Multilevel modelling approach to analysing life course socioeconomic status and compensating for missingness

Adrian Byrne*; Social Statistics, University of Manchester, UK

Natalie Shlomo; Social Statistics, University of Manchester, UK

Tarani Chandola; Social Statistics, University of Manchester, UK

*Adrian.Byrne@manchester.ac.uk

Abstract

This paper investigates the heterogeneity between individuals in relation to a finely-grained measure of socioeconomic status over the life course. We examine the extent to which parental socioeconomic status can explain this life course socioeconomic status heterogeneity between individuals using 1958 National Child Development Study data. This empirical study shows how substantial between-individual variation in life course socioeconomic status is suitably captured by a step function multilevel model. Our results highlight the significant contribution of parental socioeconomic status in explaining the divergence in achieved socioeconomic status over the life course. We also explore the issue of missing data in relation to our model of interest and show evidence of missing at random when including sex and region of residence as model covariates. We compare the empirical difference between the full information maximum likelihood approach and two methods designed to compensate for missing at random data namely multilevel multiple imputation and multiple imputation chained equations. The former method is congenial with our model of interest whereas the latter method is computationally more efficient. The two multiple imputation methods produce similarly plausible results to the full information maximum likelihood approach given the underlying data and model of interest. The results for the complete case and partially observed, both defined in terms of our life course dependent variable, proved to be less similar on average. This evidence suggests the full information maximum likelihood approach using all available cases is appropriate.

Keywords: multilevel; longitudinal; socioeconomic status; life course; missing data; multiple imputation

Introduction

This study examines the extent to which parental socioeconomic status (SES) can explain the variation between individuals in relation to their life course SES development and also interrogates this examination with respect to missing data. SES is often measured either as a combination of education, income and occupation or by investigating the aforementioned elements separately. This framework is akin to that set out by Pierre Bourdieu in his "The Forms of Capital" (Bourdieu, 1986), i.e. economic capital (income), social capital (occupation) and cultural capital (education). Moreover, there are many different measures of SES that can be used throughout the life course (Galobardes, Shaw, Lawlor, & Lynch, 2006; Galobardes, Shaw, Lawlor, Lynch, & Smith, 2006). Although some of these variables are categorical (e.g. occupational class or educational qualifications), SES is commonly conceptualised as a continuous variable, referring to the rank or social standing of an individual or group (American Psychological Association, 2007). Therefore, treating SES as a continuous variable can help reveal greater relative inequities between individuals in a more fine-grained fashion.

However, most research evaluating life course SES uses categorical measures of SES thereby producing more aggregated floor and ceiling effects when analysing SES trajectories and social mobility (Goldthorpe & Jackson, 2007; Li & Devine, 2011; Sturgis & Sullivan, 2008). Consequently, a great deal is known about mobility between large categories of occupations for example, but less is known about the heterogeneity within these large categories (Laurison & Friedman, 2016). We argue in this paper that employing a continuous, repeated measure of life course SES provides a better opportunity of finding any potentially important differences obscured by the floor and ceiling effects of these categorical measures during the life course as it is more sensitive to changes than conventional broad categorical measures. Therefore, this study investigates the heterogeneity between individuals in relation to a finely-grained continuous measure of SES over the life course.

Although the statistical methods used to analyse SES can vary (Pollitt, Rose, & Kaufman, 2005), another feature of the conventional social mobility research is examining the origin (childhood) versus destination (adulthood) transition matrix thereby only using two time points in the life course (Blanden, Goodman, Gregg, & Machin, 2004; Goldthorpe & Jackson,

2007; Li & Devine, 2011). Incorporating more time points into the life course analysis enables us to estimate the changing effect of childhood SES on adult SES over the life course by interacting childhood SES with the time points. This examination could not be undertaken with just two time points between childhood and adulthood. In this paper we examine to what extent parental SES can explain life course SES heterogeneity between individuals.

An additional methodological challenge when examining life course SES using longitudinal survey data is that missing data is an unavoidable reality. While many studies report missingness (Niedzwiedz, Katikireddi, Pell, & Mitchell, 2012), not many examine in detail how the missing data affects inference established from the model of interest (MoI). Furthermore, past analyses of life course SES have employed a direct maximum likelihood approach under certain assumptions without conducting any sensitivity analysis to ensure their inferences are robust to the missingness (Sturgis & Sullivan, 2008). In this paper we explore the issue of missing data in relation to the MoI by investigating the empirical difference between suitable missing data methods and consider how missing data affects the relationship of parental SES on life course SES heterogeneity between individuals.

We begin by introducing the longitudinal birth cohort survey data, life course SES measure and model covariates. We next motivate our choice of life course statistical model for a continuous, repeated SES outcome and adjust for parental SES, sex and region of residence. We then turn our attention to the issue of missing data in relation to our life course SES measurement. We introduce two multiple imputation (MI) methods to compensate for missing data; one being congenial with our MoI and one being computationally more efficient. Results from different methods are presented before concluding with our findings on the effectiveness of our MoI in evaluating life course SES development and the impact of compensating for missing data.

National Child Development Study life course data

Building upon the framework set out by (Schuller, Wadsworth, Bynner, & Goldstein, 2011), this paper conducts a secondary analysis of existing longitudinal birth cohort survey data to enhance our understanding of SES throughout the life course. Our chosen birth cohort

longitudinal dataset is the 1958 National Child Development Study (NCDS) (University of London. Institute of Education. Centre for Longitudinal Studies., 2008a, 2008b, 2008c, 2008d, 2012, 2014, 2015). The NCDS follows the lives of all people born in England, Scotland and Wales in one particular week of March 1958. Since the birth survey in 1958, there have been nine further 'sweeps' of all cohort members (CMs) at ages 7, 11, 16, 23, 33, 42, 46, 50 and 55. In the first three sweeps (at ages 7, 11 and 16), the target sample was augmented to include immigrants born outside of Great Britain in the same week. The survey CMs remain part of the target sample until they either die or permanently emigrate from Great Britain. The total eligible sample, including those not resident in Great Britain up to the age of 16 in 1974, is 18558. The survey data is collected at individual level across Great Britain making it a nationally representative longitudinal cohort study. There are currently no survey weights to compensate for attrition.

As our target population is based on those with known occupation data over the life course, the total eligible NCDS sample size of 18558 reduces to 14268 individuals as 4290 (23.1%) CMs have no occupation data over the life course (6 time points) between the ages of 23 and 55. Our target population has a 51:49 men to women ratio whereas the ratio is 54:46 for those CMs omitted from our target population with no occupation data. As there are a total of 14268 person-level (defined as Level 2 units "L2" hereafter) available cases in this analysis, that allows for a possible total of 85608 (14268*6 time points) occasion-level (defined as Level 1 units "L1" hereafter) observations over the life course. However, with missing data in our response variable this number is reduced to 53958 L1 units. Including the model covariates with their missingness reduces the number of L1 units to 37478 and the number of L2 units to 9622 representing 44% and 67% respectively of the sample at each level.

Life course SES

We derive our life course outcome measure of SES from the Occupational Earnings Scale (Erzsebet Bukodi, Dex, & Goldthorpe, 2011; Nickell, 1982). This measure injects a form of hierarchy by classifying occupations into Standard Occupational Classes and then by ordering each of these classes according to their mean hourly wage rate using ONS Annual Survey of Hours and Earnings data (1997-2013). By linking published earnings data to

routinely collected occupation data the problem of survey responders not declaring their income is avoided. Therefore, we choose a hierarchical occupation variable that is more observed than income and more fine-grained than other variables associated with SES such as level of education and social class. These mean hourly wages have been adjusted for inflation and wages are deflated to 1997 prices using ONS Consumer Price Index data. The 1997 prices are imposed on pre-1997 occupation wage data and post-1997 occupation wage data have been deflated to keep prices constant at 1997 to aid comparability across the life course. These hourly wage data have also been transformed onto the natural logarithmic scale to correct for positive skewness.

We use this Occupational Earnings Scale to develop NCDS occupation data as a continuous, repeated outcome measure of SES over the life course (age range 23-55 over 6 time points).The scale was originally developed in the course of research into the determinants of 'occupational success' and is based on a well-defined attribute of occupations, namely earnings (Nickell, 1982). Its construct validity is easier to appreciate than more complicated, composite measures of socioeconomic status over the life course with respect to longitudinal measurement invariance issues. (Erzsebet Bukodi et al., 2011) argue that the Occupational Earnings Scale can be used as both an explanatory variable and as dependent variable in relation to a range of individuals' life-chances and life-choices and as a basis for assessing occupational mobility and success. (Erzsébet Bukodi & Dex, 2009) demonstrated that the Occupational Earnings Scale is largely gender-neutral and converted the scale into scores ranging from 1 to 100 by way of standardising over the life course rather than adjusting for inflation as we have done in this paper thereby introducing floor and ceiling effects which are absent in this study. We denote our outcome variable as "Mean Hourly Occupational Earnings" (MHOE) hereafter. By way of a sensitivity analysis, we investigated correlations over the life course between MHOE, take-home pay and social class (NS-SEC) using our NCDS dataset. Both MHOE and NS-SEC had average correlations of 0.6 with take-home pay and had an average correlation of 0.8 with each other. Therefore, NS-SEC is similar to MHOE in terms of its approximation for income with the difference being the granularity of each measure.

Parental SES

Our main predictor of interest is the NCDS CMs' parental SES. While studies such as (Erola, Jalonen, & Lehti, 2016; Kumar, Kroon, & Lalloo, 2014) highlight the importance of parental SES in the life course, we explicitly examine the legacy effects of parental SES on adult (child of parents) SES across the life course using birth cohort longitudinal data.

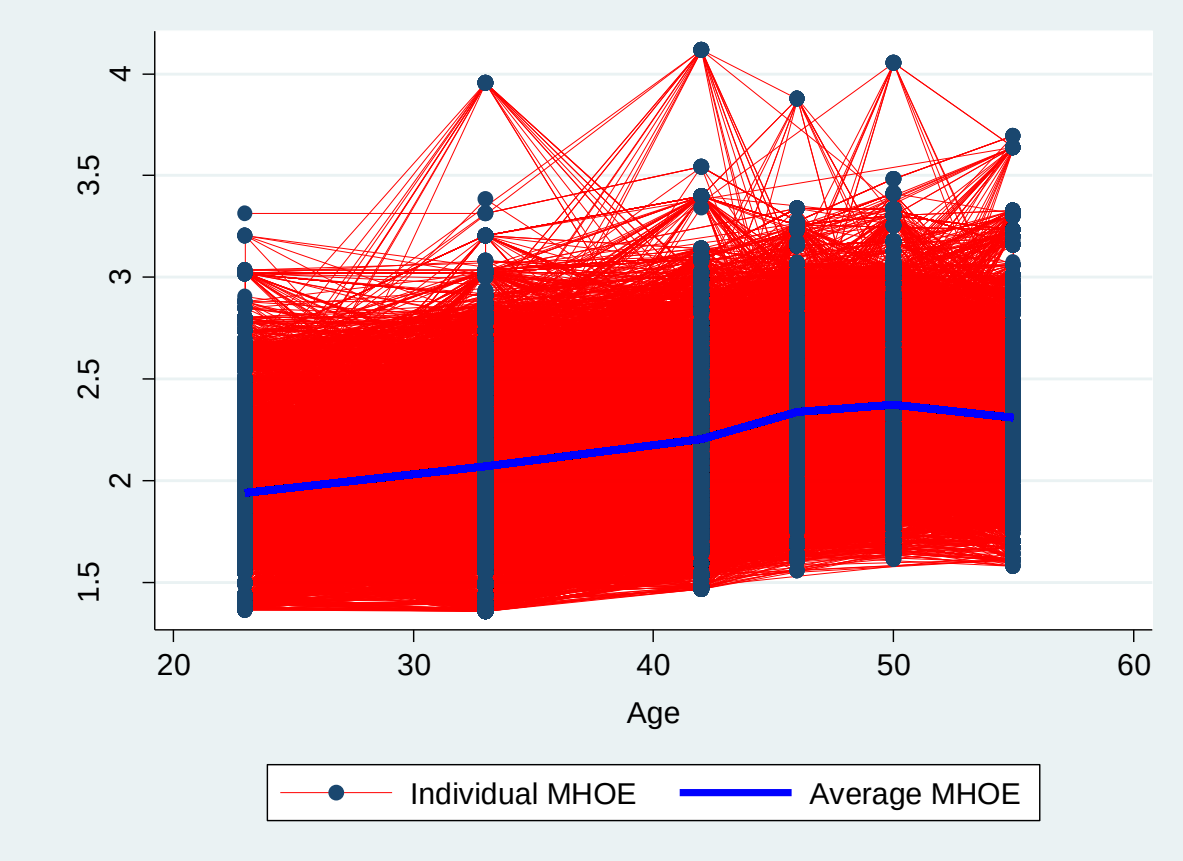
Similar to (Caro & Cortés, 2012), we treat this predictor as a formatively causal composite response to observed variables that are associated with SES; namely social class (2 variables consisting of 8 categories for both parents), age left education (2 variables consisting of 10 categories for both parents) and housing tenure status (1 variable consisting of 6 categories). These variables were collected when survey CMs were aged 16 (1974) and we chose age 16 because the same social class and education variables were available for both parents. Formatively causal composite response here means that the observed variables cause the latent SES construct and not the other way round as is the case with factor analysis (Bollen & Bauldry, 2011). We apply the ordinal Principal Component Analysis (oPCA) methodology developed by (Kolenikov & Angeles, 2009) to derive the main component which represents a linear summary of our selected parental SES variables. All five discrete variables are ordinal with higher values representing higher social class, more education and greater ownership of accommodation. The number of categories across the five variables allows for a possible 38,400 (8*8*10*10*6) unique combinations. Each of the five ordinal variables were first normalised so that their ranges were bounded between zero and one. Then oPCA was employed to uncover the main principal component which proved to be the only principal component with an eigenvalue greater than one and the individual component scores were all positive indicating that a higher weighted value implied a greater parental SES score. We refer the reader to Appendix A in the supplemental material to access the results from this oPCA estimation. The configured scores for the main principal component were estimated for each NCDS CM, transformed onto the natural logarithmic scale to account for positive skewness and then standardised with mean of zero and standard deviation of one to aid interpretation. We denote this person-level variable as "PSES16" hereafter. We adjust the relationship between PSES16 and MHOE by controlling for key socio-demographic variables namely sex and region of residence as advocated by (American Psychological Association, 2007).

Modelling life course MHOE using multilevel modelling

Multilevel modelling provides a method for analysing change over time whereby the repeated measures are viewed as outcomes that are dependent on some metric of time and predictors of interest at either level (time/individual) and may include cross-level interactions (Steele, 2008). Repeated observations over time, which need not be equally spaced out, constitute level one units nested within individuals at level two and the multilevel framework accounts for correlations of observations across time. Multilevel modelling summarises the change in the outcome variable for each individual over the observation period and each individual's summarised change can be allowed to vary in relation to the overall sample average summarised change. This individual variability can be summarised via the random effects employed within the multilevel model (MLM) set-up.

With respect to our life course outcome variable, Figure 1 displays both the average and individual life course MHOE trajectories. The average life course trend appears to show upward linearity between the ages of 23 and 42 (1981 – 2000), steeper growth between the ages of 42 and 46 (2000 – 2004), more gentle growth between the ages of 46 and 50 (2004 – 2008) and trends downward between the ages of 50 and 55 (2008 – 2013). This decline might be explained by the Great Recession which began in 2008 with the financial crisis when the survey CMs where aged 50 and/or by the effects of early retirement. However, there is substantial individual variation around this average life course trend and multilevel modelling provides a suitable way of accounting for this variation when modelling the average life course trend.

Figure 1. Life course Mean Hourly Occupational Earnings on natural log scale at 1997 prices



To model this individual life course MHOE development and variation, we adopt a multilevel framework (Steele, 2014). Without including the covariates, we explore a number of different time functions to empirically verify a suitable model given the data and the results of this analysis are presented in Table 1. We compare two different functional forms; namely orthogonal/fractional polynomial growth and step function growth whereby time is treated as categorical. Orthogonal polynomials are employed to diffuse high collinearity between time powers and fractional polynomials are investigated for extreme nonlinearities. With only 6 time points, time dummy variables were considered instead of splines. The step function multilevel model (MLM) is a multivariate linear model comprising 6 responses corresponding to the 6 waves at which MHOE is measured. The variance-covariance structure is fully specified by random terms at the person-level with no occasion-level residuals. Furthermore, we can employ a step function as the measurement occasions are the same for all CMs.

Multilevel modelling approach to analysing life course socioeconomic status and compensating for missingness

Given the life course MHOE development and variation presented in Figure 1, the evidence suggests that individuals do vary in terms of their starting points at age 23 and their life course growth rates. Therefore, the simplest MLM presented in Table 1, denoted "Orthogonal polynomial power 1", is a random intercept and slope MLM and initial modelling not presented here confirmed that more simplistic MLMs have a poorer fit.

Table 1. Life course Mean Hourly Occupational Earnings multilevel model fit statistics

MLM time function type	Log-likelihood	Deviance	Parameters	AIC	BIC
Step function/time dummies	-12807	25613	28	25669	25919
Orthogonal polynomial power 5	-12807	25613	28	25669	25919
Orthogonal polynomial power 4	-12994	25987	21	26029	26216
Fractional Polynomial inverse cubic power 4	-13083	26165	21	26207	26394
Orthogonal polynomial power 3	-13452	26905	15	26935	27068
Orthogonal polynomial power 2	-14245	28490	10	28510	28599
Orthogonal polynomial power 1	-15096	30193	6	30205	30258

Occasions = 53958; Individuals = 14268; Occasions per Individual: min=1; mean=3.8; max=6

The evidence presented in Table 1 suggests additional model complexity in terms of more parameters is merited given the underlying data. With a maximum of 6 time points over the life course, adding higher polynomial terms to the growth function up to the limit of 5 significantly improves model fit and the $4^{th}/5^{th}$ order orthogonal polynomial growth functions fit the data better than a 4^{th} order inverse cubic fractional polynomial growth function. The most suitable MLMs are equivalent in terms of model identification and fit statistics. However, we choose the step function MLM as the basis for our MoI as the multivariate formulation of MHOE life course between-individual development is simpler to interpret (Steele, 2014).

Our MoI is set up with six age dummy variables and no model intercept to aid interpretation. Each age dummy variable has a random slope attached to it allowing for individual variation at each time point in relation to the MHOE response variable. We employ cross-level

interactions between each of the age dummies (L1) and both of the L2 explanatory variables, namely PSES16 and Female, to explore life course legacy effects of these person-level predictors in relation to the MHOE response variable. To ensure model identification, the age 23 cross-level interaction effects are omitted and act as reference categories. The reference category for our nominal categorical L1 region of residence predictor is North & Midlands. Therefore, our life course step function MLM of interest is specified in Equation 1 with subscripts t for L1 time points and j for L2 individuals.

Equation 1. Life course step function multilevel model of interest

$$\begin{aligned}
&\log \text{Mean Hourly Occupational Earnings}_{tj} \\
&= \beta_1 \text{age23}_{tj} + \beta_2 \text{age33}_{tj} + \beta_3 \text{age42}_{tj} + \beta_4 \text{age46}_{tj} + \beta_5 \text{age50}_{tj} \\
&+ \beta_6 \text{age55}_{tj} + \beta_7 \text{PSES16}_j + \beta_8 \text{Female}_j + \beta_9 \text{age33}_{tj} * \text{PSES16}_j \\
&+ \beta_{10} \text{age42}_{tj} * \text{PSES16}_j + \beta_{11} \text{age46}_{tj} * \text{PSES16}_j + \beta_{12} \text{age50}_{tj} * \text{PSES16}_j \\
&+ \beta_{13} \text{age55}_{tj} * \text{PSES16}_j + \beta_{14} \text{age33}_{tj} * \text{Female}_j + \beta_{15} \text{age42}_{tj} * \text{Female}_j \\
&+ \beta_{16} \text{age46}_{tj} * \text{Female}_j + \beta_{17} \text{age50}_{tj} * \text{Female}_j + \beta_{18} \text{age55}_{tj} * \text{Female}_j \\
&+ \beta_{19} \text{South \& East}_{tj} + \beta_{20} \text{Wales}_{tj} + \beta_{21} \text{Scotland}_{tj} + u_{1j} \text{age23}_{tj} \\
&+ u_{2j} \text{age33}_{tj} + u_{3j} \text{age42}_{tj} + u_{4j} \text{age46}_{tj} + u_{5j} \text{age50}_{tj} + u_{6j} \text{age55}_{tj}
\end{aligned}$$

$$\mathbf{u} \sim N(\mathbf{0}, \boldsymbol{\Omega_u})$$

$$\boldsymbol{\Omega_u} = \mathbb{V} \begin{bmatrix} \boldsymbol{u_{0j}} \\ \vdots \\ \boldsymbol{u_{5j}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\sigma^2_{u_0}} & \cdots & \boldsymbol{\sigma_{u_0 u_5}} \\ \vdots & \ddots & \vdots \\ \boldsymbol{\sigma_{u_0 u_5}} & \cdots & \boldsymbol{\sigma^2_{u_5}} \end{bmatrix}$$

$$\text{t} = 23, 33, 42, 46, 50, 55;\ \text{j} = 1, \ldots, 9622$$

The statistical analysis was carried out using MLwiN v2.32 with the default method of estimation, Iterative Generalised Least Squares (IGLS). This estimation procedure is referred to as a Full Information Maximum Likelihood (FIML) approach and MLwiN only listwise deletes model predictors with missing data and not the outcome variable thereby establishing a complete-case analysis with respect to the predictors. Table 2 displays the fixed effect model results on the natural log scale at 1997 prices. Significance of parameter estimates is denoted as follows: $p < 0.001$ ***, $p < 0.01$ **, $p < 0.05$ * with accompanying standard errors in parentheses. We see the age dummies follow the sample MHOE means at each time point controlling for the model covariates. After controlling for sex and region of

residence, higher parental socioeconomic status at age 16 has a positive impact on life course MHOE which appears to strengthen after the age of 23 and remains significant throughout the life course. From this analysis, we infer that females with lower parental SES at age 16 residing in Wales are at a disadvantage compared to males with higher parental SES at age 16 residing in the South & East region with respect to life course SES attainment according to our measure of log mean hourly occupational earnings.

Table 2. Life course Mean Hourly Occupational Earnings multilevel model fixed effects regression coefficients (standard errors)

Age	**Coeff. (se)**	**Interaction Effects**	**Coeff. (se)**
23	2.016 (0.006)***	**Age*Parental SES at 16 (Ref: Age 23*Parental SES)**	
33	2.161 (0.007)***	33 * PSES16	0.043 (0.005)***
42	2.279 (0.007)***	42 * PSES16	0.038 (0.005)***
46	2.380 (0.007)***	46 * PSES16	0.034 (0.005)***
50	2.421 (0.007)***	50 * PSES16	0.024 (0.005)***
55	2.338 (0.007)***	55 * PSES16	0.022 (0.005)***
Parental SES at 16	0.091 (0.004)***	**Age*Female (Ref: Age 23*Female)**	
Female	-0.148 (0.007)***	33 * Female	-0.075 (0.009)***
Region of residence (Ref: North & Midlands)		42 * Female	-0.083 (0.010)***
South & East	0.025 (0.006)***	46 * Female	-0.032 (0.010)**
Wales	-0.038 (0.012)**	50 * Female	-0.029 (0.010)**
Scotland	0.011 (0.009)	55 * Female	-0.022 (0.011)*
N occasion-level units	37478	N individual-level units	9622
Min occasions per indivl	1	Avg occasions per indivl	3.9
Max occasions per indivl	6	Model deviance	14789

p-value < 0.001 ***, p-value < 0.01 **, p-value < 0.05 *

Table 3 presents the associated person-level random effects for this step function life course MLM. We see the greatest variability between individuals in terms of MHOE occurs at age 42 while the smallest variability occurs at age 23. The magnitude of the random effect covariances increases as the time between ages decreases but more interestingly all covariances are positive implying that these data are exhibiting divergent behaviour. This evidence suggests that larger deviations from the MHOE average at a preceding age are positively correlated with larger deviations from the MHOE average at a subsequent age controlling for the model predictors. This evidence is consistent with the idea of divergent MHOE growth over the life course with the already better off increasing their MHOE as they grow older at a faster rate compared to those worse off.

Table 3. Life course Mean Hourly Occupational Earnings multilevel model random effects regression coefficients (standard errors)

Random Effects	Coeff. (se)
var(age23)	0.086 (0.001)***
cov(age23,age33)	0.045 (0.001)***
var(age33)	0.139 (0.002)***
cov(age23,age42)	0.040 (0.002)***
cov(age33,age42)	0.082 (0.002)***
var(age42)	0.156 (0.003)***
cov(age23,age46)	0.038 (0.002)***
cov(age33,age46)	0.072 (0.002)***
cov(age42,age46)	0.101 (0.002)***
var(age46)	0.144 (0.003)***
cov(age23,age50)	0.035 (0.002)***
cov(age33,age50)	0.070 (0.002)***
cov(age42,age50)	0.094 (0.002)***
cov(age46,age50)	0.109 (0.002)***
var(age50)	0.145 (0.003)***
cov(age23,age55)	0.033 (0.002)***
cov(age33,age55)	0.064 (0.002)***
cov(age42,age55)	0.083 (0.002)***
cov(age46,age55)	0.091 (0.002)***
cov(age50,age55)	0.100 (0.002)***
var(age55)	0.138 (0.003)***

Occasions = 37478; Individuals = 9622; Occasions per Individual: min=1; mean=3.9; max=6

p-value < 0.001 ***, p-value < 0.01 **, p-value < 0.05 *

The case for addressing the existence of missing data

(Hawkes & Plewis, 2006) report that there are systematic differences between respondents and non-respondents at every sweep of the NCDS with the more disadvantaged survey members being more likely to be lost from the study. They conclude there is no support for the position that the NCDS data can be considered "Missing Completely At Random" (MCAR). MCAR means the probability of the data being missing does not depend on the observed or unobserved data. If the data were MCAR then complete case analysis would be sufficient with no need to account for the missing data. As both our response variable

(MHOE) and main predictor (PSES16) are related to socioeconomic advantage/disadvantage, we do not consider our missing data to be MCAR. Furthermore, mean difference statistical analysis not presented here comparing fully observed MHOE responders with those who are only partially observed over the life course confirms our data are not MCAR.

Another feature of multilevel modelling is that there is no requirement to have balanced data, so individuals who contribute fewer than the maximum number of observations, i.e. partially observed, can be retained in the analysis without further adjustment. Within a multilevel framework, these individuals can "borrow strength" from other individuals with full information. The benefit of multilevel modelling being able to handle unbalanced data is useful as longitudinal surveys may suffer from some individuals not participating in one or more waves of the study (J. Carpenter & Plewis, 2011). However, this approach is only reasonable assuming the missing data are "Missing At Random" (MAR) and a full information estimation procedure is used such as maximum likelihood assuming normally distributed data. MAR implies the probability of data being missing does not depend on the unobserved data conditional on the observed data. If these two conditions are satisfied then the actual missingness mechanism can be ignored (Rasbash, Steele, Browne, Goldstein, & Charlton, 2015).

If the probability of missingness depends on the unobserved data conditional on the observed data then the missingness can be considered "Missing Not at Random" (MNAR). However, it is not possible to distinguish between MAR and MNAR mechanisms from the observed data alone. The MAR assumption can be made more plausible by including explanatory variables in the MoI that contain predictive power in relation to the unobserved data as well as the observed data (White, Royston, & Wood, 2011). Without this, inferences on life course MHOE derived from the observed sample may be biased in relation to the parameters of interest in the target population.

In terms of the missingness contained in our life course outcome variable, Table 4 displays the decomposition of CMs in relation to number of MHOE data points across the life course excluding and including missingness in the MoI covariates. To help contrast the variation between CMs with no missing data in relation to our life course repeated MHOE measure

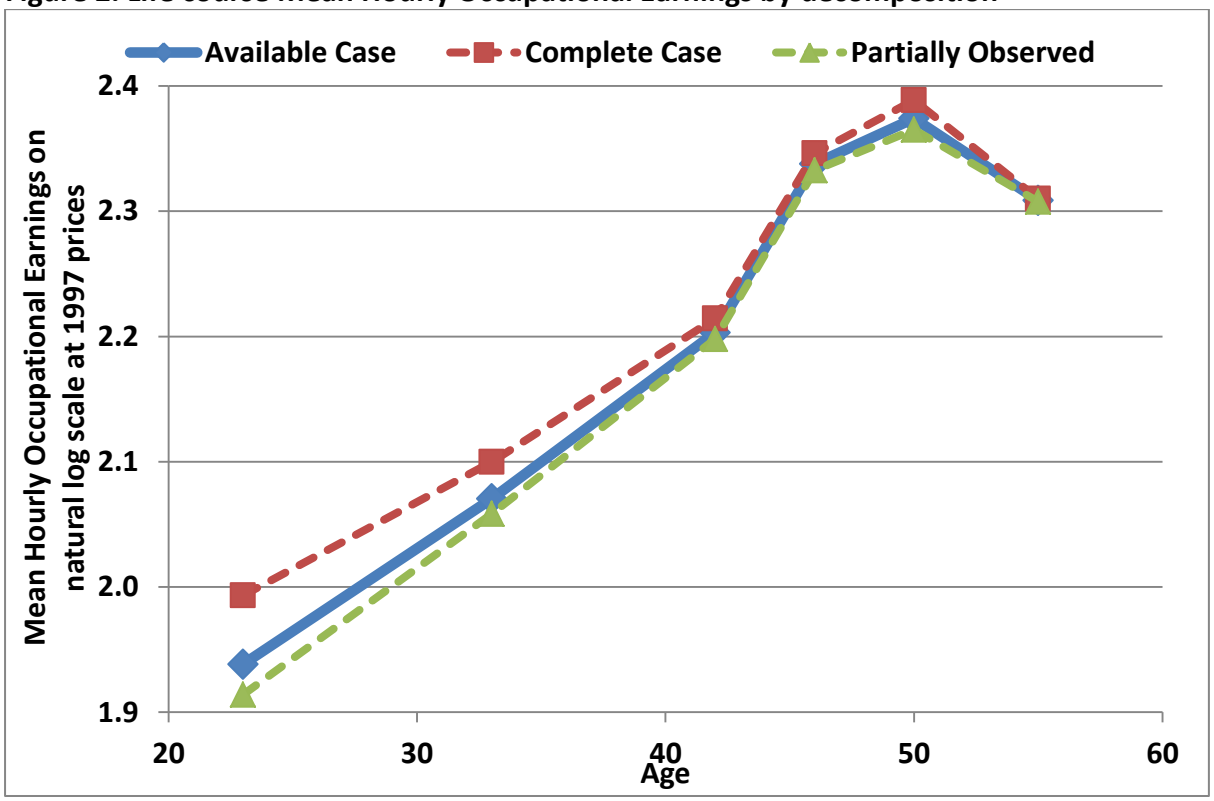
with those who do, we separate the CMs into subgroups; namely complete cases (CC; individuals with no missing life course MHOE data) and partially observed cases (PO; individuals with some missing life course MHOE data). Individuals with some or no missing life course MHOE data, i.e. CC and PO combined, are referred to as available cases (AC). Almost 22% (n = 3083) of eligible CMs can be considered CC with no missing MHOE data excluding missingness in the covariates. The remaining 78% (n = 11185) can be considered PO with 1 to 5 data time points missing. Applying the FIML approach to our MoI, as we did in the previous section, reduces our available case sample size by almost one third.

Table 4. Decomposition of cohort members with respect to Mean Hourly Occupational Earnings observations and the effect of including covariates

MHOE decomposition	N individuals excluding covariates	N individuals including covariates	% reduction with covariate missingness
Complete case	3083 (21.6%)	2263 (23.5%)	26.6
Partially observed	11185 (78.4%)	7359 (76.5%)	34.2
Available case	14268 (100%)	9622 (100%)	32.6

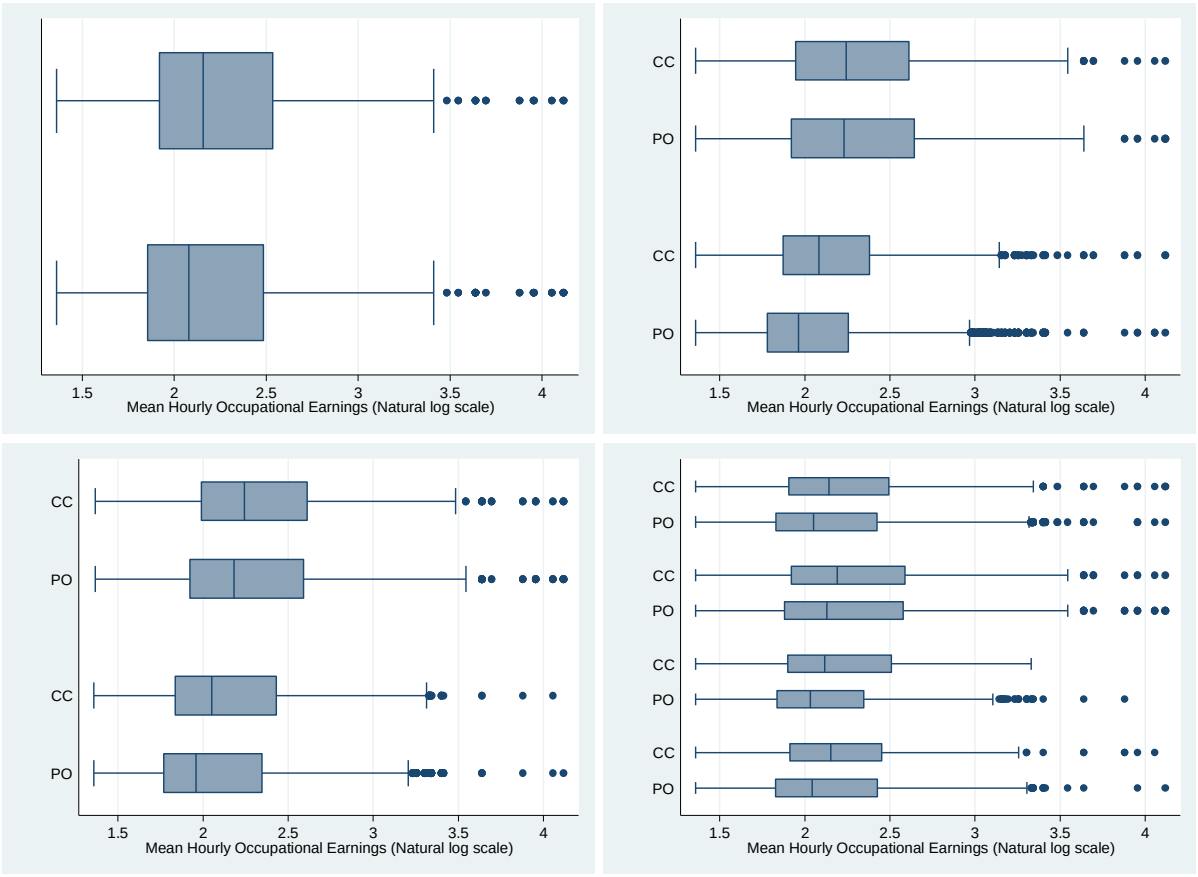
Figure 2 displays these MHOE decompositions, excluding missingness in covariates, as life course trajectory averages. The AC trajectory can be considered a weighted average of CC and PO trajectories and is located closer to the PO curve given PO cases account for 78.4% of the life course survey sample. MHOE discrepancies exist between CC and PO trajectories most prevalently at ages 23 and 33 and appear to converge thereafter. However, CC CMs attain an average MHOE at every time point that is higher than PO CMs. This shows that CC CMs tend to be more socially advantaged than CMs with some missing data. By accounting for MHOE life course development in our MoI through our step function, we are able to model the variation between CMs with full and partial information as demonstrated in Figure 2. Therefore, including this step function of MHOE life course development in our MoI adds to the plausibility of assuming our data is MAR.

Figure 2. Life course Mean Hourly Occupational Earnings by decomposition



Further to our step function being able to discriminate between fully and partially observed individuals in terms of life course MHOE development, Figure 3 shows how each MoI covariate can also discriminate between CC and PO individuals and all covariate plots reflect the overall average situation whereby CC cohort members experience higher MHOE. We argue our MoI data provides the type of evidence required to cover the likelihood of being missing at random thereby reducing the probability of data being missing due to what we do not observe.

Figure 3. Life course Mean Hourly Occupational Earnings by covariate subgroup



The evidence presented in Figures 2 and 3 strengthens the argument for assuming our data are MAR. It also suggests that a FIML approach using all available cases, like we have presented in the previous section, may provide unbiased estimation and inference (Bartlett & Carpenter, 2013). However, as our MoI predictors suffer from a non-trivial amount of missing data, we also use this evidence of MAR validity to conduct MI. We note here that the need for imputation is due to the missingness in our MoI predictors. (White & Carlin, 2010) advocate the use of MI when there exists a substantial amount of missing data due to the MoI predictors. In our life course study of MHOE, the number of eligible individuals reduces from 14268 to 9622 once we take into account the missing data in the covariates representing a 33% reduction in L2 units. When addressing our missing data situation we do not include any auxiliary or instrumental variables as there are no NCDS variables that are fully (or almost fully) observed containing information on our respondents and non-respondents. In addition, we found that there is evidence of MAR based on the covariates in our MoI. (Van Buuren, Boshuizen, & Knook, 1999) argue that auxiliary variables should be fully observed and (Enders, 2010) recommends that the correlation between auxiliary

information and any missingness be at least 0.4. By way of example, we looked at the possibility of including father's social class in 1958 as an auxiliary variable in the imputation model but it was not fully observed in relation to our MoI and proved to have a weaker association with the missingness compared to our main predictor of interest, PSES16.

MI as a possible solution to missing data

Another consideration regarding missing data is the pattern of non-response. All CMs who first contribute data to the NCDS survey and then become permanent non-responders during the life course can be said to follow a monotonic pattern of non-response. The alternative non-response pattern is referred to as non-monotonic whereby NCDS CMs have intermittent survey responses throughout the life course. Appreciating the difference between these two non-response patterns is important in terms of selecting an appropriate missing data solution, especially when dealing with repeatedly measured life course data.

In terms of monotonic MHOE response patterns whereby CMs either never leave the study or become permanent non-responders, approximately 52% of CMs follow a monotonic pattern. The remaining 48% follow a non-monotonic pattern whereby CMs have intermittent life course MHOE responses. In total, there exist 63 different MHOE life course response patterns consisting of zeros (not missing MHOE) and ones (missing MHOE). The concept of monotone missingness is important because it can simplify the application of missing data solutions. However, the development of joint modelling and chained equation approaches to MI has reduced the computational problems associated with a lack of monotonicity within the missing data pattern and we focus on these two MI approaches. Other solutions such as inverse probability weighting and selection/pattern mixture modelling are considerably more complicated when addressing intermittent, non-monotone missingness (Bartlett & Carpenter, 2013) and are not considered in this paper. (Reinecke, 2013) demonstrates the use of the latter type of statistical models when modelling panel dropouts with a much smaller number of response patterns.

MI replaces each missing value with multiple imputed values, which are random draws from the distribution of an imputation regression model that conditions on the observed data (Rubin, 1987). The end result is multiple complete datasets. We then fit the MoI to each

imputed dataset, and the parameter estimates and standard errors from these models are combined using Rubin's rules (Rubin, 1987). This takes into account the uncertainty of the estimates due to the missing data. Inferences from imputed data are valid provided the imputation model is correctly specified and data are MAR. (Goldstein, Carpenter, & Browne, 2014) highlight that there are two approaches to MI; one that uses the joint posterior distribution of all variables when sampling for missing values and the chained equation approach that uses the conditional distribution for each variable in turn. A distinct advantage of the former is the implementation of multilevel data structures and interactions in the imputation process (Goldstein, Carpenter, Kenward, & Levin, 2009). Our MoI contains both a multilevel structure and interactions therefore the joint modelling MI approach should be more suitable for our situation. Moreover, (Schafer, 1997) argues that imputations drawn from correctly specified models will result in estimates that are unbiased and are efficient in the sense that optimal use of the observed information is used.

In this paper, we investigate the empirical difference between the two MI approaches and compare results with the FIML approach as shown in previous sections. It should be noted that the two approaches will produce different imputation models. The Multilevel MI (MLMI) approach uses a joint two-level model to impute missing values whereas the MI Chained Equations (MICE) approach uses single-level regression models. We refer the reader to Appendix B in the supplemental material to appreciate the difference between the two types of imputation models where all model equations are presented. Furthermore, the two approaches require different dataset formations before beginning each process; MLMI requires the data in long format whereby a record corresponds to an occasion nested within an individual and MICE requires the data in wide format whereby a record corresponds to an individual. Despite the differences between the two approaches, we attempt to design imputation models that are faithful to the MoI as advocated by (Goldstein, 2009). Therefore, any significant differences found between either approach and the FIML results could indicate a poor choice of imputation model.

MLMI for missing data in a longitudinal setting

(J. R. Carpenter, Goldstein, & Kenward, 2011) developed the Realcom-Impute software which performs MLMI and can handle ordinal and unordered categorical data appropriately.

With this approach, a MLM is specified for the partially observed variables given the fully observed variables. This MLM is fitted to the observed data and multiple imputations of the missing data are generated using Markov Chain Monte Carlo (MCMC) methods. This approach models the joint distribution of the incomplete variables conditional on the complete variables using a multivariate latent normal model which allows for the proper handling of a two-level structure whereby L2 variables are constant over the observations at L1 and random coefficients can be applied according to the MoI. This approach can also handle cross-level interactions such as the interaction terms we have in our MoI; namely the (age * PSES16) and (age * Female) interactions. Therefore, we adopted the "Just Another Variable" (JAV) approach as advocated by (Seaman, Bartlett, & White, 2012) when employing linear regression. This means the interaction terms were produced prior to the imputation phase and entered into the imputation model as response variables in the case of the (age * PSES16) interactions given the missingness in PSES16 and predictor variables in the case of the (age * Female) interactions as both terms are fully observed. However, the resulting imputations for the (age * PSES16) interactions require correcting before the pooling stage as the structure of the (age * PSES16) interactions should be the same as before imputation, i.e. the same term on the diagonal and zeros elsewhere. Therefore, the imputed PSES16 terms, imputed in conjunction with the cross-level interactions, are interacted with the age dummies and these revised cross-level interactions are then used in the pooling stage of the process.

MICE for missing data in a longitudinal setting

A popular MI approach involves employing chained equations, also referred to as the fully conditional specification (FCS) algorithm, which specifies separate univariate imputation models for each variable with missing data conditional on all other variables (Van Buuren et al., 1999). Therefore, we can choose a model appropriate to the variable type, e.g. continuous, count, ordered categorical, unordered categorical, and characteristics measured at fixed times over the life course are treated as distinct variables. (Welch, Bartlett, & Petersen, 2014) report that this method is easier computationally than directly specifying a multivariate distribution for a mixture of continuous and categorical variables with missing data, as required in parametric MI's original form. In the same paper, they also advocate a two-fold FCS algorithm for applying MI to longitudinal data. This modified version of the

original FCS algorithm imputes missing values at a given time point from a model that only uses information from that time point and immediately adjacent time points. However, this simplification may induce bias in parameter estimates if the measurements excluded from imputation models have independent effects, i.e. there exists dependency over the life course. It is for this reason we opt for the FCS algorithm instead of the two-fold version. Furthermore, it is not possible to compute cross-level interactions prior to commencing the imputation phase when employing the MICE approach as there is no formal way to simultaneously incorporate both L1 and L2 variables with the data in wide format. Therefore, the cross-level interaction terms could only be computed after completing the imputation phase when the data was converted into long format in order to run our MLMs of interest. Moreover, no random coefficients can be utilised during the imputation process so our MICE imputation model is uncongenial with respect to our MoI leading to potential biases (Goldstein et al., 2014).

Relevant tools and computational time

All MIs were conducted on a Windows 7 64-bit operating system with 8GB of RAM. For MLMI, we used Realcom-Impute with a Windows 32-bit MATLAB runtime installer processed via Stata/SE 13.1. For MICE, we used Stata/SE 13.1. All imputed MLMs of interest were conducted using MLwiN v2.32 and Stata/SE 13.1 via the "runmlwin" command (Leckie & Charlton, 2013) and Rubin's combination rules (Rubin, 1987), as set out by (Dong & Peng, 2013), were implemented in Microsoft Excel 2010. This is because Stata/SE 13.1 could not compute our life course step function MLM of interest with occasion-level residuals set to zero so we could not use Stata's built-in "mi estimate" functionality. Before applying Rubin's combination rules, all imputed values belonging to NCDS survey members who were recorded as dead or emigrant during a given survey sweep were set to missing again as they are not part of the target sample and only their productive responses in previous sweeps were preserved for analysis purposes. This approach has been advocated by (Biering, Hjollund, & Frydenberg, 2015).

When conducting MLMI, we followed the advice of (J. R. Carpenter et al., 2011) and for MICE we adhered to the procedure as described by (Lloyd, Obradović, Carpiano, & Motti-Stefanidi, 2013). In terms of time taken to complete the imputation phases; 10 imputed

datasets using the MLMI method for 8 response variables with missingness conditional on 12 fully observed variables and the random coefficients on the age dummies with 2000 iterations burn-in phase followed by 500 iterations between each imputed dataset (7000 iterations in total) took approximately 6 days. Whereas completing 10 imputed datasets using the MICE procedure for 13 incomplete response variables plus 1 fully observed with 7000 iterations took less than 24 hours. We refer the reader to Appendix B in the accompanying supplemental material for details on the variation in imputation model setup. The level of non-monotonic missingness in the life course MHOE variable coupled with the amount of missing data contained in the model covariates and the complexity of our life course statistical model all contributed to the computational intensity and time taken to complete the MI process.

Results

In this section, we present a summary of results using MHOE MLM life course fixed effects. We refer the reader to Appendix C in the supplemental material for the MLM life course random effects results. Table 5 displays the comparison of fixed effect model results for AC, CC, PO, MLMI and MICE. Accompanying standard errors are in parentheses. The AC results are as in Table 2. With dead and emigrant cohort members removed before combining imputed datasets (i.e. 3% of L1 units omitted), both MLMI and MICE results produce more similar coefficients and inferences to the AC results compared with the CC and PO results. CC cohort members attained the highest average MHOE over the life course while the MLMI approach produced the lowest average parental SES at 16 legacy effect at age 23. Furthermore, the CC analysis underestimated the interaction effects of parental SES at age 16 on MHOE later on in the life course, with some statistically insignificant results, in comparison to the other methods of handling missing data. This result indicates that there is a differential continuing effect of parental SES at age 16 on MHOE over the life course depending on the missing data method employed.

In a sensitivity analysis not presented here regarding the MLMI JAV (age * PSES16) interactions, we examined the results without applying the correction before the pooling stage and obtained parental SES at age 16 life course effects which were negative and insignificant. It is also important to note that only cohort members who had at least one

occasion with occupational data throughout the life course (i.e. partially observed) were eligible for imputation. Those with no occupational data from age 23 to 55 were excluded from this study.

Table 5. Life course Mean Hourly Occupational Earnings multilevel model fixed effects comparisons of coefficients (standard errors)

	AC	CC	PO	MLMI	MICE
Age	**Coeff. (se)**	**Coeff. (se)**	**Coeff. (se)**	**Coeff. (se)**	**Coeff. (se)**
23	2.016 (0.006)	2.050 (0.011)	2.002 (0.007)	2.006 (0.005)	2.013 (0.005)
33	2.161 (0.007)	2.195 (0.012)	2.151 (0.008)	2.159 (0.006)	2.163 (0.006)
42	2.279 (0.007)	2.310 (0.013)	2.269 (0.008)	2.273 (0.007)	2.277 (0.006)
46	2.380 (0.007)	2.414 (0.012)	2.368 (0.009)	2.379 (0.008)	2.380 (0.006)
50	2.421 (0.007)	2.458 (0.012)	2.406 (0.009)	2.419 (0.006)	2.421 (0.006)
55	2.338 (0.007)	2.368 (0.012)	2.327 (0.009)	2.339 (0.007)	2.339 (0.007)
Parental SES at 16	0.091 (0.004)	0.085 (0.007)	0.089 (0.004)	0.077 (0.004)	0.091 (0.004)
Female	-0.148	-0.145	-0.148	-0.147	-0.147
	(0.007)	(0.013)	(0.008)	(0.020)	(0.005)
Region of residence (Ref: North & Midlands)					
South & East	0.025 (0.006)	0.010 (0.010)	0.030 (0.007)	0.030 (0.005)	0.018 (0.009)
Wales	-0.038	-0.038	-0.035	-0.003	-0.018
	(0.012)	(0.023)	(0.014)	(0.010)	(0.017)
Scotland	0.011 (0.009)	0.014 (0.018)	0.012 (0.011)	0.009 (0.009)	0.001 (0.012)
Interaction Effects					
Age*Parental SES at 16 (Ref: Age 23*Parental SES)					
33 * PSES16	0.043 (0.005)	0.024 (0.008)	0.052 (0.006)	0.040 (0.005)	0.043 (0.005)
42 * PSES16	0.038 (0.005)	0.005 (0.009)	0.054 (0.006)	0.035 (0.005)	0.039 (0.005)
46 * PSES16	0.034 (0.005)	0.016 (0.009)	0.043 (0.006)	0.031 (0.005)	0.033 (0.005)
50 * PSES16	0.024 (0.005)	0.006 (0.009)	0.033 (0.006)	0.022 (0.006)	0.022 (0.006)
55 * PSES16	0.022 (0.005)	0.008 (0.009)	0.030 (0.007)	0.023 (0.008)	0.023 (0.005)
Age*Female (Ref: Age 23*Female)					
33 * Female	-0.075	-0.075	-0.075	-0.076	-0.078
	(0.009)	(0.016)	(0.011)	(0.012)	(0.007)
42 * Female	-0.083	-0.075	-0.087	-0.076	-0.077
	(0.010)	(0.017)	(0.012)	(0.013)	(0.008)
46 * Female	-0.032	-0.021	-0.040	-0.031	-0.030
	(0.010)	(0.017)	(0.012)	(0.016)	(0.008)
50 * Female	-0.029	-0.030	-0.027	-0.026	-0.024
	(0.010)	(0.017)	(0.013)	(0.015)	(0.008)
55 * Female	-0.022	-0.005	-0.038	-0.017	-0.013
	(0.011)	(0.017)	(0.014)	(0.016)	(0.010)
Occasions	37478	13561	23917	82654	82654
Individuals	9622	2263	7359	14268	14268
Minimum	1	5	1	1	1
Average	3.9	6	3.3	5.8	5.8
Maximum	6	6	5	6	6
Deviance	14789	4689	9972	42739	28046

Discussion

This study investigated the heterogeneity between individuals in relation to a finely-grained measure of SES over the life course such as MHOE. We found employing a step function multilevel life course model most suitable for exploring this MHOE heterogeneity given the amount of variation between individuals over the life course. This step function multilevel life course model enabled us to measure the legacy effect of parental SES at age 16 on achieved MHOE in adulthood by introducing cross-level interaction dummy variables between parental SES and age over the life course. We established the significant and positive life course contribution of parental SES at age 16 which helped explain MHOE heterogeneity between NCDS cohort members. We then explored the issue of missing data in relation to our MoI by investigating the empirical difference between two MI methods and compared results with the FIML (AC), CC and PO approaches. First, we provided evidence strengthening our claim that the missingness associated with our MoI is MAR and showed how MI could be a viable solution to correct for the missing data problem that exists in this empirical study. Second, we discovered the CC analysis underestimated the effect of parental SES on MHOE later on in the life course in comparison with the other methods of handling missing data. This result was uncovered by our step function multilevel life course model which proved to be more effective at appreciating the MHOE variation between individuals compared to more parsimonious life course models.

Our life course statistical model enabled a detailed examination of between-individual variation across the life course in relation to our chosen SES measurement. This person-level variance analysis helped identify the general divergent pattern between individuals whereby larger deviations from the MHOE average at a preceding age are positively correlated with larger deviations from the MHOE average at a subsequent age. This evidence suggests positive momentum is greatest with those survey CMs who are already more advantaged compared to those who are already less advantaged. The model covariates provided explanatory power and showed significant differential effects in relation to life course MHOE depending on a survey CM's family's socioeconomic background, sex and region of residence.

We contrasted the faster MICE approach with the approach that provided a closer fit between the imputation model and analytical model, MLMI, given the structure of our data and MoI. The MICE method takes less time computationally and was found by (Romaniuk, Patton, & Carlin, 2014) to produce more plausible results than the multivariate normal imputation procedure. We found that the MICE method produces similarly plausible empirical results compared with the computationally slower MLMI method. Moreover, both sets of results were more similar to AC results compared with CC and PO results on average. We refer the reader to Appendix D in the supplemental material for the subject specific life course MHOE MLM predictions which documents these results. Our CC results represented the situation whereby missing data only affects the MoI covariates but these results are biased upwards in terms of life course MHOE. Therefore, the AC, MLMI and MICE results are more representative of the NCDS sample and the target population thereby confirming previous research findings (White & Carlin, 2010).

The evidence presented in this study suggests the FIML approach using all available cases and our step function multilevel life course model is sufficient with the MI process amounting to a robustness check for the model results and inferences similar to (Maharani & Tampubolon, 2014). However, it should be noted that the main advantage to using MI is the potential to have a more complex imputation model incorporating (auxiliary) variables which are not included in the MoI but nonetheless predict the model variables and non-response. The inclusion of such auxiliary information would make the MAR assumption more plausible and has been discussed in detail by (Collins, Schafer, & Kam, 2001). Given our use of the NCDS data, we have argued in this paper that our MoI covariates provide sufficient explanatory power in relation to the MAR assumption so the exclusion of auxiliary variables is not likely to have biased our study findings.

It is interesting to note that despite the difference in setup between the MLMI and MICE imputation models with MICE using single-level regression models and MLMI using a joint two-level model to impute missing values, the difference in model congeniality with respect to the MoI did not produce significantly different results compared to the FIML AC results. This finding is interesting for two reasons; either imputation model could be considered appropriate despite the difference in MoI congeniality plus post-imputation (age * PSES16)

interaction processing and neither imputation model enhanced the substantive inference derived from the FIML AC approach given the MoI and underlying data. Despite this finding, (Goldstein et al., 2014) have addressed the well-known biases that can arise from omitting MoI interaction terms from the imputation model as we have done with respect to our MICE imputation model which cannot accommodate cross-level interaction terms. Such terms are computationally demanding to impute using Realcom-Impute. Therefore, it remains to be seen if the more recently developed Stat-JR software (Charlton et al., 2012) can produce MLMI datasets at a similar pace to MICE given our MoI and underlying data. Recent work by (Goldstein et al., 2014) shows promise.

It also remains as further work to see if increasing the number of imputations as advocated by (Dong & Peng, 2013; Spratt et al., 2010), and hence the time taken to complete the process, changes the conclusions we have reached. One more extension arising from this empirical study would be to develop an MNAR statistical model process that can accommodate our life course model of interest so that a proper sensitivity analysis could be conducted contrasting the MAR and MNAR missingness mechanisms.

Acknowledgements

Adrian Byrne is in receipt of a School of Social Sciences PhD Studentship at the University of Manchester. NCDS data were accessed via the UK Data Service plus Annual Survey of Hours and Earnings data and Consumer Price Index data were accessed via the Office for National Statistics. The authors have no conflicts of interest to declare.

References

American Psychological Association. (2007). Task Force on Socioeconomic Status *Report of the APA task force on socioeconomic status. Washington, DC: American Psychological Association*.

Bartlett, J., & Carpenter, J. (2013). Missing Data Concepts. LEMMA VLE Module 14, 1-29. (http://www.bristol.ac.uk/cmm/learning/course.html).

Biering, K., Hjollund, N. H., & Frydenberg, M. (2015). Using multiple imputation to deal with missing data and attrition in longitudinal studies with repeated measures of patient-reported outcomes. *Clinical epidemiology, 7*, 91.

Blanden, J., Goodman, A., Gregg, P., & Machin, S. (2004). Changes in intergenerational mobility in Britain. *Generational income mobility in North America and Europe*, 122-146.

Bollen, K. A., & Bauldry, S. (2011). Three Cs in measurement models: causal indicators, composite indicators, and covariates. *Psychological methods, 16*(3), 265.

Bourdieu, P. (1986). The forms of capital Handbook of theory and research for the sociology of education (pp. 241–258): New York: Greenwood.

Bukodi, E., & Dex, S. (2009). Bad start: Is there a way up? Gender differences in the effect of initial occupation on early career mobility in Britain. *European Sociological Review*, jcp030.

Bukodi, E., Dex, S., & Goldthorpe, J. H. (2011). The conceptualisation and measurement of occupational hierarchies: a review, a proposal and some illustrative analyses. *Quality & Quantity, 45*(3), 623-639.

Caro, D. H., & Cortés, D. (2012). Measuring family socioeconomic status: An illustration using data from PIRLS 2006. *IERI Monograph Series. Issues and Methodologies in Large-Scale Assessments, 5*, 9-33.

Carpenter, J., & Plewis, I. (2011). Analysing longitudinal studies with non-response: issues and statistical methods. In M. Williams & P. Vogt (Eds.), *The SAGE handbook of Innovation in Social Research Methods* (pp. 498-523). New York: Sage.

Carpenter, J. R., Goldstein, H., & Kenward, M. G. (2011). REALCOM-IMPUTE software for multilevel multiple imputation with mixed response types. *Journal of Statistical Software, 45*(5), 1-14.

Charlton, C., Michaelides, D., Cameron, B., Szmaragd, C., Parker, R., Yang, H., . . . Browne, W. J. (2012). Stat-JR software.

Collins, L. M., Schafer, J. L., & Kam, C.-M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological methods, 6*(4), 330.

Dong, Y., & Peng, C.-Y. J. (2013). Principled missing data methods for researchers. *SpringerPlus, 2*(1), 1-17.

Enders, C. K. (2010). *Applied missing data analysis*: Guilford Press.

Erola, J., Jalonen, S., & Lehti, H. (2016). Parental education, class and income over early life course and children's achievement. *Research in Social Stratification and Mobility, 44*, 33-43.

Galobardes, B., Shaw, M., Lawlor, D. A., & Lynch, J. W. (2006). Indicators of socioeconomic position (part 2). *J Epidemiol Community Health, 60*(2), 95.

Galobardes, B., Shaw, M., Lawlor, D. A., Lynch, J. W., & Smith, G. D. (2006). Indicators of socioeconomic position (part 1). *J Epidemiol Community Health, 60*(1), 7-12.

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

Goldstein, H., Carpenter, J., Kenward, M. G., & Levin, K. A. (2009). Multilevel models with multivariate mixed response types. *Statistical Modelling, 9*(3), 173-197.

Goldstein, H., Carpenter, J. R., & Browne, W. J. (2014). Fitting multilevel multivariate models with missing data in responses and covariates that may include interactions and non‐linear terms. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 177*(2), 553-564.

Goldthorpe, J. H., & Jackson, M. (2007). Intergenerational class mobility in contemporary Britain: political concerns and empirical findings1. *Br J Sociol, 58*(4), 525-546.

Hawkes, D., & Plewis, I. (2006). Modelling non‐response in the national child development study. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 169*(3), 479-491.

Kolenikov, S., & Angeles, G. (2009). Socioeconomic status measurement with discrete proxy variables: Is principal component analysis a reliable answer? *Review of Income and Wealth, 55*(1), 128-165.

Kumar, S., Kroon, J., & Lalloo, R. (2014). A systematic review of the impact of parental socio-economic status and home environment characteristics on children's oral health related quality of life. *Health and quality of life outcomes, 12*(1), 1.

Laurison, D., & Friedman, S. (2016). The class pay gap in higher professional and managerial occupations. *American Sociological Review, 81*(4), 668-695.

Leckie, G., & Charlton, C. (2013). Runmlwin-a program to Run the MLwiN multilevel modelling software from within stata. *Journal of Statistical Software, 52*(11), 1-40.

Li, Y., & Devine, F. (2011). Is social mobility really declining? Intergenerational class mobility in Britain in the 1990s and the 2000s. *Sociological Research Online, 16*(3), 4.

Lloyd, J. E., Obradović, J., Carpiano, R. M., & Motti-Stefanidi, F. (2013). JMASM 32: Multiple Imputation of Missing Multilevel, Longitudinal Data: A Case When Practical Considerations Trump Best Practices? *Journal of Modern Applied Statistical Methods, 12*(1), 29.

Maharani, A., & Tampubolon, G. (2014). Unmet needs for cardiovascular care in Indonesia. *PloS one, 9*(8), e105831.

Nickell, S. (1982). The determinants of occupational success in Britain. *The Review of Economic Studies, 49*(1), 43-53.

Niedzwiedz, C. L., Katikireddi, S. V., Pell, J. P., & Mitchell, R. (2012). Life course socio-economic position and quality of life in adulthood: a systematic review of life course models. *BMC public health, 12*(1), 1.

Pollitt, R., Rose, K., & Kaufman, J. (2005). Evaluating the evidence for models of life course socioeconomic factors and cardiovascular outcomes: a systematic review. *BMC public health, 5*(1), 7.

Rasbash, J., Steele, F., Browne, W. J., Goldstein, H., & Charlton, C. (2015). A user's guide to MLwiN. *Centre for Multilevel Modelling, University of Bristol, UK*.

Reinecke, J. (2013). Growth curve models and panel dropouts: Applications with criminological panel data. *Netherlands Journal of Psychology, 67*(4).

Romaniuk, H., Patton, G. C., & Carlin, J. B. (2014). Multiple imputation in a longitudinal cohort study: a case study of sensitivity to imputation methods. *Am J Epidemiol*, kwu224.

Rubin, D. B. (1987). Multiple Imputation for Nonresponse in Surveys (Wiley Series in Probability and Statistics).

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*: CRC press.

Schuller, T., Wadsworth, M., Bynner, J., & Goldstein, H. (2011). The Measurement of Well-being: the Contribution of Longitudinal Studies.

Seaman, S. R., Bartlett, J. W., & White, I. R. (2012). Multiple imputation of missing covariates with non-linear effects and interactions: an evaluation of statistical methods. *BMC medical research methodology, 12*(1), 46.

Spratt, M., Carpenter, J., Sterne, J. A., Carlin, J. B., Heron, J., Henderson, J., & Tilling, K. (2010). Strategies for multiple imputation in longitudinal studies. *Am J Epidemiol, 172*(4), 478-487.

Steele, F. (2008). Multilevel models for longitudinal data. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 171*(1), 5-19.

Steele, F. (2014). Multilevel Modelling of Repeated Measures Data. LEMMA VLE Module 15, 1-62. (http://www.bristol.ac.uk/cmm/learning/course.html).

Sturgis, P., & Sullivan, L. (2008). Exploring social mobility with latent trajectory groups. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 171*(1), 65-88.

University of London. Institute of Education. Centre for Longitudinal Studies. (2008a). *National Child Development Study: Sweep 4, 1981, and Public Examination Results, 1978*. Retrieved from: http://dx.doi.org/10.5255/UKDA-SN-5566-1

University of London. Institute of Education. Centre for Longitudinal Studies. (2008b). *National Child Development Study: Sweep 5, 1991*. Retrieved from: http://dx.doi.org/10.5255/UKDA-SN-5567-1

University of London. Institute of Education. Centre for Longitudinal Studies. (2008c). *National Child Development Study: Sweep 6, 1999-2000*. Retrieved from: http://dx.doi.org/10.5255/UKDA-SN-5578-1

University of London. Institute of Education. Centre for Longitudinal Studies. (2008d). *National Child Development Study: Sweep 7, 2004-2005*. Retrieved from: http://dx.doi.org/10.5255/UKDA-SN-5579-1

University of London. Institute of Education. Centre for Longitudinal Studies. (2012). *National Child Development Study: Sweep 8, 2008-2009*. Retrieved from: http://dx.doi.org/10.5255/UKDA-SN-6137-2

University of London. Institute of Education. Centre for Longitudinal Studies. (2014). *National Child Development Study: Childhood Data, Sweeps 0-3, 1958-1974*. Retrieved from: http://dx.doi.org/10.5255/UKDA-SN-5565-2

University of London. Institute of Education. Centre for Longitudinal Studies. (2015). *National Child Development Study: Sweep 9, 2013*. Retrieved from: http://dx.doi.org/10.5255/UKDA-SN-7669-1

Van Buuren, S., Boshuizen, H. C., & Knook, D. L. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in medicine, 18*(6), 681-694.

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.

White, I. R., & Carlin, J. B. (2010). Bias and efficiency of multiple imputation compared with complete‐case analysis for missing covariate values. *Statistics in medicine, 29*(28), 2920-2931.

White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: issues and guidance for practice. *Statistics in medicine, 30*(4), 377-399.

Multilevel modelling approach to analysing life course socioeconomic status and compensating for missingness_supplemental material

Appendix A

Table 1A. Ordinal Principal Component Analysis table of components*

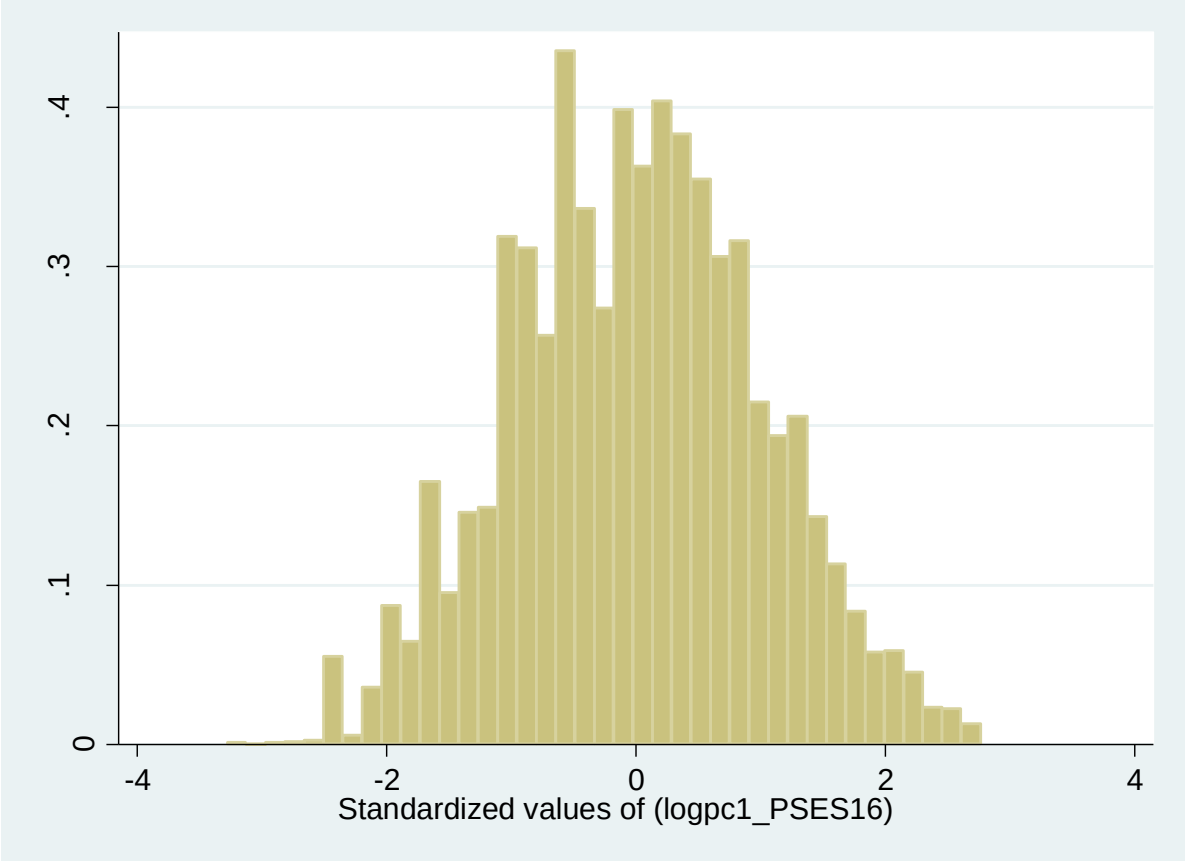
Component	Eigenvalue	Difference	Proportion	Cumulative
Component 1	2.25	1.31	0.45	0.45
Component 2	0.94	0.16	0.19	0.64
Component 3	0.78	0.16	0.16	0.79
Component 4	0.62	0.22	0.12	0.92
Component 5	0.41	.	0.08	1

* the reported sampling adequacy statistics Kaiser-Meyer-Olkin (KMO) and Cronbach's Alpha were 0.71 and 0.6 respectively

Table 2A. Ordinal Principal Component Analysis loadings

PSES16 loadings	Comp1	Comp2	Comp3	Comp4	Comp5
Father's social class	0.48	-0.18	0.23	0.80	-0.21
Mother's social class	0.24	0.95	0.08	0.09	0.15
Father's education	0.53	-0.21	-0.35	-0.07	0.74
Mother's education	0.51	0.04	-0.51	-0.31	-0.62
Housing tenure type	0.42	-0.11	0.75	-0.50	-0.03

Figure 1A. Parental SES at age 16 (PSES16)



Appendix B

Equations 1B. MLMI – joint/multivariate modelling equations

$$
\begin{aligned}
&log\ Mean\ Hourly\ Occupational\ Earnings_{tj} \\
&\quad = \beta_{11}\text{age23}_{tj} + \beta_{21}\text{age33}_{tj} + \beta_{31}\text{age42}_{tj} + \beta_{41}\text{age46}_{tj} + \beta_{51}\text{age50}_{tj} \\
&\quad + \beta_{61}\text{age55}_{tj} + \beta_{71}\text{Female}_{j} + \beta_{81}\text{age33}_{tj} * \text{Female}_{j} \\
&\quad + \beta_{91}\text{age42}_{tj} * \text{Female}_{j} + \beta_{101}\text{age46}_{tj} * \text{Female}_{j} \\
&\quad + \beta_{111}\text{age50}_{tj} * \text{Female}_{j} + \beta_{121}\text{age55}_{tj} * \text{Female}_{j} + u_{11j}\text{age23}_{tj} \\
&\quad + u_{21j}\text{age33}_{tj} + u_{31j}\text{age42}_{tj} + u_{41j}\text{age46}_{tj} + u_{51j}\text{age50}_{tj} \\
&\quad + u_{61j}\text{age55}_{tj}
\end{aligned}
$$

$$
\begin{aligned}
&Region\ of\ Residence_{tj} \\
&\quad = \beta_{12}\text{age23}_{tj} + \beta_{22}\text{age33}_{tj} + \beta_{32}\text{age42}_{tj} + \beta_{42}\text{age46}_{tj} + \beta_{52}\text{age50}_{tj} \\
&\quad + \beta_{62}\text{age55}_{tj} + \beta_{72}\text{Female}_{j} + \beta_{82}\text{age33}_{tj} * \text{Female}_{j} \\
&\quad + \beta_{92}\text{age42}_{tj} * \text{Female}_{j} + \beta_{102}\text{age46}_{tj} * \text{Female}_{j} \\
&\quad + \beta_{112}\text{age50}_{tj} * \text{Female}_{j} + \beta_{122}\text{age55}_{tj} * \text{Female}_{j} + u_{12j}\text{age23}_{tj} \\
&\quad + u_{22j}\text{age33}_{tj} + u_{32j}\text{age42}_{tj} + u_{42j}\text{age46}_{tj} + u_{52j}\text{age50}_{tj} \\
&\quad + u_{62j}\text{age55}_{tj}
\end{aligned}
$$

$$
\begin{aligned}
age33_{tj} * PSES16_{j} &= \beta_{13}\text{age23}_{tj} + \beta_{23}\text{age33}_{tj} + \beta_{33}\text{age42}_{tj} + \beta_{43}\text{age46}_{tj} + \beta_{53}\text{age50}_{tj} \\
&+ \beta_{63}\text{age55}_{tj} + \beta_{73}\text{Female}_{j} + \beta_{83}\text{age33}_{tj} * \text{Female}_{j} \\
&+ \beta_{93}\text{age42}_{tj} * \text{Female}_{j} + \beta_{103}\text{age46}_{tj} * \text{Female}_{j} \\
&+ \beta_{113}\text{age50}_{tj} * \text{Female}_{j} + \beta_{123}\text{age55}_{tj} * \text{Female}_{j} + u_{13j}\text{age23}_{tj} \\
&+ u_{23j}\text{age33}_{tj} + u_{33j}\text{age42}_{tj} + u_{43j}\text{age46}_{tj} + u_{53j}\text{age50}_{tj} \\
&+ u_{63j}\text{age55}_{tj}
\end{aligned}
$$

$$
\begin{aligned}
age42_{tj} * PSES16_{j} &= \beta_{14}\text{age23}_{tj} + \beta_{24}\text{age33}_{tj} + \beta_{34}\text{age42}_{tj} + \beta_{44}\text{age46}_{tj} + \beta_{54}\text{age50}_{tj} \\
&+ \beta_{64}\text{age55}_{tj} + \beta_{74}\text{Female}_{j} + \beta_{84}\text{age33}_{tj} * \text{Female}_{j} \\
&+ \beta_{94}\text{age42}_{tj} * \text{Female}_{j} + \beta_{104}\text{age46}_{tj} * \text{Female}_{j} \\
&+ \beta_{114}\text{age50}_{tj} * \text{Female}_{j} + \beta_{124}\text{age55}_{tj} * \text{Female}_{j} + u_{14j}\text{age23}_{tj} \\
&+ u_{24j}\text{age33}_{tj} + u_{34j}\text{age42}_{tj} + u_{44j}\text{age46}_{tj} + u_{54j}\text{age50}_{tj} \\
&+ u_{64j}\text{age55}_{tj}
\end{aligned}
$$

$$
\begin{aligned}
age46_{tj} * PSES16_{j} &= \beta_{15}\text{age23}_{tj} + \beta_{25}\text{age33}_{tj} + \beta_{35}\text{age42}_{tj} + \beta_{45}\text{age46}_{tj} + \beta_{55}\text{age50}_{tj} \\
&+ \beta_{65}\text{age55}_{tj} + \beta_{75}\text{Female}_{j} + \beta_{85}\text{age33}_{tj} * \text{Female}_{j} \\
&+ \beta_{95}\text{age42}_{tj} * \text{Female}_{j} + \beta_{105}\text{age46}_{tj} * \text{Female}_{j} \\
&+ \beta_{115}\text{age50}_{tj} * \text{Female}_{j} + \beta_{125}\text{age55}_{tj} * \text{Female}_{j} + u_{15j}\text{age23}_{tj} \\
&+ u_{25j}\text{age33}_{tj} + u_{35j}\text{age42}_{tj} + u_{45j}\text{age46}_{tj} + u_{55j}\text{age50}_{tj} \\
&+ u_{65j}\text{age55}_{tj}
\end{aligned}
$$

$$
\begin{aligned}
age50_{tj} * PSES16_{j} &= \beta_{16}\text{age23}_{tj} + \beta_{26}\text{age33}_{tj} + \beta_{36}\text{age42}_{tj} + \beta_{46}\text{age46}_{tj} + \beta_{56}\text{age50}_{tj} \\
&+ \beta_{66}\text{age55}_{tj} + \beta_{76}\text{Female}_{j} + \beta_{86}\text{age33}_{tj} * \text{Female}_{j} \\
&+ \beta_{96}\text{age42}_{tj} * \text{Female}_{j} + \beta_{106}\text{age46}_{tj} * \text{Female}_{j} \\
&+ \beta_{116}\text{age50}_{tj} * \text{Female}_{j} + \beta_{126}\text{age55}_{tj} * \text{Female}_{j} + u_{16j}\text{age23}_{tj} \\
&+ u_{26j}\text{age33}_{tj} + u_{36j}\text{age42}_{tj} + u_{46j}\text{age46}_{tj} + u_{56j}\text{age50}_{tj} \\
&+ u_{66j}\text{age55}_{tj}
\end{aligned}
$$

$$
\begin{aligned}
age55_{tj} * PSES16_{j} &= \beta_{17}\text{age23}_{tj} + \beta_{27}\text{age33}_{tj} + \beta_{37}\text{age42}_{tj} + \beta_{47}\text{age46}_{tj} + \beta_{57}\text{age50}_{tj} \\
&+ \beta_{67}\text{age55}_{tj} + \beta_{77}\text{Female}_{j} + \beta_{87}\text{age33}_{tj} * \text{Female}_{j} \\
&+ \beta_{97}\text{age42}_{tj} * \text{Female}_{j} + \beta_{107}\text{age46}_{tj} * \text{Female}_{j} \\
&+ \beta_{117}\text{age50}_{tj} * \text{Female}_{j} + \beta_{127}\text{age55}_{tj} * \text{Female}_{j} + u_{17j}\text{age23}_{tj} \\
&+ u_{27j}\text{age33}_{tj} + u_{37j}\text{age42}_{tj} + u_{47j}\text{age46}_{tj} + u_{57j}\text{age50}_{tj} \\
&+ u_{67j}\text{age55}_{tj}
\end{aligned}
$$

$$
\begin{aligned}
PSES16_{j} &= \beta_{78}\text{Female}_{j} + u_{18j}\text{age23}_{tj} + u_{28j}\text{age33}_{tj} + u_{38j}\text{age42}_{tj} + u_{48j}\text{age46}_{tj} \\
&+ u_{58j}\text{age50}_{tj} + u_{68j}\text{age55}_{tj}
\end{aligned}
$$

$$\mathbf{u} \sim N(\mathbf{0},\ \boldsymbol{\Omega_u})$$

$$\boldsymbol{\Omega_u} = \mathbb{V}\begin{bmatrix} \boldsymbol{u_{11j}} \\ \vdots \\ \boldsymbol{u_{68j}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\sigma^2_{u_{11}}} & \cdots & \boldsymbol{\sigma_{u_{11}u_{68}}} \\ \vdots & \ddots & \vdots \\ \boldsymbol{\sigma_{u_{11}u_{68}}} & \cdots & \boldsymbol{\sigma^2_{u_{68}}} \end{bmatrix}$$

$$\text{t} = 23, 33, 42, 46, 50, 55;\ \text{j} = 1, \ldots, 14268$$

Equations 2B. MICE – chained equations (RoR = Region of residence variables omitted to avoid perfect prediction problem within multinomial logit RoR models)

$$log\ MHOE23_j = \beta_0 + \beta_1 MHOE33_j + \beta_2 MHOE42_j + \beta_3 MHOE46_j + \beta_4 MHOE50_j + \beta_5 MHOE55_j + \beta_6 PSES16_j + \beta_7 Female_j + \beta_8 South\ \&\ East23_j + \beta_9 Wales23_j + \beta_{10} Scotland23_j + \beta_{11} South\ \&\ East33_j + \beta_{12} Wales33_j + \beta_{13} Scotland33_j + \beta_{14} South\ \&\ East42_j + \beta_{15} Wales42_j + \beta_{16} Scotland42_j + \beta_{17} South\ \&\ East46_j + \beta_{18} Wales46_j + \beta_{19} Scotland46_j + \beta_{20} South\ \&\ East50_j + \beta_{21} Wales50_j + \beta_{22} Scotland50_j + \beta_{23} South\ \&\ East55_j + \beta_{24} Wales55_j + \beta_{25} Scotland55_j$$

$$log\ MHOE33_j = \beta_0 + \beta_1 MHOE23_j + \beta_2 MHOE42_j + \beta_3 MHOE46_j + \beta_4 MHOE50_j + \beta_5 MHOE55_j + \beta_6 PSES16_j + \beta_7 Female_j + \beta_8 South\ \&\ East23_j + \beta_9 Wales23_j + \beta_{10} Scotland23_j + \beta_{11} South\ \&\ East33_j + \beta_{12} Wales33_j + \beta_{13} Scotland33_j + \beta_{14} South\ \&\ East42_j + \beta_{15} Wales42_j + \beta_{16} Scotland42_j + \beta_{17} South\ \&\ East46_j + \beta_{18} Wales46_j + \beta_{19} Scotland46_j + \beta_{20} South\ \&\ East50_j + \beta_{21} Wales50_j + \beta_{22} Scotland50_j + \beta_{23} South\ \&\ East55_j + \beta_{24} Wales55_j + \beta_{25} Scotland55_j$$

$$log\ MHOE42_j = \beta_0 + \beta_1 MHOE23_j + \beta_2 MHOE33_j + \beta_3 MHOE46_j + \beta_4 MHOE50_j + \beta_5 MHOE55_j + \beta_6 PSES16_j + \beta_7 Female_j + \beta_8 South\ \&\ East23_j + \beta_9 Wales23_j + \beta_{10} Scotland23_j + \beta_{11} South\ \&\ East33_j + \beta_{12} Wales33_j + \beta_{13} Scotland33_j + \beta_{14} South\ \&\ East42_j + \beta_{15} Wales42_j + \beta_{16} Scotland42_j + \beta_{17} South\ \&\ East46_j + \beta_{18} Wales46_j + \beta_{19} Scotland46_j + \beta_{20} South\ \&\ East50_j + \beta_{21} Wales50_j + \beta_{22} Scotland50_j + \beta_{23} South\ \&\ East55_j + \beta_{24} Wales55_j + \beta_{25} Scotland55_j$$

$$log\ MHOE46_j = \beta_0 + \beta_1 MHOE23_j + \beta_2 MHOE33_j + \beta_3 MHOE42_j + \beta_4 MHOE50_j + \beta_5 MHOE55_j + \beta_6 PSES16_j + \beta_7 Female_j + \beta_8 South\ \&\ East23_j + \beta_9 Wales23_j + \beta_{10} Scotland23_j + \beta_{11} South\ \&\ East33_j + \beta_{12} Wales33_j + \beta_{13} Scotland33_j + \beta_{14} South\ \&\ East42_j + \beta_{15} Wales42_j + \beta_{16} Scotland42_j + \beta_{17} South\ \&\ East46_j + \beta_{18} Wales46_j + \beta_{19} Scotland46_j + \beta_{20} South\ \&\ East50_j + \beta_{21} Wales50_j + \beta_{22} Scotland50_j + \beta_{23} South\ \&\ East55_j + \beta_{24} Wales55_j + \beta_{25} Scotland55_j$$

$$log\ MHOE50_j = \beta_0 + \beta_1 MHOE23_j + \beta_2 MHOE33_j + \beta_3 MHOE42_j + \beta_4 MHOE46_j + \beta_5 MHOE55_j + \beta_6 PSES16_j + \beta_7 Female_j + \beta_8 South\ \&\ East23_j + \beta_9 Wales23_j + \beta_{10} Scotland23_j + \beta_{11} South\ \&\ East33_j + \beta_{12} Wales33_j + \beta_{13} Scotland33_j + \beta_{14} South\ \&\ East42_j + \beta_{15} Wales42_j + \beta_{16} Scotland42_j + \beta_{17} South\ \&\ East46_j + \beta_{18} Wales46_j + \beta_{19} Scotland46_j + \beta_{20} South\ \&\ East50_j + \beta_{21} Wales50_j + \beta_{22} Scotland50_j + \beta_{23} South\ \&\ East55_j + \beta_{24} Wales55_j + \beta_{25} Scotland55_j$$

$$log\ MHOE55_j = \beta_0 + \beta_1 MHOE23_j + \beta_2 MHOE33_j + \beta_3 MHOE42_j + \beta_4 MHOE46_j + \beta_5 MHOE50_j + \beta_6 PSES16_j + \beta_7 Female_j + \beta_8 South\ \&\ East23_j + \beta_9 Wales23_j + \beta_{10} Scotland23_j + \beta_{11} South\ \&\ East33_j + \beta_{12} Wales33_j + \beta_{13} Scotland33_j + \beta_{14} South\ \&\ East42_j + \beta_{15} Wales42_j + \beta_{16} Scotland42_j + \beta_{17} South\ \&\ East46_j + \beta_{18} Wales46_j + \beta_{19} Scotland46_j + \beta_{20} South\ \&\ East50_j + \beta_{21} Wales50_j + \beta_{22} Scotland50_j + \beta_{23} South\ \&\ East55_j + \beta_{24} Wales55_j + \beta_{25} Scotland55_j$$

$$RoR23_j = \beta_0 + \beta_1 MHOE23_j + \beta_2 MHOE33_j + \beta_3 MHOE42_j + \beta_4 MHOE46_j + \beta_5 MHOE50_j + \beta_6 MHOE55_j + \beta_7 PSES16_j + \beta_8 Female_j + \beta_9 South\ \&\ East55_j + \beta_{10} Wales55_j + \beta_{11} Scotland55_j$$

$$RoR33_j = \beta_0 + \beta_1 MHOE23_j + \beta_2 MHOE33_j + \beta_3 MHOE42_j + \beta_4 MHOE46_j + \beta_5 MHOE50_j + \beta_6 MHOE55_j + \beta_7 PSES16_j + \beta_8 Female_j + \beta_9 South\ \&\ East23_j + \beta_{10} Wales23_j + \beta_{11} Scotland23_j$$

$$RoR42_j = \beta_0 + \beta_1 \text{MHOE23}_j + \beta_2 \text{MHOE33}_j + \beta_3 \text{MHOE42}_j + \beta_4 \text{MHOE46}_j + \beta_5 \text{MHOE50}_j + \beta_6 \text{MHOE55}_j + \beta_7 \text{PSES16}_j + \beta_8 \text{Female}_j + \beta_9 \text{South \& East23}_j + \beta_{10} \text{Wales23}_j + \beta_{11} \text{Scotland23}_j$$

$$RoR46_j = \beta_0 + \beta_1 \text{MHOE23}_j + \beta_2 \text{MHOE33}_j + \beta_3 \text{MHOE42}_j + \beta_4 \text{MHOE46}_j + \beta_5 \text{MHOE50}_j + \beta_6 \text{MHOE55}_j + \beta_7 \text{PSES16}_j + \beta_8 \text{Female}_j + \beta_9 \text{South \& East23}_j + \beta_{10} \text{Wales23}_j + \beta_{11} \text{Scotland23}_j$$

$$RoR50_j = \beta_0 + \beta_1 \text{MHOE23}_j + \beta_2 \text{MHOE33}_j + \beta_3 \text{MHOE42}_j + \beta_4 \text{MHOE46}_j + \beta_5 \text{MHOE50}_j + \beta_6 \text{MHOE55}_j + \beta_7 \text{PSES16}_j + \beta_8 \text{Female}_j + \beta_9 \text{South \& East23}_j + \beta_{10} \text{Wales23}_j + \beta_{11} \text{Scotland23}_j$$

$$RoR55_j = \beta_0 + \beta_1 \text{MHOE23}_j + \beta_2 \text{MHOE33}_j + \beta_3 \text{MHOE42}_j + \beta_4 \text{MHOE46}_j + \beta_5 \text{MHOE50}_j + \beta_6 \text{MHOE55}_j + \beta_7 \text{PSES16}_j + \beta_8 \text{Female}_j + \beta_9 \text{South \& East23}_j + \beta_{10} \text{Wales23}_j + \beta_{11} \text{Scotland23}_j$$

$$\begin{aligned} PSES16_j = {} & \beta_0 + \beta_1 \text{MHOE23}_j + \beta_2 \text{MHOE33}_j + \beta_3 \text{MHOE42}_j + \beta_4 \text{MHOE46}_j \\ & + \beta_5 \text{MHOE50}_j + \beta_6 \text{MHOE55}_j + \beta_7 \text{Female}_j + \beta_8 \text{South \& East23}_j \\ & + \beta_9 \text{Wales23}_j + \beta_{10} \text{Scotland23}_j + \beta_{11} \text{South \& East33}_j \\ & + \beta_{12} \text{Wales33}_j + \beta_{13} \text{Scotland33}_j + \beta_{14} \text{South \& East42}_j \\ & + \beta_{15} \text{Wales42}_j + \beta_{16} \text{Scotland42}_j + \beta_{17} \text{South \& East46}_j \\ & + \beta_{18} \text{Wales46}_j + \beta_{19} \text{Scotland46}_j + \beta_{20} \text{South \& East50}_j \\ & + \beta_{21} \text{Wales50}_j + \beta_{22} \text{Scotland50}_j + \beta_{23} \text{South \& East55}_j \\ & + \beta_{24} \text{Wales55}_j + \beta_{25} \text{Scotland55}_j \end{aligned}$$

Appendix C

Table 1C. Life course Mean Hourly Occupational Earnings multilevel model random effects comparisons of coefficients (standard errors)

	AC	CC	PO	MLMI	MICE
Random Effects	**Coeff. (se)**	**Coeff. (se)**	**Coeff. (se)**	**Coeff. (se)**	**Coeff. (se)**
var(age23)	0.086 (0.001)	0.090 (0.003)	0.082 (0.002)	0.092 (0.002)	0.085 (0.001)
cov(age23,age33)	0.045 (0.001)	0.045 (0.003)	0.044 (0.002)	0.046 (0.002)	0.046 (0.002)
var(age33)	0.139 (0.002)	0.138 (0.004)	0.140 (0.003)	0.151 (0.003)	0.142 (0.002)
cov(age23,age42)	0.040 (0.002)	0.037 (0.003)	0.041 (0.002)	0.042 (0.001)	0.041 (0.002)
cov(age33,age42)	0.082 (0.002)	0.075 (0.003)	0.085 (0.002)	0.086 (0.002)	0.082 (0.002)
var(age42)	0.156 (0.003)	0.155 (0.005)	0.155 (0.003)	0.169 (0.003)	0.157 (0.002)
cov(age23,age46)	0.038 (0.002)	0.036 (0.003)	0.037 (0.002)	0.039 (0.002)	0.038 (0.002)
cov(age33,age46)	0.072 (0.002)	0.067 (0.003)	0.074 (0.002)	0.076 (0.002)	0.070 (0.002)
cov(age42,age46)	0.101 (0.002)	0.098 (0.004)	0.102 (0.003)	0.106 (0.002)	0.101 (0.002)
var(age46)	0.144 (0.003)	0.142 (0.004)	0.144 (0.003)	0.167 (0.004)	0.146 (0.002)
cov(age23,age50)	0.035 (0.002)	0.032 (0.002)	0.037 (0.002)	0.037 (0.001)	0.035 (0.002)
cov(age33,age50)	0.070 (0.002)	0.063 (0.003)	0.072 (0.003)	0.073 (0.002)	0.069 (0.002)
cov(age42,age50)	0.094 (0.002)	0.089 (0.004)	0.097 (0.003)	0.100 (0.002)	0.095 (0.002)
cov(age46,age50)	0.109 (0.002)	0.103 (0.004)	0.112 (0.003)	0.113 (0.003)	0.110 (0.002)
var(age50)	0.145 (0.003)	0.138 (0.004)	0.148 (0.003)	0.165 (0.003)	0.145 (0.002)
cov(age23,age55)	0.033 (0.002)	0.031 (0.002)	0.033 (0.002)	0.034 (0.001)	0.033 (0.002)
cov(age33,age55)	0.064 (0.002)	0.059 (0.003)	0.067 (0.003)	0.069 (0.002)	0.064 (0.002)
cov(age42,age55)	0.083 (0.002)	0.081 (0.003)	0.083 (0.003)	0.090 (0.002)	0.084 (0.002)
cov(age46,age55)	0.091 (0.002)	0.087 (0.003)	0.093 (0.003)	0.100 (0.002)	0.093 (0.002)
cov(age50,age55)	0.100 (0.002)	0.095 (0.003)	0.102 (0.003)	0.105 (0.002)	0.102 (0.002)
var(age55)	0.138 (0.003)	0.133 (0.004)	0.139 (0.004)	0.163 (0.004)	0.142 (0.002)
Occasions	37478	13561	23917	82654	82654
Individuals	9622	2263	7359	14268	14268
Minimum	1	5	1	1	1
Average	3.9	6	3.3	5.8	5.8
Maximum	6	6	5	6	6
Deviance	14789	4689	9972	42739	28046

Appendix D

In this appendix, we present a summary of results using MHOE subject specific model predictions. Table 1D presents a comparison of average life course step function MLM predictions. The AC results are displayed as mean predictions and the other results are displayed as predicted percentage differences compared to the AC values. Relative to AC, the average life course predicted percentage difference between CC and PO is just over 2% with the gap between the groups narrowing over the life course. This finding is consistent with the descriptive statistics presented in Figure 2 in the paper. By contrast, the average life course predicted percentage difference between CC and PO is larger than the gap between MLMI and MICE by a factor greater than 20 relative to AC. The differences between MLMI and MICE, with respect to these predictions, appear to be negligible. Given the underlying data and model of interest, the evidence suggests that these two MI methods produce similar results. Furthermore, both MLMI and MICE methods produced results that were more similar to PO than CC but more similar to AC than PO.

Table 1D. Subject specific life course log Mean Hourly Occupational Earnings multilevel model predictions

Age	AC	CC	PO	MLMI	MICE
23	1.954	2.0%	-0.9%	-0.4%	-0.3%
33	2.065	1.9%	-0.8%	-0.2%	-0.3%
42	2.182	1.6%	-0.7%	-0.3%	-0.4%
46	2.322	1.2%	-0.7%	-0.8%	-1.0%
50	2.357	1.3%	-0.7%	-0.4%	-0.6%
55	2.287	1.0%	-0.7%	-0.6%	-0.8%
Average	2.194	1.5%	-0.7%	-0.5%	-0.6%