A multilevel SEM for the interrelationships between multiple latent dimensions of childhood socioeconomic circumstances, partnership transitions and midlife health

Yajing Zhu

Department of Statistics, Houghton Street, London School of Economics and Political Science, London, U.K.

E-mail: y.zhu18@lse.ac.uk

Fiona Steele

Department of Statistics, Houghton Street, London School of Economics and Political Science, London, U.K.

E-mail: f.a.steele@lse.ac.uk

Irini Moustaki

Department of Statistics, Houghton Street, London School of Economics and Political Science, London, U.K.

E-mail: i.moustaki@lse.ac.uk

Summary. We propose a multilevel structural equation model (SEM) to investigate the interrelationships between childhood socioeconomic circumstances (SECs), partnership formation and stability, and midlife health, using data from the 1958 British birth cohort. The SEM comprises latent class models that characterize the patterns of change in four dimensions of childhood SECs and a joint regression model that relates these categorical latent variables to partnership transitions in adulthood and midlife health, while allowing for informative dropout. The model can be extended to handle multiple outcomes of mixed types and at different levels in a hierarchical data structure.

Keywords: event history analysis, non-ignorable dropout, latent variable model, multilevel model, 3-step approach

1. Introduction

In recent years, there has been growing interest in the sources of social and health inequalities. A better understanding of life course influences can be particularly beneficial to policies that aim to reduce inequalities due to demographic and socioeconomic circumstances (SECs) in early life (Bartley, 2016). With the availability of rich data collected from large-scale longitudinal (in particular, birth cohort) studies, recent methodological developments in life-course epidemiology have been made to test the influence of environmental risk factors on health outcomes at different life stages (see Ben-Shlomo et al. (2016) for a comprehensive review and the references therein). Aside from genetic reasons, previous research has found that disadvantage in multiple aspects of childhood, such as poor parental socioeconomic circumstances, housing conditions, family financial difficulties and parental relationship stability, is associated with poor physical (e.g. cardiovascular disease) and mental health in later life (e.g. Poulton et al., 2002). There is also interest in the effects of cumulative exposure to unfavourable life experiences such as partnership dissolution on later health (Ploubidis et al., 2015; Arpino et al., 2018; O'Flaherty et al., 2016). Some recent studies have contributed to uncovering behavioural, physical and psycho-social pathways that mediate the effects of childhood SECs on midlife or later health (e.g. Ben-Shlomo et al., 2016). It has been hypothesized that experiences in childhood operate through influencing trajectories of life course events and functional changes in health-related behaviours that can partly explain the apparently persistent effects of childhood SECs on later health. This paper considers the influence of one of these social processes: the experiences of partnership formation and dissolution up to midlife.

Many studies have found that family characteristics during childhood have a significant impact on the experience of partnership transitions. For example, individuals from families of a relatively high social position tend to delay forming the first partnership (Berrington and Diamond, 2000; Steele et al., 2006) while individuals who have experienced parental separation during childhood are at increased risk of partnership dissolution themselves. Moreover, previous research has found a positive association between partnership instability and poor later health (Brewer and Nandi, 2014; O'Flaherty et al., 2016; Ploubidis et al., 2015). However, little is known about the interrelationships between childhood characteristics, partnership transitions and health outcomes and this requires bridging two streams of research discussed above. A recent progress was made by Arpino et al. (2018) using a mediation model in which educational attainment and life course trajectories (fertility, partnership, employment) are treated as mediators for the effect of childhood characteristics on later health (but without discussing causal assumptions). We consider the same broad substantive research question, but within a flexible joint SEM framework.

We use rich life course data collected in the National Child Development Study (NCDS) of Great Britain, including histories of partnership events collected for ages 16 to 50. To investigate the extent to which the effects of childhood SECs on midlife health are explained by partnership experiences, a number of methodological challenges need to be addressed. First, multiple aspects of childhood SECs are of substantive interest, and information from repeated measurements over four childhood waves (ages 0, 7, 11 and 16) must be combined while accounting for missing data. Second, most previous research has used simple forms of structural equation models (SEMs) with only observed variables (including multiple mediators and confounders) to explore the pathways linking childhood SECs to later health. Few studies have additionally introduced latent variables to capture trajectories of development while accounting for measurement error (Hagger-Johnson et al., 2011; Ploubidis et al., 2015; Arpino et al., 2018). In terms of the consideration of childhood SECs, previous research has used observed family characteristics at single or multiple time points (Kelly-Irving et al., 2013; Maggs et al., 2008), ad-hoc composite summaries (Hagger-Johnson et al., 2011; Kelly-Irving et al., 2013) or has derived latent summaries (Ploubidis et al., 2015). Simple inclusion of highly correlated repeated measures into the regression model ignores measurement error and can lead to inflated standard errors. Having a broad summary of childhood characteristics is not ideal, especially if it summarises information across various aspects and time. To capture life course partnership transitions, a latent variable approach (Green et al., 2012; Hagger-Johnson et al., 2011) and a combination of sequence analysis and cluster analysis (Ploubidis et al., 2015; Arpino et al., 2018) have been adopted. The derived summary of partnership trajectories is often treated as an observed variable which is problematic because it ignores classification error in assigning individuals to latent classes or clusters. These approaches also make limited use of the rich event history data available on the timing and development of life events over long follow-ups in adulthood. Studies that have used event history data have been primarily interested in identifying risk factors of time-to-event outcomes rather than relating the information captured by event histories to distal outcomes such as later health (e.g. Steele et al., 2006). Finally, attrition is a common issue in cohort studies and the reasons for dropout are likely to be associated with the outcomes of interest. To address these challenges simultaneously, the aim of this paper is to develop a joint modelling framework which makes better use of the available longitudinal and event history data collected at different life stages.

We propose a multilevel joint SEM with latent variables that can incorporate information from different life stages and domains and capture the dependencies between childhood socioeconomic experiences, partnership transitions in adulthood and midlife health. The main contributions and advantages of our model can be summarised as follows. First, the 3-step method of Zhu et al. (2017) for estimating the effects of multiple associated categorical latent variables on a distal outcome is generalised to handle mixed outcomes measured at different levels in a hierarchical structure. The method is applied in an analysis of the effects of latent variables for four dimensions of childhood SECs on a binary health outcome and on the durations between recurrent partnership transitions. Second, partnership histories are used to define not only outcomes that are predicted by childhood SECs but predictors of later health, by deriving summary variables of partnership stability for ages 16-50 (e.g. the total number of partners and percentage of time spent single). Third, the effects of childhood SECs on partnership formation and dissolution hazards for ages 16-50 and distal health at age 50 are jointly modelled by allowing for a residual association across equations via a continuous latent variable capturing unobserved time-invariant characteristics with differential effects on each response. This residual association helps to 1) mitigate the problem of endogeneity due to correlation between individual-specific unobservables and the summaries of partnership events included as predictors in the health model, 2) further allow for residual association between partnership stability and midlife health, in addition to that captured by the summary variables, and 3) account for possible additional dependence between partnership transitions and midlife health given the latent classes (see Zhu et al., 2017).

The proposed joint SEM belongs to the general latent variable modelling frameworks of Skrondal and Rabe-Hesketh (2004), and hence has the same generalizability. More specifically, life events (e.g. partnership transitions), distal outcomes (e.g. midlife health) and dropout indicators can all be viewed as indicators of one or more individual-level latent variables. This makes it possible to estimate the coefficients of the random effects in the joint model, especially as the distal outcome is not measured across time. Different specifications of the structural model can accommodate various research questions where the underlying association of individual-level variables (rather than their manifestations with measurement errors) is of primary interest. Future extensions of our model could address questions related to the dynamics between repeated measures of earlier health and life events over the life course, and to better understand the time-varying influence of childhood SECs on the longitudinal development of these quantities across life stages.

The rest of the paper is organised as follows. Section 2 discusses modelling approaches, including a review of previous research and our proposed extensions. Section 3 gives an overview of the NCDS and the event history data and other measures that are used in the analysis. The proposed methods are applied to these data to better understand the interrelationships between childhood SECs, partnership experiences and midlife health in Section 4. Finally, some implications for future work are discussed in Section 5.

2. Methodology

2.1. Categorical latent variables for childhood SECs

A large number of indicators measuring multiple dimensions of childhood SECs are available in the childhood waves of the NCDS. Simple treatments of childhood SECs in previous studies have several limitations that we outline below. First, including covariates relating to only one or two aspects of childhood circumstances that are measured at one or two time points provide a limited summary of a family's SECs in childhood, especially when repeated measures are available. Second, treating repeated measures of childhood situations at different time points as separate predictors in a regression model can lead to inflated standard errors as these predictors are likely to be highly correlated. Also, it is inevitable that there are missing values in at least one childhood measure, and naive treatment of missing data such as listwise deletion may lead to biased estimates of the effects of these variables on the outcomes of interest. Latent class models are particularly useful in such situations as they summarise patterns of the change in values of repeated measures over time with minimal loss of information (see Hagenaars and McCutcheon, 2002).

In this study, we identify four aspects of childhood SECs that are commonly considered in the literature as risk factors of poor well-being in later life: parental social status, financial difficulty, material difficulty and parental relationship breakdown. We consider financial difficulty (a proxy for economic deprivation) and material hardship (i.e. housing deprivation, a proxy for wealth) to be associated but conceptually distinct dimensions that contribute to the complexity of family SECs (e.g. Schoon et al., 2003; Shavers, 2007; Bartley et al., 2012). At each childhood wave, we construct a composite index (either ordinal or nominal) for each dimension using a set of correlated measurements, and guided by previous research (e.g. Bartley, 2003; Hobcraft, 1998). Each set of indices then forms repeated measures of an aspect of childhood SECs at ages 0, 7, 11 and 16 which measure four categorical latent variables (see section 3.2 for details). Denote by C_q the categorical latent variable for SEC dimension q with K_q categories ($q = 1, ..., 4$), measured by the 4×1 vector $\mathbf{Y}_q^{(C)}$ containing the corresponding composite index for each of the four childhood waves. We next fit a separate latent class model to each $\mathbf{Y}_q^{(C)}$ and derive four categorical latent variables, capturing potentially associated aspects of childhood SECs. The latent class models are estimated using full information maximum likelihood which allows individuals with missing information at any wave to be retained under a missing at random assumption.

2.2. Relating latent childhood SEC variables to midlife health: the 3-step approach

The next step is to relate the latent variables for childhood SECs to distal outcomes, in particular those that are temporally distal such as midlife health. Commonly used approaches are the 1-step approach and step-wise approaches. The 1-step approach estimates the measurement model for the latent variable and the regression model for the outcomes of interest simultaneously. As outcomes are essentially regarded as additional indicators for the latent variable, they influence the measurement model such that the latent variable that is assumed to predict the outcome is also partly determined by it (e.g. Vermunt, 2010; Asparouhov and Muthén, 2014). To avoid this circularity, various stepwise approaches have been proposed in which the measurement model is estimated separately from the regression model; examples include the modal class, pseudo-class, the modified Bolck-Croon-Hagenaars (BCH) (Bolck et al., 2004; Vermunt, 2010), the Bayes (Lanza et al., 2013), the 3-step maximum likelihood (Vermunt, 2010) and the two-step (Bakk and Kuha, 2018) approaches. Extensive discussions of the benefits and limitations of each method for models with one latent variable are available in Asparouhov and Muthén (2014), Bakk and Vermunt (2016) and Bakk and Kuha (2018). When the model assumptions hold, among these approaches, the 3-step maximum likelihood estimator is efficient, produces unbiased estimates for various outcome types and is highly flexible.

In a recent study, Zhu et al. (2017) extended the 3-step maximum likelihood approach to models where a distal outcome is predicted by multiple and possibly associated categorical latent variables. We begin with a review of this method as it is a major component of the multilevel joint SEM proposed in Section 2.3. For illustration purposes, consider a regression model where the distal outcome H (midlife health) is predicted by two associated categorical latent variables for childhood SECs, C_1 and C_2, and a covariate X (e.g. health-related behaviours, such as smoking). Note that C_1 and C_2 may be associated.

In steps 1 and 2, separate latent class models are fitted to each measurement vector $\mathbf{Y}_q^{(C)}$ ($q = 1, 2$), without making any assumptions about whether or not the two latent classes are related at this stage. Based on the patterns of change in each composite index across childhood waves, individuals are assigned to the most likely (modal) class M_q based on the posterior probabilities of class membership. These modal classes are regarded as imperfect measurements of true classes (Cs). The misclassification probabilities $P(M_q|C_q)$ are computed (see details in Bakk et al. (2013), Bakk and Vermunt (2016) and Zhu et al. (2017)) and, together with the class memberships, are fixed in step 3 where the latent variables and X are related to distal health H. This specification also allows for the association between latent variables via a flexible log-linear model (see Zhu et al. 2017). Denoting the parameter vector by $\boldsymbol{\theta}$, the log-likelihood of the observed data for a sample of N observations is

$$\begin{aligned} l(\boldsymbol{\theta}) &= \sum_{i=1}^{N} \log P(H_i, M_{1i}, M_{2i}|X_i; \boldsymbol{\theta}) \qquad (1) \\ &= \sum_{i=1}^{N} \log \sum_{k_2=1}^{K_2-1} \sum_{k_1=1}^{K_1-1} P(H_i, M_{1i}, M_{2i}|X_i, C_{1i}=k_1, C_{2i}=k_2) \times P(C_{1i}=k_1, C_{2i}=k_2), \end{aligned}$$

where

$$\begin{aligned} P(H_i, M_{1i}, M_{2i}|X_i, C_{1i}=k_1, C_{2i}=k_2) =& P(H_i|X_i, C_{1i}=k_1, C_{2i}=k_2) \qquad (2) \\ &\times P(M_{1i}|C_{1i}=k_1) \\ &\times P(M_{2i}|C_{2i}=k_2). \end{aligned}$$

Our interest lies in $P(H_i|X_i, C_{1i} = k_1, C_{2i} = k_2)$. Key model assumptions for step 3 of the 3-step approach are that the distal outcome H and the indicators $\mathbf{Y}_q^{(C)}$ are conditionally independent given C_1 and C_2, and that the functional form of the model is correctly specified (e.g. as a logit-linear model for a binary outcome). Simulation studies of Zhu et al. (2017) show that when all model assumptions are satisfied the performance of the most efficient 1-step approach and the 3-step ML approach is similar across various combinations of class separation quality (entropy) and sample sizes (consistent with the findings of Bakk and Vermunt 2016). The 3-step approach performs poorly in situations with poor class separation (low entropy), small sample size ($N < 500$) and the presence of local dependence between the indicators and the distal outcome given the latent classes.

2.3. *Extensions of the 3-step approach to a multilevel SEM*

We extend the model in Section 2.2 to a multilevel joint SEM where the four latent variables C_q for childhood SECs predict partnership transitions during ages 16-50 and midlife health, and outcomes of the partnership transitions (as captured by summary variables) can predict health. Our proposed model has links with the joint modelling approach commonly used in biostatistics where repeated measures of bio-markers of health are jointly modelled with a survival outcome (usually mortality). The joint modelling approach has been advocated due to its flexibility in modelling associated longitudinal and duration data (e.g. Tsiatis and Davidian, 2004). The joint model consists of two submodels, one for each of the survival and longitudinal outcomes, that are linked through shared parameters (i.e. in the mean and residual structure of the joint model). Joint models have also been widely applied in social research where they are often referred to as multiprocess models. One of the earliest examples is Lillard et al. (1995) who specified a joint model for pre-marital cohabitation (binary, repeated measures) and the time to marital dissolution, allowing for residual correlation between the two submodels to test for selection of individuals with a higher-than-average risk of divorce into pre-marital cohabitation.

With the availability of 34-year partnership histories in the NCDS, incorporating event history models into our joint SEM is particularly attractive. On the one hand, we can specify flexible multilevel event history models to understand the impacts of childhood SECs on potentially recurrent time-to-event outcomes (formation and dissolution events). On the other hand, relating multiple derived summaries of the event history to the midlife health outcome provides insights on the cumulative effects of partnership transitions between ages 16 and 50 on later health. Therefore, our key modelling interest is to treat partnership stability both as an outcome in the childhood SECs-partnership stability relationship, but also as a predictor in the partnership stability-midlife health relationship. This can be achieved via a joint model of partnership transitions and midlife health. Another motivation for a joint modelling approach is that the summaries of partnership events included as predictors of health are potentially endogenous: partnership transitions and health may share unmeasured influences. Assuming that such shared influences are time-invariant, normally distributed and independent of the covariates other than partnership history summaries, we introduce to the model a set of individual-specific random effects $u_i \sim N(0, \sigma_u^2)$, which affect all outcomes.

Extending the 3-step ML approach to estimation of a multilevel SEM with categorical latent variables requires the measurement models for latent dimensions of childhood SECs to be estimated separately from the rest of the model. In steps 1 and 2, we fit separate latent class models to each measurement vector $\mathbf{Y}_q^{(C)}, (q = 1, ..., 4)$ and save the modal classes and misclassification probabilities. These quantities are inputs in the estimation of the SEM at step 3 where the latent childhood SECs C_q are related to time-to-event outcomes and midlife health. In step 3, we estimate a joint model for partnership transitions and midlife health where C_q is included as a predictor in each submodel.

2.3.1. Submodels for partnership formation and dissolution

The NCDS partnership histories were derived from the information collected in 1991 (age 33), 2000 (age 42), 2004 (age 46) and 2008 (age 50), with respondents asked to recall the start and end date of coresidential relationships to the nearest year and month. It is therefore natural to use the discrete-time approach to event history analysis (Allison, 1984). We focus on the time to first partnership formation and time to (possibly recurrent) partnership dissolutions. The data have a hierarchical structure with partnership episodes at level 1, nested within individuals at level 2. We allow for unobserved individual heterogeneity (also known as shared frailty) in the hazards of formation and separation by including an individual-specific random effect which allows for time-invariant unmeasured influences shared by all episodes from the same person. Discrete-time random effects models have been often used to model recurrent events in the social sciences (see Barber et al. (2000) and Steele (2011) for further details). A joint model of the formation and dissolution processes is considered because a key predictor in the dissolution model, age at the start of the partnership, is highly related to the outcome of the partnership formation model (i.e. age at the start of first partnership), and may thus be endogenous. We allow for a residual association between the formation and dissolution processes by including the individual-level random effect u_i in both models.

To prepare for the discrete-time event history analysis, the duration to first partnership formation of individual i is expanded to S_i records indexed by s and, for the dissolution model, the duration of each partnership episode j ($j = 1, ..., J_i$) is expanded to T_{ij} records indexed by t. Denote by $y_i^{(F)}$ and $y_{ij}^{(S)}$ the event or censored time for formation and separation events. We define the corresponding binary responses $y_{si}^{(F)}$ and $y_{tij}^{(S)}$ such that

$$y_{si}^{(F)} = \begin{cases} 1 & y_i^{(F)} = s, \text{uncensored} \\ 0 & y_i^{(F)} = s, \text{censored} \\ 0 & y_i^{(F)} > s. \end{cases} \qquad y_{tij}^{(S)} = \begin{cases} 1 & y_{ij}^{(S)} = t, \text{uncensored} \\ 0 & y_{ij}^{(S)} = t, \text{censored} \\ 0 & y_{ij}^{(S)} > t. \end{cases}$$

Denote by $h_{si}^{(F)} = P(y_{si}^{(F)} = 1 | y_{s'<s}^{(F)} = 0)$ the hazard of the first entry into partnership in the time interval $[s, s+1)$ and $h_{tij}^{S} = P(y_{tij}^{(S)} = 1 | y_{t'<t,ij}^{(S)} = 0)$ the hazard of separation in time interval $[t, t+1)$. Given the long observation period, the monthly duration data are aggregated to six-month intervals to reduce the size of the discrete-time dataset. The likelihood function for this discrete-time model is equivalent to that for a model for multilevel binary data where $y_{si}^{(F)}$ and $y_{tij}^{(S)}$ follow binomial distributions with denominator $n_{si}^{(F)}$ and $n_{tij}^{(S)}$ equal to the exposure time in each time interval, respectively (Steele et al., 2005).

We specify a generalized linear model for the binomial responses using a logit link function:

$$\begin{aligned} \text{logit}\left(h_{si}^{(F)}\right) &= \alpha_s^{(F)} + \boldsymbol{\beta}^{(F)\prime}\mathbf{X}_{si}^{(F)} + \sum_{q=1}^{4}\sum_{k_q=1}^{K_q-1} \tau_{C_q,k_q}^{(F)} I(C_{qi} = k_q) + \lambda^{(F)} u_i \\ \text{logit}\left(h_{tij}^{(S)}\right) &= \alpha_t^{(S)} + \boldsymbol{\beta}^{(S)\prime}\mathbf{X}_{tij}^{(S)} + \sum_{q=1}^{4}\sum_{k_q=1}^{K_q-1} \tau_{C_q,k_q}^{(S)} I(C_{qi} = k_q) + \lambda^{(S)} u_i, \end{aligned} \tag{3}$$

where $\alpha_s^{(F)}$ and $\alpha_t^{(S)}$ are functions of time which specify the baseline hazard (assumed to be piecewise constant here), the $\mathbf{X}$s are event-specific vectors of predictors with coefficients $\boldsymbol{\beta}$, $I(C_{qi} = k_q)$ is a dummy variable for class k of latent childhood SEC variable q, with coefficient τ_{C_q,k_q} (taking category K_q as the reference category), and $\lambda^{(F)}$ and $\lambda^{(S)}$ are coefficients of the individual random effects u_i. The implicit assumption for this model is that both event hazards are influenced by a common set of individual-specific unobservables but differential effects of u_i are allowed through the λ parameters. Equation (3) is similar to the multiprocess model (without latent categorical variables) that has been discussed in previous research to model correlated processes with endogenous predictors (e.g. Steele et al., 2005; Aassve et al., 2006a).

2.3.2. Submodel for midlife health

We consider a scenario where the distal health outcome is the endpoint that is collected later than the event histories and only once per individual. The aim is to relate latent childhood SECs and summaries of partnership experience to midlife health. For simplicity, we consider a single health outcome denoted by H_i. The extension to more than one health outcome is straightforward, and involves constructing a multivariate response vector.

Denote by $\mathbf{X}_i^{(H)}$ a vector of exogenous health-related predictors and $\mathbf{Z}_i^{(P)}$ a vector of summary measures of partnership stability. Examples of such summaries include the total number of partners, percentage of time spent single, the latest partnership status or the longitudinal profile of partnership transitions (Ploubidis et al., 2015). For an outcome following an exponential family distribution, a generalized linear model can be specified for $E(H_i|C_1,\ldots,C_4,\mathbf{X}_i^{(H)},\mathbf{Z}_i^{(P)},u_i)$ with appropriate link function. The self-reported health status in midlife is typically measured on a binary or ordinal scale. Suppose that we have a binary measure (coded 1 for poor health), then a logit model for $P_i^{(H)} = P(H_i = 1)$ can be written:

$$\text{logit}\left(P_i^{(H)}\right) = \alpha_0^{(H)} + \boldsymbol{\beta}_{\mathbf{1}}^{(H)\prime}\mathbf{Z}_i^{(P)} + \boldsymbol{\beta}_{\mathbf{2}}^{(H)\prime}\mathbf{X}_i^{(H)} + \sum_{q=1}^{4}\sum_{k_q=1}^{K_q-1}\tau_{C_q,k_q}^{(H)} I(C_{qi}=k_q) + \lambda^{(H)}u_i, \quad (4)$$

where $\lambda^{(H)}$ is the effect of the individual random effect u_i on the log-odds of poor health.

In the submodels, we have assumed that childhood SECs are exogenous with respect to partnership transitions during adulthood and midlife health because childhood circumstances reflect characteristics of the parents rather than the individual. However, $\mathbf{Z}_i^{(P)}$ is endogenous if there are shared unmeasured influences on partnership transitions and health. If these unmeasured variables are time-invariant, there will be a correlation between $\mathbf{Z}_i^{(P)}$ and u_i. Suppose, for example, that individuals prone to poor health tend also to have less stable relationships and that $Z_i^{(P)}$ is a measure of partnership instability. In that case we expect $\text{corr}(Z_i^{(P)}, u_i) > 0$ and, if ignored, a positive $\beta_1^{(H)}$ will be overstated and the effects of variables that are associated with $Z_i^{(P)}$ (e.g. childhood circumstances) may also be biased. To handle endogeneity due to shared unmeasured individual-level influences, we allow for non-zero residual correlations across equations. This multiprocess modelling approach has been widely used (e.g. Lillard et al., 1995; Steele et al., 2005; Aassve and Billari, 2006). A more flexible random effects structure was initially considered to capture residual association across processes, where different random effect terms were included in each equation, and assumed to follow a multivariate normal distribution. However, convergence was not achieved for this model, most likely due to its complexity and the lack of repeated events for first partnership formation and having one health outcome. This led to the simplified specification of Equations (3) and (4) with the same u_i across equations, but with a different coefficient for each outcome.

Our specification has several advantages. First, the joint model defined by Equations (3) and (4) can be viewed as a factor model where the time-to-event outcomes and distal health are essentially indicators for the latent variable u_i. The loading for u_i in the dissolution submodel, $\lambda^{(S)}$, is fixed at one for scaling and identification purposes while the remaining loadings, $\lambda^{(F)}$ and $\lambda^{(H)}$, are freely estimated. Second, the interpretation of $\lambda^{(H)}$ is informative. If high values of H_i indicate poor health, then $\lambda^{(H)} > 0$ suggests that people whose unobserved time-invariant characteristics put them at high risk of dissolution tend also to have poorer health in later life. Finally, Zhu

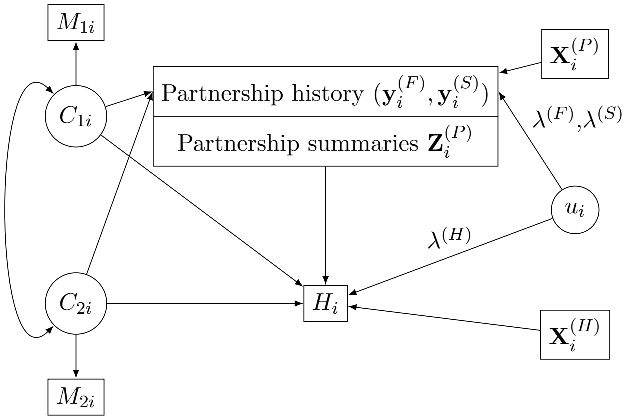

Fig. 1. A path diagram for a simplification of the multilevel SEM of Equations (3) and (4), with a factor structure for the individual-level random effects. C_1 and C_2 are two of the latent variables for childhood SECs, M_1 and M_2 are the most likely class memberships derived in step 1 for each dimension of the childhood SECs and H_i is midlife health. $\mathbf{X}_i^{(P)}$ and $\mathbf{X}_i^{(H)}$ denote the covariates for partnership events and distal health respectively. The upper half of the node with stacked rectangles is the outcome variable in the partnership submodels. The lower half contains summaries of the partnership history that are used as predictors in the health submodel.

et al. (2017) noted that the estimates from the 3-step approach may be biased due to local dependence (i.e. outcomes are dependent conditional on latent classes). The inclusion of u_i can help to mitigate this problem by capturing the additional associations between outcomes due to unmeasured time-invariant characteristics. The current model assumes that the time-varying residual association across equations is zero, i.e. unobserved time-varying influences on partnership transitions and distal health outcomes are uncorrelated. While the model could be extended to allow a non-zero correlation between time-varying residuals, identification would require instrumental variables, in our case variables that predict partnership transitions but not health. Such variables are difficult to find.

Figure 1 is a graphical illustration of the model given by Equations (3) and (4), i.e. step 3 of the 3-step procedure, that imposes a factor structure on the random effects. For clarity of presentation, we present a model with two associated latent categorical variables, each measured by a distinct set of repeated measures. The step 1 latent class models for C_1 and C_2 are not depicted for simplicity but we note that modal classes (M_1, M_2) and misclassification probabilities derived in step 2 are used in step 3 where the SEM is fitted.

The proposed multilevel SEM combines a multilevel generalized linear model (Skrondal and Rabe-Hesketh, 2004) with a multilevel factor model (Goldstein, 2010) with both continuous and categorical level 2 (individual-level) factors. Such a specification allows for both a direct (via $\mathbf{Z}_i^{(P)}$ and the latent Cs) and indirect (via the individual-level unobservables u_i) association between partnership transitions, childhood SECs and midlife health.

2.4. Missing data

Due to dropout, not all NCDS cohort members have a complete partnership history for ages 16-50. Restriction of the analysis sample to those individuals with a complete history results in a loss of nearly 50% of those cohort members present at age 16. We have full histories for all individuals who were present at the age 50 wave, regardless of whether they missed one or more previous adult waves: for individuals who did not respond at a given wave, their history was updated at the next wave. Therefore only individuals who dropout (and do not return) have incomplete histories, and we model dropout rather than intermediate non-response. Of the three missing data mechanisms in Little and Rubin's (1987) classification, we focus on the more flexible missing not at random (MNAR) or non-ignorable mechanism where missingness may depend on the unobserved values for two main reasons. First, as noted by Molenberghs et al. (2014), it is impossible, both empirically and theoretically to assess the validity of the assumption of ignorable missingness. Second, it is reasonable to expect that individuals who have a poor health status or an unstable relationship are more likely to drop out from the study.

We specify a model for the probability of dropout at each of the waves when partnership data were collected: waves 5 (age 33), 6 (age 42), 7 (age 46) and 8 (age 50). Denote by D_{ri} an indicator for dropout at wave r ($r = 5, 6, 7, 8$). Thus an individual that provides partnership information only at wave 6 and is not present in other waves (i.e. dropout at wave 7) has a dropout vector $\mathbf{D}_i = (0, 0, 1)'$. The probability of dropout at wave r is defined as $h_{ri}^{(D)} = P(D_{ri} = 1 | D_{r'<r,i} = 0)$. Considering the complexity of the current model with four latent categorical variables and random effects, we specify a logit model for the dropout probability with a wave-specific baseline hazard ($\alpha_r^{(D)}$) and the effects of predictors assumed time-invariant:

$$\text{logit}\left(h_{ri}^{(D)}\right) = \alpha_r^{(D)} + \boldsymbol{\beta}^{(D)\prime}\mathbf{X}_i^{(D)} + \sum_{q=1}^{4}\sum_{k_q=1}^{K_q-1}\tau_{C_q,k_q}^{(D)} I(C_{qi} = k_q) + \lambda^{(D)}u_i. \tag{5}$$

This specification essentially combines the selection model and the shared parameter model, which has also been considered by Muthén et al. (2011) and Washbrook et al. (2014). In addition to having childhood SECs predicting the dropout probability, indirect residual association between dropout and the outcomes of interest is accommodated by including u_i with coefficient $\lambda^{(D)}$ in the dropout model. The significance of $\lambda^{(D)}$ signals a selection effect in the tendency of partnership dissolution. For example, $\lambda^{(D)} > 0$ indicates that cohort members with unobservables that place them at a higher risk of dropout tend also to have a higher tendency to separate from a partner.

For individuals who drop out before age 50, the midlife health outcome is not recorded and the partnership information is incomplete. Consequently, in the health submodel, we have missing values for H_i and only partial (or no) information on $\mathbf{Z}_i^{(P)}$, the summaries of partnership experience. Simulation results show that using the partially observed $\mathbf{Z}_i^{(P)}$ leads to heavily biased estimates of parameters (more details are given in the supplementary material). Multiple imputation of these summary variables under an assumption of missing at random is also inappropriate because it is inconsistent with the possibility that midlife health and partnership stability may be interrelated via individual-level unobservables u_i. Alternatively, joint estimation of these interrelated processes produces unbiased estimates if the models are correctly specified. However, the results are sensitive to the specification of the dropout model (also noted in Molenberghs et al. 2014). In particular, the omission of confounders in the dropout model may lead to biased estimates of the parameters of primary interest (i.e. the $\boldsymbol{\beta}$s in Equation (4)).

2.5. Estimation and assessment of model fit

The proposed SEM with the four submodels defined by Equations (3) to (5) can be estimated using the 3-step maximum likelihood approach both manually and automatically (using the step3 option) in LatentGOLD (Vermunt and Magidson, 2015). Manual and automatic estimation using the 3-step procedure for models with one latent categorical variable is also possible in Mplus 7.1 (Muthén and Muthén, 2017) and later versions. Latent class models in step 1 were estimated using full information maximum likelihood with the iterative EM algorithm (Dempster et al., 1977) to estimate the latent class models (Hagenaars and McCutcheon, 2002; Vermunt, 2010). Before carrying out the estimation in step 3, the input dataset needs to be restructured. Denote by $\mathbf{Y}_i = (H_i, \mathbf{y}_i^{(F)}, \mathbf{y}_i^{(S)}, \mathbf{D}_i)$ a stacked response vector for partnership outcomes, binary midlife health and the dropout indicators. A binomial logit model is then fitted to the single stacked response vector with denominator $(1, \mathbf{n}_i^{(F)}, \mathbf{n}_i^{(S)}, \mathbf{1})$. Explanatory variables can be included in the model by interacting them with four binary indicators that index each response. If H_i is not binary, a different link function in the health submodel is required. In that case, we can define a bivariate response vector where partnership outcomes and dropout indicators are stacked into a single vector and H_i is expanded to the same length. For H_i, only one record per individual is retained and the remaining entries (both for H_i and its predictors) are set to zero. The part of the data with both the outcome and the predictors set to zero does not contribute to the likelihood function of the model. Examples of the data structure and LatentGOLD syntax files are available in the supplementary materials.

For a complex SEM with a measurement model and several components to the structural model, global fit indices are not instructive about the local fit of each part of the model. For this reason, we instead assess the fit of the measurement model (i.e. latent class models for each dimension of childhood SECs) using a set of model selection criteria and entropy for classification quality, and then compare nested specifications of the structural model to test for endogeneity of partnership transitions and non-ignorable dropout to mitigate model misspecification.

3. Data

3.1. Overview of the NCDS

The National Child Development Study is a birth cohort study commissioned in 1958. The target population ($N = 18,558$) contains all individuals born (including stillbirths) in one week of March 1958 in England, Scotland and Wales. In the first sweep, 17,415 individuals were interviewed and this forms the baseline sample. The cohort has been followed up at ages 7, 11, 16, 23, 33, 42, 46, 50 and 55 or until death or permanent emigration from Britain. In all childhood waves (from birth to age 16), the survey interviewed cohort members' parents (to collect, for example, information on the socio-economic situation of the family), teachers (for measures of cognitive development), school doctors and cohort members themselves (since age 7). The adult sweeps (from age 23 onwards) cover a wide range of topics, including physical (health-related, mostly self-reported) and mental well-being, life events (employment, partnership and birth histories) and cognitive developments. In this study, we use data collected up to age 50.

3.2. Measures of childhood socioeconomic circumstances

It is common to use measures of childhood SECs at one or multiple time points during childhood as covariates for partnership transitions and later health. For instance, parental social class defined by the father's or male head's social class at birth, ages 7, 11 and 16 have been considered as predictors of partnership transitions (Aassve et al., 2006b) and physical well-being in midlife (Kelly-Irving et al., 2013). The experience of parental divorce by age 16 has been found to be associated with physical disease, such as cancer (Kelly-Irving et al., 2013), and partnership decisions (Aassve et al., 2006b). Financial hardship in childhood is often constructed as a composite indicator using multiple items for living conditions collected at birth and ages 7 and 11. Previous studies have noted the importance of childhood SECs in relation to partnership transitions (Berrington and Diamond, 2000) and physical health in midlife (Bartley et al., 2012; Kelly-Irving et al., 2013). In this study, we consider changes in these four dimensions of childhood SECs, i.e. parental social class, financial difficulty, material hardship and family structure, across four childhood waves. For each childhood wave, we construct a composite index for each dimension, combining a set of variables that have been considered in previous studies to capture a similar aspect of childhood SECs (e.g. Schoon et al., 2003). Separate latent class models are then fitted to each set of these composite indexes and each derived categorical latent variable captures the pattern of change across four waves (at ages 0, 7, 11 and 16), for each dimension of childhood SECs.

The social class measure is based on the occupation of the male head of the cohort member's family and is scored on a 6-point ordinal scale: unemployed, unskilled, partially skilled, skilled, managerial and professional. The composite binary measure of financial difficulty is coded following the approach of Bartley (2003) and Schoon et al. (2003) where a value of 1 is assigned if there is at least one positive answer to the questions related to financial hardship (receiving free school meals, receiving benefits and having a father in the last two social class categories). Material hardship is measured on a 5-point ordinal scale with higher values for higher levels of material hardship, and is derived from a set of binary questions related to the following four aspects: overcrowding, shared household amenities, not owning the property and recipient of support benefits. Answers to these four questions were collected repeatedly only at ages 7, 11 and 16. The last measure of family structure is constructed as a nominal variable with five categories following Hobcraft (1998). The five situations from poor to good are as follows: in care or in foster care or in other similar situations, cared for by other blood relatives, cared for by a single parent, cared for by step parents and cared for by joint parents. More details of the construction of these measures are available in the supplementary material. Table 1 summarises the distribution of each composite index at each childhood wave.

Table 1. Percentage distribution of childhood measures by age.

Category	0 yr	7 yr	11 yr	16 yr
Social Class				
Unemployed	0.05	1.80	2.60	2.78
Unskilled	8.71	4.55	3.58	2.72
Partially skilled	11.09	11.51	11.76	7.90
Skilled	51.71	41.58	35.42	29.22
Managerial	11.18	11.48	12.64	11.19
Professional	3.90	4.03	3.89	3.04
Missing	12.64	23.79	30.37	43.17
Financial Difficulty				
Yes	8.76	12.23	17.25	13.69
No	78.60	59.38	55.77	48.19
Missing	12.64	28.39	26.99	38.11
Material Hardship				
Low	–[a]	19.67	24.18	23.72
Low to medium	–	19.66	21.63	20.75
Medium	–	18.43	19.70	13.89
Medium to high	–	6.26	6.72	2.52
High	–	1.10	1.05	0.16
Missing	–	34.89	26.71	38.96
Family Structure				
Others[b]	–	0.43	0.56	0.72
Blood relatives	–	0.25	0.29	0.34
Single parent	3.62	3.01	3.98	6.18
Step parents	0.18	1.41	2.80	2.79
Joint parents	89.99	73.91	66.79	52.87
Missing	6.21	20.99	25.58	37.11

[a] – denotes not available
[b] Others includes in care, in foster care, and other situations

3.3. Partnership histories and predictors

The merged partnership history dataset (1974-2008) records all co-residential partnership episodes which were over one month in length. Partnerships ending within one month were also recorded for completeness. A detailed description of the dataset is available in Hancock et al. (2011). After excluding cohort members who died before age 16 ($N = 881$) and individuals who have records with coding error or erroneous partnership information ($N = 3,368$), our sample consists of $14,309$ individuals, of whom $7,313$ have complete partnership histories. The remaining $6,996$ individuals have incomplete partnership histories during ages 16-50, i.e. they were not present at wave 8 (age 50). After grouping durations into 6-month intervals, the individuals with complete histories contributed 340,883 records to the process of first partnership formation and 451,639 records to the process of partnership dissolution.

The choice of the predictors of partnership transitions are guided by previous studies that have investigated the risk factors of entering into and moving out of a union (Steele et al., 2006; Aassve et al., 2006a; Aassve and Billari, 2006). Childhood circumstances (measured at birth and age 16) have often been used to control for family backgrounds. In this study, latent variables are derived to summarise the experiences in childhood. As these variables measure childhood circumstances from birth, they capture much of the information from baseline covariates. Among the other explanatory variables considered in this application, two variables were derived from the related event histories of the NCDS. "Number of post-compulsory years of education" is derived from indicators of education attainment in the economic activity history. When the activity history has missing values, we impute the indicators for being in full-time education by assuming the individual is in continuous education until the highest degree is obtained. "Number of pre-school children" is derived from the fertility history where only coresident pre-school (less than 5 years old) children are counted. Further detail on the construction of these variables is available in Steele et al. (2006). Other time-invariant unobservables are captured by individual-specific random effects. Descriptive statistics for the observed covariates included in the final models for partnership transitions are available in the supplementary material.

3.4. Measure of midlife health and predictors

The midlife health outcome at age 50 is based on a self-reported measure recorded on a 5-point scale with categories excellent, very good, good, fair and poor. To reduce measurement error due to subjective interpretations of adjacent categories (e.g. the boundary between good and very good may be unclear), we group the first three categories into "good health" (coded 0) and the last two into "poor health" (coded 1).

In terms of the predictors of midlife health, our main interest is the childhood SECs and summaries of partnership experiences. Three summaries of partnership experiences are derived. The total number of partners before age 50 is coded as 0 (2%), 1 (37.6%), 2 (8.9%) and 3+ (51.5%) (see Table 2). The percentage of time spent single during ages 16-50 is computed as the fraction of time spent between partnerships over the 34 years of follow-up. Age at the first partnership, an outcome of the history of first partnership formation, is a continuous variable that is centered around its median and log-transformed. For individuals who have never partnered, this variable is set at zero. It

Table 2. Distributions of categorical and continuous predictors of the midlife health ($N = 7,313$). For categorical predictors, percentage distributions are reported; for continuous predictors, mean and the standard deviation (SD) are reported.

Categorical predictors	Percentages
Gender	
Male	50.9
Female	49.1
Overweight at age 16	
No	57.0
Yes	43.0
Total number of partners	
0	2.0
1	37.6
2	8.9
3+	51.5
Continuous predictors	Mean (SD)
Percentage time spent single[a]	0.3 (0.2)
log(age at first partnership)	0.3 (1.3)

[a]Percentage time spent single is calculated as the proportion of total number of months spent in between relationships over the total number of months observed before dropout.

is essentially an interaction variable for the total number of partners and the age at first partnership. This also allows for a clear interpretation of the effect of the percentage of time spent in a single status on midlife health, by allowing for comparison among those who started the first partnership at the same age. We also include gender and early health as control variables, as in previous studies (e.g. Ploubidis et al., 2015). For early health, we derive a binary indicator for overweight at age 16 using World Health Organisation cut-offs where BMI (kg/m^2) > 25.0 denotes "overweight" (WHO, 2000). With respect to the other predictors of partnership transitions and midlife health that are unobserved in the survey, the influence of time-invariant unobservables is captured by the individual-level random effects in Equations (3) and (4). Descriptive statistics for the predictors of the midlife health model are summarised in Table 2.

The limited number of observed control variables is explained below. First, we have considered multiple dimensions of childhood SECs for ages 0-16 that capture much of the baseline covariates. Other variables related to early health before age 16 have a large amount of missing values that would require imputation. However, as these variables may be related to the latent childhood SECs, the imputation procedure is not straightforward. Indicators for childhood SECs may be used as auxiliary variables to impute early health variables, but these indicators contain missing values themselves. Therefore, in addition to the validity of the imputation model (discussed by Seaman and White (2013)), there are practical issues to consider. Second, although multiple variables for various aspects of adulthood situations at each adulthood wave of the study are available, they may be moderators or even mediators of the effects of childhood SECs on midlife health (e.g. education, employment transitions and health-related be-

Table 3. Proportion of respondents falling into the latent classes (modal class membership) for each aspect of childhood SECs measured at four waves, entropy and proportion of missing data.

Childhood SECs	Entropy	Missing (%)	Modal class allocation (%)		
			1	2	3
Father/male head social class	0.727	2.9	21.0 (high)	51.0 (medium)	25.1 (low)
Financial difficulty	0.700	2.2	80.7 (low)	17.1 (high)	
Material hardship	0.790	12.5	28.9 (low)	30.7 (high)	27.9 (medium)
Family structure/union	0.916	1.0	9.3 (unstable)	89.7 (stable)	

haviours). This requires an extensive discussion in order to properly incorporate them into the structural model of the SEM and is not the focus of this paper. In this paper, we study the effects of childhood SECs and partnership experiences on health using a joint modelling approach that controls for endogeneity of partnership experiences due to unmeasured time-invariant individual characteristics. An alternative approach would be to study the processes that lead to worse or better later health by decomposing the total effects of SECs into direct effects on health and indirect effects through possible mediation variables. Such a mediation study is similar to the work of Arpino et al. (2018) but further research is required to satisfy assumptions for causal interpretation in longitudinal observational settings. This is beyond the scope of this paper.

4. Results

4.1. Results from the latent class models

The results of the latent class models for childhood SECs are summarised in Table 3, where descriptive labels for each class are based on an examination of the response probabilities for each observed repeated measure (composite index) conditional on class membership. For each latent variable, the choice of the number of classes is based on a combination of goodness-of-fit test statistics and model selection criteria (including the AIC, sample-size adjusted BIC, entropy, and the Lo-Mendell-Rubin and bootstrap likelihood ratio tests), the proportion in each class (ensuring at least 10% in each category) and interpretation of the classes (see class labels in Table 3). Specifically, entropy measures the quality of classification where values closer to one indicate a clear class separation. The percentage distributions for the modal class allocations are derived by assigning each individual to the class for which their posterior probability of class membership is highest, where this probability is based on their pattern of responses on the observed time-varying indicators. In step 2, misclassification probabilities are calculated and treated as fixed loadings in the multilevel joint SEM of step 3. Further detail on the selection and labelling of latent classes for each dimension of childhood SECs is provided in the supplementary material.

4.2. Multilevel joint SEM

We next relate the latent variables for childhood SECs to partnership transitions and midlife health by estimating the multilevel joint SEM described in Section 2, taking into account the classification error in the assignment of individuals to the most likely class in step 1. Table 4 shows the estimated effects of four childhood SECs and partnership experiences on midlife health. The remaining results from the multilevel joint SEM are reported in the supplementary material. Model 1 is the health model only, and Models 2 and 3 are the multilevel joint SEM for health and partnership transitions before and after including the dropout model, respectively.

We now discuss the findings from the SEM, with a focus on the effects of childhood SECs on midlife health and partnership transitions. Controlling for gender and BMI at age 16, the results from Model 1 show significant effects of the three dimensions of childhood SECs that relate to material difficulties – parental social class, financial difficulty and material hardship – on midlife health, with cohort members from families with unfavourable conditions on these aspects significantly more likely to be in a poor health state at age 50 (see Table 4). The effect of family structure is non-significant.

Results from Models 2 and 3 in Table 4 and Table 5 show the association between partnership transitions and midlife health as explained by childhood SECs and other covariates taking also into account unobserved individual level characteristics. In particular, a comparison of Models 1 and 3 (or 2) in Table 4 show the extent to which the effects of childhood SECs can be explained by partnership transitions. We find that the experience of material difficulties has a persistent influence on poor midlife health, even after controlling for partnership experiences. In contrast, the effect of family structure on midlife health operates only indirectly by influencing partnership transitions, which in turn affect later health. Specifically, cohort members who have experienced unstable family structures in childhood form their first partnerships earlier and have a higher risk of separation across relationships (see Table 5). From the health submodel (Table 4), individuals who have formed their first partnership late have a lower risk of developing health issues at age 50. Moreover, conditional on the age at first partnership, cohort members who have spent a longer time single between ages 16 and 50 (accounting for dropout) have an increased chance of poor health in midlife. From Model 3 we also find that, compared with individuals who have only formed one partnership, those who have over three relationships are significantly more likely to have poor midlife health. These results suggest that unstable family structure is positively related to a higher chance of poor midlife health, but the influence is only indirect through partnership transitions.

To illustrate the magnitude of the effects of childhood SECs and partnership experience on midlife health, we consider four combinations of childhood circumstances and partnership experiences for different values of the random effect, and compute the predicted probabilities of poor midlife health, holding other predictors at reference values (Figure 2). We find that individuals with the most disadvantage in childhood and with unstable relationship experiences in adulthood have a higher probability of poor midlife health, but there is a large amount of between-individual unobserved heterogeneity.

As part of a sensitivity analysis we give the results of Models 1-3 when the modal class is used instead of the 3-step approach for the four dimensions of childhood SECs in the supplementary material. The results change for all models. Specifically, the

Table 4. Effects of childhood SECs and summary partnership experiences on log-odds of poor midlife health. Model 1 is the health model alone, Models 2 and 3 are the respective SEMs before and after including the dropout submodel.

Covariates	Model 1		Model 2		Model 3	
	Est.	(SE)	Est.	(SE)	Est.	(SE)
Intercept	-2.36**	(0.09)	-2.70**	(0.15)	-3.06**	(0.23)
Male (ref.= Female)	-0.05	(0.06)	-0.04	(0.06)	-0.05	(0.06)
Overweight at age 16[a] (ref.= No)	0.25**	(0.07)	0.26**	(0.07)	0.26**	(0.07)
Childhood circumstances						
Social class [b] (ref.=High)						
Low	0.40**	(0.19)	0.46**	(0.11)	0.44**	(0.12)
Medium	0.32**	(0.11)	0.31**	(0.09)	0.30**	(0.10)
Financial difficulty (ref.=Low)						
High	0.53**	(0.21)	0.46**	(0.09)	0.42**	(0.10)
Material hardship (ref.=Low)						
Medium	0.33**	(0.11)	0.32**	(0.08)	0.32**	(0.09)
High	0.35**	(0.12)	0.41**	(0.09)	0.39**	(0.10)
Family structure (ref.=Stable)						
Unstable	0.08	(0.13)	0.14	(0.13)	0.17	(0.17)
Partnership experience						
Total number of partners before age 50 (ref. =1)						
0			-0.12	(0.32)	-0.13	(0.32)
2			0.04	(0.13)	0.18	(0.14)
3+			0.15	(0.23)	0.41*	(0.24)
Age at first partnership			-0.13**	(0.05)	-0.13**	(0.05)
Percentage time spent single			1.08**	(0.37)	1.26**	(0.38)
Random effect parameters						
σ_u^2			0.93**	(0.10)	1.32**	(0.10)
$\lambda^{(H)}$			-0.20	(0.15)	-0.44**	(0.16)
$\lambda^{(F)}$			0.03	(0.09)	-0.05**	(0.02)
$\lambda^{(D)}$					1.25**	(0.12)

** $p < 0.05$, * $p < 0.1$

[a] Binary indicator for overweight at age 16.

[b] Father or male head social class.

Table 5. Effects of childhood SECs on partnership transitions. Models 2 and 3 are the respective SEMs before and after including the dropout submodel.

Childhood SECs	Model 2		Model 3	
	Est.	(SE)	Est.	(SE)
First entry into partnership				
Social class[a] (ref.=High)				
Low	0.11**	(0.04)	0.11**	(0.04)
Medium	0.10**	(0.03)	0.11**	(0.03)
Financial difficulty (ref.=Low)				
High	0.03	(0.04)	0.03	(0.04)
Material hardship (ref.=Low)				
Medium	0.04	(0.03)	0.04	(0.03)
High	0.05	(0.03)	0.05	(0.03)
Family structure (ref.=Stable)				
Unstable	0.15**	(0.04)	0.15**	(0.03)
Partnership dissolution				
Social class[a] (ref.=High)				
Low	-0.13*	(0.07)	-0.07	(0.07)
Medium	-0.04	(0.06)	0.01	(0.05)
Financial difficulty (ref.=Low)				
High	0.05	(0.07)	0.17**	(0.07)
Material hardship (ref.=Low)				
Medium	-0.09*	(0.05)	-0.09	(0.06)
High	-0.14**	(0.06)	-0.11*	(0.06)
Family structure (ref.=Stable)				
Unstable	0.23**	(0.07)	0.29**	(0.07)

**$p < 0.05$, *$p < 0.1$.

[a] Father or male head social class.

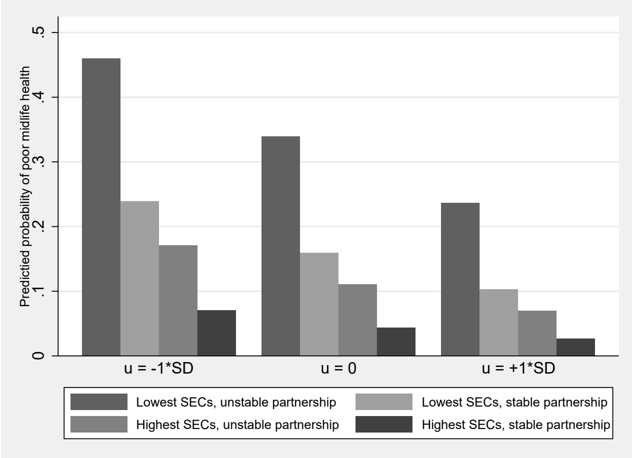

Fig. 2. Predicted probabilities of poor midlife health for various combinations of childhood SECs and partnership experiences and for different random effect values (the mean of 0 and plus or minus one standard deviation about the mean).

modal class approach finds that all four dimensions of childhood SECs are significant risk factors for midlife health and the estimated standard errors are smaller than those in the 3-step approach. Similar results were found in Zhu et al. (2017) for a continuous health outcome. The difference in estimates from the modal and 3-step approach may be substantial if the class separation in the latent class models (step 1) is poor. Extensive simulation studies (Asparouhov and Muthén, 2016; Bakk and Vermunt, 2016; Zhu et al., 2017) have confirmed that results from the modal class approach are likely to be biased, even when the classification quality is reasonably high (e.g. entropy values of 0.6 or 0.7, as in our application).

Turning to dropout (see Table 9 of the supplementary material), disadvantage in childhood is associated with a higher probability of dropout at each adult wave. We also find that males, younger mothers, and individuals with lower maths and reading marks and behaviour problems at age 16 have a significantly higher risk of dropout. Comparing the SEM with (Model 3) and without (Model 2) the dropout submodel, we find that allowing for non-ignorable dropout has minimal influence on the effects of childhood SECs on midlife health, but the effects of having three or more partners and the proportion of time single become stronger (Table 4). Although there is a significant residual association between dropout, health ad partnership transitions ($\lambda^{(D)} \neq 0$, see Table 4), its impact on the effects of primary interest is negligible in this application. We also note that the addition of the dropout submodel brings more information into the joint SEM, which helps to estimate more stably the parameters related to the random effect components. Further detail on the interpretation of the random effects part of the joint model are given in Section 6 of the supplementary materials for interested readers.

In addition to these analysis, we have also conducted a simulation study to assess the sensitivity of the parameters of interest (effects of childhood SECs on midlife health and partnership experiences) to misspecification of the dropout model, assuming the models for health and partnership transitions are correctly specified. More details are given in

Section 11 of the supplementary material. In summary, under the assumption that the models for the primary outcomes are correctly specified, predictors that are common to the missingness process and the primary outcomes (childhood SECs and early health, i.e. BMI at age 16) should be included in the specification of the dropout model where possible. Neglecting such predictors may lead to biased estimates of the parameters of interest.

5. Conclusion

In this paper, we proposed a multilevel joint SEM to explore the associations between the childhood SECs, partnership transitions and midlife health. Our model extends the 3-step approach proposed in Zhu et al. (2017) to an SEM where the measurement error and the pattern of changes in SECs across childhood waves is captured by latent categorical variables and then related to multivariate, clustered and mixed-type outcomes. The processes of substantive interest are modelled jointly with an informative dropout mechanism where the association is driven by the common predictors, including childhood SECs, and shared unobserved time-invariant characteristics. The 3-step approach corrects the bias in estimated coefficients produced by conventional methods that ignore classification error in assigning individuals to latent classes, while the joint model of the associated processes allows us to explore the life course influence on midlife health by bringing together early-life characteristics and experiences of life events during adulthood.

Our findings highlight the importance of taking a life course perspective to understand the effects of childhood SECs on later health. By incorporating longitudinal information on associated dimensions of the SECs, we find that of the four childhood SECs considered, parental social class, material hardship and financial difficulty (i.e. material disadvantages) have a persistent effect on midlife health over the life course. These results are of particular importance and have implications for the development of social inequalities. The effect of family structure, on the other hand, has no direct influence on midlife health. We also find that partnership transitions have an important role to play in the relationship between childhood SECs and later health.

The model presented here could be extended in several ways. First, we have considered a small set of control variables of midlife health. To include other variables that are indicators for early health in childhood, missing values need to be imputed. As they may be associated with childhood SECs (latent variables), the imputation procedure requires careful consideration. Although one option would be to use the modal classes for the SEC variables as auxiliary variables in the imputation model, they are subject to classification errors, particularly when class separation is poor. Alternatively, indicators for the childhood SECs may be used directly but they contain missing values themselves. Second, the proposed SEM can be extended to handle multiple health outcomes, for example distinguishing mental and physical health, which may be of mixed types and may have repeated measurements over the life course. The model could be extended to include mental and physical health as other processes that are modelled jointly with partnership transitions and mid-life health. However, we note that this may be methodologically challenging due to the mismatch of record frequency of each process in the

NCDS, and careful consideration of the specification of the extended model would be required where latent categorical predictors (childhood SECs) are involved.

We have assumed additive effects of the multiple childhood SECs on the outcomes of interest. While their effects may interact, fitting models with interactions among latent categorical variables requires non-linear constraints on the parameters (in addition to current constraints) and will further complicate the estimation, in particular when using the 3-step procedure. This is an interesting direction for future research. Alternatively, interactions between the latent categorical variables and the covariates can be introduced as a more straightforward extension of our proposed framework. This requires a modification of the syntax for our current model that are provided in the supplementary material.

Given that childhood SECs and partnership experiences cannot be manipulated by an individual, the potential outcome framework (Rubin, 2005) is not suitable for our setting and research questions. As we are not answering a causal question, and that a single-purposed prediction model for midlife health would not provide insights on the interrelationships between the processes of interest, the joint multilevel modelling framework is best placed in this context to handle the complexity of the conceptual model and the statistical issues discussed earlier simultaneously and efficiently. It also has considerable generalisability (for example to continuous, categorical and mixed type outcomes) because it belongs to the generalised latent variable modelling framework (Skrondal and Rabe-Hesketh, 2004).

Causal assumptions are essentially part of the statistical assumptions required for model-building. A good discussion of this issue is given by Bollen and Pearl (2013): "SEM is an inference engine that takes in two inputs, qualitative causal assumptions and empirical data, and derives two logical consequences of these inputs: quantitative causal conclusions and statistical measures of fit for the testable implications of the assumptions.". In our modelling context, qualitative (i.e. substantive) causal assumptions were necessary to define functional relationship between observed variables and latent variables, as well as the correlation between residuals across equations. A closely related research stream in the life course literature focuses on causal mediation analysis (Imai et al., 2010; Pearl, 2009; VanderWeele et al., 2016). Provided the sequential ignorability assumption is met, non-parametric identification of causal effects is possible. In a mediation analysis, the interest is in decomposing the total effects of childhood SECs on later health into direct and indirect effects where mediators may be summary variables of life experiences (e.g. partnership or employment transitions). For complex longitudinal settings such as that described here, careful control of time-varying confounders, time-varying mediators, prior exposures and intermediate confounders needs to be explicitly addressed. These variables may include education attainment, health states and health-related behaviors (such as drinking and smoking) in early adulthood and variables that are not collected in the survey. It is well-known that for observational longitudinal studies, these issues are particularly difficult to account for and this alone forms a separate research stream.

References

Aassve, A. and Billari, Francesco C.and Spéder, Z. (2006) Societal transition, policy changes and family formation: Evidence from hungary. *European Journal of Population / Revue européenne de Démographie*, **22**, 127–152.

Aassve, A., Burgess, S., Propper, C. and Dickson, M. (2006a) Employment, family union and childbearing decisions in Great Britain. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **169**, 781–804.

Aassve, A., Burgess, S. M., Dickson, M. and Propper, C. (2006b) Modelling poverty by not modelling poverty: an application of a simultaneous hazards approach to the UK. *LSE STICERD Research Paper No. CASE106.*

Allison, P. D. (1984) *Event history analysis: Regression for longitudinal event data: First Edition.* New York: Sage.

Arpino, B., Gumà, J. and Julià, A. (2018) Early-life conditions and health at older ages: The mediating role of educational attainment, family and employment trajectories. *PloS One*, **13**, e0195320.

Asparouhov, T. and Muthén, B. (2014) Auxiliary variables in mixture modeling: Three-step approaches using Mplus. *Structural Equation Modeling: A Multidisciplinary Journal*, **21**, 329–341.

— (2016) Structural equation models and mixture models with continuous nonnormal skewed distributions. *Structural Equation Modeling: A Multidisciplinary Journal*, **23**, 1–19.

Bakk, Z. and Kuha, J. (2018) Two-step estimation of models between latent classes and external variables. *Psychometrika*, **83**, 871–892.

Bakk, Z., Tekle, F. B. and Vermunt, J. K. (2013) Estimating the association between latent class membership and external variables using bias-adjusted three-step approaches. *Sociological Methodology*, **43**, 272–311.

Bakk, Z. and Vermunt, J. K. (2016) Robustness of stepwise latent class modeling with continuous distal outcomes. *Structural Equation Modeling: A Multidisciplinary Journal*, **23**, 20–31.

Barber, J. S., Murphy, S. A., Axinn, W. G. and Maples, J. (2000) Discrete-time multilevel hazard analysis. *Sociological methodology*, **30**, 201–235.

Bartley, M. (2003) Commentary: Relating social structure and health. *International Journal of Epidemiology*, **32**, 958–960.

— (2016) *Health inequality: an introduction to concepts, theories and methods.* John Wiley & Sons.

Bartley, M., Kelly, Y. and Sacker, A. (2012) Early life financial adversity and respiratory function in midlife: a prospective birth cohort study. *American Journal of Epidemiology*, **175**, 33–42.

Ben-Shlomo, Y., Cooper, R. and Kuh, D. (2016) The last two decades of life course epidemiology, and its relevance for research on ageing. *International Journal of Epidemiology*, **45**, 973–988.

Berrington, A. and Diamond, I. (2000) Marriage or cohabitation: A competing risks analysis of first-partnership formation among the 1958 British birth cohort. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **163**, 127–151.

Bolck, A., Croon, M. and Hagenaars, J. (2004) Estimating latent structure models with categorical variables: One-step versus three-step estimators. *Political Analysis*, **12**, 3–27.

Bollen, K. A. and Pearl, J. (2013) Eight myths about causality and structural equation models. In *Handbook of causal analysis for social research*, 301–328. Springer.

Brewer, M. and Nandi, A. (2014) Partnership dissolution: how does it affect income, employment and well-being? *ISER Working Paper Series 2014-30*, Institute for Social and Economic Research, University of Essex.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **39**, 1–38.

Goldstein, H. (2010) *Multilevel Statistical Models, 4th Edition.* New York: John Wiley & Sons.

Green, D. M., Cox, C. L., Zhu, L., Krull, K. R., Srivastava, D. K., Stovall, M., Nolan, V. G., Ness, K. K., Donaldson, S. S., Oeffinger, K. C. et al. (2012) Risk factors for obesity in adult survivors of childhood cancer: a report from the childhood cancer survivor study. *Journal of Clinical Oncology*, **30**, 246.

Hagenaars, J. A. and McCutcheon, A. L. (2002) *Applied Latent Class Analysis.* Cambridge: Cambridge University Press.

Hagger-Johnson, G., Batty, G. D., Deary, I. J. and von Stumm, S. (2011) Childhood socioeconomic status and adult health: comparing formative and reflective models in the Aberdeen Children of the 1950s Study (prospective cohort study). *Journal of Epidemiology & Community Health*, **65**, 1024–1029.

Hancock, M., Elliott, J., Johnson, J. and Bukodi, E. (2011) National Child Development Study, partnership histories 1974-2008. *A guide to datasets*, Centre for Longitudinal Studies, Institute of Education, University of London.

Hobcraft, J. (1998) Intergenerational and life-course transmission of social exclusion: Influences and childhood poverty, family disruption and contact with the police. *STICERD Research Paper No. CASE015*, London School of Economics, London.

Imai, K., Keele, L. and Tingley, D. (2010) A general approach to causal mediation analysis. *Psychological methods*, **15**, 309.

Kelly-Irving, M., Lepage, B., Dedieu, D., Lacey, R., Cable, N., Bartley, M., Blane, D., Grosclaude, P., Lang, T. and Delpierre, C. (2013) Childhood adversity as a risk for cancer: findings from the 1958 British birth cohort study. *BMC Public Health*, **13**, 767.

Lanza, S. T., Tan, X. and Bray, B. C. (2013) Latent class analysis with distal outcomes: A flexible model-based approach. *Structural Equation Modeling: A Multidisciplinary Journal*, **20**, 1–26.

Lillard, L. A., Brien, M. J. and Waite, L. J. (1995) Premarital cohabitation and subsequent marital dissolution: a matter of self-selection? *Demography*, **32**, 437–457.

Little, R. J. A. and Rubin, D. B. (1987) *Statistical Analysis with Missing Data*. New York: John Wiley & Sons.

Maggs, J. L., Patrick, M. E. and Feinstein, L. (2008) Childhood and adolescent predictors of alcohol use and problems in adolescence and adulthood in the national child development study. *Addiction*, **103**, 7–22.

Molenberghs, G., Fitzmaurice, G., Kenward, M. G., Tsiatis, A. and Verbeke, G. (2014) *Handbook of Missing Data Methodology*. London: Chapman and Hall/CRC.

Muthén, B., Asparouhov, T., Hunter, A. M. and Leuchter, A. F. (2011) Growth modeling with non-ignorable dropout: Alternative analyses of the STAR*D antidepressant trial. *Psychological Methods*, **16**, 17–33.

Muthén, L. K. and Muthén, B. O. (2017) *Mplus User's Guide. Eighth Edition.* Muthén & Muthén, Los Angeles, CA.

O'Flaherty, M., Baxter, J., Haynes, M. and Turrell, G. (2016) The family life course and health: Partnership, fertility histories, and later-life physical health trajectories in australia. *Demography*, **53**, 777–804.

Pearl, J. (2009) *Causality*. Cambridge university press.

Ploubidis, G. B., Silverwood, R. J., DeStavola, B. and Grundy, E. (2015) Life-course partnership status and biomarkers in midlife: Evidence from the 1958 British birth cohort. *American Journal of Public Health*, **105**, 1596–1603.

Poulton, R., Caspi, A., Milne, B. J., Thomson, W. M., Taylor, A., Sears, M. R. and Moffitt, T. E. (2002) Association between children's experience of socioeconomic disadvantage and adult health: a life-course study. *The Lancet*, **360**, 1640–1645.

Rubin, D. B. (2005) Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, **100**, 322–331.

Schoon, I., Sacker, A. and Bartley, M. (2003) Socio-economic adversity and psychosocial adjustment: A developmental-contextual perspective. *Social Science & Medicine*, **57**, 1001–1015.

Seaman, S. R. and White, I. R. (2013) Review of inverse probability weighting for dealing with missing data. *Statistical Methods in Medical Research*, **22**, 278–295.

Shavers, V. L. (2007) Measurement of socioeconomic status in health disparities research. *Journal of the national medical association*, **99**, 1013.

Skrondal, A. and Rabe-Hesketh, S. (2004) *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models.* New York: Chapman and Hall/CRC.

Steele, F. (2011) Multilevel discrete-time event history analysis with applications to the analysis of recurrent employment transitions. *Australian & New Zealand Journal of Statistics*, **53**, 1–20.

Steele, F., Kallis, C., Goldstein, H. and Joshi, H. (2005) The relationship between childbearing and transitions from marriage and cohabitation in Britain. *Demography*, **42**, 647–673.

Steele, F., Kallis, C. and Joshi, H. (2006) The formation and outcomes of cohabiting and marital partnerships in early adulthood: the role of previous partnership experience. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **169**, 757–779.

Tsiatis, A. A. and Davidian, M. (2004) Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica*, **14**, 809–834.

VanderWeele, T. J., Jackson, J. W. and Li, S. (2016) Causal inference and longitudinal data: a case study of religion and mental health. *Social psychiatry and psychiatric epidemiology*, **51**, 1457–1466.

Vermunt, J. K. (2010) Latent class modeling with covariates: Two improved three-step approaches. *Political Analysis*, **18**, 450–469.

Vermunt, J. K. and Magidson, J. (2015) *LG-Syntax User's Guide: Manual for Latent GOLD 5.0 Syntax Module.* Statistical Innovations Inc, Belmont, MA.

Washbrook, E., Clarke, P. S. and Steele, F. (2014) Investigating non-ignorable dropout in panel studies of residential mobility. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **63**, 239–266.

WHO (2000) Obesity: preventing and managing the global epidemic. *Report of a WHO Consultation (WHO Technical Report Series) 2000-894*, World Health Organization.

Zhu, Y., Steele, F. and Moustaki, I. (2017) A general 3-step maximum likelihood approach to estimate the effects of multiple latent categorical variables on a distal outcome. *Structural Equation Modeling: A Multidisciplinary Journal*, **24**, 643–656.