Early Skills, Environments and Lifecycle Profiles of Earnings and Health

Click Here for Latest Version

Pedro Carneiro, Sarah Cattan, Matthew Nibloe*

November 1, 2026

Abstract

Earnings and health are correlated but imperfectly, so no one variable summarises how a life has unfolded. We ask whether life-cycle profiles of earnings and health can be characterised by a small number of distinct types, and to what extent early skills and environments can predict them. Applying multi-channel sequence analysis to the British 1958 birth cohort, followed from birth to age 50, we classify individuals by their joint trajectories of employment, earnings, physical health and mental health: five types among men and six among women. The types are broadly ordered, although some in the middle exhibit trade-offs between earnings and health. One type – comprising of 5% of men and 9% of women – lies at or near the bottom of every outcome at every age, with employment falling from under 60% at 23 to around 20% at 50, compared with roughly 90% among other types. Membership of this deprived group is substantially predictable from childhood: skills and environments measured by age 16 identify 59% of deprived men and 52% of deprived women, rising to 70% and 65%, respectively, when basic labour-market information at age 23 is included.

Key words: Life-cycle trajectories; Earnings inequality; Health inequality; Early-life skills; Childhood environments; Multidimensional inequality.

JEL: I14, J31, J24, D31, C38

*Carneiro: University College London; Cattan: Institute for Fiscal Studies; Nibloe: University College London. We acknowledge the support of the ESRC through the ESRC Centre for the Microeconomic Analysis of Public Policy at IFS (ES/T014334/1). All errors are our own.

1 Introduction

Inequality is inherently multidimensional and dynamic. Individuals differ not only in their earnings and employment opportunities, but also in their physical and mental health, all of which are important determinants of well-being (Sen 1985). A substantial literature documents inequalities in economic outcomes and physical health and the strong relationships between them (Deaton and Paxson 1998; Deaton 2013; Case and Kraftman 2024), while growing attention has focused on mental health and its relationship with economic disadvantage, notably through the work of Case and Deaton (2015) and Case and Deaton (2017) on deaths of despair among less-educated Americans. Yet these dimensions are imperfectly correlated: individuals with similar economic circumstances may experience very different health outcomes, and vice versa. These different dimensions of inequality also evolve over the life-cycle, as cumulative experiences and shocks shape economic circumstances and health, potentially reinforcing or narrowing differences between individuals as they age. Understanding inequality therefore requires examining not only how outcomes are distributed at a particular point in time, but also how different dimensions of well-being evolve jointly over the life-cycle.

Characterising inequality across dimensions and/or over time is, however, challenging. One established approach is to aggregate different dimensions of well-being into a single index. Multidimensional poverty measures, for example, combine deprivations across several domains using explicit weights and thresholds (Alkire and Foster 2011), building on a broader literature on multidimensional inequality measurement (Maasoumi 1986; Bourguignon and Chakravarty 2003; Decancq and Lugo 2012). Such indices provide tractable summaries of deprivation and are widely used to monitor inequality and inform the geographical or demographic targeting of public policies. However, aggregation necessarily involves choices about the relative importance of different outcomes and may obscure meaningful differences between individuals who experience advantages in some dimensions and disadvantages in others.

Another approach preserves the joint distribution of outcomes and examines whether dis-

tributions can be ranked according to classes of social welfare functions (Atkinson and Bourguignon 1982). While this avoids imposing a single index, establishing dominance becomes increasingly difficult as the number of dimensions grows, particularly when individuals or groups experience trade-offs across outcomes or over time. A complementary literature models joint distributions directly, using copulas to separate marginal distributions from the dependence between outcomes or across periods. For example, Bonhomme and Robin (2009) use this approach to provide a flexible representation of earnings inequality across time and study earnings mobility and intertemporal inequality.

In this paper, we follow an alternative, complementary approach to characterising multidimensional inequality over the life-cycle. Our starting point is that, although individuals' joint trajectories of economic outcomes and health are potentially highly heterogeneous, this heterogeneity may be organised into a relatively small number of distinct and interpretable profiles. Rather than reducing multiple outcomes to a single measure of well-being, we identify groups of individuals who experience similar combinations of outcomes over time. This allows us to distinguish individuals who experience persistent disadvantage or advantage across several dimensions from those whose trajectories involve trade-offs between economic circumstances and health. It also provides a framework for investigating the early-life origins of different patterns of lifetime inequality. Within this framework, we ask two questions. First, what are the principal profiles of joint life-cycle trajectories of employment, earnings, physical health and mental health? Second, to what extent can membership in these profiles be predicted from individuals' childhood skills and environments?

We address these questions using the National Child Development Study, which has followed individuals born in Britain during a single week in 1958 from birth to adulthood. We observe employment, earnings, physical health and mental health at ages 23, 33, 42 and 50, alongside rich information on childhood cognitive and socio-emotional skills and environments measured at ages 7, 11 and 16. To characterise joint adult trajectories, we employ multi-channel sequence analysis, a methodology developed in the life-course literature that measures similarities between individuals' sequences of outcomes, accounting

for their timing, duration and ordering (Abbott and Tsay 2000; Studer and Ritschard 2016). We combine this with hierarchical clustering to identify groups of individuals whose joint trajectories are similar, estimating the classification separately for men and women. We then use multinomial logit models to examine how childhood skills and environments are associated with membership in different profiles. Finally, we assess the extent to which these profiles can be predicted using information available in childhood and early adulthood, employing flexible machine-learning methods evaluated on held-out data.

Our analysis identifies five distinct life-cycle profiles among men and six among women. These profiles reveal substantial and persistent inequalities across economic outcomes, physical health and mental health. The most striking finding is the presence, among both men and women, of a group experiencing severe and persistent disadvantage across all dimensions. This group constitutes 5% of men and 9% of women. Its members have very low employment rates throughout adulthood and, when employed, tend to earn wages close to the minimum wage. Among men, for example, employment rates are approximately 90% for most other groups, but fall from below 60% at age 23 to around 20% at age 50 for the most deprived group. From age 30 onwards, their average earnings decline over the life-cycle. These individuals also consistently exhibit high levels of psychological distress, with average scores exceeding the depression threshold on the Malaise Inventory, and are disproportionately represented in the lowest physical-health category throughout adulthood. They are less likely to marry or own property and more likely to experience chronic illness, disability and involvement in criminal behaviour. The joint analysis thus reveals a group facing entrenched disadvantage across multiple dimensions of well-being, whose circumstances are not adequately captured by examining earnings or health in isolation.

At the opposite end of the distribution, we identify a similarly sized group of individuals who enjoy consistently favourable economic and health outcomes. However, the remaining profiles cannot always be unambiguously ranked. This is particularly apparent among women, for whom we identify a group with relatively high earnings and employment rates

comparable to those of men, but less favourable physical and mental health outcomes. These women also tend to have fewer children. Conversely, another group of women experiences persistently better physical health alongside lower earnings and a greater likelihood of having children. These patterns illustrate the importance of preserving the multidimensionality of lifetime outcomes. Although some individuals experience consistently better or worse outcomes across all dimensions, others exhibit combinations of advantages and disadvantages that cannot readily be summarised through a single ranking. More generally, we find considerable persistence in individuals' relative positions over adulthood, with inequalities across economic and health outcomes becoming entrenched over the life-cycle.

The second part of our analysis examines how these lifetime profiles relate to childhood skills and environments. We first consider cognitive, internalising and externalising skills measured at ages 7, 11 and 16. All three domains are associated with adult profiles, although their relative importance varies across ages and outcomes. Skills measured later in childhood are generally stronger predictors, particularly internalising skills at age 16. Cognitive skills at age 7 are associated with a lower probability of belonging to the persistently deprived group, while cognitive skills at age 16 are particularly important for membership in profiles characterised by favourable economic and health outcomes. Among men, a one-standard-deviation increase in age-16 internalising skills is associated with a 10-percentage-point increase in the relative probability of belonging to the most advantaged group and an 8-percentage-point reduction in the relative probability of belonging to the most deprived group.

Childhood environments also predict lifetime profiles, even after conditioning on skills. Lower exposure to social problems at home and hospitalisation, together with greater parental time investments, are associated with more favourable lifetime outcomes, although these relationships are generally smaller than those associated with skills. The patterns also differ by sex: family socioeconomic status is particularly important for men, while parental demographic characteristics are more strongly associated with women's profiles. Conditional on the rich set of childhood skills and environments, we find limited

additional association between school quality and adult profiles.

We then investigate how early individuals at risk of persistent multidimensional disadvantage can be identified. Using a flexible neural network, we find that childhood skills and environments measured by age 16 can identify 59% of men and 52% of women belonging to the most deprived lifetime profile. Incorporating basic labour-market information observed at age 23 increases these recall rates to 70% and 65%, respectively. These results demonstrate that substantial information about individuals' subsequent economic and health trajectories is already observable in childhood and early adulthood. While predictive relationships should not be interpreted as causal effects, they suggest that early-life information may be valuable for identifying individuals at heightened risk of experiencing persistent disadvantage across multiple dimensions.

Our paper makes two principal contributions to the literature. First, we contribute to the economics literature on multidimensional inequality and the joint evolution of economic and health outcomes by providing a nonparametric, data-driven characterisation of heterogeneity over the life-cycle. A growing literature uses structural models to investigate the dynamic relationships between health and economic outcomes, including the contribution of health inequalities to lifetime earnings inequality (De Nardi et al. 2025; Hosseini et al. 2026) and the interactions between mental health and labour-market trajectories (Jolivet and Postel-Vinay 2025). Our approach complements this work by focusing on the structure of the joint distribution itself, rather than modelling the causal mechanisms linking different outcomes. In doing so, we also connect to the longstanding use of discrete types to represent persistent unobserved heterogeneity in economic models (Heckman and Singer 1984; Keane and Wolpin 1997), and to more recent econometric work on approximating complex heterogeneity through discrete groups (Bonhomme, Lamadon, et al. 2022). Unlike structural approaches in which types typically represent latent endowments or unobserved individual characteristics, we identify profiles directly from observed combinations of lifetime outcomes. Our methodology builds on sequence analysis, widely used in sociology and demography to study life-course patterns, but extends this approach to characterise the joint evolution of economic outcomes, physical

health and mental health and to investigate the childhood origins of the resulting profiles. Importantly, the profiles preserve economically meaningful trade-offs between dimensions of well-being that would be obscured by reducing outcomes to a single index.

Second, we contribute to the literature on the early-life origins of lifetime inequality. Structural models of life-cycle outcomes have emphasised the importance of initial conditions and persistent individual heterogeneity relative to shocks experienced during adulthood (Keane and Wolpin 1997; Huggett et al. 2011). A complementary literature documents the importance of childhood cognitive and socio-emotional skills and environments for later economic, health and social outcomes (Carneiro and Heckman 2003; Cunha and Heckman 2008; Cunha, Heckman, and Schennach 2010; Heckman, Humphries, et al. 2010; Heckman, Pinto, et al. 2013; Goodman et al. 2011; Gensowski 2018; Papageorge et al. 2019; Savelyev 2022; Bolt et al. 2023). Much of this research examines outcomes at particular ages or studies the evolution of individual outcomes separately. Smith (2009), for example, investigates the relationships between childhood health and several adult economic trajectories, but considers each outcome separately. We extend this literature by relating rich measures of childhood skills and environments to the joint evolution of multiple dimensions of well-being throughout adulthood. In contrast to approaches that represent initial conditions through latent endowments, we examine which observable childhood characteristics contain information about individuals' subsequent lifetime profiles. By combining this analysis with out-of-sample prediction, we provide evidence on how early persistent multidimensional disadvantage becomes statistically identifiable.

These findings have implications for policies aimed at reducing persistent disadvantage. The identification of a group experiencing severe economic, physical-health and mental-health difficulties throughout adulthood suggests that policies addressing individual dimensions of deprivation in isolation may be insufficient to meet the needs of the most disadvantaged individuals. Moreover, the substantial predictability of membership in this group from childhood characteristics points to potential opportunities for earlier identification and support. These findings are particularly relevant given recent increases in mental-health-related disability and economic inactivity, although the experiences of

the 1958 birth cohort cannot necessarily be extrapolated to younger generations. More broadly, our analysis illustrates how a multidimensional life-cycle perspective can inform the understanding of inequality and the design of policies intended to address it.

The remainder of the paper is organised as follows. Section 2 describes the data and the construction of childhood skills and environments. Section 3 presents the empirical strategy, including multi-channel sequence analysis, clustering and prediction methods. Section 4 characterises the adult life-cycle profiles. Section 5 examines their associations with childhood skills and environments, and Section 6 presents the prediction results. Section 7 concludes.

2 Data

We employ data from the first nine waves of the National Child Development Study, a rich longitudinal dataset that tracks approximately 17,000 individuals born across England, Scotland and Wales during a single week in 1958, from birth to age 50. Childhood data are measured at birth and at ages 7, 11 and 16, while adult data are collected at ages 23, 33, 42, 46 and 50. We omit Wave 7 (age 46) from the analysis, however, as it does not record mental-health outcomes and, due to an administrative error, does not correctly record wage outcomes. Our analytical sample consists of all individuals with non-missing information on all variables used to measure early-childhood skills; this yields a sample size of 7,199 men and 6,809 women.

The study has been used extensively to examine the returns to human-capital investment in the UK over the life cycle. For example, Blundell, Dearden, Goodman, et al. (2000) and Blundell, Dearden, and Sianesi (2005) evaluate the wage returns to formal education. Carneiro, Crawford, et al. (2007) examine the impact of cognitive and non-cognitive skills on adult outcomes such as earnings, employment and health. However, these prior analyses fail to account for the joint relationship between wages and health, a gap that we address in this paper. To our knowledge, this is the first application of multichannel sequence analysis using the NCDS – and, for that matter, within the economics literature

– to examine the joint relationship between early-life skills and lifetime trajectories of earnings and health.

Table 1: Adult Descriptive Statistics

	(1) Age 23 Mean (S.D.)	(2) Age 30 Mean (S.D.)	(3) Age 42 Mean (S.D.)	(4) Age 50 Mean (S.D.)
Core Outcomes				
Hourly wage	5.55 (1.78)	11.90 (5.86)	15.57 (13.33)	18.26 (13.34)
Weekly earnings	72.99 (29.57)	466.76 (305.78)	615.70 (596.40)	706.79 (594.51)
Employed	0.70 (0.46)	0.81 (0.39)	0.84 (0.37)	0.82 (0.38)
Mental health	2.86 (1.23)	2.96 (1.16)	2.55 (1.35)	2.63 (1.39)
Physical Health	3.32 (0.68)	3.16 (0.71)	3.05 (0.78)	2.90 (0.77)
Age 23				
Education				
No qualification	0.64 (0.48)			
O-level	0.13 (0.33)			
A-level	0.10 (0.31)			
Post-secondary	0.13 (0.34)			
Out of labor force	0.12 (0.33)			
Employee/self-employed	0.06 (0.24)			
Above avg. no. number of jobs	0.47 (0.50)			
Social class	3.02 (0.82)			
Drinks daily	0.21 (0.41)			
Smokes daily	0.55 (0.50)			
Age 33				
Married		0.61 (0.49)		
Has a partner		0.78 (0.41)		
Homeowner		0.75 (0.43)		
Number of children		1.35 (1.03)		
Age 42				
Employee/self-employed			0.15 (0.36)	
Questioned by police			0.21 (0.40)	
Convicted criminal			0.06 (0.23)	
Goes out once per week			0.86 (0.35)	
Partner weekly pay			300.22 (505.64)	
Age 50				
Drinks daily				0.23 (0.42)
Chronic illness				0.19 (0.39)
Disabled at age 50				0.06 (0.24)
Married				0.65 (0.48)
Divorced				0.18 (0.39)
Homeowner				0.80 (0.40)
Health problems				0.25 (0.43)
Observations	14008	14008	14008	14008

Results from multiply imputed dataset. Standard deviations are reported in parentheses. Social class is a six-point scale based on the individual's occupation (the lower score is associated with higher skilled occupations). The standard occupational classification had not been developed at this point, which is why occupation is not included.

This section proceeds in three steps. First, we document the adult labour-market and health outcomes that serve as the main outcome variables in our analysis and form the basis of our trajectories. Second, we outline our strategy for identifying early-life skills and demographic factors, which we later relate to adult trajectories. Third, we describe our approach to handling potentially non-random missing data in the NCDS.

2.1 Earnings and Health Outcomes

In our framework, trajectories are defined as a tuple of labour-market and health outcomes observed throughout an individual's life cycle. For simplicity, we categorise labour-market outcomes using wages and employment status. Employment is a dummy variable that captures whether an individual currently participates in either wage employment or self-employment. Thus, the set of individuals not in employment encompasses the unemployed, those in education and those out of the labour force.

Our measure of earnings is the natural logarithm of gross hourly earnings deflated to 2019 prices. Our methodology for analysing socioeconomic trajectories over the life cycle requires the discretisation of all outcome data (the strategy is discussed in detail in Section 3). For this reason, we divide wages into within-year quintiles. Since earnings are undefined for individuals who are not employed, we define a final labour-market variable that combines the employment indicator and discretised earnings into a single variable with six categories: the five earnings bins and one non-employment bin.

Health is defined along two dimensions: physical and mental. Physical health is self-reported using a four-point scale.[1] Mental health is proxied by the individual's Malaise Inventory score on a scale from 1–9. The Malaise score was developed by Rutter et al. (1981) and attempts to assess psychological distress, including symptoms of depression, anxiety and overall emotional discomfort. Our final mental-health variable collapses the Malaise score into a five-point scale that ensures reasonable sample sizes across bins and aligns with theoretically indicated depression thresholds. Each of our outcomes is scaled such that a higher value implies higher earnings or a higher level of overall health.

Both health measures are self-reported, raising potential concerns about reporting error and differences in how individuals interpret and respond to questions. Nevertheless, self-reported physical health is widely used in empirical research and has been shown to contain meaningful information about underlying health, although reporting differences may affect

[1] Possible responses are "poor", "fair", "good" and "excellent". Wave 8 (age 50) adds a further category ("very good"), which we combine with "excellent" in our analysis.

comparisons across individuals (**baker˙what˙2004**). The Malaise Inventory is an established measure of psychological distress, with evidence supporting its validity in general population samples (**rodgers˙validity˙1999**). Moreover, **ploubidis˙longitudinal˙2019** provide evidence on the longitudinal comparability of mental health assessments in British birth cohorts, including the NCDS.

Table 1 documents the raw distributions of earnings, physical health and mental health across the adult waves that define our lifetime trajectories. Although our main outcome of interest is hourly wages, we include annual earnings in the sample to show that our general trends hold after accounting for intensive-margin labour-supply effects. There is a clear increase in hourly and annual earnings alongside employment. In contrast, mental and physical health appear to deteriorate over the life cycle, although mental health appears to increase slightly between the ages of 42 and 50.[2]

The main stage of our analysis involves clustering the trajectories derived from the wage, employment and health data described above. To describe the characteristics of workers following particular trajectories, we collect demographic data on individuals across their life cycle, which we report in the bottom half of Table 1. The NCDS contains a wealth of data on individuals at each wave, so we select information within each wave that reflects key decisions individuals make at that stage of their lives. At age 23, we record data on standard labour-market variables such as education, labour-force participation and job spells. We also collect information on lifestyle choices that are likely to affect health, including alcohol consumption and smoking.

At age 33, we select variables that reflect individuals' family circumstances, including whether they are married and whether they have children. At age 42, we supplement these with information on the individual's criminal record, social life and partner's income. Finally, at age 50, we collect broad information on individuals' socioeconomic outcomes that are likely either to be outcomes of, or factors influencing, health and wage trajectories. These factors include alcohol consumption, chronic illnesses, health problems, disabilities,

[2]The increase in mental health at age 50 relative to age 42 reflects the change in mental-health questioning from a 21- to a 9-point Malaise scale.

marital status and home-ownership status.[3]

2.2 Childhood Skills and Demographic Factors

Next, we outline our selection of childhood variables that we suspect determine lifetime trajectories of earnings and health. We derive two key sets of variables separately from data at ages 7, 11 and 16. The first comprises a set of individual-level skills that have each been shown to have positive and significant effects on labour-market outcomes: cognitive skill, internalising behaviour and externalising behaviour (see, for example, Deming (2017), who argues for the increasing importance of social skills in the labour market). To fix ideas, we define cognitive skills as an individual's mental abilities that enable them to process information, reason and solve problems. Internalising behaviour is associated with withdrawal, unforthcomingness and depression, while externalising behaviour is associated with anxiety about acceptance by peers, hostility towards peers, hostility towards adults, anxiety, restlessness and impulsive acting out.

Since skills are not readily observed in the data, we adopt a latent-factor approach to estimating them that is commonly used in the literature (Carneiro, Crawford, et al. 2007; Bolt et al. 2023). The primary assumption underlying these models is as follows: skills and environments are described by a vector F of unobserved latent factors, and each factor, θ^f, is associated with M^f imperfect but observed proxy measures, T_j^f:

$$T_j^f = \mu_j^f + \alpha_j^f \theta^f + \epsilon_j^f, \qquad j = 1, \ldots, M^f. \tag{1}$$

where α_j^f denotes the estimated factor loading, μ_j^f is a constant and ϵ_j^f is a mean-zero error term assumed to be independent across all individuals and measures, as well as independent of all latent factors in F. For example, cognition at age 7 is defined by the latent factor θ^f, for which the M^f measures comprise maths, reading and design/copying scores. Externalising and internalising skills draw on relevant items from the Bristol Social

[3]The variables reported in Table 1 constitute the final specification chosen for our analysis. Many variables, such as the highest level of education, are recorded at all waves. We have repeated the analysis using permutations of variables recorded at different ages and find no substantive changes to our results.

Adjustment Guide (BSAG) and the Rutter Scale, both of which are frequently used in the economics literature and have previously been validated by Carneiro, Crawford, et al. (2007).[4] Typically, internalising and externalising skills yield measures of maladjustment across various domains. We therefore invert their rankings such that our non-cognitive factors capture positive skills. Table A01 outlines the observed measures used to define each skill.

It should be noted that, due to data limitations, not all relevant items are available at each age. Thus, care is required when making inter-age comparisons, as some factors are missing at certain ages, while the sets of items used to measure other factors vary by age. These limitations are discussed in more detail in the results section.

The latent-factor method has the dual advantages of correcting for measurement error and achieving dimensionality reduction in a flexible way.[5] By contrast, the alternative of simply averaging over the relevant set of measures offers far less flexibility, as it imposes equal weighting on all measurements for a given factor.

The second step is to obtain measures of the environments to which children are exposed in early life. We undertake this part of the analysis in two stages. First, we identify broad categories of childhood environments used in the related literature; these include family socioeconomic status, school quality, housing quality, family demographic characteristics and parental investments (see, for example, Bolt et al. (2023)). In the case of variables for which we have no clear precedent for measurement, notably those capturing exposure to various types of socioeconomic household shocks at age 7, we use exploratory factor models to select appropriate measures. This step is explained in more detail in Appendix B, and the full list of environments and the variables used to characterise each is presented in Table A02.[6]

[4] Other studies, including McDermott and Watkins (1981) and Behar and Stringfield (1974), have demonstrated the validity of these factors in other samples.

[5] The framework described above is appropriate for continuous measures; we modify it accordingly for use with categorical and binary measures using item response theory models.

[6] As the majority of the measures that we include are categorical or binary in nature, we use an item response theory framework to estimate latent environmental characteristics at each age. The results of this estimation process are shown in Tables A03 - A17.

2.3 Multiple Imputation by Chained Equations (MICE)

Attrition is substantial in the adult waves of the NCDS and is unlikely to be completely random, so restricting the analysis to complete cases risks both biased estimates and reduced precision. We therefore use multiple imputation by chained equations (MICE), following Silverwood et al. (2024), to impute missing observations. We generate 30 imputed datasets using the variables included in our analysis alongside auxiliary variables that predict missingness, and assume that missingness is random conditional on this rich set of observed characteristics. All results reported throughout the analysis are based on the imputed data; Appendix C provides full details of the imputation procedure and compares the raw and imputed adult outcomes.

3 Empirical Strategy

3.1 Multi-Channel Sequence Analysis

The next step is to describe the heterogeneity in the joint trajectories that characterise individuals' experiences of physical and mental health, wages and employment over the life cycle. This is where our paper departs from the existing literature, which either takes a reduced-form approach by focusing on the impact of early-life skills on different aspects of life-cycle outcomes *independently* (Carneiro, Crawford, et al. 2007) or employs a structural framework to model explicitly the joint distributions of earnings and health (CITE).

We address this gap in the literature by employing sequence analysis, which aligns closely with the reduced-form approach outlined above but allows us to incorporate the joint distribution of earnings and health into a regression framework. The goal of sequence analysis is to cluster individuals based on their time series, or "trajectories", of labour-market earnings and health outcomes. Since the labour-market opportunities available to men and women differed significantly in the UK during the twentieth century, we conduct our analysis separately for men and women.

To motivate the importance of accounting for this joint relationship in our analysis, Table 2 presents a correlogram of hourly wages, physical health and mental health over the life cycle. Not only are individual channels highly correlated over time, but earnings and health outcomes are also correlated with one another over the life cycle. For example, the correlation between wages and mental health at age 50 is 0.35. Studies that do not account for these joint relationships – or that simply study unidimensional outcomes such as wages – when examining lifetime inequalities are therefore unlikely to achieve a complete understanding of well-being.

Table 2: Correlation Between Outcomes Over Time

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Log Hourly Wages												
(1) Age 23	1.00											
(2) Age 30	0.29	1.00										
(3) Age 42	0.27	0.48	1.00									
(4) Age 50	0.25	0.43	0.54	1.00								
Physical Health												
(5) Age 23	0.12	0.15	0.15	0.16	1.00							
(6) Age 30	0.15	0.19	0.19	0.21	0.40	1.00						
(7) Age 42	0.14	0.18	0.26	0.27	0.32	0.45	1.00					
(8) Age 50	0.15	0.21	0.27	0.35	0.28	0.40	0.52	1.00				
Mental Health												
(9) Age 23	0.19	0.22	0.22	0.21	0.36	0.26	0.25	0.24	1.00			
(10) Age 30	0.18	0.20	0.21	0.24	0.25	0.38	0.32	0.30	0.55	1.00		
(11) Age 42	0.13	0.14	0.19	0.20	0.20	0.28	0.39	0.32	0.45	0.56	1.00	
(12) Age 50	0.13	0.15	0.19	0.23	0.19	0.23	0.30	0.37	0.40	0.46	0.55	1.00

Pairwise correlation of trajectory outcomes over the lifecycle. Unemployed earnings are set to zero. The correlation is constructed using the first multiple imputation dataset.

The first step in our sequence analysis is to characterise the dissimilarity between two trajectories. As Studer and Ritschard (2016), alongside Settersten and Mayer (1997), observe, there are three main dimensions along which one life trajectory may differ from another. First, spell duration refers to the number of consecutive periods in which an individual experiences a given state. For example, two individuals who both enter the top quartile of the earnings distribution at least once in their lives are likely to have more similar lifetime trajectories to one another than either is to an individual who never does so.

Second, sequencing refers to the order in which individuals experience distinct states. For example, one could compare an individual who enters tertiary education, which may limit earnings in the short run in return for higher earnings in the long run, with an individual who enters the labour market after completing secondary school, leading to higher earnings in the short run but potentially lower wages in the long run.

Third, timing refers to the age at which an individual experiences a given state; in other words, *when* two individuals enter the top quartile of the earnings distribution.

Each of these dimensions is incorporated into a dissimilarity matrix through optimal matching (Abbott and Forrest 1986). Its distinguishing feature, compared with other forms of sequence analysis, is its use of "edit distance" as a measure of dissimilarity between sequences. The edit distance between two individual sequences is the minimum total cost of the operations – insertions, deletions and substitutions – required to convert one sequence into another. For example, editing the sequence "C B A" to become "A B B C" requires one insertion and two substitutions. If the cost of each operation is set to unity, the edit distance between the two sequences is three.

Computing the dissimilarity between two such sequences therefore requires us to specify the costs of these operations. In practice, these choices are shaped by which aspects of the empirical trajectories we consider relatively more important. In particular, setting insertion and deletion (indel) costs high relative to substitution costs places greater weight on the timing of spells, while the reverse places relatively greater weight on spell duration. In our balanced-panel setting, all sequences are of equal length; thus, substitutions are the only relevant operation to consider. Following the literature, we allow substitution costs to depend on the empirical transition probabilities between states observed in our data. Specifically, we define the substitution cost between states i and j as follows:

$$C_{ij} = 2 - p\left(s_i \mid s_j\right) - p\left(s_j \mid s_i\right). \tag{2}$$

where $p\left(s_j \mid s_i\right)$ denotes the transition probability from state i to state j. This may be

computed as:

$$p\left(s_j \mid s_i\right) = \frac{\sum_{t=1}^{L-1} n_{t,t+1}\left(s_i, s_j\right)}{\sum_{t=1}^{L-1} n_t\left(s_i\right)}. \tag{3}$$

where $n_t\left(s_i\right)$ is the number of non-terminal sequences in which state s_i is experienced at time t, $n_{t,t+1}\left(s_i, s_j\right)$ is the number of sequences in which state s_i is experienced at time t and state s_j is experienced at time $t+1$, and L is the maximum sequence length.

Table 3 presents the Markov transition matrices used to define the cost matrices in Equation 2. For simplicity, we report the results for males, as their labour-market trajectories more closely resemble those of a representative worker today.[7] Focusing on wages, the cost increases with distance from the initial state. For example, the probability of moving from unemployment to the first wage quintile is 0.25, compared with 0.9 for moving from unemployment to the top quintile. This reflects limited social mobility and the importance of initial conditions in determining life-cycle earnings (Adda and Dustmann 2023). We observe a similar pattern for mental and physical health, consistent with the results in Table 2, which suggest strong path dependence in life-cycle health outcomes.

The set of individual sequences, combined with this characterisation of substitution costs, allows us to derive a dissimilarity matrix that quantifies the relative distance between each pair of sequences for a single socioeconomic outcome. In our application, however, we are interested in measuring dissimilarities between individual life trajectories along multiple dimensions rather than a single dimension: namely, wages, physical health and mental health. Consequently, we employ a variant of this technique termed multi-channel sequence analysis. This approach involves adding an additional step to the analysis described above. In particular, we first compute dissimilarity matrices separately for each channel using optimal matching and then sum these matrices across channels to obtain a single multi-channel measure of sequence dissimilarity. Thus, with K channels, the aggregated dissimilarity measure between sequences a and b, D_{ab}, is given by:

[7] Table **??** reports the corresponding results for women. The direction of the patterns is consistent with that observed for men; for earnings, this is partly because we define wage quintiles within gender.

Table 3: Male Transition Matrices

Panel A: Wages

	Destination state					
Origin state	Unemp.	Q1	Q2	Q3	Q4	Q5
Unemp.	0.281	0.246	0.156	0.124	0.107	0.086
Q1	0.142	0.282	0.251	0.160	0.103	0.063
Q2	0.091	0.222	0.227	0.224	0.149	0.088
Q3	0.074	0.145	0.208	0.214	0.213	0.145
Q4	0.067	0.093	0.144	0.217	0.236	0.243
Q5	0.065	0.074	0.116	0.193	0.243	0.309

Panel B: Physical Health

	Destination state			
Origin state	Q1	Q2	Q3	Q4
Q1	0.415	0.415	0.148	0.022
Q2	0.125	0.330	0.473	0.072
Q3	0.028	0.224	0.495	0.253
Q4	0.007	0.078	0.546	0.369

Panel C: Mental Health

	Destination state				
Origin state	Q1	Q2	Q3	Q4	Q5
Q1	0.352	0.236	0.213	0.130	0.068
Q2	0.209	0.187	0.263	0.218	0.123
Q3	0.107	0.147	0.224	0.296	0.226
Q4	0.044	0.085	0.218	0.262	0.390
Q5	0.017	0.040	0.134	0.330	0.479

Notes: Entries report transition probabilities between adjacent observed states using the multiply imputed dataset. Rows denote the origin state and columns the destination state. Physical and mental health states are ordered such that higher values correspond to better health.

$$D_{ab} = \sum_{k=1}^{K} D_{ab}^{k}. \tag{4}$$

Since missingness in our data is addressed using MICE, we have multiple observations for individuals at any given time period. We must account for this appropriately when constructing the distance matrix, and the literature does not provide a clear solution. Our approach is to calculate the distance matrix using all 30 imputations simultaneously, appended into a single large dataset. Under this approach, the distance between complete cases will be zero, but positive differences may arise between imputed trajectories, reflecting uncertainty about their true values.[8]

3.2 Clustering

The next step in our analysis is to cluster the observed multi-channel sequences using the dissimilarity measure computed in the previous step in order to identify distinct "types" of socioeconomic trajectories. Armed with this classification, we can explore the early-life determinants of outcomes across the life cycle.

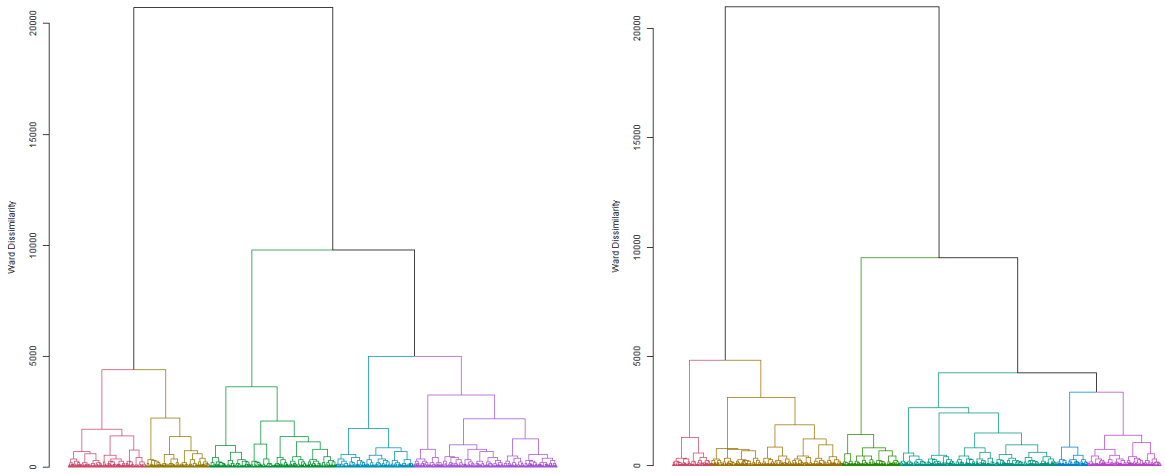
For this purpose, we employ Ward's hierarchical clustering, an unsupervised classification technique widely used in the economics and sociology literature (CITE). We favour hierarchical clustering over alternatives such as k-means, which require the number of clusters to be specified ex ante, because we hold no priors over the distribution of trajectories. Other classification methods, such as random forests (see Cengiz et al. 2022), are also unsuitable in our context because they rely on a large vector of features to predict cluster membership, whereas our methodology relies solely on the distance matrix defined above.

The algorithm proceeds as follows. We begin with n clusters, with each sequence in a

[8] We refrain from calculating the distance measure separately for each imputed dataset because the next step is to conduct a clustering exercise that classifies individual sequences into a small number of discrete groups. If the distance matrices were defined separately at the imputation level, we would need to compute the clustering algorithm separately for each imputation. Since the definition of a cluster is a purely data-driven process that aims to minimise within-group variance and maximise between-group variance, clusters have little ex ante economic meaning, much like latent factors in factor models, and are responsive to changes in the underlying dataset. Hence, it is likely that, for example, the first cluster would not be comparable across imputed datasets, rendering subsequent analysis inaccurate.

cluster of its own, and then identify the two "closest" clusters and merge them into a single cluster, leaving $n-1$ clusters. Identifying the two closest clusters involves computing the merging cost of all possible cluster mergers. The merging cost of any two clusters, A and B, is simply the increase in the sum of squared distances caused by the merge, as measured using our distance matrix. We repeat this process until only a single cluster containing all n observations remains.

Figure 1 represents the clustering process as a dendrogram. The x-axis records each of our observations, while the y-axis represents the within-cluster sum of squared distances (Ward distance). Initially, each observation is represented by a node in a separate cluster and the Ward distance is zero. As observations are progressively merged into clusters, the Ward distance increases. A large increase in Ward distance between two successive splits is indicative of an economically meaningful cluster structure.



(a) Male (b) Female

Figure 1: Cluster Dendrograms

Notes: Results from multiply imputed dataset clustered as one object.

The final step is to select the optimal number of clusters. A larger number of clusters implies that a greater fraction of the underlying variation in the data is explained. However, adding additional clusters comes at the cost of interpretability. Our goal is to trade off these considerations by examining four measures of cluster quality commonly employed in the literature.[9]

[9]We avoid calculating silhouette measures in the context of multidimensional sequence analysis because, in high-dimensional problems, the measure quickly loses explanatory power.

First, we calculate the point-biserial correlation (PBC), which measures how well the clustering structure aligns with the observed distances in the dataset. This metric computes the correlation between cluster assignments (whether a pair of observations belongs to the same or different clusters) and pairwise distances, thereby assessing whether points within the same cluster tend to be closer together than points in different clusters. A higher PBC value indicates more compact and well-separated clusters.

Next, we compute Hubert's C index, which evaluates clustering quality by comparing observed pairwise distances with the clustering assignments. A lower value of Hubert's C indicates that the clustering successfully groups closer observations together while keeping different clusters well separated.

To complement Hubert's C, we report Hubert's Gamma (HG), which assesses whether the ranking of pairwise distances is consistent with the clustering structure. Unlike Hubert's C, which considers the magnitude of distances, Hubert's Gamma assesses whether pairs of observations that are closer together are more likely to be assigned to the same cluster. A higher value of Hubert's Gamma therefore indicates that the clustering better preserves the natural ordering of distances.

Finally, we calculate R^2, which captures the share of total variation accounted for by the clustering. This is the only measure that does not directly penalise the addition of new clusters, but it provides a familiar and tractable metric for assessing how well our clusters fit the data.

Given the high-dimensional nature of our problem and the inherent lack of structure in socioeconomic data, we define a strong clustering solution, in line with the literature, as one in which the PBC and Hubert's Gamma exceed 0.3 (both range from -1 to 1) and Hubert's C is below 0.3. Many combinations of clusters between 2 and 10 satisfy these criteria, so we select the solution that best balances interpretability and cluster strength. We select five clusters for males and six clusters for females. Throughout our analysis, we refer to these clusters as trajectories. Figure A01 plots these metrics across different numbers of clusters for males and females, while Appendix D formally

defines each cluster-quality metric. The coloured branches of the dendrograms in Figure 1 represent the optimal individual cluster allocations. We also observe that any merges beyond our chosen splits lead to large increases in Ward distance, suggesting that our selected clusters are well chosen.

3.3 Multinomial Logit

The next step in our analysis is to understand the early-life factors influencing later-life trajectories in earnings and health. Given our discrete number of clusters, K^s, for each sex, s, we choose to employ multinomial logits to estimate this relationship. The model is grounded in the assumption that individuals, i, enjoy linear utility from being in a cluster, such that $U_{ik} = X_i\beta_k + \epsilon_{ik}$, where ϵ_{ik} is i.i.d. and follows a type I extreme-value distribution. In our analysis, X_i refers to the set of early-childhood skills and factors outlined in the data section. The coefficients are identified in our exploratory analysis under the assumption that the error terms are uncorrelated with cluster membership. While we do not pose our analysis as causal, our rich set of early-life skills and demographic controls lends support to this assumption.

We are in the fortunate position of observing individuals born in a specific year and following them over the same life cycle, implying that cohort-specific effects are irrelevant. While we would ideally control for individual fixed effects in our analysis, our sequence-analysis approach means that we only have one outcome per individual. However, standard multinomial models suffer from the incidental-parameter problem, making estimation of fixed effects difficult unless T is very large. Moreover, our rich data broadly capture the underlying characteristics that most studies aim to capture by including individual fixed effects. We are in the advantageous position of being able to quantify the effects of factors that are typically unobserved on life-cycle outcomes.

A limitation of using the multinomial logit model is its reliance on the independence of irrelevant alternatives (IIA) assumption, which states that the odds ratio between any two outcome categories is unaffected by the inclusion or exclusion of other categories. In

our case, although the clusters representing trajectories of health and earnings are derived from inherently ordered outcomes that might exhibit some degree of correlation, we argue that this potential violation is mitigated by conditioning on an extremely rich set of covariates—such as cognitive and non-cognitive skills, along with detailed demographic factors. By capturing the major observable determinants of cluster membership, we assume that the remaining unobserved factors are independent across clusters, ensuring that the conditional odds ratios are invariant to the presence of additional clusters. This conditional independence justifies the use of the multinomial logit model in our analysis despite the inherent ordering in the original outcomes.

3.4 Prediction Exercise

To examine how well early-life skills and environments predict lifetime trajectories, we compare predictive models using a 60:20:20 training, validation and test split. The training sample is used to estimate the models, the validation sample to assess model choices, and the held-out test sample to evaluate predictive performance. We begin with the multinomial logit outlined above, but also consider non-linear machine-learning models that can capture relationships not accommodated by the multinomial specification. In particular, we focus on neural networks.

Neural networks are non-linear function approximators designed to mimic the human brain. The model consists of an input layer, which in our case contains early-life skills and environments. The input layer is fed through hidden layers before being mapped to an output layer. In our case, the output layer has six nodes for females and five nodes for males, each corresponding to one of our clusters.

The structure of the hidden layers is treated as a set of hyperparameters and is therefore chosen by the researcher. Each hidden layer consists of a series of neurons that are activated when data pass through them. For example, if we have a neural network with one hidden layer consisting of an arbitrary number of nodes and a given node represents a certain characteristic of an individual, such as intelligence, then that node would return

a high value when data on an individual with high cognitive test scores are input into the system. Strong activation of this neuron makes it more likely that the system predicts that the individual belongs to a cluster with other individuals with high cognitive test scores. In practice, the meaning of the neurons is data-driven and cannot be labelled as "intelligence", much as one cannot easily ascribe meaning to latent factors in factor models.

After choosing the desired number of hidden layers, $l = 1, \ldots, L$, and nodes, $\alpha^{(l)}$, the model is estimated via gradient descent. To this end, we define a loss function to measure the prediction accuracy of the neural network and an activation function for each of the nodes. Our activation function of choice is the sigmoid function (identical to that used in the logit model). The first hidden layer takes a linear combination of the covariates and passes it through the sigmoid function to activate that neuron to a value between zero and one, determined by the set of weights $\omega^{(1)}$, including a constant (referred to as a "bias" term in the machine-learning literature):

$$\alpha^{(1)} = \Lambda\big(\omega^{(1)} X\big)\,. \tag{5}$$

For each subsequent hidden layer, the activation of nodes is defined by:

$$\alpha^{(l)} = \Lambda\big(\omega^{(l)} \alpha^{(l-1)}\big)\,. \tag{6}$$

The final layer, $\alpha^{(L)}$, consists of a number of nodes equal to the number of clusters. The prediction of the neural network is the node with the highest activation. The loss function, referred to in the literature as categorical cross-entropy, penalises incorrect predictions and is defined in the draft as follows:

$$\text{Loss} = \sum_{i} \sum_{k \in K^s} \mathbb{1}[i \in k] \log(y_{ik}). \tag{7}$$

Here, y_{ik} is the predicted probability that individual i belongs to cluster k. The model aims to minimise the loss function by sequentially updating $\omega^{(l)}$ until the function reaches a minimum. In practice, gradient descent is likely to reach local minima because neural networks are non-linear functions; we seek to address this by employing stochastic gradient descent.

To account for uncertainty introduced by multiple imputation, we train a separate neural network on each of the 30 imputed datasets. Each fitted model produces a predicted trajectory class for a given individual. We combine these classifications using a modal vote: the final predicted class is the class most frequently selected across the imputation-specific models. Writing $\widehat{k}_i^{(m)}$ for the predicted class of individual i from the network fitted to imputation m, the ensemble prediction is

$$\widehat{k}_i^{\text{ens}} = \arg\max_{k \in K^s} \sum_{m=1}^{30} \mathbb{1}\left[\widehat{k}_i^{(m)} = k\right]. \tag{8}$$

This is a multiple-imputation ensemble with hard (majority or plurality) voting: it aggregates predictions rather than pooling the networks' estimated weights or coefficients.

4 Results

4.1 Relating Early-Life Skills to Later-Life Outcomes

To illustrate why it is important to evaluate the impact of early-life skills across trajectories, we first present results from a more standard unidimensional approach, regressing each outcome at age 42 on skills at age 11 and early-life environments. We choose outcomes at age 42 because this is prime working age, when individuals are likely to have their strongest attachment to the labour force and to be near the peak of their earnings profiles. Figure 2 presents the results from these regressions graphically. In each case, both the independent and dependent variables are standardised to have mean zero and variance one, except for employment, which we specify using a linear probability model. For the purposes of this introductory analysis, we do not present separate results for men

and women.

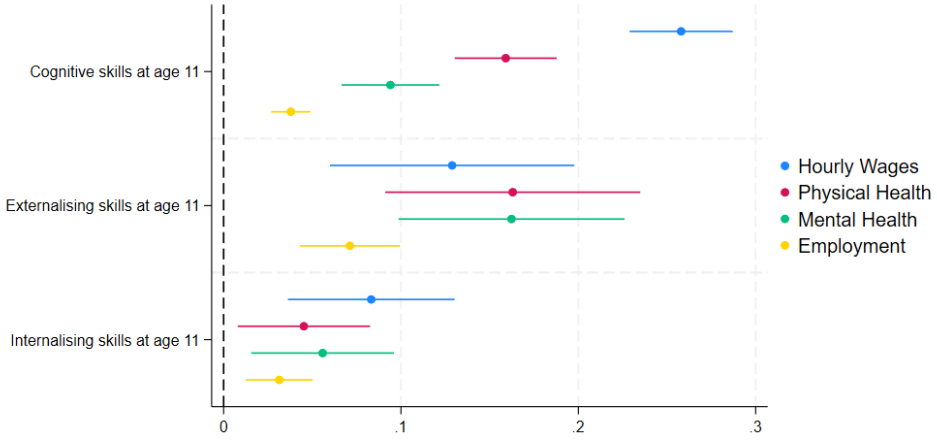
Panel A of Figure 2 presents the associations between cognitive, externalising and internalising skills and wages, employment and health. Each skill is positively and significantly associated with each outcome. The strongest relationship is between cognitive skills and wages, which likely reflects selection into higher education. However, cognitive skills have a more muted association with health outcomes and employment than with wages. The relatively large association between externalising skills and employment, compared with cognitive skills, speaks to the growing literature suggesting that academic skills are becoming weaker predictors of labour-market outcomes.

Internalising skills have the smallest association with later-life outcomes across all dimensions. However, their associations remain positive and significant: for example, a one-standard-deviation increase in age-11 internalising skills is associated with wages that are 0.09 standard deviations higher. Overall, these results suggest that early-life skills play an important role in shaping later-life outcomes.

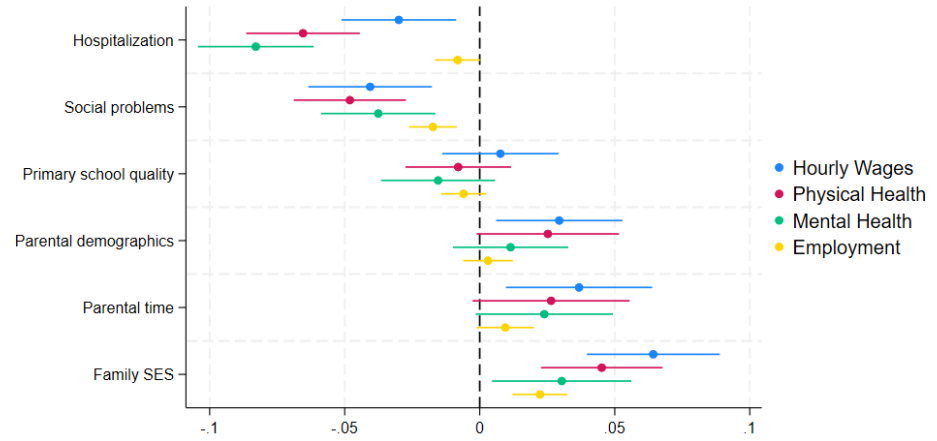
Turning to early-life environments in Panel B, we observe a more mixed picture: some environments play a pronounced role in determining outcomes, whereas others are statistically insignificant. Across environmental factors, those with larger absolute associations with one later-life outcome also tend to have larger associations with other later-life outcomes.

The factors most strongly associated with better later-life outcomes include family socioeconomic status and parental time devoted to children. The relationship between parental investment and children's later-life earnings has been widely documented, and the magnitudes of our estimates align with previous empirical work (Cunha and Heckman 2008). Together, these results provide a strong foundation for further analysis using our imputed dataset.

Not all environmental shocks are associated with higher earnings, employment and health. Social problems in the household, such as divorce or the death of a parent, are unam-



(a) Skills



(b) Environments

Figure 2: Skill and Environment Regressions at Age 42.

Notes: All skills and environments are standardised to have mean 0 and variance 1. Wages, physical health and mental health are also standardised to have mean 0 and variance 1. Employment is a binary indicator and the regression is interpreted as a linear probability model. The regressions are also conditional on a large set of demographic characteristics. In the wage regressions, non-employed individuals are set to have wage 0.

biguously negatively associated with all later-life outcomes. The factor with the largest absolute association with later-life health outcomes relates to childhood hospitalisation. If a child is hospitalised for an extended period or referred to a medical professional for specialist treatment, this is associated with long-lasting differences in adult outcomes.

In Figure A02, we plot the associations between skills and earnings, employment and health outcomes over the life cycle. The results confirm that our findings at age 42 are broadly representative of the associations observed at other ages. We also demonstrate that different skills vary in importance over the life cycle. For example, cognitive skills are much more strongly associated with mental health in early adulthood than in later life. In contrast, the association between cognitive skills and earnings becomes stronger over the life cycle.[10]

Nonetheless, these regressions fail to account for correlations between adult outcomes and the manner in which gains along one dimension may be reinforced or attenuated by losses along another. By modelling the joint distribution of earnings and health trajectories in adulthood, our methodology can shed light on these relationships and capture nuances that would otherwise be obscured.

4.2 Describing Trajectories

As a first step towards relating early-life environments to trajectories and deprivation, we describe and categorise the trajectory types defined in the previous section. Our clustering strategy suggests that there are five male and six female types. Table 4 presents the trajectory types alongside the relative rank of each outcome used to define the trajectories at age 42. The first salient feature is the strong association between outcomes: individuals of a given type who achieve a high rank in one outcome have a high probability of ranking highly in the others. This is consistent with the correlations between wages and health reported in Table 2 and is particularly pronounced among men, for whom outcomes are

[10]Figure A03 examines whether choosing age 11 as the focal point for skills provides either an upper or lower bound. We observe that age-11 skills are broadly representative of age-7 skills and age-16 cognitive skills. However, age-16 internalising skills are more strongly correlated with later-life outcomes, while age-16 externalising skills are less strongly correlated with them.

ranked almost uniformly across clusters. This also validates our clustering methodology, suggesting that we have isolated economically meaningful relationships.

Table 4: Trajectory Classification

Male						
Type	(1)	(2)	(3)	(4)	(5)	(6)
Classification	All-rounder	High Earner	Healthy	Relative Poverty	Deprived	–
Rank						
Wages	1	**2**	3	4	5	–
Mental Health	2	3	**1**	4	5	–
Physical Health	1	3	**2**	4	5	–
% Sample	6.0	29.8	47.0	12.0	5.2	–

Female						
Type	(1)	(2)	(3)	(4)	(5)	(6)
Classification	All-rounder	High Earners	Mentally Healthy	Physically Healthy	Relative Poverty	Deprived
Rank						
Wages	2	**1**	5	3	4	6
Mental Health	1	3	**2**	4	5	6
Physical Health	1	3	3	**2**	4	6
% Sample	9.6	28.2	11.0	17.5	24.8	9.0

Note: Rank refers to the mean rank of the type at age 42 within sex.

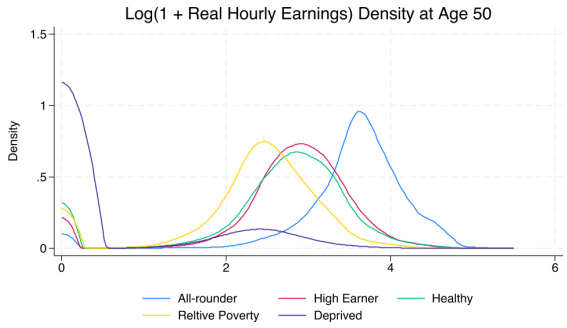
The strong rank correlation within types points to substantial inequality across individuals. Among both males and females, there is a small group that ranks either first or second across outcomes. As we show later, these "all-rounders" have significantly higher early-life skills than individuals in the remaining clusters, and their wage and health distributions almost always dominate those of the other types. In contrast, there are two types of men and women who rank below the median across all outcomes. We consider these individuals to be living in relative poverty. Together, these types account for 17.2% of men and 33.8% of women. Within this group, there is a specific type that ranks at the bottom across all outcomes. We define these individuals as living in a state of deprivation. They account for a non-trivial fraction of the sample, comprising 5.2% of men and 9.0%

of women. Explaining the origins of this deprived group is a focal point of our subsequent analysis.

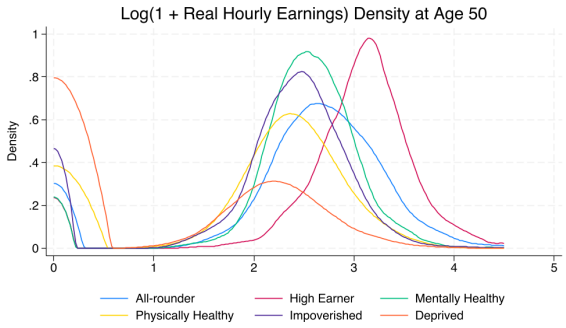
Women exhibit greater variation in outcome rankings within types than men, with some types ranking particularly highly in wages, mental health or physical health. This potentially reflects the socioeconomic opportunities available to women in our cohort. For example, the female type with the highest earnings is not the all-rounder but a larger group of women who achieve only median levels of health. These individuals are likely to be women in full-time employment, and we refer to them as "high earners". Likewise, we refer to women whose mental-health (physical-health) rank is substantially higher than their rank in the other outcomes as "mentally healthy" ("physically healthy"). In the male sample, we are unable to distinguish between separate mentally and physically healthy types. Instead, we distinguish only between "high earners", whose earnings rank exceeds their health ranks, and "healthy" individuals, for whom the reverse is true.

The ranks in Table 4 obscure differences in outcome levels and distributions in favour of a parsimonious ranking. Figure 3 presents the distributions of wages, physical health and mental health across trajectories at age 50, further motivating our choice of ranking. We deliberately focus on age 50 because, as we later, socioeconomic differences widen over the life cycle. Studying outcomes at age 50 therefore provides particularly clear insight into the extent of life-cycle inequality.

The left column of Figure 3 presents the results for men. We observe that the wage distribution of the all-rounder trajectory stochastically dominates those of all other types. To avoid mass points in the distributions, we code zero earnings as missing for this exercise. Moreover, the wage distribution of the all-rounder type is shifted substantially to the right relative to those of the other types, suggesting that this trajectory captures a well-defined subset of individuals who are likely to be in full-time, high-paying occupations. Their mean wage is approximately 0.5 log points higher than that of any other trajectory across both males and females. In contrast, the wage distributions of those in deprivation and relative poverty are shifted to the left relative to the other trajectories, illustrating the



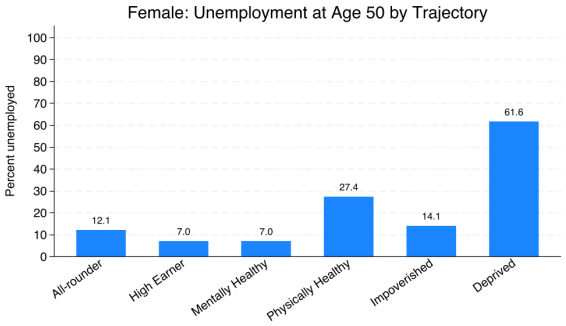
(a) Male: Log Hourly Wage



(b) Female: Log Hourly Wage



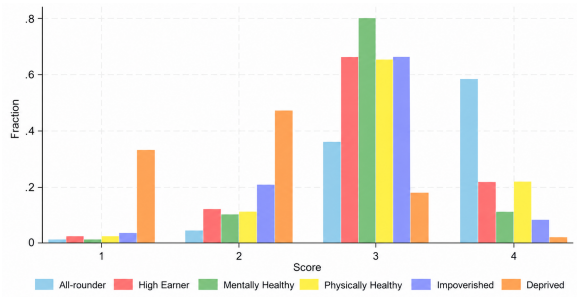
(c) Male: Unemployment



(d) Female: Unemployment



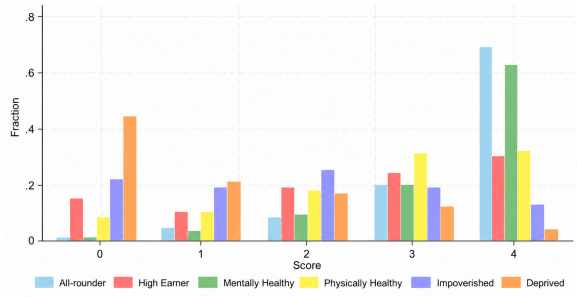
(e) Male: Physical Health



(f) Female: Physical Health



(g) Male: Mental Health



(h) Female: Mental Health

Figure 3: Distribution of Outcomes at Age 50 Across Trajectories

Notes: Results show distributions of outcomes by cluster for males (left) and females (right). The densities are defined treating the multiply imputed dataset as one object. Wages are dropped for individuals who do not receive a labour income.

extent of inequality. There is nevertheless overlapping support across the wage distributions of all types for two reasons. First, the kernel plots show outcomes at a specific point in time, so workers with high wages who belong to the deprived trajectory may be individuals who have only recently transitioned into employment. Second, workers with high wages may nevertheless be classified within the deprived trajectory because their health outcomes are substantially worse than those of other individuals.

For physical health, we observe a similar but more pronounced pattern. More than 40% of deprived workers report the lowest health category, while no other trajectory has a comparably large share of individuals reporting this outcome. Likewise, while approximately 40% of all-rounders report the highest health category, no deprived workers do so. Even workers in relative poverty report meaningfully higher levels of physical health than the deprived group. Although the mean wage is only slightly higher among high earners than among individuals in the healthy trajectory, there are large differences in health: 29% of the healthy type report the highest level of physical health, compared with only 16% of high earners.

The results for women (right column) are directionally similar to those for men, although the earnings of the deprived group are substantially lower both relative to those of other female trajectories and in absolute terms. Moreover, the highest-earning female trajectory is the specialised high-earner trajectory rather than the all-rounder trajectory, marking an important difference from the male classification. Female all-rounders nevertheless have substantially higher levels of health than individuals in any other trajectory. This makes comparisons between female trajectories less straightforward, as whether an individual would prefer the high-earner trajectory to the all-rounder trajectory depends on the relative weight placed on health and earnings in lifetime utility.

Next, we examine the dynamic outcomes associated with lifetime trajectories by plotting mean outcomes over the life cycle by trajectory in Figure 4. At this stage, we reintroduce employment because labour-force participation is an important indicator of economic activity and of socioeconomic choices over the life cycle, including childrearing. Wages

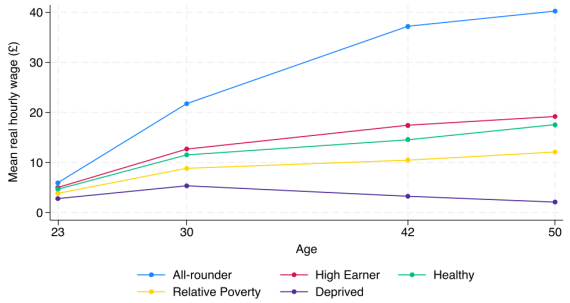
for individuals who are not employed are therefore coded as missing. The structure of Figure 4 mirrors that of Figure 3.[11]

Our first observation is that trajectories are associated with large and persistent differences in outcomes over the life cycle that widen with age. This is particularly true for wages, where wage growth is especially high among male all-rounders, who earn on average approximately 0.7 log points per hour more than individuals in the next-highest trajectory. As in the previous results, individuals in relative poverty earn substantially less than the overall mean. Among men, there is strikingly little difference in the hourly wages of deprived workers and those in relative poverty. However, the employment rates of deprived men and women are substantially lower than those of any other group. This is particularly pronounced among women at age 23, whose average employment rate is 31%. For both deprived men and women, employment falls markedly between ages 33 and 50. By age 50, only 20% of deprived men are employed. This suggests that the distinction between those living in relative poverty and those experiencing deprivation reflects not only the types of jobs in which individuals are employed, but also the fraction of their working lives that they spend in employment.

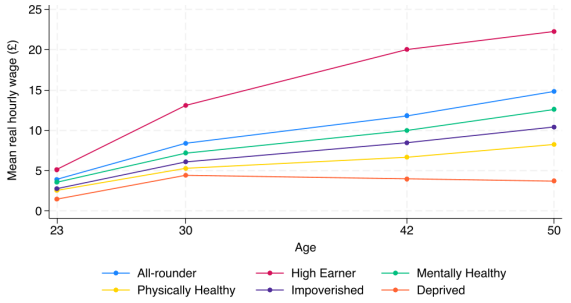
Physical health declines over the life cycle for both men and women. As before, the ordering is consistent with our classification, with all-rounders exhibiting the highest levels of health throughout the life cycle and the smallest percentage decline by age 50. Average physical health is next highest among the healthy trajectories, followed by those in relative poverty. Again, the deprived trajectory experiences a particularly large percentage decline in physical health for both men and women.

A similar pattern emerges for mental health, although healthy men exhibit slightly higher levels of mental health than all-rounders, despite earning substantially less. Of particular note is the extremely low level of mental health among deprived women, even relative to deprived men. Moreover, healthy men exhibit very little decline in mental health over

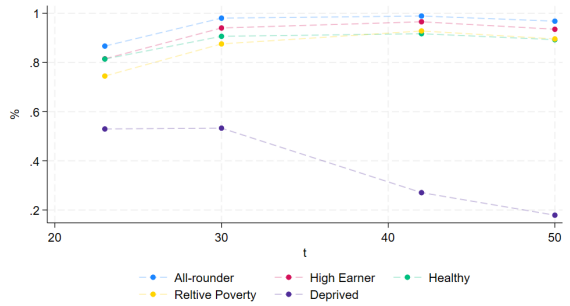
[11] The full state-distribution of outcomes (wages and employment, mental health and physical health) in each cluster over the lifecycle can be seen in Figures A04 - A06 for males and Figures A07 - A09 for females.



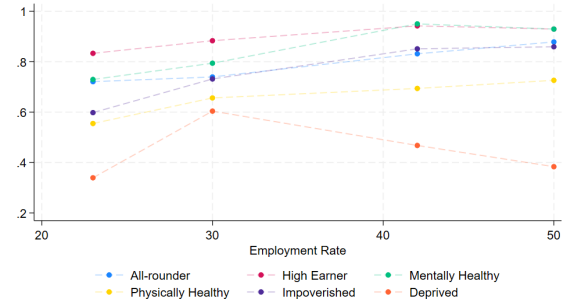
(a) Male: Hourly Wage (including zeroes)



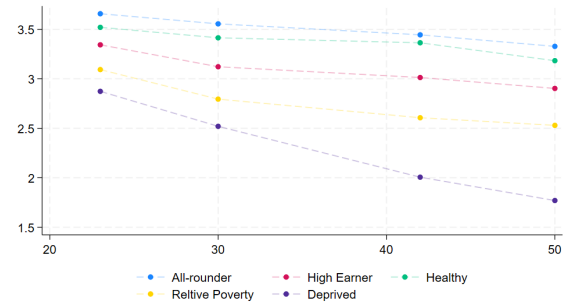
(b) Female: Hourly Wage (including zeroes)



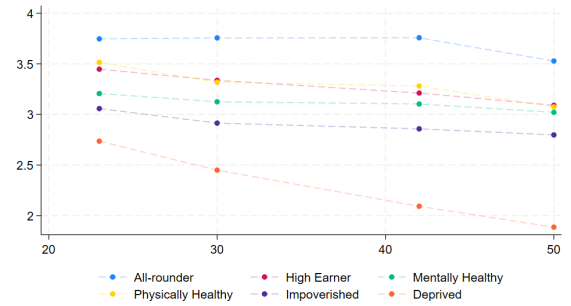
(c) Male: Employment



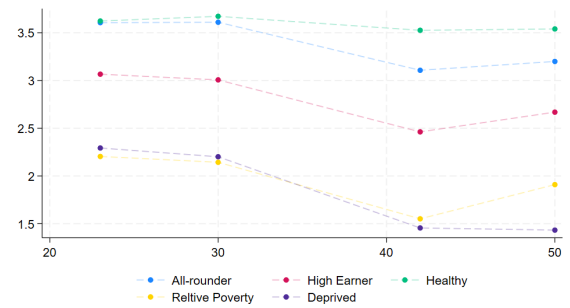
(d) Female: Employment



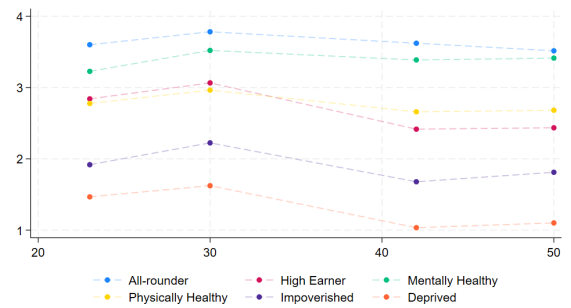
(e) Male: Physical Health



(f) Female: Physical Health



(g) Male: Mental Health



(h) Female: Mental Health

Figure 4: Average Outcomes Over the Life Cycle Across Clusters

Notes: Results show average outcomes by cluster for males (left) and females (right) over the lifecycle. The densities are defined treating the multiply imputed dataset as one object. Wages are set to zero for individuals who do not receive a labour income.

the life cycle. This suggests that the positive association between mental and physical health may hold only up to a point, beyond which there may be a trade-off between health and earnings. A final difference between male and female mental-health trajectories is that female mental health improves between ages 23 and 33, whereas male mental health declines. Given that this is a prime childbearing period for women, these results may suggest an association between childbirth and female mental health.

As a final step towards understanding the nature of these trajectories, Tables 5 and 6 report a set of descriptive statistics for individuals across trajectories. Education is strongly associated with cluster membership for both men and women. Although the average level of education is higher for men than for women, all-rounders have the highest share of individuals with post-secondary education in both groups, while 84% of deprived workers have no qualifications. Individuals in relative poverty and deprivation are not only less likely to be employed but are also more likely to be out of the labour force. In the deprived female trajectory, approximately half of all individuals are out of the labour force. Individuals with high levels of earnings and health are less likely to be self-employed, reflecting the typical wage premium earned by employees relative to self-employed workers. In addition to having lower employment rates, deprived individuals have held fewer jobs, suggesting that they face both fewer employment opportunities and weaker labour-market attachment. Social class also varies systematically across trajectories. In our context, social class is measured on a six-point scale based on occupation. We are unable to record exact occupations at age 23 because the relevant Standard Occupational Classification had not yet been introduced.

Examining health-related behaviours at age 23, we find that individuals in the deprived clusters are more likely to smoke on a daily basis. We also observe that all-rounders are the least likely to smoke, even compared with physically healthy individuals. Surprisingly, the proportion of individuals who drink alcohol daily is positively associated with earnings and health outcomes. However, this gradient is small, and the more notable difference is in the mean level of alcohol consumption between males and females. This may partly explain why average physical health at age 50 is substantially higher among females than

Table 5: Male Trajectory Descriptives

	(1) All-rounder Mean (S.D.)	(2) High Earner Mean (S.D.)	(3) Healthy Mean (S.D.)	(4) Relative Poverty Mean (S.D.)	(5) Deprivation Mean (S.D.)
Age 23					
Education					
No qualification	0.23 (0.42)	0.54 (0.50)	0.67 (0.47)	0.81 (0.40)	0.84 (0.37)
O-level	0.10 (0.30)	0.13 (0.33)	0.14 (0.35)	0.09 (0.29)	0.08 (0.28)
A-level	0.20 (0.40)	0.14 (0.34)	0.09 (0.29)	0.06 (0.23)	0.05 (0.22)
Post-secondary	0.46 (0.50)	0.19 (0.39)	0.10 (0.30)	0.04 (0.20)	0.02 (0.14)
Employed	0.87 (0.33)	0.84 (0.37)	0.80 (0.40)	0.68 (0.47)	0.54 (0.50)
Out of labor force	0.01 (0.08)	0.01 (0.10)	0.01 (0.12)	0.02 (0.16)	0.06 (0.24)
Self-employed	0.05 (0.21)	0.08 (0.28)	0.09 (0.29)	0.11 (0.31)	0.05 (0.21)
Above avg. no. jobs	0.27 (0.45)	0.42 (0.49)	0.48 (0.50)	0.62 (0.49)	0.66 (0.48)
Social class	2.33 (0.84)	2.81 (0.83)	3.09 (0.85)	3.37 (0.88)	3.49 (0.88)
Drinks daily	0.34 (0.48)	0.32 (0.47)	0.31 (0.46)	0.29 (0.45)	0.30 (0.46)
Smokes daily	0.35 (0.48)	0.48 (0.50)	0.55 (0.50)	0.70 (0.46)	0.76 (0.43)
Age 33					
Married	0.72 (0.45)	0.64 (0.48)	0.60 (0.49)	0.48 (0.50)	0.39 (0.49)
Has a partner	0.86 (0.35)	0.81 (0.39)	0.78 (0.42)	0.69 (0.46)	0.60 (0.49)
Homeowner	0.94 (0.24)	0.86 (0.35)	0.75 (0.43)	0.57 (0.50)	0.42 (0.49)
Number of children	1.20 (1.03)	1.22 (1.02)	1.26 (1.03)	1.17 (1.03)	1.06 (1.06)
Age 42					
Employed	0.99 (0.09)	0.97 (0.17)	0.91 (0.29)	0.81 (0.39)	0.27 (0.44)
Questioned by police	0.20 (0.40)	0.30 (0.46)	0.29 (0.45)	0.39 (0.49)	0.38 (0.49)
Convicted criminal	0.03 (0.18)	0.08 (0.27)	0.08 (0.28)	0.14 (0.35)	0.16 (0.37)
Partner's weekly salary	335.52 (414.55)	332.91 (457.63)	340.38 (453.90)	387.02 (1167.40)	386.92 (490.05)
Age 50					
Drinks daily	0.40 (0.49)	0.30 (0.46)	0.25 (0.43)	0.24 (0.43)	0.21 (0.41)
Chronic illness	0.04 (0.20)	0.13 (0.33)	0.12 (0.33)	0.30 (0.46)	0.78 (0.41)
Disabled	0.00 (0.06)	0.02 (0.15)	0.04 (0.20)	0.11 (0.31)	0.45 (0.50)
Divorced	0.10 (0.30)	0.15 (0.36)	0.17 (0.38)	0.23 (0.42)	0.27 (0.44)
Health problems	0.17 (0.38)	0.23 (0.42)	0.16 (0.37)	0.32 (0.47)	0.51 (0.50)
Observations	429	2151	3384	860	375

Results from multiply imputed dataset. Social class is a six-point scale based on the individual's occupation (the lower score is associated with higher-skilled occupations).

among males.

At age 33, we observe that deprived men are approximately half as likely to be married as all-rounders. The same pattern is observed among women, although to a lesser extent. A similar relationship holds for whether an individual has a partner, although, by construction, partnership rates are higher than marriage rates. Given the poorer earnings opportunities of individuals following deprived trajectories, we also observe that these individuals are less likely to own a home and are more likely to rent.

Table 6: Female Trajectory Descriptives

	(1) All-rounder	(2) High Earner	(3) Mentally Healthy	(4) Physically Healthy	(5) Relative Poverty	(6) Deprived
	Mean (S.D.)	Mean (S.D.)	Mean (S.D.)	Mean (S.D.)	Mean (S.D.)	Mean (S.D.)
Age 23						
Education						
No qualification	0.51 (0.50)	0.54 (0.50)	0.65 (0.48)	0.76 (0.43)	0.67 (0.47)	0.84 (0.37)
O-level	0.12 (0.32)	0.16 (0.37)	0.11 (0.31)	0.11 (0.32)	0.13 (0.33)	0.08 (0.28)
A-level	0.14 (0.35)	0.12 (0.33)	0.10 (0.30)	0.07 (0.26)	0.11 (0.32)	0.03 (0.17)
Post-secondary	0.23 (0.42)	0.18 (0.38)	0.13 (0.34)	0.06 (0.23)	0.09 (0.29)	0.04 (0.20)
Employed	0.74 (0.44)	0.66 (0.48)	0.75 (0.43)	0.50 (0.50)	0.66 (0.47)	0.32 (0.47)
Out of labor force	0.13 (0.33)	0.19 (0.40)	0.16 (0.36)	0.33 (0.47)	0.20 (0.40)	0.46 (0.50)
Self-employed	0.03 (0.16)	0.04 (0.18)	0.02 (0.14)	0.04 (0.20)	0.04 (0.20)	0.04 (0.19)
Above avg. no. jobs	0.38 (0.49)	0.46 (0.50)	0.41 (0.49)	0.50 (0.50)	0.49 (0.50)	0.51 (0.50)
Social class	2.78 (0.68)	2.92 (0.78)	2.90 (0.67)	3.23 (0.72)	3.02 (0.69)	3.39 (0.71)
Drinks daily	0.11 (0.31)	0.11 (0.31)	0.10 (0.29)	0.09 (0.28)	0.09 (0.29)	0.06 (0.24)
Smokes daily	0.38 (0.49)	0.55 (0.50)	0.53 (0.50)	0.59 (0.49)	0.56 (0.50)	0.74 (0.44)
Age 33						
Married	0.69 (0.46)	0.61 (0.49)	0.67 (0.47)	0.62 (0.49)	0.64 (0.48)	0.49 (0.50)
Has a partner	0.83 (0.38)	0.78 (0.41)	0.85 (0.36)	0.79 (0.41)	0.82 (0.38)	0.69 (0.46)
Homeowner	0.91 (0.29)	0.78 (0.42)	0.84 (0.37)	0.69 (0.46)	0.77 (0.42)	0.45 (0.50)
Number of children	1.40 (1.00)	1.38 (1.03)	1.42 (0.98)	1.65 (0.97)	1.51 (0.98)	1.73 (1.01)
Age 42						
Employed	0.83 (0.38)	0.83 (0.38)	0.95 (0.23)	0.65 (0.48)	0.86 (0.34)	0.45 (0.50)
Self-employed	0.11 (0.31)	0.11 (0.31)	0.08 (0.27)	0.10 (0.31)	0.10 (0.30)	0.09 (0.28)
Questioned by police	0.09 (0.29)	0.11 (0.31)	0.10 (0.30)	0.09 (0.29)	0.11 (0.31)	0.10 (0.30)
Convicted criminal	0.01 (0.11)	0.03 (0.16)	0.00 (0.07)	0.02 (0.15)	0.02 (0.13)	0.03 (0.17)
Goes out once per week	0.89 (0.31)	0.84 (0.37)	0.87 (0.34)	0.83 (0.38)	0.83 (0.37)	0.75 (0.43)
Partner's weekly salary	611.44 (946.95)	564.27 (622.78)	482.75 (543.86)	523.96 (954.93)	522.55 (720.26)	439.69 (501.70)
Age 50						
Drinks daily	0.21 (0.41)	0.20 (0.40)	0.18 (0.38)	0.16 (0.37)	0.18 (0.39)	0.11 (0.31)
Chronic illness	0.05 (0.22)	0.18 (0.38)	0.07 (0.25)	0.20 (0.40)	0.17 (0.37)	0.66 (0.47)
Disabled	0.02 (0.14)	0.04 (0.20)	0.01 (0.12)	0.06 (0.23)	0.03 (0.17)	0.26 (0.44)
Married	0.75 (0.43)	0.63 (0.48)	0.70 (0.46)	0.63 (0.48)	0.67 (0.47)	0.53 (0.50)
Divorced	0.12 (0.32)	0.19 (0.39)	0.18 (0.38)	0.22 (0.41)	0.20 (0.40)	0.27 (0.45)
Homeowner	0.94 (0.25)	0.84 (0.37)	0.91 (0.29)	0.75 (0.43)	0.84 (0.37)	0.54 (0.50)
Health problems	0.10 (0.30)	0.27 (0.44)	0.17 (0.37)	0.27 (0.44)	0.28 (0.45)	0.61 (0.49)
Observations	652	1920	748	1192	1686	611

Results from multiply imputed dataset. Social class is a six-point scale based on the individual's occupation (the lower score is associated with higher-skilled occupations). Cluster sample sizes are not reported in the supplied age-specific exports.

Up to this point, outcomes have generally followed similar patterns for men and women. However, a clear distinction emerges when we consider the number of children. Among men, the number of children increases systematically with earnings and health, whereas

among women it decreases. In part, this may reflect women's greater likelihood of leaving the labour market during periods of pregnancy and childrearing, as well as temporary constraints on physical activity associated with pregnancy and childbirth. Importantly, this result suggests that women in the high-earner and all-rounder categories are more likely to be career-oriented and to have a greater propensity to forgo childbirth.

At age 42, we observe that men are substantially more likely to belong to the deprived or relative-poverty trajectories if they have been questioned by the police or convicted of a criminal offence. The rate of police questioning is substantially higher among men than among women, peaking at 38% for deprived men compared with only 10% for deprived women. There is also relatively little variation in conviction rates among women, although this likely reflects their lower overall rates of criminal offending.

We find relatively evidence that individual trajectories are compensated by differences in partner earnings. For men, we observe that partner income is higher for individuals in the deprived cluster than in the all-rounder or high earning clusters, but the difference is less than $55 per week. Among females, the pattern is reversed, suggesting that females in deprived households are more likely to live with partners on deprived trajectories.

By age 50, we continue to observe that the proportion of individuals consuming alcohol daily is negatively correlated with earnings and health, although the average level of alcohol consumption approximately doubles. In contrast, the proportion of deprived individuals suffering from chronic illnesses, disabilities and health problems is substantially higher among both men and women. These results serve two purposes: they validate our measure of physical health and quantify the nature of the health problems faced by individuals following trajectories associated with poverty. Finally, we observe that deprived individuals are less likely to own a home at age 50, less likely to be married and more likely to be divorced.

Taken together, the substantial differences in outcomes across trajectories confirm the effectiveness of our clustering techniques in isolating distinct patterns of socioeconomic outcomes. The direction of these differences is consistent with our trajectory labels across

a broad range of outcomes and at all points in the life cycle, providing a strong foundation for our core analysis of the role of early-life skills and environments in shaping these trajectories.

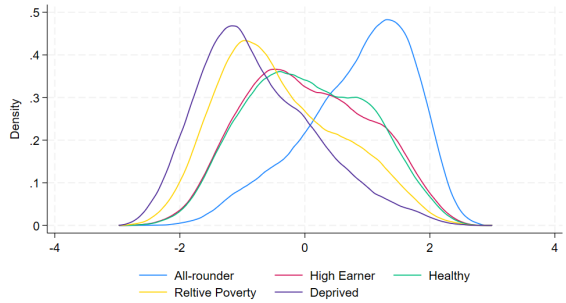
5 Relating Skills and Environments to Trajectories

The core of our analysis involves relating early-life skills to lifetime trajectories. For skills to be informative about trajectory membership, their distributions must vary across clusters. Figure 5 addresses this question by plotting the distributions of cognitive, externalising and internalising skills at age 11 across clusters and sexes. Our analysis sample is restricted to individuals for whom age-11 skills are observed, so the densities plotted in Figure 5 are based entirely on observed values and do not rely on imputation.

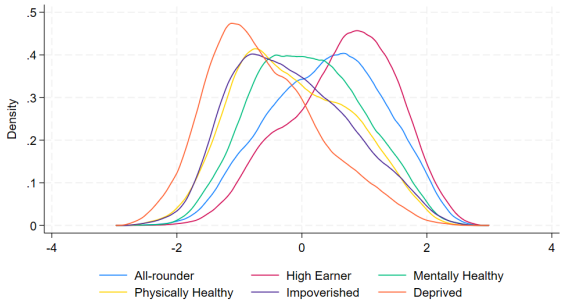
With respect to cognitive skills, all-rounders and high earners have the highest mean levels, while the deprived groups have the lowest endowments. These results are intuitive given the strong link between the mathematical, reading and writing abilities that underpin our cognitive-skill measure and the tasks performed in the labour market. Among the health-oriented trajectories, individuals with higher mental-health ranks also tend to have higher cognitive-skill endowments. Individuals in relative poverty have a slightly higher average level of cognition than those in deprivation, but remain below the healthy trajectories.

The same broad ordering across trajectories holds for internalising and externalising skills. However, there are notable differences in the distributions of these skills across sexes. Specifically, the distributions of externalising and internalising skills appear to be bi-modal for males but unimodal for females. Their variances are also larger among males, suggesting greater heterogeneity in non-cognitive skills among men. Despite these differences in distributional shape, mean skill levels are approximately equal across sexes. This suggests that the male distributions contain two distinct groups: those with externalising or internalising skills above the female mean and those with skills below it.

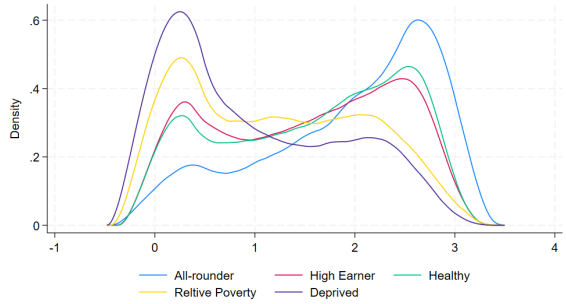
To relate these skills to lifetime trajectories, we employ a multinomial logit framework.



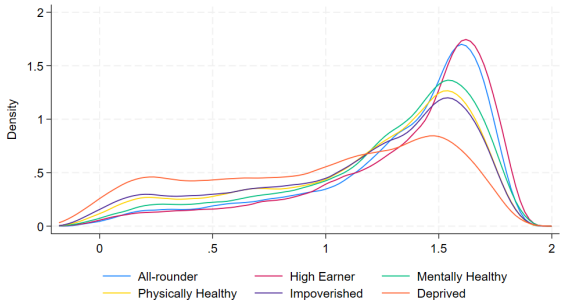
(a) Male: Cognitive



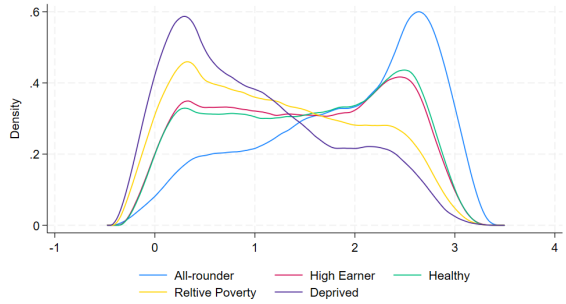
(b) Female: Cognitive



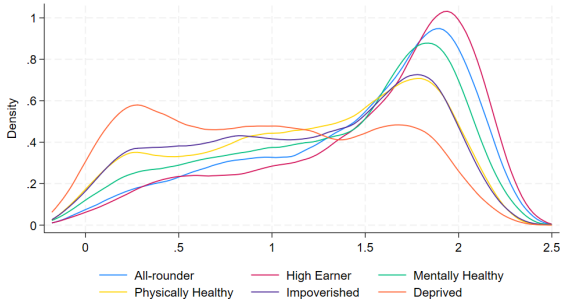
(c) Male: Externalising



(d) Female: Externalising



(e) Male: Internalising



(f) Female: Internalising

Figure 5: Distribution of Age 11 Skills Across Types

Notes: Results show distributions of age 11 skills by cluster for males (left) and females (right). The densities are defined treating the multiply imputed dataset as one object. Internalising and externalising skills are exponentiated.

For ease of interpretation, we convert the estimated coefficients into marginal effects and scale them by 100, such that each reported coefficient represents the percentage-point change in the probability of selecting into a particular trajectory associated with a one-standard-deviation increase in the relevant variable. We also reintroduce our measures of early-life environments to examine how skills and environments jointly relate to lifetime trajectories.

Table 7 reports the marginal effects.[12] Cognitive, internalising and externalising skills are all positively associated with trajectories characterised by higher earnings and health. For males, cognitive skills are particularly strongly associated with the high-earner and all-rounder trajectories. A one-standard-deviation increase in cognitive skills reduces the probability of selecting into the deprived cluster by 2.83 percentage points. Similarly, a one-standard-deviation increase in internalising skills raises the probability that males select into the all-rounder trajectory by 1.6 percentage points and reduces the probability of deprivation by 0.8 percentage points; a similar pattern holds for females. Externalising skills are particularly negatively associated with relative poverty and deprivation across both sexes. Taken together, these results imply that, for both males and females, higher skill levels increase the probability of following trajectories characterised by better health and earnings outcomes. Our findings therefore emphasise the importance of early-life skill formation and suggest that skills are associated with socioeconomic outcomes across multiple dimensions throughout adulthood.

The associations between early-life environments and trajectories are more heterogeneous across sexes. Childhood hospitalisation is a particularly strong predictor of trajectory membership: hospitalisation at a young age significantly increases the probability of selecting into the deprived cluster and decreases the probability of selecting into the physically healthy cluster. Likewise, social problems in the household are associated with a higher probability that an individual's lifetime trajectory falls into relative poverty or deprivation. All-rounders are particularly likely to come from environments in which par-

[12]To show that each variable is distinct and captures a unique aspect of childhood development Table A19 presents a correlogram of all skills and environments included in the multinomial logits stratified by sex.

Table 7: Marginal Effects (x100) of Skills and Environments on Lifetime Trajectories

Panel A: Male					
	(1) All-rounder	(2) High Earner	(3) Healthy	(4) Relative Poverty	(5) Deprived
Cognitive skills	6.14***	7.83***	-6.51***	-4.63***	-2.83***
	(0.470)	(0.732)	(0.827)	(0.595)	(0.415)
Internalising skills	1.60**	1.34	-0.144	-1.94***	-0.852**
	(0.644)	(1.000)	(1.040)	(0.523)	(0.355)
Externalising skills	-0.314	3.10*	3.05*	-4.47***	-1.37**
	(1.080)	(1.610)	(1.670)	(0.800)	(0.533)
Hospitalization	-0.331	0.803	-2.61***	1.56***	0.582**
	(0.332)	(0.617)	(0.720)	(0.381)	(0.248)
Social problems	0.0377	-1.37**	-0.466	1.05***	0.747***
	(0.312)	(0.584)	(0.634)	(0.392)	(0.261)
Parental time	0.990***	0.882	-0.122	-1.23**	-0.516
	(0.364)	(0.682)	(0.742)	(0.512)	(0.355)
Family SES	0.398	1.84***	0.296	-1.33***	-1.21***
	(0.358)	(0.637)	(0.711)	(0.411)	(0.284)
Primary school quality	0.461	0.457	0.0278	-0.677*	-0.269
	(0.281)	(0.565)	(0.653)	(0.410)	(0.271)
Parental demographics	0.459	0.604	-0.734	-0.0648	-0.263
	(0.339)	(0.649)	(0.705)	(0.436)	(0.315)
Observations	7199	7199	7199	7199	7199

Panel B: Female						
	(1) All-rounder	(2) High Earner	(3) Physically Healthy	(4) Mentally Healthy	(5) Relative Poverty	(6) Deprived
Cognitive skills	3.68***	2.64***	0.959*	-5.17***	2.64***	-4.74***
	(0.504)	(0.833)	(0.535)	(0.706)	(0.775)	(0.579)
Internalising skills	2.19**	-1.99	1.77*	-2.84***	1.34	-0.457
	(1.08)	(1.29)	(1.04)	(0.940)	(1.27)	(0.614)
Externalising skills	4.32**	-0.0495	2.31	-0.151	-3.07	-3.36***
	(2.03)	(2.41)	(1.89)	(1.70)	(2.29)	(1.02)
Hospitalization	-1.04**	0.199	-1.05**	0.505	0.635	0.754**
	(0.442)	(0.588)	(0.454)	(0.448)	(0.528)	(0.304)
Social problems	-0.500	-0.226	-1.87***	0.656	0.381	1.56***
	(0.420)	(0.621)	(0.468)	(0.495)	(0.582)	(0.349)
Parental time	0.882**	-0.255	0.759	-0.376	-0.438	-0.573
	(0.434)	(0.812)	(0.495)	(0.649)	(0.690)	(0.479)
Family SES	0.545	0.0256	0.187	-0.839	0.678	-0.597
	(0.448)	(0.717)	(0.463)	(0.533)	(0.657)	(0.404)
Primary school quality	-0.407	0.777	-0.641	-0.232	0.142	0.361
	(0.386)	(0.591)	(0.420)	(0.514)	(0.575)	(0.378)
Parental demographics	0.915**	0.656	-0.177	-0.0680	-0.114	-1.21***
	(0.399)	(0.672)	(0.447)	(0.536)	(0.628)	(0.431)
Observations	6809	6809	6809	6809	6809	6809

Coefficients derived from multivariate regressions multinomial logit models estimated separately for males and females. Skills and environments are measured at age 11. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

ents invest more time in their children. However, there is limited evidence that parental investment significantly reduces the probability of following other lifetime trajectories. The associations of family socioeconomic status, age-11 school quality and parental demographics also vary across sexes. Family socioeconomic status is a particularly strong predictor of deprivation among men but has no significant association with female trajectories, while parental demographics are significantly associated only with female trajectories. We find little evidence that primary-school quality is a strong predictor of lifetime trajectories.

There is a substantial policy debate surrounding the timing of early-childhood human-capital investment (Cunha, Heckman, and Schennach 2010). We address this issue by examining the marginal effects of cognitive, externalising and internalising skills measured at ages 7, 11 and 16. This analysis uses the same multinomial logit framework and is again conditional on our set of early-childhood environments. Table 8 reports the results.

For males, cognitive skills consistently play an important role in predicting later-life trajectories. At age 7, higher cognitive skills reduce the probability of following trajectories associated with relative poverty. From age 11 onwards, cognitive skills are strongly associated with selection into trajectories characterised by high earnings and health. Internalising skills play a relatively limited role in predicting lifetime trajectories until age 16, when they become a particularly strong predictor of trajectory membership.

As noted in Section 2, our measure of internalising skills changes from the BSAG to the Rutter Scale between ages 11 and 16, so the measures are not directly comparable across these ages. Nevertheless, internalising skills as measured by the Rutter Scale are strongly associated with trajectory membership and are positively related to both earnings and health. Externalising skills display weaker associations with trajectories over the life cycle, marginally reducing the probability of relative poverty at ages 7 and 11. Overall, the results suggest that skills measured at age 16 are the strongest predictors of lifetime trajectories.

Cognitive skills appear to be substantially weaker predictors of lifetime trajectories for

Table 8: Marginal Effects (x100) of Childhood Skills on Lifetime Trajectories

Panel A: Male					
	(1) All-rounder	(2) High Earner	(3) Healthy	(4) Relative Poverty	(5) Deprived
Cognitive Skills					
Age 7	-1.38	2.11	2.76	-1.94*	-1.56**
	(0.91)	(1.67)	(1.78)	(1.10)	(0.78)
Age 11	3.01***	1.70	-3.82**	0.18	-1.07
	(0.76)	(1.40)	(1.60)	(1.07)	(0.71)
Age 16	4.15***	5.49***	-5.50***	-3.97***	-0.17
	(0.86)	(1.64)	(1.87)	(1.27)	(0.81)
Internalising Skills					
Age 7	1.21*	0.56	-0.38	-0.71	-0.67
	(0.71)	(1.19)	(1.28)	(0.67)	(0.44)
Age 11	1.25*	0.83	-0.07	-1.60***	-0.41
	(0.66)	(1.05)	(1.10)	(0.56)	(0.37)
Age 16	10.00***	11.00**	-6.40	-6.62**	-7.99***
	(3.35)	(5.38)	(5.40)	(2.92)	(1.81)
Externalising Skills					
Age 7	1.60	1.67	-1.28	-1.80*	-0.20
	(1.22)	(1.87)	(1.95)	(1.05)	(0.66)
Age 11	-1.60	1.60	4.34**	-3.32***	-1.01
	(1.19)	(1.78)	(1.86)	(0.93)	(0.62)
Age 16	-3.27	-3.66	2.87	1.80	2.25
	(2.74)	(4.46)	(4.51)	(2.37)	(1.65)
Observations	7199	7199	7199	7199	7199

Panel B: Female						
	(1) All-rounder	(2) High Earner	(3) Physically Healthy	(4) Mentally Healthy	(5) Relative Poverty	(6) Deprived
Cognitive Skills						
Age 7	-0.15	-2.03	0.61	-0.43	3.56**	-1.56
	(1.17)	(1.88)	(1.29)	(1.42)	(1.69)	(1.05)
Age 11	1.84*	2.04	-1.00	-3.72***	0.39	0.44
	(0.98)	(1.58)	(1.08)	(1.32)	(1.50)	(1.04)
Age 16	1.27	3.81*	1.53	-1.90	0.90	-5.62***
	(1.34)	(2.09)	(1.51)	(1.75)	(1.97)	(1.29)
Internalising Skills						
Age 7	1.05	-2.18	1.12	0.08	0.42	-0.49
	(1.24)	(1.59)	(1.23)	(1.22)	(1.57)	(0.80)
Age 11	1.45	-1.16	1.28	-2.53**	0.81	0.15
	(1.12)	(1.36)	(1.09)	(0.99)	(1.34)	(0.64)
Age 16	16.30***	-8.22	5.29	-5.99	0.38	-7.75**
	(5.03)	(5.99)	(4.78)	(4.51)	(5.67)	(3.03)
Externalising Skills						
Age 7	2.38	-3.25	1.64	2.82	-4.99*	1.40
	(2.34)	(2.87)	(2.20)	(2.30)	(2.73)	(1.50)
Age 11	1.91	1.69	0.81	-0.01	-1.88	-2.52**
	(2.19)	(2.70)	(2.06)	(1.96)	(2.58)	(1.19)
Age 16	-0.78	-2.17	1.20	-0.28	0.57	1.46
	(6.19)	(6.94)	(5.59)	(5.17)	(6.46)	(3.38)
Observations	6809	6809	6809	6809	6809	6809

Notes: Coefficients derived from multivariate regressions multinomial logit models estimated separately for males and females. Regressions also condition on the set of environments included in Table 7. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

females than for males, reducing the probability of deprivation at age 16 but slightly increasing the probability of relative poverty at age 7. The comparatively small association between cognitive skills and female trajectories is not entirely surprising given that relatively few women in our cohort acquired post-secondary education and that women's career opportunities were more limited than men's. As with males, internalising skills are a particularly strong predictor of lifetime trajectories at age 16, increasing the probability of being an all-rounder by 16.3 percentage points while significantly decreasing the probability of being in the deprived cluster. Externalising skills appear to have only a small effect, reducing the probability of deprivation at ages 7 and 11, with little evidence of an association at age 16.

6 Predicting Trajectories

Our data contain information on a rich set of predictors that are strongly associated with lifetime trajectories. Our next goal is to assess how well skills and environments can predict trajectory membership. In a standard regression context, we might evaluate predictive performance using R^2. However, R^2 is not directly interpretable in a multinomial logit framework. Instead, we use a 60:20:20 training, validation and test split, as described in Section 3.4. Models are fitted on the training sample, with the validation sample used for model selection and the test sample reserved for final out-of-sample evaluation. For the neural network, we fit one model per imputed dataset and classify each individual using the modal prediction across models. Accuracy is defined as the proportion of observations that are correctly classified.

From a policy and redistributive perspective, the most important subpopulation to identify is individuals living in deprivation. If our model can accurately predict lifetime deprivation using a subset of skills, environments and demographic characteristics at relatively low computational cost, policymakers may be able to identify individuals at greater risk of deprivation and target interventions accordingly. We therefore also report the deprivation recall of our model, defined as the proportion of individuals in the deprived trajectory

who are correctly identified by the model.

Since a central objective of this exercise is to predict deprivation, we complement the multinomial logit model with a neural network containing two hidden layers. Neural networks are flexible non-linear models that can capture complex relationships between predictors and outcomes. Our primary measure of performance for this model is deprivation recall.

Table 9 reports our results. We present overall accuracy and deprivation recall across four specifications. The first includes skills only; the second supplements these with environmental characteristics; the third adds age-23 demographic information, including occupation, proxied by social class, and employment status; and the final specification adds information that could be obtained from a standard labour-force survey. Today, many of these variables are available for a large share of individuals born and educated in the United Kingdom, for example through administrative tax records or the Longitudinal Education Outcomes dataset, and could therefore potentially be used in a policy context. We stress, however, that our prediction exercise relies on survey data from a relatively small sample of individuals born in 1958. It is therefore unreasonable to expect exceptionally high recall rates. Given the greater scope and quality of data available today, we view our results as a conservative benchmark for the predictive performance that might be achievable using contemporary data.

Column 1 of Table 9 reports the results from the specification that includes skills only. The multinomial logit performs poorly in predicting deprivation for both males and females, despite achieving an overall prediction accuracy of approximately 50% for males. This poor performance partly reflects the imbalance in trajectory sizes: deprivation accounts for at most 9% of individuals, whereas, in the male sample, individuals following the healthy trajectory account for 47% of observations. The model therefore disproportionately predicts individuals to belong to the healthy trajectory. We address this imbalance in the neural network by reweighting the clusters so that they receive equal weight in the training sample. This substantially increases deprivation recall to 59% for males and 52%

for females. These results provide initial evidence that skills have meaningful predictive power for lifetime trajectories.

Table 9: Deprivation Prediction Accuracy and Recall

	(1) Skills	(2) Environment	(3) Age 23 Demographics	(4) Age 23 Labor Force Survey Variables
Male				
Multinomial Logit				
Deprivation Recall	0.01	0.01	0.03	0.09
Overall Accuracy	0.50	0.50	0.50	0.51
Neural Network				
Deprivation Recall	0.59	0.62	0.69	0.70
Overall Accuracy	0.36	0.39	0.42	0.53
Female				
Multinomial Logit				
Deprivation Recall	0.10	0.13	0.20	0.23
Overall Accuracy	0.28	0.29	0.30	0.34
Neural Network				
Deprivation Recall	0.52	0.59	0.60	0.65
Overall Accuracy	0.29	0.33	0.34	0.44

Models are trained using a 60% training, 20% validation and 20% held-out test split. For each of the 30 multiply imputed datasets, we train a separate neural network with two hidden layers containing 64 and 16 nodes, respectively. The training sample is reweighted so that each trajectory receives equal total weight, despite differences in relative sample size. Predictions are combined across imputations using the modal predicted trajectory (hard-voting ensemble). Recall is reported for the respective 'deprived' trajectory, while overall accuracy measures the proportion of all correctly classified observations in the held-out test sample.

Adding environmental variables increases deprivation recall by 3 percentage points for males and 7 percentage points for females. This suggests that environmental characteristics provide greater incremental predictive power for female than for male trajectories. Likewise, adding age-23 demographic characteristics increases recall by a further 7 percentage points for males, compared with only 1 percentage point for females. One possible explanation is the lower labour-force participation of women at age 23, which reduces variation in occupation-based social class relative to men and therefore limits its predictive content. Our final specification also includes age-23 labour-force information, namely wages. Under this specification, deprivation recall reaches 70% for males and 65% for

females. Given the limited range of covariates, the age of the data and the relatively small sample size, we view these recall rates as promising evidence that relatively simple prediction models may help identify individuals at risk of persistent deprivation.

7 Conclusion

In this paper, we use machine-learning tools to describe patterns of heterogeneity in the joint trajectories of earnings, employment, physical health and mental health among a cohort of men and women born in Great Britain in 1958. Our methodology sheds light on the nuanced relationships between these different aspects of socioeconomic life in a manner that is not easily replicated using more standard regression frameworks.

We estimate separate models for men and women and show that they follow distinctly different trajectories. Common to both is the presence of substantial inequality across all dimensions and a particularly deprived group that consistently experiences the worst labour-market and health outcomes throughout the life cycle. These individuals are likely to spend much of their lives out of the labour force and are disproportionately likely to suffer from chronic illnesses and disabilities and to have poor housing outcomes. From a policy perspective, this is an important group to identify when designing targeted interventions. In contrast, there is a similarly sized group of men and women who excel in both labour-market and health outcomes throughout their lives. The coexistence of these groups highlights the substantial degree of multidimensional inequality across trajectories. Among the remaining individuals, we document a potential trade-off between labour-market success and health outcomes. Differences across trajectories emerge early in life and remain deeply entrenched throughout the life cycle.

In a second step, we relate socioeconomic trajectories to early-life cognitive and non-cognitive skills and environments measured throughout childhood. First, we find that trajectories are strongly associated with skills, with these relationships being most pronounced for skills measured at age 16 rather than earlier in childhood. For males, these skills are important predictors of earnings and health across the distribution, whereas for

females they are particularly important predictors of avoiding extreme deprivation and relative poverty. Second, we find that early-life environments are important predictors of lifetime trajectories, but their associations differ by sex. Males are particularly responsive to family socioeconomic status, while females are more responsive to family demographic characteristics. For both males and females, parental time, social problems within the household and childhood hospitalisation are associated with trajectory membership. We find limited evidence that school quality predicts lifetime trajectories after conditioning on the other factors included in our analysis.

Finally, we conduct a prediction exercise to evaluate the predictive power of early-life skills and environments for lifetime trajectories. We show that a simple neural network can identify cases of deprivation among males with a recall rate of 70%. Given the age of our data and the relatively small sample size, we view this as a promising result. Contemporary policymakers have access to substantially richer and more comprehensive administrative data, suggesting that relatively simple prediction models may provide a useful starting point for identifying individuals at risk of persistent deprivation and informing targeted policymaking.

Taken together, the results of this analysis, although descriptive in nature, contribute to our understanding of the complex relationships between different dimensions of socioeconomic well-being, the importance of persistence and multidimensionality in considering deprivation and inequality, and the ways in which early-life skills and experiences are associated with later-life outcomes.

Our results highlight that the importance of multidimensionality is not limited to considerations of deprivation but also characterises the relationships between early-childhood factors and adult socioeconomic trajectories. This is consistent with findings elsewhere in the literature that emphasise the importance of both the type and timing of interventions designed to improve childhood outcomes.

References

Abbott, A. and J. Forrest (1986). "Optimal Matching Methods for Historical Sequences". In: *The Journal of Interdisciplinary History* 16.3, pp. 471–494. ISSN: 0022-1953. DOI: 10.2307/204500. URL: https://www.jstor.org/stable/204500 (visited on 02/24/2025).

Abbott, A. and A. Tsay (2000). "Sequence Analysis and Optimal Matching Methods in Sociology: Review and Prospect". In: *Sociological Methods and Research* 29.1, pp. 3–33.

Adda, J. and C. Dustmann (Feb. 2023). "Sources of Wage Growth". In: *Journal of Political Economy* 131.2, pp. 456–503. ISSN: 0022-3808. DOI: 10.1086/721657. URL: https://www.journals.uchicago.edu/doi/abs/10.1086/721657 (visited on 03/12/2025).

Alkire, S. and J. Foster (2011). "Counting and Multidimensional Poverty Measurement". In: *Journal of Public Economics* 95.7-8, pp. 476–487. DOI: 10.1016/j.jpubeco.2010.11.006.

Atkinson, A. B. and F. Bourguignon (1982). "The Comparison of Multi-Dimensioned Distributions of Economic Status". In: *Review of Economic Studies* 49.2, pp. 183–201.

Azur, M. J., E. A. Stuart, C. Frangakis, and P. J. Leaf (2011). "Multiple imputation by chained equations: what is it and how does it work?" en. In: *International Journal of Methods in Psychiatric Research* 20.1. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/mpr.3 pp. 40–49. ISSN: 1557-0657. DOI: 10.1002/mpr.329. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/mpr.329 (visited on 02/22/2024).

Behar, L. and S. Stringfield (1974). "A behavior rating scale for the preschool child". In: *Developmental Psychology* 10.5. Place: US, pp. 601–610. ISSN: 1939-0599. DOI: 10.1037/h0037058.

Blundell, R., L. Dearden, A. Goodman, and H. Reed (Feb. 2000). "The Returns to Higher Education in Britain: Evidence From a British Cohort". In: *The Economic Journal* 110.461, F82–F99. ISSN: 0013-0133. DOI: 10.1111/1468-0297.00508. URL: https://doi.org/10.1111/1468-0297.00508 (visited on 03/12/2025).

Blundell, R., L. Dearden, and B. Sianesi (July 2005). "Evaluating the Effect of Education on Earnings: Models, Methods and Results from the National Child Development Survey". en. In: *Journal of the Royal Statistical Society Series A: Statistics in Society* 168.3, pp. 473–512. ISSN: 0964-1998, 1467-985X. DOI: 10.1111/j.1467-985X.2004.00360.x. URL: https://academic.oup.com/jrsssa/article/168/3/473/7084510 (visited on 03/12/2025).

Bolt, U., E. French, J. Hentall-MacCuish, and C. O'Dea (2023). "Intergenerational altruism and transfers of time and money: a life cycle perspective". eng. In: *IFS Working Paper*. DOI: 10.1920/wp.ifs.2023.1123. URL: https://www.econstor.eu/handle/10419/284237 (visited on 02/24/2025).

Bonhomme, S., T. Lamadon, and E. Manresa (2022). "Discretizing Unobserved Heterogeneity". In: *Econometrica* 90.2, pp. 625–643.

Bonhomme, S. and J.-M. Robin (2009). "Assessing the Equalizing Force of Mobility Using Short Panels: France, 1990–2000". In: *Review of Economic Studies* 76.1, pp. 63–92. DOI: 10.1111/j.1467-937X.2008.00521.x.

Bourguignon, F. and S. R. Chakravarty (2003). "The Measurement of Multidimensional Poverty". In: *Journal of Economic Inequality* 1, pp. 25–49.

Brock, D. and M. Aitkin (1981). "Marginal Maximum Likelihood Estimation of Item Parameters: Application of an EM Algorithm". In: *Psychometrica* 46, pp 443 459. DOI: https://doi.org/10.1007/BF02293801. URL: https://www.cambridge.org/core/journals/psychometrika/article/abs/marginal-maximum-likelihood-

estimation-of-item-parameters-application-of-an-em-algorithm/ABCDBCF23ECBBC837C2F3F (visited on 03/14/2025).
Carneiro, P., C. Crawford, and A. Goodman (2007). "The impact of early cognitive and non-cognitive skills on later life outcomes". In.
Carneiro, P. M. and J. J. Heckman (July 2003). *Human Capital Policy*. en. SSRN Scholarly Paper. Rochester, NY. DOI: 10.2139/ssrn.434544. URL: https://papers.ssrn.com/abstract=434544 (visited on 03/16/2025).
Case, A. and A. Deaton (2015). "Rising Morbidity and Mortality in Midlife among White Non-Hispanic Americans in the 21st Century". In: *Proceedings of the National Academy of Sciences* 112.49, pp. 15078–15083. DOI: 10.1073/pnas.1518393112.
– (2017). "Mortality and Morbidity in the 21st Century". In: *Brookings Papers on Economic Activity* 2017.1, pp. 397–476.
Case, A. and L. Kraftman (2024). "Health inequalities". In: *Oxford Open Economics* 3.Supplement 1. Citation key retained for compatibility; publication year is 2024, pp. i499–i528. DOI: 10.1093/ooec/odad060.
Cengiz, D., A. Dube, A. Lindner, and D. Zentler-Munro (Apr. 2022). "Seeing beyond the Trees: Using Machine Learning to Estimate the Impact of Minimum Wages on Labor Market Outcomes". In: *Journal of Labor Economics* 40.S1, S203–S247. ISSN: 0734-306X. DOI: 10.1086/718497. URL: https://www.journals.uchicago.edu/doi/full/10.1086/718497 (visited on 03/27/2024).
Charrad, M., N. Ghazzali, V. Boiteau, and A. Niknafs (2014). "NbClust : An R Package for Determining the Relevant Number of Clusters in a Data Set". en. In: *Journal of Statistical Software* 61.6. ISSN: 1548-7660. DOI: 10.18637/jss.v061.i06. URL: http://www.jstatsoft.org/v61/i06/ (visited on 03/14/2025).
Cunha, F. and J. J. Heckman (Oct. 2008). "Formulating, Identifying and Estimating the Technology of Cognitive and Noncognitive Skill Formation". en. In: *Journal of Human Resources* 43.4, pp. 738–782. ISSN: 0022-166X, 1548-8004. DOI: 10.3368/jhr.43.4.738. URL: https://jhr.uwpress.org/content/43/4/738 (visited on 02/22/2024).
Cunha, F., J. J. Heckman, and S. M. Schennach (2010). "Estimating the Technology of Cognitive and Noncognitive Skill Formation". en. In: *Econometrica* 78.3. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA6551, pp. 883–931. ISSN: 1468-0262. DOI: 10.3982/ECTA6551. URL: https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA6551 (visited on 02/22/2024).
De Nardi, M., S. Pashchenko, and P. Porapakkarm (2025). "The Lifetime Costs of Bad Health". In: *Review of Economic Studies* 92.3, pp. 1987–2026. DOI: 10.1093/restud/rdae080.
Deaton, A. (2013). "The Great Escape : Health, Wealth, and the Origins of Inequality". en. In: *Princeton University Press*, pp. 1–384. URL: https://www.torrossa.com/en/resources/an/5793096 (visited on 03/16/2025).
Deaton, A. S. and C. Paxson (Jan. 1998). "Health, Income, and Inequality over the Life Cycle". In: *Frontiers in the Economics of Aging*. University of Chicago Press, pp. 431–462. URL: https://www.nber.org/books-and-chapters/frontiers-economics-aging/health-income-and-inequality-over-life-cycle (visited on 03/15/2025).
Decancq, K. and M. A. Lugo (2012). "Inequality of Wellbeing: A Multidimensional Approach". en. In: *Economica* 79.316. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1468-0335.2012.00929.x, pp. 721–746. ISSN: 1468-0335. DOI: 10.1111/j.1468-0335.2012.00929.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1468-0335.2012.00929.x (visited on 02/25/2024).

Deming, D. J. (Nov. 2017). "The Growing Importance of Social Skills in the Labor Market*". In: *The Quarterly Journal of Economics* 132.4, pp. 1593–1640. ISSN: 0033-5533. DOI: 10.1093/qje/qjx022. URL: https://doi.org/10.1093/qje/qjx022 (visited on 03/12/2025).

Gensowski, M. (Apr. 2018). "Personality, IQ, and lifetime earnings". In: *Labour Economics* 51, pp. 170–183. ISSN: 0927-5371. DOI: 10.1016/j.labeco.2017.12.004. URL: https://www.sciencedirect.com/science/article/pii/S0927537117303287 (visited on 03/16/2025).

Goodman, A., R. Joyce, and J. P. Smith (Apr. 2011). "The long shadow cast by childhood physical and mental problems on adult life". EN. In: *Proceedings of the National Academy of Sciences* 108.15. Company: National Academy of Sciences Distributor: National Academy of Sciences Institution: National Academy of Sciences Label: National Academy of Sciences, pp. 6032–6037. DOI: 10.1073/pnas.1016970108. URL: https://www.pnas.org/doi/abs/10.1073/pnas.1016970108 (visited on 03/15/2025).

Harvey, R. J. and A. L. Hammer (May 1999). "Item Response Theory". In: *The Counseling Psychologist* 27.3, pp. 353–383. ISSN: 0011-0000. DOI: 10.1177/0011000099273004. URL: https://doi.org/10.1177/0011000099273004 (visited on 03/14/2025).

Heckman, J., R. Pinto, and P. Savelyev (Oct. 2013). "Understanding the Mechanisms through Which an Influential Early Childhood Program Boosted Adult Outcomes". en. In: *American Economic Review* 103.6, pp. 2052–2086. ISSN: 0002-8282. DOI: 10.1257/aer.103.6.2052. URL: https://www.aeaweb.org/articles?id=10.1257/aer.103.6.2052 (visited on 03/16/2025).

Heckman, J. J., J. E. Humphries, S. Urzua, and G. Veramendi (2010). "The effects of educational choices on labor market, health, and social outcomes". en. In: *Working Papers*. Number: 2011-002. URL: https://ideas.repec.org//p/hka/wpaper/2011-002.html (visited on 03/16/2025).

Heckman, J. J. and B. Singer (1984). "A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data". In: *Econometrica* 52.2, pp. 271–320.

Hosseini, R., K. Kopecky, and K. Zhao (2026). "How Important Is Health Inequality for Lifetime Earnings Inequality?" In: *Review of Economic Studies* 93.1, pp. 556–599. DOI: 10.1093/restud/rdaf030.

Huggett, M., G. Ventura, and A. Yaron (2011). "Sources of Lifetime Inequality". In: *American Economic Review* 101.7, pp. 2923–2954.

Jolivet, G. and F. Postel-Vinay (2025). "A Structural Analysis of Mental Health and Labour Market Trajectories". In: *Review of Economic Studies* 92.3, pp. 1920–1954. DOI: 10.1093/restud/rdae071.

Keane, M. P. and K. I. Wolpin (1997). "The Career Decisions of Young Men". In: *Journal of Political Economy* 105.3, pp. 473–522.

Maasoumi, E. (1986). "The Measurement and Decomposition of Multi-Dimensional Inequality". In: *Econometrica* 54.4, pp. 991–997.

Milligan, G. W. and M. C. Cooper (June 1985). "An examination of procedures for determining the number of clusters in a data set". en. In: *Psychometrika* 50.2, pp. 159–179. ISSN: 1860-0980. DOI: 10.1007/BF02294245. URL: https://doi.org/10.1007/BF02294245 (visited on 03/14/2025).

Papageorge, N. W., V. Ronda, and Y. Zheng (Feb. 2019). *The Economic Value of Breaking Bad: Misbehavior, Schooling and the Labor Market*. Working Paper. DOI: 10.3386/w25602. URL: https://www.nber.org/papers/w25602 (visited on 03/16/2025).

Rubin, D. (Dec. 1976). "Inference and missing data". In: *Biometrika* 63.3, pp. 581–592. ISSN: 0006-3444. DOI: 10.1093/biomet/63.3.581. URL: https://doi.org/10.1093/biomet/63.3.581 (visited on 02/29/2024).

Rubin, D. B. (1996). "Multiple Imputation After 18+ Years". In: *Journal of the American Statistical Association*.

Rutter, M., J. Tizard, and K. Whitmore (1981). *Education, health, and behaviour*. R.E. Krieger Pub. Co. URL: https://cir.nii.ac.jp/crid/1130000795322914304 (visited on 02/24/2025).

Savelyev, P. A. (Sept. 2022). "Conscientiousness, Extraversion, College Education, and Longevity of High-Ability Individuals". en. In: *Journal of Human Resources* 57.5, pp. 1526–1565. ISSN: 0022-166X, 1548-8004. DOI: 10.3368/jhr.58.1.0918-9720R2. URL: https://jhr.uwpress.org/content/57/5/1526 (visited on 03/15/2025).

Sen, A. (1985). *Commodities and Capabilities*. North-Holland.

Settersten, R. and K. Mayer (Aug. 1997). "The Measurement of Age, Age Structuring, and the Life Course". en. In: *Annual Review of Sociology* 23.Volume 23, 1997, pp. 233–261. ISSN: 0360-0572, 1545-2115. DOI: 10.1146/annurev.soc.23.1.233. URL: https://www.annualreviews.org/content/journals/10.1146/annurev.soc.23.1.233 (visited on 02/24/2025).

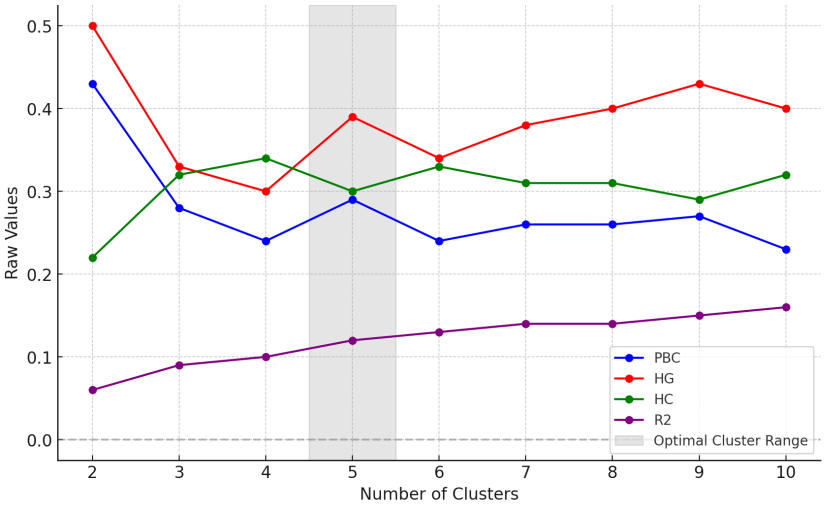
Silverwood, R., M. Narayanan, B. Dodgeon, M. Katsoulis, and G. Ploubidis (2024). "Handling missing data in the CLS cohort studies". In: *UCL Centre for Longitudinal Studies*.

Smith, J. (2009). "The Impact of Childhood Health on Adult Labor Market Outcomes". In: *The Review of Economics and Statistics* 91.3, pp. 478–489. DOI: https://doi.org/10.1162/rest.91.3.478. URL: https://direct.mit.edu/rest/article/91/3/478/57778/The-Impact-of-Childhood-Health-on-Adult-Labor (visited on 03/16/2025).

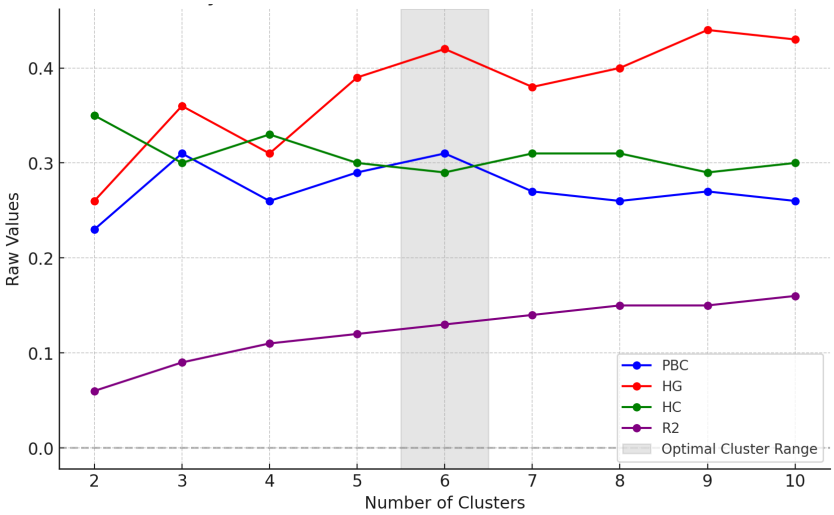
Studer, M. and G. Ritschard (2016). "What Matters in Differences Between Life Trajectories: A Comparative Review of Sequence Dissimilarity Measures". In: *Journal of the Royal Statistical Society*. URL: https://academic.oup.com/jrsssa/article/179/2/481/7068212 (visited on 02/24/2025).

A Additional Figures and Tables

Appendix Figures



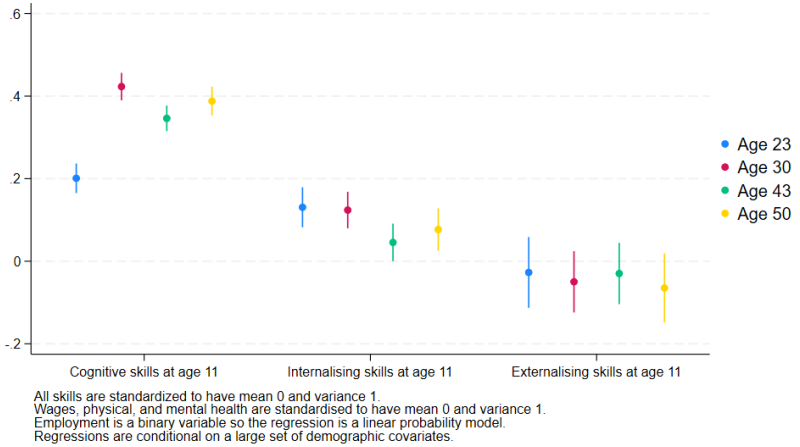
(a) Male



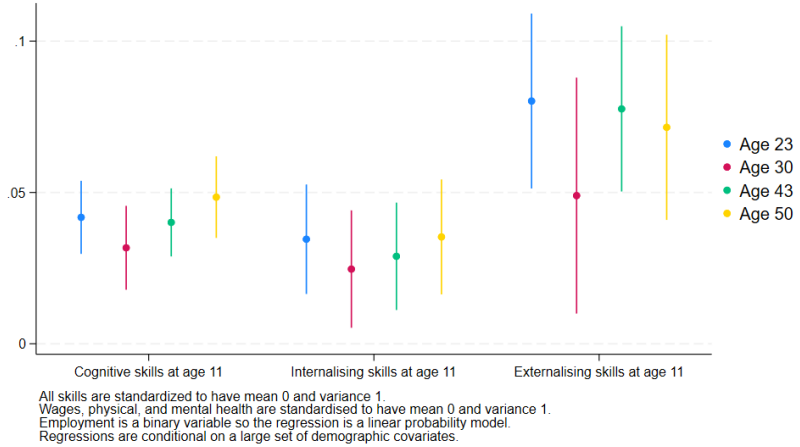
(b) Female

Figure A01: Optimal clusters

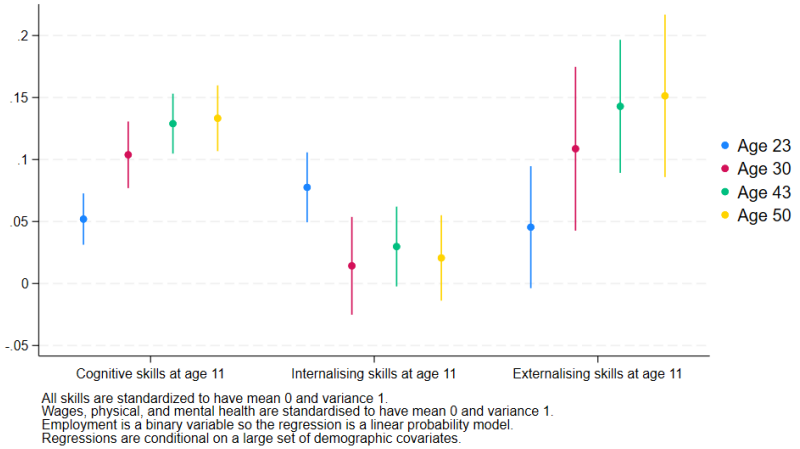
Notes: Cluster quality computed on a random subsample of observations equivalent in size to five imputed datasets.



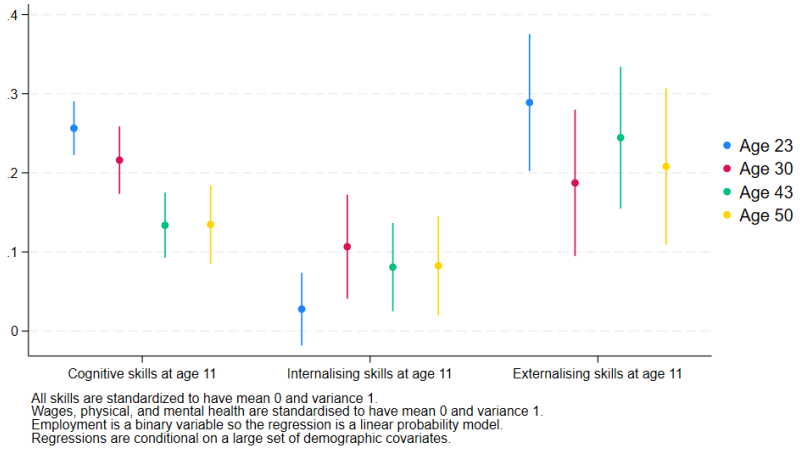
(a) Earnings



(b) Employment



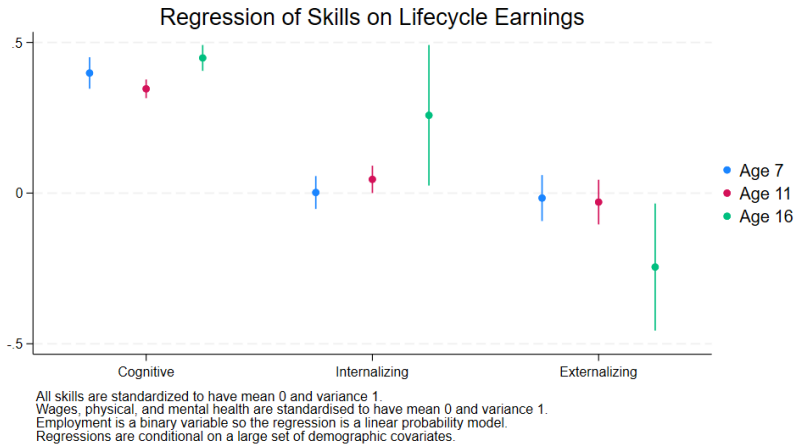
(c) Physical Health



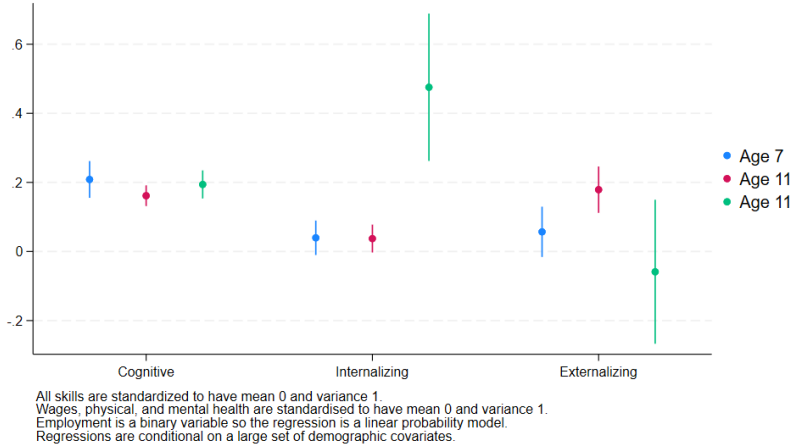
(d) Mental Health

Figure A02: Regressions of Outcomes Over the Lifecycle on Skills at Age 11

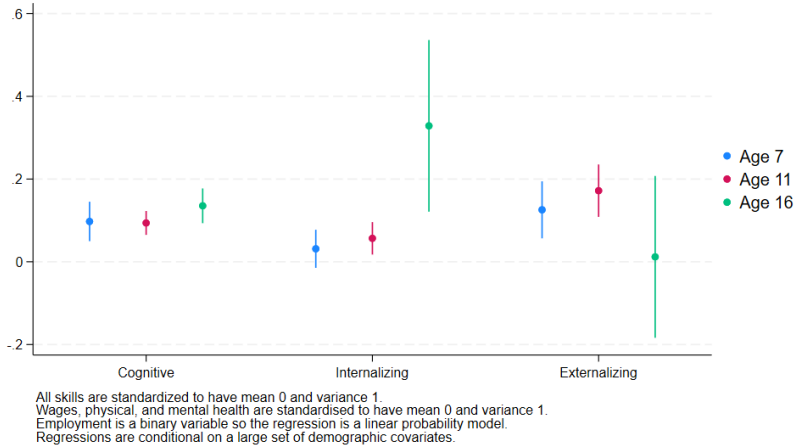
Notes: Each coefficient of interest represents the results from the regression of that outcome at that age on skills at age 11. All skills and environments are standardised to have mean 0 and variance 1. Wages, physical health and mental health are also standardised to have mean 0 and variance 1. Employment is a binary indicator and the regression is interpreted as a linear probability model. The regressions are also conditional on a large set of demographic characteristics.



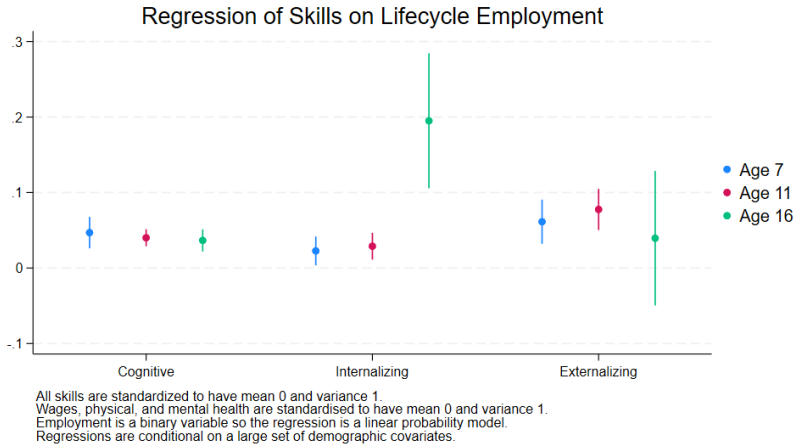
(a) Earnings



(b) Physical Health



(c) Mental Health



(d) Employment

Figure A03: Regressions of Outcomes at Age 42 on Skills Over Childhood

Notes: Each coefficient of interest represents the results from the regression of the outcome at age 42 on skills at the age of interest. All skills and environments are standardised to have mean 0 and variance 1. Wages, physical health and mental health are also standardised to have mean 0 and variance 1. Employment is a binary indicator and the regression is interpreted as a linear probability model. The regressions are also conditional on a large set of demographic characteristics.

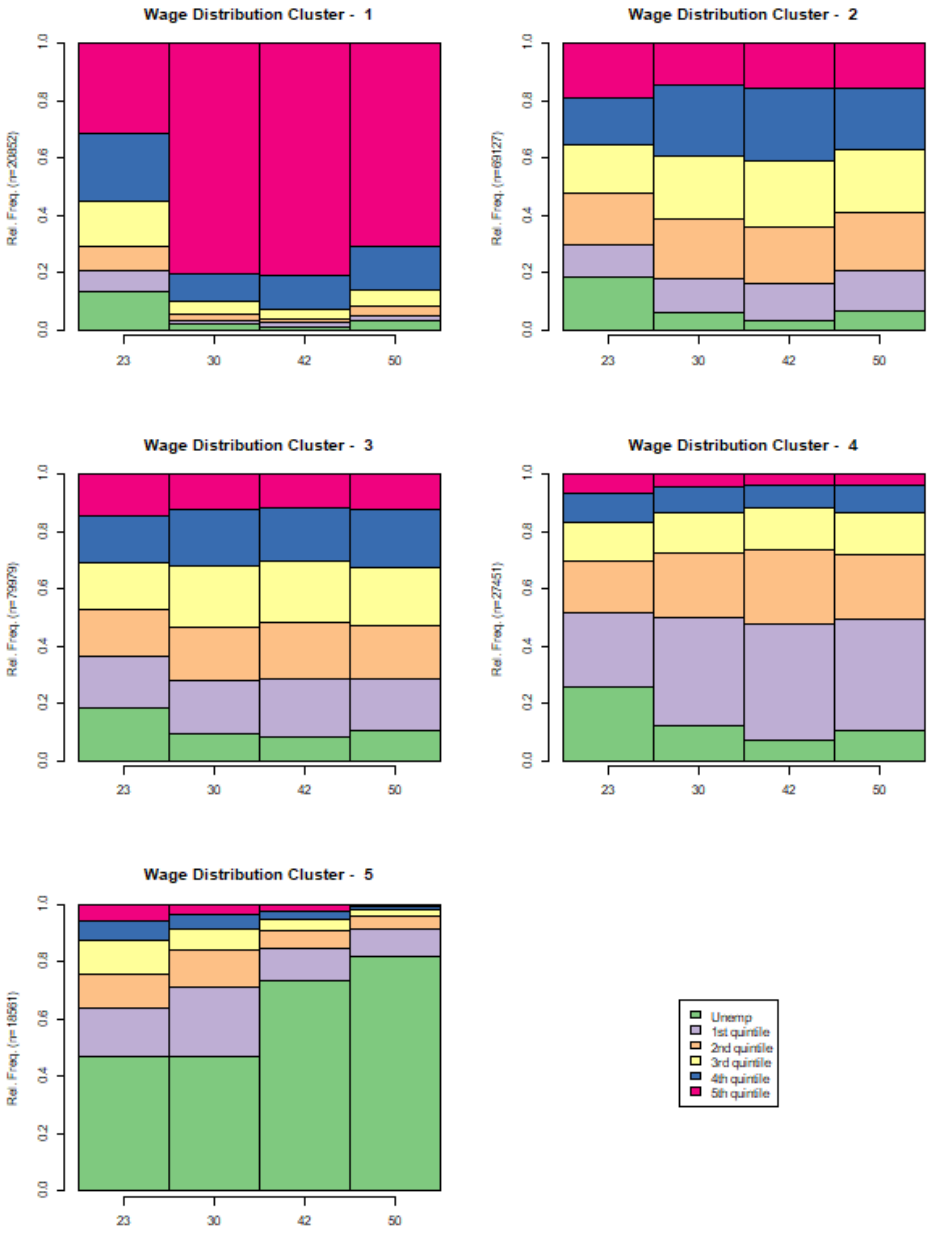


Figure A04: Male Wage State Distributions Over the Lifecycle

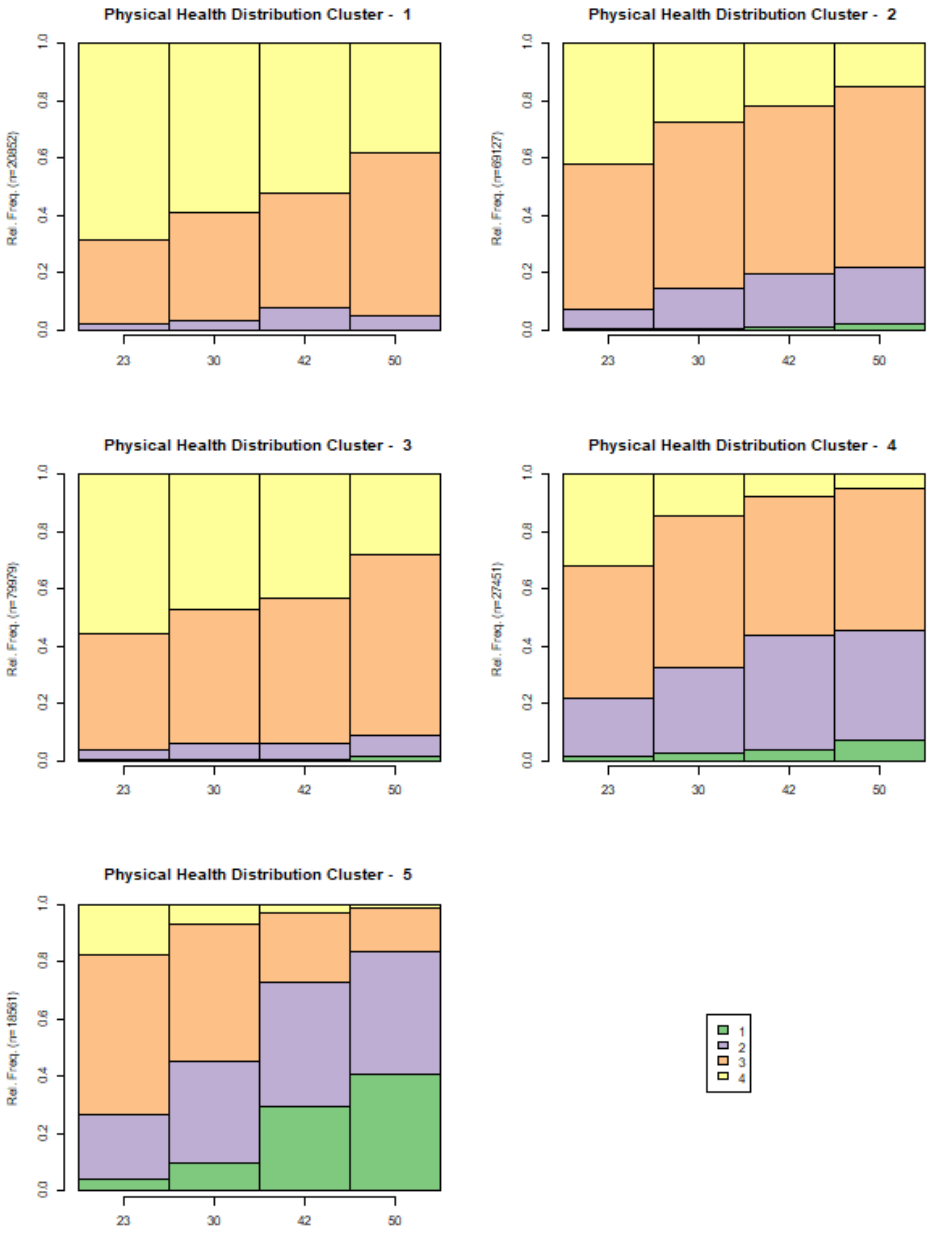


Figure A05: Male Physical Health State Distributions Over the Lifecycle

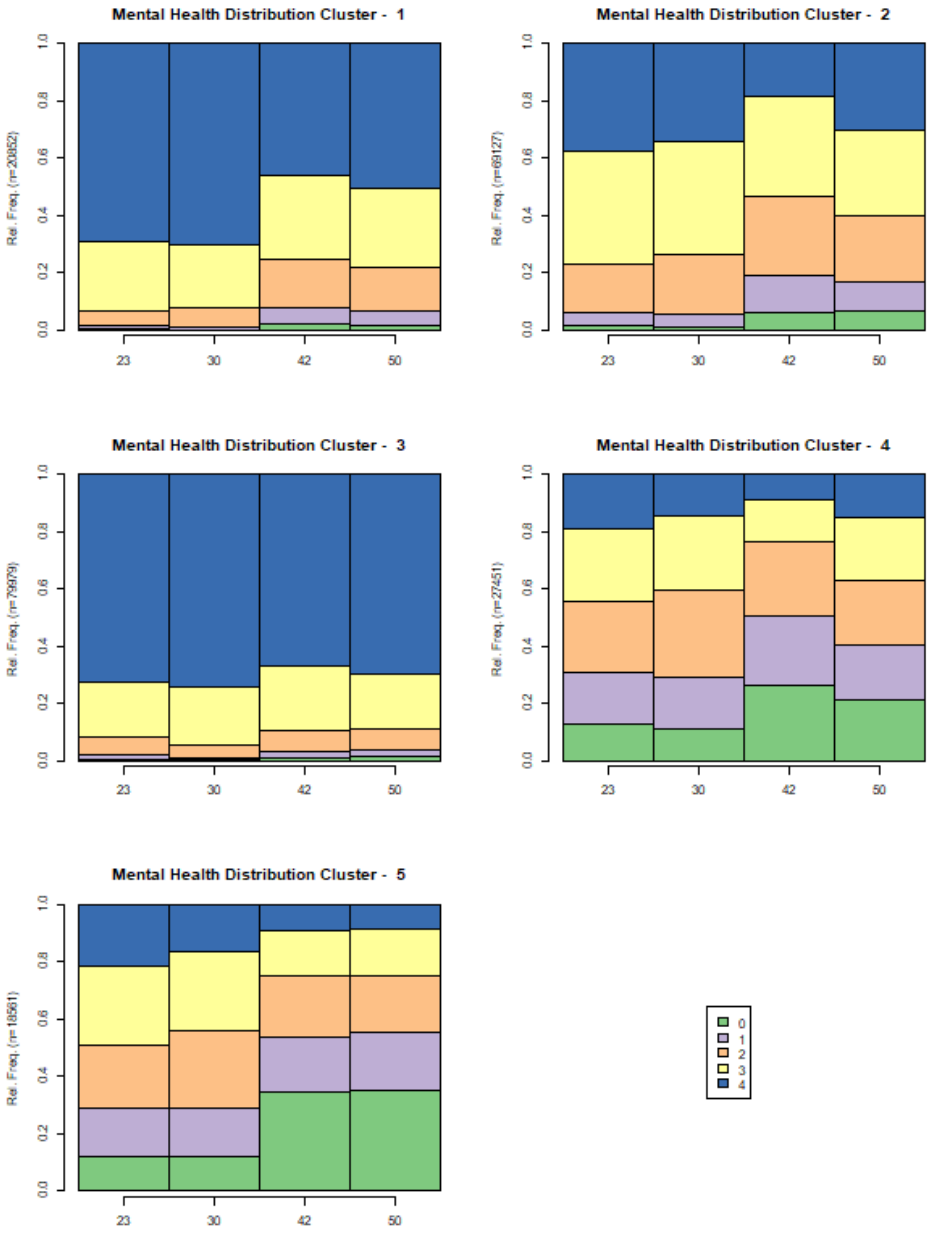


Figure A06: Male Mental Health State Distributions Over the Lifecycle

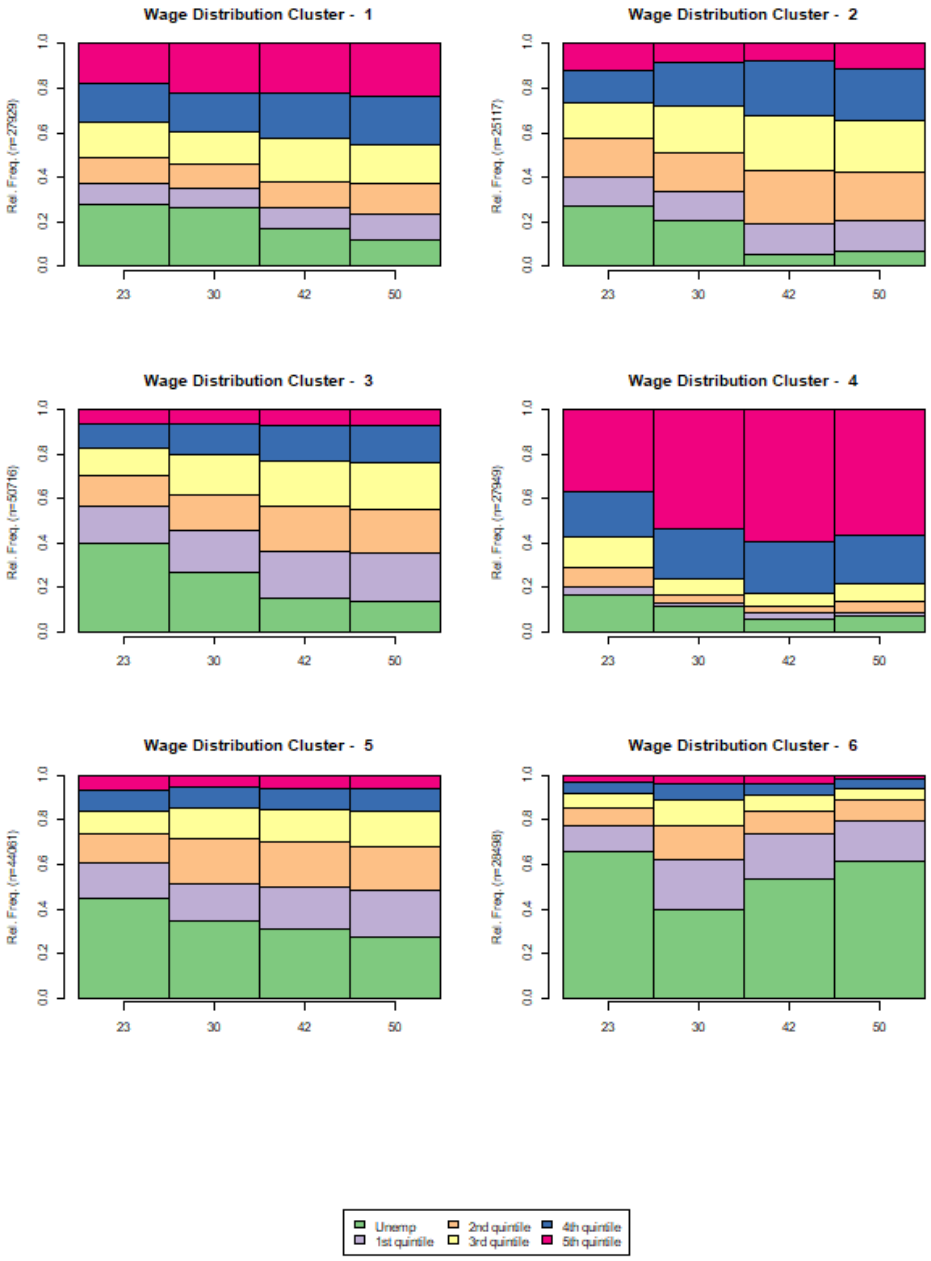


Figure A07: Female Wage State Distributions Over the Lifecycle

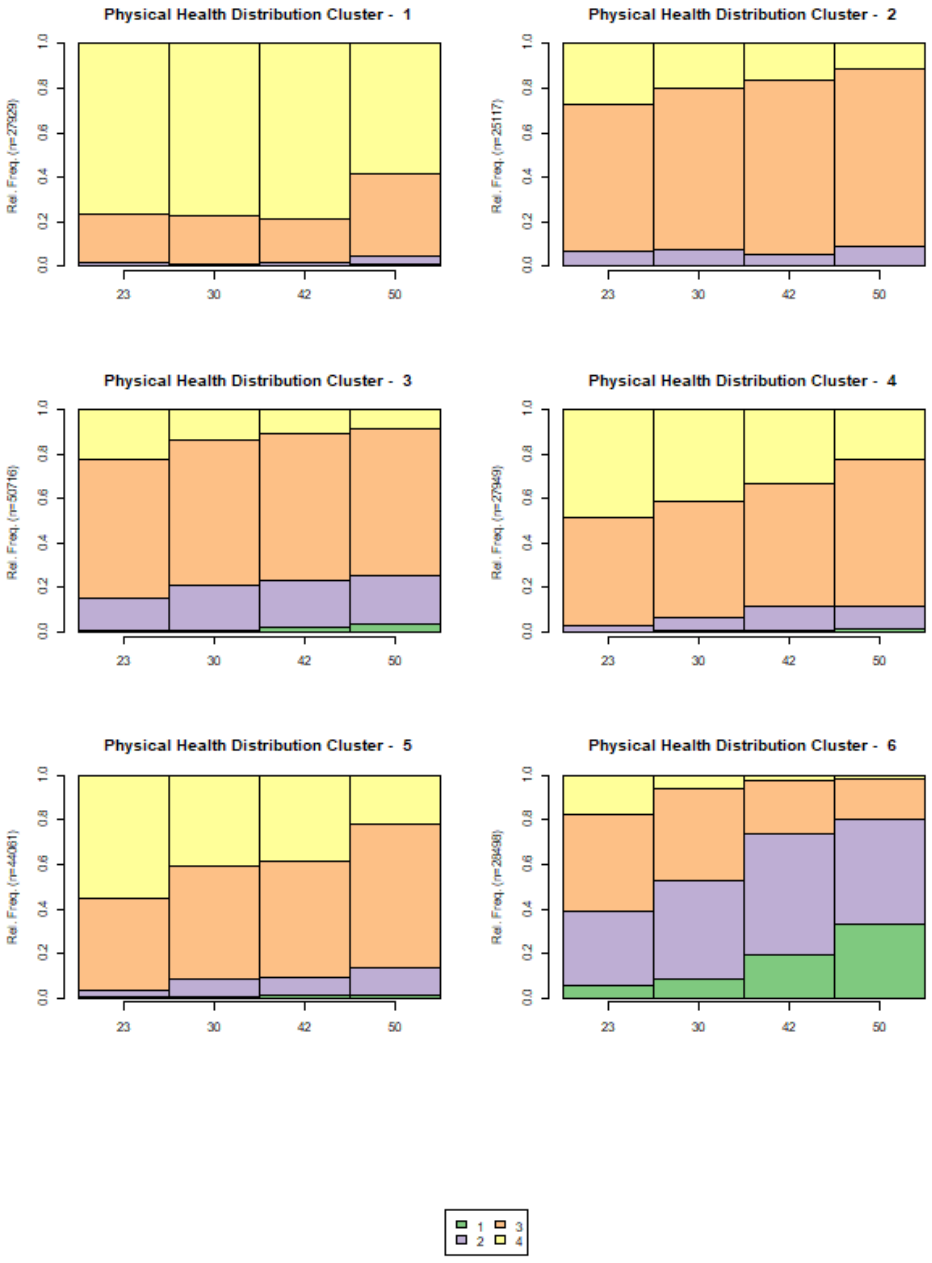


Figure A08: Female Physical Health State Distributions Over the Lifecycle

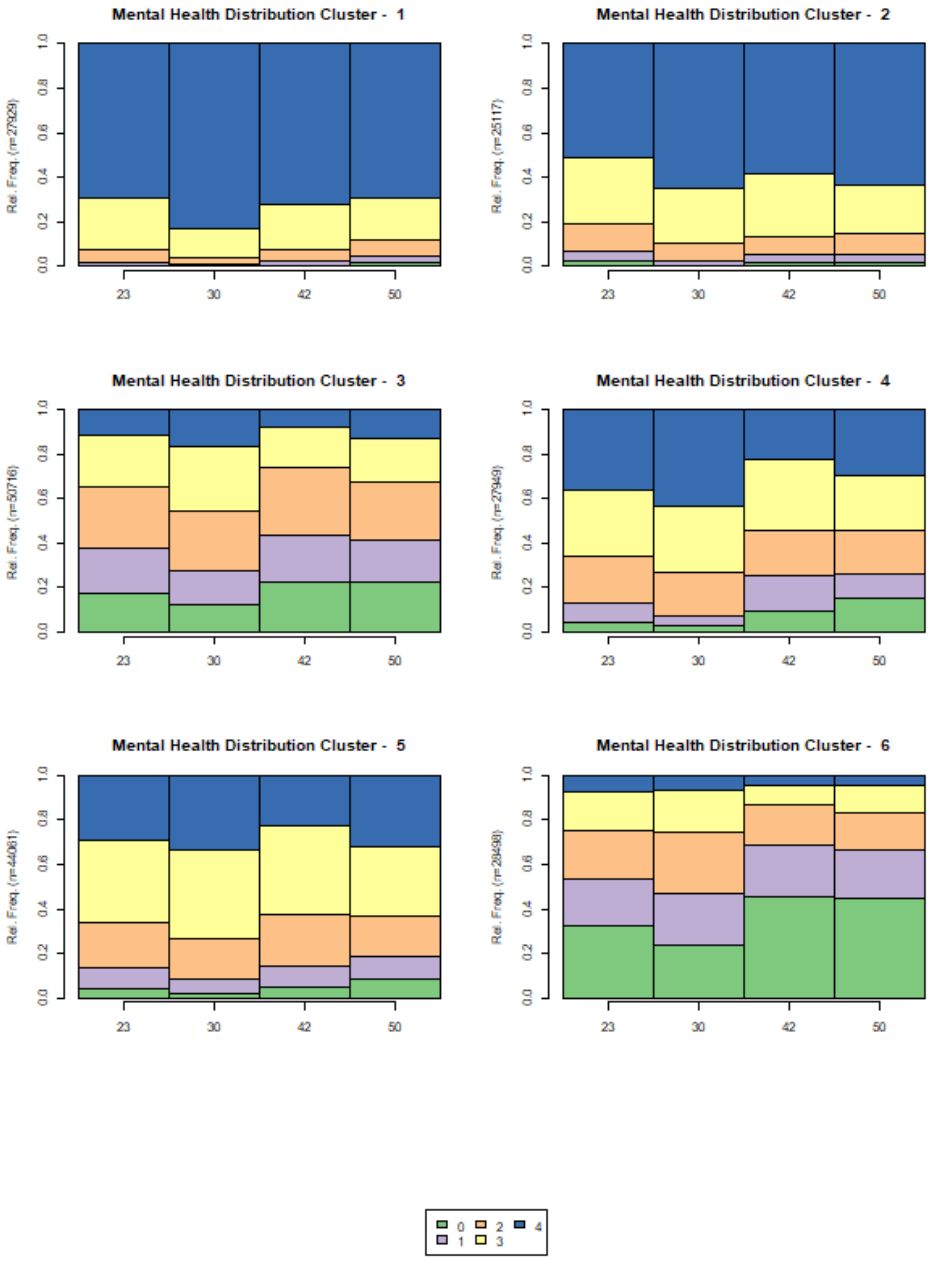


Figure A09: Female Mental Health State Distributions Over the Lifecycle

Appendix Tables

Table A01: List of childhood skills assessed at different ages.

Age	Factor	Items
7	Cognition	Maths score, Reading score, Design-copying score
	Internalising behaviour	BSAG items: Withdrawal, Unforthcomingness, Depression, Writing off adults and standards
	Externalising behaviour	BSAG items: Anxiety for acceptance by peers, Hostility towards peers, Hostility towards adults, Writing off adults and standards, Anxiety, Restlessness, and Impulsive acting out
11	Cognition	Maths score, Reading score, Design-copying score
	Internalising behaviour	BSAG items: Withdrawal, Unforthcomingness, Depression, Writing off adults and standards
	Externalising behaviour	BSAG items: Anxiety for acceptance by peers, Hostility towards peers, Hostility towards adults, Writing off adults and standards, Anxiety, Restlessness, and Impulsive acting out
16	Cognition	Maths score, Reading score
	Internalising behaviour	Rutter scale items: Past-year truancy, Not much liked by other children, Often worries about many things, Tends to be on their own/rather solitary, Often appears miserable/unhappy, Has twitches/mannerisms/tics (face or body), Frequently sucks thumb or finger, Frequently bites nails or fingers, School absences for trivial reasons, Fearful of new situations and things, Fussy or over-particular, Unresponsive/inert/apathetic, Often complains of aches/pains, Tears on arrival/refuses to enter school, Has a stutter or stammer
	Externalising behaviour	Rutter scale items: Restless/has difficulty staying seated, Squirmy/fidgety, Destroys/damages own or others' property, Frequently fights/very quarrelsome, Irritable/touchy/flies off the handle, Is often disobedient, Cannot settle for more than a few moments, Often tells lies, Has stolen at least once in the past year, Resentful/aggressive when corrected, Bullies other children

Notes: The BSAG is replaced by the Rutter scale at age 15.

Table A02: Childhood Environments

Factor	**Age**	**Items**
Family Demographics	–	Whether first-born child, Father's education, Mother's education, Mother's age at birth, Number of siblings at age 16
School Quality	7	Has PTA, Has education meetings for PTA, Has social events for parents, Ability streaming, School type (LEA/non-LEA), Class size, School size
	11	School size, School type (LEA junior, LEA junior-infants, other/non-LEA), Student-teacher ratio, Ability streaming
	16	School size, Ability streaming, School type (LEA comp, LEA non-comp, non-LEA), Student-teacher ratio, Has PTA
Family Socio-Economic Status	7	Dad's class, Had housing problems, Had financial problems, Had unemployment problems
	11	Dad's class, Free school meals, Had financial problems
	16	Dad's class, Free school meals, Had financial problems, Main income per week
Parental Time	7	Mum reads to child, Dad reads to child, Outings with mum, Outings with dad, Dad's involvement with child
	11	Outings with dad, Dad's involvement with child
Family Shock: Tension	7	Domestic tension, In-law conflicts, Alcoholism, Other family problems
Family Shock: Illness	7	Physical illness, Mental illness, Mental subnormality
Family Shock: Death	7	Death of father, Death of mother
Housing	7	Number of rooms, Has indoor lavatory, Has hot water, Has garden
	11	Number of rooms, Has indoor lavatory
	16	Number of rooms, Has indoor lavatory, Has garden
Parents' Interest in Education	7	Mother's interest in education, Father's interest in education
	11	Mother's interest in education, Father's interest in education

Notes: Any variation in factor content over time owes to the changing structure of the survey.

Table A03: Family Problems: Death and Divorce at Age 7

	Coef.	S.E.
Discrimination		
	0.85	0.18
Difficulty		
Divorce, separation, or desertion	4.05	0.71
Death of father	5.58	1.02
Death of mother	6.85	1.28

Table A04: Family Problems: Illness, Age 7

	Coef.	S.E.
Discrimination		
	1.74	0.08
Difficulty		
Physical illness or disability	2.40	0.06
Mental illness or neurosis	2.67	0.07
Mental subnormality	3.10	0.09

Table A05: Family Problems: Tension or Conflict, Age 7

	Coef.	S.E.
Discrimination		
	2.64	0.10
Difficulty		
Domestic tension	1.83	0.04
Other family shock	2.21	0.05
In-law conflicts	2.42	0.06
Alcoholism	2.74	0.07

Table A06: Family SES, Age 7

	Coef.	S.E.
Discrimination		
Dad's socio-economic status	0.95	0.04
No unemployment issues	2.71	0.08
No housing issues	2.71	0.08
No financial issues	2.71	0.08
Difficulty		
Dad's SES: Level 2	-3.12	0.11
Dad's SES: Level 3	-1.48	0.05
Dad's SES: Level 4	-1.36	0.05
Dad's SES: Level 5	1.07	0.04
Dad's SES: Level 6	1.72	0.06
Dad's SES: Level 7	3.44	0.12
No unemployment issues	-2.10	0.03
No housing issues	-1.68	0.03
No financial issues	-1.62	0.03

Table A07: Housing Quality, Age 7

	Coef.	S.E.
Discrimination		
Number of rooms	0.76	0.03
Has garden	2.03	0.05
Has hot water	2.03	0.05
Has indoor lavatory	2.03	0.05
Difficulty		
Number of rooms: 2	-8.47	0.41
Number of rooms: 3	-5.44	0.22
Number of rooms: 4	-2.74	0.11
Number of rooms: 5	-0.58	0.03
Number of rooms: 6	2.15	0.08
Number of rooms: 7	3.57	0.14
Number of rooms: 8	4.58	0.18
Number of rooms: 9	5.57	0.23
Number of rooms: 10 or more	6.60	0.28
Has garden	-2.26	0.04
Has hot water	-1.91	0.03
Has indoor lavatory	-1.65	0.03

Table A08: Parental Time, Age 7

	Coef.	S.E.
Discrimination		
Outings with mum	2.20	0.08
Outings with dad	2.21	0.08
Mum reads to child	1.40	0.04
Dad's involvement $\in (0, \text{mum's})$	1.00	0.05
Dad's involvement $\geq$ mum's	1.46	0.05
Dad reads to child	1.90	0.06
Difficulty		
Outings with mum: occasionally	-2.77	0.06
Outings with mum: weekly	-1.34	0.03
Outings with dad: occasionally	-1.98	0.04
Outings with dad: weekly	-0.70	0.02
Mum reads to child: occasionally	-1.57	0.03
Mum reads to child: weekly	0.03	0.02
Dad's involvement $\in (0, \text{mum's})$	-1.56	0.05

Table A09: School Quality, Age 7

	Coef.	S.E.
Discrimination		
School type: non-LEA school	-0.05	0.01
Streaming: ability	-0.05	0.01
Has PTA	-0.05	0.01
Class size	2.98	0.31
School size: second quartile	1.57	0.11
School size: third quartile	1.95	0.13
School size: fourth quartile	2.06	0.13
Has social events for parents	-0.05	0.01
Has educational meetings for parents	-0.05	0.01
Difficulty		
School type: non-LEA school	-71.72	21.17
Streaming: ability	-53.22	15.71
Has PTA	-33.98	10.03
Class size: second quartile	-0.69	0.02
Class size: third quartile	0.15	0.02
Class size: fourth quartile	0.89	0.03
Has social events for parents	-1.28	0.05
Has educational meetings for parents	-1.04	0.04

Table A10: Family SES, Age 11

	Coef.	S.E.
Discrimination		
Free school meals	-0.00	0.03
Dad's socio-economic class	0.55	2.98
No financial issues	-0.00	0.03
Difficulty		
Free school meals	-2634.40	94499.86
Dad's SES: Level 2	-5.22	25.62
Dad's SES: Level 3	-2.27	10.84
Dad's SES: Level 4	1.35	6.37
Dad's SES: Level 5	2.23	10.61
Dad's SES: Level 6	5.35	26.29
No financial issues	2517.57	90308.85

Table A11: Housing Quality, Age 11

	Coef.	S.E.
Discrimination		
Number of rooms	2.14	0.23
Has indoor lavatory	0.54	0.05
Difficulty		
Number of rooms: 2	-4.39	0.29
Number of rooms: 3	-3.08	0.16
Number of rooms: 4	-1.76	0.08
Number of rooms: 5	-0.49	0.02
Number of rooms: 6	0.98	0.04
Number of rooms: 7	1.71	0.07
Number of rooms: 8	2.22	0.10
Number of rooms: 9	2.64	0.12
Number of rooms: 10 or more	3.00	0.15
Has indoor lavatory	-4.37	0.36

Table A12: Parental Time, Age 11

	Coef.	S.E.
Discrimination		
Outings with dad	1.99	0.23
Dad's involvement $\in (0, \text{mum's})$	0.93	0.10
Dad's involvement $\geq$ mum's	1.56	0.14
Difficulty		
Outings with dad: occasionally	-1.74	0.09
Outings with dad: weekly	-0.04	0.01
Dad's involvement $\in (0, \text{mum's})$	-1.54	0.08
Dad's involvement $\geq$ mum's	-1.57	0.08

Table A13: School Quality, Age 11

	Coef.	S.E.
Discrimination		
Student-teacher ratio	1.16	0.05
School type: LEA junior and infants	-1.06	0.06
School type: LEA other or non-LEA	-2.71	0.13
School size: second quartile	1.71	0.09
School size: third quartile	2.19	0.10
School size: fourth quartile	1.95	0.11
Streaming: ability	1.07	0.04
Difficulty		
Student-teacher ratio: second quartile	-1.12	0.04
Student-teacher ratio: third quartile	0.08	0.02
Student-teacher ratio: fourth quartile	1.25	0.04
School type: LEA junior and infants	0.07	0.02
School type: LEA other or non-LEA	-1.11	0.03

Table A14: Family SES, Age 16

	Coef.	S.E.
Discrimination		
Free school meals	-0.01	0.03
Father's socio-economic class	0.92	0.99
No financial issues	-0.01	0.03
Difficulty		
Free school meals	-219.32	683.60
Dad's SES: Level 2	-3.42	2.90
Dad's SES: Level 3	-1.82	1.47
Dad's SES: Level 4	-1.71	1.37
Dad's SES: Level 5	0.79	0.63
Dad's SES: Level 6	1.37	1.10
Dad's SES: Level 7	3.51	2.98
No financial issues	214.05	667.17

Table A15: Housing Quality, Age 16

	Coef.	S.E.
Discrimination		
Number of rooms	0.39	0.05
Has hot water	3.07	0.16
Has indoor lavatory	3.07	0.16
Difficulty		
Number of rooms: 2	-18.14	2.25
Number of rooms: 3	-13.28	1.56
Number of rooms: 4	-6.38	0.73
Number of rooms: 5	-1.01	0.12
Number of rooms: 6	3.24	0.36
Number of rooms: 7	5.89	0.66
Number of rooms: 8	7.95	0.90
Number of rooms: 9	9.91	1.14
Number of rooms: 10 or more	11.46	1.33
Has hot water	-2.45	0.06
Has indoor lavatory	-2.03	0.05

Table A16: School Quality, Age 16

	Coef.	S.E.
Discrimination		
Student-teacher ratio	0.31	0.02
Streaming: mixed + setting	-0.22	0.02
Streaming: mixed ability	-0.37	0.03
Streaming: other	-0.20	0.06
Has PTA	0.36	0.02
School type: grammar/LEA/other	-2.18	0.07
School type: non-LEA school	-2.75	0.10
School size: second quartile	4.26	0.37
School size: third quartile	8.51	0.70
School size: fourth quartile	11.47	0.80
Difficulty		
Student-teacher ratio: second quartile	-3.49	0.24
Student-teacher ratio: third quartile	0.10	0.06
Student-teacher ratio: fourth quartile	3.66	0.24

Table A17: Family Demographics

	Coef.	S.E.
Discrimination		
Total siblings (inverted)	0.24	0.02
Mum's age at birth (inverted)	0.85	0.03
Mum's years of education	2.98	0.14
Dad's years of education	1.94	0.05
First-born child	0.88	0.03
Difficulty		
Has 3 siblings	-9.95	0.97
Has 2 siblings	-5.17	0.51
Has 1 sibling	0.58	0.10
Has no siblings	10.70	1.05
Mum's age at birth: 31–35	-2.85	0.08
Mum's age at birth: 25–30	-0.97	0.03
Mum's age at birth: 20–24	0.84	0.03
Mum's age at birth: 19 or less	3.61	0.10

Table A18: Female Transition Matrices

Panel A: Wages

	Destination state					
Origin state	Unemp.	Q1	Q2	Q3	Q4	Q5
Unemp.	0.288	0.224	0.178	0.130	0.104	0.076
Q1	0.225	0.251	0.231	0.157	0.092	0.044
Q2	0.189	0.186	0.199	0.214	0.145	0.067
Q3	0.170	0.107	0.176	0.205	0.218	0.125
Q4	0.175	0.069	0.104	0.175	0.230	0.247
Q5	0.168	0.058	0.076	0.145	0.238	0.316

Panel B: Physical Health

	Destination state			
Origin state	1	2	3	4
1	0.350	0.413	0.224	0.013
2	0.126	0.316	0.489	0.068
3	0.033	0.214	0.497	0.255
4	0.007	0.079	0.549	0.365

Panel C: Mental Health

	Destination state				
Origin state	1	2	3	4	5
1	0.363	0.246	0.204	0.131	0.056
2	0.253	0.182	0.241	0.205	0.118
3	0.138	0.173	0.222	0.269	0.198
4	0.069	0.104	0.221	0.263	0.342
5	0.030	0.052	0.150	0.337	0.430

Notes: Entries report transition probabilities between adjacent observed states using the multiply imputed dataset. Rows denote the origin state and columns the destination state. Physical and mental health states are ordered such that higher values correspond to better health.

Table A19: Correlations among skills and environments in the final multinomial logit

Panel A: Men

	C7	I7	E7	C11	I11	E11	C16	I16	E16	Hosp.	Social	P. time	SES	School	P. demo.
C7	1.000														
I7	0.583	1.000													
E7	0.499	0.374	1.000												
C11	0.743	0.389	0.382	1.000											
I11	0.355	0.358	0.199	0.415	1.000										
E11	0.354	0.178	0.432	0.475	0.466	1.000									
C16	0.688	0.362	0.378	0.853	0.374	0.459	1.000								
I16	0.365	0.242	0.280	0.445	0.295	0.403	0.594	1.000							
E16	0.283	0.147	0.310	0.360	0.198	0.437	0.499	0.791	1.000						
Hosp.	-0.064	-0.049	-0.017	-0.070	-0.089	-0.054	-0.079	-0.085	-0.056	1.000					
Social	-0.203	-0.140	-0.123	-0.183	-0.117	-0.120	-0.210	-0.187	-0.155	0.063	1.000				
P. time	0.233	0.168	0.151	0.262	0.139	0.181	0.312	0.281	0.246	-0.008	-0.195	1.000			
SES	0.215	0.129	0.098	0.218	0.090	0.115	0.234	0.193	0.130	-0.015	-0.147	0.222	1.000		
School	-0.018	-0.049	-0.035	0.016	-0.019	-0.020	0.034	0.011	0.005	-0.011	0.029	0.001	0.025	1.000	
P. demo.	0.207	0.119	0.060	0.289	0.128	0.122	0.312	0.176	0.132	-0.039	-0.095	0.272	0.158	-0.001	1.000

Panel B: Women

	C7	I7	E7	C11	I11	E11	C16	I16	E16	Hosp.	Social	P. time	SES	School	P. demo.
C7	1.000														
I7	0.625	1.000													
E7	0.494	0.462	1.000												
C11	0.736	0.431	0.366	1.000											
I11	0.405	0.381	0.230	0.469	1.000										
E11	0.326	0.204	0.427	0.404	0.460	1.000									
C16	0.697	0.408	0.367	0.854	0.426	0.391	1.000								
I16	0.362	0.268	0.273	0.416	0.333	0.367	0.584	1.000							
E16	0.263	0.173	0.292	0.319	0.235	0.412	0.463	0.774	1.000						
Hosp.	-0.070	-0.067	-0.060	-0.098	-0.087	-0.069	-0.089	-0.063	-0.019	1.000					
Social	-0.193	-0.149	-0.122	-0.187	-0.119	-0.119	-0.204	-0.168	-0.121	0.053	1.000				
P. time	0.206	0.169	0.144	0.221	0.124	0.126	0.252	0.212	0.163	-0.041	-0.208	1.000			
SES	0.176	0.144	0.100	0.192	0.116	0.075	0.201	0.159	0.076	-0.048	-0.139	0.179	1.000		
School	-0.013	-0.029	0.005	0.018	-0.014	0.003	0.027	0.026	0.006	0.013	0.050	0.057	0.051	1.000	
P. demo.	0.198	0.130	0.067	0.260	0.127	0.095	0.293	0.179	0.149	-0.019	-0.097	0.218	0.135	0.009	1.000

Notes: Entries are Pearson correlations calculated separately in each of the 30 completed imputation datasets and averaged across imputations. The samples contain 7,199 men and 6,809 women. All variables are those included in the final multinomial-logit specification. C7 = Cognitive skill, age 7; I7 = Internalizing skill, age 7; E7 = Externalizing skill, age 7; C11 = Cognitive skill, age 11; I11 = Internalizing skill, age 11; E11 = Externalizing skill, age 11; C16 = Cognitive skill, age 16; I16 = Internalizing skill, age 16; E16 = Externalizing skill, age 16; Hosp. = Hospitalization; Social = Social problems; P. time = Parental time; SES = Family SES; School = Primary-school quality; P. demo. = Parental demographics.

B Item Response Models

Item response models are a family of mathematical models that explain the relationship between latent variables and observed outcomes. They are similar in spirit to the factor models outlined in the main text but are adapted to handle discrete outcome data, as is the case for a large fraction of our outcome variables. The following outlines the specific item response models that we use to obtain the early-life environments described in the main text. Our notation follows Harvey and Hammer (1999).

Assume that we have a set of j binary outcomes, x, which describe a true early-life environment, θ, for an individual, where θ is continuous and normally distributed with mean zero and variance one. For example, family socioeconomic status is described by the child's access to free school meals, the father's socioeconomic class and whether the family has experienced financial difficulties. The model begins by assuming a functional relationship between θ and the probability that a given element of x takes the value one. This relationship is referred to as the item characteristic curve and is assumed to be monotonically increasing in θ. We choose to model the characteristic curve using the logistic function, meaning that, as socioeconomic status, θ, increases, the probability of, say, not experiencing financial difficulties also increases. The red line in Figure B01 represents the item characteristic curve.

It is unlikely that the probability of x_j taking a value of one scales with θ in the same way across all j. To this end, we introduce a *difficulty* parameter, b_j, which shifts the characteristic curve to the left or right depending on whether the level of θ required to activate x_j is high or low. Second, we allow different characteristic curves to have varying slopes by introducing a *discrimination* parameter, a_j. The blue and green lines in Figure B01 illustrate the effects of changing the difficulty and discrimination parameters, respectively. This specification is referred to in the literature as a two-parameter item response model. The model and its parameters can then be estimated by maximum likelihood (Brock and Aitkin 1981). The results of our estimation are reported in Tables A03 - A17.

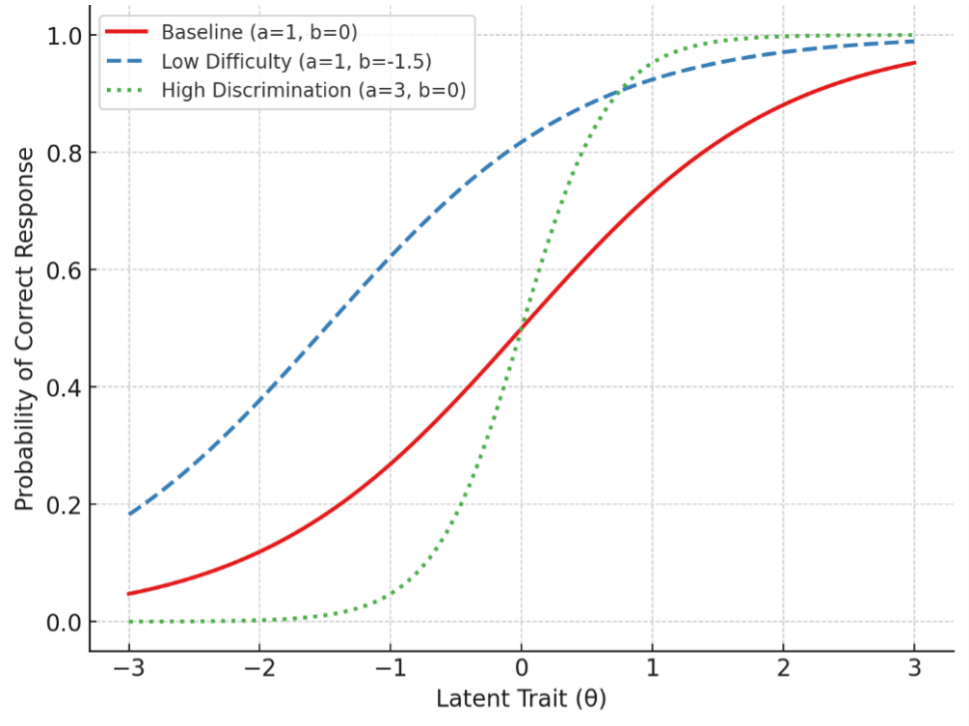


Figure B01: Item Characteristic Curves

C Multiple Imputation by Chained Equations (MICE)

A significant challenge when working with the NCDS data, particularly the adult waves, is the high degree of attrition, which is unlikely to be completely random. For instance, individuals from disadvantaged socioeconomic backgrounds in childhood, those with worse mental health and those with lower early cognitive ability are disproportionately likely to be missing in later waves (Silverwood et al. 2024). Restricting our analysis to complete cases is therefore likely to bias our estimates and inflate our standard errors due to the resulting reduction in sample size.[13]

To address this issue, we use the method of multiple imputation by chained equations (MICE) outlined by Silverwood et al. (2024). The goal is to reproduce the variance-covariance matrix of our dataset that would be observed in the absence of missingness. Developed by Rubin (1976), MICE is a Bayesian approach that estimates multiple possible values for each missing observation, which are then combined for analysis. The uncertainty surrounding the imputed elements represents the remaining uncertainty about the "true" parameter value. This uncertainty is omitted in simpler approaches, such as mean imputation.

MICE is a semi-parametric approach, meaning that we do not need to impose restrictive distributional relationships between observed and missing variables that may bias our analysis. We follow Azur et al. (2011) and construct each imputed dataset in six steps:

1. Use a naïve mean imputation to choose priors for each of the variables with missing values.
2. Randomly set one variable, x_j, from the set of variables with missing data, J, back to missing.
3. Use a linear model if the variable is continuous, and a multinomial model if it is categorical, to predict values for the missing variable based on the imputed data, yielding $\widehat{x}_j^{(1)}$.
4. Retain $\widehat{x}_j^{(1)}$, then randomly reset another variable, $x_{j'}^{(1)}$, where $j' \neq j$ and $j' \in J$, and calculate $\widehat{x}_{j'}^{(1)}$.
5. Repeat steps 2–4 for all remaining variables in J.
6. Repeat steps 2–5 a total of 10 times (as recommended by CITE) to obtain the final imputed dataset, $\widehat{X}_1 = \{\widehat{x}_1^{(10)}, \ldots, \widehat{x}_J^{(10)}\}$.

The procedure above produces a single imputed dataset. We repeat this process 30 times to generate 30 imputed values for each missing observation. The total number of imputations required for convergence is a subjective choice, with recommendations typically ranging from 20 to 50 (Rubin 1996; Silverwood et al. 2024). Since the variance of our estimates depends inversely on the number of imputations, we opt for 30 – the maximum number that is computationally feasible for our analysis.[14] Rubin (1976) shows

[13] Figure C01 illustrates the non-random nature of missingness by plotting missingness in earnings, employment, mental health and physical health over the life cycle. In each case, lower earnings at age 23 are associated with a higher fraction of missing observations at age 50.

[14] The variance of any parameter, V, estimated via MICE is a function of the average within-imputation

that, when the data are missing at random (MAR) and the imputation equations are correctly specified, analysis using the imputed dataset will be unbiased. To this end, we include the variables used in our analysis alongside a range of auxiliary variables recommended by Silverwood et al. (2024) that predict missingness. Our final assumption is that, conditional on the variables outlined in Table 1, childhood skills at ages 7, 11 and 16, the rich set of childhood factors and the auxiliary variables, missingness is random.[15]

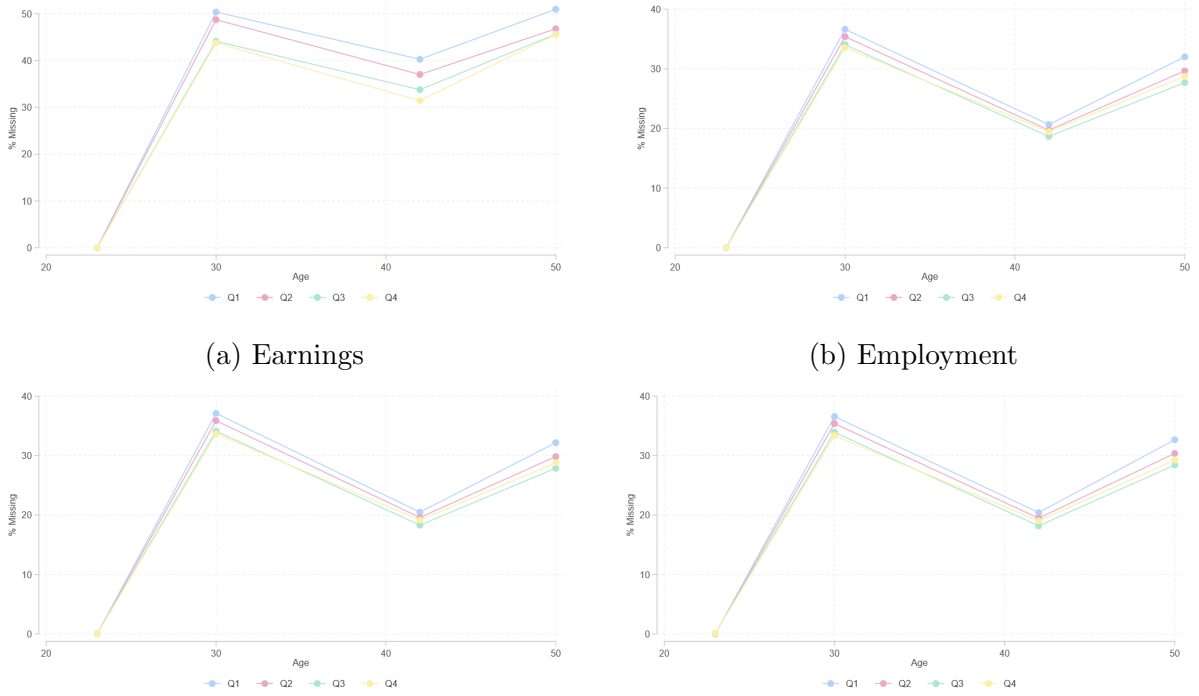
Throughout our analysis, the results we present are based on the imputed data. Table C02 presents summary statistics comparing imputed and non-imputed adult outcomes. For wages, employment, physical health and mental health, the mean is lower in the imputed dataset than in the non-imputed dataset.[16] For each outcome, the difference between the raw and imputed means becomes increasingly negative over the life cycle, supporting our prior that individuals from lower socioeconomic backgrounds are more likely to exit the sample and providing confidence in our subsequent estimates.

variance, $\overline{W}$, and the between-imputation variance, B. For a dataset with m imputations, the variance can be expressed as

$$V = \overline{W} + \left(1 + \frac{1}{m}\right) B.$$

[15]The auxilliary variables that we choose for the analysis are described in Table C01.

[16]In contrast to Table 1, wages for individuals not in employment are coded as missing in this dataset rather than zero. Hence, wages are higher over the life cycle.



(a) Earnings

(b) Employment

(c) Physical Health

(d) Mental Health

Figure C01: Missingness Over the Life Cycle by Wage

Notes: Results show percentage missingness in each wave for a given outcome conditional on being observed at age 23. Missingness is stratified by quartile of the age 23 wage distribution.

Table C01: Multiple Imputation Model: Auxiliary Variables

Age	Variable	Description	Included in Final Model
0	n0region	Region	✓
	n512	Number of persons per room	✓
	n236	Social class of mother's husband	✓
	n622	Sex of child	✓
7	n99	Number of kids under 21 in the household	✓
	age7dv1	Common difficulties age 7 (mother)	✗
	age7dv5	Hospital admissions	✓
	CogAbil7	Cognitive ability summary	✗
	OUTCME01	Non-response at sweep 1	✓
	maw5	Mother worked from birth to 5	✗
	n222	Ever breastfed	✗
	n95	Family moves since child's birth	✓
	n180	Dad reads to child	✓
	age7dv10	Social problems (alcoholism etc.)	✓
	MedExSum7	Summary of medical conditions	✗
11	OUTCME02	Non-response at sweep 2	✓
	n1150	Number of family moves since child's birth	✗
	genability11	Cognitive ability summary	✗
	n1434	Area of world in which mother born	✗
	Amens	Number of household amenities	✓
	age11dv32	Child's positive activities outside school	✓
	n1436	Area of the world in which father born	✗
	Ext11Dec18	Conduct problems	✗
16	n2392	Number of family moves since child's birth	✓
	age16dv46	Sum of favourable learning environments/outcomes re sex educ etc.	✗
	Ext16MTDec18	Conduct problems	✗
	OUTCME03	Non-response at sweep 3	✓
	n2888	How long since child drank alcohol	✓
	n2930	Test 2 - Mathematics comprehension	✗
	n5318	Type of current accommodation	✗
	age16dv47	Sum of good activities performed outside school	✗
	n1721	Child's school attendance	✗
23	n5960	Voted in 1979 general election	✓
	econgstrg	Economic status	✓
	dv5	Number of voluntary activities (youth club, church, etc.)	✗
	OUTCME04	Non-response at sweep 4	✓
	n5113	Legal marital status	✓
33	n502940	Type of accommodation	✓
	n504656	Current member of a trade union/staff association	✓
	dvsoccapital	Social capital score (people turn to for advice, support)	✗
	dvcontentment	Life contentment score	✓
	OUTCME05	Non-response at sweep 5	✓
	n504636	Voted in 1987 general election	✓
42	tenure2	Is current accommodation owned or rented	✓
	dmpart	Participated in NCDS V	✗
	mthimp	Ever wanted to improve their maths	✓
	Org42	Membership in organisations	✓

The list of imputed variables includes all variables that predict missing data, as recommended by the official NCDS handbook. Including too many variables in the imputation procedure will lead to non-convergence and instability. To circumvent this issue we only keep variables that predict missiningess in early life skills, wages, earnings or health outcomes. We also only select variables predicting missingness that themselves have less than 50% missingness. We also omit skills that predict missingness as they are themselves an outcome.

Table C02: Raw and Imputed Summary Statistics

	Raw Mean	Imputed Mean
Log Hourly Wages		
Age 23	1.67	1.67
Age 30	2.40	2.37
Age 42	2.54	2.51
Age 50	2.74	2.71
Employment		
Age 23	0.71	0.70
Age 30	0.83	0.81
Age 42	0.86	0.83
Age 50	0.85	0.81
Physical Health		
Age 23	3.35	3.31
Age 30	3.21	3.12
Age 42	3.09	3.03
Age 50	2.95	2.88
Mental Health		
Age 23	2.91	2.86
Age 30	3.09	2.95
Age 42	2.62	2.54
Age 50	2.72	2.63

Physical and mental health are both categorical variables. The former has 4 categories while the latter has 5. Wages for the non-employed are coded as missing.

D Cluster Choice

Our choice of cluster-quality measures is grounded in the sociology and machine-learning literature (see Milligan and Cooper (1985) and Charrad et al. (2014) for excellent reviews). This appendix describes exactly how each measure is calculated and explains how to interpret Figure A01. We use a combination of four measures to select the number of clusters: the point-biserial correlation (PBC), Hubert's Gamma (HG), R^2 and Hubert's C.

The first two of these measures range between -1 and 1, where 1 represents perfect clustering and -1 represents the perfect misallocation of individuals to clusters. Zero represents the score that would be achieved if the data were perfectly uniformly distributed. As usual, R^2 ranges between 0 and 1. Hubert's C also ranges between 0 and 1, but a score closer to 0 implies higher cluster quality.

Considering the first two metrics, it is rare to observe scores below 0. While clustering scores exceeding 0.7 are common in the natural sciences, scores in sociological data tend to be lower because the underlying relationships are less systematic. This is particularly true when conducting multi-channel sequence analysis, where imperfect correlations between channels systematically reduce cluster quality. For this reason, we refrain from calculating common measures of cluster quality, including silhouette width. Consequently, we search for clustering solutions in which the PBC and Hubert's Gamma exceed 0.3 and Hubert's C falls below the same threshold. The following subsections explain mathematically how each measure of cluster quality is defined.

D.1 Point-Biserial Correlation

The point-biserial correlation measures the capacity of the clustering to reproduce the original distance matrix. PBCs are commonly used to measure the correlation between categorical and continuous variables and play the same role in this analysis. The PBC is defined as:

$$PBC = \frac{\overline{D}_1 - \overline{D}_0}{s}\sqrt{\frac{n_1 n_0}{N^2}}. \tag{9}$$

where $\overline{D}_1$ is the mean distance between points in different clusters and $\overline{D}_0$ is the mean distance between points within clusters. The difference is scaled by the standard deviation of distances, s. The correlation is then adjusted by the relative numbers of between-cluster (n_1) and within-cluster (n_0) pairwise distances, where $N = n_1 + n_0$ is the total number of pairwise distances. The purpose of this final term is to adjust for imbalances in cluster size, which can inflate $\overline{D}_1 - \overline{D}_0$.

D.2 Hubert's Gamma

As a correlational measure, Hubert's Gamma shares many properties with the PBC. The metric is calculated using within- and between-cluster dissimilarities for all points in the dataset. The distance between any two points is judged to be concordant if the within-cluster dissimilarity is strictly less than the between-cluster dissimilarity. The numbers of

concordant data points, $s(+)$, and discordant data points, $s(-)$, are calculated, and the metric is constructed as follows:

$$HG = \frac{s(+) - s(-)}{s(+) + s(-)}. \tag{10}$$

Concordance implies that data points are assigned to suitable clusters. If all data points are concordant, the metric takes a value of 1, representing strong cluster quality. Likewise, if no points are assigned to clusters concordantly, the metric equals -1.

D.3 R2

Our final correlational measure, R^2, is defined as one minus the ratio of within-cluster squared distances (the within-cluster sum of squared distances) to the total sum of squared distances. The calculation is analogous to the standard R^2 used in regression analysis, except that residuals are replaced by distances.

$$R^2 = 1 - \frac{\sum_{k \in K} \sum_{i \in k} d^2\left(i, \mu_k\right)}{\sum_i d^2\left(i, \mu\right)}. \tag{11}$$

where d is the distance function, μ_k is the medoid of a given cluster and μ is the global medoid. If the clustering fits the data perfectly, then the sum of squared distances within clusters will equal zero and R^2 will equal 1.

D.4 Hubert's C

Hubert's C intuitively measures the gap in quality between the current clustering partition and the best partition available for a given distance matrix and fixed number of clusters. Milligan and Cooper (1985) report that Hubert's C is consistently able to recover the correct number of clusters. The index ranges from 0 to 1, where 0 represents perfect clustering. The statistic is defined as:

$$HC = \frac{S_w - S_{\min}}{S_{\max} - S_{\min}}, \qquad S_{\min} \neq S_{\max}. \tag{12}$$

where S_w is the sum of within-cluster distances, $S_{\min}$ is the sum of the N_k smallest distances between all points and $S_{\max}$ is the corresponding sum of the N_k largest distances. Intuitively, if the clusters are well defined, then S_w will equal $S_{\min}$, as the clustering assigns observations such that within-cluster distances are as small as possible. In that case, the numerator will equal 0, indicating a well-defined clustering solution.