



Multilevel modelling approach to analysing life course socioeconomic status and understanding missingness

Adrian Byrne¹ · Natalie Shlomo² · Tarani Chandola³

Received: 17 May 2021 / Accepted: 17 May 2022
© The Author(s) 2022

Abstract

This paper investigated the extent to which parental socioeconomic status was associated with life course socioeconomic status heterogeneity between adult cohort members of the 1958 National Child Development Study and how this association differed depending on methods used to address longitudinal missing data. We compared three variants of the full information maximum likelihood approach, namely available case, complete case and partially observed case and two methods designed to compensate for missing at random data, namely multilevel multiple imputation and multiple imputation chained equations. Our results highlighted the important contribution of parental socioeconomic status in explaining the divergence in achieved socioeconomic status over the adult life course, how the available case approach increasingly overestimated socioeconomic attainment as age increased and survey sample size decreased and how the complete case approach downwardly biased the effect of parental socioeconomic status on adult socioeconomic status.

Keywords Multilevel models · Longitudinal data · Socioeconomic status · Life course · Missing data · Multiple imputation

1 Introduction

Conventional intergenerational social mobility research often examines transitions between origin (parental) class and destination (adult) class/SES using two time points in the life course (Blanden et al. 2004; Goldthorpe and Jackson 2007; Li and

✉ Adrian Byrne
adrian.byrne@nottingham.ac.uk

¹ Division of Rehabilitation, Ageing and Wellbeing School of Medicine, Queen's Medical Centre, University of Nottingham, Nottingham NG7 2UH, UK

² Department of Social Statistics, School of Social Sciences, University of Manchester, Humanities Bridgeford Street Building, Manchester M13 9PL, UK

³ Department of Sociology, University of Hong Kong, 9.07, 9/F., The Jockey Club Tower, Centennial Campus, Pokfulam Road, Hong Kong, People's Republic of China

Devine 2011). However, occupations and socioeconomic status (SES¹) change over the life course so estimates of social mobility using only one time point in adulthood may be biased. This has proven to be the case with respect to income elasticity as the proxy for intergenerational economic dependence and is referred to as the “lifecycle bias” (Nybom and Stuhler 2016, 2017). Moreover, the analysis of changes in occupational class and socioeconomic status in adulthood from longitudinal surveys may also be biased because of non-random missing data mechanisms in longitudinal studies. In other words, socioeconomically disadvantaged survey respondents are more likely to be lost to follow-up resulting in potentially biased estimates of the association between origin and destination class or SES using complete case approaches. Therefore, this study addressed these methodological limitations of existing studies on social mobility. By doing so, this study contributes to the research literature on life course inequality and the serious issue of survey non-response throughout the life course.

SES is often measured either as a combination of education, income and occupation or by investigating these elements separately. Moreover, there are many different measures of SES that can be used throughout the life course (Galobardes et al. 2006a, 2006b). In contrast to categorical social class measures, SES is commonly conceptualised as a continuous variable, referring to the rank or social standing of an individual or group (American Psychological Association 2007).

Most research on social mobility have used categorical measures of social class thereby producing more aggregated floor and ceiling effects when analysing social mobility (Goldthorpe and Jackson 2007; Li and Devine 2011; Sturgis and 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 and Friedman 2016). Jonsson et al. (2009) have tackled this issue by introducing the microclass approach to examine intergenerational social mobility. This approach requires large sample sizes to ensure adequate statistical power when covering a large number of microclass occupational categories. However, we argue in this paper that employing a continuous, repeated measure of life course SES provided a better opportunity of finding any potentially important differences obscured by the floor and ceiling effects of these categorical measures during the life course. We argue that our approach is more sensitive to changes than conventional broad categorical measures. Therefore, this study investigated the heterogeneity between individuals in relation to a finely grained continuous measure of SES over the life course.

¹ Abbreviations used in this paper: SES=socioeconomic status; MCAR=missing completely at random; MI=multiple imputation; MAR=missing at random; MNAR=missing not at random; NCDS=National Child Development Study; CAMSIS=Cambridge Social Interaction and Stratification Scale; SIOPS=Standard International Occupational Prestige Scale; ISEI=International Socio-Economic Index; MHOE=Mean Hourly Occupational Earnings; oPCA=ordinal Principal Component Analysis; PSES16=parental SES at age 16; ML=multilevel; FIML=Full Information Maximum Likelihood; MLMI=multilevel multiple imputation; MICE=multiple imputation chained equations; IGLS=iterative generalised least squares; AC=available case; CC=complete case; PO=partially observed.

With their analysis of a longitudinal birth cohort survey, Hawkes and Plewis (2006) reported that there were systematic differences between respondents and non-respondents at every wave with the more disadvantaged survey members being more likely to be lost from the study. They concluded that there was no support for the position that particular longitudinal datasets (and indeed any life course longitudinal dataset) can be considered “missing completely at random” (MCAR). MCAR implies that the probability that a cohort member becomes a non-responder in the study for at least one wave does not depend on the factors/data/variables that are accounted for in the statistical model or the unobserved factors/data/variables. If the data were MCAR, then complete case analysis, excluding those with any missing data, would not be affected by non-response bias with no need to account for the missing data.

Existing studies on social origin and destination associations have not explored whether different approaches to compensating for missing data result in different estimates of these associations. These studies either ignored the problem of missing data (Li and Devine 2011) or employed a method without much consideration of the relevance for their substantive purposes (Gugushvili et al. 2017; Sturgis and Sullivan 2008). Conversely, studies that focused on missing data methods prioritised showcasing the methods through theory and simulation techniques (Goldstein et al. 2014). This study attempted to bridge the gap between these two areas of research that so far have been kept distinct.

While many studies on life course SES reported missingness (Niedzwiedz et al. 2012), not many examined in detail how the missing data affected inference established from the model of interest. Furthermore, past analyses of life course SES had employed an estimation procedure that could handle missing data under certain assumptions without conducting any sensitivity analysis to ensure their inferences are robust to the missingness (Sturgis and Sullivan 2008). More recently, studies such as Gugushvili et al. (2017) dealt with missing data through multiple imputation (MI) analyses. However, these studies did not provide much detail on how the missing data affected their variables/models of interest so it was hard to evaluate how the imputed data impacted upon the subsequent analysis and inference. Furthermore, the issue of whether to impute longitudinal data that respects or ignores the underlying data structure, i.e. long (multiple lines of data for each individual clustered by wave) vs wide (each line of data represents an individual) format, remains unexplored in social mobility studies but has been covered elsewhere (Young and Johnson 2015).

Multilevel modelling is one of the most common methods for analysing trajectories of growth and change in time through repeated measurements (such as socioeconomic status). One benefit of using multilevel modelling in relation to missing data is that there is no requirement to have balanced data, so individuals who contribute fewer than the maximum number of observations, i.e. who are 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 (Carpenter and Plewis 2011). However, this approach is only reasonable assuming the missing data are “missing at random”

(MAR) (Rasbash et al. 2015). MAR implies that the probability that a cohort member becomes a non-responder in the study for at least one wave does not depend on the unobserved factors/data/variables so long as the observed factors/data/variables are accounted for in the statistical model.

Further to the unbalanced data capability, we chose multilevel modelling over fixed effects modelling as we wanted to target the between-individual variation at the higher level within our set-up despite running the risk our estimator could be inconsistent due to some correlation existing between the random effects and model regressors. Despite this potential risk, there is evidence that it may still be preferable to select the multilevel model over the fixed effects model due to sufficient gains in efficiency, i.e. smaller standard errors (Bell and Jones 2015; Clark and Linzer 2015; Bell et al. 2019).

If the probability that a cohort member becomes a non-responder in the study for at least one wave does depend on the unobserved factors/data/variables, even if the observed factors/data/variables are accounted for in the statistical model, 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 is made plausible by including explanatory variables in the model of interest that contain predictive power in relation to the unobserved data as well as the observed data (White et al. 2011). Without this, inferences on life course socioeconomic status derived from the observed sample may be biased in relation to the parameters of interest in the target population.

In this study, we modelled the effect of parental SES on attained SES in adulthood. We also compared different approaches to accounting for missing data in multilevel models of life course SES including complete case analysis. We examined the extent to which different approaches to missing data compensate for non-response bias in the estimates of the effect of childhood SES on adulthood SES in a longitudinal life course survey. Our fine-grained approach to SES coupled with our investigation of different modelling solutions to missing data over the life course forms the basis of our contribution to the research literature on life course inequality.

2 Materials and methods

2.1 National Child Development Study life course data

Our chosen birth cohort longitudinal dataset was the 1958 National Child Development Study (NCDS) (University of London. Institute of Education. Centre for Longitudinal Studies. 2008a, 2008b, 2008c, 2008d, 2012, 2014, 2015). We chose this dataset because it is internationally renowned and tracks a sample of people born in 1 week in 1958 with surveys from birth, through childhood and adolescence and into later life adulthood. The survey data is collected at the individual level across Great Britain making it a nationally representative longitudinal cohort study of British children born in 1958. Data from the parents was collected from 1958 to 1974 with the children becoming the main survey responders from 1981 until 2013 in this study.

At the most recent sweep in this study (i.e. NCDS9 in 2013 when cohort members were aged 55), both the cross-sectional and longitudinal response rates were 58% (Johnson et al. 2015). As our target population was based on those with known occupation data over the life course, the total eligible NCDS sample size of 18,558 reduced to 14,268 individuals as 4290 (23.1%) cohort members had no occupation data over the life course (6 time points) between the ages of 23 and 55. Our target population had a 51:49 men-to-women ratio whereas the ratio was 54:46 for those cohort members omitted from our target population with no occupation data. As there were a total of 14,268 person-level (defined as level 2 units “L2” hereafter) available cases in this analysis, that allowed for a possible total of 85,608 ($14,268 \times 6$ 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 was reduced to 53,958 L1 units. Including the model covariates with their missingness reduced the number of L1 units to 37,478 and the number of L2 units to 9622 representing 44% and 67% respectively of the sample at each level.

2.2 Life course outcome measure

We derived our life course outcome measure of SES from the Occupational Earnings Scale (Bukodi et al. 2011; Nickell 1982; Byrne et al. 2018) for NCDS survey responders aged between 23 (in 1981) and 55 (in 2013). This measure injected 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). A similar social stratification scale was developed by Sobek (1995). We acknowledged that other gradational scales exist such as the Cambridge Social Interaction and Stratification Scale (CAMSIS) (Prandy 1990; Stewart et al. 1980), the Standard International Occupational Prestige Scale (SIOPS) (Ganzeboom and Treiman 1996, 2003) and the International Socioeconomic Index (ISEI) (Ganzeboom et al. 1992), but we chose the Occupational Earnings Scale as it provided an indication of the market value for an occupation at a given point in time. By linking published earnings data to routinely collected occupation data, the problem of survey responders not declaring their income was avoided. Therefore, we chose a hierarchical occupation variable that was 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 were adjusted for inflation, and wages were deflated to 1997 prices using ONS Consumer Price Index data. The 1997 prices were imposed on pre-1997 occupation wage data (for NCDS survey sweeps in 1981 and 1991), and post-1997 occupation wage data (for NCDS survey sweeps in 2000, 2004, 2008 and 2013) were deflated to keep prices constant at 1997 to aid comparability across the life course. These hourly wage data were also transformed onto the natural logarithmic scale to correct for positive skewness. We denoted our outcome variable as “mean hourly occupational earnings” (MHOE) and referred to it in the text as “occupational earnings”.

2.3 Parental SES predictor

Similar to Caro and Cortés (2012), we treated this predictor as a formatively causal composite response to observed variables that are associated with SES, namely Registrar General's 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). Formatively causal composite response here implied that the observed variables cause the latent SES construct and not the other way round as is the case with factor analysis (Bollen and Bauldry 2011). We applied the ordinal principal component analysis (oPCA) methodology developed by Kolenikov and Angeles (2009) to derive the main component which represented a linear summary of our selected parental SES variables. We calculated that a linear summary of all five variables would be a more parsimonious measure of parental SES which avoided the issue of multicollinearity while being more potent than any one of the underlying variables with respect to our life course outcome measure. These variables were collected when survey cohort members were aged 16 (1974), and we chose age 16 because the same social class and education variables were available for both parents.

All five discrete variables were ordinal with higher values representing higher social class, more education and greater ownership of accommodation. The number of categories across the five variables allowed for a possible 38,400 ($8 \times 8 \times 10 \times 10 \times 6$) 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. 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 supplementary material to access the results from this oPCA estimation. The configured scores for the main principal component were estimated for each NCDS cohort member and then transformed onto the natural logarithmic scale to account for positive skewness. The scores were then standardised with mean of zero and standard deviation of one to aid interpretation. We denoted this person-level variable as "PSES16" hereafter. We adjusted the relationship between PSES16 and MHOE by controlling for key sociodemographic variables namely sex at birth and region of residence as advocated by the American Psychological Association (2007).

2.4 Modelling life course occupational earnings using multilevel modelling

Multilevel modelling provided a method for analysing change over time whereby the repeated measures are viewed as outcomes that are dependent on some metric of time. There can be predictors of interest at either level (time/individual) and may include cross-level interactions (Steele 2008). 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 (ML) model set-up.

Repeated observations over time, which need not be equally spaced out, constitute level 1 units nested within individuals at level 2. The multilevel framework can account for correlations of observations across time. However, this may prove not to be the case when dealing with many time points (Bolger and Laurenceau 2013), and in such instances, one could test for the level of autocorrelation between observations using a more recently developed framework such as dynamic structural equation modelling (Asparouhov and Muthén 2020).

2.5 Compensating for missing data

As both our response variable (MHOE) and main predictor (PSES16) are related to socioeconomic advantage/disadvantage and as Hawkes and Plewis (2006) had concluded previously, we did 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 confirmed our data are not MCAR.

We evaluated our substantive model through employing different approaches to compensating for missing longitudinal survey data. A consideration regarding missing data was the pattern of non-response. All cohort members who first contributed data to the NCDS survey and then become permanent non-responders during the life course are described as following a monotonic pattern of non-response. The alternative non-response pattern is referred to as non-monotonic whereby NCDS cohort members had intermittent survey responses throughout the life course. Appreciating the difference between these two non-response patterns was important in terms of selecting an appropriate missing data solution.

In terms of monotonic MHOE response patterns, whereby NCDS cohort members either never left the study or became permanent non-responders, approximately 52% of cohort members followed a monotonic pattern. The remaining 48% followed a non-monotonic pattern whereby cohort members had intermittent life course MHOE responses. In total, there existed 63 different MHOE life course response patterns. The concept of monotone missingness was important because it could simplify the application of missing data solutions. However, the development of joint modelling and chained equation approaches to multiple imputation had reduced the computational problems associated with a lack of monotonicity within the missing data pattern. We focus on these two MI approaches. Other solutions such as inverse probability weighting and selection/pattern mixture modelling were considerably more complicated when addressing intermittent, non-monotone missingness (Bartlett and Carpenter 2013) and were not considered in this paper.

MI replaces each missing value with multiple imputed values, which are random draws from the joint conditional probability distributions derived from the selected imputation model for all variables affected by missingness conditional on available data (Rubin 1987). The end result is multiple complete datasets. We then fitted the model of interest to each imputed dataset, and the parameter estimates and standard

errors from these models were combined using Rubin's rules (Rubin 1987). This took 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 et al. (2014) highlighted that there are two approaches to MI, one that uses the joint conditional probability distribution of all variables with missing data conditional on available data and the chained equation approach that uses the conditional probability distribution for each variable with missing data in turn conditional on available data. A distinct advantage of the former is the implementation of multilevel data structures and interactions in the imputation process (Goldstein et al. 2009). Our model of interest contains both a multilevel structure and interactions. Therefore, the joint modelling MI approach should be more suitable for our situation.

In this paper, we first created an artificial missing data scenario using only complete case data to compare available case and the two MI methods against "ground truth". We then investigated the empirical difference between the following five approaches, all assuming that the underlying survey data were MAR given the substantive model of interest:

1. Available case (with respect to life course MHOE)—using an estimation procedure that could handle missing data namely the full information maximum likelihood (FIML) approach as advocated by Bartlett and (Bartlett and Carpenter 2013)
2. Complete case (with respect to life course MHOE)
3. Partially observed (with respect to life course MHOE)
4. Multilevel MI (MLMI)—computationally slower MI approach but the imputation model was more similar to the substantive model of interest
5. MI chained equations (MICEs)—computationally faster MI approach but the imputation model was less similar to the substantive model of interest

These five approaches allowed us to investigate two distinct inquiries:

1. 1 vs 2 vs 3 highlighted any bias arising from a complete case analysis.
2. 1 vs 4 vs 5 highlighted the difference between not imputing and imputing as well as the difference between imputing in long format (same data structure as 1–3) vs wide format.

The MLMI approach used a joint two-level model to impute missing values whereas the MICE approach used single-level regression models. Furthermore, the two approaches required different dataset formats before beginning each process; MLMI required the data in long format whereby a record corresponds to an occasion nested within an individual (similar to our FIML approach), and MICE required the data in wide format whereby a record corresponds to an individual. By comparing an imputation model that was faithful to the substantive FIML model of interest as advocated by Goldstein (2009) with one that was not, any significant differences found between the MI approaches and the FIML results could have indicated a poor choice of imputation model.

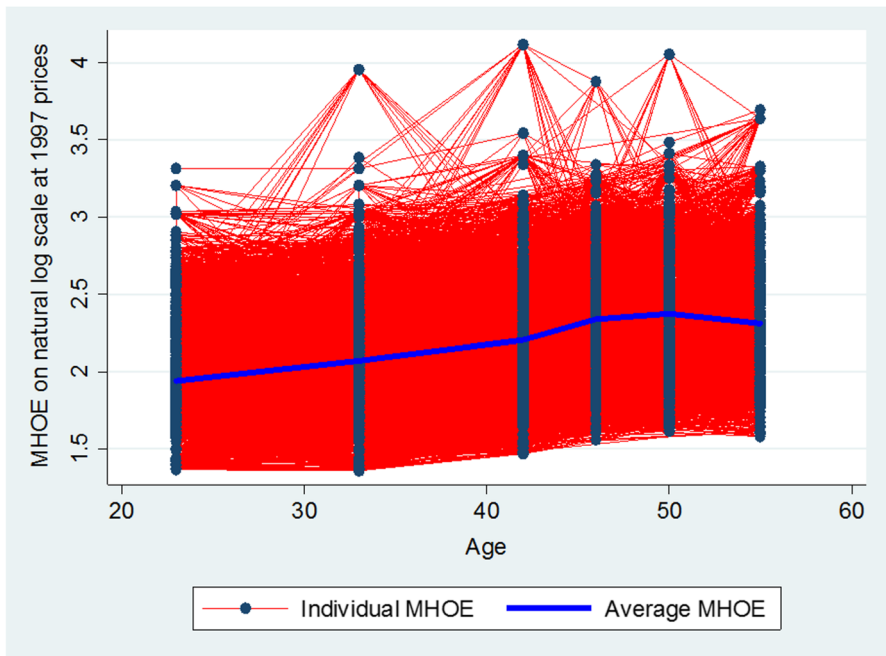


Fig. 1 Life course mean hourly occupational earnings on natural log scale at 1997 prices. Data sources are NCDS occupations (1981–2013) and ONS CPI (1997–2013). Single column figure shows individual scatter points around the average trend between ages 23 and 55

3 Results

3.1 Descriptive statistics and model set-up

With respect to our life course outcome variable, Fig. 1 displays both the average and individual life course MHOE trajectories. The average life course trend appeared 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 cohort members were aged 50 and/or by the effects of early retirement. However, the mean curve is not too dissimilar from the expected income curve in the life cycle hypothesis developed by Modigliani and Brumberg (1954) whereby income rises through the 20 s, 30 s, 40 s peaking around age 50 before declining into retirement. We note that there was substantial individual variation around this average life course trend and multilevel modelling provided a suitable way of accounting for this variation when modelling the average life course trend.

To model this individual life course occupational earnings development and variation, we adopted a multilevel framework (Steele 2014). Without including

the covariates, we explored a number of different time functions to empirically verify a suitable model given the data and chose to model time using a step function. Our step function multilevel model was a multivariate linear model comprising six responses corresponding to the six waves at which MHOE was measured. The variance–covariance structure was fully specified by random terms at the person-level with no occasion-level residuals. Furthermore, we could employ a step function as the measurement occasions were the same for all NCDS cohort members. Our analysis suggested that additional model complexity in terms of more parameters was merited given the variation in the underlying data. We chose the step function ML model as the basis for our model of interest as the multivariate formulation of MHOE life course between-individual development was simpler to interpret (Steele 2014).

Our model of interest set-up included six age dummy variables and no model intercept to aid interpretation. A different random intercept was deployed at each age to allow for individual variation at each time point in relation to the MHOE response variable. We employed 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 were omitted and acted as reference categories. The reference category for our nominal categorical L1 region of residence predictor was North & Midlands. See Eq. 1 for our life course step function ML model of interest specification with subscripts *t* for L1 time points and *j* for L2 individuals.

Equation 1 Life course step function multilevel model of interest

$$\begin{aligned}
 \log \text{ Mean Hourly Occupational Earnings}_{ij} = & \beta_1 \text{age}23_{ij} + \beta_2 \text{age}33_{ij} \\
 & + \beta_3 \text{age}42_{ij} + \beta_4 \text{age}46_{ij} + \beta_5 \text{age}50_{ij} + \beta_6 \text{age}55_{ij} + \beta_7 \text{PSES}16_j \\
 & + \beta_8 \text{Female}_j + \beta_9 \text{age}33_{ij} \times \text{PSES}16_j + \beta_{10} \text{age}42_{ij} \times \text{PSES}16_j \\
 & + \beta_{11} \text{age}46_{ij} \times \text{PSES}16_j + \beta_{12} \text{age}50_{ij} \times \text{PSES}16_j + \beta_{13} \text{age}55_{ij} \times \text{PSES}16_j \\
 & + \beta_{14} \text{age}33_{ij} \times \text{Female}_j + \beta_{15} \text{age}42_{ij} \times \text{Female}_j + \beta_{16} \text{age}46_{ij} \times \text{Female}_j \\
 & + \beta_{17} \text{age}50_{ij} \times \text{Female}_j + \beta_{18} \text{age}55_{ij} \times \text{Female}_j + \beta_{19} \text{South \& East}_{ij} \\
 & + \beta_{20} \text{Wales}_{ij} + \beta_{21} \text{Scotland}_{ij} + u_{1j} \text{age}23_{ij} + u_{2j} \text{age}33_{ij} \\
 & + u_{3j} \text{age}42_{ij} + u_{4j} \text{age}46_{ij} + u_{5j} \text{age}50_{ij} + u_{6j} \text{age}55_{ij} \\
 \mathbf{u} \sim & N(\mathbf{0}, \mathbf{\Omega}_u)
 \end{aligned} \tag{1}$$

$$\mathbf{\Omega}_u = \mathbb{V} \begin{bmatrix} \mathbf{u}_{1j} \\ \vdots \\ \mathbf{u}_{6j} \end{bmatrix} = \begin{bmatrix} \sigma_{u_1}^2 & \cdots & \sigma_{u_1 u_6} \\ \vdots & \ddots & \vdots \\ \sigma_{u_1 u_6} & \cdots & \sigma_{u_6}^2 \end{bmatrix}$$

$t = 23, 33, 42, 46, 50, 55; j = 1, \dots, 9622$

The statistical analysis was carried out using MLwiN v2.32 with the default method of estimation, iterative generalised least squares (IGLS), representing the FIML approach. MLwiN only listwise deleted model predictors with missing data and not the outcome variable thereby establishing a complete-case analysis with

Table 1 Missingness with respect to variables used in the analysis ($n = 14,268$)

Variable	Missing	Not missing	Unique values	Min values	Max values
MHOE at age 23	4219	10,049	266	1.366	3.316
MHOE at age 33	3591	10,677	303	1.358	3.955
MHOE at age 42	4656	9612	300	1.465	4.117
MHOE at age 46	5984	8284	300	1.559	3.879
MHOE at age 50	6032	8236	302	1.614	4.056
MHOE at age 55	7168	7100	321	1.578	3.696
PSES at age 16	4621	9647	> 500	-3.279	2.769
Female	0	14,268	2	0	1
RoR at age 23	2125	12,143	4	1	4
RoR at age 33	2992	11,276	4	1	4
RoR at age 42	3058	11,210	4	1	4
RoR at age 46	4829	9439	4	1	4
RoR at age 50	4640	9628	4	1	4
RoR at age 55	5478	8790	4	1	4

Data source is NCDS (1974–2013). The table displays how observed each study variable is as well as the design of each variable

MHOE mean hourly occupational earnings, *PSES* parental socioeconomic status, *RoR* region of residence

respect to the predictors. This approach represented our “available case” analysis and was the most standard way of analysing missing data using multilevel modelling.

Before presenting the ML model results, we return to our NCDS study sample size of 14,268 cohort members with known occupation data over the life course. Table 1 introduces all the model variables used in this study and provides some detail about the amount of missing data for each variable over the life course. Only one variable in our analysis was fully observed, namely our sex at birth variable “female”. The other sociodemographic variable, namely region of residence (“RoR”), was also treated as a categorical variable while our main predictor of interest (“PSES”) and outcome variable (“MHOE”) were treated as continuous variables.

3.2 Multilevel model results

Table 2 displays the coefficients of two life course ML models with two different outcome variables: the linear “MHOE” fixed effects for the available cases on the natural log scale at 1997 prices and the binary “missing MHOE” (missing MHOE = 1) fixed effects for the multilevel logistic model examining which of the model covariates are predictive of missing MHOE over the life course. Both ML models used the same set of predictor variables. Starting with the former linear ML model, we saw the age dummies followed the sample occupational earnings means at each time point controlling for the model covariates. After controlling for sex at birth and region of residence, higher parental socioeconomic status at age 16,

Table 2 Life course mean hourly occupational earnings multilevel model fixed effects regression coefficients (standard errors) for available cases and binary missing MHOE

	MHOE	Missing MHOE
Age	Coeff. (SE)	Coeff. (SE)
23	2.016 (0.006)***	-1.496 (0.048)***
33	2.161 (0.007)***	-3.076 (0.085)***
42	2.279 (0.007)***	-2.414 (0.066)***
46	2.380 (0.007)***	-2.514 (0.074)***
50	2.421 (0.007)***	-2.132 (0.064)***
55	2.338 (0.007)***	-1.624 (0.057)***
Parental SES at 16	0.091 (0.004)***	0.315 (0.032)***
Female	-0.148 (0.007)***	-0.053 (0.062)
Region of residence (ref: North and Midlands)		
South and East	0.025 (0.006)***	-0.167 (0.04)***
Wales	-0.038 (0.012)**	0.183 (0.079)*
Scotland	0.011 (0.009)	0.074 (0.061)
Interaction effects		
Age × parental SES at 16 (ref: age 23 × parental SES)		
33 × PSES16	0.043 (0.005)***	-0.514 (0.058)***
42 × PSES16	0.038 (0.005)***	-0.481 (0.046)***
46 × PSES16	0.034 (0.005)***	-0.484 (0.05)***
50 × PSES16	0.024 (0.005)***	-0.544 (0.047)***
55 × PSES16	0.022 (0.005)***	-0.429 (0.044)***
Age × female (ref: age 23 × female)		
33 × Female	-0.075 (0.009)***	0.572 (0.117)***
42 × Female	-0.083 (0.010)***	1.034 (0.094)***
46 × Female	-0.032 (0.010)**	0.962 (0.102)***
50 × Female	-0.029 (0.010)**	0.694 (0.094)***
55 × Female	-0.022 (0.011)*	0.705 (0.088)***
Occasions	37,478	43,466
Individuals	9622	9635
Minimum	1	1
Average	3.9	4.5
Maximum	6	6

Data sources are NCDS (1974–2013) and ONS CPI (1997–2013). The table displays life course multilevel model fixed effects for linear MHOE and binary Missing MHOE outcome variables using the same predictor variables

* p value < 0.05

** p value < 0.01

*** p value < 0.001

in terms of standard deviation units, had 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 inferred that females with lower parental

SES at age 16 residing in Wales were at a disadvantage compared to males with higher parental SES at age 16 residing in the South and East region with respect to life course SES attainment according to our measure of log mean hourly occupational earnings. In terms of the person-level random effects, we saw the greatest variability between individuals in terms of occupational earnings occurred at age 42 while the smallest variability occurred at age 23. The magnitude of the random effect covariances increased as the time between ages decreased, and all covariances were positive implying that NCDS cohort members were becoming more divergent. This evidence suggested that larger deviations from the MHOE average at a preceding age were positively correlated with larger deviations from the MHOE average at a subsequent age controlling for the model predictors. This evidence was consistent with the idea of divergent MHOE growth over the life course with the already better off increasing their occupational earnings as they grow older at a faster rate compared to those worse off (or alternatively, the already better off not decreasing their occupational earnings as fast as those worse off through the life course). We refer the reader to Appendix B of the supplementary material for the person-level random effects estimated variances and covariances for this available case step function life course ML model.

Moving onto the latter logistic ML model, the results indicated that the model covariates were predictive of both MHOE and missing MHOE over the life course. For missing MHOE, the life course interaction effects had a different sign from the age 23 main effects suggesting a “university” effect whereby children from better-off socioeconomic backgrounds (as measured by PSES16) were more likely to be missing at age 23 but then they were less likely to be missing thereafter. There was no sex distinction at age 23, but females were more likely to be missing from age 33 onwards. The analysis also suggested broad geographical disparities with the survey responders from the South and East less likely to be missing MHOE compared to North and Midlands survey responders. Survey responders from Wales were more likely to be missing, and Scotland responders appeared to be statistically similar to North and Midlands responders so no discernible differences with respect to missing MHOE.

The life course model results presented in Table 2 show statistically significant findings for all model covariates. Figure 2 represents these significant fixed effects diagrammatically, and the patterns suggested an inverse relationship between the likelihood of missing MHOE and the likelihood of obtaining more MHOE over the life course.

3.3 A closer look at the missing data situation

As our available case (AC) cohort could be separated into complete case (CC) and partially observed (PO), we also examined MHOE differences between CC and PO for all model covariates. Figure 3 shows how each model predictor could also discriminate between CC and PO individuals, and all covariate plots reflected the overall average situation whereby CC cohort members experienced higher occupational earnings.

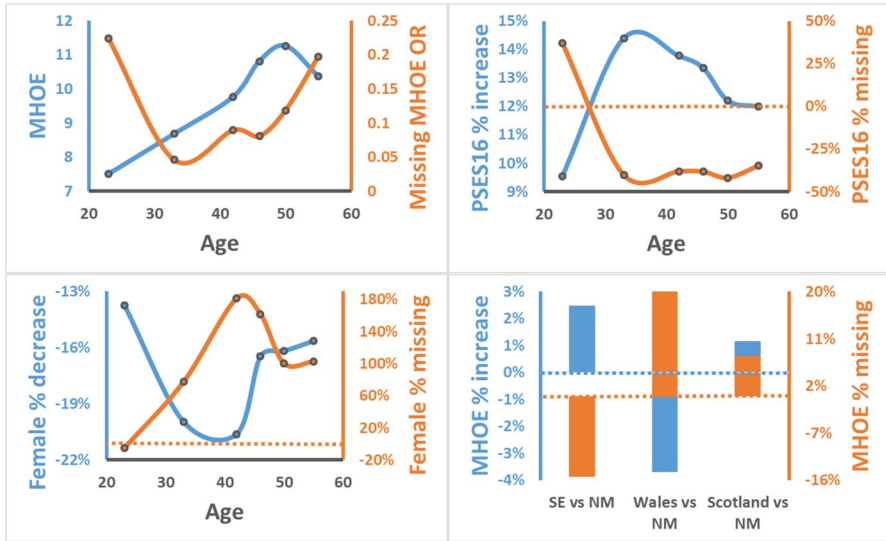


Fig. 2 Life course mean hourly occupational earnings and missing MHOE by model covariate (OR=odds ratios). Data sources are NCDS (1974–2013) and ONS CPI (1997–2013). Double column figure showing exponentiated/percentage change life course occupational earnings (blue y-axis) and percentage likelihood of missing MHOE (orange y-axis) for a one unit increase in all model covariates. The dotted lines represent the zero bound (NM=North and Midlands; SE=South and East).

A FIML approach using all available cases, as we have presented in the previous section, may provide unbiased estimation, and inference provided the data are MAR (Bartlett and Carpenter 2013). Apart from our “female” variable, our model of interest predictors suffered from a non-trivial amount of missing data, hence the need for imputation with respect to the missingness in our model of interest predictors. White and Carlin (2010) advocated the use of MI when there exists a substantial amount of missing data due to the model of interest predictors. In our life course study of occupational earnings, the number of eligible individuals reduced from 14,268 to 9622 once we took into account the missing data in the covariates, representing a 33% reduction in L2 units. When addressing our missing data situation, we did not include any auxiliary or instrumental variables in the imputation model as there were very few NCDS variables that were fully (or almost fully) observed and contained information on our respondents and non-respondents. Country of birth and ethnicity were two fully observed variables, but given the nature of the 1958 NCDS birth cohort, less than 4% of our sample were born outside of Great Britain and over 97% of our sample were recorded as white. Therefore, we deemed our sample cohort too homogeneous with respect to these variables to make a significant impact in the imputation phase. Van Buuren et al. (1999) argued that auxiliary variables should be fully observed, and Enders (2010) recommended that the correlation

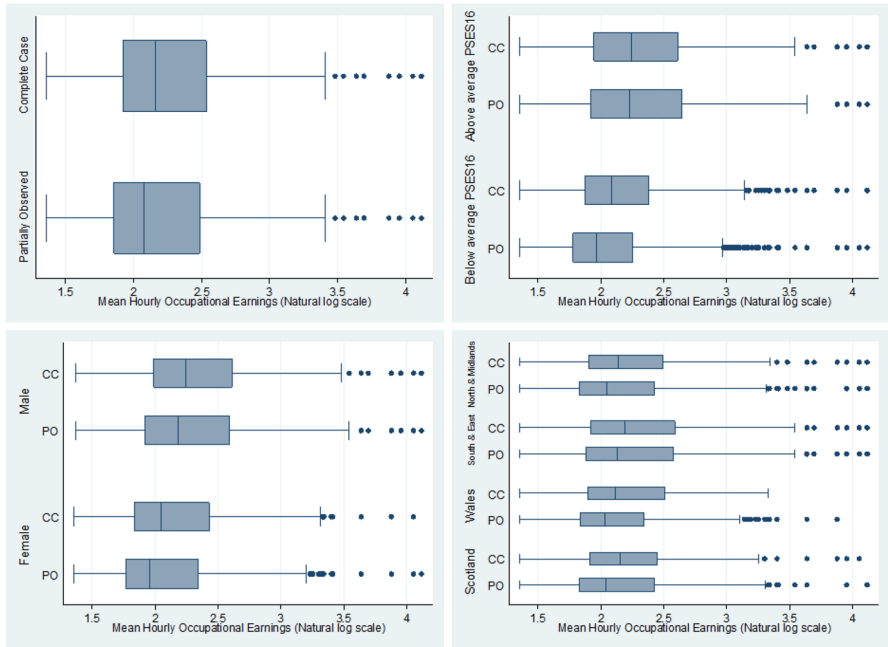


Fig. 3 Life course mean hourly occupational earnings by model covariate. Data sources are NCDS (1974–2013) and ONS CPI (1997–2013). Double column figure showing life course occupational earnings (averaged between ages 23 and 55) for all model covariates stratified by complete case and partially observed cohort members

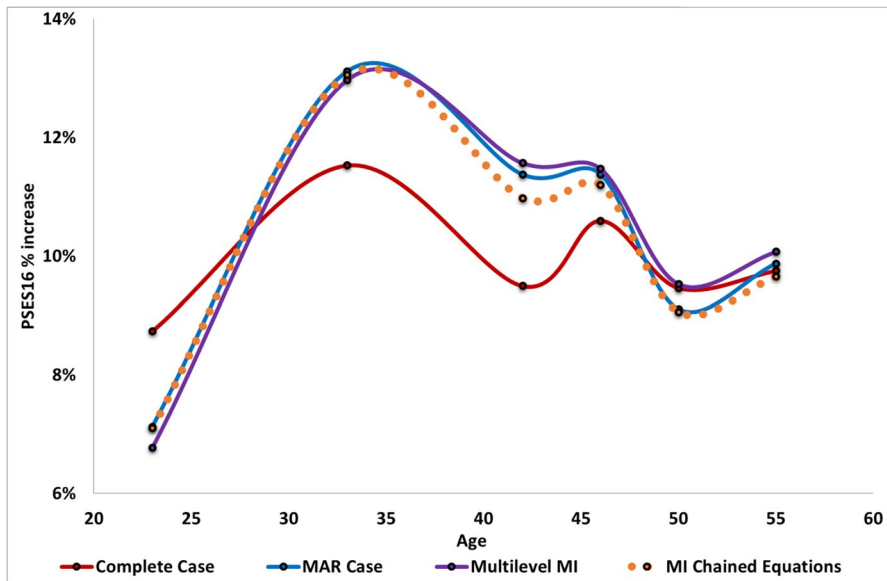
between auxiliary information and any missingness be at least 0.4. By way of example, we looked at the possibility of including father’s social class in 1958 as an auxiliary variable in the imputation model, but it was not fully observed in relation to our model of interest and proved to have a weaker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

3.4 Comparing the different approaches—artificial missingness example

In this section, we present a summary of results using MHOE subject specific model predictions incorporating both fixed and random effects. First, we present results from our artificial missingness investigation which we undertook to compare the two MI methods against a ground truth dataset. We then present the results of our empirical life course examination. We refer the reader to Appendix C (Tables C1–C4) in the supplementary material for the ML model results that underpin these model predictions, and Appendix C also contains a description of how we created the artificial missingness via the missing at random (MAR) dataset to initially test the two MI methods. Table 3 presents a comparison of average life course step function ML model predictions. The CC (i.e. ground truth) results are displayed as mean

Table 3 Subject specific life course log mean hourly occupational earnings multilevel model predictions—artificial missingness example

Age	CC	MAR	MLMI	MICE
23	1.993	-0.5%	-0.6%	-0.5%
33	2.103	-0.3%	-0.3%	-0.3%
42	2.217	-0.4%	-0.5%	-0.4%
46	2.349	-0.3%	-0.3%	-0.2%
50	2.385	0.4%	0.3%	0.4%
55	2.309	2.1%	1.9%	2.1%
Average	2.226	0.2%	0.1%	0.2%

**Fig. 4** Percentage increase in life course mean hourly occupational earnings for a one unit (standard deviation) increase in PSES16—artificial missingness example. Data sources are NCDS (1974–2013) and ONS CPI (1997–2013). Single column figure showing percentage change life course occupational earnings for a one unit (standard deviation) increase in PSES16

predictions, and the other results are displayed as predicted percentage differences compared to the CC values. Relative to CC, the life course predicted percentage differences across MAR, MLMI and MICE model results proved to be similar with the MLMI results appearing closest, on average, to recapturing the CC results mainly due to less overestimation at age 55.

Next, we examined how this artificial missing data example impacted upon our main predictor of interest. Figure 4 tracks the effect of a one standard deviation increase in PSES16 on life course MHOE for the four scenarios. We refer the reader to Table C1 within Appendix C in the supplementary material for the fixed effects coefficients that underpin the data used in this figure. Figure 4 shows the two

Table 4 Subject specific life course log mean hourly occupational earnings multilevel model predictions—empirical missingness example

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%

Data sources are NCDS (1974–2013) and ONS CPI (1997–2013). The table displays life course multilevel model subject specific (fixed+random effects) MHOE predictions for available cases and percentage differences for other approaches (CC, PO, MLMI and MICE)

MI curves staying closer to the artificially created MAR curve over the life course, but there is evidence of convergence towards the CC ground truth curve from age 46 onwards. This convergence occurred with double the amount of missing MHOE data for ages 46, 50 and 55 compared to ages 23, 33 and 42. We refer the reader to Appendix C in the supplementary material for a description of how we created the artificial missingness via the MAR dataset to initially test the two MI methods.

3.5 Comparing the different approaches—empirical missingness example

Overall, the two MI methods displayed similar disparities in relation to the ground truth. We next present their performance in relation to the bigger empirical life course investigation. Table 4 displays a comparison of average life course step function ML model predictions. Similar to Table 3, the AC results are presented 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 that percentage difference between CC and PO was just over 2% with the gap between the groups narrowing over the life course. This narrowing over the life course suggested some evidence of convergence between the two groups in terms of the degree of similarity between the cohort members who return to the NCDS (i.e. PO) and the cohort members who never left the NCDS (i.e. CC). By contrast, the average life course predicted percentage difference between CC and PO was 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, appeared to be negligible. Given the underlying data and model of interest, the evidence suggested that these two MI methods produced 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. This evidence suggested the AC analysis increasingly overestimated MHOE as age increases and survey sample size decreases.

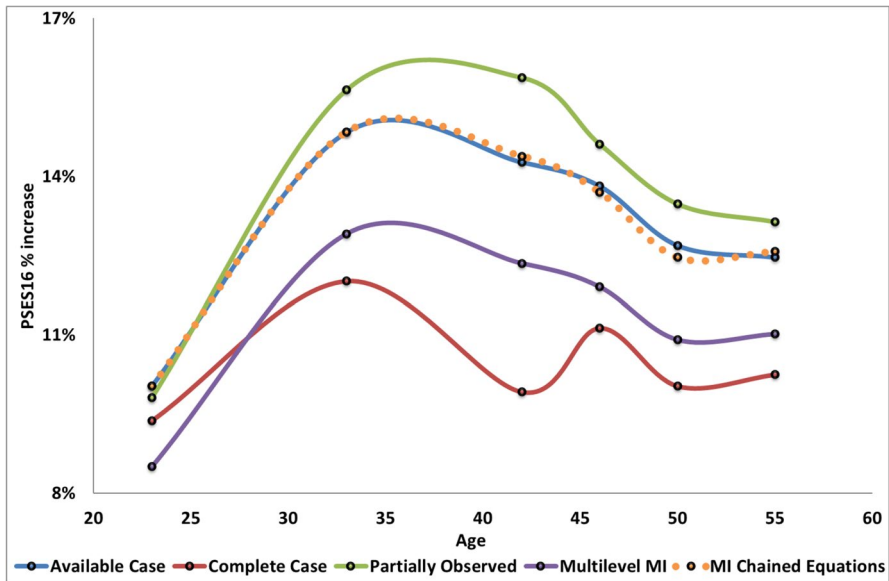


Fig. 5 Percentage increase in life course mean hourly occupational earnings for a one unit increase in socioeconomic background—empirical missingness example. Data sources are NCDS (1974–2013) and ONS CPI (1997–2013). Single column figure showing percentage change life course occupational earnings for a one unit (standard deviation) increase in PSES16

Focusing specifically on our main predictor of interest, Fig. 5 tracks the effect of a one standard deviation increase in PSES16 on life course MHOE for the five scenarios. We refer the reader to Table C3 within Appendix C in the supplementary material for the fixed effects coefficients that underpin the data used in this figure. The biggest contrast was between PO and CC, and interestingly, the MLMI results appeared to deviate from the AC/MICE results. The MICE results here appeared to add nothing to the AC results, but the MLMI results reflected the fact that the interaction between age and PSES16 could be factored into the MI process unlike the MICE approach. The CC analysis appeared to underestimate the interaction effects of parental SES at age 16 on occupational earnings later on in the life course. Whereas increments in PSES16 for the PO subgroup appeared to provide a bigger boost to later life MHOE attainment by comparison. This evidence indicated inferential divergence across the five scenarios with respect to the effect of parental SES at age 16 on occupational earnings over the life course.

4 Discussion

We found employing a step function multilevel life course model most suitable for exploring the changes in adult occupational earnings 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

occupational earnings in adulthood by introducing cross-level interaction dummy variables between parental SES and age over the life course.

With this model, we were able to examine the between-individual variation in SES across the life course. This person-level variance analysis identified the general divergent pattern between individuals whereby larger deviations from the occupational earnings average at a preceding age are positively correlated with larger deviations from the MHOE average at a subsequent age. This evidence suggests that positive momentum is greatest with those survey cohort members who are already more advantaged compared to those who are already less advantaged. In a study about income inequality and intergenerational mobility, Jerrim and Macmillan (2015) reported that advantage is transmitted across the generations with the help of unequal access to financial resources. In this study, we have shown that unequal socioeconomic backgrounds is an important predictor with respect to later life occupational earnings attainment adjusting for sex at birth and region of residence. Moreover, both these controls also demonstrated significant differential effects in relation to life course occupational earnings.

This study investigated whether we observe similar patterns of associations between childhood SES and adult occupational earnings using different methods of compensating for missing data. The complete case analysis underestimated the effect of parental SES on occupational earnings later on in the life course in comparison with the other methods of handling missing data. Moreover, the two multiple imputation approaches to compensating for longitudinal missing data based on wide versus long panel data indicated that the available case approach increasingly overestimates socioeconomic attainment as age increases and survey sample size decreases.

We compared the computationally faster MICE approach to the approach with a higher degree of similarity between the imputation model and analytical model, MLMI, given the structure of our data and model of interest. The MICE method took less time computationally and was found by Romaniuk et al. (2014) to produce more plausible results than the multivariate normal imputation procedure. We found that the MICE method produced similarly plausible empirical results compared with the computationally slower MLMI method. The biggest difference between the two approaches was the effect of socioeconomic background on occupational earnings particularly at age 23. Moreover, both sets of results were more similar to AC results compared with CC and PO results on average. White and Carlin (2010) found FIML to be similar to MI when the analysis explicitly models the missing covariates. Our CC results represented the situation whereby missing data only affects the model of interest covariates but these results are biased upwards in terms of life course occupational earnings. Therefore, the AC, MLMI and MICE results are more representative of the NCDS sample and the target population.

The evidence presented in this study suggests the FIML approach using all available cases and our step function multilevel life course model is more similar to the MI approaches than the CC and PO approaches. Likewise, the artificial missing data example displayed the MI approaches being more similar to the MAR results than the CC results. This amounted to a robustness check for the model results and inferences similar to Maharani and 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 model of interest 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 et al. (2001).

However, this MAR assumption cannot be empirically verified so one way to strengthen the validity of this assumption would have been to assume that the missing data were MNAR and develop an appropriate statistical model process that can accommodate our step function life course multilevel model of interest. Comparing results from both MAR and MNAR processes would have helped evaluate if the missing data did depend on the unobserved data given the observed data. However, we were not able to conduct an MNAR analysis that could handle the number of non-response patterns that existed within the NCDS cohort over the life course due to computational difficulties. Failure to undertake a complementary MNAR analysis represents a limitation of this study.

Goldstein et al. (2014) have addressed the well-known biases that can arise from omitting model of interest 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 statistical software; 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 model of interest and underlying data. Recent work by Goldstein et al. (2014) shows promise.

In relation to omitted variable bias, another limitation of this study was not having the capability to check for possible omitted confounding variable bias. Omission of such a confounder may have biased the estimates derived for parental socioeconomic status. This potential bias may have led us to overestimate the effect accruing to parental socioeconomic status on life course occupational earnings. Furthermore, a related aspect to this was repeating the analysis presented in this study on other birth cohorts to ensure the effects identified in this paper are not simply a manifestation of the NCDS cohort, i.e. validation through age, period and cohort controls. However, by not conducting this cross-cohort comparison to validate the results uncovered by this research represents a further limitation of this study.

5 Conclusion

Using 1958 NCDS birth cohort longitudinal survey data, we empirically evaluated the intergenerational transmission of SES between individuals over the life course in Great Britain. The key finding from this research was that there was no evidence of the occupational earnings gap narrowing in adulthood and parental SES up to the age of 16 only serves to widen such a gap over the life course. The implication is an intergenerational cumulative process whereby those with more socioeconomic advantages in early life start from a much stronger position which allows them access to much greater opportunities resulting in being relatively better off in adulthood compared to those without the same initial advantages.

In terms of missing data, restricting ourselves to complete cases led to underestimating the effect of socioeconomic background on later life occupational earnings attainment. Addressing the issue artificially revealed that the MI approaches produced results that were more similar to the artificially created available case results than the complete case results. This finding carried over to the empirical setting with the better aligned, in terms of the substantive model of interest, MLMI approach producing similar results to the computationally faster but worse aligned MICE approach.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s43253-022-00081-8>.

Funding Dr. Adrian Byrne: School of Social Sciences PhD Studentship, University of Manchester. Professor Natalie Shlomo: Economic and Social Research Council grant number ES/L008351/1. Professor Tarani Chandola: Economic and Social Research Council grant number ES/J019119/1.

Data availability National Child Development Study 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.

Code availability Software used: Stata/SE 13.1, MLwiN v2.32 and REALCOM-Impute.

This research did not involve human participants or animals so no informed consent or statement on animal welfare was required to undertake this research.

Declarations

Conflict of interest The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- American Psychological Association (2007) Task force on socioeconomic status. In Report of the APA task force on socioeconomic status, Washington, DC: American Psychological Association
- Asparouhov T, Muthén B (2020) Comparison of models for the analysis of intensive longitudinal data. *Struct Equ Model Multidiscip J* 27(2):275–297
- Bartlett J, and Carpenter J (2013) Missing data concepts. LEMMA VLE Module 14, 1–29. <http://www.bristol.ac.uk/cmm/learning/course.html>. Accessed 21 May 2018
- Bell A, Jones K (2015) Explaining fixed effects: random effects modeling of time-series cross-sectional and panel data. *Polit Sci Res Methods* 3(1):133–153
- Bell A, Fairbrother M, Jones K (2019) Fixed and random effects models: making an informed choice. *Qual Quant* 53(2):1051–1074

- Blanden J, Goodman A, Gregg P, Machin S (2001) Changes in intergenerational mobility in Britain. University of Bristol, Centre for Market and Public Organisation
- Bolger N, Laurenceau JP (2013) Intensive longitudinal methods: an introduction to diary and experience sampling research. Guilford Press Google Scholar, New York
- Bollen KA, Bauldry S (2011) Three Cs in measurement models: causal indicators, composite indicators, and covariates. *Psychol Methods* 16(3):265
- Bukodi E, Dex S, Goldthorpe JH (2011) The conceptualisation and measurement of occupational hierarchies: a review, a proposal and some illustrative analyses. *Qual Quant* 45(3):623–639
- Byrne A, Chandola T, Shlomo N (2018) How does parental social mobility during childhood affect socio-economic status over the life course? *Research in Social Stratification and Mobility* 58:69–79
- Caro DH, 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: Williams M, Vogt P (eds) *The SAGE handbook of Innovation in Social Research Methods*. Sage, New York, pp 498–523
- Charlton CMJ, Michaelides DT, Cameron B, Szmargd C, Parker RMA, Yang H, Zhang Z, Browne WJ (2012) Stat-JR software
- Clark TS, Linzer DA (2015) Should I use fixed or random effects? *Polit Sci Res Methods* 3(2):399–408
- Collins LM, Schafer JL, Kam C-M (2001) A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychol Methods* 6(4):330
- Enders CK (2010) *Applied missing data analysis*. Guilford Press
- Galobardes B, Shaw M, Lawlor DA, Lynch JW (2006a) Indicators of socioeconomic position (part 2). *J Epidemiol Community Health* 60(2):95
- Galobardes B, Shaw M, Lawlor DA, Lynch JW, Smith GD (2006b) Indicators of socioeconomic position (part 1). *J Epidemiol Community Health* 60(1):7–12
- Ganzeboom HBG, Treiman DJ (1996) Internationally comparable measures of occupational status for the 1988 international standard classification of occupations. *Soc Sci Res* 25(3):201–239
- Ganzeboom HBG, De Graaf PM, Treiman DJ (1992) A standard international socio-economic index of occupational status. *Soc Sci Res* 21(1):1–56
- Ganzeboom HB, Treiman DJ (2003) Three internationally standardised measures for comparative research on occupational status. *Advances in cross-national comparison*. Springer, Boston, pp 159–193
- Goldstein H, Carpenter J, Kenward MG, Levin KA (2009) Multilevel models with multivariate mixed response types. *Stat Model* 9(3):173–197
- Goldstein H, Carpenter JR, Browne WJ (2014) Fitting multilevel multivariate models with missing data in responses and covariates that may include interactions and non-linear terms. *J R Stat Soc A Stat Soc* 177(2):553–564
- Goldstein H (2009) Handling attrition and non-response in longitudinal data. *Longitudinal and Life Course Studies* 1(1)
- Goldthorpe JH, Jackson M (2007) Intergenerational class mobility in contemporary Britain: political concerns and empirical findings. *Br J Sociol* 58(4):525–546
- Gugushvili A, Bukodi E, Goldthorpe JH (2017) The direct effect of social origins on social mobility chances: 'glass floors' and 'glass ceilings' in Britain. *Eur Sociol Rev* 33(2):305–316
- Hawkes D, Plewis I (2006) Modelling non-response in the national child development study. *J R Stat Soc A Stat Soc* 169(3):479–491
- Jerrim J, Macmillan L (2015) Income inequality, intergenerational mobility, and the Great Gatsby Curve: Is education the key? *Soc Forces* 94(2):505–533
- Johnson JMMI, Brown M, Hancock M (2015) *National child development study: user guide to the response and deaths datasets*. Centre for Longitudinal Studies
- Jonsson JO, Grusky DB, Di Carlo M, Pollak R, Brinton MC (2009) Microclass mobility: social reproduction in four countries. *Am J Sociol* 114(4):977–1036
- 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
- Laurison D, Friedman S (2016) The class pay gap in higher professional and managerial occupations. *Am Sociol Rev* 81(4):668–695
- 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

- Maharani A, Tampubolon G (2014) Unmet needs for cardiovascular care in Indonesia. *PLoS ONE* 9(8):e105831
- Modigliani F, Brumberg R (1954) Utility analysis and the consumption function: An interpretation of cross-section data. *Post-Keynesian Economics* 1:338–436
- Nickell S (1982) The determinants of occupational success in Britain. *Rev Econ Stud* 49(1):43–53
- Niedzwiedz CL, Katikireddi SV, Pell JP, 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
- Nyblom M, Stuhler J (2016) Heterogeneous income profiles and lifecycle bias in intergenerational mobility estimation. *Journal of Human Resources* 51(1):239–268
- Nyblom M, Stuhler J (2017) Biases in standard measures of intergenerational income dependence. *Journal of Human Resources* 52(3):800–825
- Prandy K (1990) The revised Cambridge scale of occupations. *Sociology* 24(4):629–655
- Rasbash J, Steele F, Browne WJ, Goldstein H, Charlton C (2015) A user's guide to MLwiN. University of Bristol, UK, Centre for Multilevel Modelling
- Romaniuk H, Patton GC, Carlin JB (2014) Multiple imputation in a longitudinal cohort study: a case study of sensitivity to imputation methods. *Am J Epidemiol* 180(9):920–932
- Rubin DB (1987) *Multiple Imputation for Nonresponse in Surveys* (Wiley Series in Probability and Statistics)
- Sobek M (1995) The comparability of occupations and the generation of income scores. *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 28(1):47–51
- Steele F (2008) Multilevel models for longitudinal data. *J R Stat Soc A Stat Soc* 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>. Accessed 21 May 2018
- Stewart A, Prandy K, Blackburn RM (1980) *Social stratification and occupations*. Springer
- Sturgis P, Sullivan L (2008) Exploring social mobility with latent trajectory groups. *J R Stat Soc A Stat Soc* 171(1):65–88
- University of London. Institute of Education. Centre for Longitudinal Studies. 2012. National Child Development Study: Sweep 8, 2008–2009. UK Data Service
- University of London. Institute of Education. Centre for Longitudinal Studies. 2014. National Child Development Study: Childhood Data, Sweeps 0–3, 1958–1974. UK Data Service.
- University of London. Institute of Education. Centre for Longitudinal Studies. 2015. National Child Development Study: Sweep 9, 2013. UK Data Service.
- University of London. Institute of Education. Centre for Longitudinal Studies. 2008a. National Child Development Study: Sweep 4, 1981, and Public Examination Results, 1978. UK Data Service
- University of London. Institute of Education. Centre for Longitudinal Studies. 2008b. National Child Development Study: Sweep 5, 1991. UK Data Service
- University of London. Institute of Education. Centre for Longitudinal Studies. 2008c. National Child Development Study: Sweep 6, 1999–2000. UK Data Service
- University of London. Institute of Education. Centre for Longitudinal Studies. 2008d. National Child Development Study: Sweep 7, 2004–2005. UK Data Service
- Van Buuren S, Boshuizen HC, Knook DL (1999) Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in medicine* 18(6):681–694
- White IR, Carlin JB (2010) Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Stat Med* 29(28):2920–2931
- White IR, Royston P, Wood AM (2011) Multiple imputation using chained equations: issues and guidance for practice. *Stat Med* 30(4):377–399
- Young R, Johnson DR (2015) Handling missing values in longitudinal panel data with multiple imputation. *J Marriage Fam* 77(1):277–294