What happens to bright 5-year-olds from poor backgrounds? Longitudinal evidence from the Millennium Cohort Study.

John Jerrim[1] (UCL Social Research Institute)

Maria Palma Carvajal (UCL Social Research Institute)

Abstract

High-achieving children from low-income families have an opportunity to break through the glass ceiling and achieve upwards social mobility. Yet there have been relatively few studies investigating how key outcomes for this group develop throughout childhood, and how this compares to their equally able but more socio-economically advantaged peers. This paper draws upon Millenium Cohort Study data from the UK to provide new evidence on this issue. We find that the cognitive skills of bright 5-year-olds from low-income families keep pace with those of children from high-income families through to the end of primary school. However, the transition into secondary is a critical period, with high-achieving children from poor families experiencing a particularly sharp relative decline in their attitudes towards school, behaviour, mental health and academic achievement between age 11 and 14. The failure to fully capitalise on the early potential of this group is likely to be a key reason why the UK is failing to become a more socially fluid society.

Key Words: High-ability children, socio-economic disadvantaged, educational transitions, social mobility, educational disadvantage.

[1] Social Research Institute, University College London, 20 Bedford Way London, WC1H 0AL. E-mail: j.jerrim@ucl.ac.uk (John Jerrim).

1. Introduction

Social mobility and the later lifetime outcomes of individuals from disadvantaged socio-economic backgrounds have become topics of widespread academic and public policy interest. It is now well-established that young people raised in more deprived socio-economic environments go on to obtain worse educational, social, economic and health outcomes than their more advantaged peers (Institute for Fiscal Studies, 2023). In the United Kingdom – the empirical setting for this paper – particular interest has been shown in the capacity of young people from low-income backgrounds to break the "glass ceiling" – such as obtaining a degree from an elite university and entering a top professional job (Wakeling & Savage, 2015). Although there are many factors driving this association between family background and later outcomes, it is widely recognised that what happens during early childhood plays an important role (Stewart & Waldfogel, 2017). For instance, numerous studies have documented how sizeable socio-economic differences can be observed in children's cognitive and socio-emotional skills by age 5 (Bradbury et al., 2019), with some arguing that these gaps further increase during young people's time at school (Bradbury et al., 2015).

The theory of lifecycle skill formation has become a widely cited framework used to explain such patterns (Cunha et al., 2006). This postulates that there are "sensitive" and "critical" periods in children's development, when their cognitive and socio-emotional skills develop rapidly. Many of these periods occur during the first years of life and – if not suitably nurtured – will lead to a child falling behind. This gives rise to the substantial socio-economic gaps in children's outcomes when they start school (Bradbury et al., 2019). It is then both challenging and costly to help these children catch-up (Heckman, 2008). Indeed, within this framework, having more advanced baseline skills is thought to help young people to develop said skills further and faster than others – i.e. that skills beget skills (OECD, 2015). This then puts children with strong cognitive abilities during the early years in pole position to push on and excel at school (Bradbury et al., 2019).

Given this theory, one group of particular interest are initially high-achieving children from low-income backgrounds (Loft & Danechi, 2020). These are young people who – despite their disadvantaged upbringing – have managed to navigate their way through their first years of life to develop great potential. They have the core baseline skills needed to flourish at school and are thus the group who have perhaps the best opportunity to become upwardly mobile. In other words, these young people are key to policymaker's hopes of increasing socio-economic diversity amongst its top universities and most prestigious jobs. Yet, at the same time, these children still have a long way to go before their potential is fulfilled. This includes making a successful transition into school, navigating their way through adolescence, potentially entering higher education, finding an appropriate job and then progressing upwards within their chosen career.

A relatively small number of studies have previously considered the progress that initially high-achieving (or high potential) children from disadvantaged socio-economic backgrounds make along this journey, and how this compares to their equally-able – but more socio-economically advantaged – peers. In an influential study, Feinstein (2003) reported that children from disadvantaged backgrounds who were amongst the most developmentally advanced at 22 months had fallen behind socio-economically advantaged children with developmental challenges before age 5. This claim was countered, however, by Jerrim & Vignoles (2013) who illustrated how this was likely to be a statistical artifact driven by regression to the mean. They went on to show how initially high-achieving children from disadvantaged backgrounds broadly maintained their position relative to their peers from more advantaged backgrounds (in terms of their cognitive skills) between the ages of 3 and 7. Focusing on high-achieving children at older ages (age 11 – the end of primary school in England), Crawford et al. (2017) found that those from disadvantaged homes lost some ground in their academic achievement relative to those from more advantaged upbringings during secondary school, particularly between ages 11 and 14. Recent research from Holt-White & Cullinane (2023) supports this view, arguing that only 40% of disadvantaged children who were high-achievers at the end of primary school in England (age 11) went on to achieve top grades (five or more GCSEs at grades 7-9) at age 16, compared to 62% of their equally high-achieving but socio-economically advantaged peers. Interestingly, they also used smaller scale survey data to explore the attitudes of these groups during the COVID-19 pandemic, finding that a fifth of high-achieving disadvantaged children agreed with the statement that "*people like me don't have much of a chance in life*", compared to a tenth of high-achieving children from more advantaged backgrounds. It should be noted, however, that this study made no attempt to adjust for Kelley's paradox (Wainer & Brown, 2007) – a form of regression to the mean (a point which we return to in section 3).

The empirical evidence presented in this paper seeks to further extend the evidence base regarding this important group. Specifically, it will be one of the first studies – either in the UK or internationally – to track the outcomes of high-achieving disadvantaged children defined at a young age (age 5 - around the point they start school) through to the end of compulsory education. This not only includes consideration of their educational outcomes, but a broader set of measures as well. We thus contribute to the existing literature by not just investigating whether initially high-achieving children from low-income backgrounds fall behind their higher-income peers, but also possible clues as to why. Do they, for instance, lose motivation at school, and if so, at what age? Are they more likely to fall-in with the "wrong crowd", and start getting into trouble with the law? By being higher achieving than their school peers, are they more likely to be bullied or develop mental health problems? Or do they simply lack the same drive and ambition to achieve? Our analysis provides important new insights into such issues, thus providing the most comprehensive analysis of the outcomes of initially high achieving children from low-income backgrounds to date. In summary, our research questions are:

- Research question 1. How do the cognitive skills and school grades of bright 5-year-olds from rich and poor backgrounds compare through to age 16?
- Research question 2. To what extent do bright 5-year-olds from poor backgrounds become disinterested or unmotivated at school relative to their peers from more affluent backgrounds?
- Research question 3. How do the socio-emotional outcomes of bright 5-year-olds from low-income backgrounds develop compared to their high-income peers?
- Research question 4. Are bright 5-year-olds from poor backgrounds more likely to get into trouble as teenagers compared to 5-year-olds from rich backgrounds?
- Research question 5. Are bright 5-year-olds from poor backgrounds more likely to get bullied at school than bright 5-year-olds from richer backgrounds? How does this change during primary and secondary school?
- Research question 6. Are bright 5-year-olds from low-income backgrounds more likely to develop mental health problems than their equally able peers from high-income backgrounds?

In answering these research questions, we also contribute methodologically to the study of high-achieving children from rich and poor backgrounds. Jerrim & Vignoles (2013) discussed the issue of regression to the mean in studies that attempt to track test scores of high-achieving children over time, offering a simple solution. This is to use one test score to classify children into higher/lower achievement groups, and then using a separate test to track changes in their skills. Such an approach works well when one’s interest is in exploring trajectories in the same outcome over time, such as children’s (subject specific) test scores. It is however of less use when one is interested in a range of different outcomes, such as whether high-achieving pupils from advantaged and disadvantaged backgrounds get into trouble with the law or develop mental health problems. Indeed, in such situations, our reading of the existing literature is that very few studies attempt to consider the impact of Kelley’s paradox on the results. We hence discuss this issue at length, illustrate how it essentially causes the same problem as regression to the mean, and propose a simple, transparent way that one can adjust estimates to produce (under stated assumptions) credible upper and lower bounds.

The paper now proceeds as follows. In section 2 we describe the Millennium Cohort Study (MCS) data. Section 3 discusses Kelley’s paradox as applied to the study of high-achieving children from high- and low-income backgrounds, and then turns to our methodological approach. Answers to our research questions are presented in section 4, with conclusions following in section 5.

2. Data

The Millennium Cohort Study (MCS) is a rich, nationally representative longitudinal survey that follows 18,818 children born in the UK in 2000–2002. The sample is geographically clustered and is stratified to over-represent areas of England with relatively high proportions of ethnic minorities and high levels of child poverty, as well as areas in the three smaller countries of Wales, Scotland, and Northern Ireland. Baseline interviews were conducted when the children were approximately nine months old, and follow-up interviews were conducted when the children were around three, five, seven, 11, 14 and 17 years old. Parents and their children have been interviewed and completed questionnaires across the various sweeps. Our baseline sample are the 15,808 children that participated in the second (age 3) wave – the first time when their cognitive skills were measured. Of the 18,818 cohort members that participated at 9 months, 10,757 remained in the study at age 17. This reflects a 43% attrition rate. The MCS response weights are applied in our analysis to adjust for non-random non-response.

As part of the age 3 and age 5 MCS sweeps, children completed a set of cognitive tests. At age 3 this encompassed:

(a) The Bracken school readiness test. An assessment covering children’s knowledge of colours, letters, numbers, sizes, shapes, and comparisons. For instance, children were asked to point to a specific colour, number or letter, or which rope is long or short. Together, it captures children's crystalised abilities, visual processing and quantitative knowledge. The test took around 10-15 minutes to complete.

(b) British Ability Scale (BAS) vocabulary. A 36-item test capturing children's emerging language skills. This involved children correctly identifying various objects or animals, such as a shoe, horse, jar or igloo.

While at age 5 the tests covered:

(c) BAS vocabulary. A 25-item test measuring children's verbal knowledge. The test took around 5 minutes to complete.

(d) BAS pattern construction. A non-verbal spatial reasoning assessment involving 23 items. This involved a pattern being presented to the child, who was then asked to repeat the pattern using plastic cubes. The test took around 15 minutes to complete.

(e) BAS picture similarities. A 23-question test of children's non-verbal reasoning skills. Children were shown a row of four pictures and were then asked to place a fifth card below one of the cards (the "stimulus card") which provided the best match. The test took around 8 minutes to complete.

Further details are available from CLOSER (2023). The correlation between children's scores on these five test scores are reported in Appendix D. These correlations are moderate (between approximately 0.2 and 0.6) indicating that – as is common with tests of young children – they are subject to a non-trivial amount of noise.

The five measures described above are used to operationalise our "high early achievement" group. Specifically, we standardise each of the five scores to mean zero and standard deviation one. The average across the five scores is then taken, with this then used to divide children into four equal groups (quartiles). The top quartile – the 25% of children that performed the best across the five tests – are the group of primary interest. Table 1 provides further information about the knowledge that this group displayed at age 3, relative to the average child.

<< Table 1 >>

The other key measure we utilise is information on household income. Across the first six MCS sweeps, parents were asked a battery of questions about their income from various sources, including earnings, self-employment and benefits. The survey organisers have then derived a measure of family income within each of the six sweeps. To create family income groups, we first standardise these six measures and then take the average across them. This is consistent with the economic literature on permanent family income, where families make spending and investment decisions (including regarding their children) based upon their long-run financial capacity (Francesconi & Heckman, 2016). The top/bottom quartile of this permanent family income measure is then used to define high/low-income groups.

The primary comparison we are then interested in is how the later lifetime outcomes of the following two groups compare:

- High-achieving children from high-income backgrounds (n = 1,392).
- High-achieving children from low-income backgrounds (n = 389).

With figures in brackets providing the relevant sample size. Table 2 presents a descriptive comparison of these groups[2].

<< Table 2 >>

3. Methodology – Kelley's paradox

Intuition

It is easy to underestimate the challenge of describing how future outcomes differ between initially high-achieving children from high- and low-income backgrounds. This is due to widespread underappreciation of how Kelley's paradox may impact the results.

As noted by Wainer & Brown (2007), Kelley's paradox emerges whenever one divides members of a given population into groups (e.g. identifying children of higher achievement levels) based upon a measure with less than perfect reliability. To understand why, consider Figure 1 below. This presents the hypothetical distribution of "true" early life ability for two groups of children – those from low-income backgrounds (left) and those from high-income backgrounds (right). Also plotted is a dashed vertical line, above which a child is classified as being of high ability (this has been placed at the average of the ability distribution for high-income children purely for ease of illustration). Note that, in this fictious example, around half

[2] These comparisons do not make any attempt to account for Kelley's paradox, which will be discussed in section 3. The intuition of this table is to provide a simple overview of how the groups of interest – as measured – compare.

of high-income children are of high true ability, compared to no child from a low-income background[3].

<< Figure 1 >>

Of course, one is unable to observe children’s “true ability”. Rather, one only observes measure(s) of their achievement in early-life tests. Such tests are subject to a degree of measurement error, making them less than perfectly reliable indicators of children’s early ability. The impact this has upon one’s data and subsequent inferences is illustrated in Figure 2. Separate charts are now provided for children from low-income (left) and high-income (right) backgrounds to help illustrate the key points. These charts now provide both the hypothetical true ability distribution (solid line) and the measured test score distribution (dashed line) of low- and high-income children, along with the threshold above which children are classified as “high-ability”.

<< Figure 2 >>

Starting with low-income children, note that some of this group get classified as being of high ability when they are not, due to their test score falling above the relevant threshold. Those low-income children who achieve a score above the threshold will have all experienced a large, positive random error draw (i.e. had an unusually good day / a lot of luck on the test(s)).

The graph on the right presents the same information for the high-income group[4]. Note that those high-income children who achieve a test score above the relevant threshold will have had varying degrees of luck (random error draws). For some, the random error will be large and positive, but for others it will be around zero (i.e. they were neither lucky or unlucky on the test), or even negative.

What this implies is that low-income children who managed to achieve a score above the high ability classification threshold will have had – ***on average*** - more luck (larger positive random error draws) than their high-income peers. In other words, despite all children above the threshold being classified as “high ability”, low- and high-income pupils who fall into this group will still differ in terms of their true ability. This holds true whenever the classification test is measured with error. It will then lead one to observe a “gap” in the later outcomes

[3] A somewhat extreme hypothetical example has been chosen here for ease of explanation.
[4] Note that the high-ability classification threshold is around the average of the high income group, as illustrated in Figure 1.

between initial high-ability children from low- and high-income backgrounds, even when no such difference really exists (or, rather, where this difference has been driven by unmeasured early-life ability rather than future events). The magnitude of this statistical artifact will be driven by two factors: (a) how much measurement error there is in the test(s) used to identify children of high initial ability[5] and (b) the strength of the association between the unmeasured part of their initial ability and the outcome of interest[6].

How much will their "true ability" differ by?

As noted by Wainer & Brown (2007), Kelley's paradox has been a long-known problem amongst statisticians, but one that has often been overlooked in empirical research. They, however, note how a simple formula can be used to estimate the true ability of different groups under different levels of test reliability:

$$\tau_i = \rho(x_i) + (1 - \rho).\mu_g \qquad (1)$$

Where:

τ = Child i's true initial ability.

ρ = The reliability of the test being used to classify children as high ability.

x_i = Child i's score on the test that is used to determine whether a child is high ability.

μ_g = The average test score of the income group (g) to which the child belongs (e.g. the average test score of children from low-income families).

Equation (1) thus "*tells us that the best estimate* [of a child's true ability] *is obtained by regressing the observed score in the direction of the mean score (μ) of the group that the examinee came from*." (Wainer & Brown, 2007). The amount of regression is determined by ρ – the reliability of the test used to classify children into ability groups. An important implication of this equation is that the true ability (τ) of low-income children with test scores above the high ability classification threshold will be lower than the average true ability of high-income children with test scores above the threshold.

[5] This will, in-turn, determine how much low- and high-income children classified as high-ability differ in their unmeasured true ability.

[6] Such associations will be particularly strong when the same test score is used in the future – giving rise to the phenomena of regression to the mean.

How can one correct estimates to account for Kelley's paradox?

A four-step procedure is used to correct estimates of the difference in future outcomes between initially high-achieving children from low- and high-income backgrounds, under different assumptions of ρ (test reliability).

To begin, equation (1) is used to estimate the true ability (τ) of each child in the dataset under an assumed value of ρ. The average value of τ is then calculated across (a) high-achieving children from low-income backgrounds and (b) high-achieving children from high-income backgrounds. The difference between these values can be estimated via the regression model:

$$\tau_i = \alpha + \delta . Ab\_Inc\_Grp_i + \varepsilon_i \qquad (2)$$

Where:

τ_i= Child i's true ability as estimated via the formula presented in equation (1).

$Ab\_Inc\_Grp_i$ = A set of dummy variables indicating whether child i is measured as a high-ability child from a high-income background (coded 0 = reference group), a high-ability child from a low-income background (coded 1), or a child not of high ability (coded 2).

ε_i = Random error term.

i = Child i.

The δ parameter from (2) then captures the difference in true ability between initially high-achieving children from low- and high-income backgrounds, under the assumed value of ρ.

In the second step, the strength of the association between τ and the outcome under investigation is estimated via a regression model:

$$O_i = \alpha + \gamma . \tau_i + \varepsilon_i \qquad (3)$$

Where:

O_i = The outcome one wishes to compare across high-ability children from low- and high-income backgrounds.

τ_i = The measure of true ability as estimated under equation (1) above.

The parameter of interest from this model is γ. This captures the strength of the link between true ability (τ) and the outcome of interest.

Next, we estimate the "raw" (uncorrected) difference in the outcome between high-ability children from low- and high-income backgrounds. This is, in essence, the magnitude of the difference in the outcome under the assumption that the test used to identify high-ability children is perfectly reliable (i.e. that ρ = 1), meaning Kelley's paradox is effectively ignored. The model used in this third step is thus:

$$O_i = \alpha + \beta.Ab\_Inc\_Grp_i + \varepsilon_i \qquad (4)$$

With all variables defined as under equations (2) and (3) above. The β parameter from this model captures the raw, uncorrected difference in the outcome between high-achieving children from high- and low-income backgrounds.

Finally, we adjust the value of β to account for Kelley's paradox – i.e. that the estimate of β from equation (4) will be too large as it will be partly capturing differences in unmeasured ability across income groups. This correction is of the form:

$$\widehat{\beta_\iota} - \widehat{\delta_\iota}.\widehat{\gamma_\iota} \qquad (5)$$

Using the parameters derived from equations (2), (3) and (4) above. Note that the term $\widehat{\delta_\iota}.\widehat{\gamma_\iota}$ is the product of the estimated difference in the unmeasured part of true ability across high-ability children from low- and high-income families ($\widehat{\delta_\iota}$) with how strongly true ability is associated with the outcome of interest ($\widehat{\gamma_\iota}$). If either of these terms are zero, then equation (5) reduces to $\widehat{\beta_\iota}$ – as estimated directly from equation (4).

The key parameter under this approach is ρ – the reliability of the test used to classify children as high initial ability. Although providers of tests often report their "reliability", what they actually provide is usually a measure of internal consistency (e.g. Cronbach's alpha) – i.e. how well children's answers to the different test questions correlate with one another. Ideally, one would rather base ρ upon the correlation in test scores achieved by the same group of children in two tests taken a short time apart (i.e. test-retest reliability). These tests may focus upon high achievement within a specified subject (e.g. mathematics) if one's primary interest is domain specific, or across various different tests if one is more broadly interested in a global indicator of high early ability. The closest available evidence for such inter-test correlations within the MCS (between the age 3 and age 5 measured used to define our high-ability group) are reported in Appendix D. As noted previously, these range between around 0.2 and 0.6, and thus appear subject to a non-trivial amount of noise.

The intuition behind us averaging scores across the five age 3/5 tests (as discussed previously in section 2) is to attempt to broadly measure high levels of early achievement while minimising the amount of noise (e.g. by averaging scores from across different tests that are taken at more than one time-point). Yet, given the well-known noise inherent in tests of young children, substantial measurement error is likely to remain (Goldstein et al, 2018). We hence investigate how our results vary under the following assumptions of ρ:

- $\rho = 1$. Initial high-ability children identified with perfect reliability (upper bound)[7].
- $\rho = 0.7$. Initial high-ability children identified with a moderate-degree of reliability (central estimate).
- $\rho = 0.5$. Initial high-ability children identified with low reliability (lower bound).

Finally, we note how sometimes one wishes to estimate the difference between high-ability children from low- and high-income backgrounds conditional upon other factors. Appendix B discusses this issue, providing an extension to the approach outlined above that can be used to produce estimates of such conditional differences. Likewise, in Appendix C we further extend the approach outlined above to consider more complex forms of measurement error. This includes (i) allowing the test scores of low-income children to be "noisier" (i.e. subject to more measurement error) than the scores of high-income children and (ii) assuming that there is systematic bias in the test scores (i.e. that the test scores of low-income children are artificially lower than the scores for high-income children).

4. Results

Research question 1. How do the cognitive skills and school grades of bright 5-year-olds from rich and poor backgrounds compare through to age 16?

To begin, Table 3 presents estimates of how future test scores and academic achievement outcomes differ between bright 5-year-olds from rich and poor backgrounds. Figures refer to standardised mean differences (effect sizes) for continuous outcomes, and percentage point differences for binary/categorical outcomes. Our discussion focuses upon our central estimates

[7] This is equivalent to assuming that the measure captures children's ability with perfect reliability, and that it does not contain any measurement error. It is the figure that most existing studies – which fail to adjust for Kelley's paradox – report.

(ρ = 0.7), though in the results tables we also present our upper and lower bounds. These are unconditional estimates, without the inclusion of any controls.

<< Table 3 >>

The first four rows of Table 3 refer to tests taken at ages 7 and 11. Overall, we find little clear evidence of a difference once Kelley's paradox has been accounted for; our central estimates are mostly small (typically less than 0.1 standard deviations) and fail to reach statistical significance at conventional thresholds. It thus seems that highly able 5-year-olds from poor backgrounds manage to keep pace academically with their more affluent peers through to the end of primary school (age 11).

A rather different pattern emerges, however, when one turns to results from high-stakes national examinations (GCSEs) taken at age 16[8]. While bright 5-year-olds from both rich and poor backgrounds are equally likely to achieve a "good pass" (C grade) in their mathematics and English language exams, they differ significantly in the likelihood of achieving a top A or A* grade. Take mathematics, for example. Our central estimate is that highly able children from affluent homes have (at least) 65% chance of achieving an A/A* grade, compared to 40% for their equally able, low-income peers. Indeed, we continue to observe a substantial 17 percentage point difference in achieving an A/A* grade even under our lower bound[9]. Thus, the evidence suggests that while highly-able 5-year-olds from poor backgrounds manage to keep pace academically with their rich peers through to age 11 (end of primary school), their performance – on average – tails off during secondary school.

A difference in future educational choices can also be observed in the bottom row of Table 3. Our central estimates indicate that highly able 5-year-olds from poor backgrounds are 21 percentage points less likely to be studying for A-Levels (or equivalent) at age 17 than their equally able, high-income peers. This is being largely driven by the differences in test scores and examination performance during secondary school discussed above. Indeed, once differences in GCSE exam performance have been controlled, the gap in taking A-Levels falls from 21 to 3 percentage points (estimates are not shown in Table 3).

[8] The equivalent qualifications are considered for children in Scotland.

[9] In additional analysis, we have found a difference of 20 percentage points in high-achieving rich and poor children achieving an A/A* mathematics grade after conditioning upon test scores measures at ages 7 and 11.

Consistent with the work of Crawford et al. (2017), Table 3 indicates that the first three years of secondary education (ages 11 to 14) are likely to be critical. In particular, note how the test score measure available at age 14 (capturing young people's verbal abilities) is the first point where a substantial difference in a cognitive outcome is observed (our central estimate putting the difference at 0.56 standard deviations)[10]. Although we are reliant upon data from a single test score at age 14, this nevertheless provides a first suggestion that the transition highly able young people from low-income backgrounds make into secondary school – and how they navigate this period of early adolescence – may be key.

Research question 2. To what extent do bright 5-year-olds from poor backgrounds become disinterested or unmotivated at school relative to their peers from more affluent backgrounds?

In Table 4 we turn to a factor that may be related to this relative decline in performance of highly-able 5-year-olds from poor backgrounds during early adolescence – their work ethic[11] and attitudes towards school. This includes responses to statements such as "*how often do you feel school is a waste of time*" and "*how important is it to you to work hard?*" (see Appendix A for further details). The results presented in panel (a) refer to the raw, unconditional estimates, while those in panel (b) are conditional upon test scores through to age 11.

<< Table 4 >>

The results in panel (a) illustrate how highly able 5-year-olds from poor homes consistently have a somewhat more negative attitude towards school from age 7 onwards. However, the difference is most apparent at age 14. In particular, note how the standardised mean difference increases from around 0.2 standard deviations at age 7/11 to around 0.4 standard deviations in the age 14 results. This is further emphasised by responses to the statement "*how important is it to you to work hard?*"; our central estimate is that at least 62% of highly able children from rich backgrounds felt that it was very important to work hard at age 14, compared to 51% of their equally able peers from poor backgrounds.

These results are re-iterated in panel (b), which includes controls for test scores of children at ages 7 and 11. These confirm that highly-able 5-year-olds with low-income parents – and who

[10] Although the difference at age 7 is statistically significant (0.16 standard deviations) the magnitude is relatively small.

[11] The results for age 14 work ethic are presented as percentage point differences, as this is based upon responses to a single question, rather than a scale formed from multiple questions.

go on to achieve the same test scores at ages 7 and 11 as their wealthier peers – have a significantly worse attitude to school and work ethic at age 14.

Research question 3. How do the socio-emotional outcomes of bright 5-year-olds from low-income backgrounds develop compared to their high-income peers?

Figure 3 turns to differences in children's socio-emotional outcomes as they age, as captured by (standardised) Strengths and Difficulty Questionnaire (SDQ) scores. These measure five aspects of children's behaviour (emotional symptoms, conduct problems, hyperactivity, peer problems, prosocial behaviour) through questions asking about (for example) whether the child "*often has temper tantrums or hot tempers*", "*is kind to younger children*" and "*has at least one good friend*". Figures along the vertical axis illustrate the gap in SDQ scores between bright 5-year-olds from rich and poor families in terms of an effect size, under our central assumption that $\rho = 0.7$. Note that these results are conditional upon SDQ scores at age 3, and hence capture how the initial gap in socio-emotional outcomes changes as children age.

<< Figure 3 >>

It appears that, throughout childhood and into early adolescence, the gap in socio-emotional outcomes between highly able rich and poor 5-year-olds progressively widens. The difference is less apparent at ages 5 and 7, where the effect size is relatively modest (~0.1 to 0.15 standard deviations) and on the boarder of statistical significance at the 5% level (indeed – as illustrated in Appendix E – our lower bound estimates at age 5 is not statistically significant at conventional thresholds). Yet the gap grows and becomes much more apparent at ages 11 (~0.25 standard deviations) and, particularly, age 14 (~0.35 standard deviations). This is consistent with the results presented in the previous sub-sections, which highlight how the first few years of secondary school is a period of particular concern for highly able children from low-income families.

Research question 4. Are bright 5-year-olds from poor backgrounds more likely to get into trouble as teenagers compared to 5-year-olds from rich backgrounds?

Building upon the analysis above, Table 5 turns to the issue of young people's behaviour. Panel (a) presents the raw, unconditional estimates, while panel (b) illustrates differences in teenage behaviour controlling for socio-emotional, behaviour and test score outcomes through to age 11.

<< Table 5 >>

Clear evidence emerges that highly able 5-year-olds from poor backgrounds go on to display significantly worse behaviour as 11-year-olds and through the teenage years than their equally able but more affluent peers. The estimates in panel (a) demonstrate how, by age 11, high-achieving low-income children have a weaker “moral compass” (e.g. less likely to believe it is wrong to vandalise, steal things or start a fight) and are more likely to engage in bad behaviour (e.g. been antisocial in public, stolen or damaged something). This continues through into the teenage years, though manifests in more serious ways. For instance, by age 17, approximately 27% of bright children from poor backgrounds have had some encounter with the police (ever been stopped, cautioned or arrested) compared to just 14% of their bright, high-income peers. This may well be related to the peer-group that they associate themselves with; Table 5 demonstrates a substantial 0.42 standard deviation socio-economic gap in high-ability children reporting that their friends get into trouble (e.g. drink, take drugs, get into trouble at school)[12].

Consistent with the results in the preceding sub-sections, panel (b) of Table 5 provides some evidence of an increasing socio-economic disparity between high-achieving children during the end of primary school and through to the end of secondary school. Even after controlling for differences in SDQ scores, moral compass, engagement in bad behaviour and test scores at age 11, highly able low-income children remain much more likely to report having a troublesome peer group (0.39 standard deviation) and having a brush with the law (12 percentage points) than their equally able but high-income peers.

Research question 5. Are bright 5-year-olds from poor backgrounds more likely to get bullied at school than bright 5-year-olds from richer backgrounds? How does this change during primary and secondary school?

Table 6 turns to the related issue of bullying, with estimates based upon children’s own reports in panel (a), and parental reports in panel (b). Note that different questions were asked to children and parents meaning their responses cannot be directly compared (see Appendix A for the precise wording of the questions used).

[12] Note that positive values on this scale indicate that the low-income group are more likely to report their friends get into trouble than their high-income peers.

<< Table 6 >>

Compared to previous sub-sections, evidence of a meaningful differences across high-achieving children from rich and poor backgrounds is more mixed. With respect to children's own responses in panel (a), most of our central estimates are small and/or statistically insignificant. The one exception is at age 11, when high-achieving low-income children are 8 percentage points more likely to say that "other children hurt or pick on them" at least once a month than the high-achieving high-income group. On the other hand, throughout primary school (up to age 11) low-income parents of bright 5-year-olds are substantially more likely to report that their child is "picked on or bullied" than high-income parents. Yet, interestingly, this difference shrinks from around 15 percentage points at age 11 to around five percentage points at age 14 and is no longer statistically significant at conventional thresholds.

Hence, overall, we conclude that there is some evidence that high-achieving disadvantaged children are more likely to experience bullying towards the end of primary school than equally-able high-income children. However, at other ages – particularly during secondary school – the results in this regard are inconclusive.

Research question 6. Are bright 5-year-olds from low-income backgrounds more likely to develop mental health problems than their equally able peers from high-income backgrounds?

To conclude, Table 7 considers how a selection of mental health and wellbeing outcomes compare across bright high and low-income children between the ages of 7 and 17.

<< Table 7 >>

During primary school (ages 7 and 11) differences in wellbeing outcomes appear relatively modest, and only occasionally reach statistical significance at conventional levels. Indeed, the lower bound for each of the age 11 wellbeing measures is around 0.1 or below, suggesting that any difference across high-achieving children from different income groups at these ages are likely to be relatively small.

The situation at age 14 is rather different. Now, high-ability low-income children score ~0.3 standard deviations lower on the happiness scale (e.g. how happy they are with their friends, schoolwork, life as a whole), ~0.45 standard deviations lower on the feelings scale (e.g. that they hated themselves, cried a lot, felt miserable or unhappy) and ~0.5 standard deviations lower on the self-esteem scale (e.g. that they felt good about themselves, they could do things as well as others, that they were a person of value). Indeed, similar substantive findings emerge across our upper and lower bounds. The age 11 to 14 period – encompassing the transition into secondary school – thus seems to be a critical time when many high-achieving children from low-income backgrounds suffer a sharp deterioration in their wellbeing and mental health in comparison to their high-income peers.

Interestingly, of the measures available, we only continue to observe a consistent, sizeable and statistically significant difference across income groups in self-esteem at age 17 (0.4 standard deviations). The difference in mental wellbeing scores at this age is comparatively small (0.11 standard deviations) and not statistically significant. For the Kessler scale – which asks young people questions such as whether they felt depressed, hopeless or restless over the last month – the results are somewhat inconclusive. Thus, while we consistently find high-achieving disadvantaged children to have lower-levels of self-esteem as teenagers than their high-income peers, it is only during early adolescence (age 14) we find convincing evidence of a sizable gap across a wide array of mental wellbeing measures.

5. Conclusions

Social mobility has been near the top of the political agenda in the UK since the turn of the 21st century. Successive governments have sought to improve the later lifetime outcomes of individuals from disadvantaged socio-economic backgrounds, with a particular focus on boosting this group's educational achievement at school. Enhancing the prospects of bright children from low-income families is a key part of creating a more socially-fluid society; this group should have amongst the best chances of breaking through the glass ceiling, having the raw potential to develop the necessary skills to access a high-status university and enter a professional job. Yet these children also have a long way to go before their potential gets fulfilled, needing to successfully navigate difficult periods such as adolescence and young adulthood before entering the world of work.

This paper has thus presented new evidence on how the lives of bright 5-year-olds from poor backgrounds develop through to age 17, and how this compares to their more affluent peers. In doing so, it provides perhaps the most detailed analysis on this group's development throughout childhood, both internationally and in the UK. We find that the cognitive skills of high-ability low-income children keep pace with their more affluent peers through to the end of primary school (age 11), but then sharply diverge during secondary education. While they are still almost certain to achieve a good pass (C grade) in key GCSE subjects (English Language and mathematics), many bright children from poor backgrounds fail to push on to achieve the highest grades – those which are desired by elite universities and the most prestigious employers.

The early part of secondary school (ages 11 to 14) is found to be a particularly important period for this group. Our analysis has shown how this coincides with a clear, rapid decline in high-ability low-income children's outcomes relative to equally able children from more affluent homes. This includes developing a more negative attitude towards school, increasingly troublesome behaviour, falling into a poorly behaved friendship group and declining levels of mental health. It is hence likely that such challenges combine to lead school performance – and ultimately their high-stakes examination grades – to decline.

How do these results compare to those within the existing literature? Our finding of no decline in high-achieving disadvantaged children's cognitive skills relative to their more affluent peers during primary school is consistent with – and builds upon – the work of Jerrim & Vignoles (2013). While they found little evidence of divergence in these groups' skills between ages 3 and 7, we extend this through to age 11, before illustrating how a notable gap then develops thereafter. We have also built upon the research of Crawford et al. (2017) who "*identified early secondary as a key transition period*" with respect high-ability disadvantaged children's academic achievement at school. Our research has first confirmed and then enriched this picture by demonstrating the much broader challenges highly able low-income children face during the critical age 11-14 period. In doing so, we provide important new evidence regarding the potential correlates and consequences of the relative decline of high-ability low-income children's academic achievement and skills.

We also recognise that our work has its limitations, with five specific issues standing out. First, sample sizes for our group of interest are relatively modest, meaning it has not been possible to explore further differences across sub-groups (e.g. high achieving disadvantaged boys compared to girls or high achievers from different ethnic backgrounds). Second, as with all studies in this line of research, only a limited array of test score data are available to identify low-income children with strong early-life skills. Although we draw upon five separate tests taken at two different ages, we recognise that those who fall into the top quartile may to some extent vary depending on the precise measures used. Third, relatedly, we have discussed how it is important to consider and adjust one's estimates for Kelley's paradox when drawing comparisons across high achieving children from different income groups. The adjustments we have made to our estimates assume that measurement error is classical, essentially meaning that early-life test scores capture children’s abilities with a degree of random noise. It is however possible that the measurement error takes a different form. We discuss this issue at length in Appendix C, including considering how this issue may impact upon our results. Fourth, the latest available MCS sweep captures information about individuals at age 17, before they have entered university and the labour market. Future research should seek to extend our analysis to consider how high-ability young people from rich and poor backgrounds navigate the transition into adulthood. Finally, the analyses we have presented in this paper are descriptive and are not intended to establish cause and effect.

What, then, do our results imply for public policy? Unfortunately, evidence about “what works” for disadvantaged high-attainers remains rather limited. Given this, it seems advisable that organisations such as the Education Endowment Foundation in England and What Works Clearing House in the United States conduct rapid evidence reviews investigating the interventions and policies that could help meet this group’s particular needs. At the same time, such organisations should also consider how interventions designed to target disadvantaged high achievers can be implemented and evaluated, given the small number likely to be found within individual schools. In England, the Department for Education and Social Mobility Commission might also consider how further evidence on this important group can be generated in the future. In England, the reception baseline test – taken by children within the first six weeks of starting school – could be a particularly valuable resource, assuming it adequately measures the skills of children towards the top end of the distribution. To provide further evidence on the development of high-ability disadvantaged children, these data should be made available for research purposes. Likewise, schools should be regularly collecting information from young people – particularly during the transition into secondary school – to ensure they are able to quickly respond to the challenges faced by their high-achieving low-income cohorts.

References

Bradbury, B., Corak, M., Waldfogel, J., & Washbrook, E. (2015). *Too Many Children Left Behind: The U.S. Achievement Gap in Comparative Perspective*. Russell Sage Foundation. http://www.jstor.org/stable/10.7758/9781610448482

Bradbury, B., Waldfogel, J., & Washbrook, E. (2019). Income-related gaps in early child cognitive development: why are they larger in the United States than in the United Kingdom, Australia, and Canada? *Demography 56*(1): 367–390. https://doi.org/10.1007/s13524-018-0738-8

CLOSER. (2023). Cognitive measures in the Millenium Cohort Study. https://closer.ac.uk/cross-study-data-guides/cognitive-measures-guide/mcs-cognition/.

Crawford, C., Macmillan, L., & Vignoles, A. (2017). When and why do initially high-achieving poor children fall behind? *Oxford Review of Education*, *43*(1), 88–108. https://doi.org/10.1080/03054985.2016.1240672

Cunha, F. Heckman, J. Lochner, L and Masterov, D. (2006). 'Interpreting the Evidence on Life Cycle Skill Formation'. In *Handbook of the Economics of Education* (eds E. Hanushek and F. Welch), pp. 697-812. Amsterdam: Holland North.

Feinstein, L. (2003). Inequality in the Early Cognitive Development of British Children in the 1970 Cohort, *Economica* 70: 73-97.

Francesconi, M., & Heckman, J. J. (2016). Child Development and Parental Investment: Introduction. *The Economic Journal* 126(596): F1–F27. doi:10.1111/ecoj.12256

Goldstein, H.; Moss, G.; Sammons, P.; Sinnott, G. & Stobart, G. (2018). *A baseline without basis the validity and utility of the proposed reception baseline assessment in England*. London: British Educational Research Association. https://www.bera.ac.uk/researchers-resources/ publications/a-baseline-without-basis

Heckman J. J. (2008). Schools, Skills, and Synapses. *Economic inquiry 46*(3): 289. https://doi.org/10.1111/j.1465-7295.2008.00163.x

Holt-White, E. & Cullinane, C. (2023). *Social Mobility: The Next Generation Lost potential at age 16*. London: Sutton Trust. https://www.suttontrust.com/our-research/social-mobility-the-next-generation-lost-potential-at-age-16/#:~:text=This%20report%20looks%20at%20the,the%20end%20of%20primary%20school

Institute for Fiscal Studies. (2023). Inequality. The IFS Deaton Review. https://ifs.org.uk/inequality/directory/

Jerrim, J., & Vignoles, A. (2013). Social mobility, regression to the mean and the cognitive development of high ability children from disadvantaged homes. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 176: 887–906.

Loft, P., & Danechi, S. (2020). Support for more able and talented children in schools (UK). House of Commons Briefing Paper Number 9065. https://dera.ioe.ac.uk/id/eprint/37230/1/CBP-9065%20Redacted.pdf

OECD. (2015). Learning contexts that drive skill formation. In *Skills for Social Progress: The Power of Social and Emotional Skills*. OECD Publishing, Paris. https://doi.org/10.1787/9789264226159-7-en

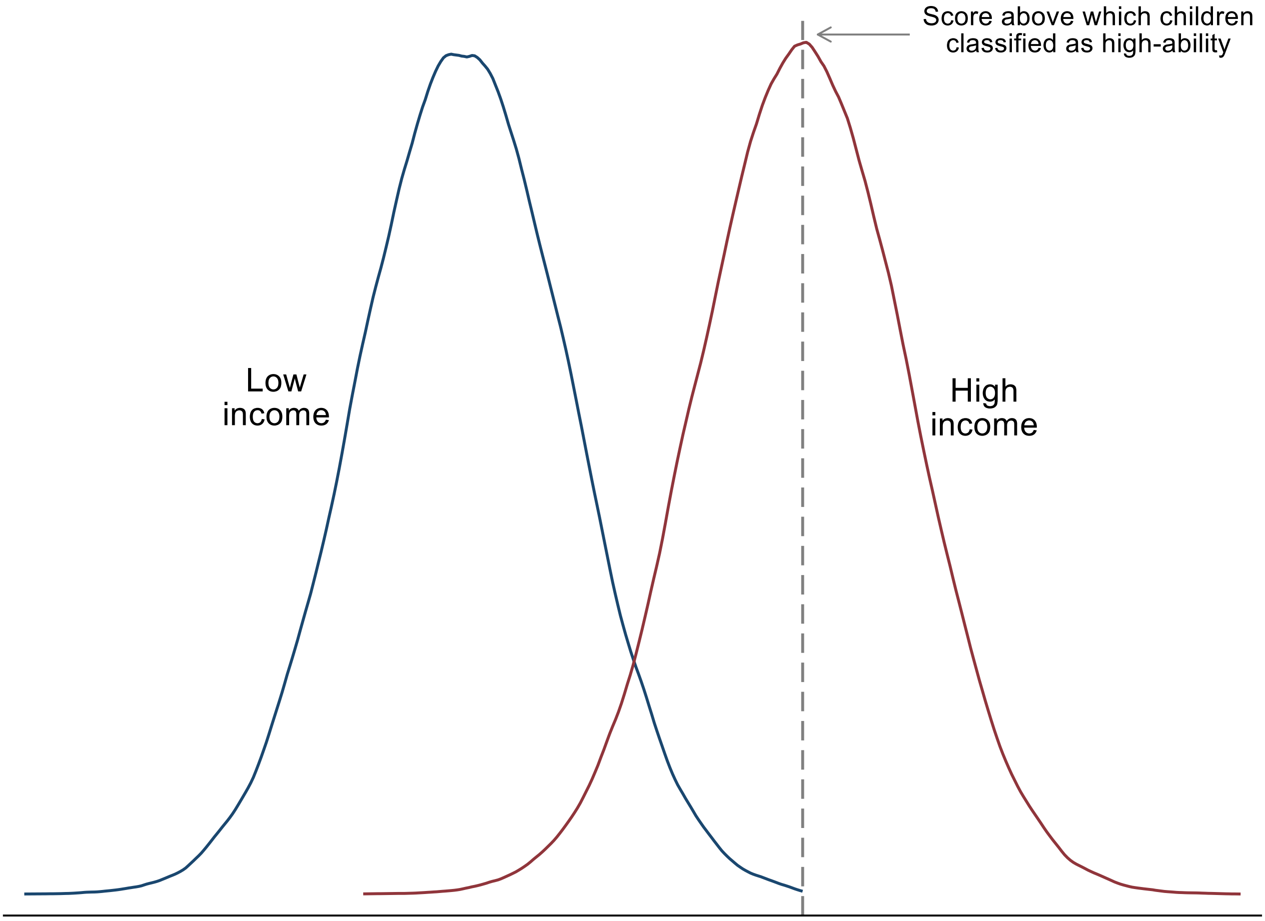
Plewis, I. (2007). Millennium Cohort Study First Survey: Technical Report on Sampling. https://cls.ucl.ac.uk/wp-content/uploads/2017/07/Technical-Report-on-Sampling-4th-Edition-August-2007.pdf

Stewart, K. & Waldfogel, J. (2017). *Closing gaps early. The role of early years policy in promoting social mobility in England*. London: Sutton Trust. https://www.suttontrust.com/wp-content/uploads/2019/12/Closing-Gaps-Early_FINAL.pdf

Wainer, H., & Brown, L. M. (2006). Three Statistical Paradoxes in the Interpretation of Group Differences: Illustrated with Medical School Admission and Licensing Data. In C. R. Rao & S. Sinharay (Eds.), Handbook of Statistics (Vol. 26, pp. 893-918). Elsevier. ISSN 0169-7161. ISBN 9780444521033. https://doi.org/10.1016/S0169-7161(06)26028-0.

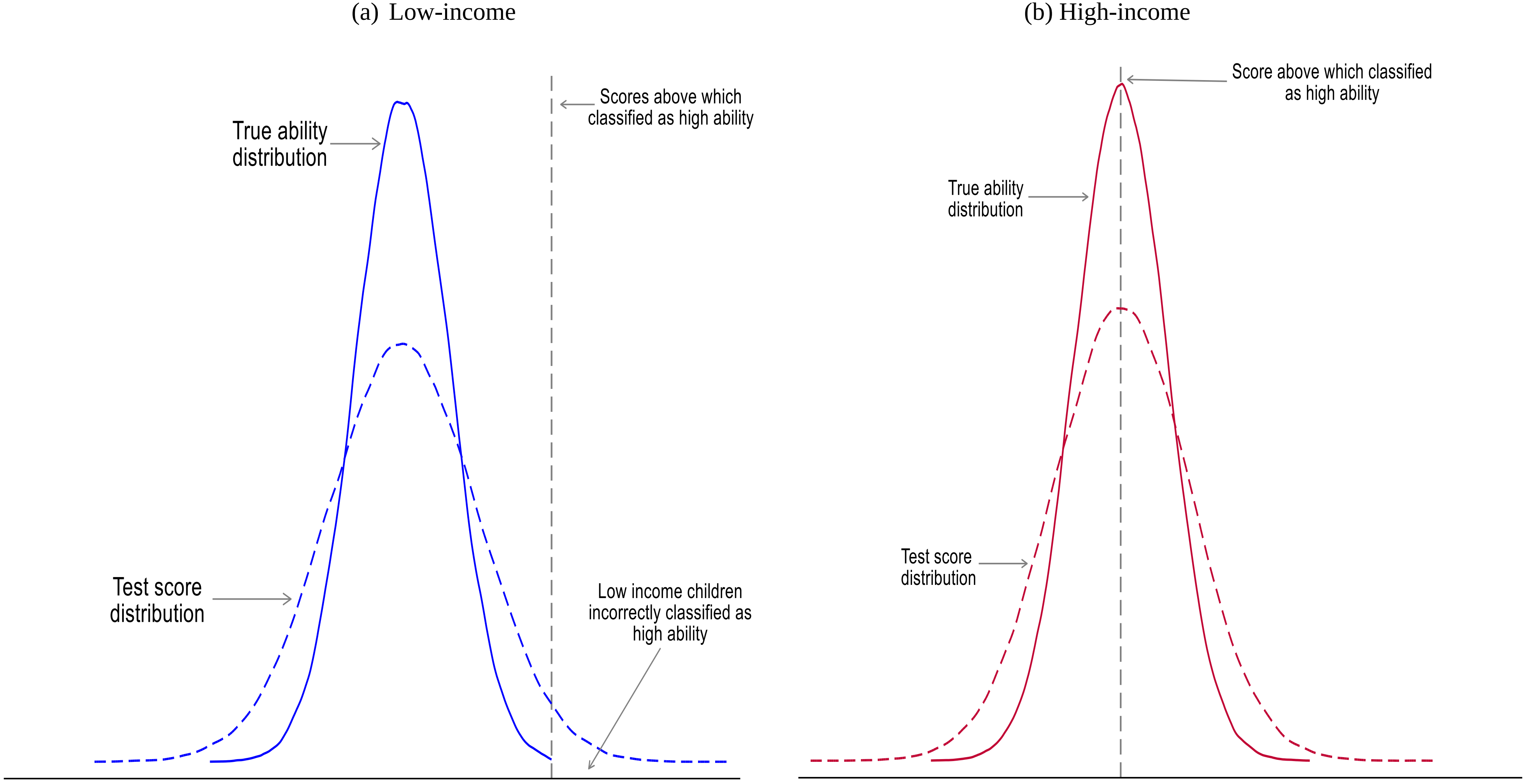
Wakeling, P., & Savage, M. (2015). Entry to Elite Positions and the Stratification of Higher Education in Britain. *The Sociological Review* 63(2): 290–320. https://doi.org/10.1111/1467-954X.12284

Figure 1. Hypothetical example of differences in true ability between children from low- and high-income backgrounds



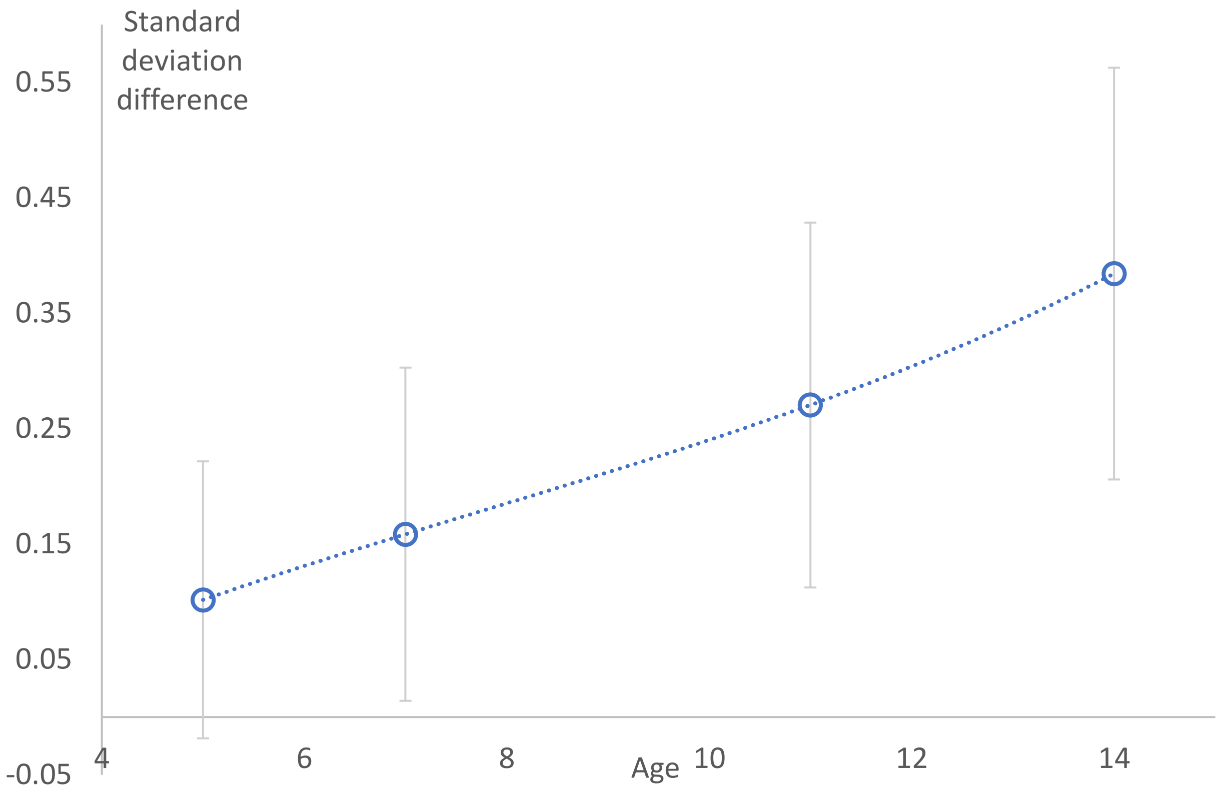
Notes: Hypothetical simulated data. Vertical grey dashed line indicates the point above which a child is classified as "high ability" based upon their test scores. Blue distribution on the left refers to low-income children. Red distribution on the right refers to high-income children. In this hypothetical example, no low-income children have true-ability above the high-ability threshold, while half of high-income children do.

Figure 2. Hypothetical example of differences in true ability between children from low- and high-income backgrounds



Notes: Hypothetical simulated data. Vertical grey dashed line indicates the point above which a child is classified as "high ability" based upon their test scores. Solid distribution refers to the (unobserved) true ability distribution. The dashed distribution refers to the (observed) test score distribution.

Figure 3. The difference in SDQ scores between initially high-achieving children from high- and low-income backgrounds. Change with age.



Notes: Figures refer to the difference in total SDQ scores between initially high-achieving children from high- and low-income backgrounds. Estimates are conditional upon differences in SDQ scores at age 3. These refer to our central estimates where we assume test reliability of 0.7. SDQ scores have been standardised to mean zero and standard deviation one at all ages. See Appendix E for all estimates. Capped vertical lines around the point estimates represent the estimated 95% confidence interval.

Table 1. Performance of children with low, average and high early-life test scores in selected test questions asked at age 3.

	Low scores	Average	High scores
Can recognise nine bees	32%	41%	52%
Can recognise the number three	19%	36%	56%
Can recognise depth of water	26%	42%	56%
Can tell which pair of shoes match	23%	32%	43%
Can recognise the letter A	17%	21%	29%
Can recognise the letter X	12%	23%	38%
Recognise which shape is a diamond	18%	45%	63%
Recognise which shape is an oval	18%	36%	50%

Notes: Figures refer to the percent of children who were asked the question and answered it correctly. These questions were asked as part of the Bracken school readiness test taken during the age 3 wave of the MCS. Children were on average 38 months old when taking the test. Low (high) scores refers to child who were in the bottom (top) quartile of the distribution of the age 3 / 5 tests.

Table 2. Background characteristics of children with high early-life test scores by income group

	Low-income		High-income	
	All pupils	High test scores	All pupils	High test scores
White ethnicity	74%	84%	93%	94%
Mother age at birth	24.9	25.5	32.5	32.4
Father age at birth	28.9	28.6	34.3	34.2
Mother verbal ability scores	-0.53	-0.11	0.87	1.06
Father verbal ability scores	-0.76	-0.28	0.66	0.77
Mother mental health (Kessler) score	-0.39	-0.34	0.25	0.28
Father mental health (Kessler) score	-0.60	-0.43	0.16	0.20
SDQ total scores age 3	-0.52	-0.13	0.40	0.51

Notes: All scale scores have been standardised to mean 0 and standard deviation 1, with lower values indicating worse outcomes.

Table 3. Difference in future educational outcomes between highly able children from low-and high-income backgrounds

	Upper bound (rho = 1)	Central estimate (rho = 0.7)	Corrected difference (rho = 0.5)	Standard error
Age 7 Word reading	-0.42*	-0.16*	0.04	0.07
Age 7 Pattern construction	-0.31*	-0.03	0.16*	0.06
Age 7 Mathematics	-0.34*	-0.07	0.11	0.07
Age 11 verbal similarities	-0.30*	-0.06	0.12	0.07
Age 14 Word activity	-0.78*	-0.56*	-0.39*	0.08
% GCSE maths grade A	-37%*	-26%*	-17%*	5.6%
% GCSE maths grade C	-7%*	1%	7%*	3.4%
% GCSE English grade A	-31%*	-21%*	-13%*	5.4%
% GCSE English grade C	-9%*	-4%	1%	4.6%
% A-Levels age 17	-33%*	-21%*	-10%	5.8%

Notes: The upper bound refers to where one assumes there is no measurement error in the test used to identify high-ability children. Our central estimate assumes test reliability of 0.7. Figures refer to the difference in the outcome between high-ability children from high-income backgrounds and high-ability children from low-income backgrounds. * indicates difference statistically significant at the 5% level. Figures refer to effect sizes or percentage point differences.

Table 4. Difference in attitudes towards school between highly able children from low- and high-income backgrounds

(a) Unconditional

	Upper bound (rho = 1)	Central estimate (rho = 0.7)	Lower bound (rho = 0.5)	Standard error
Age 7 attitude to school	-0.27*	-0.19*	-0.12	0.08
Age 11 attitude to school	-0.27*	-0.21*	-0.16*	0.08
Age 14 attitude to school	-0.43*	-0.38*	-0.32*	0.10
Age 14 work ethic	-13%*	-11%*	-9%*	4%

(b) Conditional (age 14)

	Upper bound (rho = 1)	Central estimate (rho = 0.7)	Lower bound (rho = 0.5)	Standar d error
Age 14 attitude to school - conditional	-0.46*	-0.44*	-0.38*	0.10
Age 14 work ethic - conditional	-14%*	-13%*	-12%*	5%

Notes: The upper bound refers to where one assumes there is no measurement error in the test used to identify high-ability children. Our central estimate assumes test reliability of 0.7. Figures refer to the difference in the outcome between high-ability children from high-income backgrounds and high-ability children from low-income backgrounds. * indicates difference statistically significant at the 5% level. Figures refer to effect sizes or probability differences. Figures in panel (b) are conditional upon achievement during primary school at ages 7 and 11.

Table 5. Difference in poor behaviour between highly able children from low-income backgrounds.

(a) Unconditional

	Upper bound (rho = 1)	Central estimate (rho = 0.7)	Lower bound (rho = 0.5)	Standard error
Age 11. Engaged in bad behaviour	16%*	13%*	10%*	4%
Age 11. Moral compass	-0.38*	-0.30*	-0.23*	0.09
Age 14. Engaged in bad behaviour	0.27*	0.25*	0.22*	0.11
Age 14. Moral compass	-0.19*	-0.19*	-0.17*	0.09
Age 14. Troublesome peer-group	0.41*	0.42*	0.41*	0.12
Ever trouble with police	15%*	13%*	10%*	5%

(b) Conditional

	Upper bound (rho = 1)	Central estimate (rho = 0.7)	Lower bound (rho = 0.5)	Standard error
Engaged bad behaviour (age 14)	10%	11%	11%	10%
Age 14. Moral compass	-0.16	-0.16	-0.15	0.10
Age 14. Troublesome peer-group	0.36*	0.39*	0.40*	0.14
Ever trouble with police	13%*	12%*	11%*	5%

Notes: The upper bound refers to where one assumes there is no measurement error in the test used to identify high-ability children. Our central estimate assumes test reliability of 0.7. Figures refer to the difference in the outcome between high-ability children from high-income backgrounds and high-ability children from low-income backgrounds. * indicates difference statistically significant at the 5% level. Figures refer to effect sizes. Conditional estimates control for SDQ scores, moral compass scale, engagement in bad behaviour and verbal similarities test scores at age 11.

Table 6. Difference in experiences of being bullied between highly able children from low-income backgrounds.

(a) Child report

	Upper bound (rho = 1)	Central estimate (rho = 0.7)	Lower bound (rho = 0.5)	Standard error
Bullied age 7 (child report)	3%*	1%	-1%	2%
Left out at school age 7 (child report)	5%*	2%	0%	2%
Bullied age 11 (child report)	9%*	8%*	8%	4%
Bullied age 14 (child report)	-2%	-2%	-2%	3%
Bullied age 17 (child report)	9%	7%	6%	5%

(b) Parent report

	Upper bound (rho = 1)	Central estimate (rho = 0.7)	Lower bound (rho = 0.5)	Standard error
Age 3 (parent report)	9%*	7%*	5%*	2%
Age 5 (parent report)	17%*	14%*	11%*	3%
Age 7 (parent report)	21%*	17%*	14%*	4%
Age 11 (parent report)	19%*	15%*	12%*	4%
Age 14 (parent report)	10%*	6%	3%	4%
Age 17 (parent report)	10%	7%	4%	6%

Notes: The upper bound refers to where one assumes there is no measurement error in the test used to identify high-ability children. Our central estimate assumes test reliability of 0.7. Figures refer to the difference in the outcome between high-ability children from high-income backgrounds and high-ability children from low-income backgrounds. * indicates difference statistically significant at the 5% level. Figures refer to differences in the probability of the child being bullied; positive values indicate high-ability children from low-income backgrounds are more likely to be bullied than their high-ability high-income peers.

Table 7. Difference in mental health outcomes between highly able children from low-income backgrounds.

	Upper bound (rho = 1)	Central estimate (rho = 0.7)	Lower bound (rho = 0.5)	Standard error
Feelings (age 7)	-0.25*	-0.21*	-0.19	0.10
Feelings (age 11)	-0.09	-0.05	-0.02	0.09
Self-esteem (age 11)	-0.21*	-0.17	-0.14	0.10
Happiness (age 11)	-0.17*	-0.13	-0.09	0.08
Feelings (age 14)	-0.44*	-0.45*	-0.43*	0.10
Self-esteem (age 14)	-0.52*	-0.52*	-0.50*	0.09
Happiness (age 14)	-0.41*	-0.32*	-0.23*	0.10
Kessler (age 17)	-0.21	-0.20	-0.18	0.14
Mental wellbeing (age 17)	-0.14	-0.11	-0.09	0.12
Self-esteem (age 17)	-0.42*	-0.40*	-0.37*	0.14

Notes: The upper bound refers to where one assumes there is no measurement error in the test used to identify high-ability children. Our central estimate assumes test reliability of 0.7. Figures refer to the difference in the outcome between high-ability children from high-income backgrounds and high-ability children from low-income backgrounds. * indicates difference statistically significant at the 5% level. Figures refer to effect sizes. Lower values indicate worse outcomes on all scales.

Appendix A. Full list of questions used to construct each outcome scale

Cognitive test score measures included in Table 3.

Age 7 Word reading. A measure capturing children's ability to decode and read single words. Test length of around 5 minutes. For further information see https://closer.ac.uk/cross-study-data-guides/cognitive-measures-guide/mcs-cognition/mcs-age-7-bas-ii-word-reading/

Age 7 Pattern construction. A measure of children's non-verbal spatial ability, capturing their visual processing skills. The test typically took 10 minutes to complete. For further information see https://closer.ac.uk/cross-study-data-guides/cognitive-measures-guide/mcs-cognition/mcs-age-7-bas-ii-pattern-construction/

Age 7 Mathematics. An adapted version of the Progress in Maths test, designed to measure children's quantitative knowledge. Children were asked to complete up to 20 questions. For further information see https://closer.ac.uk/cross-study-data-guides/cognitive-measures-guide/mcs-cognition/mcs-age-7-nfer-progress-in-maths-adapted/

Age 11 verbal similarities. A measure of children's verbal knowledge and reasoning, involving around 12 test questions. For further information see https://closer.ac.uk/cross-study-data-guides/cognitive-measures-guide/mcs-cognition/mcs-age-11-bas-ii-verbal-similarities/

Age 14 Word activity. A test of children's vocabulary, including understanding the meaning of words. It took approximately four minutes to complete. For further information, see https://closer.ac.uk/cross-study-data-guides/cognitive-measures-guide/mcs-cognition/mcs-age-14-apu-vocabulary-test-applied-psychology/

Attitudes towards school in Table 4.

Age 7 attitudes towards school

- How often do you behave well in class?
- How often do you get fed up at school?
- How often is school interesting?
- How often do you feel unhappy at school?
- How often do you try to do your best at school?

Age 11 attitudes towards school

- How often do you misbehave or cause trouble in class?
- How often do you get tired at school?
- How often do you feel school is a waste of time?
- How often do you try your best at school?
- How often do you find school interesting?
- How often do you try your best at school?

Age 14 attitudes towards school

- How often do you misbehave or cause trouble in lessons?
- How often do you feel school is a waste of time?
- How often do you get tired at school?
- How often do you feel unhappy at school?
- How often do you find school interesting?
- How often do you try your best at school?

Age 14 work ethic

- How important is it to you to work hard?

Indicators of poor behaviour in Table 5.

Age 11. Engaged in bad behaviour

- Have you ever been noisy or rude in a public place?
- Have you ever taken something from a shop without paying for it?
- Have you ever written things or sprayed paint on a building?
- Have you ever on purpose damaged anything in a public place?

Age 11. Moral compass

- How wrong it is for someone your age to start a fight?
- How wrong it is for someone your age to graffiti?
- How wrong it is for someone your age to take some thing?
- How wrong it is to copy or download music?

Age 14. Engaged in bad behaviour

- In the last 12 months have you stolen something from someone.
- In the last 12 months have you written things or spray painted on a building, fence or train or anywhere else where you shouldn't have?
- In the last 12 months have you on purpose damaged anything in a public place that didn't belong to you, for example by burning, smashing or breaking things like cars, bus shelters and rubbish bins?
- Have you ever carried a knife or other weapon for your own protection because someone else asked you to or in case you get into a fight?
- In the last 12 months have you pushed or shoved/hit/slapped/punched someone?
- In the last 12 months have you used or hit someone with a weapon?
- In the last 12 months have you stolen something from someone. e.g. a mobile phone, money etc?

Age 14. Moral compass

- How wrong do you think it is for someone your age to start a fight with someone?
- How wrong do you think it is for someone your age to write things or spray paint on a building, fence or train?
- How wrong do you think it is for someone your age to take something from a shop without paying for it?
- How wrong do you think it is for someone your age to copy or download music, games or films without paying for them, when they should have done?

Age 14. Troublesome peer-group

- How many of your close friends work hard at school?
- How many of your close friends get into a lot of trouble at school?
- How many of your friends drink alcohol?
- Do any of your friends take cannabis (weed) or any other illegal drugs?

Ever trouble with police

Responding "yes" to any of the below at *either* ages 14 or 17:

- Have you ever been stopped and questioned by the police?

- Have you ever been given a formal warning or caution by a police officer?
- Have you ever been arrested by a police officer and taken to a police station?

Indicators of poor behaviour in Table 6

Bullied age 7 (child report)

- How often do other children bully you?

Left out at school age 7 (child report)

- How often do you feel left out of things by other children?

Bullied age 11 (child report)

- How often do other children hurt you or pick on you on purpose?

Bullied age 14 (child report)

- How often do **other children** hurt you or pick on you on purpose?

Bullied age 17 (child report)

- Other children or young people pick on me or bully me

Parent report of the child being bullied (all ages)

Question from the Strengths and Difficulties Questionnaire (SDQ) – whether the child us "Picked on or bullied by other children" (Not true, somewhat true, certainly true)

Indicators of mental health and wellbeing in Table 7

Feelings (age 7)

- How often do you feel happy?
- How often do you get worried?
- How often do you laugh?
- How often do you feel sad?
- How often do you lose your temper?

Feelings (age 11)

- In the last 4 weeks, how often did you feel happy?
- In the last 4 weeks, how often worried about what would happen?
- In the last 4 weeks, how often did you feel sad?
- In the last 4 weeks, how often did you feel afraid or scared?
- In the last 4 weeks, how often did you laugh?
- In the last 4 weeks, how often did you get angry?

Self-esteem (age 11)

- On the whole, I am satisfied with myself.
- I feel that I have a number of good qualities.
- I am able to do things as well as most other people.
- I am a person of value.
- I feel good about myself.

Happiness (age 11)

On a scale of 1 to 7 where '1' means completely happy and '7' means not at all happy, how do you feel about…

- ….Your school work
- ….The way you look
- ….Your family
- ….Your friends
- ….The school you go to
- ….Your life as a whole

Feelings (age 14)

The next few questions are about how you have been feeling or acting recently. For each question please select the answer which reflects how you have been feeling or acting in the past two weeks. (1. Not true, 2. Sometimes, 3. True).

- I felt miserable or unhappy
- I didn't enjoy anything at all
- I felt so tired I just sat around and did nothing
- I was very restless

- I felt I was no good any more
- I cried a lot
- I found it hard to think properly or concentrate
- I hated myself
- I was a bad person
- I felt lonely
- I thought nobody really loved me
- I thought I could never be as good as other kids
- I did everything wrong

Self-esteem (age 14)

- On the whole, I am satisfied with myself.
- I feel that I have a number of good qualities.
- I am able to do things as well as most other people.
- I am a person of value.
- I feel good about myself.

Happiness (age 14)

On a scale of 1 to 7 where ‘1’ means completely happy and ‘7’ means not at all happy, how do you feel about the following parts of your life?

- Your school work?
- The way you look?
- Your family?
- Your friends?
- The school you go to?
- Your life as a whole?

Kessler (age 17)

- During the last 30 days, about how often did you feel so depressed that nothing could cheer you up?
- During the last 30 days, about how often did you feel hopeless?
- During the last 30 days, about how often did you feel restless or fidgety?
- During the last 30 days, about how often did you feel that everything was an effort?
- During the last 30 days, about how often did you feel worthless?
- During the last 30 days, about how often did you feel nervous?

Mental wellbeing (age 17)

Below are some statements about feelings and thoughts. Please select the answer that best describes your experience of each over the last two weeks.

- I’ve been feeling optimistic about the future
- I’ve been feeling useful
- I’ve been feeling relaxed
- I’ve been dealing with problems well
- I’ve been thinking clearly
- I’ve been feeling close to other people
- I’ve been able to make up my own mind about things

Self-esteem (age 17)

- On the whole, I am satisfied with myself.
- I feel that I have a number of good qualities.
- I am able to do things as well as most other people.
- I am a person of value.
- I feel good about myself.

Appendix B. Extending the methodological approach to control for covariates

In section 3 of the paper we outlined our approach for correcting estimates for Kelley's paradox. Recall that the reason why this problem emerges is that measurement error inherent in test scores means that there will remain some unaccounted-for differences in the abilities between "high achieving" (top test score quartile) children from high- and low-income backgrounds. The intuition behind the approach presented in the main text is that we adjust estimates for differences in this unmeasured ability, under different assumptions of the amount of measurement error presence.

In this appendix we extend this approach to allow for the inclusion of covariates. The challenge of introducing controls is that they will themselves "soak up" some of the difference in unmeasured ability between high-achieving advantaged and disadvantaged children. This needs to be taken into account when an adjustment for Kelley's paradox is made.

The first point to consider is then how large is the difference in "true ability" (τ) between high- and low-income children who fall into the top test score quartile, over and above the part that is explained by the control(s)? This can be estimated by a revised version of equation (2), which becomes:

$$\tau_i = \alpha + \delta . Ab\_Inc\_Grp_i + \Omega . X_i + \varepsilon_i \qquad \text{(2b)}$$

Where:

X_i = A vector of background controls the analyst wishes to include in the model.

With all other variables defined as under equation (2) above.

The parameter of interest from this model is $\widehat{\delta_i}$ as previously – which now captures the difference in true ability (τ_i) across high- and low-income groups, over and above the part explained by the controls.

An analogous change is made to equation (3) investigating the association between true ability (τ_i) and the outcome of interest. In particular, we are now interested in the strength of this association over and above the part explained by the controls:

$$O_i = \alpha + \gamma . \tau_i + \Omega . X_i + \varepsilon_i \qquad \text{(3b)}$$

With the parameter of interest from equation (3b) continuing to be the parameter $\widehat{\gamma_\iota}$.

Finally, equation (4) – estimating the unadjusted difference between high ability high and low income groups – is also revised to include the controls:

$$O_i = \alpha + \beta.Ab\_Inc\_Grp_i + \Omega.X_i + \varepsilon_i \qquad \text{(4b)}$$

With the $\widehat{\beta_\iota}$ parameter now capturing the difference in the outcomes (over and above the controls) between high- and low-income groups under the assumption that the test scores used to divide children into ability groups is measured with perfect reliability.

The $\widehat{\delta_\iota}$ $\widehat{\gamma_\iota}$ and $\widehat{\beta_\iota}$ parameters form equations 2b, 3b and 4b above are then plugged into equation (5) presented in section 3 as previously. This then provides estimates of the *conditional* difference in outcomes between high-achieving low- and high-income groups, correcting for Kelley's paradox under a stated assumption of ρ.

Appendix C. Kelley’s paradox in the context of non-classical measurement error

In the main body of the paper we discussed Kelley’s paradox and suggested an approach to correct comparisons between high-achieving children from high and low income backgrounds. The approached used essentially assumes that measurement error is “classical” – that children’s abilities are subject to random noise and thus have less than perfect reliability. In this appendix we extend the approach to consider specific forms of non-random measurement errors.

Evidence of differences in test conditions from the age 3 MCS sweep

Before doing so, we consider why measurement error in the test scores of young children may differ between those from different socioeconomic backgrounds. To do so, we present empirical evidence from the age 3 sweep of the MCS, when children took the Bracken School Readiness and BAS vocabulary tests. When taking these assessments, the individual conducting the survey recorded details about the environment the child took the test, and their reactions to it. This included potential distractions (e.g. noise from the TV or background conversations) and various factors that could have impacted upon children’s performance on the assessment (e.g. the child being off-task or disinterested during the assessments). Note that this information was only recorded in the age 3 MCS and not in future waves.

Table C1 below illustrates how these distractions differ across our high- and low-income groups. It illustrates how low-income children faced many more distractions than their high-income peers when they were sitting the age 3 MCS tests. For instance, it was much more likely there was noise from the TV (43% versus 21%), the child was off-task (36% versus 21%) and that there were background conversations going on (22% versus 10%).

In additional analysis, we have explored to what extent these differences can explain the difference in age 3 Bracken School Readiness test scores between children from high- and low-income backgrounds. The raw, unconditional difference in their test scores is 1.06 standard deviations. This difference declines to 0.91 standard deviations once the factors in Appendix Table C1 have been controlled. Hence it seems that distractions on the age 3 MCS tests can explain around 15% of the difference in performance of children from low- and high-income groups.

Appendix Table C1. Differences in distractions across income groups when children were completing the age 3 assessments.

	High-income	Low-income	Difference
Parent introduced child to interviewer	16%	46%	-30%
Noise from TV / Radio during test	21%	43%	-22%
Parent not at ease during test	15%	34%	-18%
Child off-task during test	21%	36%	-16%
Child disinterested during test	38%	52%	-14%
Anyone enter/leaving home during test	22%	35%	-13%
Background conversations during test	10%	22%	-12%
Child resistant to suggestions during test	17%	28%	-12%
Interruption from another adult during test	13%	21%	-8%
Interior of home is dark	1%	6%	-5%
Interruption from another child during test	29%	34%	-5%
Parent didn't keep child in vision during test	6%	11%	-5%
Child seemed fearful during test	29%	33%	-4%
Child had negative response during test	39%	43%	-4%

Extension 1. Difference levels of noise / reliability for high- and low-income groups

Having considered how distractions differ between income groups when completing early-life tests, we turn to how this conceptually may impact upon children's test performance. If one group (e.g. children from low-income families) face more distractions than another group (e.g. children from high-income families) then one may argue that their test scores are subject to different amounts of noise. In other words, while average scores are not affected, the greater number/frequency of distractions leads to more random error in the scores for some income groups than others.

We introduce this possibility be re-writing equation (1) with ρ now allowed to vary across income groups. This will allow us to consider how results are impacted when, for instance, one assumes that there is more noise (and thus ρ will be lower) in the test scores of low-income children. Equation (1) is thus modified to:

$$\tau = \rho_g(x) + (1 - \rho_g).\mu \qquad (1b)$$

Where ρ_g refers to the group-specific level of reliability, with all other terms specified as under equation (1) in the main text.

Note that, if it is the case that test scores are noisier for low-income children then their high-income peers, then the correction made for Kelley's paradox will be bigger (i.e. meaning our central estimates will be too high).

In Appendix Table C1 we consider how estimates of the difference in the probability of obtaining an A/A* grade in GCSE mathematics between high-ability high- and low-income groups varies under different assumptions of ρ_g. In this table we assume that the test scores of children from high-income children are reliably measured ($\rho_{high} = 0.9$), but vary the reliability assumed for low-income children's test scores (ρ_{low} is allowed to vary between 0.5 and 0.9).

Appendix Table C1. Estimated difference in high-achieving 5-year-olds from rich and poor background obtaining a GCSE mathematics A/A* grade: allowing test reliability to differ across income groups.

Rho		Difference in probability
Low income	High income	of GCSE A/A* grade
0.9	0.9	-32.2%
0.8	0.9	-27.9%
0.7	0.9	-23.1%
0.6	0.9	-18.0%
0.5	0.9	-12.5%

As Table C1 illustrates, as the noise in the test scores of low-income children *increases* (i.e. ρ_{low} declines), we find the difference in the probability of high-ability children from rich and poor backgrounds of achieving an A/A* grade *decreases*. In other words, the central estimates we present in the main text (where the reliability is assumed to be 0.7 for both high and low income groups) will generally be slightly too big if the reliability of the scores for low-income children is lower (e.g. 0.5).

Extension 2. Bias in average scores for high- and low-income groups.

It is possible that the greater distractions low-income children face when taking tests does more than just introduce a greater amount of random noise. For instance, it could mean that low-income children, on average, do not show their true potential in their tests – i.e. that their scores are downwardly biased relative to their high-income peers. This may be particularly true for those low-income children who did not perform well on the tests due to these distractions[13].

We thus make a further adaption to the equation presented in equation (1b), introducing an income-group specific shift in μ. The intuition here is that – if the extra distractions low-income children face leads to a downward bias on average in their test scores - μ should be shifted upwards so that their scores will regress towards a higher mean. This will in-turn lead to a smaller correction for Kelley's paradox than under assumption of random error.

Equation (1b) is thus modified to:

$$\tau = \rho_g(x) + (1 - \rho_g).(\mu_g \mp s_g) \qquad (1c)$$

Where:

μ_g = The income-group average test score

s_g = An income-group specific shift in the average test scores of income group g.

g = Income group

Note that equation 1c now allows us to both allow the amount of noise in the test scores to differ by income group (through the parameter ρ_g) and for there to be income-group specific bias in average scores (through the parameter s_g).

Appendix Table C2 illustrates how altering assumptions surrounding s_g impact upon one of our results – the difference in the probability of high-ability children from rich and poor backgrounds achieving an A or A* grade in GCSE mathematics. In this table, we assume – as per our central estimates – that the reliability (ρ_g) is 0.7 for both high and low income groups. However, as the middle column illustrates, we assume different amounts of downwards bias in low-income children's average test scores, ranging from nothing (top row) down to 0.5 standard deviations (bottom row). The right-hand most column illustrates how this impacts the results. As anticipated, as we assume a greater bias in low-income children's average scores, the smaller the correction that is made for Kelley's paradox. Hence the estimated difference

[13] An alternative argument that one could make on a related matter is that, if the tests used had measured other aspects of children's skills, the average scores for low and high income children may be closer together.

between high-ability low- and high-income children in the probability of achieving an A/A* grade in GCSE mathematics increases as s_g increases.

Appendix Table C2. Estimated difference in high-achieving 5-year-olds from rich and poor background obtaining a GCSE mathematics A/A* grade: allowing for average test scores of low-income children to be downwardly biased.

Rho (ρ_g)	Downward bias in average scores of low-income pupils (s_g)	Difference in probability of GCSE A/A* grade
0.7	0	-25.6%
0.7	0.1SD	-26.6%
0.7	0.2SD	-27.6%
0.7	0.3SD	-28.6%
0.7	0.4SD	-29.6%
0.7	0.5SD	-30.7%

Notes: ρ_g refers to the reliability of the test score used to identify high-ability children. s_g refers to amount of bias assumed in children's test scores (in terms of standard deviations). The right-hand most column than illustrates the estimated difference in the probability of high-ability low and high-income children achieving an A or A* grade in GCSE mathematics.

Appendix D. The correlation between the age 3 and age 5 cognitive test score measures

	Age 3 BAS naming vocabulary	Age 3 Bracken	Age 5 BAS picture similarities	Age 5 BAS naming vocabulary	Age 5 BAS pattern construction
Age 3 BAS naming vocabulary					
Age 3 Bracken	0.58				
Age 5 BAS picture similarities	0.21	0.24			
Age 5 BAS naming vocabulary	0.55	0.50	0.35		
Age 5 BAS pattern construction	0.25	0.31	0.37	0.37	

Appendix E. The difference in SDQ scores between initially high-achieving children from high- and low-income backgrounds. Estimates accompanying Figure 3.

	Raw difference (rho = 1)	Corrected difference (rho = 0.7)	Lower bound (rho = 0.5)	Standard error
Age 3 (unconditional)	0.65*	0.43*	0.24*	0.07
Age 5 (unconditional)	0.58*	0.35*	0.17*	0.08
Age 5 (conditional)	0.17*	0.10	0.03	0.06
Age 7 (conditional)	0.22*	0.16*	0.08	0.07
Age 11 (conditional)	0.34*	0.27*	0.19*	0.08
Age 14 (conditional)	0.45*	0.38*	0.29*	0.09
Age 17 (conditional)	0.30*	0.21	0.11	0.12

Notes: The upper bound refers to where one assumes there is no measurement error in the test used to identify high-ability children. Our central estimate assumes test reliability of 0.7. Figures refer to the difference in the outcome between high-ability children from high-income backgrounds and high-ability children from low-income backgrounds. * indicates difference statistically significant at the 5% level. Figures refer to effect sizes.