Department of Quantitative Social Science

The use (and misuse) of statistics in understanding social mobility: regression to the mean and the cognitive development of high ability children from disadvantaged homes

John Jerrim
Anna Vignoles

DoQSS Working Paper No. 11-01
March 2011

Leading education and social research
Institute of Education University of London

DISCLAIMER

Any opinions expressed here are those of the author(s) and not those of the Institute of Education. Research published in this series may include views on policy, but the institute itself takes no institutional policy positions.

DoQSS Workings Papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be available directly from the author.

Department of Quantitative Social Science. Institute of Education, University of London. 20 Bedford way, London WC1H 0AL, UK.

The use (and misuse) of statistics in understanding social mobility: regression to the mean and the cognitive development of high ability children from disadvantaged homes

John Jerrim*, **Anna Vignoles**†‡

Abstract. Social mobility has emerged as one of the key academic and political topics in Britain over the last decade. Although economists and sociologists disagree on whether mobility has increased or decreased, and if this is a bigger issue in the UK than other developed countries, both groups recognise that education and skill plays a key role in explaining intergenerational persistence. This has led academics from various disciplines to investigate how rates of cognitive development may vary between children from rich and poor backgrounds. A number of key studies in this area have reached one particularly striking (and concerning) conclusion – that highly able children from disadvantaged homes are overtaken by their rich (but less able) peers before the age of 10 in terms of their cognitive skill. This has become a widely cited "fact" within the academic literature on social mobility and child development, and has had a major influence on public policy and political debate. In this paper, we investigate whether this finding is due to a spurious statistical artefact known as regression to the mean (RTM). Our analysis suggests that there are serious methodological problems plaguing the existing literature and that, after applying some simple adjustments for RTM, we obtain dramatically different results.

JEL classification: C01, C54, I2, I28.

Keywords: Educational mobility, socio-economic gap, disadvantaged children, regression to the mean.

*Department of Quantitative Social Science, Institute of Education, University of London, 20 Bedford Way London, WC1H 0AL. E-mail: (J.Jerrim@ioe.ac.uk)

†Department of Quantitative Social Science, Institute of Education, University of London. 20 Bedford Way, London WC1H 0AL, UK. E-mail: a.vignoles@ioe.ac.uk

‡We would like to thank John Micklewright for particularly helpful comments on an initial draft. We also acknowledge the extremely helpful feedback we received from participants at a seminar at the University of Southampton, particularly the comments of Patrick Sturgis, along with the assistance provided by Paul Clarke. This work has been produced as part of the ESRC ALSPAC large grant scheme.

1. Introduction

There is a widespread belief that in the UK ‘bright’ children from poor homes rapidly fall behind their rich (but less able) peers, in terms of their cognitive skill. This view is based on a number of influential studies that have provided valuable insights into the much broader problem of social mobility (Feinstein, 2003; Schoon, 2006 and Blanden and Machin, 2007, 2010). The notion that able children from poor backgrounds have only limited chances to succeed has, understandably, led to serious alarm amongst policymakers. For instance, the Higher Education Minister David Willetts MP (2010) recently reflected on this literature:

“There is a shocking destruction of talent as the cognitive skills of bright children from modest backgrounds steadily decline during their years at school compared with more affluent children who start off with lower cognitive skills......”

In this paper we assess whether it is indeed the case that poor but able children fall behind their richer but less able peers or if this apparent trend is caused by a misinterpretation of the data via the well known statistical problem of regression towards the mean. This issue has occasionally been recognised by authors whose work has informed this debate (e.g. Blanden et al. 2010, Schoon 2006), but no research has explored the extent to which results (and the substantive inferences one draws) change after trying to take this problem into account.

To illustrate the issues we raise, our analysis focuses on two groups of children (one born in 1991 the other in 2000) using the Avon Longitudinal Study of Parents And Children (ALSPAC) and the Millennium Cohort Study (MCS). To preview our findings, we initially replicate previous work which indicates that high ability children from disadvantaged homes are quickly overtaken by their less able, but affluent, peers. However, once we apply a common correction for the aforementioned regression to the mean problem, we no longer find this to be the case. As such, we believe that there is little evidence that disadvantaged children who score highly on early cognitive tests fall behind low ability children from affluent backgrounds during their school years, and that more work is needed to assess the genuine progress made by this very important group.

We begin in section 2 by reviewing the existing literature. In section 3 we discuss what is meant by regression to the mean, how it can emerge as a result of selecting children into ability groups based on a single test, and potential ways of correcting for this problem. We then demonstrate the implications of these statistical problems in section 4 using simulated data. Section 5 moves on to the related problem of regression to the mean due to the use of non-comparable tests. We then provide examples using the ALSPAC and MCS datasets in sections 6 and 7, before concluding in section 8.

2. Existing literature and the methodology being used

The most well known study to investigate the cognitive development of high ability disadvantaged children is Feinstein's (2003) analysis of the British Cohort Study. In these data, children were examined at four time points (22 months, 42 months, 60 months and 120 months). Warning the reader to carefully interpret the results and explicitly acknowledging that the tests used measure different abilities at the different ages[1], he defines high ability as those children scoring in the top quartile of the 22 month assessment. Then, for this and the following three test points, he assigns each child a score between 1 and 100 based on their percentile of the test distribution (1 being the lowest scoring 1% of children, 100 the highest). He then calculates an average score at each of the ages, for the following four groups (having defined SES on the basis of parental occupation)[2]:

1. High ability-high SES
2. High ability-low SES
3. Low ability-high SES
4. Low ability-low SES

The main finding of the Feinstein analysis is presented in Figure 1.

Figure 1 about here

At 22 months, both high ability-high SES and high ability-low SES children sit at the same point (roughly the 88th percentile) of the test distribution[3]. But, by 42 months, the latter group has slipped to the 55th percentile and, by 120 months, to the 40th percentile. On the other hand, high ability children from advantaged homes remain much higher (sitting above the 70th percentile through to 120 months). Even more strikingly, low ability children from advantaged homes have moved up from the 12th to the 60th percentile over the same time period. Thus, in the words of Deputy Prime Minister Nick Clegg (2010):

"By the time they hang up their coats for their first day at school, bright children from poor backgrounds have fallen behind [low ability] children from affluent homes"

[1] For instance, the tests applied at 22 months are based on a combination of cognitive, personal and locomotive skill, whereas at 120 months measurement tends to focus on the first of these three traits (via reading, language and maths assessments). We discuss this issue further in Section 5.

[2] So a value of, say, 90 at 22 months for the high ability-low SES group would indicate that the average child with these characteristics (high ability-low SES) occupies the 90th percentile of the test distribution (at that age). A figure below 90 at subsequent time points would indicate that their relative position in the test distribution has declined.

[3] This will always (roughly) happen because "high ability" and "low ability" groups are being defined at this first time point (22 months).

Schoon (2006) undertakes a similar analysis using the 1958 and 1970 birth cohorts. Similar to Feinstein, she faces the problem that different skills were assessed at different ages (reading at ages 5/10 and national examinations in a range of subjects taken at age 16). She also standardises her tests in a slightly different way – rather than using percentile rank, marks at each age are transformed into a standardised z-score. Her results are presented in Figure 2.

Figure 2

The similarities with Figure 1 are striking. In particular, the cognitive skills of initially able children from disadvantaged homes decline rapidly between the ages of 5 and 16, whereas the scores of their equally able but advantaged peers show much greater stability.

Blanden and Machin (2007) perform a similar analysis using the Millennium Cohort Study[4]. As their results are based on only two points (age 3 and age 5), high ability children from poor homes have yet to be overtaken. Nevertheless, over the short period they consider, results tally with those from the other two studies (see Figure 3)[5].

Figure 3

Blanden and Machin, however, put a short (but very important) caveat on their results; because there is a random element to test scores, those who do well on an initial assessment are unlikely to perform as well on future re-tests (i.e. their scores will *regress towards the mean*). Schoon includes a similar warning shortly after presenting her results. We have similar concerns. These concerns are evident from the auxiliary analysis presented by Feinstein. Figure 4 illustrates his results again, but now on the basis of ability groupings based on the second test assessment (at 42 months) rather than the first (22 months) assessment.

Figure 4

Notice the 'V' pattern that we highlight with large dotted circles. It seems that there is not only a sharp decline in performance on later tests (i.e. those taken after 42 months) but also on those taken before this point (i.e. the test at 22 months). Feinstein did not explicitly comment on this as providing evidence of regression towards the mean, but Campbell and Kenny (1999) suggest that this is a classic sign of such a statistical artefact taking place. If we think the test scores of high ability children from disadvantaged homes genuinely decline as they get older, we would not expect to see that they show an *increase* in their test scores between age 22 and 42 months. We develop this argument further in the sections that follow.

[4] This is the most recent British birth cohort study, which follows children who were born in 2000/2001.

[5] We in fact show later in the paper that, when one includes the recently available age 7 wave, the lines in the diagram cross about the same time as found by Feinstein (2003) – i.e. roughly 76 months.

3. Regression to the mean due to selection

Lohman and Korb (2006) point out that regression to the mean can occur through many channels, including statistical error, changes to the content (and scale of) the tests being used and genuine differential rates of development. In this section, we focus on the first of these issues (i.e. regression to the mean that is caused by the selection of children into ability groups based on a single test). We return to the use of non-comparable tests later in the paper.

Regression to the mean caused by selection

Regression to the mean due to selection is a statistical phenomenon that occurs when taking repeated measures on the same individual(s) over time. Due to random error, those with a relatively high (or low) score on an initial examination are likely to receive a less extreme mark on subsequent tests. In the context of the results presented above, children defined as "high ability" based on one single exam are not necessarily the most talented in the population. Rather assignment to this group is actually based on children's true ability and the "luck" that the child happened to have when sitting that particular assessment (i.e. random error).

Figure 5 provides a graphical example for one particular child, whose "true" ability is average (we label this as T and set it equal to zero as would be the case for the mean of a standardised test). However, as researchers can not directly observe this child's true ability, it must be estimated from how they perform on an assessment. Moreover, there is a cutpoint (C) on this exam, above which children are defined as "high ability" (in this example it is at one standard deviation above the mean). The distribution presented in Figure 5 illustrates the set of possible scores that the child may receive, though on that particular day they happen to have good fortune and end up with a mark at point A. Figure 5 clearly shows that, even though the child's true ability is average (T = 0), it is still possible that they get mistaken as a high achiever (their score on the initial test is point A, which is greater than the cut-off C). What, then, would we expect to happen if this child were re-tested a short time after this initial assessment (e.g. the next week or month)? They would be unlikely to have such good fortune, and hence would probably receive a lower mark (point B) that is a better reflection of their true ability (T). In other words, they suffer "regression towards the mean".

Figure 5

The same problem occurs when classifying children into ability groups across a population. By using a set cut-off on a single test (e.g. scores greater than 1 standard deviation above the mean or the top performing quartile), our selection will be partly based upon those who experienced good fortune on the day of the assessment. What happens when this "high ability" group gets reassessed? Just as for the individual illustrated in Figure 5, they are unlikely to have such good fortune, and hence the average score will move closer towards the group's "true" value (i.e. it will regress towards the group's true mean of 0).

This would suggest that groups identified as "high ability" and observed over time would exhibit apparently falling levels of achievement due to the way we have selected individuals into the "high ability" classification. This phenomenon does not, however, solely explain the pattern seen in the existing literature (which shows that the test scores of high ability children from poor homes drop at an appreciably faster rate than high ability pupils from advantaged homes - taken as a sign that progress of the former is stunted compared to the latter)[6]. There is an additional problem. There are likely to be genuinely large gaps in early test scores between children from advantaged and disadvantaged homes. Hence SES is not something that is randomly assigned within this "high ability" subset and low SES children who get defined as "high ability" have probably had a particularly large random positive error (i.e. a lot of luck) during the initial test (and more so than their high SES peers). Under such circumstances, we would expect regression to the mean to be greater for "high ability" low SES children than for their "high ability" high SES peers. This has not been fully recognised as a possible reason for low SES children's striking decline in test scores observed in the literature and is one of the main issues we pursue in this paper.

We further illustrate our argument with the use of a statistical model[7]. To start, let:

$Y_{it} = A_{it} + \varepsilon_{it}$

Where:

A_{it} = the child's "true" ability or cognitive achievement at time t and this can change over time

ε_{it} = Error in measuring the child's true ability at time t

Y_{it} = Measured test score of individual i at time t

Assume:

$A_{it} \sim N(\mu_t , \delta_t)$

$\varepsilon_{it} \sim N(0, \gamma_t)$

and that corr $Y_{it,,} \varepsilon_{it}$ =0 and corr ε_{it} , ε_{it+1} =0.

[6] We agree that this would be the case if socio-economic status was a trait one could randomly assign to children within this academically talented group. Under this scenario, regression to the mean would still occur – but it would happen at the same rate (and tend towards the same point) for both groups.

[7] In doing so, we shall focus our discussion on children whose test scores sit above some pre-specified cut-off, with similar arguments following for children defined as "low ability" if they fall below some pre-specified cut-off

Davis (1976), citing Tallis (1961), shows that the mean test score at time t=1 for those who score above the cut point equals[8]:

$$E(Y_{i1}|Y_{i1} > k_1) \quad = \mu_1 + C_1\sigma_1 \quad = \mu_1 + C_1\sqrt{\delta_1^2 + {\gamma_1}^2}$$

Where:

Y_{i1} = Test score of individual i achieved on the first test at time=1

μ_1 = The population average score at time t=1

$C_1 = \frac{\emptyset(a_1)}{[1 - \Phi(a_1)]}$ = Mills ratio of the standardised cut-point

$$\emptyset(a_1) = \frac{\exp(-0.5.a_1^2)}{\sqrt{2\pi}}$$

$$\Phi(a_1) = \int_{-\infty}^{a_1} \emptyset(x).dx$$

$a_t = \frac{(K_t - \mu_t)}{\sigma_t}$ = Standardised cut-point at time t

K_t = Cut point used to divide children into ability groups at time t

$\sigma_1 = \sqrt{\delta_1^2 + \gamma_1^2}$ = Measured standard deviation of the first test

δ_1^2 = Variance of "true" ability at time t=1

${\gamma_1}^2$ =Variance of the error term on the first test at time t=1

Davis also shows that the expected value of scores on subsequent tests for those who score above the cut-point of the first test equals:

$$E(y_{it}|y_{i1} > k_1) = \quad \mu_t + \rho_{1t}C_1\sigma_t \quad = \quad \mu_t + \rho_{1t}C_1\sqrt{\delta_t^2 + {\gamma_t}^2} \qquad (1)$$

Where:

${\gamma_t}^2$ =Variance of the error term on the t[th] test

$\rho_{12} = \frac{\delta^2}{\delta^2+\gamma^2}$ = The accuracy of the test

[8] In other words, the average scores at time 1 for those that get defined as high ability.

The total regression to the mean effect between test points 1 and 2 is equal to the expected value of the test score in period 1 less the expected value of the test score in period 2:

$$E(y_{i1}|y_{i1} > k_1) - E(y_{i2}|y_{i2} > k_1) =$$

$$RTME_{12} = \quad (\mu_1 - \mu_2) + C_1\sqrt{\delta_1^2 + {\gamma_1}^2} - \rho_{12}C_1\sqrt{\delta_2^2 + {\gamma_2}^2} \qquad (2)$$

While the regression to the mean effect between the SECOND and SUBSEQUENT tests is equal to:

$$RTME_{2t} = E(y_{i2}|y_{i1} > k_1) - E(y_{it}|y_{i1} > k_1) = (\mu_2 - \mu_t) + \rho_{2t}C_1\sqrt{\delta_2^2 + {\gamma_2}^2} - \rho_{1t}C_1\sqrt{\delta_t^2 + {\gamma_t}^2} \qquad (3)$$

Equations (2) and (3) have important implications for our understanding of regression to the mean and, in particular, imply that it does not occur through a single channel. Firstly, there could be a genuine change in the trait (cognitive achievement for example) across the entire population over the two time periods such that the population average of the test score changes over time (μ). Secondly, there could be a change in the variance of the trait within the population between time points (δ_2). And, finally, there is the influence of the error variance γ_t. The first two of these components one might think of as “real” effects; they are substantive reasons for a change in test scores amongst the extreme group that we are interested in. The latter is the change due to statistical error or random noise. It is this last factor that we wish to purge from our estimates.

Now assume that the parameters γ and δ are relatively stable over time (i.e. the variance of the error and the variance of the underlying trait stay relatively constant). It follows that:

$$RTME_{12} > RTME_{2t}$$

In other words, this implies that most of the regression to the mean effect that is due to error comes between the first and the second test:

RULE 1: With similarly reliable tests, the regression to the mean effect due to error is usually greatest between the first and second tests

Moreover, notice from equation 2 that:

$$|RTME_{12}| \rightarrow \infty \quad as \quad \rho_{12} \rightarrow 0$$

And that:

$$\rho_{12} \rightarrow 0 \quad as \quad \gamma \rightarrow \infty$$

Thus:

RULE 2: The regression towards the mean effect is larger when the variance of the error is larger (and hence when the accuracy of the test is lower)[9]

One can also see from formulae 2 and 3 that:

$$RTME_{12} \to \infty \quad as \quad C_1 \to \infty \tag{3}$$

And that:

$$C_1 \to \infty \quad \text{as} \quad |K_1 - \mu| \to \infty$$

and hence:

RULE 3: The regression towards the mean effect is larger when the cutpoint used to divide individuals into extreme groups is further from the average mark achieved in that particular population/group

Now assume there are two types of children – Low SES (L) and High SES (H). Many studies from the UK and US (Cunha and Heckman 2006, Feinstein 2003, Goodman et al 2009) have shown that even at a very young age (e.g. ages 2-3) cognitive skill test scores differ dramatically between high and low SES groups. In other words:

$$\mu_1^H > \mu_1^L$$

When using a single cutpoint (K_1) to identify children with high (or low) early cognitive test scores, this means that:

$$\left|K_1 - \mu_1^H\right| \quad < \quad \left|K_1 - \mu_1^L\right|$$

And hence:

$$C_1^H < C_1^L$$

Under the assumptions that:

$\delta_{12}^H \approx \delta_{12}^L$ The variance of true ability is roughly the same amongst low and high ability groups

[9] However, as noted by Lohman and Korb (2006) and shown by equations 2 and 3 above, measurement error is not the only reason why regression to the mean may occur. Genuine changes in the mean or variability of the children's true ability could also be important (i.e. the δ and μ parameters will also play a role).

$\gamma_t^H \approx \gamma_t^L$ The variance in the error term in test scores is similar amongst low and high ability groups

Then:

$$RTME_{12}^H < RTME_{12}^L$$

In other words, there will be more regression to the mean for high ability – low SES individuals than the high ability – high SES group. It is then this phenomenon that could give rise to the patterns found in the existing literature, namely a steeper fall in the test scores of low SES initially high ability children than for high SES initially high ability children.

Methods of accounting for regression to the mean that is due to statistical error

We now describe two methods that attempt to correct for the problem set out above. The first was initially proposed by Ederer (1972), extended by Davis (1974), and lies at the heart of modern equivalents, such as those suggested by Marsh and Hau (2002) in the context of multi-level modelling. It requires that one has two initial measures of the construct of interest. The first of these measures should be used to divide children into ability groups and then change should be measured from the *second* test onwards. The intuition behind this comes from "Rule 1" above (that regression to the mean effects due to selection are usually greatest between the first and the second test). By using one test to classify children into ability groups and another to measure change from, one is hoping to miss most of the regression to the mean effect.

To see how this method works algebraically, we draw again upon the work of Davis. Recall from section 3.1 (and assuming that the true variance of each group's ability is unchanging δ^2):

$$E(y_2|y_1 > k_1) = \quad \mu_2 + \rho_{12}C_1\sigma_2 \quad = \quad \mu_2 + \rho_{12}C_1\sqrt{\delta^2 + \gamma_2{}^2}$$

And hence more generally that:

$$E(y_t|y_1 > k_1) = \quad \mu_t + \rho_{1t}C_1\sigma_t \quad = \quad \mu_t + \rho_{1t}C_1\sqrt{\delta^2 + \gamma_t{}^2}$$

The regression to the mean effect between the second and third test, given that the FIRST test is used to split children into different ability groups is:

$$RTME_{23} = E(y_2|y_1 > k_1) - E(y_3|y_1 > k_1) = (\mu_2 - \mu_3) + \rho_{12}C_1\sqrt{\delta^2 + \gamma_2{}^2} - \rho_{13}C_1\sqrt{\delta^2 + \gamma_3{}^2}$$

assuming that the variance of the underlying trait (δ) does not change over time. This then further simplifies to:

$$(\mu_2 - \mu_3) + \frac{\delta^2}{\delta^2+\gamma_2^2} C_1\sqrt{\delta^2 + \gamma_2{}^2} - \frac{\delta^2}{\delta^2+\gamma_3^2} C_1\sqrt{\delta^2 + \gamma_3{}^2}$$

$$= (\mu_2 - \mu_3) + \frac{\delta^2}{\sqrt{\delta^2+\gamma_2^2}} C_1 - \frac{\delta^2}{\sqrt{\delta^2+\gamma_3^2}} C_1$$

If one then also assumes that the error variance does not change over time (i.e $\gamma_2 = \gamma_3$), implying the two tests used are similarly reliable, then the final two terms cancel one another out and there is no further regression to the mean due to selection[10].

What are the limitations of this method? Firstly, it will only rid our estimates of regression to the mean (due to selection error) if the accuracy of the tests we use remains the same throughout the study (i.e $\gamma_2 = \gamma_3$). It this does not hold, then the problem of RTM may only be reduced. Secondly, there is the implicit assumption that errors are uncorrelated between the "screening" test used to divide children into ability groups and the "baseline" test that one uses as the first time point from which to measure change from. This may be a problem if the two tests one has available are taken on the same day (or in close proximity to one another)[11]. Finally, this method will still mean that we end up "misclassifying" many children as high ability when they are not. In essence, we are still only partially identifying and following the group we are actually interested in (i.e. we want to know about the development of high ability kids, but are actually following some mixture of high ability children and others who are not genuinely above the cut point).

The second method of reducing regression to the mean effects was initially suggested by Gardner and Heady (1973) and developed in the paper by Davis (1976). Assume there are now *multiple* baseline measures at your disposal (i.e. children are assessed several times on the skill(s) we are interested in before the point that we wish to measure change from). This method proposes that the *average* of (n – 1) of these measures should be used to divide children into ability groups, with change measured from the remaining one. The intuition is that the variance of the random error (that is at the heart of the regression to the mean due to selection problem) is substantially reduced when you average scores across several baseline assessments. This hence lessens the chance of defining children as "high ability" when they are not (i.e. lowers the probability of miss-classification) and thus overcomes one of the key limitations of the Ederer method described above.

[10] If, however, $\gamma_2 < \gamma_t$ then there will be further regression to the mean in future periods. This method will then reduce, rather than rid, the regression to the mean effects due to selection from our estimates.

[11] There may, for instance, be temporary factors (e.g. illness on the test day) that will lead to correlated errors across these tests. If this is the case, one should still expect to see traces of regression effects (although reduced) in the following estimates.

Davis (1974) provides the algebra behind this idea. He shows that:

$$E(\bar{y}|\bar{y} > k_1) \quad = \quad \mu_t + C\sqrt{\delta^2 + \frac{\gamma_t{}^2}{n}}$$

And

$$E(y^*|\bar{y} > k_1) \quad = \quad \mu_t + C.\delta^2\sqrt{\delta^2 + \frac{\gamma_t{}^2}{n}}$$

And thus that:

$$RTME = E(\bar{y} - y^*|\bar{y} > k_1) \quad = \quad \mu_t + C.\frac{\gamma^2}{n}\sqrt{\delta^2 + \frac{\gamma_t{}^2}{n}}$$

Where:

$\bar{y}$= Average scores on the tests used to divide children into ability groups

y^*= The first time point used to measure change from

N = Number of tests used to divide children into ability groups

$C = \frac{\emptyset(a)}{[1-\Phi(a)]}$ = Mills ratio of the standardised cut-point

$$\emptyset(a_t) = \frac{\exp(-0.5.a_t^2)}{\sqrt{2\pi}}$$

$\Phi(a_t) = \int_{-\infty}^{a_1} \emptyset(x).dx$

$a_t = \frac{(K_t - \mu)}{\sigma}$ = Standardised cut-point at time t

K_t = Cut point used to divide children into ability groups at time t

$\sigma_1 = \sqrt{\delta^2 + \frac{\gamma^2}{n}}$ = Measured standard deviation of the first test

δ^2 = Variance of "true" ability, assumed to be constant

γ_1^2 =Variance of the error term on the first test

One can clearly see that as the number of tests used to divide children into ability groups increases, the influence of the error variance on our estimates declines.

4. Simulation model

We now turn to a simulation to illustrate the implications of the model set out in section 3. Our goal in doing so is to show the reader that one can generate similar results to those found in the existing literature simply due to problems with measurement. This will help us to illustrate that the methodology applied in the current literature does not enable us to distinguish between statistical noise and genuine (policy relevant) change.

To begin, assume there is a population of 200,000 children. "True" ability across this population is assumed to be normally distributed with a mean equal to 0 and a standard deviation of 1. We call half of the population "high SES" and the other half "low SES". By the time we come to first test these children (e.g.at 22 months in the case of Feinstein) there are already large differences in "true" ability[12]. This is incorporated into the simulation by allowing the mean of true ability to differ between advantaged and disadvantaged groups (i.e. we set $\mu^H > \mu^L$). We then simulate 100,000 random draws from the following normal distributions for the two groups.

$A_1^L \sim N(\mu^L, \delta^L)$ = Distribution of true ability in period 1 for low SES children

$A_1^H \sim N(\mu^H, \delta^H)$ = Distribution of true ability in period 1 for high SES children

In the examples that follow, we set $\delta^L = \delta^H = 1$, $\mu^L = -0.5$ and $\mu^H = 0.5$. We call any child who has true ability in the top quarter (across the WHOLE population of 200,000 children) "true high ability".

This quantity (children's "true ability") is obviously something that researchers can not directly observe. We must instead rely on children's test scores as an indicator. These scores will incorporate some degree of random error[13]. Recall from the previous section that the greater the variance of this random noise, the more our estimates will suffer from regression to the mean (Rule 2). This is incorporated in our simulations via a second series of random draws, where:

$\varepsilon_1 \sim N(0, \gamma_1)$

We then add this random draw onto the child's true ability to give their OBSERVED ability (i.e. their OBSERVED test score) in period 1.

$Y_{i1} = A_{i1} + \varepsilon_{i1}$

[12] Evidence for such a gap stems from Feinstein (2003), Goodman et al (2009) and Cunha et al (2006) – to name but a few. We do not make any statement here as to how much of this early differential is due to environmental or genetic factors, but point the reader towards Cunha et al (2006) for some discussion.

[13] It is, of course, also possible that said tests have an element of non-random error. We do not consider this possibility here.

In a similar manner to before, we call any child who has an observed test score in the top quarter of the population observed "high ability".

Finally, we generate scores on two further tests following a similar process. To begin, we will assume that the child's true ability does not change over time. We then take two more random error draws (assumed to be independent of the first random error draw) and add these to the child's simulated "true ability" at time points 2 and 3:

$$Y_{i2} = A_{i2} + \varepsilon_{i2}$$

$$Y_{i3} = A_{i3} + \varepsilon_{i3}$$

$$A_{i1} = A_{i2} = A_{i3}$$

$$\varepsilon_2 \sim N(0, \gamma_2)$$

$$\varepsilon_3 \sim N(0, \gamma_3)$$

We begin by illustrating results from this base model, where there is no change in the underlying characteristic we are trying to measure over time (this will be built on later when we allow true ability to vary over time), and hence the real cognitive trajectory for all groups is completely flat. This is equivalent to the situation where the variance of the error term in the model above is always equal to zero ($\gamma_1 = \gamma_2 = \gamma_3 = 0$), with an example given in Figure 6 panel A.

Figure 6

There is, of course, no such test in the real world that has complete accuracy (i.e. suffers no random error) particularly with tests administered to young children in a non-clinical setting. We therefore let the error variance be non-zero in panel B. Specifically, we set the error variance so the correlation between observed and true ability is roughly 0.9 (i.e. that 10% of the total variation in observed test scores is due to error). In other words, although ability is now not observed, we nevertheless have quite accurate tests[14].

One can see that there is now a marked difference between what we observe and the true trajectory. Instead of a flat, constant trend over the period, we observe a sharp decline between test 1 and 2, before flattening out between tests 2 and 3. Note that we also see a significant gap emerge

[14] Recall the formula $\rho_{12} = \frac{\delta^2}{\delta^2+\gamma^2}$ = The accuracy of the test

between high SES and low SES groups[15]. This pattern is exacerbated in panel C, where we set the tests being used to have lower levels of accuracy (ρ is now set to 0.25 at each time point)[16]. Indeed, we have reduced the accuracy of tests far enough for the high observed ability – low SES and low observed ability – high SES lines to cross. We know, however, that this is not "real" change in this instance (recall from panel A that we have set the true gradient to be flat). Rather we are finding this pattern simply as the result of statistical error.

We explore the implications of this further in Table 1, where we illustrate the proportion of children who get misclassified into the "high ability" group. The far right hand column refers to the "truth". One can see that in our simulated model, only 4,444 (4%) of low SES children should get defined as high ability, compared to 45,556 (46%) of high SES children. Yet as the error variance of our test measure increases, more and more children get misclassified. Take, for instance, a test that has quite high levels of accuracy (0.8). Table 1 reveals that 8,862 low SES children (8.9%) get defined as "high ability", twice as many as the number we would classify as "high ability" if we could observe their ability perfectly (i.e. than "should" be the case). On the other hand, *fewer* high SES children (41,138 or 41.1%) make it into this group (i.e. *fewer* get defined as high ability than should be the case). Moreover, note that the error term on the first test for those who get defined as "high ability" is larger for those from low SES backgrounds, while on the second test, the error for both groups is roughly zero. The implication is that scores for the former will fall more by those than the latter due to them loosing this larger random draw – giving rise to the patterns illustrated in Figure 6[17].

Table 1

We build on this initial simulation in Figure 7. In particular, we now allow there to be true change in ability over time (the term A is now sub-scripted with t):

$$Y_{it} = A_{it} + \varepsilon_{it} \qquad \varepsilon \sim N(0, \gamma) \qquad (5)$$

Specifically, in our simulation we now let true ability to be constant between period 1 and 2 for all groups, but that (true) high ability – low SES children suffer a marked decline in their cognitive skill between periods 2 and 3 (see panel A of Figure 7).

Figure 7

[15] Socio-economic status is the only dimension across which we allow true ability to vary in this simulation
[16] In other words, 75% of the total variation in observed test scores is due to error
[17] Table 1 also reveals this becomes an increasing problem the less reliable the measure used to capture the underlying trait (tallying with the comparison made between panels B and C of Figure 6).

Following a similar logic to before, we show the patterns that one would observe if one did not allow for RTM. Panel B once more refers to when using a test with high accuracy. Two key points emerge. Panel A show a genuine decline in test scores for low SES "high ability" children between periods two and three. Yet this genuine pattern is not observed in Panel B, even when using high quality tests. The methodology being used in the existing literature therefore potentially identifies a very different pattern to what occurs in reality – it suggests there is a big decline between the first two periods and only a shallow change thereafter where as in this case the reverse is true. By implication, if one were to use this methodology to advise policymakers (as has been done consistently in the UK), it is likely the problem at hand would be exaggerated, and we *could* end up telling them to invest in the wrong place (i.e. by saying our real problem occurs between periods 1 and 2, when infact the big drop seen in the simulation is between periods 2 and 3).

The second key point comes from comparing panel B in Figure 6 to panel B in Figure 7. Recall that we simulated no change in *true* ability (for any group) in the former, but a sharp decline for high true ability – low SES children in the latter. It seems, however, that (when applying current methodology) one is unable to distinguish between these two quite different situations. In other words, we are unable to tell whether the patterns we observe are "real" or not, and would end up reaching the same substantive conclusion no matter what the "truth" might be.

Methods to account for regression to the mean due to selection

We now illustrate how our methods for correcting this problem perform in our simulated data. Specifically, we begin by assuming there is a (single) auxiliary test available in period 1. We set "reality" to be exactly the same as in Figure 7 panel A – "true ability" remains stable between period one and two, but then declines dramatically for the high true ability – low SES group between period two and three. The new "auxiliary" test is then used to divide children into ability quartiles, with all other aspects of the simulation unchanged. Our goal is to investigate whether we are now able to accurately identify the big decline in test performance for high ability-low SES children between periods 2 and 3. Results can be found in Figure 8.

Figure 8

When using a high accuracy test (panel B) results are reasonably encouraging (certainly in comparison to the existing methodology as presented in Figure 7). In particular, we correctly find the gradient to be flat between period one and two, and that there is a decline for the high ability-low SES group between time point two and three. There does, however, seem to be some attenuation in our estimates, with the drop in test scores lower than in "reality" (this occurs due to the fact that our observed "high ability" group contains many children who have been classified as highly able when they are not –recall Table 1). This problem is exacerbated in panel C, when one uses rather less accurate tests. Indeed, it becomes extremely difficult to say anything meaningful about the progress of the high ability – low SES groups when using low quality tests.

As we discuss in section 4, Davis notes one may improve on this method by using the average of multiple auxiliary tests to assign children into ability groups. This will, in particular, help to reduce the problem of misclassifying children as "high ability" when they are not. We investigate this by repeating the analysis above, but now defining ability groups based upon the average of five auxiliary tests rather than just one[18]. Results can be found in Figure 9.

Figure 9

The results are quite encouraging. As with the previous method, we correctly observe that the gradient is flat between periods 1 and 2, while also seeing a clear decline for the high ability – low SES group between periods 2 and 3. Regarding the later, it also seems we are able to make a reasonable estimate of the size of the decline (i.e. there is less evidence of attenuation). Hence it seems that, when one has multiple baseline measures, it is possible to significantly reduce regression to the mean effects due to selection while also being able to detect substantive changes to the socio-economic gradient.

5. Regression to the mean due to the non-comparability of tests

Regression to the mean due to selection can explain a substantial part of the findings in the existing literature. In particular, it explains why such a large fall in test scores occurs between the first and second tests. But this is not the whole story; Figures 1 and 2 suggest that the relative performance of low SES children continues to fall past the first re-test (albeit at a slower pace). We are concerned, however, that this could be the result of the authors measuring different skills at different ages[19]. Feinstein, for instance, uses a measure at 22 months that is a combination of cognitive, motor, personal and locomotive skill, while at 120 months the variable is rather more geared to the first of these abilities (via reading, language and maths assessments). Likewise, Schoon moves from explicitly using a measure of language skill at age 10 to scores on national examinations (across a number of different subjects) at age 16.

We proceed by giving the intuition as to why this may lead to additional regression to the mean. Individuals who do well on a specific test are those who excel in a specific area or skill (assuming we are using a test that is a reasonably accurate measure of this skill). If we then go on to measure a different skill (or set of skills) on a follow-up test, it is unlikely that these individuals are

[18] All tests are assumed to have reasonable levels of accuracy (the correlation between true ability and the test measure is set to 0.85).

[19] Indeed, it is harder to argue that this drop in later tests is occurring due to selection.

also the most talented in this other skill, and will thus (as a group) look more like average members of the population. So, for instance, say we identify the top quarter of children with advanced skills in mathematics via an aptitude test, but when it comes to re-assessment, the children are examined in their language ability. There will inevitably be some children who are very good at the former, but unspectacular at the later. Hence, the average mark for the "high ability" group will be noticeably lower on the re-assessment. In other words, what we believe is change is actually regression towards the mean from measuring a different skill. In essence, if you select a group of children who excel in one area, but then test their ability in another, is it really any surprise to find that their average mark declines?

The use of different tests over time can explain why we see a decline in the test scores of high ability children. This problem can also, however, explain why the decline is greater for high ability – low SES children than for their high ability – high SES peers. Assume we have tests at two time points measuring children's skill in different areas (e.g. reading and maths). Evidence from the existing literature suggests that there will be socio-economic gaps in both domains (e.g. high SES children score higher marks than low SES children on both reading and maths tests even from an early age). The implication of this is that "high ability" children (as defined on one of these skills e.g. reading) will revert towards these different means, depending on whether they are from an advantaged or disadvantaged home. In particular, "high ability" low SES children will be reverting to a lower group average score than their high SES peers. This, consequently, gives rise to the larger fall in test scores that one observes for the former compared to the latter. But this is again not "real" or policy relevant change; rather it emerges as a result of the children being assessed in different skills at different ages.

We also illustrate this problem with our simulated data. Say that over the three time periods there is no change in the ability we wish to measure for any group:

$$A_{i1} = A_{i2} = A_{i3}$$

The first two tests that we have available are quite accurate measures of this skill.

$$Y_{i1} = A_{i1} + \varepsilon_{i1}$$

$$Y_{i2} = A_{i2} + \varepsilon_{i2}$$

We do not, however, have a measure of the same ability at the third period. Instead, there is a test (Z) of another ability (A*):

$$Z_{i3} = \mathrm{A}^{*}_{\mathrm{i}3} + \varepsilon_{i3}$$

Assume that the first two moments of this other ability (A*) are the same as that of the ability (A) we are interested in. That is:

${A^*}_3^L \sim N({\mu^*}^L, {\delta^*}^L)$ = Distribution of A* for low SES children

${A^*}_3^H \sim N({\mu^*}^H, {\delta^*}^H)$= Distribution of A* for high SES children

$A_3^L \sim N(\mu^L, \delta^L)$ = Distribution of A (the skill we are interested in) for low SES children

$A_3^H \sim N(\mu^H, \delta^H)$ = Distribution of A (the skill we are interested in) for high SES children

Where:

${\mu^*}^L = \mu^L = -0.5$ (true ability low SES children 0.5 below the mean for both A and A*)

${\mu^*}^H = \mu^H = 0.5$ (true ability high SES children 0.5 above the mean for both A and A*)

$\delta^H = {\delta^*}^H = \delta^L = {\delta^*}^L = 1$ (variance of true ability for low and high SES children is 1 for both A and A*).

One implication of the above is that A and A* will have the same sized socio-economic gap (i.e. low SES children are, on average, just as far behind their high SES peers in terms of A* as they are A).

Also assume that, although A and A* are different skills, there is a reasonable correlation between them (for instance reading and maths are different skills, but there is nevertheless likely to be a correlation between them). We call this ρ*:

A* = ρ*.A

ρ* = The (unobservable) correlation between the skill we are interested in (A) and the skill that we measure (A*) at the third time point

Note that the higher the value of ρ*, the less that this form of regression to the mean becomes a problem (in the extreme, where ρ*=1, we are measuring the same skill over time and hence do not face the problems discussed in this section at all).

We proceed in our simulation by generating a new test score (Z_{i3}) in period 3, which is a measure of the skill A* that contains some error (ε*). We set ρ* (described above) to equal 0.6. The error is assumed to have a mean of 0 and variance γ* (in the results below, we assume γ* is relatively small and hence reasonably accurate tests).

Results from this simulation can be found in Figure 10. Panel A illustrates what actually happens to the skill, or set of skills, (A) we are interested in (in this example, it is set to be flat for all groups). Panel B, on the other hand, is what we as researchers would observe when using the existing methodology in the literature, assuming that we measure a different skill at time 3 to at times 1 and 2, and that we have quite accurate tests throughout.

Figure 10

Notice (in panel B) that the pattern seen between the first two periods in very similar to that shown previously (reflecting regression to the mean due to selection). The important point of note now, however, is that regression to the mean continues to occur in the right hand panel between periods 2 and 3 due to the measurement of a different skill at the final time-point. Again, this leads us to a very different conclusion to "reality" in panel A on the left. We observe that initially highly able children from poor homes get overtaken by their less able but affluent peers, whereas the reality is that there is actually no change in the gradient for any socio-economic group.

The implication of this result should be clear. When exploring cognitive gradients for "high ability" children from disadvantaged homes, it is particularly important to use tests that measure the same skill over time. In the context of the existing literature, by measuring different skills at different ages, it is impossible to substantiate whether the sharp decline for the high ability – low SES group is representing genuine change or simply an artefact of the data.

6. Examples from actual datasets - ALSPAC

The previous section highlighted the problem of regression to the mean using simulated data. We now explore whether similar findings hold in our analysis of two well-known UK datasets.

We turn first of all to the Avon Longitudinal Study of Parents and Children (ALSPAC)[20]. This resource has been widely used to explore child development from both medical and social science perspectives, and is one of the richest datasets (in terms of the information it has collected on respondents) available in the UK. This resource is particularly suited for our purposes due to the number of test measures it contains at various points in children's lives.

To begin, we set out the ALSPAC sample design and the measures it contains. All women who lived within the former English district of Avon and expected to give birth between April 1991 and December 1992 were asked to take part in the ALSPAC study. In total, roughly 14,000 women agreed to take part, approximately 85% of all births in the area between these two time points. The non-response that did occur was not random, and the dataset generally under-represents young

[20] See http://www.bristol.ac.uk/alspac/sci-com/ for more details on the ALSPAC data resource.

mothers, ethnic minorities and lower socio-economic groups. We do not dwell on this issue in this paper, as it is not our intention to get the best possible estimate of the socio-economic gradient, but rather illustrate the difficulties that are caused by regression to the mean. Our sample is further restricted to those children who have full information available on their key stage 1 -3 test scores (age 7, 11 and 14 national test scores that have been linked into ALSPAC from administrative education records) and those who attended a special clinic that a selection of survey participants attended at age 7. This leaves us with a working sample of 3,776 children.

The main outcomes that we shall focus upon are children's scores on Key Stage English exams (at ages 7, 11 and 14). We also have available a number of additional indicators of children's skill from the ALSPAC clinic. The measures that we use are described in detail in Appendix 1, and include indicators of children's reading, spelling and language ability along with two assessments of their motor skills. We proceed as per the existing literature, and assign each child a score between 1 and 100 on each of the assessments based on their percentile rank in the test distribution. The other key covariate, family background, is measured by the highest level of education achieved by the child's mother or father, which is reported by the child's parents in one of the background questionnaires. We reduce responses into three groups:

Low = Neither parent has more than O-levels (i.e. no post-compulsory schooling)

Medium = At least one parent holds A-levels, but neither holds any higher qualification

High = At least one parent holds a degree.

We note that one may debate whether this is the best way to divide children into "advantaged" and "disadvantaged" backgrounds. Again we abstract from this discussion here, although an overview of the importance of (and difficulties with) such definitional issues is provided in Appendix 1.

To begin, we demonstrate the importance of using comparable tests over time. In particular, we consider the situation where our initial test measures "development" in a broad sense (as per the 22 and 42 month tests used by Feinstein) but then follow children's progress through school with language based achievement tests. Specifically, we take a simple average of children's scores on two tests of their motor ability (taken from the age 7 clinic) and their total point score on Key Stage 1 assessments as our measurement at time 1 (which is, in this instance, 84 months). Follow up tests (at 132 and 168 months) are, on the other hand, are based solely upon children's performance in Key Stage 2 and Key Stage 3 English exams. Results from this analysis can be found in Figure 11 panel A. One can see the common pattern found in the existing literature. There is a dramatic decline for the disadvantaged high ability group between the first two test points, and a continuing (but significantly shallower decline) thereafter. This leads the lines for high ability - low SES and low ability - high SES children to cross somewhere between the second and third test point.

We perform exactly the same analysis again in panel B with one important difference. Now instead of using a composite test measure (i.e. a combination of motor and language skills) at the first time point, we rely solely upon children's performance in the Key Stage 1 exams (i.e. just their language skills). This means we are now comparing children's total points score on very similar national examinations in English throughout the study period. The change in our results (see panel B) is dramatic. In particular, there is not such a sharp decline in test scores for high ability children from disadvantaged homes, and no longer any evidence that the "lines cross". Yet there is still some evidence of a decline for the high ability – low SES group. This seems to emerge between the ages of 7 and 11 (the average percentile rank for this group drops from the 85th percentile to the 72nd percentile), with only a very slight subsequent decline (down to the 68th percentile) thereafter.

Figure 11

Although panel B now shows results based on measures of the same skill (broadly speaking) over time, we have yet to take into account regression to the mean that occurs due to error from selection. In other words, the estimates in panel B still use just a single measure (Key Stage 1 test results) to divide children in ability groups *and* to measure change from. We now attempt to correct for this problem in panel C, using the method proposed by Gardner and Heady (1973). We do this by dividing children into ability groups based upon the <u>three</u> auxiliary clinic tests (that are quite strongly correlated with children's achievement on national English assessments), and then measuring change in English ability from key stage 1 onwards[21].

The estimated gradient for all groups is now rather gentle. In particular, note that we now find there to be essentially no decline between tests taken at Key Stage 1 and Key Stage 2 for the high ability groups. We do see some movement, however, between Key Stage 2 and 3 for initially high ability children from low SES homes (they decline from the 68th to the 61st percentile). Note that this is quite the opposite conclusion to the unadjusted estimates in panel B (where the main decline seemed to be occurring between Key Stage 1 and 2, and not between Key Stage 2 and 3)[22]. Most importantly, there is no longer support for the "crossing lines" phenomenon that was found in panel A.

[21] The ALSPAC clinic tests we are referring to assessed children's spelling, reading and language skills. Scores correlate quite highly with children's performance on national exams. For further details see Appendix 1
[22] Indeed, now we have applied a method to take account of regression to the mean due to selection in our estimates, the decline in rank position for bright children from poor homes seems rather modest (particularly given we are looking over a seven year time horizon).

7. Example using the MCS data

The MCS also offers the opportunity to investigate many of the methodological concerns laid out in previous sections of this paper. This is a nationally representative dataset of children born in 2000/2001, who have been surveyed at four ages (roughly at age 1, 3, 5 and 7). Others (e.g. Blanden and Machin) have investigated the progress of initially high ability children from poor homes using these data and applying the methodology that prevails in the existing literature. These authors note that regression to the mean may be causing some difficulty in their estimates. We hence attempt to take their work a step further by considering how results change once we try to take this problem into account.

As with any longitudinal survey, there is an element of non-response and attrition in the MCS. Although 19,488 children were included in the initial study, only 14,043 remain by wave 4. However, as part of the MCS, the survey organisers have produced a set of high quality response weights to take into account longitudinal non-response over the four waves currently available (we apply these weights throughout our analysis). Of course, some individuals have missing data on key variables, leaving us with a working sample of 10,049 individuals. In the analysis that follows, we measure household income as our measure of family background. Specifically, we define:

Low SES = Bottom quartile of household income

Middle SES = Second or third quartile of household income

High SES = Top quartile of household income

Again, we do not dwell on issues of non-response and whether income is the appropriate measure of "advantage" here, as this is not the primary concern of this paper[23]. Rather we want to show that we can obtain the distinct pattern found in the existing literature, and how this changes once the problem of regression to the mean is taken into account.

As part of the MCS study children took two types of developmental assessment at age 3 – the naming vocabulary sub-set of the British Ability Scale and the Bracken School Readiness Test. The former has been designed to assess children's expressive language and, as such, was only administered to children who speak English (thus our sample includes English speakers only)[24]. The latter assessment (Bracken) measures concepts that parents and teachers traditionally teach children in preparation for formal education. This is based on a set of six sub-tests from which standardised scores are calculated based on the child's combined performance. Each child is then categorised into

[23] We have, nevertheless, checked that all our substantive results hold when using alternative measures of family background and when non-response has been taken into account.

[24] This has been described by Hansen et al (2008) as tests of cognitive ability suitable for children between 3 and 17 years of age, and was administered by a third party (the MCS interviewer) in a computer aided interview.

one of five groups (very delayed, delayed, average, advanced and very advanced) based on their total Bracken score (i.e. the summation of their marks on the six sub – domains). This measure has been validated against various other indicators of childhood abilities and intelligence, including the WPPSI-R measure of child IQ (Laughlin 1995). Indeed, this is now widely used in early intellectual screening and identification of "high ability" children at a young age. It is, for instance, one of the tests used by the city of New York in gaining access to its "Gifted and Talented" scheme. This therefore seems to be a particularly useful measure for studying the topic at hand.

Having two cognitive measures at age 3 is of obvious appeal given the arguments laid out in previous sections. Specifically, we take children who have been defined as "delayed" or "very delayed" on the Bracken assessment as our indicator of low starting test scores ("low ability") and those classified as "advanced" or "very advanced" as our indicator of "high ability". The other assessment (the vocabulary subset of BAS) will be used as the first observation point from which we will measure change from. Regarding follow-up tests, children were re-examined on the BAS vocabulary sub-domain at age 5, and the reading subscale of BAS at age 7. The latter is a test of children's receptive language skill, and has obvious similarities with the BAS vocabulary assessments that took place at ages 3 and 5. Yet it does measure a slightly different skill (children's receptive, rather than their expressive, language). It has, nevertheless, been used to compare change in children's language skills over time (Hansen et al 2010 p 161). This is therefore taken as our indicator of children's language ability at age 7.

It is important to recognise that the MCS data we use does have its limitations. Firstly, while the two age 3 tests are clearly designed to assess early cognitive abilities, they do not measure exactly the same skill (BAS involves spoken language while Bracken is a non-verbal assessment). Secondly, these two tests were taken by children on the same day. Recalling our discussion in section 3, this could mean that errors on the two assessments are correlated (for instance, the child is feeling ill on the day of the test, and so performs below his/her ability level on both assessments). If this is the case, regression to the mean due to selection error will not be completely purged from our estimates. Finally, as we note above, slightly different aspects of children's language skills are being assessed at different points in time.

We begin by presenting a cross-tabulation of the two tests that the children sat at age 3 in Table 2 (BAS vocabulary and Bracken School Readiness). Our intention is to illustrate that many children change classification even when using assessments taken on the same day (i.e. they make it into a "high ability" group defined on one test but not another) and that the pattern of this change differs by socio-economic group. One reason for this might be that tests are capturing different skills, and that children who excel in one area do not necessarily do so in another. However, an alternative explanation (and one that we emphasise in this paper) is that, by using a single test, many children will be misclassified on a single assessment due to random error.

Table 2

Focusing on the right hand column, notice that half of the low SES children who score in the top quartile of the age 3 BAS vocabulary distribution are defined as "average" on the Bracken test. This is in comparison to less than a third of those from high SES backgrounds. Analogous findings emerge at the bottom of the distribution. For instance, 43% of low SES children who score in the bottom quartile of the BAS distribution are defined as delayed or very delayed under the Bracken scale, compared to only 12% of the highest income group. This is consistent with our simulation results that suggest that many children will end up being misclassified on the basis of a single test, and that there is a difference in the proportion misclassified by socio-economic group. In Tables 3 and 4 we show this is not due to the particular tests we are using; one reaches the same conclusion when comparing children's age 5 BAS (vocabulary) and foundation stage profile[25] (language and communication) scores.

Next, we turn to our substantive findings with regards to change over time. In panel A of Figure 12, we present results based on the methodology used in the literature, using scores on the age 3 BAS vocabulary test to both divide children into ability groups *and* provide the initial observation from which to measure change from. By contrast, in the right hand panel, ability groups are defined using a separate age 3 test (the Bracken test), in an attempt to correct for regression to the mean due to error from selection.

Figure 12

The pattern in the left hand panel should be familiar. Children with scores in the top quartile of the age 3 BAS assessment see a rapid decline between ages 3 and 5 – particularly those from low income backgrounds. Specifically, they move (on average) from roughly the 90th to the 50th percentile. On the other hand, high income children who were initially in the bottom quartile move (on average) from roughly the 10th to 40th percentile. By the last time point (age 7) initially high scoring children from poor homes have been overtaken by their less able, but affluent, peers.

The regression to the mean adjusted estimates in the right hand panel tells quite a different story[26]. There is now no suggestion that the cognitive skills of bright children from poor homes

[25] A nationally comparable test taken by children on entry into primary school.

[26] Note that the average test scores on the first (age 3) BAS assessment of those children within the high ability group are now much lower than the previous estimates. For instance, the average score of those defined as high ability- low SES is at the 90th percentile in the left hand panel and the 60th on the right, while the analogous figures for high ability-high SES is the 90th percentile and the 70th. The reason for this is that the extent of miss-classification into high ability groups (based on a single test) is likely greater for low SES children.

rapidly decline between 3 and 7 years of age. In fact, the estimated gradients between ages 3 and 7 in the MCS for the "high ability" groups now seem to be essentially flat. There is, on the other hand, some evidence that those defined as delayed or very delayed improve over the study period - although this is true for both low SES and high SES groups[27].

Given the discussion in previous section, we urge that care needs to be taken when interpreting this result. In particular, the incline in test scores for initially low scoring children could be evidence of residual regression to the mean effects (e.g. due to the errors in test scores being correlated due to age 3 BAS and Bracken assessments taking place on the same day). Likewise, we illustrated how there may be attenuation in estimates because the MCS data do not provide multiple auxiliary tests to define children into ability groups.

Nevertheless, the main conclusion to emerge from our analysis of the MCS data is clear. If we divide children into ability groups using an auxiliary test in an attempt to combat error caused by regression to the mean, we reach a very different conclusion to that which prevails in the existing literature. In particular, we do *not* find any evidence that the cognitive skills of initially able children from poor homes rapidly decline.

8. Discussion and conclusion

In this paper, we have considered one particular methodological difficulty in studying the academic progress of initially high ability children from poor homes, namely regression to the mean. By dividing children into ability groups on the basis of a single assessment (which is subject to a certain amount of error) one is likely to encounter the problem of regression to the mean when subjects are given a re-test. This can induce substantial bias and cause wrong conclusions to be drawn from trends in the data. Our simulation evidence shows clearly how, using existing methodologies, one can find a large apparent decline in test performance for bright children from poor homes even when no real decline is taking place. Statistical error can therefore explain why we see bright children from poor homes apparently falling behind their affluent high ability peers, a result that is not confirmed using ALSPAC and MCS data and applying two common ways of taking thes problem of RTM into account. Contrary to the findings in the existing literature, when we attempt to take account of regression to the mean, we find little evidence of a striking decline in the cognitive skill of poor high ability children, There is likely to still be some attenuation bias in our estimates, as discussed above, and hence even the decline we do observe in the cognitive skill of poor high ability children is likely to be an overestimate. .

[27] Of course,

What then can we conclude from these findings? Firstly, the methodology currently being used to study this topic is inadequate. Our simulation study makes it quite clear that when we apply existing methodology, we are able to observe a change in gradient even when there is none. Equally it is also possible to miss a change in gradient when there is one. Different methodological approaches are urgently needed in this area of research before any further statements can be made about the academic progress of this high ability low SES group. There is also a limit to how well researchers can deal with the problems associated with RTM due to data limitations. Relying on a single test to define children into ability groups leads to a significant proportion of misclassified individuals. Multiple measures are needed therefore and most data sets (bar ALSPAC) do not have them. Likewise, tests that measure different skills at different ages will confound any attempt by researchers to separate statistical artefacts from genuine change. Ideally therefore we need to have data that contains multiple high-quality measures of children's cognitive skills over time. The British 2012 cohort study may provide an opportunity to collect data of this kind, with which to study the academic progress of highly able but disadvantaged pupils.

Finally, what then does this methodological problem imply for the current consensus that highly able children from poor homes get overtaken by their affluent (but less able) peers before the end of primary school? Although this empirical finding is treated as a stylised fact in UK policymaking, and indeed the academic literature, our results provide little evidence that this is actually the case. In our estimates we show the cognitive gradient to be essentially flat (after applying two simple methods to account for regression to the mean). We are however, acutely aware that our simulations illustrate the difficulty of correctly classifying children as "high ability", and the possibility of attenuation bias remains in our empirical estimates of trends using MCS and ALSPAC. What we can say is that by not considering the problem of regression to the mean, current estimates of the decline in cognitive skill of high ability poor children is likely to be overstated. Further research, using methodologies that account for RTM, is needed so that we can learn more about the progress made by this particular group.

References

Bracken, B. (1998) “Bracken Basic Concept Scale – Revised”

Hansen, K. and Joshi, H. (2007) “Millennium Cohort Study Second Survey – A User’s Guide to Initial Findingers”

Laughlin, T. (1995) “The School Readiness Composite of the Bracken Basic Concept Scale as on Intellectual Screening Instrument”

Blanden, J. and Machin, S. (2010) “Changes in inequality and intergenerational mobility in early years assessments” In Hansen, K., Joshi, H. Dex, S. (eds) Children of the 21st century (Volume 2) The first five years

Campbell, D. and Kenny, D. (1999) *A Primer on Regression Artifacts*

Clegg, N. (2010) "Putting a Premium on Fairness", Speech delivered at Spires Junior School, Chesterfield, October 15th 2010

Cunha, Flavio & Heckman, James J. & Lochner, Lance, 2006. "Interpreting the Evidence on Life Cycle Skill Formation," Handbook of the Economics of Education, Elsevier.

Davis, C (1976) “The effect of regression to the mean in epidemiologic and clinical studies”, American Journal of Epidemiology 104, 493-498

Ederer, F. (1972) “Serum cholesterol: effects of diet and regression toward the mean”, Journal of Chronic Disorders, 25, pp 277-289

Feinstein, Leon, Inequality in the Early Cognitive Development of British Children in the 1970 Cohort. Economica, Vol. 70, pp. 73-97, 2003

Galton, F. (1886). "Regression towards mediocrity in hereditary stature". The Journal of the Anthropological Institute of Great Britain and Ireland **15**: 246–263

Gardner, M. & Heady, J. (1973) “Some effects of within person variability in epidemiological studies”, Journal of Chronic Disorders, 26, pp 781-795

Goodman A., Sibieta L. & Washbook E. (2009) “Inequalities in educational outcomes among children aged 3 to 16”, Report for the National Equality Panel

Kavsek, M. (2004) “Predicting later IQ from infant visual habituation and dishabituation: A meta-analysis”, Applied Developmental Psychology, volume 25, page 369–393

Lohman, D. F., & Korb, K. A. (2006). Gifted today but not tomorrow? Longitudinal changes in ability and achievement during elementary school. *Journal for the Education of the Gifted, 29(4),* 451-486

Marsh, Herbert W. and Hau, Kit-Tai(2002) 'Multilevel Modeling of Longitudinal Growth and Change: Substantive Effects or Regression Toward the Mean Artifacts?', Multivariate Behavioral Research, 37: 2, 245 — 282

Slater, A. (1997) “Can measures of infant habituation predict later intellectual ability?”, *Arch Dis Child* 1997;**77**:474-476

Schoon, I. (2006) *Risk and Resilience. Adaptations in changing times*. Cambridge University Press.

Tallis, G. (1961) "The Moment Generating Function of the Truncated Multinormal Distribution", Journal of the Royal Statistical Society B, 23, pp 223 – 229

Willetts, D. (2010) "The Pinch: How the Baby Boomers Took Their Children's Future – And Why They Should Give It Back", Atlantic Books

ZEANAH, C. H., BORIS, N. W. and LARRIEU, J. A. (1997). Infant development and developmental risk: a review of the past ten years. Journal of the American Academv of Child and Adolescent Psychiatry, 36(2), 165–78.

Table 1. Descriptive statistics drawn from simulated data

	Accuracy of the test			
	1 ("Truth")	**0.8**	**0.5**	**0.25**
Number of children <u>observed</u> as high ability				
High SES	4,444	8,862	14,079	20,074
Low SES	45,556	41,138	35,921	29,926
Proportion of children MISSCLASSIFIED as high ability (i.e. defined as high ability when they are not)				
High SES	0	45%	85%	91%
Low SES	0	15%	34%	42%
Average error on first test (ε1) for those defined as high ability				
High SES	0	0.7	1.9	3.1
Low SES	0	1.3	2.8	6.9
Average error on second test (ε2) for those defined as high ability				
High SES	0	0	0	0
Low SES	0	0	0	0

Notes:

Table refers to data from our simulation. It illustrates: (a) the number of children we define as high ability, (b) the proportion of children who are mistakenly classified as high ability and (c) the average size of the residual on the first and second test for those who get defined as high ability. This is done separately for our simulated high and low SES groups. We show how results change when using tests of different "accuracy". The first column on the left (labelled "truth") refers to when we are able to perfectly observe children's true ability. The columns to the right of this illustrate how more children are wrongly classified (and the average residual for the high ability group gets bigger) as tests of lower accuracy are used.

Table 2. Cross-tab of BAS quartile by Bracken classification for low and high SES groups (column percentages)

(a) Low SES

		Age 3 BAS Classification			
		Bottom Q	2nd Q	3rd Q	Top Q
Age 3 Bracken Classification	Very delayed	7	1	1	0
	Delayed	36	18	9	1
	Average	54	72	68	50
	Advanced	4	8	19	39
	Very advanced	1	1	4	10
	TOTAL	100	100	100	100

(b) High SES

		Age 3 BAS Classification			
		Bottom Q	2nd Q	3rd Q	Top Q
Age 3 Bracken Classification	Very delayed	2	1	0	0
	Delayed	10	4	2	0
	Average	71	67	57	32
	Advanced	14	23	33	47
	Very advanced	3	5	9	21
	TOTAL	100	100	100	100

Notes:

Table illustrates cross-tabulation between quartiles of children's score on the age 3 BAS vocabulary assessment and the classification they were assigned based the age 3 Bracken test. This is presented separately for low SES (top panel) and high SES (bottom panel) children. Figures refer to column percentages.

Table 3. Cross-tabulation of children's age 5 BAS vocabulary quartile against their foundation stage profile language and communication quartile (column percentages)

Low SES

		Age 5 BAS Vocab Classification			
		Bottom Q	**2nd Q**	**3rd Q**	**Top Q**
Age 5 Foundation Stage Profile (Language and Communication)	**Bottom Q**	52	39	28	16
	2nd Q	28	31	25	29
	3rd Q	13	19	28	29
	Top Q	6	11	19	25
	TOTAL	100	100	100	100

Notes: See notes to Table 2 above

High SES

		Age 5 BAS Vocab Classification			
		Bottom Q	**2nd Q**	**3rd Q**	**Top Q**
Age 5 Foundation Stage Profile (Language and Communication)	**Bottom Q**	24	13	11	6
	2nd Q	34	32	24	21
	3rd Q	26	27	27	30
	Top Q	16	28	37	43
	TOTAL	100	100	100	100

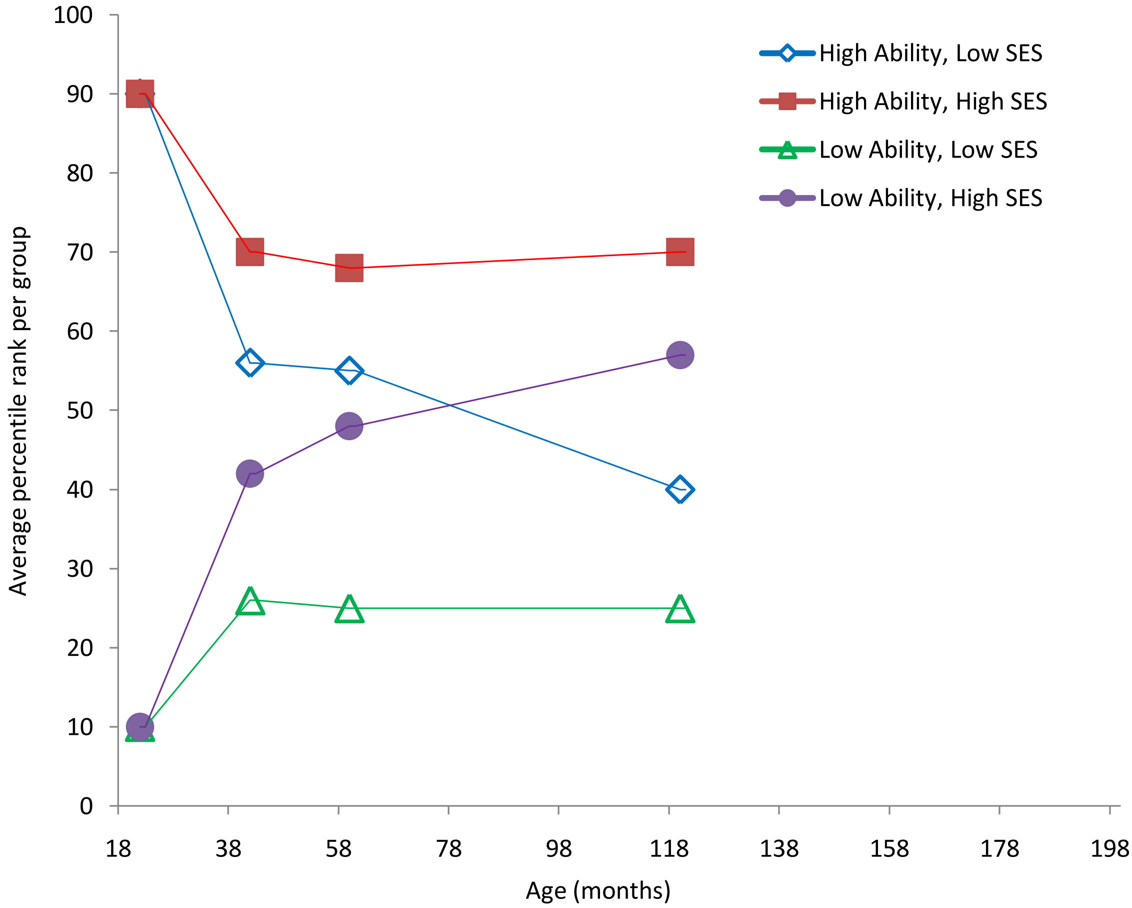
Table 4. Cross-tabulation of children's quartile on test 1 versus their quartile on test 2 using simulated data (when allowing no "real" change to take place between the two tests)

		Simulated test 1			
		Bottom Q	**2nd Q**	**3rd Q**	**Top Q**
Simulated test 2	**Bottom Q**	44	35	30	25
	2nd Q	28	29	29	27
	3rd Q	19	22	25	27
	Top Q	9	14	17	22
	TOTAL	**100**	**100**	**100**	**100**

Notes: See notes to Table 2 above

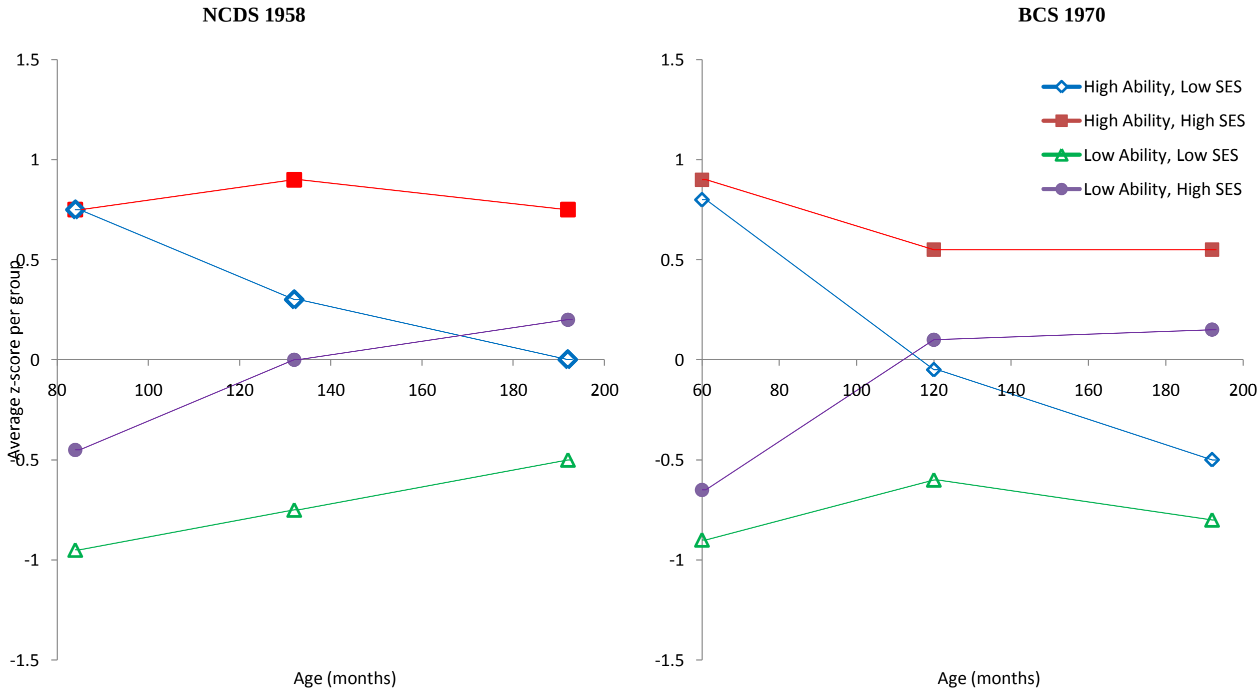
		Simulated test 1			
		Bottom Q	**2nd Q**	**3rd Q**	**Top Q**
Simulated test 2	**Bottom Q**	22	17	14	10
	2nd Q	26	24	22	19
	3rd Q	27	28	28	28
	Top Q	25	31	36	43
	TOTAL	**100**	**100**	**100**	**100**

Figure 1. The development of high and low ability children by socio-economic group – Feinstein (2003)



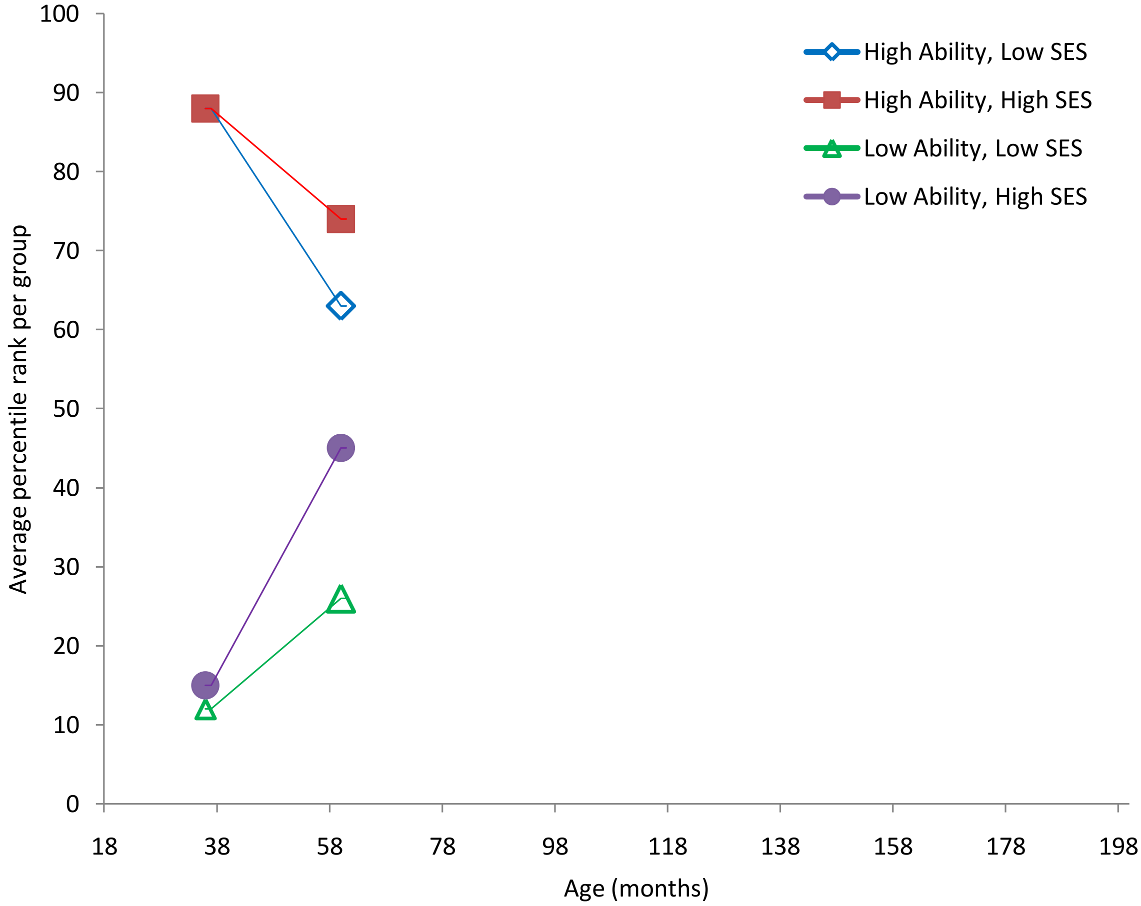
Notes: 1 Figure adapted from Feinstein (2003) Figure 2

Figure 2. The development of high and low ability children by socio-economic group – Schoon (2006)



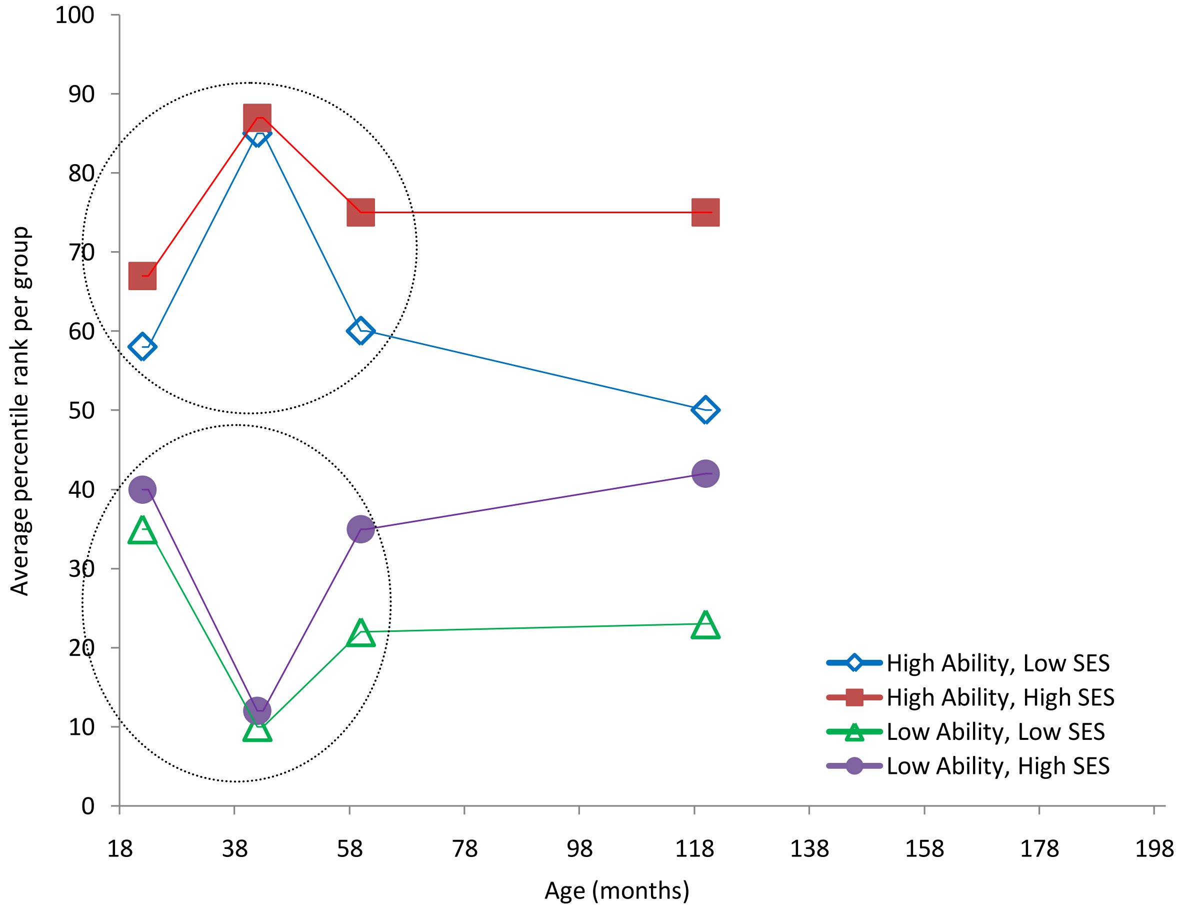
Notes: 1 Adapted from Schoon (2006) page 99

Figure 3. The development of high and low ability children by socio-economic group – Blanden and Machin (2007/2010)



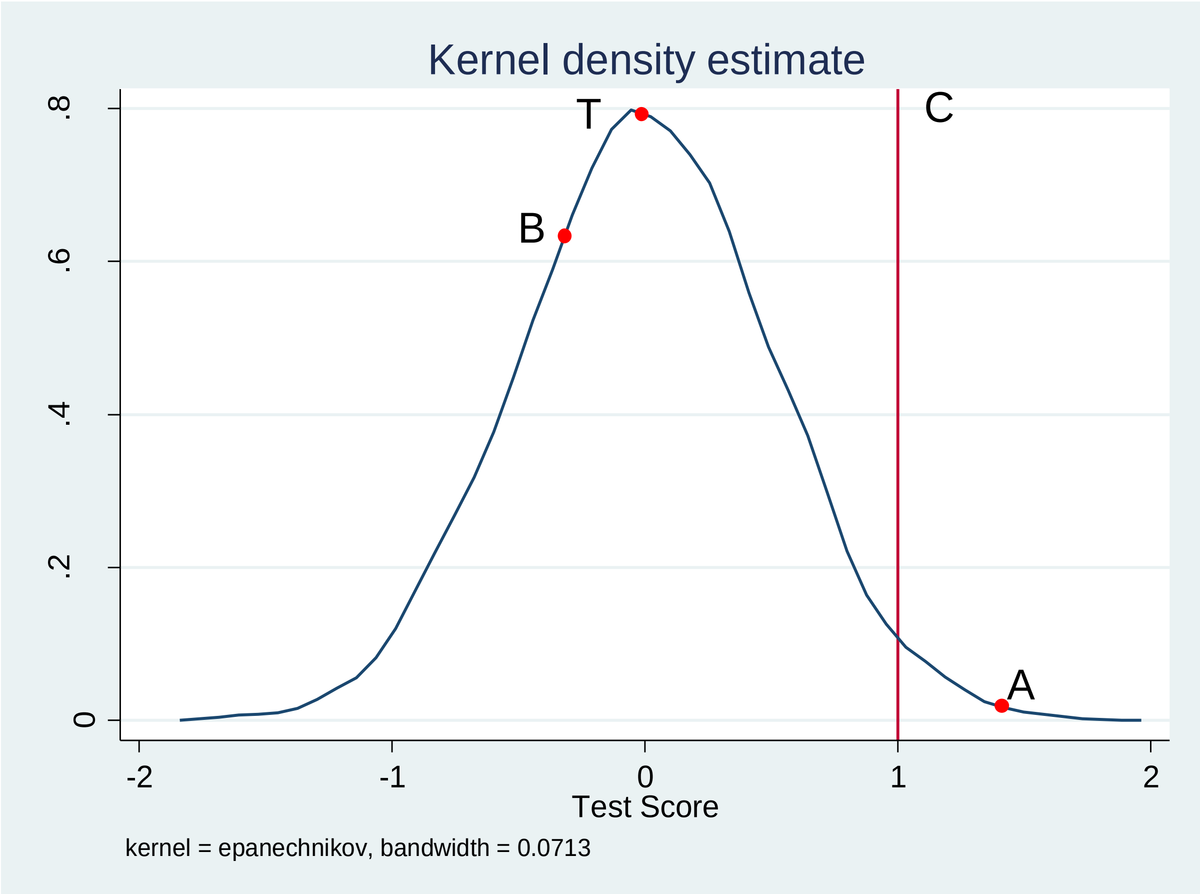
1 Notes: Adapted from Blanden and Machin (2007/2010)

Figure 4. The development of high and low ability children by socio-economic group – Feinstein (2003) when defining ability at 42 months



Notes: 1 Figure adapted from Feinstein (2003) Figure 3

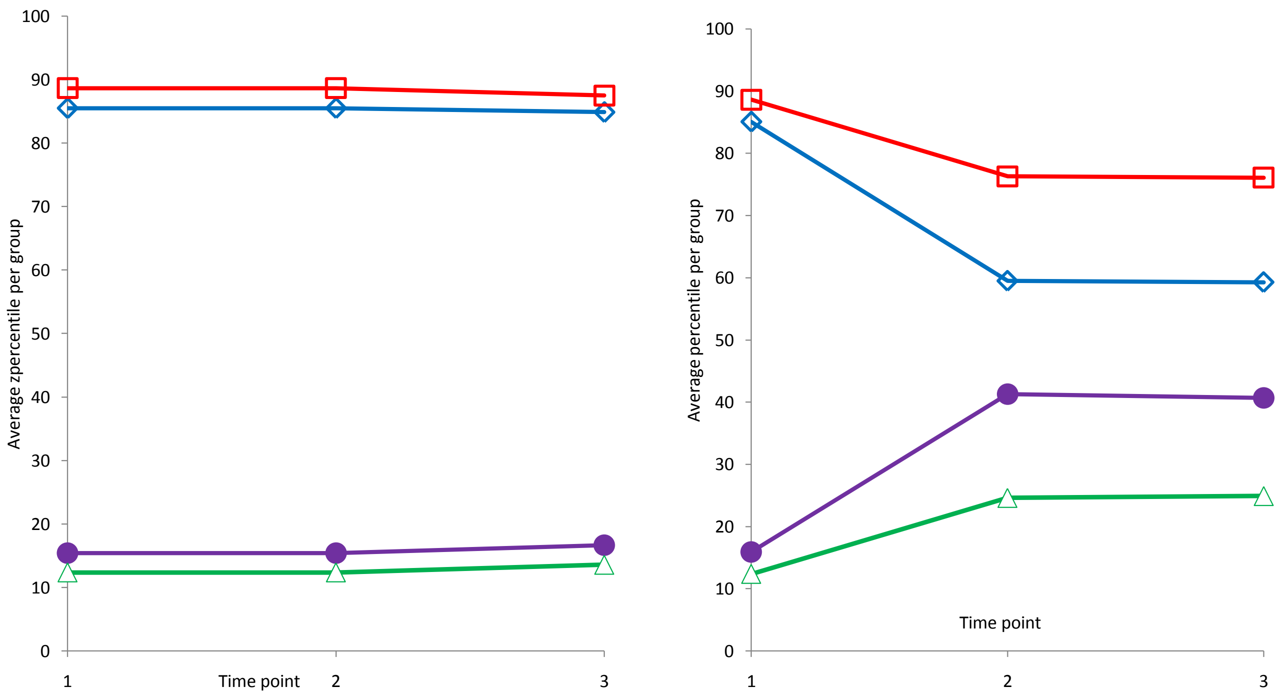
Figure 5. Hypothetical test scores for a child who has experienced “good luck” a test



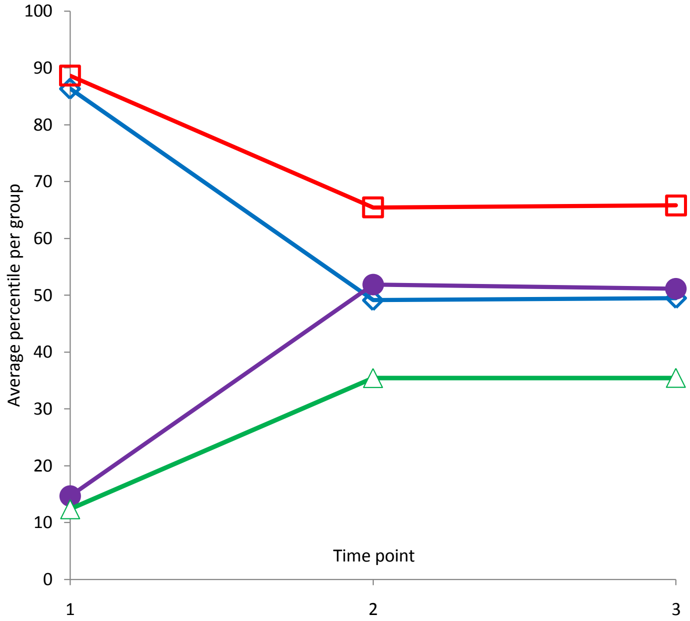
Notes: Figure refers to hypothetical test scores a particular child could receive on a test. T signifies the child’s “true” ability (this is unobserved by the researcher) – which in this example is at the population average (T=0). There is a cut point C, above which children are defined as “high ability”. This child is not really part of this “high ability” group, but happens to have a lot of good “luck” on the day a screening assessment is taken place and scores a mark at point A. As this is higher than the cut-point C, they are mistaken as being of “high ability”. If they were to take a re-test, however, they would be unlikely to score such a high mark (point B). Hence their scores “regress towards the mean”.

Figure 6. Results from our simulation model using existing methodology, when children's true ability does not change over time

(a)"REALITY" | **(b) High test reliability at all 3 points (rho ≈ 0.8)**

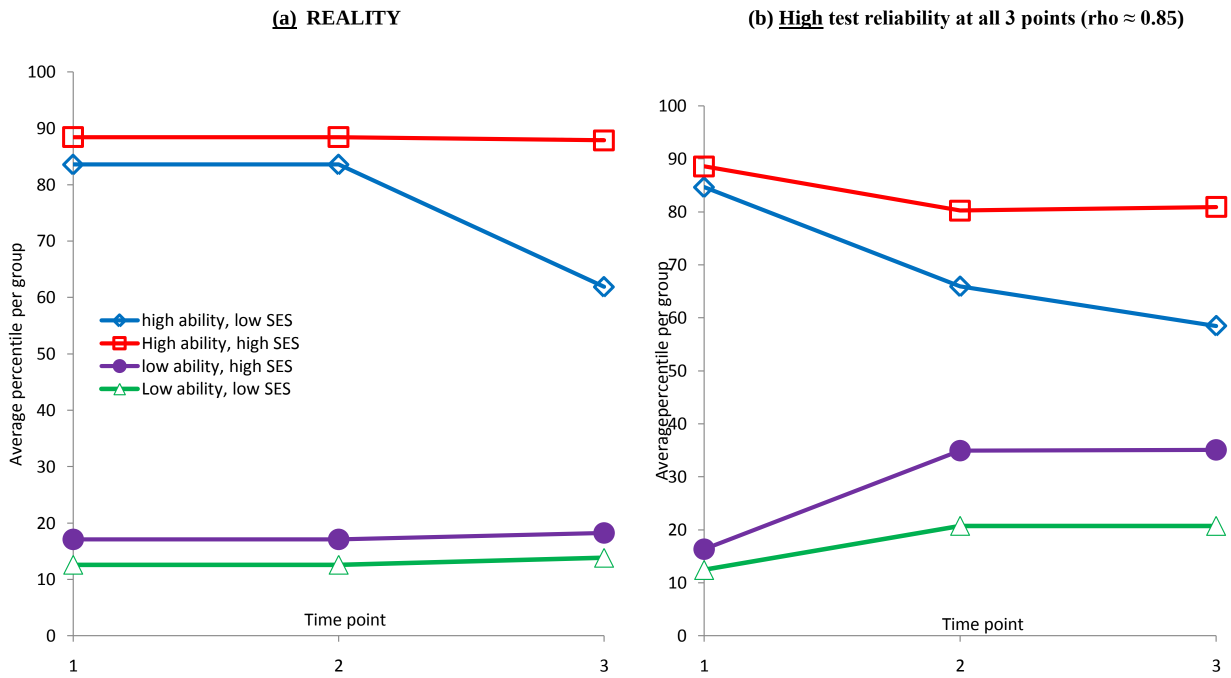


(c) <u>Low</u> test reliability at all 3 points (rho ≈ 0.25)

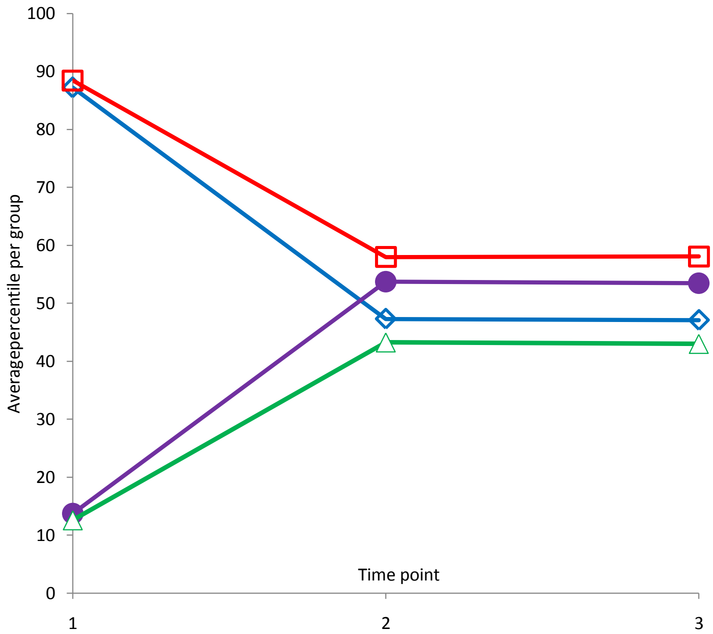


Notes: Diagram produced from our simulated data, described in detail in section 5. Children's (hypothetical) age runs along the x-axis, while the average percentile rank for each group is on the y-axis. Panel A on the left refers to when we can observe children's true ability perfectly (i.e. it is the actual cognitive trajectory of their skill). Panel B refers to what we as researchers observe when applying the methodology that prevails in the existing literature, assuming one is using reasonably accurate tests (panel C extends this by considering what we would observe if using only low accuracy tests).

Figure 7a. Results from our simulation model using existing methodology, when there is a sharp fall in true ability for high ability – low SES children between time points 2 and 3

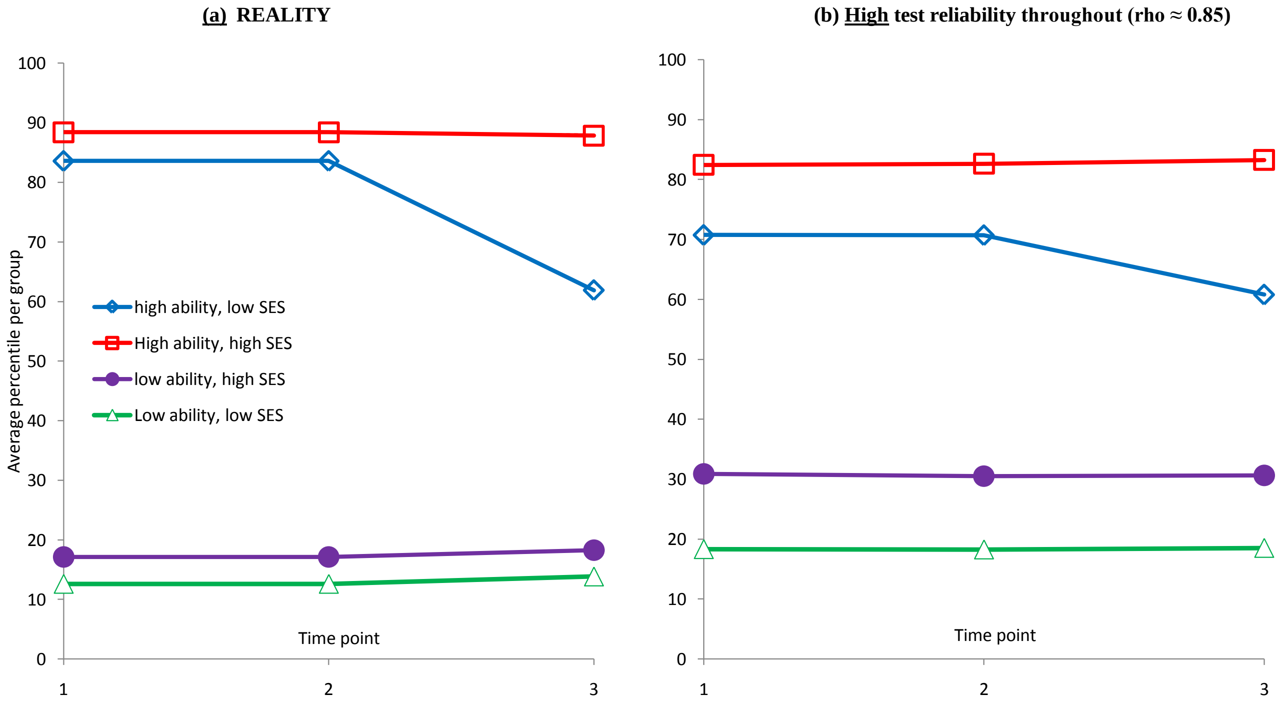


© <u>Low</u> test reliability at all 3 points (rho ≈ 0.25)

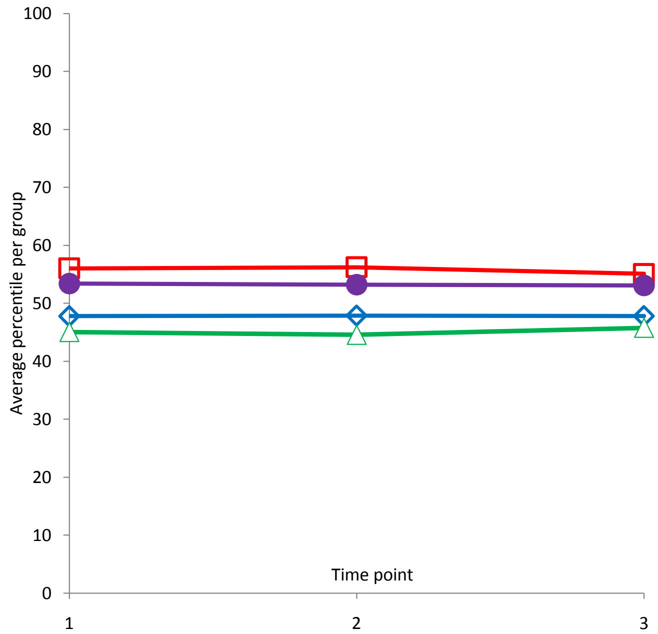


Notes: Diagram produced from our simulated data, described in detail in section 5. Children's (hypothetical) age runs along the x-axis, while the average percentile rank for each group is on the y-axis. Panel A on the left refers to when we can observe children's true ability perfectly (i.e. it is the actual cognitive trajectory of their skill). Panel B refers to what we as researchers observe when applying the methodology that prevails in the existing literature, assuming one is using reasonably accurate tests (panel C extends this by considering what we would observe if using only low accuracy tests).

Figure 8. Results from our simulation model using a single auxiliary test to divide children into ability groups

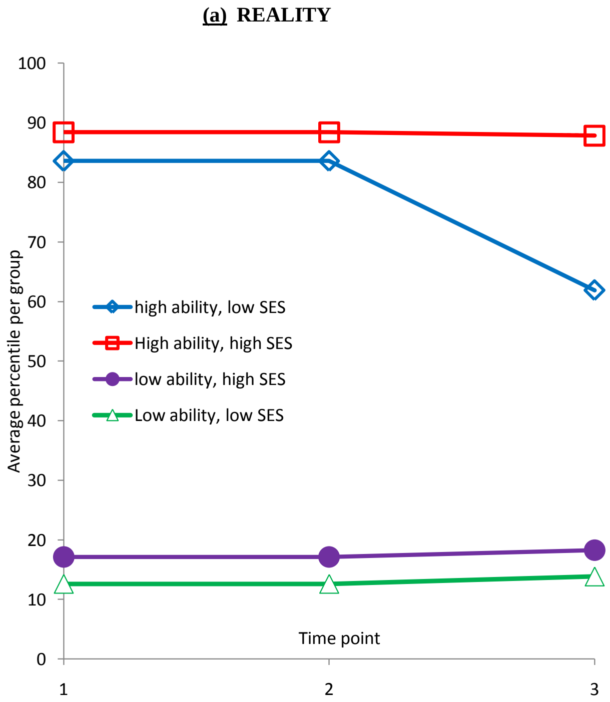


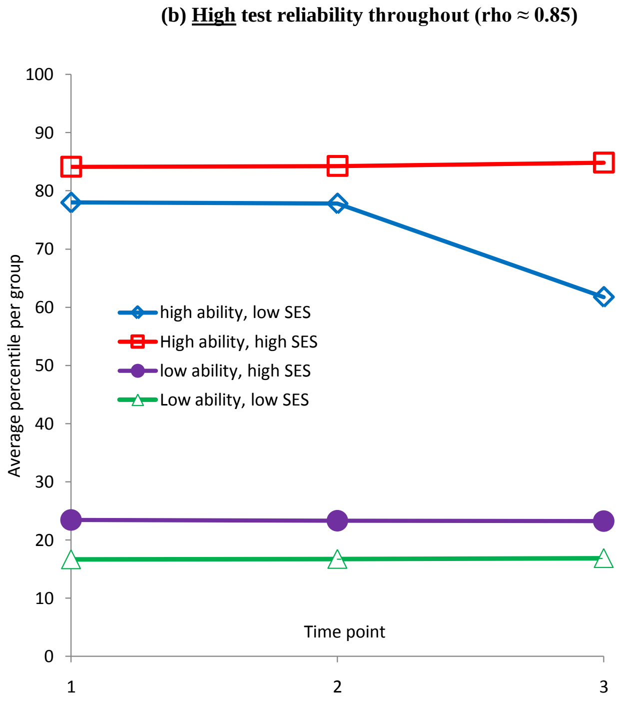
© Low test reliability at all 3 points (rho ≈ 0.25)



Notes: Diagram produced from our simulated data, described in detail in section 5. Children's (hypothetical) age runs along the x-axis, while the average percentile rank for each group is on the y-axis. Panel A on the left refers to when we can observe children's true ability perfectly (i.e. it is the actual cognitive trajectory of their skill). Panel B refers to what we as researchers observe when one has an auxiliary test available which is used to divide children into ability groups and a separate test score from which to measure change from, assuming one is using reasonably accurate tests (panel C extends this by considering what we would observe if using only low accuracy tests).

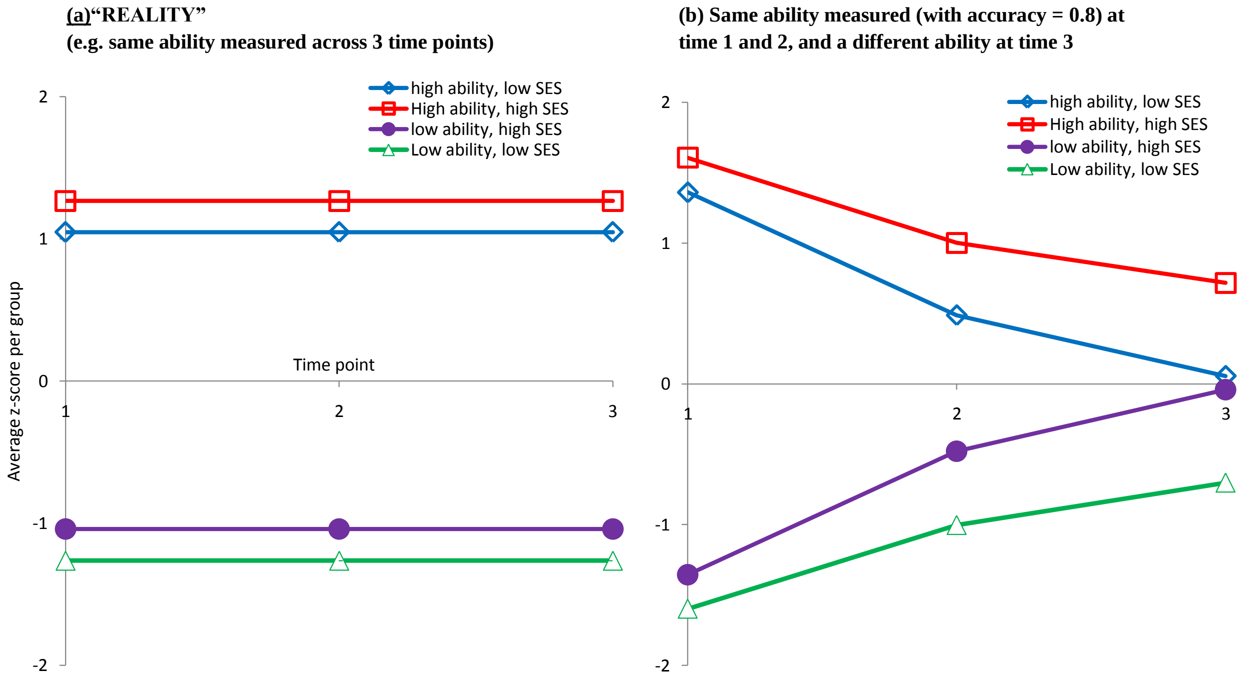
Figure 9. Results from our simulation model using multiple auxiliary tests to divide children into ability groups





Notes: See notes to Figure 7 above, but now the right hand panel are the results for when we have *multiple* auxiliary tests which we can use to divide children into ability groups.

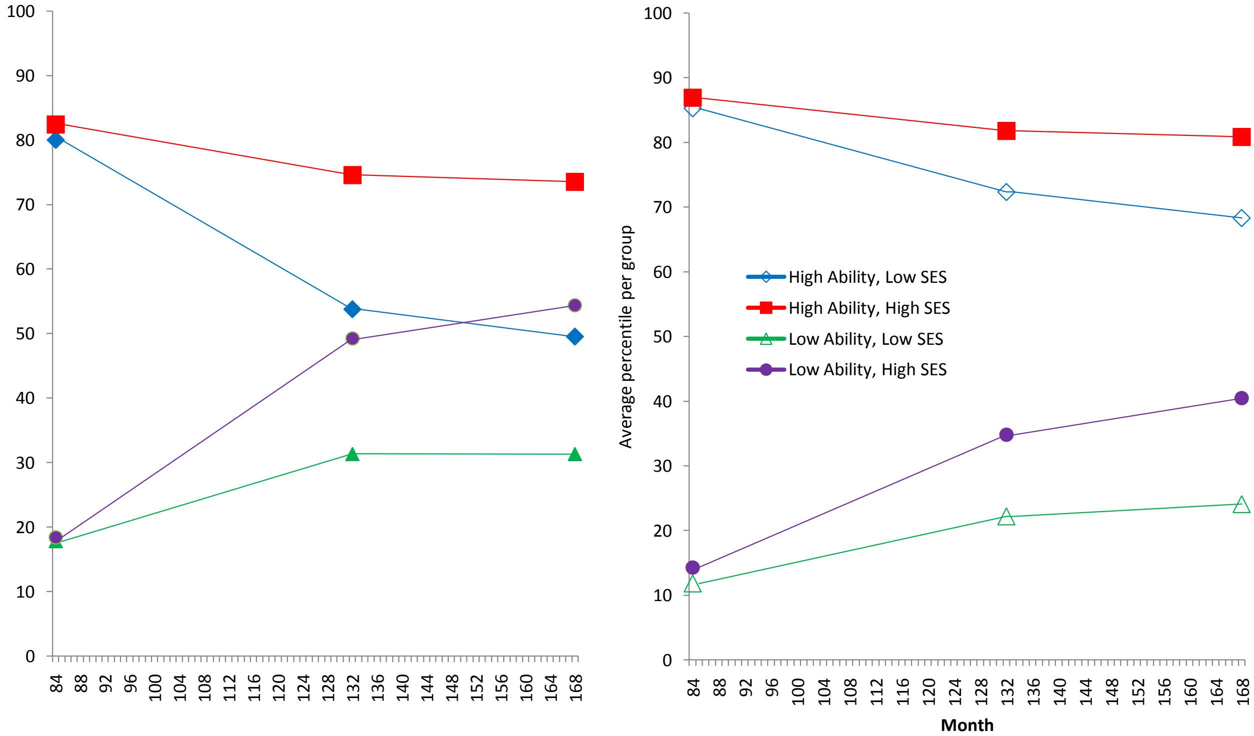
Figure 10. Results from our simulation model when a non-comparable test is used at time 3



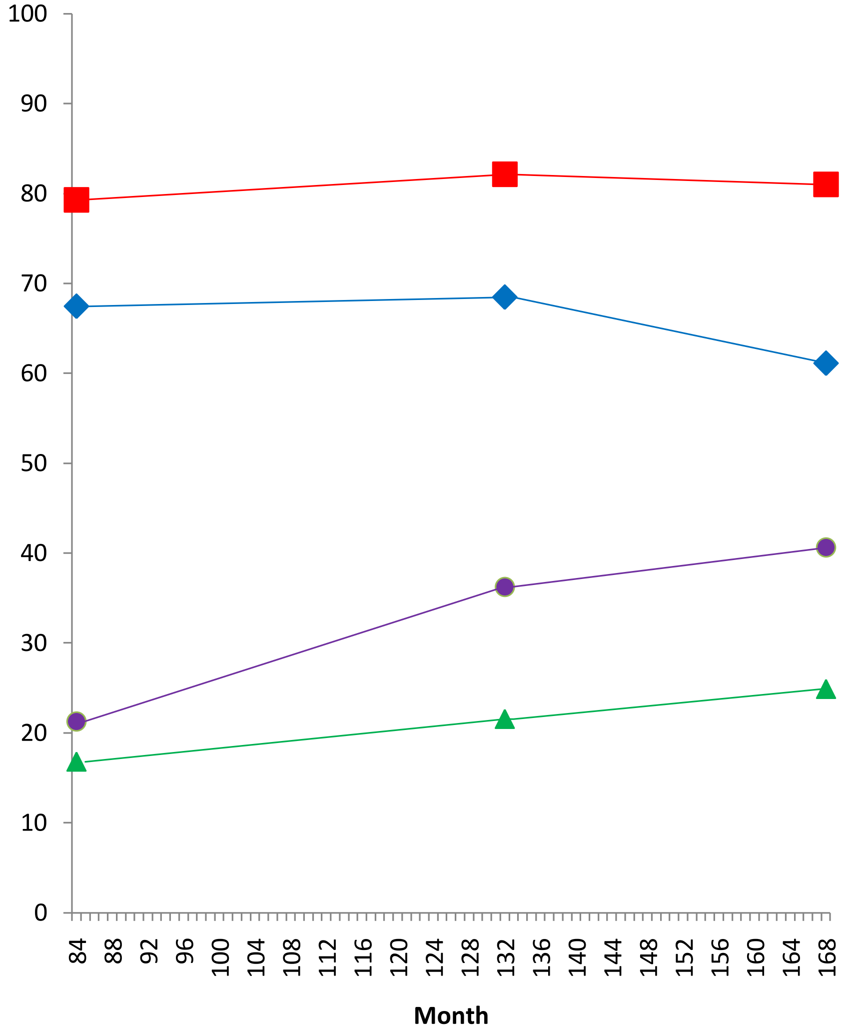
Notes: Diagram produced from our simulated data, described in detail in section 5. Children's (hypothetical) age runs along the x-axis, while the average z-scores for each group is on the y-axis. Panel A on the left refers to when we can observe children's true ability (*in the area we are interested in*) perfectly. Panel B refers to what we as researchers observe when applying the methodology that prevails in the existing literature, assuming one is using reasonably accurate tests, but that one ends up measuring a different skill at time point 3 (i.e. we have a non-comparable test)

Figure 11. Estimated cognitive gradients in ALSPAC using three different methodologies

(a)Combination of Key Stage 1 and motor skills (no RTM adjustment) **(b) Key Stage 1 only (no RTM adjustment)**

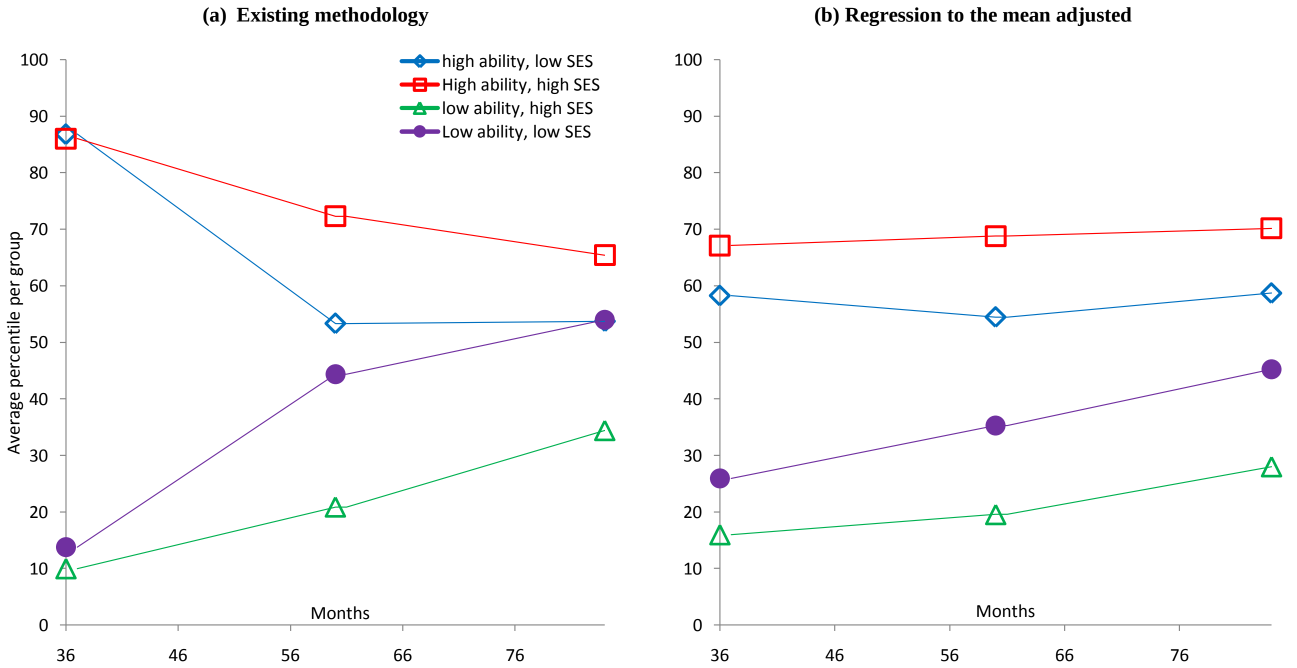


(c) Key Stage English tests used only (RTM adjusted)



Notes: Diagram produced from the ALSPAC data, described in detail in section 6. Children's age in months runs along the x-axis, while the average percentile rank for each group is on the y-axis. Panel A refers to when we use a general indicator of development (a mixture of their Key Stage 1 performance and motor skills) as our first test, and then performance in national English exams (Key Stage 2 and 3) to follow children's performance. In panel B is the same as panel A, except now our first test is based solely upon performance in Key Stage 1 English exams. Finally, panel C is where we use a series of auxiliary tests to divide children into ability groups, with development tracked by their performance on key stage 1 – key stage 3.

Figure 12. . Estimated cognitive gradients in MCS when using different methodologies



Note: Figure 11 provides a set of cognitive trajectories from the MCS. The left hand panel refers to estimates using existing methodology. The right hand panel is the equivalent figures when using Age 3 Bracken test scores as an auxiliary to separate children into high and low ability groups.

Appendix 1. Definitional issues that researchers working in this area face

In this Appendix, we briefly summerise a number of definitional issues that researchers working in this area have faced. Firstly, it is important to make clear what we mean by a child being of "high ability", both in terms of how we measure ability and how we define "high"? One may conceptualise a child being of high ability if they are some pre-determined distance above the population average in a specific skill (e.g. maths, communication, strength, speed) or a more general, multi-dimensional combination of these traits within a given domain (e.g. high levels of maths and communication within a cognitive function domain)[28]. Yet it is unlikely that we can create a single measure that encompasses all children's talents (e.g. cognitive, physical, emotional) without severe information loss. Hence the meaning of "high ability" or "talent" in empirical analysis is often restricted to mean a high achiever in a given domain – such as cognition as measured by an IQ test. If we are interested in the *development* of this group, it is thus important that the same skill is measured over time (e.g. via repeated measures of IQ at different ages). It would be unclear, for instance, what defining high ability on a cognitive measure in period 1, then assessing the child on physical skill from period 2 onwards, would tell us about development[29]. In practise, however, this ideal can rarely be achieved. As such, researchers who have estimated socio-economic differences in cognitive gradients have tended to use whatever measures they have available (e.g. Goodman et al 2009, Feinstein 2003).

Likewise, the age at which to define a child as "high ability" is far from clear cut. Studies of habituation (infants response to a visual stimuli) from the psychological literature suggest that one can collect a key predictor of later IQ from children as young as 6 months old (Kavsek 2004) – although others are less confident (Slater 1997). Feinstein (2003), citing Zeanah et al. (1997), notes that there are three periods of "structural reorganisation" during infancy, the last of which occurs at 20 months, after which time changes are more suitable for quantitative assessment. In a similar manner Cunha et al (2006), drawing on the neuroscience literature, suggests that cognitive tests (such as IQ) are only suitable when children are around age 4 or 5[30]. Consequently, researchers in this area face a trade-off. Defining a child as "high ability" with a very early measure means one can capture changes from a young age, but results from such tests are often unstable and (some would claim) unreliable. Alternatively, one can measure children's progress from later ages but, in doing so, potentially miss

---

[28] The 'g-factor' is a well-known version of the latter, which combines children's scores on different forms of cognitive tests to generate an overall measure of "intelligence".
[29] One also requires that such follow-up tests are conducted regularly and over a long range of time.
[30] Cunha et al (2006) suggest IQ measures recorded before age 4 or 5 are poor indicators of intelligence in adulthood. Nevertheless, IQ tests have been developed for use on children from as young as 2.5 years. The Wechsler Preschool and Primary Scale of Intelligence (WPPSI) is one example.

out on a key stage of their development (i.e. the early years that Cunha et al (2006) and others stress are the most important).

Another issue that researchers face is how to define “advantaged” and “disadvantaged” backgrounds; is a single variable (such as parental income, education or social class) sufficient or do we require a multi-dimensional measure that attempts to capture this concept in a broader sense (Chowdry et al, 2009)? We shall not describe the merits of these different approaches here, but simply note that there is again no universally accepted convention in the literature. The consequence is that how one measures “advantage” and “disadvantage” is not straightforward, and open to debate.

What we hope this short description has highlighted is that even defining our primary group of interest (high ability – disadvantaged children) is not trivial, and whatever one settles on could be disputed by others working in the field. This problem is exacerbated by the fact that datasets containing all the necessary information are extremely rare. The existing literature is, consequently, rather inconsistent on the definition of “high ability” and “disadvantage”, the age at which children are followed-up and the tests that have been used. Indeed, in most studies it would seem the choices made have been largely dictated by the availability of the data.

Appendix 2. A description of the test measures we use from the ALSPAC age 7 clinic

As part of the ALSPAC study, all participants were invited to attend a special clinic session at roughly 7.5 years of age. As part of this clinic, children were examined in their basic reading, phoneme deletion and spelling skills by trained psychologists and speech therapists. The exact details of these tasks are given below:

(a) Reading test

This was assessed using the basic reading subtest of the "WORD" (Wechsler Objective Reading Dimensions). Firstly, the child was shown a series of four pictures, which had four short, simple words underneath them. They then had to point to the word which had the same beginning or ending sound as the picture. Following this, the child was shown a series of three further pictures, each with four words beneath, each starting with the same letter as the picture. The child was asked to point to the word that correctly named the picture. Finally, the child was asked to read aloud a series of 48 unconnected words which increases in difficulty. This reading task was stopped after the child had made six consecutive errors.

(b) Spelling test

Children were given 15 words to spell. The words were chosen specifically for this age group after pioleting on several hundred children in Oxford and London. They were put in order of increasing difficulty based on results from the pilot study. For each word, the member of staff first read the word out alone to the child, then within a specific sentence incorporating the word, and finally alone again. The child was then asked to write down the spelling.

Children were awarded three points for a correct answer, two points if their response was incorrect but they spelt the word phonetically, one point if the spelling had one "sound" (a vowel sound) wrong, and zero otherwise.

(c) Language test (phoneme deletion task)

The phoneme deletion task, known as the word game in the session, was the Auditory Analysis Test developed by Rosner and Simon (1971). The task comprised 2 practise and 40 test items of increasing difficulty. The task involved asking the child to repeat a word and then to say it again but with part of the word (a phoneme or number of phonemes) removed. For example, the child was asked to say 'sour' and then say it again without the /s/ to which the child should respond 'our'. There were seven categories of omission: omission of a first, a medial or a final syllable; omission of the initial, of the final consonant of a one syllable word and omission of the first consonant or consonant blend of a medial consonant. Words from the different categories were mixed together but were placed in order of increasing difficulty.

As part of the clinic, children's motor abilities were also assessed via the movement ability assessment for children. When referring to children's "motor skills", we are using two tests of children's manual dexterity that were contained within ALSPAC:

(d). The peg game

In the placing pegs task (known in the clinic as the peg game), the child had to insert twelve pegs, one at a time, into a peg board, holding the board with one hand and inserting the pegs with the other, as quickly as possible. The task was carried out with the preferred and the non-preferred hand, after it had been described and demonstrated by the tester, and after a practice attempt with each hand. The time it took them to complete this task with their better hand is taken as our first indicator of children's motor skills/manual dexterity.

(e). The string game

This task involved children threading lace through a wooden board. The exact task was demonstrated by the tester and the child was given a practice attempt. The time it took them to complete this task is taken as our second indicator of children's motor skills/manual dexterity.

The correlation matrix between each of these tests and children's key stage 1 total points score can be found in Appendix Table 1 below. One can see that the three clinical tests of children's reading, spelling and language ability all correlate reasonably highly with their key stage 1 score. There is, on the other hand, almost no association between children's motor skills at this age and their outcome on national exams.

Appendix Table 1. Correlation matrix of ALSPAC language based tests and children's key stage 1 points score

	Reading	Spelling	Language	Motor (String)	Motor (Peg)	KS 1
Reading	1					
Spelling	0.75	1				
Language	0.68	0.60	1			
Motor (String)	-0.09	-0.09	-0.07	1		
Motor (Peg)	-0.01	-0.02	0.00	0.03	1	
KS 1	0.70	0.62	0.54	-0.14	-0.02	1