analysis adjusts for the characteristics which impact upon the probability to respond, inferences will be unreliable. The impact of subject loss may be compounded when we consider the combined effect of these characteristics. We now consider this possibility by using logistic regression analyses for each wave at a time.

What follows in table 3 are the results of a logistic regression of obtaining a response or otherwise for each wave using the birth characteristics as a set of covariates. Response as an outcome is a binary variable taking the value of 1 for those who participated (first category in table 2) and 0 for all other categories including those who died or migrated. Note that sample size is relatively smaller (i.e. 15,270 instead of 17,284) than in table 2 because some CMs had missing birth characteristics.

Table 3. Odds ratios based on logistic regressions of binary response outcome for successive BCS70 waves

	Wave 2	Wave 3	Wave 4	Wave 5	Wave 6	Wave 7	Wave 8	Wave 9
Age	5	10	16	26	30	34	38	42
Gender (reference: men)								
Women	1.00	1.08	1.26[***]	1.80[***]	1.49[***]	1.48[***]	1.48[***]	1.44[***]
	(0.040)	(0.049)	(0.044)	(0.060)	(0.052)	(0.049)	(0.049)	(0.048)
Marital status (reference: single)								
Married	1.47[***]	2.18[***]	1.67[***]	1.85[***]	1.89[***]	1.89[***]	1.79[***]	1.42[***]
	(0.140)	(0.218)	(0.151)	(0.174)	(0.171)	(0.174)	(0.169)	(0.128)
Mother lives in London in 1970 (reference: not in London)								
In London	0.57[***]	0.55[***]	0.47[***]	0.71[***]	0.61[***]	0.62[***]	0.70[***]	0.67[***]
	(0.032)	(0.034)	(0.024)	(0.037)	(0.031)	(0.032)	(0.036)	(0.034)
Parity (reference: 0)								
1	0.97	1.02	0.87[**]	0.92	0.94	0.89[**]	0.93	0.92[*]
	(0.050)	(0.059)	(0.039)	(0.039)	(0.042)	(0.038)	(0.039)	(0.039)
2	0.82[**]	0.89	0.81[***]	0.79[***]	0.84[**]	0.74[***]	0.75[***]	0.81[***]
	(0.053)	(0.065)	(0.046)	(0.042)	(0.047)	(0.040)	(0.040)	(0.044)
3+	0.72[***]	0.90	0.70[***]	0.58[***]	0.65[***]	0.58[***]	0.54[***]	0.61[***]
	(0.053)	(0.076)	(0.045)	(0.036)	(0.041)	(0.036)	(0.033)	(0.038)
Breastfeeding (reference: attempted)								
Not attempted	0.82[***]	0.84[***]	0.85[***]	0.85[***]	0.92[*]	0.87[***]	0.87[***]	0.80[***]
	(0.036)	(0.041)	(0.032)	(0.031)	(0.034)	(0.031)	(0.031)	(0.029)
Mother's age at Delivery (reference: less than 20)								
[20-24]	1.42[***]	1.17	1.20[**]	1.31[***]	1.23[**]	1.33[***]	1.28[***]	1.26[***]
	(0.105)	(0.098)	(0.080)	(0.085)	(0.080)	(0.085)	(0.083)	(0.081)
[25-29]	1.51[***]	1.27[**]	1.28[***]	1.46[***]	1.35[***]	1.50[***]	1.45[***]	1.35[***]
	(0.121)	(0.115)	(0.092)	(0.102)	(0.096)	(0.103)	(0.101)	(0.093)
[30-34]	1.63[***]	1.36[**]	1.30[**]	1.62[***]	1.44[***]	1.66[***]	1.59[***]	1.39[***]
	(0.151)	(0.143)	(0.106)	(0.129)	(0.117)	(0.131)	(0.125)	(0.109)
35 or more	1.81[***]	1.56[***]	1.40[***]	1.69[***]	1.51[***]	1.81[***]	1.73[***]	1.45[***]
	(0.204)	(0.198)	(0.140)	(0.164)	(0.149)	(0.175)	(0.167)	(0.139)

(Table 3 cont'd)

Mother's age at completion of education (reference: 14 or less)

15	1.56***	1.81***	1.29**	1.38***	1.20*	1.32***	1.15	1.04
	(0.141)	(0.179)	(0.106)	(0.114)	(0.098)	(0.107)	(0.094)	(0.084)
16	1.63***	1.73***	1.50***	1.50***	1.37***	1.51***	1.34***	1.21*
	(0.164)	(0.190)	(0.137)	(0.135)	(0.124)	(0.134)	(0.119)	(0.107)
17	1.47***	1.42**	1.32**	1.56***	1.26*	1.45***	1.32**	1.18
	(0.172)	(0.180)	(0.138)	(0.160)	(0.131)	(0.148)	(0.134)	(0.120)
18 or more	1.31*	1.34*	1.30*	1.48***	1.14	1.33**	1.24*	1.05
	(0.147)	(0.164)	(0.133)	(0.149)	(0.116)	(0.134)	(0.124)	(0.105)
Father's social class (reference: SC 1)								
Professional	0.94	0.98	0.85	0.94	0.93	0.99	0.95	0.97
	(0.102)	(0.116)	(0.084)	(0.087)	(0.090)	(0.092)	(0.088)	(0.090)
Clerical, non-manual	1.06	1.20	1.04	1.07	1.10	1.13	0.99	1.00
	(0.122)	(0.151)	(0.107)	(0.102)	(0.111)	(0.108)	(0.094)	(0.095)
Skilled manual	0.90	0.94	0.79*	0.79*	0.83*	0.84	0.74***	0.77**
	(0.097)	(0.109)	(0.076)	(0.071)	(0.078)	(0.076)	(0.067)	(0.070)
Unskilled manual	0.87	0.85	0.75**	0.70***	0.76**	0.75**	0.68***	0.69***
	(0.101)	(0.108)	(0.079)	(0.068)	(0.077)	(0.073)	(0.066)	(0.068)
Lowest grade workers	0.70**	0.77	0.69**	0.56***	0.64***	0.65***	0.56***	0.59***
	(0.091)	(0.111)	(0.081)	(0.063)	(0.074)	(0.072)	(0.063)	(0.065)
Other	0.34***	0.60***	0.70**	0.70**	0.69**	0.76*	0.65***	0.70**
	(0.044)	(0.085)	(0.086)	(0.082)	(0.083)	(0.088)	(0.075)	(0.081)
Father's age at completion of education (reference: 14 or less)								
15	1.20*	1.24*	1.11	1.02	1.19*	1.03	1.11	1.03
	(0.102)	(0.119)	(0.083)	(0.076)	(0.089)	(0.076)	(0.082)	(0.076)
16	1.09	1.00	1.14	1.07	1.13	1.00	1.10	0.99
	(0.107)	(0.108)	(0.098)	(0.090)	(0.096)	(0.084)	(0.092)	(0.082)
17	0.92	1.04	1.25*	1.21	1.27*	1.10	1.29*	1.08
	(0.107)	(0.136)	(0.131)	(0.122)	(0.132)	(0.111)	(0.130)	(0.108)
18 or more	0.79*	0.82	0.98	0.96	1.05	1.00	1.06	0.92
	(0.083)	(0.094)	(0.092)	(0.088)	(0.097)	(0.091)	(0.097)	(0.083)
N	15270	15270	15270	15270	15270	15270	15270	15270
pseudo R^2	0.036	0.034	0.026	0.040	0.028	0.031	0.033	0.025

*Notes. Exponentiated coefficients; Standard errors in parentheses, * p < 0.05, ** p < 0.01, *** p < 0.001*

The pseudo R-squared values for the regressions in table 3 are consistently very small in magnitude, dropping from 3.6 per cent in wave one to 2.5 per cent in wave 9. This indicates that the combined predictive power of birth characteristics is weak even for the early waves. This happens because a large number of variables which affect the probability of response are not accounted for in the model. Such variables may include the characteristics of interviewers and the conditions surrounding the collection of the data. However, metadata are not available in BCS70 and any underlying theory of response as a social process is absent as a means of informing the selection of covariates. On a practical note, of course including additional covariates from other waves will inevitably lead to a reduction in the observations remaining available to model response for a particular wave, due to missingness as the result of previous attrition or item non-response.

Turning our focus to the parameter estimates for the regression results in table 3 we obtain some indicative understanding of the response process. As

expected, women are more likely to respond than men. The effect of gender becomes stronger after wave 3 (CM age 10). Interestingly, the sample becomes more skewed towards women at a time when the responsibility for co-operation is transferred from parents to CMs. Individuals whose parents were married at birth are also more likely to respond than those whose parents were single. In contrast, individuals whose mothers were living in London in 1970, and those whose mothers did not attempt breast-feeding, are less likely to respond. Further, the probability of response drops with parity. The higher the number of older siblings a CM has the less likely he or she is to respond. The probability of response is strictly increasing in relation to the age of mothers at delivery for all waves, and is higher for CMs whose mothers had a longer formal education.

The higher the social class of the CM's father at birth the more likely the CM will respond. However, the effect of social class is only significant for the lowest three social classes (e.g. skilled manual, unskilled manual and lowest grade workers). Father's age at completion of education does not appear to have a significant effect on the likelihood of response. One should note that the effects of the covariates are highly significant in statistical terms except father's social class and father's age at completion of education. However, the explanatory power of these models remains very weak, suggesting that the

missingness mechanism is driven by a more complex set of influences than birth characteristics alone.

The results from the logistic regression analysis confirm the findings from the descriptive analysis. In other words, attrition is not a random process and dropout will most likely depend on some of the CM characteristics. Hence, working with only the productive cases from any sweep without any adjustments will lead to bias unless we have MCAR.

In order to make the illustration in the next section, we construct IPWs based on the response model for wave 4 (fourth column in table 3). These weights are to adjust for unit non-response in wave 4 but not for item missingness or unit non-response from previous waves.

Table 4 presents a set of point estimates for the birth characteristics used as covariates in the response probability models (table 3), comparing adjusted and unadjusted estimates with those obtained at birth in wave 1. In wave 1 the sample consists of 15,270 CMs which drops to 10,059 by wave 4 due to unit non-response. The first row gives the percentages for each category at birth (N=15,270), the second row gives the percentages at wave 4 (without the use of IPWs, N=10,059) and the third row gives the percentages at sweep four after adjustment using the IPWs we have constructed (N=10,059).[i]

Table 4. The impact of non-response weights in wave 4

Variables	Men	Father's occupation (skilled manual)	Mother finishing education at 15	Parents are single in 1970
Wave 1 (at birth)	51.87	59.14	45.61	4.78
Wave 4 without weights	49.93	58.59	45.43	3.63
Wave 4 with weights	51.91	59.17	45.59	4.78

Table 4. (continued)

Variables	Mother lives in London	Parity=+3	Number of observations
Wave 1 (at birth)	12.34	13.75	15,270
Wave 4 without weights	9.53	12.44	10,059
Wave 4 with weights	12.31	13.72	10,059

When there is no adjustment for unit non-response the percentage estimates for wave 4 deviate substantially from their original value at birth. This indicates that the sample is biased according to these

birth characteristics. In contrast, when the IPWs are applied, the percentages are almost identical to their original values despite the loss of observations due to non-response. Hence, non-response weights are

effective in reducing bias in produced descriptive estimates for our selected birth characteristics. Obviously, the efficacy of these IPWs will be highest when applied to substantive analyses which include the covariates used in the construction of these weights in the explanatory part of the model. Given their weak explanatory power, this will be most unlikely in models which draw upon wider range of predictors.

Next, we will explore the performance of our IPWs and MI for an artificial example based upon BCS70 data. The purpose of the illustration is to assess the effectiveness of weights and imputation techniques in dealing with statistical bias in regression analyses (in terms of both estimates and their standard errors).

An illustration to examine the effect-iveness of reweighting and imputations?

The generation of BCS70 datasets used to illustrate the application of re-weighting and multiple imputation

We generate a BCS70 dataset imagining that the analyst wishes to predict vocabulary scores at the age of 16 years (Parsons, 2014) as an outcome based upon gender, the gross family income per week (measured at wave 3 when the CM is aged 10 years) and highest parental qualification (measured at wave 4) as explanatory variables. Income was chosen from a previous wave (wave 3) to the outcome variable in order to illustrate the complexity of working with data collected at more than one time point. Figure 4 below summarises the history of the sample sizes that ultimately generate our example. Sample A refers to the number of CMs in wave 1 (15,270), sample B those CMs available in wave 4 (10,059). The difference between samples A and B is due to cumulative attrition and unit non-response between wave 1 and wave 4. Sample C is a set of complete cases (CC) selected from sample B for the four variables used in our illustration. Finally, we generate sample D from sample C to contain an artificial amount of missing items in a number of distinct steps:

1. For the vocabulary scores we introduce missing values for 10 per cent of the cases in a random or haphazard manner unknowingly related to the values of any other variable.
2. We recode the father's occupation into a binary variable with two categories, manual and non-manual.

3. On income and highest qualification we introduce 40 per cent missing values at random if the father is from the manual group and 10 per cent at random if he is from the non-manual group. We do this for each of the two variables separately. The absolute difference in terms of missing values between the two categories is then 30 per cent. This difference is considered large enough to reveal a bias in the analysis when compared to the CC analysis based on sample C. Father's occupational group is not included in any of our subsequent models.
4. We do not introduce any missing values for gender since it is unlikely to suffer from any item missingness in practice (especially as it is recorded at the birth of a CM).

Finally, the sample size reported for sample D (1926) in figure 4 is that for a reduced set of CC obtained by applying listwise deletion to the modified version of sample C (C_{miss}) after imposing the missingness reported above.

The generated pattern of missingness attempts to provide a semblance of the reality of working with longitudinal data. Even if the application of IPWs were to be efficient in dealing with unit non-response bias, item missingness will still lead to additional bias if it is not completely at random. This bias will vary according to the magnitude of missingness and how much it deviates from MCAR. In this example we introduced item-missingness on income and highest qualification. The magnitude of missingness varied according to the father's social class in 1970 – manual vs. non-manual – with those CMs with fathers from the manual group being less likely to answer the questions. Hence, the item missingness we have introduced is not MCAR, and the extent to which it can be assumed to be MAR or MNAR depends on whether the father's social class is treated as observable (MAR) or unobservable (MNAR) in the imputation procedures. In figure 4, the reduced sample C will artificially represent the 'truth' for purposes of comparing various analyses.

Note that the difference between samples B and C is due to the combination of missing values on the three variables included in the substantive model: vocabulary scores (49.92 per cent item missing for 5,021 CMs), income (17.47 per cent item missing or 1,757), parents highest qualification (11.81 per cent item missing for 1,188 CMs). Note that item missingness is high on vocabulary scores because not all CMs had undertaken this test.

Figure 4. The size of the sample
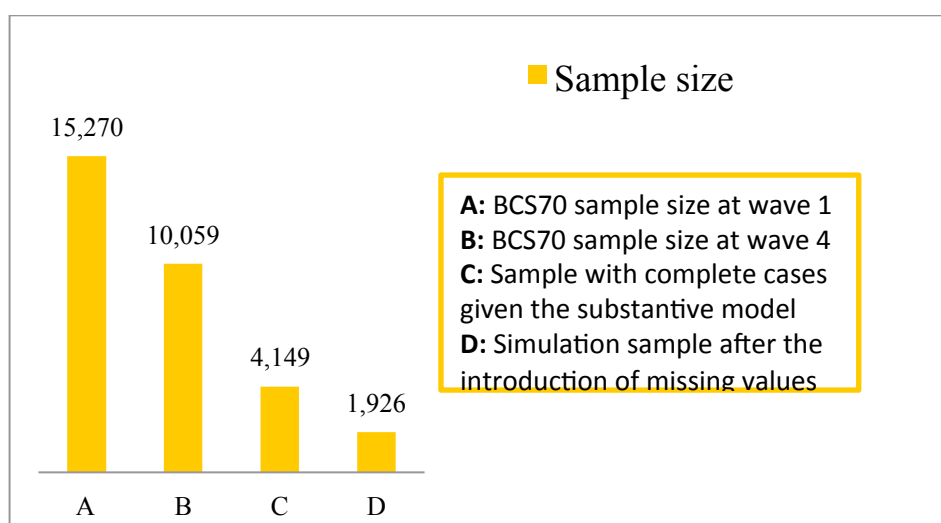
Modelling strategy

After introducing item missingness into sample data C, we estimate the following models:

Model 1: this model is estimated using sample C (with CC) while applying the non-response weights to adjust for the bias resulting from the unit non-response that is reported between waves 1 and 4 (A-B).This model has no item-missingness because sub-sample C is fully observed or complete for our substantive example. The parameter estimates and standard errors provide a 'benchmark' model to which all other model estimates can be compared.

Model 2: this model is estimated using sample C again but without applying the non-response weights. By comparing this model with model 1 we will be able to ascertain by how much the application of non-response weight adjustment affects the findings irrespective of the need to fill-in any item missingness.

Model 3: this model is estimated using the reduced sample D (with listwise deletion applied after our hypothetical pattern of missingness is imposed on sample C). This model will suffer from both biases (i.e. unit non-response and item missingness) and we are not applying any adjustment technique.

Model 4: this model is estimated using sample D with IPWs applied to adjust for the loss of observations (across samples A to B).

Model 5: this model is estimated under MI (averaging under Rubin's Rules across 20 imputed datasets) which restores the sample size back to that reported for C (4,149). The imputations adjust for the bias resulting

from the generation of item missingness (in sample C) but not for any unit non-response.

Model 6: this model is fully adjusted and combines the MI results based upon model 5 in conjunction with the application of IPWs based on the sample loss between waves 1 and 4 (A-B).

It is important to note that none of the models adjusts for the bias resulting from the existing item missingness already present at wave 4 (sample B cases minus C cases). To that extent all models suffer from the same degree of bias, which does not affect their comparability.

The imputations in models 5 and 6 are carried out using MI under a Markov-Chain-Monte-Carlo procedure (Gilks, Richardson & Spiegelhalter, 1996) and chained equations in STATA (Royston (2009) and Royston & White (2011)). We use the MI command with a linear procedure to impute vocabulary scores because the variable is continuous, and ordinal-logit to impute income and highest qualification (the two variables are ordinal). Following the example of Goldstein (2009) we produce 20 imputed datasets.[ii] The explanatory variables used in the imputation model are based on the birth characteristics listed in table 3, with the exception of father's social class at CM's birth: the variables used are gender of the CM, parental marital status, parity, breast-feeding, mother's age at delivery, mother's age at completion of education, and father's age at completion of education. In other words, we treated father's social class as unobservable in the imputation model because we introduced item missingness based on the values of the variable father's social class. Our motivation was to ensure that the data

was more likely to deviate from being fully MAR. In other words, missingness will depend on observable and unobservable factors. In substantive terms, it could be argued that in the analysis model we introduced a proxy for social class by including parental income and education as influences upon a CMs vocabulary score.

We expect that models 5 and 6 will generate the closest results to our benchmark-model 1 in terms of the magnitude of the estimates and their standard errors. Model 3 is expected to generate the least similar results to model 1 since it does not adjust for any type of bias.

Modelling Results

Table 5. Results for regression modelling of vocabulary scores at age 16 years under varying adjustments for unit and item non-response

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Gender						
Women	8.99**	8.41**	5.50	5.93	9.91**	10.4***
	(2.921)	(2.921)	(4.256)	(4.248)	(3.098)	(3.099)
Age 10 gross family income per week (reference: under £50)						
£50 - £99	3.61	2.22	0.90	2.93	1.97	2.78
	(7.464)	(7.538)	(12.054)	(11.919)	(9.508)	(9.532)
£100 - £149	9.40	7.64	2.24	4.45	8.19	8.92
	(7.443)	(7.507)	(11.994)	(11.865)	(8.989)	(9.132)
£150 - £199	14.4	12.9	5.54	7.54	10.0	10.6
	(7.941)	(7.989)	(12.641)	(12.536)	(9.913)	(10.073)
£200 - £249	13.7	11.9	10.5	12.8	17.7	18.7
	(9.184)	(9.195)	(13.898)	(13.820)	(11.240)	(11.468)
£250 or more	28.2**	27.2**	23.4	26.3	27.0*	28.0*
	(9.523)	(9.502)	(14.131)	(14.078)	(11.140)	(11.478)
Parental highest qualification (reference: no qualification)						
Other	24.3*	26.4*	30.6	29.9	20.7	19.8
	(11.816)	(11.897)	(16.687)	(16.410)	(13.989)	(13.807)
Vocational	16.9***	17.2***	23.4**	23.6***	13.8**	14.2**
	(4.502)	(4.545)	(7.256)	(7.163)	(5.262)	(5.265)
O level	32.8***	34.1***	40.1***	38.9***	28.4***	27.9***
	(4.250)	(4.246)	(6.485)	(6.465)	(5.286)	(5.235)
A level	51.3***	51.5***	56.9***	57.3***	44.1***	44.5***
	(5.533)	(5.477)	(8.135)	(8.153)	(7.180)	(7.320)
Nurse	54.5***	56.5***	46.3**	43.1**	47.6***	46.3***
	(9.536)	(9.381)	(14.076)	(14.272)	(10.854)	(11.129)
Teacher	70.3***	69.1***	74.6***	75.4***	59.1***	60.1***
	(9.115)	(8.994)	(12.006)	(12.071)	(9.470)	(9.563)
Higher degree	85.6***	84.5***	90.7***	91.1***	73.1***	74.0***
	(5.288)	(5.250)	(7.434)	(7.444)	(6.000)	(6.030)
Constant	-55.3***	-52.1***	-44.4***	-48.0***	-49.6***	-52.3***
	(8.428)	(8.517)	(13.288)	(13.142)	(9.940)	(9.885)
N	4,149	4,149	1,926	1,926	4,149	4,149

Notes. Standard errors in parentheses * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. **Model 1**: complete cases dataset for 'benchmarking' with IPWs applied. **Model 2**: complete cases dataset but no IPW adjustment. **Model 3**: listwise deleted dataset after missing data pattern imposed. **Model 4**: as above with an IPW adjustment. **Model 5**: dataset with missing data pattern imposed with MI adjustment. **Model 6**: as above together with an IPW adjustment.

The findings show that models 1 and 2 generate almost identical results even though model 2 does not adjust for unit non-response. Both models use the sample with complete cases (C) and do not suffer from item missingness. This similarity between the two models is a first indication that non-response weights do not improve the parameter estimates from regression analysis by much. That may quite possibly be because the weights have been constructed using a selection of birth characteristics which have very low predictive power (see pseudo R-squared in table 3).

In terms of standard errors, model 1 generates the lowest standard errors on all estimates, as we expected. Comparing model 1 with model 2, the use of IPWs generates slightly higher standard errors for the latter, as is to be expected. Models 5 and 6, which both involve MI with and without the application of IPWs, generate the closest results to model 1. This is the case for all variables without exception. Notably, the standard errors on the estimates in model 6 are almost identical to those in model 5. In contrast models 3 and 4, based upon listwise deletion with and without weight adjustment, generate parameter estimates that deviate substantially from model 1, together with larger standard errors (as a result of a halving of the sample size reported for models 5 and 6) indicating that they are least reliable.

For the parameter estimates, the picture is mixed. Models 1 and 2 provide very similar estimates since neither suffers from item missingness. Model 2 slightly deviates from model 1 because it does not adjust for unit non-response. Models 5 and 6 generate the closest findings to model 1 on gender and most of the modalities of income. However, for parental highest qualification, models 3 and 4 generate closer results to model 1 on 3 out of 7 modalities. Hence, one can say that MI appears to bring the estimates closer to their original values, but with some exceptions.

Discussion

Our illustration used unit non-response weights to adjust for lost cases and multiple imputations to adjust for any item missingness in our substantive model. These imputations adopt an imputation model based on auxiliary birth characteristics. The efficacy of both our adjustment methodologies depend on a number of conditions. The efficacy of the application of IPWs will depend on the

predictive power of the response models used in their construction. In our case, the predictive power of the models was weak. Further, the use of wave specific weights is also undermined when variables from previous waves are used in our regression analysis (as in the case of parental income). This would therefore be an instance where the application of longitudinal weights might improve the adjustments. The efficacy of reweighting could also be improved by considering other variables that would better predict unit non-response, in particular metadata which is not currently collected in BCS70 waves.

Considering the impact of item missingness, our findings show that the application of MI improves the precision (standard errors) on all variable estimates, without exception. This happens because imputations increase sample size to its former level and therefore improve accuracy. Further, the analysis of imputed data generated parameter estimates that were closest to our benchmark model (1) with the exception of parental highest educational qualification. Ideally, this would require further investigation perhaps by repeating the generation of our chosen levels of missingness across many analyses (i.e. a simulation study) to assess the stability of these findings. The levels of item missingness were introduced on income and parental qualifications based on father's social class (manual vs. non-manual) and thought to be quite substantial at a 30 per cent difference in loss between the non-manual and manual social class groupings. In this manner, we would expect, as evidenced, that imputation would outperform any estimation based on listwise deletion. In contrast, if the difference in terms of the levels of missingness between the two categories had been very small then the researcher can assume that item missingness is almost MCAR and that carrying out MI prior to analysis will not make much difference to the results. By and large, the efficacy of imputations will depend on the extent of item missingness, how much it appears to deviate from MCAR, and whether or not the researcher is able to anticipate the missingness mechanism under appropriate assumptions (MAR or MNAR). The application of IPW and MI either in combination or as stand-alone adjustments are based on MAR. Under MI, it is important that the imputation procedure is appropriate for the type of the variables to be imputed (e.g. linear procedures for

continuous variables, ordinal and multinomial logit procedures for ordinal and multinomial variables, multilevel models for nested observations), in order to generate robust and valid inference. However, it is also advisable to conduct sensitivity analyses under MNAR as described by Carpenter and Kenward (2013). Instances of such checks are all too rare in the literature. Finally, for further information about missing data that may include interaction or non-linear terms, see Carpenter and Plewis (2011) and Goldstein et al., (2014) as these issues have not been covered in our illustration.

Conclusion

In this paper, we examine the extent of non-response in the British Cohort Study begun in 1970 (BCS70) and its effect on sample composition over the nine available data waves (1970 to 2012). The findings are based on BCS70, but their relevance extends to a wide range of birth cohort studies such as the National Child Development Study of 1958, the 1946 MRC National Survey of Health and Development, and the Millennium Cohort Study (for further information see:

 http://www.closer.ac.uk/data-resources/explore
 (Cohort & Longitudinal Studies Enhancement Resources, 2014).

We analyse the determinants of non-response using binary logistic regression models with selected birth characteristics as explanatory variables. We find that men from lower social backgrounds and with less educated parents are less likely to respond. However, despite the statistical significance of the estimated regression coefficients the predictive power of the models is weak. In the second section of the paper, we develop a hypothetical substantive model in order to illustrate the impact of the use non-response weights and imputation techniques on our inference.

The construction of IPW is itself a challenge. In our illustration there is an intuitive approach drawing upon CM's birth characteristics simply because these characteristics are available for almost all of the cohort and loss can be traced over time. Using these items to provide weights does not actually improve the efficacy of our models, largely because they have weak predictive power. The judicious inclusion of further items requires a better-developed theory of response co-operation and access to metadata. What is clearly attractive about MI is that it enables the researcher to restore the sample size to include cases with partial information. In longitudinal analysis this implies that it is possible to regard any wave-specific non-response as a set of missing items in a longitudinal record spanning the life of the cohort. In our illustration we were able to demonstrate that model estimates based on MI came closest to our benchmark model. However, not all estimates were in close agreement.

As the life of a cohort continues it will be increasingly important for analysts to make best use of the available data, which implies that it will be unwise to ignore cases over time with partial information arising from attrition, wave-specific dropout and item non-response. In order to exploit the availability of powerful software tools and the range of approaches now available under MAR and MNAR, assumptions will require that sufficient time and effort is made to understand and model missingness mechanisms as an integral part of the research process.

Acknowledgements

References

Arbuckle, J. L. (1996). Full information estimation in the presence of incomplete data. In G. A. Marcoulides & R. E. Schumaker (Eds.), *Advanced structural equation modelling: issues and techniques* (pp. 243–277). Mahwah, NJ: Lawrence Erlbaum Associates.

Carpenter, J. R., & Plewis, I. (2011). Analysing longitudinal studies with non-response: issues and statistical methods. In M. Williams & W. P. Vogt (Eds.) *The SAGE Handbook of innovation in social research methods* (pp.498-523). London: SAGE Publications Ltd. http://dx.doi.org/10.4135/9781446268261.n28

Carpenter, J. R. & Kenward, M. G. (2013). *Multiple imputation and its application*. Chichester: John Wiley & Sons Ltd. http://dx.doi.org/10.1002/9781119942283

Centre for Longitudinal Studies (2014). Technical Report of the 1970 British Cohort Study: Age 42 Survey 2012-2013). TNS British Market Research Bureau Ltd.: London

Cohort and Longitudinal Studies Enhancement Resources (CLOSER). (2014). URL:http://www.closer.ac.uk/data-resources/explore/ accessed 15.12 on 23.10.2014.

Davey, A. & Savla, J. (2010). *Statistical power analysis with missing data: a structural equation modelling approach.* New York: Taylor Francis Group.

Enders, C. (2008). A note on the use of missing auxiliary variables in Full Information Maximum Likelihood-based Structural Equation Models. *Structural Equation Modelling: A Multidisciplinary Journal*, *15*(3), 434-448. http://dx.doi.org/10.1080/10705510802154307

Gilks, W., Richardson, S., & Spiegelhalter, D. (1996). *Markov chain Monte Carlo in practice*. London: Chapman and Hall.

Goldstein, H. (2009). Handling attrition and non-response in longitudinal data. *Longitudinal and Life Course Studies, 1*(1), 63-72.

Goldstein, H., Carpenter, J. R., & Browne, W. J. (2014). Fitting multilevel and multivariate models with missing data in responses and covariates that may include interactions and non-linear terms. *Journal of the Royal Statistical Society A,* 177, Part II, 553-564. http://dx.doi.org/10.1111/rssa.12022

Hawkes, D., & Plewis, I. (2006). Modelling non-response in the National Child Development Study. *Journal of the Royal Statistical Society A, 169*(3), 479-492. http://dx.doi.org/10.1111/j.1467-985X.2006.00401.x

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica, 47*(1), 149-74. http://dx.doi.org/10.2307/1912352

Kalton, G. (1986). Handling wave nonresponse in panel surveys. *Journal of Official Statistics, 2*(3), 303-314.

Lepkowski, J. M. (1989). Treatment of wave nonresponse in panel surveys. In D. Kasprzyk, G. Duncan, G. Kalton and. M. P. Singh (Eds.) *Panel Surveys* (pp. 348-374). New York: Wiley.

Little, R., & Rubin, D. (2002). *Statistical analysis with missing data,* 2nd ed. Hoboken, N.J: Wiley. http://dx.doi.org/10.1002/9781119013563

McDonald, J. W. and Ketende, S. Nonresponse weight adjustments using multiple imputation for the UK Millennium Cohort Study. Statistics Canada International Symposium Series - Proceedings, Symposium 2009 : Longitudinal Surveys: from Design to Analysis, 27-30 October, Gatineau, Canada.

McDonald, J. W., & Ketende, S. C. (2009). Nonresponse weight adjustment using multiple imputation for the UK Millennium Cohort Study. *Proceedings of Statistics Canada Symposium: Longitudinal Surveys from Design to Analysis*, 27-30 October, Ottawa: Statistics Canada.

Micklewright, J., Skinner, C., & Schnepf, S. (2012). Non-response biases in surveys of school children: the case of the English PISA samples. *Journal of the Royal Statistical Society A*, *175*(4), 915-938. http://dx.doi.org/10.1111/j.1467-985X.2012.01036.x

Nathan, G. (1983). A simulation comparison of estimators for a regression coefficient under differential non-response. *Communications in Statistics: Theory and Methods, 12*(6) 645-659.

Nathan, G., & Holt, D. (1980). The effect of survey design on regression analysis. *Journal of the Royal Statistical Society B*, *42*(3), 377-386. http://dx.doi.org/10.1080/03610928308828485

Nevalainen, J., Kenward, M. G., & Virtanen, S. M. (2009). Missing values in longitudinal dietary data: a multiple imputation approach based on fully conditional specification. *Statistics in Medicine 28*(29), 3657-3669. http://dx.doi.org/10.1002/sim.3731

Parsons, S. (2014). British Cohort Study Data Note. Centre for Longitudinal Studies, University of London.

Pfeffermann, D., Moura, F., & Silva, P. (2001). *Multilevel modelling under informative probability sampling*. S3RI Methodology Working Paper M04/09. University of Southampton.

Plewis, I. (2007a). Non-response in a birth cohort study: the case of the Millennium Cohort Study. *International Journal of Social Research Methodology*, *10*(5), 325-334. http://dx.doi.org/10.1080/13645570701676955

Plewis, I. (2007b). *The Millennium Cohort Study: technical report on sampling.* Centre for Longitudinal Studies, the Institute of Education, University of London.

Plewis, I., Calderwood, L., & Ketende, S. (2009). Sample loss from cohort studies: patterns, characteristics and adjustments. *Proceedings of Statistics Canada Symposium: Longitudinal Surveys: from Design to Analysis.* 27-30 October, Ottawa: Statistics Canada.

Plewis, I. (2011). *Auxiliary variables and adjustments for missingness in longitudinal studies*. Cathie Marsh Institute for Social Research, CCSR working paper No. 2011-04.

Royston, P. (2009). Multiple imputation of missing values: further update on ICE, with an emphasis on categorical variables. *The Stata Journal*, *9*(3), 466-477.

Royston, P., & White, I. R. (2011). Multiple imputation by chained equations (MICE): implementation in Stata. *Journal of Statistical Software, 45*(4), 1-20.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(6), 581-92. http://dx.doi.org/10.1093/biomet/63.3.581

Rubin, D. B. (1987) *Multiple imputation for non-response in surveys*. New York: J. Wiley & Sons. http://dx.doi.org/10.1002/9780470316696

Rubin, D. B. (2004). The design of a general and flexible system for handling non-response in sample surveys. *American Statistician, 58*(4), 298-302. http://dx.doi.org/10.1198/000313004X6355

Seaman, S. R., & White, I. R. (2011). Review of inverse probability weighting for dealing with missing data. *Statistical Methods in Medical Research*, *22*(3) 278-295. http://dx.doi.org/10.1177/0962280210395740

Seaman, S. R., White, I. R., Copas, A. J. & Li, L. (2012). Combining multiple imputation and inverse-probability weighting. *Biometrics, 68*(1), 129-137. http://dx.doi.org/10.1111/j.1541-0420.2011.01666.x

Schafer, J., & Olsen, M. (1997). Multiple imputation for multivariate missing-data problems: a data analyst's perspective. *Multivariate Behavioural Research, 33*(4) 545-571. http://dx.doi.org/10.1207/s15327906mbr3304_5

Schouten, B., & de Nooij, G. (2005). *Non-response adjustment using classification trees.* Voorburg: Statistics Netherlands.

Taylor, M., Brice, J., Buck, N., & Prentice-Lane, E. (2010). *British Household Panel Survey User Manual, Volume A: Introduction, technical report and appendices.* Colchester: University of Essex.

Welch, C., Bartlett, J., & Petersen, I. (2014). Application of multiple imputation using the two-fold fully conditional specification algorithm in longitudinal clinical data. *The Stata Journal*, *14*(2), 418-431.

Wiggins, R. D., Schofield, P., Bartley, M., Sacker, A, & Head, J. (2004). Social determinants of changes in minor psychiatric morbidity over time in the British Household Panel Study 1991-1998. *Journal of Epidemiology and Community Health*, *58*, 779-791. http://dx.doi.org/10.1136/jech.2003.015958

Endnotes

[i] Note that the number of observations in wave 1 (15,270) and wave 2 (10,059) deviate from those in table 2 because some of the CMs included in the category 'participated' have missing birth characteristics. Hence, the observations included in the computation of descriptive statistics and in the logit models are those with non-missing birth characteristics.

[ii] Note that we repeated the same analysis with 100 imputations. The difference in magnitude of the results was very limited which suggests that 20 imputations are enough to generate valid inference.