Attanasio, Orazio P.; Blundell, Richard W.; Conti, Gabriella; Mason, Giacomo

Working Paper

Inequality in socioemotional skills: A cross-cohort comparison

IFS Working Papers, No. W18/22

Provided in Cooperation with:
Institute for Fiscal Studies (IFS), London

Suggested Citation: Attanasio, Orazio P.; Blundell, Richard W.; Conti, Gabriella; Mason, Giacomo (2018) : Inequality in socioemotional skills: A cross-cohort comparison, IFS Working Papers, No. W18/22, Institute for Fiscal Studies (IFS), London

This Version is available at:
https://hdl.handle.net/10419/200311

Institute for Fiscal Studies

Inequality in socioemotional skills: a cross-cohort comparison

IFS Working Paper W18/22

Orazio Attanasio
Richard Blundell
Gabriella Conti
Giacomo Mason

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

Inequality in socioemotional skills: a cross-cohort comparison

Orazio Attanasio Richard Blundell
Gabriella Conti Giacomo Mason

Institute for Fiscal Studies
Department of Economics, University College London

October 2, 2018

Abstract

We examine changes in inequality in socio-emotional skills very early in life in two British cohorts born 30 years apart. We construct socio-emotional scales comparable across cohorts for both boys and girls, using two validated instruments for the measurement of child behaviour. We identify two dimensions of socio-emotional skills for each cohort: 'internalising' and 'externalising', related to the ability of children to focus their concentration and to engage in interpersonal activities, respectively. Using recent methodological advances in factor analysis, we establish comparability in the inequality of these early skills across cohorts, but not in their average level. We document for the first time that inequality in these early skills has increased across cohorts, especially for boys and at the bottom of the distribution. We also document changes in conditional skills gaps across cohorts. We find an increase in the socio-emotional skills gap in the younger cohort for children born to mothers with higher socio-economic status (education and employment), and to mothers who smoked during pregnancy. The increase in inequality in early socio-emotional skills is particularly pronounced for boys. On the other hand, we find a decline in the skills gradient for children without a father figure in the household. Lastly, we document that socio-emotional skills measured at a much earlier age than in most of the existing literature are significant predictors of outcomes both in adolescence and adulthood, in particular health and health behaviours. Our results show the importance of formally testing comparability of measurements to study skills differences across groups, and in general point to the role of inequalities in the early years for the accumulation of health and human capital across the life course.

Keywords: Inequality, Socio-emotional skills, Cohort studies, Measurement invariance
JEL Classification: J13, J24, I14, I24, C38

Corresponding Author: Orazio Attanasio, Department of Economics, University College London, Gordon Street, London WC1H0AX, UK. Email: o.attanasio@ucl.ac.uk. We thank participants to the NBER/LSE Trans-Atlantic Public Economics Seminar (TAPES) and our two discussants Hilary Hoynes and Gabriel Ulyssea for excellent comments. We are grateful to The Centre for Longitudinal Studies, UCL Institute of Education, for the collection of the data, and to the UK Data Archive and UK Data Service for making them available. However, they bear no responsibility for the analysis or interpretation of the data. Giacomo Mason is funded by an Economics and Social Research Council scholarship.

1 Introduction

Human capital is a key determinant of economic growth and performance and of the resources an individual creates and controls over the life cycle. Human capital is also important for various determinants of individual well-being, ranging from health to life satisfaction. In recent years, the process of human capital accumulation has received considerable attention. There is growing consensus on the fact that human capital is a multidimensional object, with different domains playing different roles in labour market as well as in the determination of other outcomes, including the process of human development. It is also recognised that human capital is the output of a very persistent process, where early years inputs play an important and persistent role.

And yet, there are still large gaps in our knowledge of the process of human capital development. These gaps are partly driven by the scarcity of high quality longitudinal data measuring the evolution over the life cycle of different dimension of human capital. Moreover, there is a lack of consensus on the best measures and on the tools to collect high quality data. As a consequence, even when data are available in different contexts, their comparability is problematic.

In this paper, we focus on an important dimension of human capital, which, so far, has received limited attention: socio-emotional skills. Evidence has shown that gaps in socio-emotional skills emerge at very young ages, and that in the absence of interventions are very persistent across the life cycle (Cunha et al., 2006). However, there is surprisingly little evidence on how inequality in this important dimension of human capital has changed across cohorts. In this paper, we start addressing this gap and focus on the measurement of these skills in two British cohorts: the one of children born in 1970 and the one of children born in 2000. We consider the measurement of socio-emotional skills during early childhood, as these skills have been shown, in a variety of contexts (Almlund et al., 2011) to have important long-run effects. Our goal is to characterise the distribution of socio-emotional skills in these cohorts and compare them. In the last part of the paper, we also consider the predictive power of different socio-emotional skills for health and socioeconomic outcomes.

The main contributions of the paper are four. First, we use two validated scales of childhood behavioural traits and select those items which are comparable across the two cohorts. By performing an exploratory factor analysis, we determine that we need at least two dimensions to characterise socio-emotional skills. We label them as 'internalising' and 'externalising' skills, the former relating to the ability of children to focus their drive and determination and the latter relating to their ability in engaging in interpersonal activities.

Second, we study the comparability of the measures in the two cohorts. In particular, we test for *measurement invariance* of the items we use to estimate the latent factors. Intuitively, if one assumes that a set measures is related to a latent unobserved factor of interest, one can think of this relationship as being driven by the saliency of each measure and the level. If one uses a given measure as the relevant metric for the relevant factor, its saliency will determine the scale of the factor, while some other parameters, which could be driven by the difficulty of a given test or the social norms and attitudes towards a certain type of behaviour, determine the *average level* of the factor. Comparability of estimated factors across different groups (such as different cohorts) assumes that both the parameters that determine the saliency of a given set of measures and the level of the factors do not vary across groups. We find that, for the measures we use and for both factors, we cannot reject measurement invariance for the saliency parameters. However, we strongly reject measurement invariance for the level parameters. These results imply that while the level of inequality across the two cohorts in the skills we consider is comparable, we cannot determine whether the *average levels* of the two factors are larger or smaller in one of the two cohorts.

Third, given the results we obtain on measurement invariance, we proceed to compare the level of inequality in the two types of socio-emotional skills across the two cohorts, for both boys and girls. We find that the most recent cohort is more unequal in both dimensions of socio-emotional skills than the 1970 cohort. This result is particularly apparent for boys, and when looking at differences by maternal background. Fourth, we study whether the socio-emotional

skills we observe at a young age are an important determinant of a variety of adolescent (and adult, for the older BCS cohort) outcomes. We find that socio-emotional skills at age five are more predictive than cognitive skills for unhealthy behaviours like smoking and measures of health capital such as body mass index. The effect of cognition, instead, dominates for educational and labour market outcomes.

The rest of the paper is organised as follows. In section 3, we briefly discuss the data we use in the analysis. In section 4, we discuss the methods we use to identify the number of dimensions in socio-emotional skills and how we estimate the latent factors that represent them. In section 5, we discuss the comparability of factors estimated with a given set of measures from different groups and the *measurement invariance* tests we use. Section 6 reports our empirical results, while section 7 concludes the paper.

2 Literature

The importance of cognition in predicting life course success is well established in the economics literature. However, in recent years the role played by 'non-cognitive' traits is being increasingly investigated. These traits include constructs as different as psychological and preference parameters such as social and emotional skills, locus of control and self-esteem, personality traits (e.g. conscientiousness), and risk aversion and time preferences. Given the vastity of this literature, we briefly review below the main papers on the determinants and consequences of socio-emotional traits which are more directly related to our work, and we refer to other sources for more exhaustive reviews (Borghans et al., 2008; Almlund et al., 2011; Goodman et al., 2015; Kautz et al., 2014).

Consequences of socio-emotional traits One of the first papers to show the importance of non-cognitive personality variables for wages was Bowles et al. (2001). Heckman et al. (2006) suggested that non-cognitive skills are at least as important as cognitive abilities in determining a variety of adults outcomes. Lindqvist and Vestman (2011), using data based on personal interviews conducted by a psychologist during the Swedish military enlistment exam, show that both cognitive and noncognitive abilities are important in the labour market, but for different outcomes: low noncognitive abilities are more correlated with unemployment or low earnings, while cognitive ability is a stronger predictor of wages for skilled workers. Segal (2013), using data on young men from the US National Education Longitudinal Survey, shows that eight-grade misbehaviour is important for earnings over and above eight-grade test scores. Layard et al. (2014) find that childhood emotional health (operationalised using the same mother-reported Rutter scale we use in the 1970 British cohort study) at ages 5, 10 and 16 is the most important predictor of adulthood life satisfaction and life course success.

There are only few studies in economics specifically studying "non-cognitive" traits and health behaviors. Conti et al. (2010) and Conti et al. (2011) were the first to consider three early endowments, including child socio-emotional traits and health in addition to cognition, using rich data from the 1970 British cohort study. They find strong evidence that non-cognitive traits promote health outcomes and healthy behaviors, and than not accounting for them overestimates the effects of cognition; additionally, they document that child cognitive traits are more important predictors of employment and wages than socio-emotional traits or early health. Chiteji (2010) used the US Panel Study of Income Dynamics (PSID) and found that future orientation and self-efficacy (related to emotional stability) are associated with less alcohol consumption and more exercise. Cobb-Clark et al. (2014) used the Australian HILDA data and found that an internal locus of control (also related to emotional stability, perceived control over one's life) is related to better health behaviours (diet, exercise, alcohol consumption anwhad smoking). Mendolia and Walker (2014) used the Longitudinal Study of Young People in England and found that individuals with external locus of control, low self-esteem, and low levels of work ethics, are more likely to engage in risky health behaviours. Savelyev and Tan

(2017) show that the association between personality traits and health behaviours also holds in a high-IQ sample (the Terman Sample). Heckman et al. (2016) use, instead, early risky behaviours to measure socio-emotional traits, and confirm their predictive power for health behaviours and health outcomes.

Very few papers attempt to make cross-cohorts comparisons about the importance of socio-emotional skills. Blanden et al. (2007) – one of the closes study to ours – examine cognitive skills, non-cognitive traits, educational attainment and labour market attachment as mediators of the decline in inter-generational mobility in UK between the 1958 and the 1970 cohorts. The authors take great care in selecting non-cognitive items to be as comparable as possible across cohorts, from the Rutter scale at age 10 for the 1970 cohort and from the Bristol Social Adjustment Guide for the 1958 cohort; however, they do not carry out a formal test of measurement invariance and the do not construct factor scores fully comparable across cohorts as we do. Another paper related to ours is the one by Reardon and Portilla (2016), who study recent trends in income, racial, and ethnic school gaps in several dimensions of school readiness, including academic achievement, self-control, and externalizing behavior, at kindergarten entry, using comparable data from the Early Childhood Longitudinal Studies (ECLS-K and ECLS-B) for cohorts born from the early 1990s to the 2000–2010 period in the US. They find that readiness gaps narrowed modestly from 1998 to 2010, particularly between high- and low-income students and between White and Hispanic students. Lastly, Deming (2017) uses a comparable set of skill measures and covariates across survey waves for the NLSY79 and the NLSY97, and finds that the labour market return to social skills was much greater in the 2000s than in the mid-1980s and 1990s.

Determinants of socio-emotional traits Equally flourishing has been the literature on the determinants of child socio-emotional skills, which ranges from reduced-form, correlational or causal estimates, to more structural approaches. One of the first papers by Segal (2008) has shown that a variety of family and school characteristics predict classroom behaviour. Carneiro et al. (2013) study the intergenerational impacts of maternal education, using data from the NLSY79 and an instrumental variable strategy; they find strong effects in terms of reduction in children's behavioural problems. Cunha et al. (2010) and Attanasio et al. (2018) both estimate production functions for child cognitive and socio-emotional development (in US and Colombia, respectively), and find an important role played by parental investments.

Interventions for improving Social and Emotional Learning (SEL) in a school setting have shown significant improvements in socio-emotional skills, attitudes, behaviours, and academic performance (Durlak et al., 2011), and a substantial positive return on investments (Belfield et al., 2015); after-school programs have been shown to be equally effective (Durlak et al., 2010).

Additionally, it has been shown that a key mechanism through which early childhood interventions improve adult socioeconomic and health outcomes is by boosting socio-emotional skills, such as four teacher-reported behavioural outcomes in the project STAR[1] (Chetty et al., 2011), reductions in externalising behaviour (from the Pupil Behavior Inventory) at ages 7-9 in the Perry Preschool Project (Heckman et al., 2013; Conti et al., 2016), or improvements in task orientation at ages 1-2 in the Abecedarian Project (Conti et al., 2016).

In sum, even if the literature on the determinants and consequences of socio-emotional skills has been booming, most papers use skills measured in late childhood or in adolescence; and no paper in economics formally tests for invariance of measurements across different groups and constructs fully comparable scores. In this paper, we use measures of child socio-emotional development at age 5, hence before formal schooling starts; and we construct comparable scales across the two cohorts we study (the 1970 and the 2000 British cohorts), so that we can investigate changes in inequality in early development, their determinants, and consequences, in a parallel fashion.

[1] Student's effort, initiative, non-participatory behavior, and how the student is seen to 'value' the class.

3 Data

We use information from two nationally representative longitudinal studies in the UK. The studies follow the lives of children born approximately 30 years apart: the British Cohort Study (BCS) surveys individuals born in 1970, and the Millennium Cohort Study (MCS) includes births between 2000 and 2002. The British Cohort Study includes all individuals born across Great Britain in a single week (April 4-11, 1970). Cohort members' families – and subsequently members themselves – were surveyed on multiple occasions. For this paper we augment the information at the five-year survey with data from birth, adolescence (16), and adulthood (30, 38, 42). The Millennium Cohort Study follows individuals born in the UK between September 2000 and January 2002. We use the birth survey, and the sweeps at around 5, 11, and 14 years.[2]

Our main focus is on socio-emotional skills of children around age five. We take advantage of the longitudinal nature of the cohorts by merging information from surveys before and after age five. From the birth survey, we include information on gestational age and weight at birth, previous stillbirths, parity, maternal smoking in pregnancy, maternal age, height, and marital status. From the five year survey, we extract maternal education, employment status, and the father's occupation. All the above variables are transformed or recoded to maximise comparability between the two studies. Furthermore, we add some adolescent risky behaviours such as smoking and BMI, with the caveat that these are surveyed at 16 and 14 in BCS and MCS respectively. Finally, for the 1970 cohort we also include measures of adult educational attainment, BMI, and income. Variable definitions are available in Table A2.

Ideally, we would compare socio-emotional skills alongside cognitive skills. However, the cognitive tests administered to each cohort have no overlap, even at the item level. We thus use the available cognitive tests in each cohort to estimate simple confirmatory factor model with a single latent dimension, separately by cohort (see Table A2 for the tests used). Unlike the other indicators in our analysis, cognitive skills are thus not comparable across cohort.

Another complication arises from the fact that, differently from the British Cohort Study, the Millennium Cohort Study has a stratified design. It oversamples children living in administrative areas characterised by higher socioeconomic deprivation and larger ethnic minority population (Plewis et al., 2007). We rebalance the MCS sample to make it nationally representative by excluding from the analysis a fraction of observations from the oversampled areas, proportionally to their sampling probability.[3] We also restrict our sample to individuals born in England. Finally, we restrict the sample to cases where the respondent in the five-year followup was the natural mother, and where there is complete information on socioemotional skills. The final sample contains 9,545 individuals from the British Cohort Study, and 5,436 from the Millennium Cohort Study.

4 Dimensions of socio-emotional skills

Child socio-emotional skills are an unobservable and difficult to measure construct. Over recent years, the measurement of such skills has evolved and, over time, different measures have been used. As we discuss below, this makes the comparison of socio emotional skills across different groups, assessed with different tools, difficult.

A common approach to infer a child's socio-emotional development is based on behavioural screening scales. As part of these tools, mothers (or teachers) indicate whether their children exhibit a series of behaviours – the *items* of the scale. In the British and Millennium Cohort Studies, two different scales were employed. In the BCS, the

[2] All data is publicly available at the UK Data Service (Chamberlain, 2013; Butler, 2016a,b, 2017; University Of London. Institute Of Education. Centre For Longitudinal Studies, 2016a,b,c, 2017a,b,c,d).

[3] See Table 5.5 in Plewis et al. (2007). This choice is mainly driven by software limitations. The `lavaan` package in **R** (Rosseel, 2012) is the most suitable tool for our invariance analysis, but it does not allow to use weights when outcomes are categorical, as it is the case for the socio-emotional measurements.

Rutter A Scale was used (Rutter et al., 1970) while in the MCS cohort, mothers were administered the Strengths and Difficulties Questionnaire (SDQ, Goodman, 1994, 1997). The SDQ was created as an update to the Rutter scale. It encompasses more recent advances in child psychopathology, and emphasises positive traits alongside undesirable ones (Stone et al., 2010). The Rutter and SDQ scales are reproduced in Table A1; they have 23 and 25 items each, respectively. In the child psychiatry and psychology literatures, the Rutter and SDQ behavioural screening scales are regarded as measures of behavioural problems and mental health. However, in our analysis we follow the economics literature, and - after having recoded them accordingly - we interpret them as measures of positive child development (Goodman and Goodman, 2011).

While the Rutter and SDQ scales are similar in their components, there is no a priori reason to expect them to be directly comparable. First, the overlap of behaviours described in the two scales is only partial. Second, the wording of each item is slightly different, both in the description and in the options that can be selected as answers. Third, the different ordering of the items within each scale might lead to order effects. Fourth, and no less importantly, the interpretation of each behaviour by respondents living 30 years apart (1975 vs 2005/6) might differ due to a host of evolving societal norms.

As our goal is to compare socio-emotional skills across the two cohorts, we construct a new scale by retaining the items that are worded in a similar way across the two original Rutter and SDQ scales, and making some slight coding adjustments to maximise comparability. In what follows, we will consider the included items to be the same *measure* in the two cohorts. The wording of the items we will be using in the analysis is presented in Table 1: we retain 13 items for the BCS (two of them are grouped) and 11 for the MCS with high degree of comparability.[4]. Item-level prevalence by cohort and gender is in Table A4. We see that, in general, item prevalence is more similar across genders within the same cohort, than across cohorts. For the majority of items, there is a lower prevalence of problematic behaviours in the MCS than in the BCS; however, four items (distracted, tantrums, fearful, aches) show a higher prevalence in 2005 than in 1975. Regardless, a simple cross-cohort comparison of item-level prevalence is misleading because of changing perceptions and norms about what constitutes problematic behaviour in children. The analysis in section 5 tackles this issue.

In the remainder of this section, we analyse the properties of the new scale. In particular, we study the *factor structure* of our scale. Namely, we establish how many latent dimensions of socio-emotional skills the scale is capturing, and which items are measuring which dimension. We then estimate the parameters of the factor models that corresponds to our choice of dimension and attribution of specific items to factors. In the following section, we investigate to what extent socio-emotional skills are measured in the same way across cohorts.

4.1 Exploratory analysis

The original Rutter scale, used for the BCS cohort, distinguishes behaviours into two subscales: *anti-social* and *neurotic* (Rutter et al., 1970). This two-factor conceptualisation has been validated using data from multiple contexts, and the latent dimensions have been broadly identified as externalising and internalising behaviour problems.[5] The Strength and Difficulties Questionnaire, used for the MCS cohort, was conceived to have five subscales of five items each. The five subscales are: *hyperactivity*, *emotional symptoms*, *conduct problems*, *peer problems*, and *prosocial*. This five-factor structure has been validated in many contexts (Stone et al., 2010); lower-dimensional structures have

[4]We exclude from the analysis items that were completely different between the two questionnaires, although we could have included them in the factor analysis and treated them as missing in the cohort were they were not administered. While this could have improved efficiency, we decided to rely on a more coherent set of measures to maximise comparability between the two cohorts.

[5]See for example Fowler and Park (1979); Venables et al. (1983); Tremblay et al. (1987); Berglund (1999); Klein et al. (2009). However, in some cases a three-factor structure was found to better fit the data, with the externalising factor separating into two factors seemingly capturing aggressive and hyperactive behaviours (Behar and Stringfield, 1974; McGee et al., 1985).

been also suggested (Dickey and Blumberg, 2004). Recent research has shown that there are some benefits to using broader subscales that correspond to the externalising and internalising factors in Rutter, especially in low-risk or general population samples (Goodman et al., 2010).

We use exploratory factor analysis (EFA) to assess the factors structure of our 11-item scale combining Rutter and SDQ.[6] We start by investigating the number of latent constructs that are captured by the scale, using different methods developed in the psychometric literature, and now also used in the economics literature. The results are displayed in Table A6. As pointed out in Conti et al. (2014), there is relatively little agreement among procedures; this is the case especially for the Rutter items in the BCS data, where different methods suggest to retain between 1 and 3 factors, while most methods suggest to retain 2 factors in the MCS. In our analysis, we adopt two factors and a dedicated measurement system, where each measure reflects only one factor. This choice is justified both by the child psychology literature cited above, and as compromise to work with the same number of factors in the two cohorts. The two-factor EFA delivers a neat and sensible separation between items, as shown in Table A7: reassuringly, similarly-worded items load on the same factor across the two cohorts, and also the magnitude of the respective loadings (measuring the strength of the association between the item and the factor) is very smilar. Following previous research, we name the first dimension *Externalising skills* (EXT, comprising the items restless, squirmy/fidgety, fights/bullies, distracted, tantrums, and disobedient) and the second dimension *Internalising skills* (INT, comprising the items worried, fearful, solitary, unhappy, and aches).

4.2 Factor model

Equipped with the factor structure inferred in the previous section, we specify a multiple-group factor analysis model to formally quantify the strength of the relationship between the observed items in our scale and the two latent socio-emotional skills. We specify two groups of children $c = \{BCS, MCS\}$, corresponding to the two cohorts. Each individual child is denoted by $j = 1 \dots N_c$, where N_c is the number of children in cohort c. For each child j in cohort c, we observe categorical items X_{ijc} with $i = 1, \dots, 11$, corresponding to the eleven maternal reports in Table 1. We assume that each child is characterised by a latent bi-dimensional vector of externalising and internalising socio-emotional skills $\boldsymbol{\theta}_{jc} = (\theta_{jc}^{EXT}, \theta_{jc}^{INT})$, as shown by the EFA in the previous section.

Children are assumed to have a latent continuous propensity X_{ijc}^* for each item $i = 1, \dots, I$. We model this propensity as a function of item- and cohort-specific intercepts ν_{ic} and loadings $\boldsymbol{\lambda}_{ic}$, and the child's latent skills $\boldsymbol{\theta}_{jc}$, plus an independent error component u_{ijc}. The propensity for each item can be written as follows:

$$X_{ijc}^* = \nu_{ic} + \boldsymbol{\lambda}_{ic}\boldsymbol{\theta}_{jc} + u_{ijc} \qquad \text{for} \quad i = 1, \dots, 11$$

or more compactly:

$$\boldsymbol{X}_{jc}^* = \boldsymbol{\nu}_c + \boldsymbol{\Lambda}_c\boldsymbol{\theta}_{jc} + \boldsymbol{u}_{jc} \tag{4.1}$$

We make the common assumption of a dedicated (or congeneric) factor structure, where each measure is assumed to load on only one latent dimension (Heckman et al., 2013; Conti et al., 2010; Attanasio et al., 2018). We mirror the structure found in the exploratory factor analysis (see Table A7), and assume that items 1-6 load exclusively on the

[6] Factor-analytic methods have long been used in psychology, and in recent years they have become increasingly popular in economics, especially to meaningfully aggregate high-dimensional items measuring different aspects of common underlying dimensions of human development. The EFA is performed decomposing the polychoric correlation matrix of the items and using weighted least squares, and the solution is rescaled using oblique factor rotation (*oblimin*). We use the **R** package `psych`, version 1.8.4 (Revelle, 2018).

externalising factor and items 7-11 on the internalising factor.[7]

The discrete ordered nature of the observed measures X_{ijc} is incorporated by introducing item- and cohort-specific threshold parameters τ_{ic} (Muthén, 1984). The observed measures as a function of the propensities X^* can be then written as follows:

$$X_{ijc} = s \qquad \text{if } \tau_{s,ic} \leq X^*_{ijc} < \tau_{s+1,ic} \qquad \text{for } s = 0, 1, 2 \tag{4.2}$$

with $\tau_{0,ic} = -\infty$ and $\tau_{3,ic} = +\infty$. Notice that we recode all ordered items to have higher values for *better* behaviours, so that our latent vectors can be interpreted as favourable skills and not behavioural problems.

The model implies the following expression for the mean and covariance structure of the latent propensities:

$$\boldsymbol{\mu}_c = \boldsymbol{\nu}_c + \boldsymbol{\Lambda}_c \boldsymbol{\kappa}_c \qquad \text{and} \qquad \Sigma_c = \boldsymbol{\Lambda}_c \boldsymbol{\Phi}_c \boldsymbol{\Lambda}_c' + \boldsymbol{\Psi}_c.$$

The model restrictions in (4.1) and (4.2) do not identify the parameters without additional assumptions. As per the traditional factor analysis approach, we impose a normal distribution on the latent skills and error terms:[8]

$$\boldsymbol{\theta}_{jc} \sim N(\boldsymbol{\kappa}_c, \boldsymbol{\Phi}_c) \qquad \text{and} \qquad \boldsymbol{u}_{jc} \sim N(\mathbf{0}, \boldsymbol{\Psi}_c). \tag{4.3}$$

Even with these assumptions, there are infinite equivalent parameterisations through which the model can be identified – the well-known issue of factor indeterminacy. We follow common practice and identify the model by setting the mean $\boldsymbol{\kappa}$ and variance $\boldsymbol{\Phi}$ of the latent skill factor in both cohorts to zero and one, respectively. Furthermore, we set intercepts to zero and error variances to one. Loadings λ and thresholds τ are instead allowed to vary across cohorts.

$$\text{diag}(\boldsymbol{\Phi}_c) = \boldsymbol{I}, \quad \boldsymbol{\kappa}_c = \mathbf{0}, \quad \boldsymbol{\nu}_c = \mathbf{0}, \quad \text{and } \text{diag}(\boldsymbol{\Psi}_c) = \boldsymbol{I} \qquad \forall c \in \{BCS, MCS\}. \tag{4.4}$$

The restrictions in (4.1), (4.2), (4.3), and (4.4) define the so-called *configural* model. This is a 'minimum' identifiable model, in that it places the least possible restrictions on how parameters are allowed to vary across cohorts. It serves as a basis for our measurement invariance analysis in the next section.[9]

5 Measurement invariance

Any comparison between socioemotional skills across the two cohorts requires that the measures at our disposal have the same relationship with the latent constructs of interest in both cohorts. In other words, the items in our new scale must measure externalising and internalising socioemotional skills in the same way in the BCS and MCS data. This property is denominated measurement invariance (Vandenberg and Lance, 2000; Putnick and Bornstein, 2016).

In the framework of factor analysis, measurement invariance is a formally testable property. In this paper, we follow the recent identification methodology by Wu and Estabrook (2016). The configural model defined in the previous section serves as the starting point. Measurement invariance is then assessed by comparing the configural

[7] The dedicated factor structure corresponds to a sparse loading matrix, i.e.:

$$\boldsymbol{\Lambda}_c := \begin{bmatrix} \lambda_{1c}, \ldots, \lambda_{6c} & \mathbf{0} \\ \mathbf{0} & \lambda_{7c}, \ldots, \lambda_{11c} \end{bmatrix}.$$

[8] Recent work has also used mixtures of normals for the latent factors distribution, e.g. Conti et al. (2010).

[9] This set of identifying restrictions is known as *Theta* parameterisation (Wu and Estabrook, 2016). See Appendix B for statistically equivalent alternative parameterisations.

model to a series of hierarchically nested models. These models place increasing restrictions on the item parameters, constraining them to be equal across groups. Their fit is then compared to that of the configural model. Intuitively, if the additional cross-group restrictions have not significantly worsened model fit, one can conclude that a certain level of invariance is achieved. The hierarchy of restrictions is detailed in Table A3.

Let's consider examples from our application. A *loading and threshold invariance* model restricts every item's loading λ and threshold τ parameters to have the same value in the two cohorts. It assumes that the items in our scale have the same relationship with latent skills across the two cohorts. In other words, items have the same salience, or informational content relative to skills. If this model fits as well as the configural model, we can be confident that the socioemotional skills of children in the two cohorts can be placed on the same scale, and their *variances* can be compared. To see why, consider equation (4.1). If the loading matrix $\mathbf{\Lambda}$ is the same across cohorts, any difference in latent skills $\Delta\boldsymbol{\theta}$ will correspond to the same difference in latent propensities $\Delta\boldsymbol{X}^*$. Equality of thresholds τ ensures that propensities $\boldsymbol{X}^*$ map into observed items $\boldsymbol{X}$ in the same way.

A *loading, threshold, and intercept invariance* model additionally restricts every item's intercept ν across cohorts. A good relative fit of this model indicates that socioemotional skills can be compared across cohorts in terms of their *means* as well. To see why, consider the following. Since the λ and ν parameters are the same across cohorts, a child in the BCS cohort with a given level of latent skills $\bar{\boldsymbol{\theta}}$ will have the same expected latent item propensities $\boldsymbol{X}^*$ as a child with the same skills in the MCS cohort. Again, equality of thresholds τ fixes the mapping between $\boldsymbol{X}^*$ and $\boldsymbol{X}$.[10]

We estimate the sequence of models detailed in Table A3 by Weighted Least Squares.[11] For the purposes of the analysis, we define groups c as cohort-gender cells, with the reference group being males in the BCS cohort. We then compare the fit of each model against the configural model.

Comparison of χ^2 values across models is a common likelihood-based strategy. However, tests based on $\Delta\chi^2$ are known to display high Type I error rates with large sample size and more complex models such as our own (Sass et al., 2014). In fact, for all invariance levels in our applications a chi-squared difference would point to a lack of measurement invariance. The use of approximate fit indices (AFIs) is therefore recommended alongside χ^2. These indices do not have a known sampling distribution, thus making it necessary to rely on rules of thumb to assess what level of ΔAFI indicates invariance. Nevertheless, AFIs are widely used in empirical practice to assess model fit.[12]

The fit of each model is compared in Panel A of Table A8. The model with restricted thresholds and loadings exhibits a comparable fit to the configural model, according to all the AFIs. Invariance of loadings and thresholds across cohorts implies that items in our scale are equally salient in their informational content, and that the latent propensities have equal mapping into the observed items. However, further restricting intercepts results in a model

[10]We recognise that simultaneous invariance of *all* items is not the minimum requirement for comparability. In theory, the availability of just one invariant item (known as 'anchor') would suffice to fix the scale and location of the system. However, partial invariance approaches are hard to implement in practice. Its validity hinges on selecting one (or more) truly invariant anchor, which is challenging on an a priori basis. The full procedure, restricting all parameters of a certain type across groups, does not identify which items are at the source of the invariance. Algorithms have been proposed to deal with this issue (Yoon and Millsap, 2007; Cheung and Lau, 2012), however there are still doubts on their robustness and their applicability to the categorical case (Vandenberg and Morelli, 2016).

[11]Parameters are estimated by mean- and variance-adjusted weighted least squares (WLSMV) – see Muthen et al. (1997); estimation starts from the items' polychoric correlation matrix, uses diagonally weighted least squares (DWLS), and exploits the full weight matrix to compute robust standard errors and test statistics. Robust WLS has proved in simulation studies to be moderately robust to small violations of the normality assumption in the latent underlying measures (Flora and Curran, 2004), and generally outperforms maximum likelihood in large samples (Beauducel and Herzberg, 2006; Li, 2016). All estimates are computed using the `lavaan` package (version 0.6-2) in **R** (Rosseel, 2012).

[12]The root mean squared error of approximation (RMSE) and the Tucker-Lewis index (TLI) are traditionally the most used AFIs in empirical practice. Simulation evidence by Cheung and Rensvold (2002) shows that these indices can show correlation between overall and relative fit, and suggest relying on additional indices, such as the comparative fit index (CFI, Bentler, 1990), McDonald non-centrality index (MFI, McDonald, 1989), and Gamma-hat index (Steiger, 1989) for the case of ordered measures. Commonly accepted thresholds for rejection are $\Delta CFI < -0.01$, $\Delta MFI < -0.02$, and $\Delta Gammahat < -0.001$. Meade et al. (2008), using the results from a simulation study, suggests stricter thresholds that should apply in a variety of conditions. For CFI, a single cutoff value of .002 is proposed, while cutoffs for MFI depend on the problem's characteristics; in our case (2 factors, 11 items), they suggest .0066. Sass et al. (2014) however cast some doubts of the generalisability of these cutoffs to WLSMV estimators.

where invariance is rejected across the board.[13] In other words, intercept parameters in our model (ν) are estimated to be different between maternal reports in the British and Millennium Cohort Studies. This means that, for a given level of latent skills, mothers in MCS tend to assess behaviours differently from mothers in BCS. Thus, cohort differences in scores on our scale cannot be unequivocally interpreted as differences in the underlying skills, since they might also reflect differences in reporting.

This is an important finding, which has to our knowledge never been acknowledged in this literature. How can this lack of comparability be explained? A possible interpretation is connected with secular evolution of social and cultural norms about child behaviours. For example, commonly held views of what constitutes a restless, distracted, or unhappy child might have changed between 1975 and 2005/6.[14]

To summarise, our measurement invariance analysis shows partial comparability of socioemotional skills across cohorts. In particular, the variance of skills can be compared across cohorts, but mean cohort differences do not necessarily reflect differences in skills. We can use scores from our scale to compare children within the same cohort, but not across cohorts. However, we can also compare within-cohort differences between groups of children, across cohorts. As an example, consider two groups of children A and B in the BCS cohort, and two groups of children C and D in the MCS. We cannot compare the mean level of skills between groups A and C, but we can compare the mean difference between groups A and B with the mean difference between groups C and D. This is the approach we take for the rest of the paper. Refraining from direct cross-cohort comparisons, we interpreting significance and magnitude of within-cohort differences across the cohorts.

6 Results

Parameter estimates from our factor model are presented in Table A9. As discussed in the previous section, loadings and thresholds are constrained to have the same value across groups. Intercepts are normalised to zero, and error variances to one, for the reference group – males in the BCS cohort. We use the estimates from this model to predict a score for each child in our sample along the latent externalising and internalising socio-emotional skill dimensions.[15] We plot the distribution of the scores in Figure 1. The unit of measurement is standard deviations of the distribution in the subsample of males in the BCS. Given our measurement invariance results in section 5, we stress that the *location* of these scores should not be directly compared across cohorts. However, the shape of the distribution can be given a cross-cohort interpretation.[16] It is immediately visible that there is more mass in the tails of the distribution in the 2000 than in the 1970 cohort.

[13] We do not present fit results for the threshold-only invariance model, as it is statistically equivalent to the configural model and thus its fit is mathematically the same – see Table 3 in Wu and Estabrook, 2016. The ages at which socio-emotional skills are observed varies slightly between BCS and MCS, due to different sampling and fieldwork schedules. In the MCS cohort, the age distribution has significantly higher variance. In Panel B of Table A8, we restrict the sample to 59 to 61 months, where the overlap between BCS and MCS is maximised. In Panels C and D, we restrict to male and female children respectively. In all these cases, invariance of thresholds and loadings is achieved, but invariance of intercepts is rejected. We can thus rule out that the lack of intercept invariance comes from differences in ages or invariance across child gender.

[14] Calibrating the Rutter and SDQ using a contemporary sample of children cannot rule out this issue. For example, Collishaw et al. (2004) administered both Rutter and SDQ items to parents of a small sample of adolescents in London. They use the mapping between the two questionnaires to impute Rutter scores for mothers who answered the SDQ. This can correct for contemporaneous reporting differences between questionnaires, but cannot tackle reporting differences between samples collected at different times in history.

[15] We use an empirical Bayes modal (EBM) approach to estimate the scores. The parameters are estimated using three sources of information. The first is the distribution of the latent variables $\boldsymbol{\theta}$, treated as random parameters with a prior $h(\boldsymbol{\theta}, \boldsymbol{\Omega})$, conditional on the parameters $\boldsymbol{\Omega}$. This prior is assumed to be multivariate normal. The second is the observed data $\boldsymbol{X}$, and the third is the estimated parameters $\hat{\boldsymbol{\Omega}}$. Data and prior are combined into the posterior distribution $w(\boldsymbol{\theta}|\boldsymbol{X}, \hat{\boldsymbol{\Omega}})$. For further details, see Chapter 7 in Skrondal and Rabe-Hesketh (2004).

[16] The density of the scored factors can be contrasted with the distribution of sum scores in Figure A1. Using raw scores, instead, shows an increase in mass only at the top of the distribution.

6.1 Inequality in socioemotional skills

We find that, both unconditionally and for specific groups, inequality in socio-emotional skills at age five has increased between 1975 and 2005/6. Table 2 shows unconditional inequality statistics, using quantile differences in the distribution of skills by gender and cohort. With the exception of internalising skills in female children, all distributions have widened substantially between the BCS and MCS cohorts. The gap for both externalising and internalising skills between the 90th and the 10th percentiles for males has increased by approximately half a standard deviation. The increase in the gap is more pronounced in the bottom half of the distribution. For females, we see a narrowing at the top (90-50), but a widening at the bottom (50-10) of the distribution, again for both externalising and internalising skills.

Inequality has also increased conditional on socioeconomic status. Figure 2 shows mean skills by maternal education. We compare mothers who continued education past the compulsory age with mothers who left school at the compulsory leaving age, according to their year of birth. Given lack of comparability in the level of skills across cohort, we normalise the mean in the 'Compulsory' group to zero for both cohorts. For both males and females, and for both externalising and internalising skills, the difference in the socio-emotional skills of their children between more and less educated mothers has increased. The size of the increase is around .1 to .15 of a standard deviation. The increase is particularly pronounced for males, for whom it goes from .20 to .30 for externalising and from .12 to .24 for internalising.

Figure 3 shows an even starker pattern when comparing children of mothers who smoked in pregnancy with non-smoking mothers. The fact that maternal smoking during pregnancy is a risk factor for offspring behavioural problems is well known in the medical literature (Gaysina et al., 2013); less evidence there is however, on whether and to which extent these associations have changed across cohorts. The difference in child skills has increased, from less than .2 to around .4 of a standard deviation, again with the biggest increase experienced by the boys. There is also a significant increase in the gradient by paternal occupation based on social class (Figure 4), although this is less pronounced if compared to the one based on maternal characteristics. In particular, male children with no father figure living in their household have worse skills compared to children with blue collar fathers in the MCS cohort. Otherwise, skill differences in father's occupation are mostly constant across the two cohorts.[17] These patterns are in stark contrast with the findings of Reardon and Portilla (2016) for the US, who have found a narrowing of the readiness gaps from 1998 to 2010.

We then examine the same patterns as in the previous figures, but conditional on other family background indicators. The aim is to disentangle the relative contribution of each indicator to socio-emotional skills, and how it has changed in the thirty years between the two cohorts. Table 3 shows coefficients from linear regressions of socio-emotional skills at five on contemporaneous and past socioeconomic indicators, by cohort and gender. Coefficients for indicators in BCS and MCS are presented side by side, together with the p-value of the hypothesis that coefficients are the same in the two cohorts.[18]

Overall, the importance of maternal socioeconomic status (education and in particular employment) in determining socio-emotional skills has increased from the BCS to the MCS children. The 'premium' in skills for children of better educated and employed mothers is significantly larger, for both boys and girls, internalising and externalising skills. At the same time, the penalty for having a blue-collar father, or not having a father figure at all in the household, has significantly declined across the two cohorts, especially for girls. Being born to an unmarried mother, and to a mother

[17] Figures A2, A3, and A4 show inequality in the scale items underlying the factor scores used in this section. The increase in inequality across cohorts is still present, but less marked when looking at these single items. This shows the importance of the factor analysis step in aggregating items, explicitly modelling the measurement error, and testing and accounting for (loadings and thresholds) invariance across the two cohorts.

[18] We also estimated Tobit models to account for the right truncation of the distribution of skills – see Figure 1. Tobit estimates are extremely similar to the linear estimates in Table 3, and are available from the authors upon request.

who smoked during pregnancy, is associated with a higher penalty for both dimensions of socio-emotional skills in the latter cohort, but only for males. This is consistent with recent evidence which shows that family disadvantage disproportionately impedes the pre-market development of boys, in terms of higher disciplinary problems, lower achievement scores, and fewer high-school completions (Autor et al., 2016). Girls of non-white ethnicity, instead, have worse internalising and externalising skills in the MCS, a penalty not suffered by 5-year old non-white girls in the BCS. Firstborn boys and girls in the BCS have worse skills, but this difference disappears in the MCS. Lastly, we document an increase in the returns to birth weight, which is more pronounced for boys' internalising skills.

These changes in the relative importance of pregnancy factors and family background characteristics for child socio-emotional skills at age 5 need to be interpreted in the light of the significant changes in the prevalence of such characteristics across cohorts. As shown in Table A5, the age of the mother at birth, proportion of mothers non-smoking in pregnancy, with post-compulsory education and in employment at the age 5 of the child has substantially increased; at the same time, the proportion of households with no father figure has increased, and so the proportion of women unmarried at birth is much higher in the 2000 than in the 1970 cohort. Also, the ethnic structure of the population has changed, with a higher proportion of non-white children in the MCS than in the BCS. In general, this has been a period of significant societal changes, with an almost continual rise in the proportion of women in employment, an older age at first birth and a rise in dual-earning parents families (Roantree and Vira, 2018). However, here we do not attempt to disentangle whether and to which extent the observed changes in inequality in socio-emotional skills across the two cohorts can be attributed to changes in returns (or penalties) to maternal characteristics (such as education and employment) or to compositional changes, like it has been done for the analysis of wage inequality (Blundell et al., 2007).

6.2 Socio-emotional skills and adolescent/adult outcomes

In this last section, we study the predictive power of socio-emotional skills for adolescent and adult outcomes. We contribute to a vast interdisciplinary literature by examining medium- and long-term impacts of skills measured at an earlier age than usually examined in previous studies, well before the start of formal education. We do so by regressing health and socioeconomic outcomes measured in adolescence and adulthood on the socio-emotional skills scores at age five obtained by our factor model, controlling for the harmonised family background variables at birth and age five (see Table A2). We present results with and without controlling for cognitive skills. As detailed in Section 3, the available cognitive measures are not comparable across cohorts. Still, we control for a factor score that summarises all information on cognitive skills that is available in each cohort, regardless of their comparability.

Socio-emotional skills at five years of age are predictive of adolescent health behaviour and outcomes in both cohorts.[19] Table 4 examines adolescent smoking and BMI for both cohorts; Table A10 reports the results for the same outcomes in adulthood (at age 42), for the BCS only. Externalising skills are negatively correlated to subsequent smoking and BMI in both cohorts, for both genders. Recall that a child with high externalising skills exhibits less restless and hyperactive behaviours, and has less anti-social conduct. Our findings are consistent with the body of evidence reviewed in section 2, which shows that better socio-emotional skills (measured using different scales and at various points during childhood and adolescence) are negatively associated with smoking. At the same time, internalising skills are positively correlated with smoking (only in the 1970 cohort) and BMI (only for girls), although less strongly than externalising skills. This apparently counterintuitive result makes sense in light of the items in our internalising scale shown in Table 1. A child with better internalising skills is less solitary, neurotic, and worried. From this perspective, he is likely more sociable and subject to peer influence in health behaviours. This is consistent

[19] Unfortunately the strength of the association cannot be directly compared, since the outcomes are measured at different ages: 16 and 14 years for BCS and MCS, respectively.

with the evidence in Goodman et al. (2015), who find a positive association between child emotional health (measured with items for the internalizing behaviour subscale of the Rutter scale at age 10 in the BCS) and smoking at age 42. Furthermore, in recent work Hsieh and van Kippersluis (2018) have shown personality to be a key mechanism through which peers affect smoking behaviour.

Conditional on socio-emotional skills, cognition has limited predictive power for these behaviours, and only for girls.[20] This is in line with the evidence in Conti and Heckman (2010), who show that not accounting for non-cognitive traits (a self-regulation factor measured at age 10) overestimates the importance of cognition for predicting health and health behaviours, using data from the British cohort study. Conti and Hansman (2013) use rich data on child personality and socio-emotional traits collected at ages 7, 11 and 16 in the 1958 British birth cohort,[21] and show that these traits rival the importance of cognition in explaining the education gradient in health behaviours (including smoking and BMI). We show that child socio-emotional skills have greater predictive power for health outcomes and behaviours even when measured at an earlier age.

For the British Cohort Study, we last examine the association between socio-emotional skills at age five and adult education and labour market outcomes. The structure of Table 5 is similar to Table 4, but it considers educational achievement, employment, and earnings (conditional on being in paid employment) for the BCS cohort members. For these outcomes, the predictive power of cognitive skills outweighs that of socio-emotional skills, whose predictive power diminishes over time (between the ages 34 and 42), and is driven to insignificance after controlling for cognition. This is consistent with the evidence in Conti et al. (2011), who show that cognitive endowments at age 10 are more predictive (than socio-emotional and health ones) for employment and wage outcomes in the BCS. Again, we show that the greater predictive power of cognition for socioeconomic outcomes holds even when considering earlier-life measures of child development.

7 Conclusion

In this paper we have studied inequality in a dimension of human capital which has received limited attention in the literature so far: socio-emotional skills very early in life. In particular, we have focused on the measurements of these skills at age 5 in two British cohorts born 30 years apart: the one of children born in 1970 (British Cohort Study, BCS) and the one of children born in 2000/1 (Millennium Cohort Study, MCS). We have provided several contributions to the recent but flourishing literature on the determinants and consequences of early human development.

We have taken very seriously the issue of comparability of measurements of socio-emotional skills across cohorts. First, we have selected 11 comparable items across two related scales: the Rutter scale in the BCS, and the Strength and Difficulties Questionnaire (SDQ) in the MCS. After examining the latent structure underlying the items, we have identified by means of exploratory factor analysis two dimensions of socio-emotional skills. We have labeled them 'internalising' and 'externalising' skills, the former related to the ability of children to focus their concentration and the latter to engage in interpersonal activities.

Second, we have formally tested for measurement invariance of the 11 items across the two externalising and internalising scales, following recent methodological advances in factor analysis with categorical outcomes. We have found only partial support for measurement invariance, with the implication that we have only been able to compare how inequality in these socio-emotional skills across the two cohorts has changed, but not whether their average level is larger or smaller in one of the two cohorts. These results sound a warning to research in this area which routinely

[20] We do not observe significant associations between early socio-emotional skills and other risky behaviours like drug-taking and alcohol consumption. One possible reason might be the relatively young age at which these skills are measured. Results are available upon request.

[21] They use the Rutter scale and the Bristol Social Adjustment Guide.

compares levels of skills across different groups (at different times, or of different gender), without first establishing their comparability.

Third, after having computed comparable scores for both externalising and internalising skills, and for both boys and girls, we have compared how inequality in these skills has changed across the 1970 and the 2000 cohort. We have documented for the first time that inequality in these early skills has increased across cohorts, especially for boys. The cross-cohort increase in the gap is more pronounced at the bottom of the distribution (50-10 percentiles). We have also documented changes in conditional skills gaps across cohorts. In particular, the difference in the socio-emotional skills of their children between mothers of higher and lower socio-economic status (education and employment) has increased. The increase in cross-cohort inequality is even starker when comparing children born to mothers who smoked during pregnancy. In both cases, the increase in inequality is particularly pronounced for boys. On the other hand, the skills penalty arising from the lack of a father figure in the household has substantially declined, especially for girls.

Fourth, we have contributed to the literature on the predictive power of socio-emotional skills by showing that even skills measured at a much earlier age than in previous work are significantly associated with outcomes both in adolescence and adulthood. In particular, socio-emotional skills are more significant predictors of health and health behaviours (smoking and BMI), while cognition has greater predictive power for socioeconomic outcomes (education, employment and wages). Our results show the importance of inequalities in the early years development for the accumulation of health and human capital across the life course.

Table 1: Subscale of comparable items

Itm.	Factor	Cat.	Title	Rutter Wording (BCS 1970)	SDQ Wording (MCS 2000/1)
1	EXT	3	*Restless*	Very restless. Often running about or jumping up and down. Hardly ever still	Restless, overactive, cannot stay still for long
2	EXT	3	*Squirmy/fidgety*	Is squirmy or fidgety	Constantly fidgeting or squirming
3	EXT	3	*Fights/bullies*	Frequently fights other children + Bullies other children	Often fights with other children or bullies them
4	EXT	3	*Distracted*	Cannot settle to anything for more than a few moments	Easily distracted, concentration wanders
5	EXT	2	*Tantrums*	Has temper tantrums	Often has temper tantrums or hot tempers
6	EXT	2	*Disobedient*	Is often disobedient	(+) Generally obedient, usually does what adults request
7	INT	3	*Worried*	Often worried, worries about many things	Many worries, often seems worried
8	INT	3	*Fearful*	Tends to be fearful or afraid of new things or new situations	Nervous or clingy in new situations, easily loses confidence
9	INT	3	*Solitary*	Tends to do things on his/her own, rather solitary	Rather solitary, tends to play alone
10	INT	3	*Unhappy*	Often appears miserable, unhappy, tearful or distressed	Often unhappy, down-hearted or tearful
11	INT	2	*Aches*	Complains of headaches + Complains of stomach-ache or has vomited	Often complains of head- aches, stomach-ache or sickness

Notes: *Itm.* is item number. *Factor* is the latent construct to which the item loads – EXT is Externalising skills, INT is Internalising skills. *Cat.* is the number of categories in which the item is coded – 2 denotes a binary item (applies/does not apply) and 3 denotes a 3-category item. *Title* is a short label for the item. *Wording* columns show the actual wording in the scales used in each of the cohort studies. Items denoted by (+) are positively coded in the original scale.

Table 2: Quantile differences in scores

	Males		Females	
	BCS (1970)	MCS (2000/1)	BCS (1970)	MCS (2000/1)
Externalising				
90-10	2.08	2.47	2.09	2.22
75-25	1.08	1.36	1.12	1.23
90-50	1.00	1.14	1.02	0.95
50-10	1.08	1.33	1.07	1.27
Internalising				
90-10	1.71	2.28	1.86	1.86
75-25	0.91	1.12	1.01	0.92
90-50	0.73	0.89	0.82	0.72
50-10	0.97	1.38	1.04	1.14

Notes: The table shows differences between quantiles of the distribution of socioemotional skills, by gender and cohort. The distribution is a factor score obtained from the factor model in Section 4.

Table 3: Determinants of Socioemotional Skills

	Externalising						Internalising					
	Males			Females			Males			Females		
	(1) BCS	(2) MCS	(3) p-value	(4) BCS	(5) MCS	(6) p-value	(7) BCS	(8) MCS	(9) p-value	(10) BCS	(11) MCS	(12) p-value
Maternal education (5)												
Post-compulsory	0.089***	0.141***	[0.246]	0.099***	0.147***	[0.270]	0.072***	0.133***	[0.127]	0.048**	0.075**	[0.491]
	(0.026)	(0.041)		(0.027)	(0.037)		(0.022)	(0.039)		(0.025)	(0.032)	
Maternal employment (5)												
Employed	0.018	0.111***	[0.030]	−0.009	0.076**	[0.041]	0.041**	0.148***	[0.005]	0.031	0.127***	[0.011]
	(0.024)	(0.040)		(0.025)	(0.037)		(0.020)	(0.037)		(0.021)	(0.031)	
Father occ. (5) - White collar = 0												
Blue collar	−0.195***	−0.112***	[0.074]	−0.117***	−0.072**	[0.312]	−0.076***	−0.053	[0.588]	−0.101***	−0.021	[0.047]
	(0.027)	(0.043)		(0.027)	(0.039)		(0.023)	(0.040)		(0.025)	(0.034)	
No father figure	−0.280***	−0.225***	[0.512]	−0.380***	−0.179***	[0.010]	−0.200***	−0.193***	[0.926]	−0.244***	−0.123**	[0.083]
	(0.063)	(0.057)		(0.059)	(0.053)		(0.053)	(0.057)		(0.054)	(0.048)	
Maternal background (0)												
Age	0.014***	0.012***	[0.711]	0.014***	0.019***	[0.290]	0.008***	0.011***	[0.430]	0.009***	0.011***	[0.474]
	(0.002)	(0.004)		(0.003)	(0.003)		(0.002)	(0.003)		(0.002)	(0.003)	
Unmarried	0.065	−0.107**	[0.017]	0.025	−0.066*	[0.183]	0.115**	−0.029	[0.025]	0.036	−0.040	[0.221]
	(0.059)	(0.044)		(0.060)	(0.041)		(0.050)	(0.042)		(0.055)	(0.035)	
Nonwhite child	−0.161**	−0.110*	[0.602]	−0.029	−0.131**	[0.238]	0.025	−0.050	[0.391]	0.081	−0.140**	[0.005]
	(0.076)	(0.063)		(0.067)	(0.054)		(0.063)	(0.057)		(0.062)	(0.049)	
Pregnancy												
Firstborn	−0.121***	0.006	[0.009]	−0.070**	0.036	[0.025]	−0.186***	−0.085**	[0.020]	−0.161***	−0.032	[0.002]
	(0.029)	(0.045)		(0.031)	(0.041)		(0.024)	(0.042)		(0.028)	(0.037)	
Mother smoked in pregnancy	−0.144***	−0.229***	[0.092]	−0.110***	−0.156***	[0.354]	−0.077***	−0.176***	[0.028]	−0.036	−0.060	[0.592]
	(0.026)	(0.051)		(0.025)	(0.048)		(0.021)	(0.048)		(0.023)	(0.041)	
(log) Birthweight	0.146**	0.311***	[0.190]	0.186**	0.362***	[0.145]	0.095	0.387***	[0.009]	0.123*	0.078	[0.674]
	(0.073)	(0.113)		(0.078)	(0.102)		(0.061)	(0.109)		(0.070)	(0.088)	
Adj. R^2	0.062	0.081		0.056	0.090		0.042	0.068		0.042	0.052	
Num. obs.	4565	2759		4313	2620		4565	2759		4313	2620	

Notes: The table shows coefficients from linear regressions of children's socioemotional skills at five years of age on family background information. The dependent variable is a factor score obtained from the factor model in Section 4. Col. (1) and (2) show coefficients and standard errors in parentheses, for male children in the BCS and MCS cohorts separately. The latter are obtained using 1,000 bootstrap repetitions, taking into account the factor estimation stage that precedes the regression. Col. (3) shows the p-value of a test that the coefficient is the same in the two cohorts. Col. (4) to (6) repeat for female children. Col. (7) to (12) repeat for internalising skills. All estimates additionally control for region of birth, mother height, number of previous stillbirths at child's birth, preterm birth, a dummy for missing gestational age, and number of other children in the household at child age 5. See Table A2 for a description of the variables used.

Table 4: Predictors of adolescent outcomes

	Males			Females		
	Mean	Coefficients		Mean	Coefficients	
	(1)	(2)	(3)	(4)	(5)	(6)
Tried smoking (BCS - 16)	.524			.586		
Externalising skills (5)		−.073***	−.081***		−.068***	−.081***
		(.023)	(.023)		(.021)	(.023)
Internalising skills (5)		.055**	.060**		.039*	.045*
		(.027)	(.028)		(.023)	(.024)
Cognitive skills (5)			.010			.012
			(.019)			(.017)
Adj. R^2		0.032	0.032		0.048	0.046
Observations		1197	1123		1693	1581
BMI (BCS - 16)	20.9			21.2		
Externalising skills (5)		−.178	−.227*		−.225*	−.227*
		(.119)	(.124)		(.124)	(.124)
Internalising skills (5)		.036	.062		.280**	.234*
		(.141)	(.148)		(.138)	(.142)
Cognitive skills (5)			.021			−.093
			(.110)			(.104)
Adj. R^2		0.017	0.018		0.023	0.023
Observations		1640	1531		1873	1757
Tried smoking (MCS - 14)	.125			.151		
Externalising skills (5)		−.043***	−.041***		−.031**	−.041***
		(.012)	(.012)		(.012)	(.012)
Internalising skills (5)		.017	.018		.009	.012
		(.012)	(.012)		(.014)	(.014)
Cognitive skills (5)			.000			−.031**
			(.010)			(.012)
Adj. R^2		0.056	0.054		0.050	0.051
Observations		1959	1936		1986	1982
BMI (MCS - 14)	20.7			21.7		
Externalising skills (5)		−.327**	−.311**		−.405***	−.311**
		(.138)	(.139)		(.141)	(.139)
Internalising skills (5)		.064	.091		.354**	.366**
		(.147)	(.149)		(.157)	(.157)
Cognitive skills (5)			.014			−.186
			(.119)			(.152)
Adj. R^2		0.025	0.024		0.045	0.045
Observations		1965	1936		1893	1886

Notes: The table shows coefficients from linear regressions of cohort members' adolescent outcomes on their externalising and internalising socioemotional skills at five years of age. Col. (1) shows the mean of the outcome for males. Col. (2) regresses the outcome on the scores obtained from the factor model in Section 4. Col. (3) additionally controls for cognitive ability at age five. This is a simple factor score obtained by aggregating the available cognitive measures. All standard errors in parentheses are obtained using 1,000 bootstrap repetitions, taking into account the factor estimation stage that precedes the regression. Col. (4) to (6) repeat for female cohort members. All estimates additionally control for region of birth, maternal education (5), maternal employment (5), father occupation (5), maternal background (age, height, nonwhite ethnicity, number of children in HH), pregnancy (firstborn child, number of previous stillbirths, mother smoked in pregnancy, preterm birth, (log) birth weight). See Table A2 for a description of the variables used.

Table 5: Predictors of adult outcomes – BCS

	Males			Females		
	Mean	Coefficients		Mean	Coefficients	
	(1)	(2)	(3)	(4)	(5)	(6)
Higher education (34)	.430			.426		
Externalising skills (5)		.043**	.024		.068***	.024
		(.021)	(.022)		(.021)	(.022)
Internalising skills (5)		−.032	−.026		−.017	−.028
		(.027)	(.027)		(.023)	(.024)
Cognitive skills (5)			.089***			.113***
			(.017)			(.017)
Adj. R^2		0.083	0.099		0.101	0.120
Observations		1320	1237		1691	1589
Employed (42)	.932			.828		
Externalising skills (5)		.012	.010		.014	.010
		(.011)	(.011)		(.016)	(.011)
Internalising skills (5)		.022*	.020		.024	.017
		(.014)	(.014)		(.018)	(.019)
Cognitive skills (5)			.023**			.037***
			(.010)			(.014)
Adj. R^2		0.056	0.052		0.010	0.014
Observations		1294	1216		1677	1571
(log) Gross weekly pay (42)	6.474			5.775		
Externalising skills (5)		.047	.047		.009	.047
		(.037)	(.036)		(.042)	(.036)
Internalising skills (5)		−.044	−.082*		.051	.041
		(.044)	(.043)		(.046)	(.047)
Cognitive skills (5)			.064**			.137***
			(.029)			(.033)
Adj. R^2		0.057	0.068		0.046	0.061
Observations		918	865		1198	1122

Notes: The table shows coefficients from linear regressions of BCS cohort members' adult outcomes on their externalising and internalising socioemotional skills at five years of age. Col. (1) shows the mean of the outcome for males. Col. (2) regresses the outcome on the scores obtained from the factor model in Section 4. Col. (3) additionally controls for cognitive ability at age five. This is a simple factor score obtained by aggregating the available cognitive measures. All standard errors in parentheses are obtained using 1,000 bootstrap repetitions, taking into account the factor estimation stage that precedes the regression. Col. (4) to (6) repeat for female cohort members. All estimates additionally control for region of birth, maternal education (5), maternal employment (5), father occupation (5), maternal background (age, height, nonwhite ethnicity, number of children in HH), pregnancy (firstborn child, number of previous stillbirths, mother smoked in pregnancy, preterm birth, (log) birth weight). See Table A2 for a description of the variables used.

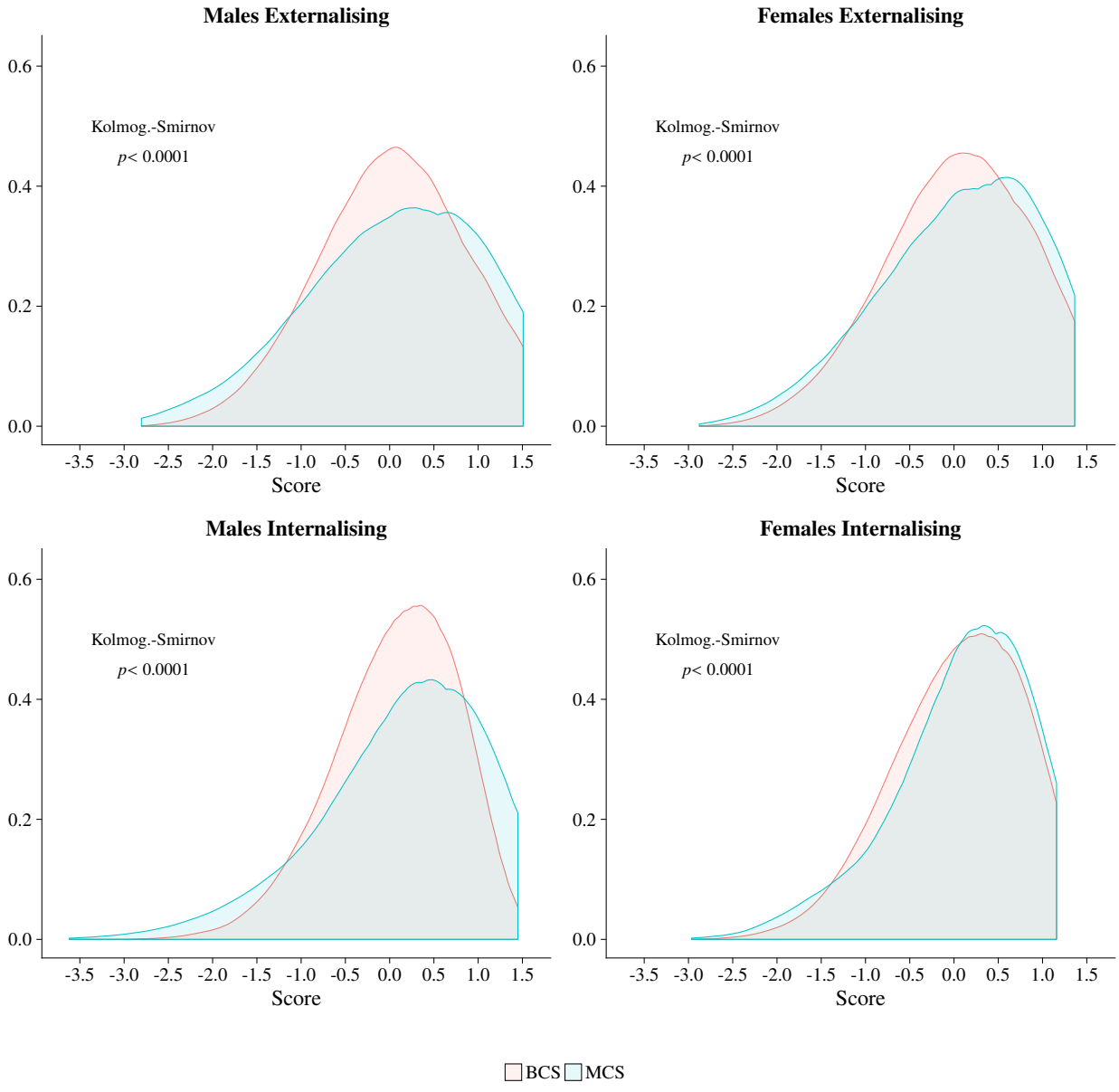

Figure 1: Distribution of factor scores

Notes: The figure shows the distribution of the externalising and internalising socioemotional skills scores at age five obtained from the factor model, by gender and cohort. The scores are estimated from parameter estimates in Table A9, using an Empirical Bayes Modal approach. Higher scores correspond to *better* skills. The distribution is estimated nonparametrically, using an Epanechnikov kernel. The figure also reports the p value from Kolmogorov-Smirnov tests of equality between the distribution in BCS and MCS.

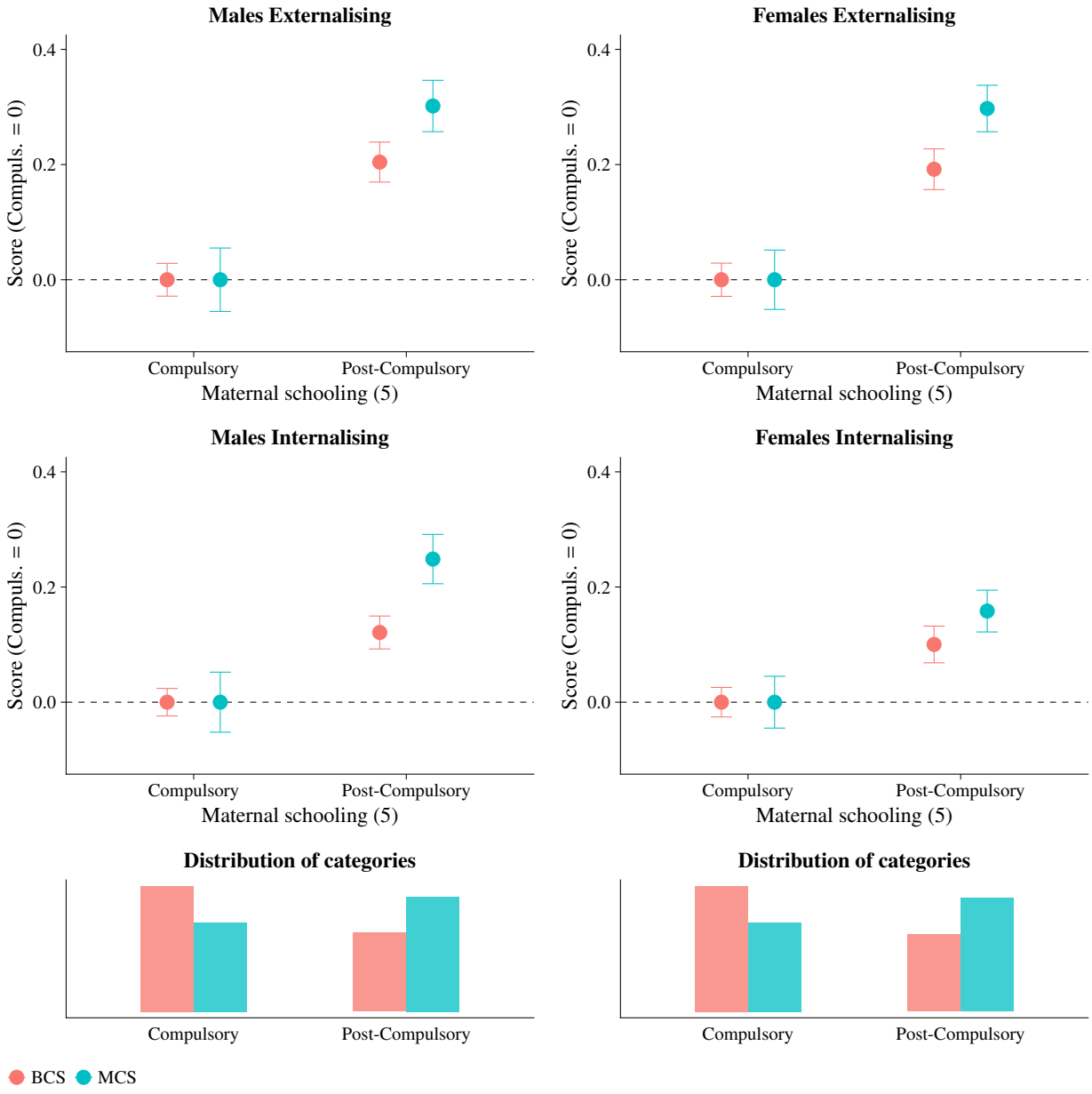

Figure 2: Skill inequality by mother's education

Notes: The figure shows unconditional mean values of socioemotional skills scores by gender, cohort, and mother's education at age five. Mother's education is a dummy for whether the mother continued schooling past the minimum leaving age, based on her date of birth. The four panels on top present mean and 95% confidence intervals. Given that we cannot compare means of skills, all scores are normalised to take value zero for the 'Compulsory' category, so that the gradient is emphasised. The bottom two panels present the unconditional distribution of mother's education.

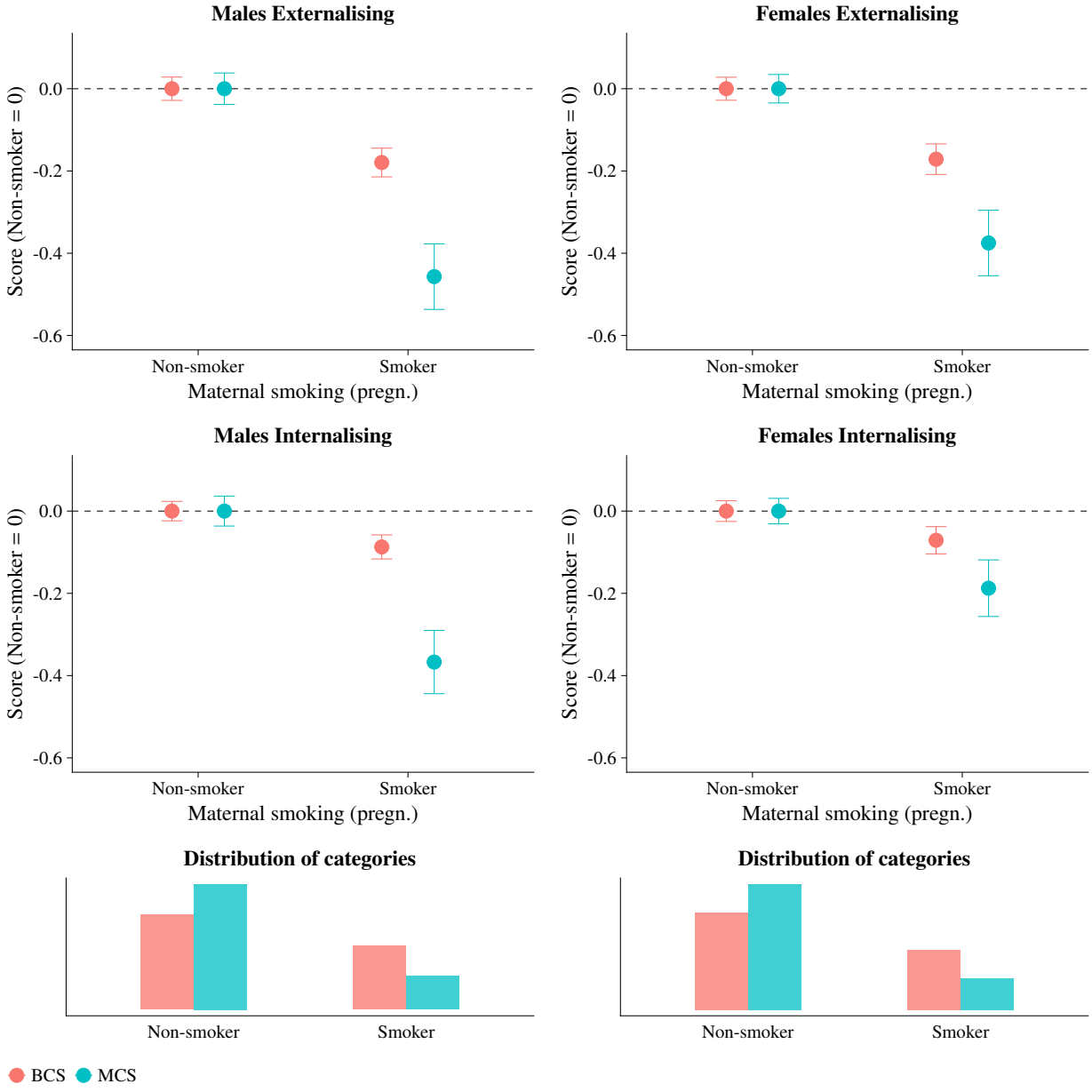

Figure 3: Skill inequality by mother's pregnancy smoking

Notes: The figure shows unconditional mean values of socioemotional skills scores by gender, cohort, and mother's pregnancy smoking. Mother's education is a dummy for whether the mother reported smoking during pregnancy. The four panels on top present mean and 95% confidence intervals. Given that we cannot compare means of skills, all scores are normalised to take value zero for the 'Non-smoker' category, so that the gradient is emphasised. The bottom two panels present the unconditional distribution of mother's pregnancy smoking

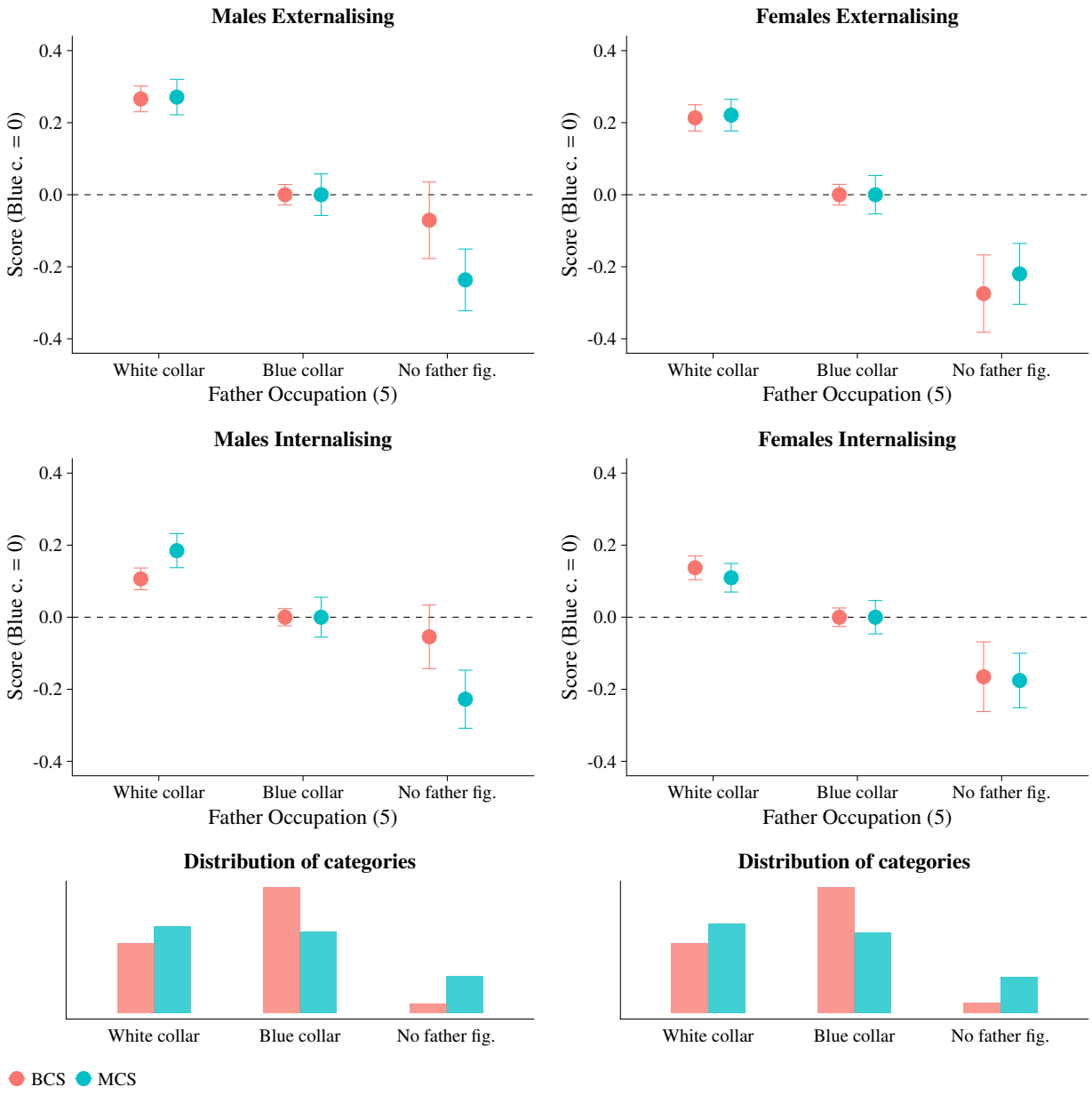

Figure 4: Skill inequality by father's occupation

Notes: The figure shows unconditional mean values of socioemotional skills scores by gender, cohort, and father's occupation at age five. Father's occupation is based on Registrar General's social class, with classes I to III Non Manual being 'White collar' and classes III Manual to V (plus 'other') being 'Blue collar'. 'No father figure' is defined as absence of a male figure living in the household. The four panels on top present mean and 95% confidence intervals. Given that we cannot compare means of skills, all scores are normalised to take value zero for the 'Blue collar' category, so that the gradient is emphasised. The bottom two panels present the unconditional distribution of father's occupation.

8 Bibliography

Almlund, M., A. L. Duckworth, J. Heckman, and T. Kautz (2011). Personality Psychology and Economics. In *Handbook of the Economics of Education*, Volume 4, pp. 1–181. Elsevier.

Attanasio, O., S. Cattan, E. Fitzsimons, C. Meghir, and M. Rubio-Codina (2018). Estimating the Production Function for Human Capital: Results from a Randomized Controlled Trial in Colombia. Technical Report 1987R2, Cowles Foundation.

Autor, D., D. Figlio, K. Karbownik, J. Roth, and M. Wasserman (2016, May). Family Disadvantage and the Gender Gap in Behavioral and Educational Outcomes. Technical Report w22267, National Bureau of Economic Research, Cambridge, MA.

Beauducel, A. and P. Y. Herzberg (2006, April). On the Performance of Maximum Likelihood Versus Means and Variance Adjusted Weighted Least Squares Estimation in CFA. *Structural Equation Modeling: A Multidisciplinary Journal 13*(2), 186–203.

Behar, L. and S. Stringfield (1974). A behavior rating scale for the preschool child. *Developmental Psychology 10*(5), 601–610.

Belfield, C., A. B. Bowden, A. Klapp, H. Levin, R. Shand, and S. Zander (2015). The Economic Value of Social and Emotional Learning. *Journal of Benefit-Cost Analysis 6*(03), 508–544.

Bentler, P. M. (1990). Comparative Fit Indexes in Structural Models. *Psychological Bulletin 107*(2), 238–46.

Berglund, L. (1999, December). Latent Variable Analysis of the Rutter Children's Behaviour Questionnaire. *Scandinavian Journal of Educational Research 43*(4), 433–442.

Blanden, J., P. Gregg, and L. Macmillan (2007). Accounting for intergenerational income persistence: Noncognitive skills, ability and education. *The Economic Journal 117*(519).

Blundell, R., A. Gosling, H. Ichimura, and C. Meghir (2007, March). Changes in the Distribution of Male and Female Wages Accounting for Employment Composition Using Bounds. *Econometrica 75*(2), 323–363.

Borghans, L., A. L. Duckworth, J. J. Heckman, and B. ter Weel (2008). The Economics and Psychology of Personality Traits. *Journal of Human Resources 43*(4), 972–1059.

Bowles, S., H. Gintis, and M. Osborne (2001, December). The Determinants of Earnings: A Behavioral Approach. *Journal of Economic Literature 39*(4), 1137–1176.

Butler, N. B. (2016a). 1970 British Cohort Study: Five-Year Follow-Up, 1975.

Butler, N. B. (2016b). 1970 British Cohort Study: Ten-Year Follow-Up, 1980.

Butler, N. B. (2017). 1970 British Cohort Study: Sixteen-Year Follow-Up, 1986.

Carneiro, P., C. Meghir, and M. Parey (2013, January). Maternal education, home environments, and the development of children and adolescents. *Journal of the European Economic Association 11*, 123–160.

Chamberlain, R. C. (2013). 1970 British Cohort Study: Birth and 22-Month Subsample, 1970-1972.

Chetty, R., J. N. Friedman, N. Hilger, E. Saez, D. W. Schanzenbach, and D. Yagan (2011, November). How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project Star. *The Quarterly Journal of Economics 126*(4), 1593–1660.

Cheung, G. W. and R. S. Lau (2012, April). A Direct Comparison Approach for Testing Measurement Invariance. *Organizational Research Methods 15*(2), 167–198.

Cheung, G. W. and R. B. Rensvold (2002, April). Evaluating Goodness-of-Fit Indexes for Testing Measurement Invariance. *Structural Equation Modeling: A Multidisciplinary Journal 9*(2), 233–255.

Chiteji, N. (2010, May). Time Preference, Noncognitive Skills and Well Being across the Life Course: Do Noncognitive Skills Encourage Healthy Behavior? *American Economic Review 100*(2), 200–204.

Cobb-Clark, D. A., S. C. Kassenboehmer, and S. Schurer (2014, February). Healthy habits: The connection between diet, exercise, and locus of control. *Journal of Economic Behavior & Organization 98*, 1–28.

Collishaw, S., B. Maughan, R. Goodman, and A. Pickles (2004, November). Time trends in adolescent mental health. *Journal of Child Psychology and Psychiatry 45*(8), 1350–1362.

Conti, G., S. Frühwirth-Schnatter, J. J. Heckman, and R. Piatek (2014, November). Bayesian exploratory factor analysis. *Journal of Econometrics 183*(1), 31–57.

Conti, G. and C. Hansman (2013, March). Personality and the education–health gradient: A note on "Understanding differences in health behaviors by education". *Journal of Health Economics 32*(2), 480–485.

Conti, G., J. Heckman, and S. Urzua (2010, May). The Education-Health Gradient. *American Economic Review 100*(2), 234–238.

Conti, G. and J. J. Heckman (2010, September). Understanding the Early Origins of the Education–Health Gradient: A Framework That Can Also Be Applied to Analyze Gene–Environment Interactions. *Perspectives on Psychological Science 5*(5), 585–605.

Conti, G., J. J. Heckman, and R. Pinto (2016). The Effects of Two Influential Early Childhood Interventions on Health and Healthy Behaviour. *The Economic Journal 126*(596), F28–F65.

Conti, G., J. J. Heckman, and S. Urzua (2011). Early endowments, education, and health. Technical Report 2011-01, Human Capital and Economic Opportunity Global Working Group.

Cunha, F., J. J. Heckman, L. Lochner, and D. V. Masterov (2006). Interpreting the Evidence on Life Cycle Skill Formation. In *Handbook of the Economics of Education*, Volume 1, pp. 697–812. Elsevier.

Cunha, F., J. J. Heckman, and S. Schennach (2010). Estimating the Technology of Cognitive and Noncognitive Skill Formation. *Econometrica 78*(3), 883–931.

Deming, D. J. (2017, November). The Growing Importance of Social Skills in the Labor Market*. *The Quarterly Journal of Economics 132*(4), 1593–1640.

Dickey, W. C. and S. J. Blumberg (2004, September). Revisiting the Factor Structure of the Strengths and Difficulties Questionnaire: United States, 2001. *Journal of the American Academy of Child & Adolescent Psychiatry 43*(9), 1159–1167.

Durlak, J. A., R. P. Weissberg, A. B. Dymnicki, R. D. Taylor, and K. B. Schellinger (2011, January). The Impact of Enhancing Students' Social and Emotional Learning: A Meta-Analysis of School-Based Universal Interventions: Social and Emotional Learning. *Child Development 82*(1), 405–432.

Durlak, J. A., R. P. Weissberg, and M. Pachan (2010, June). A Meta-Analysis of After-School Programs That Seek to Promote Personal and Social Skills in Children and Adolescents. *American Journal of Community Psychology 45*(3-4), 294–309.

Flora, D. B. and P. J. Curran (2004). An Empirical Evaluation of Alternative Methods of Estimation for Confirmatory Factor Analysis With Ordinal Data. *Psychological Methods 9*(4), 466–491.

Fowler, P. C. and R. M. Park (1979, October). Factor Structure of the Preschool Behavior Questionnaire in a Normal Population. *Psychological Reports 45*(2), 599–606.

Gaysina, D., D. M. Fergusson, L. D. Leve, J. Horwood, D. Reiss, D. S. Shaw, K. K. Elam, M. N. Natsuaki, J. M. Neiderhiser, and G. T. Harold (2013, September). Maternal Smoking During Pregnancy and Offspring Conduct Problems: Evidence From 3 Independent Genetically Sensitive Research Designs. *JAMA Psychiatry 70*(9), 956.

Goodman, A. and R. Goodman (2011, January). Population mean scores predict child mental disorder rates: Validating SDQ prevalence estimators in Britain: SDQ prevalence estimators. *Journal of Child Psychology and Psychiatry 52*(1), 100–108.

Goodman, A., H. E. Joshi, B. Nasim, and C. Tyler (2015). Social and emotional skills in childhood and their long-term effects on adult life. Technical report, Institute of Education.

Goodman, A., D. L. Lamping, and G. B. Ploubidis (2010, November). When to Use Broader Internalising and Externalising Subscales Instead of the Hypothesised Five Subscales on the Strengths and Difficulties Questionnaire (SDQ): Data from British Parents, Teachers and Children. *Journal of Abnormal Child Psychology 38*(8), 1179–1191.

Goodman, R. (1994). A modified version of the Rutter parent questionnaire including extra items on children's strengths: A research note. *Journal of Child Psychology and Psychiatry 35*(8), 1483–1494.

Goodman, R. (1997). The Strengths and Difficulties Questionnaire: A research note. *Journal of child psychology and psychiatry 38*(5), 581–586.

Heckman, J., J. E. Humphries, and G. Veramendi (2016, May). Returns to Education: The Causal Effects of Education on Earnings, Health and Smoking. Technical Report w22291, National Bureau of Economic Research, Cambridge, MA.

Heckman, J., R. Pinto, and P. Savelyev (2013). Understanding the Mechanisms Through Which an Influential Early Childhood Program Boosted Adult Outcomes. *American Economic Review 103*(6), 2052–2086.

Heckman, J. J., J. Stixrud, and S. Urzua (2006, July). The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior. *Journal of Labor Economics 24*(3), 411–482.

Horn, J. L. (1965, June). A rationale and test for the number of factors in factor analysis. *Psychometrika 30*(2), 179–185.

Hsieh, C.-S. and H. van Kippersluis (2018, July). Smoking initiation: Peers and personality. *Quantitative Economics 9*(2), 825–863.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and psychological measurement 20*(1), 141–151.

Kautz, T., J. J. Heckman, R. Diris, B. Ter Weel, and L. Borghans (2014). Fostering and measuring skills: Improving cognitive and non-cognitive skills to promote lifetime success. Technical report, National Bureau of Economic Research.

Klein, J. M., A. Gonçalves, and C. F. Silva (2009). The Rutter Children Behaviour Questionnaire for teachers: From psychometrics to norms, estimating caseness. *Psico-USF 14*(2), 157–165.

Layard, R., A. E. Clark, F. Cornaglia, N. Powdthavee, and J. Vernoit (2014, November). What Predicts a Successful Life? A Life-course Model of Well-being. *The Economic Journal 124*(580), F720–F738.

Li, C.-H. (2016, September). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods 48*(3), 936–949.

Lindqvist, E. and R. Vestman (2011, January). The Labor Market Returns to Cognitive and Noncognitive Ability: Evidence from the Swedish Enlistment. *American Economic Journal: Applied Economics 3*(1), 101–128.

McDonald, R. P. (1989). An index of goodness-of-fit based on noncentrality. *Journal of classification 6*(1), 97–103.

McGee, R., S. Williams, J. Bradshaw, J. L. Chapel, A. Robins, and P. A. Silva (1985). The Rutter scale for completion by teachers: Factor structure and relationships with cognitive abilities and family adversity for a sample of New Zealand children. *Journal of Child Psychology and Psychiatry 26*(5), 727–739.

Meade, A. W., E. C. Johnson, and P. W. Braddy (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology 93*(3), 568–592.

Mendolia, S. and I. Walker (2014, September). The effect of noncognitive traits on health behaviours in adolescence. *Health Economics 23*(9), 1146–1158.

Millsap, R. E. and J. Yun-Tein (2004, July). Assessing Factorial Invariance in Ordered-Categorical Measures. *Multivariate Behavioral Research 39*(3), 479–515.

Muthén, B. (1984, March). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika 49*(1), 115–132.

Muthen, B. O., S. H. C. du Toit, and D. Spisic (1997). Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes. Technical report, Unpublished manuscipt.

Plewis, I., L. Calderwood, D. Hawkes, G. Hughes, and H. Joshi (2007). Millennium cohort study: Technical report on sampling. Technical report, Centre for Longitudinal Studies, Institute of Education, London, UK.

Putnick, D. L. and M. H. Bornstein (2016, September). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review 41*, 71–90.

Raîche, G., T. A. Walls, D. Magis, M. Riopel, and J.-G. Blais (2013, January). Non-Graphical Solutions for Cattell's Scree Test. *Methodology 9*(1), 23–29.

Reardon, S. F. and X. A. Portilla (2016, July). Recent Trends in Income, Racial, and Ethnic School Readiness Gaps at Kindergarten Entry. *AERA Open 2*(3), 233285841665734.

Revelle, W. (2018). Psych: Procedures for Personality and Psychological Research. Northwestern University.

Revelle, W. and T. Rocklin (1979). Very simple structure: An alternative procedure for estimating the optimal number of interpretable factors. *Multivariate Behavioral Research 14*(4), 403–414.

Roantree, B. and K. Vira (2018). The rise and rise of women's employment in the UK. Technical Report BN234, Institute for Fiscal Studies.

Rosseel, Y. (2012). Lavaan : An R Package for Structural Equation Modeling. *Journal of Statistical Software 48*(2), 1–36.

Rutter, M., J. Tizard, and K. Whitmore (1970). *Education, Health, and Behaviour*. London, UK: Prentice Hall.

Sass, D. A., T. A. Schmitt, and H. W. Marsh (2014, April). Evaluating Model Fit With Ordered Categorical Data Within a Measurement Invariance Framework: A Comparison of Estimators. *Structural Equation Modeling: A Multidisciplinary Journal 21*(2), 167–180.

Savelyev, P. and K. T. K. Tan (2017, December). Socioemotional Skills, Education, and Health-Related Outcomes of High-Ability Individuals. *American Journal of Health Economics*, 1–73.

Segal, C. (2008). Classroom behavior. *Journal of Human Resources 43*(4), 783–814.

Segal, C. (2013, August). Misbehavior, Education, and Labor Market Outcomes. *Journal of the European Economic Association 11*(4), 743–779.

Skrondal, A. and S. Rabe-Hesketh (2004). *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*. Chapman & Hall/CRC interdisciplinary statistics series. Boca Raton: Chapman & Hall/CRC.

Steiger, J. H. (1989). EzPATH Causal modeling.

Stone, L. L., R. Otten, R. C. M. E. Engels, A. A. Vermulst, and J. M. A. M. Janssens (2010, September). Psychometric Properties of the Parent and Teacher Versions of the Strengths and Difficulties Questionnaire for 4- to 12-Year-Olds: A Review. *Clinical Child and Family Psychology Review 13*(3), 254–274.

Tremblay, R. E., L. Desmarais-Gervais, C. Gagnon, and P. Charlebois (1987, December). The Preschool Behaviour Questionnaire: Stability of its Factor Structure Between Cultures, Sexes, Ages and Socioeconomic Classes. *International Journal of Behavioral Development 10*(4), 467–484.

University Of London. Institute Of Education. Centre For Longitudinal Studies (2016c). 1970 British Cohort Study: Forty-Two-Year Follow-Up, 2012.

University Of London. Institute Of Education. Centre For Longitudinal Studies (2016b). 1970 British Cohort Study: Thirty-Eight-Year Follow-Up, 2008-2009.

University Of London. Institute Of Education. Centre For Longitudinal Studies (2016a). 1970 British Cohort Study: Twenty-Nine-Year Follow-Up, 1999-2000.

University Of London. Institute Of Education. Centre For Longitudinal Studies (2017c). Millennium Cohort Study: Fifth Survey, 2012.

University Of London. Institute Of Education. Centre For Longitudinal Studies (2017a). Millennium Cohort Study: First Survey, 2001-2003.

University Of London. Institute Of Education. Centre For Longitudinal Studies (2017d). Millennium Cohort Study: Sixth Survey, 2015.

University Of London. Institute Of Education. Centre For Longitudinal Studies (2017b). Millennium Cohort Study: Third Survey, 2006.

van de Schoot, R., P. Lugtig, and J. Hox (2012, July). A checklist for testing measurement invariance. *European Journal of Developmental Psychology 9*(4), 486–492.

Vandenberg, R. J. and C. E. Lance (2000, January). A Review and Synthesis of the Measurement Invariance Literature: Suggestions, Practices, and Recommendations for Organizational Research. *Organizational Research Methods 3*(1), 4–70.

Vandenberg, R. J. and N. A. Morelli (2016). A contemporay update on testing for measurement equivalence and invariance. In *Handbook of Employee Commitment*. Edward Elgar Publishing.

Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika 41*(3), 321–327.

Venables, P. H., R. P. Fletcher, J. C. Dalais, D. A. Mitchell, F. Schulsinger, and S. A. Mednick (1983, April). Factor structure of the Rutter 'Children's Behaviour Questionnaire' in a primary school population in a developing country. *Journal of Child Psychology and Psychiatry 24*(2), 213–222.

Wu, H. and R. Estabrook (2016, December). Identification of Confirmatory Factor Analysis Models of Different Levels of Invariance for Ordered Categorical Outcomes. *Psychometrika 81*(4), 1014–1045.

Yoon, M. and R. E. Millsap (2007, July). Detecting Violations of Factorial Invariance Using Data-Based Specification Searches: A Monte Carlo Study. *Structural Equation Modeling: A Multidisciplinary Journal 14*(3), 435–463.

Appendices

Appendix A Deriving a common scale of socioemotional skills

In the BCS data, maternal reports on child socioemotional skills are measured using the Rutter A Scale (Rutter et al., 1970) – see Panel A of Table A1. The Rutter items are rated on three levels: 'Does not apply', 'Somewhat applies', 'Certainly applies'. Since they are all behaviours indicating lower skills, we encode all of them in reverse, i.e. 'Certainly applies' = 0, 'Somewhat applies' = 1, 'Does not apply' = 2. We augment the 19-item Rutter Scale with four additional parent-reported questions from the parental questionnaire, items A to D. These are rated on 4 levels: 'Never in the last 12 months', 'less than once a month', 'at least once a month', 'at least once a week'. we recode these into binary indicators, with 'Never' and 'Less than once a month' to 1 and zero otherwise. To increase comparability between the two scales, we merge together two pairs of items: 4 and 19 (to mirror SDQ item 12 "Often fights with other children or bullies them"), and A and B (to mirror SDQ item 3 "Often complains of head-aches, stomach-ache or sickness"). We assign the lowest category among the two original items to the newly obtained item. We also recode items 5 and 14 to binary instead of three categories. These items are recorded with a positive phrasing in SDQ, so a 3-category split would be harder to compare.

In MCS, we use the 25-item strengths and difficulties questionnaire (Goodman, 1997) – see Panel B of Table A1. All items are recorded on a 4-point scale: 'Not true', 'Somewhat true', 'Certainly true', 'Can't say'. We set the latter option to missing and recode the rest in ascending order of skill as for the BCS items, i.e. 'Certainly true' = 0, 'Somewhat true' = 1, 'Not true' = 2. For comparability with the BCS Rutter scale, we dichotomise items 3 and 5 to make them comparable with , and dichotomise and invert items 7, and 14.

Appendix B Measurement invariance details

B.1 Alternative parameterisations for the configural model

There are infinite ways to parameterise the configural model defined by (4.1) and (4.2). Widely used parameterisations are:

◇ Delta parameterisation [WEΔ] (Wu and Estabrook, 2016)
For all groups:

$$\text{diag}(\boldsymbol{\Phi}) = \boldsymbol{I}, \quad \boldsymbol{\kappa} = \mathbf{0}, \quad \boldsymbol{\nu} = \mathbf{0}, \quad \text{and } \text{diag}(\boldsymbol{\Sigma}) = \boldsymbol{I}.$$

◇ Theta parameterisation [WEΘ] (Wu and Estabrook, 2016)
For all groups:

$$\text{diag}(\boldsymbol{\Phi}) = \boldsymbol{I}, \quad \boldsymbol{\kappa} = \mathbf{0}, \quad \boldsymbol{\nu} = \mathbf{0}, \quad \text{and } \text{diag}(\boldsymbol{\Psi}) = \boldsymbol{I}.$$

◇ Anchored parameterisation [MT] (Millsap and Yun-Tein, 2004)
- For all groups, normalise a reference loading to 1 for each factor
- Set invariant across groups one threshold per item (e.g. $\tau_{0,Ai} = \tau_{0,Bi}$), and an additional threshold in the reference items above
- In the first group: $\boldsymbol{\kappa_A} = \mathbf{0}$, $\text{diag}(\boldsymbol{\Sigma}_A) = \boldsymbol{I}$
- Set all intercepts $\boldsymbol{\nu}$ to zero

The first two parameterisations (WEΔ and WEΘ) normalise the mean and variance of factors to the same constants in both groups, and they leave all loadings and intercepts to be freely estimated; they only differ in whether the additional

required normalisation is imposed on the variances of the error terms ($\boldsymbol{\Psi}$) or on the diagonal of the covariance matrix of the measures ($\boldsymbol{\Sigma}$). The MT parameterisation instead proceeds by identifying parameters in one group first, and then imposing cross-group equality constraints to identify parameters in other groups (Wu and Estabrook, 2016).

B.2 Identification of models with different levels of invariance

In the case where available measures are continuous, MI analysis is straightforward (van de Schoot et al., 2012). The hierarchy of the nested models usually proceeds by testing loadings first, and then intercepts (to establish *metric* and *scalar* invariance – see Vandenberg and Lance, 2000).

Invariance of systems with categorical measures, such as the scale we examine in this paper, is less well understood. In particular, the lack of explicit location and scale in the measures introduces an additional set of parameters compared to the continuous case (thresholds τ). This makes identification reliant on more stringent normalisations. A first comprehensive approach for categorical measures was proposed by Millsap and Yun-Tein (2004). New identification results in Wu and Estabrook (2016) indicate that, in the categorical case, invariance properties cannot be examined by simply restricting one set of parameters at a time. This is because the identification conditions used in the configural baseline model, while being minimally restrictive on their own, become binding once certain additional restrictions are imposed. In light of this, they propose models that identify structures of different invariance levels. They find that some restrictions cannot be tested alone against the configural model, because the models they generate are statistically equivalent. This is true of loading invariance, and also of threshold invariance in the case when the number of categories of each ordinal item is 3 or less. Furthermore, they suggest that comparison of both latent means and variances requires invariance in loadings, thresholds, and intercepts. A summary of the approach by Wu and Estabrook (2016) is available in Table A3.

Appendix C Appendix tables

Table A1: Behavioural screening scales in the BCS and MCS five-year surveys

Panel A: Rutter A Scale (Rutter et al., 1970) – British Cohort Study (1975) five-year survey

1. Very restless. Often running about or jumping up and down. Hardly ever still.†
2. Is squirmy or fidgety.†
3. Often destroys own or others' belongings.
4. Frequently fights other children.†
5. Not much liked by other children.
6. Often worried, worries about many things.†
7. Tends to do things on his/her own, is rather solitary.†
8. Irritable. Is quick to fly off the handle.
9. Often appears miserable, unhappy, tearful or distressed.†
10. Sometimes takes things belonging to others.
11. Has twitches, mannerisms or tics of the face or body.
12. Frequently sucks thumb or finger.
13. Frequently bites nails or fingers.
14. Is often disobedient.†
15. Cannot settle to anything for more than a few moments.†
16. Tends to be fearful or afraid of new things or new situations.†
17. Is over fussy or over particular.
18. Often tells lies.
19. Bullies other children.†

A. Complains of headaches.†
B. Complains of stomach-ache or has vomited.†
C. Complains of biliousness.
D. Has temper tantrums (that is, complete loss of temper with shouting, angry movements, etc.).†

Panel B: Strength and Difficulties Questionnaire (Goodman, 1997) – Millennium Cohort Study (2000/1) five-year survey

1. Considerate of other people's feelings.
2. Restless, overactive, cannot stay still for long.†
3. Often complains of head- aches, stomach-ache or sickness.†
4. Shares readily with other children (treats, toys, pencils, etc.).+
5. Often has temper tantrums or hot tempers.†
6. Rather solitary, tends to play alone.†
7. Generally obedient, usually does what adults request.†+
8. Many worries, often seems worried.†
9. Helpful if someone is hurt, upset or feeling ill.+
10. Constantly fidgeting or squirming.†
11. Has at least one good friend.+
12. Often fights with other children or bullies them.†
13. Often unhappy, down-hearted or tearful.†
14. Generally liked by other children.+
15. Easily distracted, concentration wanders.†
16. Nervous or clingy in new situations, easily loses confidence.†
17. Kind to younger children.+
18. Often lies or cheats.
19. Picked on or bullied by other children.
20. Often volunteers to help others (parents, teachers, other children).+
21. Thinks things out before acting.+
22. Steals from home, school or elsewhere.
23. Gets on better with adults than with other children.
24. Many fears, easily scared.
25. Sees tasks through to the end, good attention span.+

Notes: Items denoted by + are positively coded in the original scale. Items denoted by † are retained in the new comparable scale.

Table A2: Description of harmonised variables

Variable Group	Age	Variable	Note
Maternal education	5	Post-compulsory schooling[d]	Whether mother continued schooling past the compulsory age, based on her year of birth. School leaving age in England was changed from 14 to 15 in 1947 and from 15 to 16 in 1972.
Maternal employment	5	Employed[d]	Includes full time and part time
Father occupation	5	White collar (I-IIINM)[d] Blue collar (IIIM-V-other)[d] No father figure[d]	Based on father's Registrar General Social Class classification of occupations. White collar includes I (Professional), II (Managerial/technical), IIINM (Skilled non-manual). Blue collar includes IIIM (Skilled manual), IV (Partly skilled), V (Unskilled), Other, Unemployed, and Armed forces. No father figure is a dummy for children whose father does not live in the same household.
Maternal background	0/5	Mother's age at birth Mother's height (cm) Mother unmarried at birth[d] Child nonwhite ethnicity[d] Number of children in HH	All variables are self-reported by the mother at birth, except for number of children in household (at five years old). Unmarried is only based on marital status, and includes cohabitation.
Pregnancy	0	Child is firstborn[d] Number of previous stillbirths Mother smoked in pregnancy[d] Preterm birth (under 37 weeks gestation)[d] (log) birth weight (kg)	Parity, stillbirths, and smoking are self-reported by the mother. Gestational length and birth weight are from hospital records.
Cognitive skills	5		Based on test batteries administered to the cohort member at five. Three tests are used for BCS children: Copy Designs child is asked to copy simple designs adjacently), Human Figure Drawing (child draws an entire human figure), English Picture Vocabulary Test (child identifies the picture referring to a word among four pictures). Three *different* tests are used in the MCS: BAS Naming Vocabulary (child is shown a series of pictures and asked to name it), BAS Picture Similarity (child is shown a row of 4 pictures on a page and places a card with a fifth picture under the one most similar to it), BAS Pattern Construction (child constructs a design by putting together flat squares or solid cubes with patterns on each side).
Adolescent outcomes	16 (BCS) 14 (MCS)	Child tried smoking[d] Body Mass Index (BMI)	Smoking is self reported by the child. Height and weight are taken as part of a medical examination.
Adult outcomes (BCS only)	34 42 34, 42	Higher education[d] Employed[d] (log) gross weekly pay	Higher education is defined on having a university degree or its vocational equivalent. It corresponds to level 4 or 5 in the National Vocational Qualification (NVQ) equivalence. Employed is a dummy for being in paid employment or self-employment, either full or part time. Gross weekly pay is weekly pre-tax pay from the respondent's main activity, conditional on being a paid employee.

Notes: Variables denoted by [d] are binary or categorical.

Table A3: Parameterisations for measurement invariance

Invariance level	Description	Restrictions
Configural (WEΘ)	· Minimally restrictive model for identification	For all groups: $\mathrm{diag}(\boldsymbol{\Phi}) = \boldsymbol{I}$; $\boldsymbol{\kappa} = \boldsymbol{0}$; $\boldsymbol{\nu} = \boldsymbol{0}$; $\mathrm{diag}(\boldsymbol{\Psi}) = \boldsymbol{I}$
Threshold invariance	· Restricts thresholds τ to be equal across groups · Statistically equivalent to configural (when measures have 3 categories or less)	$\tau_{1,ci} = \tau_{1,c'i}$ for all items, $\forall c, c'$ $\tau_{2,ci} = \tau_{2,c'i}$ for non-binary items, $\forall c, c'$ For all groups: $\mathrm{diag}(\boldsymbol{\Phi}) = \boldsymbol{I}$; $\boldsymbol{\kappa} = \boldsymbol{0}$ For ref. group A: $\boldsymbol{\nu}_A = \boldsymbol{0}$; $\mathrm{diag}(\boldsymbol{\Sigma}_A) = \boldsymbol{I}$
Threshold and Loading invariance	· Restricts thresholds τ and loadings λ to be equal across groups · Allows comparison of latent factor variances	$\tau_{1,ci} = \tau_{1,c'i}$ for all items, $\forall c, c'$ $\tau_{2,ci} = \tau_{2,c'i}$ for non-binary items, $\forall c, c'$ $\lambda_{ci} = \lambda_{c'i}$ for all items, $\forall c, c'$ For all groups: $\boldsymbol{\kappa} = \boldsymbol{0}$ For ref. group A: $\boldsymbol{\nu}_A = \boldsymbol{0}$; $\mathrm{diag}(\boldsymbol{\Sigma}_A) = \boldsymbol{I}$; $\mathrm{diag}(\boldsymbol{\Phi}_A) = \boldsymbol{I}$
Threshold, Loading, and Intercept invariance	· Restricts thresholds τ and loadings λ to be equal across groups · Restricts intercepts ν to zero in both groups · Allows comparison of latent factor variances *and* means	$\tau_{1,ci} = \tau_{1,c'i}$ for all items, $\forall c, c'$ $\tau_{2,ci} = \tau_{2,c'i}$ for non-binary items, $\forall c, c'$ $\lambda_{ci} = \lambda_{c'i}$ for all items, $\forall c, c'$ For all groups: $\boldsymbol{\nu} = \boldsymbol{0}$ For ref. group A: $\boldsymbol{\kappa}_A = \boldsymbol{0}$; $\mathrm{diag}(\boldsymbol{\Sigma}_A) = \boldsymbol{I}$; $\mathrm{diag}(\boldsymbol{\Phi}_A) = \boldsymbol{I}$

Notes: Adapted from Wu and Estabrook (2016).

Table A4: Item prevalence, by cohort and gender

				Males						Females					
				BCS			MCS			BCS			MCS		
Itm.	Factor	Cat.	Title	Cert. Appl. (%)	Smtm. Appl. (%)	Appl. (%)	Cert. True (%)	Smwt. True (%)	True (%)	Cert. Appl. (%)	Smtm. Appl. (%)	Appl. (%)	Cert. True (%)	Smwt. True (%)	True (%)
1	EXT	3	Restless	32.0	40.4		17.3	29.0		25.0	40.4		13.1	24.2	
2	EXT	3	Squirmy/fidgety	12.3	31.8		11.3	29.4		11.3	32.1		9.1	25.4	
3	EXT	3	Fights/bullies	6.6	39.3		1.7	9.1		3.1	28.1		0.9	4.9	
4	EXT	3	Distracted	8.0	30.1		16.0	44.1		6.1	25.6		9.7	38.7	
5	EXT	2	Tantrums			26.3			51.1			19.6			47.1
6	EXT	2	Disobedient			73.7			48.9			64.9			41.7
7	INT	3	Worried	5.6	29.4		2.4	11.8		5.8	31.3		1.5	11.9	
8	INT	3	Fearful	7.0	29.2		11.1	34.3		6.6	30.0		9.8	38.0	
9	INT	3	Solitary	9.6	37.4		6.4	26.1		8.5	35.3		5.0	24.3	
10	INT	3	Unhappy	2.3	18.3		1.6	9.0		3.0	22.5		1.6	8.3	
11	INT	2	Aches			13.3			17.0			14.8			22.2

Notes: The table shows the prevalence by gender and cohort for each item of our novel subscale. *Itm.* is item number. *Factor* is the latent construct to which the item loads – EXT is Externalising skills, INT is Internalising skills. *Cat.* is the number of categories in which the item is coded – 2 denotes a binary item (applies/does not apply) and 3 denotes a 3-category item. *Title* is a short label for the item. *Cert. / Smtm. Appl.* = Certainly / sometimes applies. *Cert. / Smwt. True* = Certainly / somewhat true.

Table A5: Summary statistics

	Males		Females	
	BCS	MCS	BCS	MCS
Mother age (0)	25.935	29.469	25.902	29.423
	(5.413)	(5.564)	(5.277)	(5.658)
Mother height (m)	1.613	1.645	1.614	1.645
	(0.063)	(0.068)	(0.065)	(0.070)
Unmarried (0)	0.049	0.363	0.055	0.353
	(0.216)	(0.481)	(0.227)	(0.478)
Nonwhite child	0.027	0.100	0.035	0.101
	(0.163)	(0.301)	(0.185)	(0.301)
Number of children in HH (5)	1.560	1.352	1.541	1.309
	(1.138)	(0.995)	(1.125)	(0.967)
Firstborn child	0.369	0.412	0.385	0.423
	(0.483)	(0.492)	(0.487)	(0.494)
Number previous stillbirths	0.023	0.010	0.021	0.011
	(0.156)	(0.098)	(0.147)	(0.113)
Mother smoked in pregnancy	0.401	0.209	0.382	0.200
	(0.490)	(0.407)	(0.486)	(0.400)
Preterm birth	0.040	0.077	0.032	0.068
	(0.197)	(0.267)	(0.175)	(0.252)
Missing gest. age	0.191	0.008	0.180	0.008
	(0.393)	(0.091)	(0.384)	(0.089)
Birthweight (kg)	3.369	3.443	3.254	3.317
	(0.544)	(0.587)	(0.509)	(0.568)
Mother has post-compulsory education (5)	0.386	0.563	0.381	0.560
	(0.487)	(0.496)	(0.486)	(0.496)
Mother is employed (5)	0.426	0.618	0.413	0.614
	(0.495)	(0.486)	(0.492)	(0.487)
Father occupation: blue collar	0.614	0.398	0.610	0.391
	(0.487)	(0.490)	(0.488)	(0.488)
No father figure	0.046	0.179	0.051	0.174
	(0.209)	(0.384)	(0.220)	(0.379)

Notes: The table shows mean (standard deviation) of harmonised variables used in the analysis.

Table A6: Suggested number of factors to retain

	BCS (1970)			MCS (2000/1)		
Approach	All	Males	Females	All	Males	Females
Optimal Coordinates	3	3	3	2	2	2
Acceleration Factor	1	1	1	1	1	1
Parallel Analysis	3	3	3	2	2	2
Kaiser	3	3	3	2	2	2
VSS Compl. 1	2	2	1	1	1	1
VSS Compl. 2	2	2	2	2	2	2
Velicer MAP	1	1	1	2	2	2

Notes: The table compares the optimal number of factors suggested by different aprroaches for our novel scale; scree test based approaches (optimal coordinates, acceleration factor – Raîche et al., 2013), parallel analysis (Horn, 1965), Kaiser's eigenvalue rule (Kaiser, 1960), Very Simple Structure (VSS, Revelle and Rocklin, 1979), Velicer Minimum Average Partial test (MAP, Velicer, 1976).

Table A7: Loadings from exploratory factor analysis

		BCS (1970)		MCS (2000/1)	
Item	Title	Factor 1 (EXT)	Factor 2 (INT)	Factor 1 (EXT)	Factor 2 (INT)
1	Restless	**0.79**	-0.113	**0.924**	-0.077
2	Squirmy/fidgety	**0.67**	0.021	**0.746**	0.037
3	Fights/bullies	**0.499**	0.046	**0.507**	0.237
4	Distracted	**0.629**	0.05	**0.661**	0.036
5	Tantrums	**0.484**	0.177	**0.485**	0.216
6	Disobedient	**0.598**	0.066	**0.563**	-0.011
7	Worried	-0.037	**0.729**	-0.079	**0.782**
8	Fearful	-0.064	**0.595**	-0.028	**0.512**
9	Solitary	0.075	**0.312**	0.003	**0.459**
10	Unhappy	0.25	**0.507**	0.112	**0.745**
11	Aches	0.135	**0.268**	0.036	**0.45**

Notes: The table displays the factor loadings obtained from exploratory factor analysis (EFA) on our novel scale, separately by cohort. The EFA is based on the decomposition of the polychoric correlation matrix, and uses weighted least squares and oblimin rotation.

Table A8: Measurement invariance fit comparison

Model	Num. params	Absolute fit						Relative fit					
		χ^2	RMSE	SRMR	MFI	CFI	G-hat	χ^2 p	ΔRMSE	ΔSRMR	ΔMFI	ΔCFI	ΔG-hat
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
A: Entire sample													
Configural	124	1887.3	0.0516	0.0647	0.9443	0.9361	0.9796						
Threshold + Loading Inv	97	2217.7	0.0520	0.0679	0.9348	0.9310	0.9761	0.0000	0.0004	0.0033	-0.0095	-0.0051	-0.0035
Threshold, Loading, + Intercept Inv	70	7198.5	0.0908	0.0746	0.7923	0.7693	0.9220	0.0000	0.0392	0.0099	-0.1520	-0.1668	-0.0576
B: 59-61 months sample													
Configural	124	1551.5	0.0527	0.0656	0.9420	0.9285	0.9788						
Threshold + Loading Inv	97	1753.9	0.0520	0.0684	0.9349	0.9268	0.9761	0.0000	-0.0007	0.0028	-0.0071	-0.0017	-0.0026
Threshold, Loading, + Intercept Inv	70	4640.1	0.0822	0.0738	0.8261	0.8004	0.9351	0.0000	0.0295	0.0082	-0.1160	-0.1280	-0.0437
C: Males only													
Configural	62	987.1	0.0522	0.0650	0.9432	0.9388	0.9792						
Threshold + Loading Inv	53	1118.1	0.0529	0.0670	0.9357	0.9328	0.9764	0.0000	0.0007	0.0020	-0.0074	-0.0060	-0.0028
Threshold, Loading, + Intercept Inv	44	3673	0.0944	0.0731	0.7931	0.7708	0.9223	0.0000	0.0423	0.0081	-0.1500	-0.1681	-0.0569
D: Females only													
Configural	62	900.2	0.0510	0.0644	0.9456	0.9324	0.9801						
Threshold + Loading Inv	53	1087.6	0.0536	0.0686	0.9341	0.9211	0.9758	0.0000	0.0026	0.0043	-0.0115	-0.0113	-0.0043
Threshold, Loading, + Intercept Inv	44	3347.5	0.0926	0.0749	0.8003	0.7488	0.9251	0.0000	0.0416	0.0105	-0.1453	-0.1835	-0.0550

Notes: The table presents fit indices for models of different invariance levels, following Wu and Estabrook (2016) Col. (1) displays the number of estimated parameters for each model. Col. (2) and (8) present the value of the χ^2 statistic and the pvalue of the test of equality with respect to the configural model. Col. (3)-(7) and (9)-(13) present alternative fit indices (AFIs), in absolute values and differences from the configural model respectively. RMSE = Root mean squared error of approximation; SRMR = standardised root mean residual; MFI = McDonald non-centrality index; CFI = comparative fit index; G-hat = gamma-hat. Panel A shows results for the whole sample of children in the BCS and MCS cohorts. Panel B is restricted to a subsample of children in the age range of maximum overlap between the two cohorts (59-61 months). Panel C and D are limited to the samples of male and female children respectively.

Table A9: Parameter estimates from factor model with threshold and loading invariance

Panel A: Measurement parameters

		Loadings	**Thresholds**		**Intercepts** (BCS M = 0)			**Variances** (BCS M = 1)		
		λ	τ_1	τ_2		ν			diag(Ψ)	
Item	Factor	All	All	All	BCS F	MCS M	MCS F	BCS F	MCS M	MCS F
X1	EXT	1.218	-0.716	0.894	0.329	1.154	1.503	1.141	0.690	0.902
X2	EXT	1.011	-1.641	-0.215	0.014	0.201	0.394	0.872	0.718	0.785
X3	EXT	0.637	-1.725	-0.156	0.486	1.345	1.951	1.116	0.977	1.248
X4	EXT	0.781	-1.843	-0.375	0.181	-0.556	-0.190	0.880	0.785	0.801
X5	EXT	0.665	-0.787		0.260	-0.730	-0.597	1.000	1.000	1.000
X6	EXT	0.683	0.746		0.303	0.830	1.112	1.000	1.000	1.000
X7	INT	0.759	-1.995	-0.501	-0.101	1.210	0.969	0.858	1.298	0.845
X8	INT	0.511	-1.716	-0.367	0.013	-0.131	-0.282	0.981	1.192	0.890
X9	INT	0.384	-1.400	-0.104	0.059	0.478	0.585	0.941	0.909	0.973
X10	INT	1.135	-3.034	-1.257	-0.207	1.149	1.241	1.030	0.850	1.256
X11	INT	0.420	-1.247		-0.094	-0.021	-0.329	1.000	1.000	1.000

Panel B: Latent variable parameters

	Mean		**Covariance**				Correlation	
	κ		Φ					
	BCS	MCS	BCS		MCS		BCS	MCS
Males								
θ^{EXT}	0.000	0.000	1.000		1.334			
θ^{INT}	0.000	0.000	0.420	1.000	0.878	1.934	0.420	0.546
Females								
θ^{EXT}	0.000	0.000	0.985		1.199			
θ^{INT}	0.000	0.000	0.478	1.012	0.607	1.418	0.478	0.465

Notes: The table presents estimates for the factor model with loadings λ and thresholds τ restricted to be equal across cohorts. Panel A shows estimates of the measurement parameters. Loadings and thresholds are the same across all cohorts. Intercepts are restricted to zero in the reference group, i.e. males in BCS (not shown). Variances of the error terms are restricted to one in the reference group, i.e. males in BCS (not shown), and for the items that only have two categories (5, 6, 11). Panel B shows estimates of the latent variable parameters. Means are restricted to zero in all cohort-gender groups, while variances are restricted to one only in the reference group, i.e. males in BCS.

Table A10: Predictors of adult behaviours, BCS

	Males				Females			
	Mean	Coefficients			Mean	Coefficients		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Daily smoker (42)	.180				.147			
Externalising skills (5)		−.062***	−.059***			−.050***	−.059***	
		(.017)	(.018)			(.015)	(.018)	
Internalising skills (5)		.027	.025			.043**	.048***	
		(.020)	(.021)			(.017)	(.017)	
Externalising (sum score)				−.040***				−.027***
				(.012)				(.010)
Internalising (sum score)				.004				.023***
				(.011)				(.009)
Cognitive skills (5)			−.022	−.022			−.032***	−.032***
			(.015)	(.014)			(.012)	(.012)
Adj. R^2		0.044	0.045	0.045		0.037	0.041	0.041
Observations		1294	1216	1216		1678	1572	1572
BMI (42)	27.5				26.1			
Externalising skills (5)		−.267	−.138			−.386	−.138	
		(.221)	(.229)			(.261)	(.229)	
Internalising skills (5)		.400	.316			.102	−.035	
		(.266)	(.272)			(.289)	(.300)	
Externalising (sum score)				−.041				−.204
				(.159)				(.176)
Internalising (sum score)				.129				−.038
				(.149)				(.153)
Cognitive skills (5)			−.235	−.235			−.729***	−.728***
			(.194)	(.192)			(.223)	(.214)
Adj. R^2		0.028	0.024	0.024		0.034	0.047	0.047
Observations		1149	1078	1078		1399	1317	1317

Notes: The table shows coefficients from linear regressions of cohort members' adolescent and adult outcomes on their externalising and internalising socioemotional skills at five years of age. Col. (1) shows the mean of the outcome for males. Col. (2) regresses the outcome on the scores obtained from the factor model in Section 4. Col. (3) additionally controls for cognitive ability at age five. This is a simple factor score obtained by aggregating the available cognitive measures. Col. (4) uses sum scores (see Figure A1) instead of factor scores. All standard errors in parentheses are obtained using 1,000 bootstrap repetitions, taking into account the factor estimation stage that precedes the regression. Col. (5) to (8) repeat for female cohort members. All estimates additionally control for region of birth, maternal education (5), maternal employment (5), father occupation (5), maternal background (age, height, nonwhite ethnicity, number of children in HH), pregnancy (firstborn child, number of previous stillbirths, mother smoked in pregnancy, preterm birth, (log) birth weight). See Table A2 for a description of the variables used.

Appendix D Appendix figures

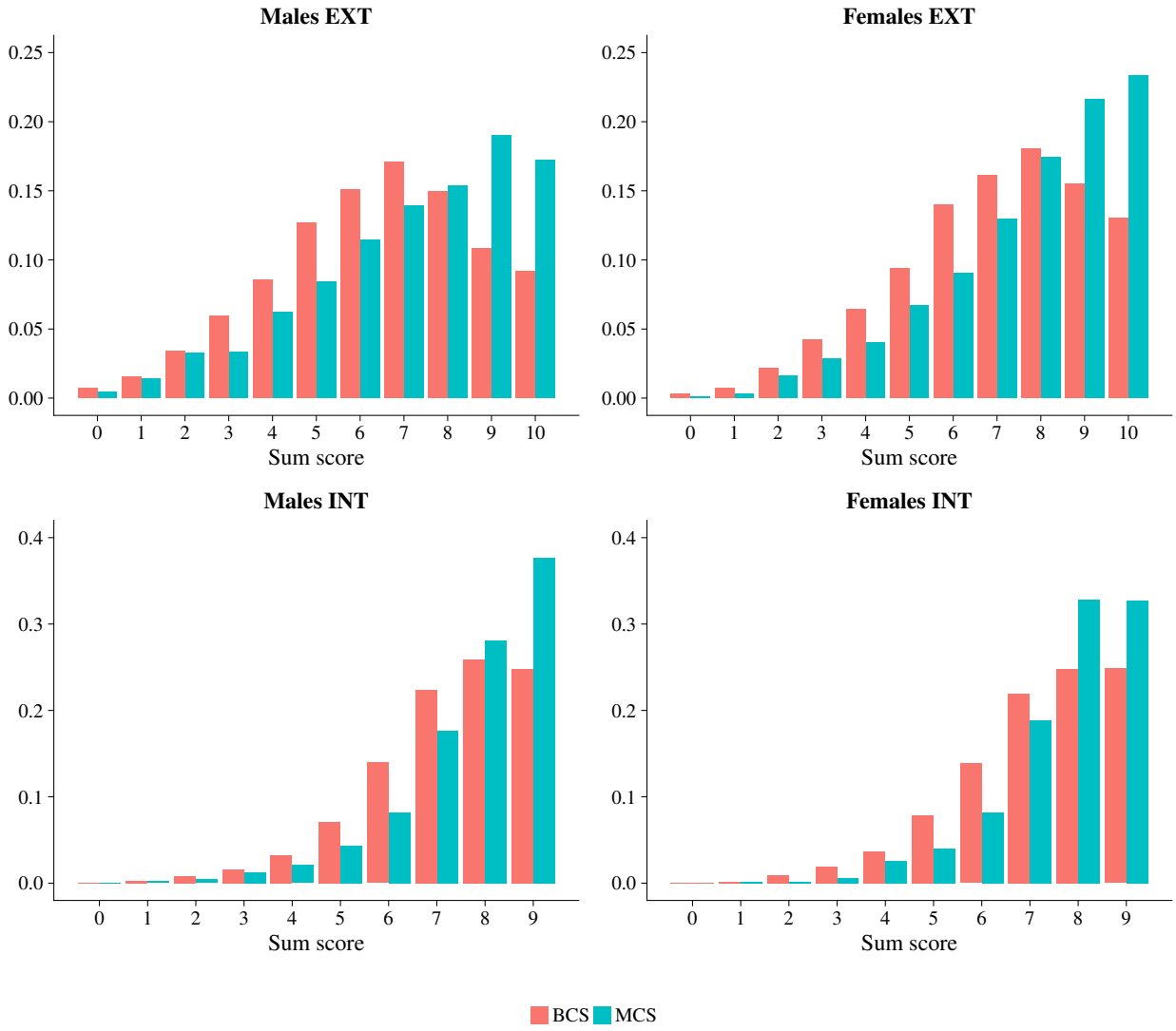

Figure A1: Distribution of sum scores

Notes: The figure shows the distribution of the externalising and internalising sum scores at age five, by gender and cohort. The scores are obtained by assigning 0, 1, or 2 points for each item in the scale in Table 1. Zero points are assigned for 'Certainly Applies / True' responses, one point for 'Sometimes applies / somewhat true', and two points for 'Doesn't apply'. Only 0 or 1 points are assigned for items that are coded as having two categories (5,6, and 11). Higher scores correspond to *better* skills.

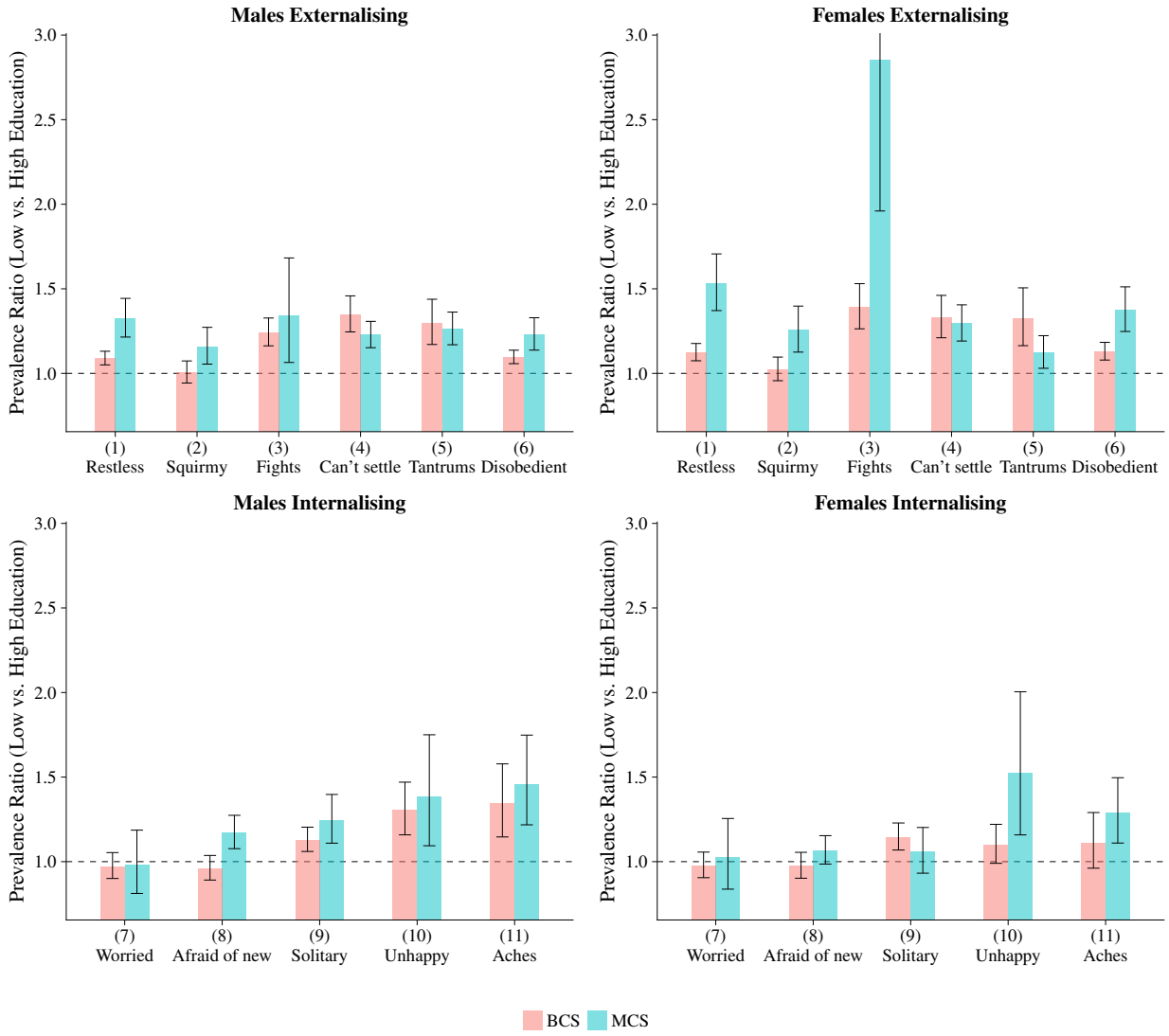

Figure A2: Item-level inequality by mother's education

Notes: The graph displays the ratio between the prevalence of each item in our scale in children of educated vs uneducated mothers, by cohort and gender. All items that have three categories in the scale have been dichotomised. For example, if the prevalence of the 'Restless' behaviours among children of mothers with compulsory schooling in the BCS cohort is 7.5%, and 5% among mothers with post-compulsory schooling, the ratio will be 1.5. The error bars represent 95% confidence intervals.

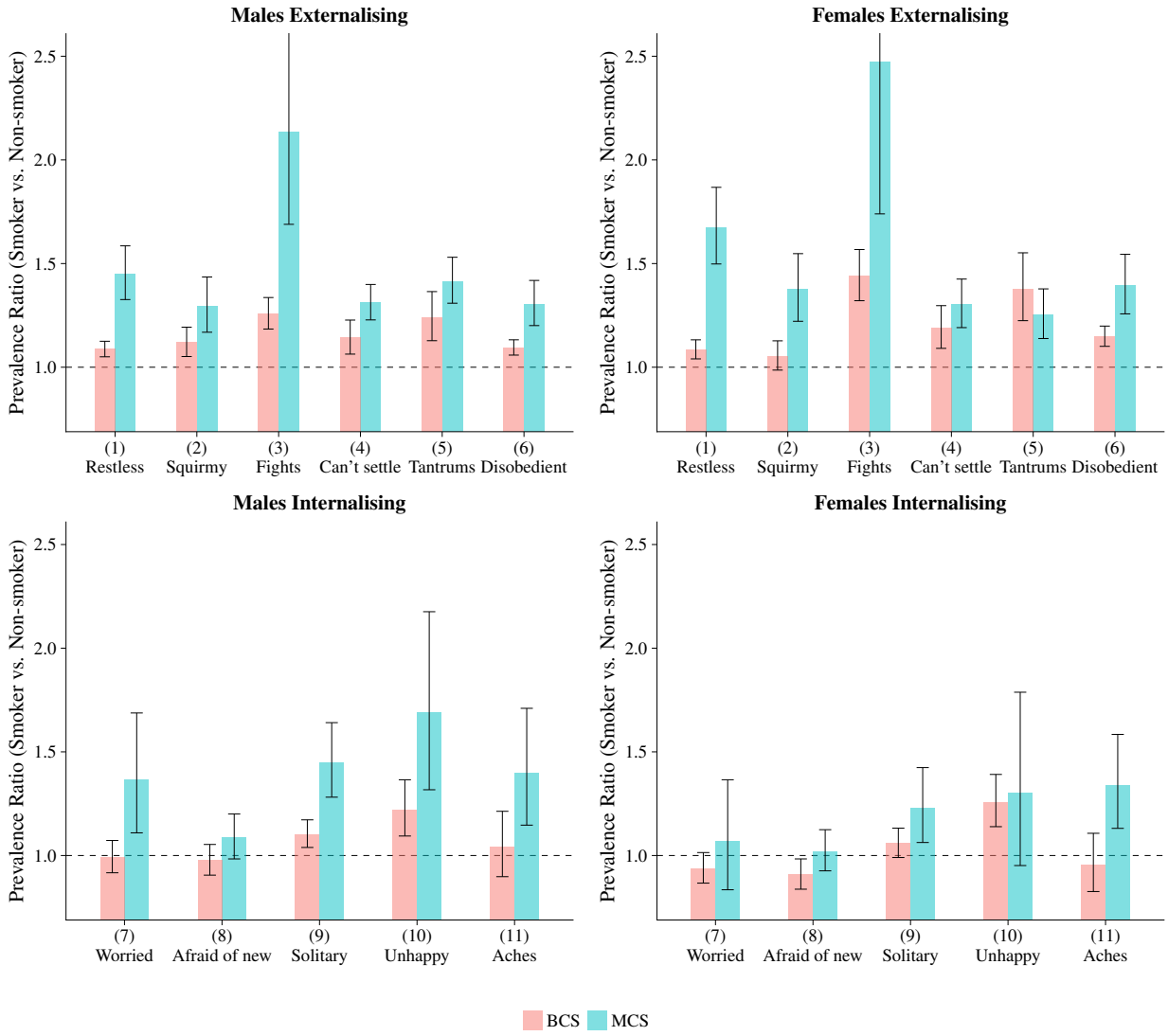

Figure A3: Item-level inequality by mother's pregnancy smoking

Notes: The graph displays the ratio between the prevalence of each item in our scale in children of mothers who smoked in pregnancy vs non-smokers, by cohort and gender. All items that have three categories in the scale have been dichotomised. For example, if the prevalence of the 'Restless' behaviours among children of smoker mothers in the BCS cohort is 7.5%, and 5% among non-smoker mothers, the ratio will be 1.5. The error bars represent 95% confidence intervals.

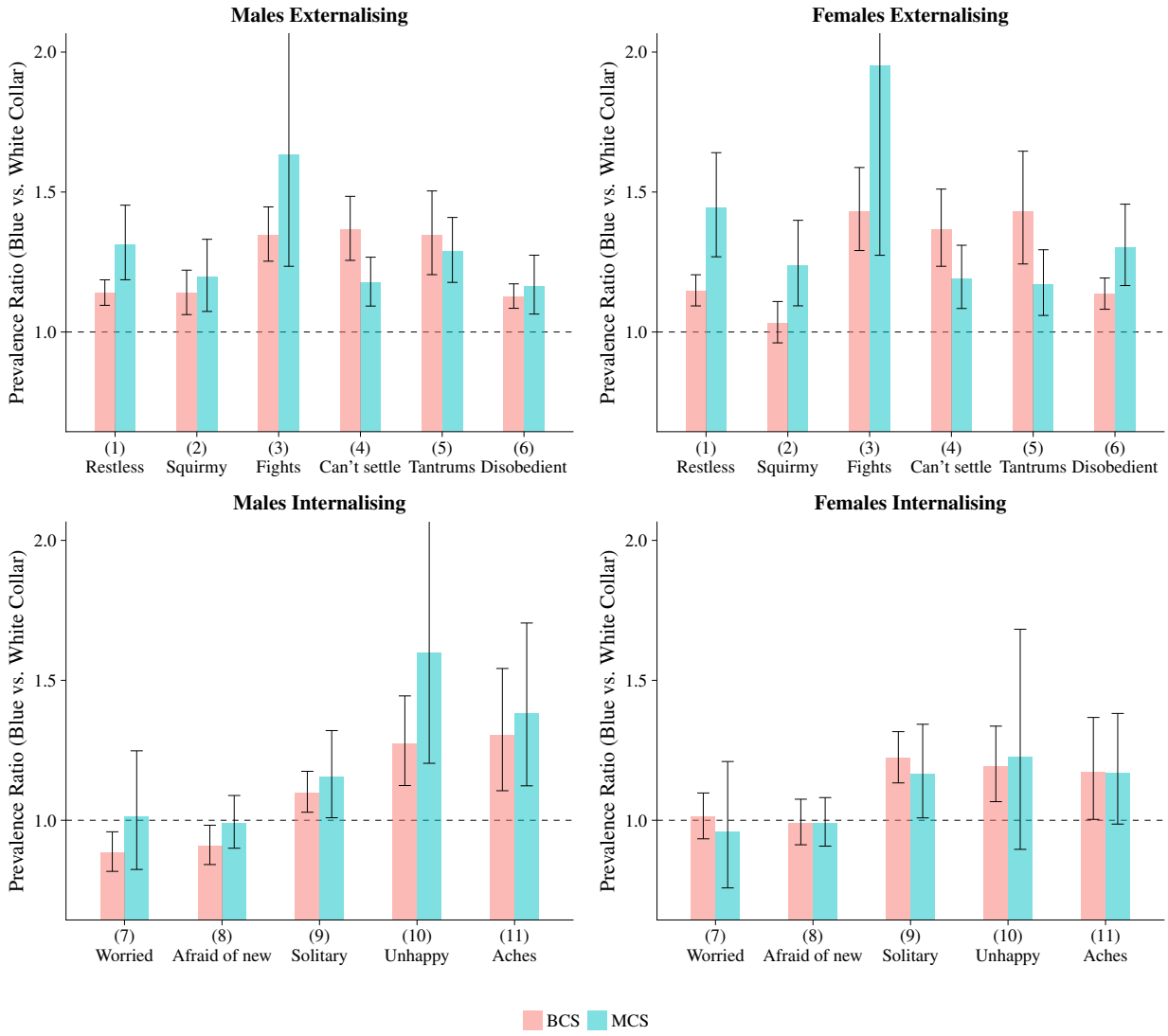

Figure A4: Item-level inequality by father's occupation

Notes: The graph displays the ratio between the prevalence of each item in our scale in children of white collar vs blue collar fathers, by cohort and gender. All items that have three categories in the scale have been dichotomised. For example, if the prevalence of the 'Restless' behaviours among children of blue collar fathers in the BCS cohort is 7.5%, and 5% among white collar fathers, the ratio will be 1.5. The error bars represent 95% confidence intervals.