

Social mobility, regression to the mean and the cognitive development of high ability children from disadvantaged homes

John Jerrim

Anna Vignoles

Institute of Education, University of London

July 2012

Abstract:

A gap in cognitive skill between richer and poorer children is evident from a very early age. Some studies have also suggested 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. The latter finding has become a widely cited “fact” within the academic literature, and has had a major influence on public policy and political debate. We show that this finding is vulnerable to a spurious statistical artefact known as regression to the mean (RTM) and we propose the application of an alternative methodology to address this problem. After applying some simple adjustments for RTM to data from the Millennium Cohort Study, we no longer find convincing evidence that able but disadvantaged pupils fall behind their more advantaged but less able peers.

Key Words: disadvantaged children, educational mobility, Millennium Cohort Study, regression to the mean, socio-economic gap.

Acknowledgements: We would like to thank John Micklewright for particularly helpful comments on an initial draft, along with the journal editor, associate editor and two anonymous referees. We also acknowledge the extremely helpful feedback we received from participants at seminars at the University of Bristol, University of Sussex, Institute of Education and the University of Southampton. This work has been produced as part of the ESRC ALSPAC large grant scheme.

Contact Details: John Jerrim (J.Jerrim@ioe.ac.uk) and Anna Vignoles (A.Vignoles@ioe.ac.uk), Department of Quantitative Social Science, Institute of Education, University of London, 20 Bedford Way London, WC1H 0AL

Children from disadvantaged backgrounds have poorer cognitive skills than their more advantaged peers, and such differences are apparent from a very early age. On this point the empirical evidence seems conclusive (Cunha et al 2006, Goodman et al 2009). There has been much debate, however, as to whether this socio-economic achievement gap widens as children age. This may occur if inputs into children's development are complementary (see Cunha et al 2006), such that greater early investment by high SES parents leads to a greater return on subsequent investments. It may also happen if the better social skills of high SES children impact positively on the development of their cognitive skills, allowing children from affluent backgrounds to extend their early cognitive lead over their disadvantaged peers. A number of studies have found that socio-economic achievement gradients do indeed increase through childhood (e.g. Goodman et al 2009, Feinstein 2003) although others suggest that there is actually rather little change (Blanden and Machin 2010, Reardon 2011, Duncan and Magnuson 2011). The empirical evidence on this matter is far from conclusive.

There is one area, however, where socio-economic differences have been consistently found to have become more pronounced as children age. Feinstein (2003), Feinstein (2004), Schoon (2006), Blanden and Machin (2007), Blanden and Machin (2010) and Parson et al (2011) have all shown that initially able children from disadvantaged backgrounds quickly lose ground in terms of their cognitive skills to their rich but less able peers. These studies adopt a similar empirical method in reaching this conclusion. Drawing upon longitudinal birth cohort data, children are separated into "ability" groups based on a test taken at approximately 2 or 3 years of age. For this test and in a series of follow-up tests, each child is given a score between 1 and 100 based on their percentile of the test score distribution. For some studies standardised z-scores have been used instead. An average score is then calculated at each age for the following four groups, having first defined socio-economic status (SES) on the basis of parental education, occupation or permanent household income:

- (a) High ability-high SES (b) High ability-low SES
- (c) Low ability-high SES (d) Low ability-low SES.

A graph similar to Figure 1 is then often produced.

Figure 1

Notice how at the first time point, 22 months in this example, both high ability-high SES and high ability-low SES children are at the same point, roughly the 88th percentile of

the achievement distribution. But, by the time of the second assessment, taken at 42 months, the latter group has slipped to the 55th percentile and, at the final time point, 120 months, to the 40th percentile. On the other hand, high ability children from advantaged homes remain much higher up the test score distribution, 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 period. Thus the conclusion often reached is that initially able children from poor backgrounds are overtaken in terms of their cognitive skill by low ability high SES children before they start secondary school, or earlier in some data.

This finding, having been replicated in a number of studies using the same method, has had a significant impact on both academic research and public policy in Britain over the last decade. It has become routinely cited in fields as diverse as economics, sociology, medicine and child development, with graphics like Figure 1 being prominent in major national reviews of Poverty and Life Chances by Frank Field (Field 2010), the Marmot Review of Inequalities in Health (Marmot 2010) and the recently released Social Mobility Strategy from the coalition government (Cabinet Office 2011). Even the Deputy Prime Minister Nick Clegg (April 2011) has described how:

“By the age of five, bright children from poorer backgrounds have been overtaken by less bright children from richer ones—and from this point on, the gaps tend to widen still further”.

Nick Clegg in a House of Commons debate announcing the launch of the UK government’s social mobility

But is the statistical methodology lying behind this result robust? In this paper we argue that the method being used to study this topic does not allow one to separate out statistical error from genuine policy relevant change due to the well-known, but often misunderstood, problem of regression to the mean (RTM) – see Galton (1886). After discussing problems with the existing method, we draw upon the work of Ederer (1972) and Davis (1974) to propose an alternative. This is followed by an empirical application to the Millennium Cohort Study (MCS) dataset. We contribute to the existing literature by:

- (a) Demonstrating that the method that has been used to study this topic is flawed, and cannot therefore inform public policy;
- (b) Proposing an alternative method that is original in its application to this substantive topic;

(c) Discussing the limitations of this alternative method and setting out what this implies for the measurement of the educational achievement trajectories of different SES and ability groups;

(d) Producing, to the best of our knowledge, the first piece of empirical evidence on this important issue that has explicitly taken the problem of regression to the mean into account.

The paper proceeds as follows. We begin in section 2 by discussing what is meant by regression to the mean, how it can emerge, and potential ways of correcting for this problem. In section 3, we demonstrate how existing research in this area is plagued by the RTM issue using simulated data, before describing a simple solution. An empirical example using the Millennium Cohort Study (MCS) dataset follows in section 4, with conclusions in section 5.

2. The problem of regression to the mean

In this paper, we focus on regression to the mean that is caused by conditioning children's baseline test scores on initial test error. We will only briefly discuss regression to the mean that can occur through other routes, such as the use of non-comparable tests, missing data and sample selection. An on-line working paper (Jerrim and Vignoles 2011) does, however, provide further discussion of these issues.

2.1. Regression to the mean caused by conditioning on initial test error

Regression to the mean due to test error 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 test score 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).

Consider, for instance, a child whose true ability is average. We cannot directly observe this, but rather must estimate it from how they perform on a test. Even though the child is only of average ability, it is still possible that they can obtain a good mark on this assessment and get mistaken for a high achiever. What, then, would we expect to happen if this child was to be re-tested a short time later? They would probably receive a lower mark

that is a better reflection of their true ability or, in other words, their score would regress towards the mean. The same problem occurs when classifying children into ability groups across a population. By using a set cut-off point, such as the top quartile, on a single test, this selection will include some individuals with a large positive error who are unlikely to have such good fortune when it comes to re-assessment, and will hence subsequently score, on average, a mark closer to their true value.

This suggests that groups initially identified as “high ability” will exhibit apparently falling levels of achievement over time due to statistical error. 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. This is due to an additional problem. On average, there are genuinely large gaps in early cognitive skill between children from advantaged and disadvantaged homes, and hence low SES children who get defined as “high ability” are more likely to have 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. Similarly, low ability children from high SES backgrounds have had a particularly large random negative error, and thus experience greater upward regression to the mean than their low ability low SES peers. This has not been fully recognised as a possible reason for the low SES children’s striking decline in test scores observed in the literature.

2.2. Statistical model

We further illustrate this argument with the use of a statistical model. In doing so, discussion shall focus upon 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. To start, let:

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

Where:

A_{it} = the child’s “true” ability at time t

ξ_{it} = Error in measuring the child’s true ability at time t

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

Assume:

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

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

and that $\text{corr}(A_{it}, \xi_{it})=0$ and $\text{corr}(\xi_{it}, \xi_{it+1})=0$.

Now say we want to divide children into ability groups at time point 1. Ideally, we would be able to observe children's true ability (A_{i1}) which could then be used to divide children into ability groups. The average level of true ability, within this selected high ability group, would then be:

$$E(A_{i1}|A_{i1} > K_1) = \mu_1 + C_1\delta_1 \quad (1)$$

Where:

A_{i1} = True ability of individual i at time $t=1$

μ_1 = The population average ability at time $t=1$

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

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

$$\Phi(a_1) = \int_{-\infty}^{a_1} \phi(x) \cdot dx$$

$a_1 = \frac{(K_1 - \mu_1)}{\delta_1}$ = The standardised threshold above which children are defined as high ability

K_1 = The threshold above which children are defined as “true” high ability

δ_1 = Standard deviation of “true” ability at time $t=1$.

We do not, of course, observe whether children's true ability sits above a certain threshold. Rather one can only observe their score on a test (Y_{i1}). Even though this test maybe unbiased $E(\xi_i)=0$, there is still variability (γ) in its error. Rather than assigning children into a high ability group based on their “true ability”, we do so based on their test score. They get labelled high ability if $Y_{i1} > K_1$.

The expected average score on this test for the group we now define as “high ability” is:

$$E(Y_{i1}|Y_{i1} > K_1) = \mu_1 + C_1 \sqrt{\delta_1^2 + \gamma_1^2} \quad (2)$$

Notice that this expectation contains the parameter γ_1^2 , the variance of the error of the test. Now consider what the average true ability level of this group is, namely of those who we define as high ability based upon their initial test score:

$$E(A_{i1}|Y_{i1} > K_1) = \mu_1 + \Omega_1 \cdot C_1 \cdot \sqrt{\delta_1^2 + \gamma_1^2}$$

Where:

$$\Omega_1 = \frac{\delta_1^2}{\delta_1^2 + \gamma_1^2} = \text{The accuracy of the test}$$

which simplifies to:

$$E(A_{i1}|Y_{i1} > K_1) = \mu_1 + \frac{C_1 \cdot \delta_1^2}{\sqrt{\delta_1^2 + \gamma_1^2}} \quad (3)$$

Now consider the difference between equations (2) and (3). This represents the difference between average true and average observed ability for the “high ability” group, i.e. for those whose tests scores are above the threshold K :

$$\begin{aligned} & E(Y_{i1}|Y_{i1} > K_1) - E(A_{i1}|Y_{i1} > K_1) \\ &= \mu_1 + C_1 \sqrt{\delta_1^2 + \gamma_1^2} - \mu_1 - \frac{C_1 \cdot \delta_1^2}{\sqrt{\delta_1^2 + \gamma_1^2}} \\ &= C_1 \cdot \gamma_1^2 \end{aligned} \quad (4)$$

Equation 4 is the difference between what we observe the average ability level amongst the “high ability” group to be and their actual ability. Notice that the variance of the error on the test in the first period (γ_1^2) is one of the key parameters, and represents the fact that the high ability group has been partly based upon those who had a good luck draw on the day of the test.

Now consider children's scores on a follow-up test. What is the expected value of scores on this second assessment, given that their first test was above the cut-off? If one assumes that errors between tests are uncorrelated, $\text{corr}(\xi_{it}, \xi_{it+1}) = 0$, then:

$$E(A_{i2}|A_{i1} > K_1) = \mu_2 + C_1 \cdot \rho_{12} \cdot \delta_1 \quad (5)$$

Where:

ρ_{12} = The correlation between children's true ability in period 1 and period 2

Hence the true change in children's ability between time 1 and 2 is:

$$\begin{aligned} E(A_{i1}|A_{i1} > K_1) - E(A_{i2}|A_{i1} > K_1) &= \\ &= \mu_1 + C_1 \delta_1 - \mu_2 - C_1 \cdot \rho_{12} \cdot \delta_1 = (\mu_1 - \mu_2) + C_1 \delta_1 (1 - \rho_{12}) \end{aligned} \quad (6)$$

But we can only observe the change in their test scores:

$$\begin{aligned} E(Y_{i1}|Y_{i1} > K_1) - E(Y_{i2}|Y_{i1} > K_1) &= \\ &= (\mu_1 - \mu_2) + C_1 \cdot \sqrt{\delta_1^2 + \gamma_1^2} - C_1 \cdot \rho_{12}^* \cdot \delta_1 \end{aligned} \quad (7)$$

Where:

ρ_{12}^* = The observed correlation between children's test scores in period 1 and period 2

Under the assumption that $\rho_{12} = \rho_{12}^*$, such that the correlation between the tests we use is an accurate reflection of the correlation between children's true ability over time, then:

$$\begin{aligned} \{E(A_{i1}|A_{i1} > K_1) - E(A_{i2}|A_{i1} > K_1)\} - \{E(Y_{i1}|Y_{i1} > K_1) - E(Y_{i2}|Y_{i1} > K_1)\} &= \\ &= \mu_1 + C_1 \delta_1 - \mu_2 - C_1 \cdot \rho_{12} \cdot \delta_1 - (\mu_1 - \mu_2) - C_1 \cdot \sqrt{\delta_1^2 + \gamma_1^2} + C_1 \cdot \rho_{12}^* \cdot \delta_1 \\ &= C_1 \delta_1 - C_1 \cdot \sqrt{\delta_1^2 + \gamma_1^2} \end{aligned} \quad (8)$$

This is the difference between what we want to know, the true change in “high ability” children's skill over time, and what we actually observe, the change in their test scores over time. Note the presence of the error variance from the first test (γ_1^2) in equation (8) above, which causes RTM to be a problem.

From this equation, we can also see that there will be no RTM effect due to conditioning upon initial test error when using the above strategy if either of the following conditions holds:

$$\gamma_1=0$$

or

$$C_1=0.$$

The first ($\gamma_1=0$) requires the error variance on the first test which we use to select children to be equal to zero. If, however, the error variance on this test is non-zero, as is always the case in real life analysis, regression to the mean due to statistical error will always occur, and we will observe a decline in the high ability group's test scores even if no genuine change has taken place.

Also note from this that:

RULE 1: The regression towards the mean effect due to conditioning upon initial test error gets bigger as the variance of this error increases.

As most cognitive ability tests of young children are not particularly powerful, and thus the error variance tends to be reasonably big, the above is likely to be a significant problem when assessing children's early cognitive development.

Now consider the second condition above (the parameter C_1). Note that:

$$RTME_{12} \rightarrow \infty \quad \text{as} \quad C_1 \rightarrow \infty$$

And that:

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

Regression towards the mean will be greater when the cut-point used to divide individuals into extreme groups is further from the population average. Now assume there are two types of children – Low SES (L) and High SES (H). Many studies from the UK and US have shown empirically that cognitive skill test scores differ between socio-economic groups even at a very early age. This reflects a combination of factors, including the role of genetic/environment interactions, the fact that the early years appears to be a critical and sensitive period for investing in children's cognitive development and more generally the

evidence that family background has the greatest impact on children's outcomes during the first years of life (Cunha et al 2006, Goodman et al 2009). 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:

$$|K_1 - \mu_1^H| < |K_1 - \mu_1^L|$$

And hence:

$$C_1^H < C_1^L$$

Under the assumption that:

$\gamma_t^H \approx \gamma_t^L$ 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 for the high ability – high SES group.

RULE 2: 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.

2.3. A method of accounting for regression to the mean that is caused by initial test error

We now describe a method that attempts to correct for the problem set out above. This 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). It requires that one has at least two measures of the construct of interest at the baseline time point (t_1). The first of these tests (Y_{i1}) is used to divide children into ability groups. The second (Y_{i1}^*) is then taken as the baseline observation from which change is measured from.

Why does this overcome the RTM caused by conditioning on initial test error? One can see that the RTM is being caused entirely by the parameter γ . This enters the expected value given in equation (7) which, in turn, also enters equation (8). Recall that γ is the “luck” that the child happened to have on the day of the test. However, if a second baseline test score (Y_{i1}^*) is available, and the error on this test is independent of the error on the first test (Y_{i1}), then the expected value of the second baseline test given that the first one is above a specific value is:

$$E(Y_{i1}^* | Y_{i1} > K_1)$$

And thus, the expectation analogous to that given in equation (7) becomes:

$$\begin{aligned} & E(Y_{i1}^* | Y_{i1} > K_1) - E(Y_{i2} | Y_{i1} > K_1) \\ &= (\mu_1 - \mu_2) + C_1 \cdot \sqrt{\delta_1^2} - C_1 \cdot \rho_{12}^* \cdot \delta_1 \\ &= (\mu_1 - \mu_2) + C_1 \cdot \delta_1 \cdot (1 - \rho_{12}^*) \end{aligned} \tag{9}$$

Notice that, as one is no longer conditioning the expected value on test error, γ does not enter equation (9). In turn equation 8, the difference between the true change in “high ability” children’s skill over time and the change we observe in their test scores, becomes:

$$\begin{aligned} & \{E(A_{i1} | A_{i1} > K_1) - E(A_{i2} | A_{i1} > K_1)\} - \{E(Y_{i1}^* | Y_{i1} > K_1) - E(Y_{i2} | Y_{i1} > K_1)\} \\ &= [(\mu_1 - \mu_2) + C_1 \cdot \delta_1 \cdot (1 - \rho_{12})] - [(\mu_1 - \mu_2) + C_1 \cdot \delta_1 \cdot (1 - \rho_{12}^*)] \\ &= 0 \quad (\text{under the previously stated assumption that } \rho_{12} = \rho_{12}^*) \end{aligned}$$

The fact that the change in high ability children’s true skill minus the change observed in their test scores is now equal to 0 illustrates that the bias from regression to the mean due to test error has been eliminated, and that this is no longer a problem affecting the estimates of how the cognitive skill of high ability – low SES children changes over time.

There are, of course, some limitations to this approach. The first is the assumption that the errors on the two baseline tests (Y_{i1} and Y_{i1}^*) are uncorrelated. This may not be the case in real life analysis if examinations are held close together, where short term factors influences the marks on both tests. Under such conditions the method proposed will only reduce RTM due to initial test error and not entirely eliminate it. Thus one will in effect be

calculating an upper bound for the amount of change there is over time rather than a point estimate. Appendix 1 discusses this issue in detail, including the sensitivity of results to violation of the independent test error assumption.

The second limitation is that there remains a “misclassification” problem; that children continue to be labelled as “high ability” based on a single test meaning that some individuals are placed into the wrong “ability” group. The implication of this is that the “high ability” group that is tracked over time is actually a mix of true high ability children and those who have been wrongly given this label. The impact this has on estimates depends on how the developmental trajectories of the children mistakenly identified as high ability differs to those of the true high ability children. In section 3.3 we provide an example via simulation of how this can, under certain assumptions, lead to attenuation. It is important to understand, however, that this is always likely to be a problem as one is unlikely to ever be able to correctly classify all children into the right ability group, and that this remains a problem even after RTM has been taken into account. We will discuss this issue further in section 3.3.

2.4. Regression to the mean beyond period 2

Thus far, we have described how the problem of regression to the mean from conditioning on initial test error can lead to the mistaken conclusion that there is a big decline in high ability – low SES children’s cognitive skills between time point 1 and time point 2 when this may not actually be the case. Can this also explain why some studies (e.g. Feinstein 2003, Schoon 2006) have found a continuing decline past the first follow-up test? Following similar logic to that presented in the sub-section above, one can see that the change in children’s true ability between time 2 and 3 equals:

$$\begin{aligned}
& E(A_{i2}|A_{i1} > K_1) - E(A_{i3}|A_{i1} > K_1) \\
&= (\mu_2 + C_1 \cdot \rho_{12} \cdot \delta_1) - (\mu_3 + C_1 \cdot \rho_{13} \cdot \delta_1) \\
&= (\mu_2 - \mu_3) + C_1 \cdot \delta_1 \cdot (\rho_{12} - \rho_{13})
\end{aligned} \tag{10}$$

While the change we observe in their test scores, assuming errors on the tests at time 2 and time 3 are independent, is:

$$\begin{aligned}
& RTME_{23} = E(Y_{i2}|Y_{i1} > K_1) - E(Y_{i3}|Y_{i1} > K_1) \\
&= (\mu_2 - \mu_3) + C_1 \cdot \rho_{12}^* \cdot \delta_1 - C_1 \cdot \rho_{13}^* \cdot \delta_1
\end{aligned}$$

$$= (\mu_2 - \mu_3) + C_1 \cdot \delta_1 \cdot (\rho_{12}^* - \rho_{13}^*) \quad (11)$$

In other words, so long as the correlation between the test measures used at t and $t+1$ equals the correlation between children's true ability at t and $t+1$, then (10) and (11) are equal and we correctly identify change in children's ability over this period.

RULE 3: When errors between tests are uncorrelated, the regression to the mean effect due to initial test error occurs completely between the first and second test.

This is not to say, however, that any change observed beyond the first follow-up test is necessarily genuine – it could still potentially be due to other statistical problems. Here we briefly overview three reasons why this might occur.

Firstly, we return to the problem of correlated errors between different tests taken at different time points - $(\text{corr } \varepsilon_{it}, \varepsilon_{it+1}) \neq 0$. This could be an issue when the same person is assessing the child across test periods; when using parental reports for example. Now recall that regression to the mean due to initial test error is driven by the fact that one is conditioning children's baseline observation on the random error they happened to have on the first test. Then when these children are followed up they receive a completely different random draw, which is on average zero, leading to the large decline in their test scores. Now, however, say that the error that children receive on the second test (at $t=2$) is correlated with the error they receive on the first test (at $t=1$). The expected value of children's test scores at time 2 will therefore still be partially conditional on the error at time point 1, meaning that this error will continue to have an impact on one's estimates. Thus, if the correlation between test errors decays over time, one will see a continual drop in high ability – low SES children's test scores even when no genuine change is taking place. Under this situation, namely of decaying correlation between test errors over time rather than independence, rule 3 above no longer holds and regression to the mean due to initial test error can continue into future periods.

A second difficulty relates to missing data and the selectivity this introduces into the sample. All resources used to study socio-economic trajectories in children's achievement suffer from this problem, with the well known potential bias induced from non-response. How might this influence the finding that initially able children from disadvantaged homes fall behind their less able but affluent peers? It could be that the parents of academically able disadvantaged children are the most likely to drop out of a study, particularly those who

would have continued to perform well on academic achievement tests. In this situation, the test score trajectory observed for high ability – low SES children would be a downwardly biased estimate of that in the population. This is, of course, just one hypothetical example of how non-response and sample selection may influence the results obtained. It nevertheless remains a possible explanation as to why a change in the socio-economic test score gradient is observed when this may really be an artefact of the underlying data.

Finally, and perhaps most importantly, there may be some artificial factor that is weakening the correlation between test scores over time. In other words, the correlation between test measures is weaker than the correlation between children’s true ability at two given time points (e.g. $\rho_{2,3} > \rho_{2,3}^*$). From (7) and (9), one can see that this will lead to a change in the socio-economic test score gradients, but not for substantive reasons that one is trying to observe. This would be the case, for instance, when one has longitudinal data that in fact measures slightly different skills at different ages, or if the same test captures different dimensions of skill as the child ages. Moreover, the decline observed may well be greater for high ability – low SES children than their high ability – high SES peers. This is due to the fact that socio-economic groups differ in terms of their average level of cognitive skill, and hence will revert towards different means (i.e. high ability low SES children will have further to fall). Further explanation of this problem, including examples, can be found in Jerrim and Vignoles (2011).

3. Simulation model

We now turn to a simulation to illustrate that one can generate similar results to those found in the existing literature simply because of problems with measurement, and hence show that the methodology currently being applied does not allow statistical noise to be separated from genuine, policy relevant change. We then illustrate the performance of the alternative method proposed in section 2.3, discussing its strengths and weaknesses to study the topic at hand. Throughout this section we focus on the problem of conditioning on initial test error that leads to regression to the mean between time point 1 and time point 2. An online working paper (Jerrim and Vignoles 2011) provides further examples regarding the use of non-comparable tests and correlated errors that can cause further regression to the mean in future periods.

3.1. Simulation set-up

To begin, assume there is a population of 200,000 children across which true ability is normally distributed with a mean zero. We call half of the population “high SES” and the other half “low SES”. By the time we come to first test these children there are already differences in “true” cognitive ability (i.e. $\mu^H > \mu^L$). As already noted, much empirical evidence supports this assumption (Goodman et al 2009, Cunha et al 2006). 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 $\delta^L = \delta^H = 1$, $\mu^L = -0.7$ and $\mu^H = 0.7$. We call any child who has true ability in the top quartile, across the whole population of 200,000 children, a true high ability child.

The fraction of high ability children is something we cannot directly observe and must instead rely on children’s test scores, which contain some degree of random error. This is incorporated in the simulations via a second series of random draws, where:

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

In all the following results, we assume tests are reasonably accurate. Specifically, we set the error variance so the correlation between observed and true ability is approximately 0.75. Jerrim and Vignoles (2011) provide additional results where different levels of error variance are used. We then add this random draw onto the child’s true ability to give their observed ability, their test score, in period 1.

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

Any child who has an observed test score in the top quarter of the population is then defined as 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:

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

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

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

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

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

Jerrim and Vignoles (2011) discusses alternative models and results where we allow the error terms to be positively correlated $\text{corr}(\varepsilon_{it}, \varepsilon_{it+1}) \neq 0$.

3.2. Simulation results

We begin by illustrating results from this base model, assuming the reality is that there is no real change in the underlying characteristic we are trying to measure over time (Figure 2 panel A). We do not observe this, however, but instead see what is presented in panel B.

Figure 2

One can see that there is a marked difference between what we observe and the true trajectory. Instead of a flat, constant trend over time, we observe a sharp decline between test 1 and 2, before flattening out between tests 2 and 3. In panel C we illustrate that, if one reduces the accuracy of tests far enough, the high observed ability – low SES line and the low observed ability – high SES line can cross. In this specific example, it is when the correlation between observed test scores at time $t=1$ (Y_{i1}) and true ability at time $t=1$ (A_{i1}) is approximately 0.4. Yet in this simulation model this is not real change but simply the result of statistical error.

In Figure 3 we allow there to be a change in true ability over time. For all groups true ability is to be constant between period 1 and 2, but between periods 2 and 3, true high ability – low SES children suffer a marked decline (see panel A of Figure 3 for “the truth”). Following a similar logic as before, we show in panel B the patterns that one would observe when applying the methodology that prevails in the existing literature.

Figure 3

Two key points emerge. Firstly, we do not accurately capture the genuine change that occurs. In fact, the existing methodological approach identifies something very different to what

happens in reality – it suggests there is a big decline between the first two periods and only a shallow decline thereafter. By implication, if one were to use this methodology to advise policymakers, it is likely that a) the problem at hand would be exaggerated, and b) that it would appear that between periods 1 and 2 there is a decline, when in fact the real fall is between periods 2 and 3.

The second key point comes from comparing panel B in Figures 2 and 3. 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 when using the method that has been applied in the existing literature.

3.3. Methods to account for regression to the mean due to initial test error

We now illustrate the alternative method proposed in section 2.3, namely using one test to identify the high ability children and another test as the initial condition. Specifically, we generate a second baseline test score (Y_{i1}^*), following the same simulation procedure as before. The first baseline test (Y_{i1}) is used to assign children into ability groups, with the second test (Y_{i1}^*) being the initial observation point from which change is measured from. We set reality to be the same as that shown in Figure 3 panel A; true ability remains stable between period one and two, but then declines for the high true ability – low SES group between period two and three. All other aspects of the simulation are unchanged. Results can be found in Figure 4.

Figure 4

The regression to the mean problem has been purged from the estimates, and we correctly identify the flat gradient between period one and two, along with the decline for the high ability-low SES group between period two and three. There does, however, seem to be some evidence of attenuation; the drop in test scores for the high ability – low SES group between period 2 and 3 is less pronounced in these estimates than occurs in reality. Recall from section 2.3 that, although we have overcome the RTM problem, some children are still defined as high ability when they are not actually part of this group. We referred to this earlier as a problem of misclassification. The extent of this problem in the simulation model is shown in Table 1. The left hand column refers to “reality”. Only 6,043 (6%) of low SES children should get defined as high ability, compared to 43,957 (44%) of high SES children.

Yet due to initial test error, we identify more low SES children as high ability (9,986 or 10%) than should be the case, while the opposite is true for high SES children (i.e. 40,014 or 40%).

Table 1

This misclassification of children may continue to influence estimates even after correcting for RTM due to initial test error, if the developmental trajectories of the children we mistakenly classify as high ability are very different to those for the true high ability group. We have deliberately set up the simulation model so that this is the case, in order to highlight this potential limitation of the method we propose to overcome RTM. In reality the impact of this misclassification is unlikely to be as extreme as is shown here. Nevertheless, researchers using the method we propose should bear in mind that estimates will only be completely unbiased under the assumption that misclassified high ability children have, on average, the same developmental trajectory as their true high ability peers. Experimentations with simulated data suggest, however, that moderate violations of this assumption are unlikely to dramatically alter the substantive conclusions that one draws using the method proposed.

Thus far, we have assumed that just two tests have been conducted at the baseline time point (Y_{i1}) and (Y_{i1}^*). If there are more tests measures available (i.e. say there are N such tests), then the researcher can use the average mark obtained on $N - 1$ of these tests to divide children into ability groups, with the remaining unused test score taken as the baseline observation from which to measure change from. This will help to limit the misclassification of children into the wrong ability group, thus limiting the potential problem described in the paragraph above, while also eliminating regression to the mean due to test error. Jerrim and Vignoles (2011) extend the simulation model to illustrate that, in this situation of multiple tests available at time point 1, one is able to overcome the attenuation in Figure 4.

4. Example using the Millennium Cohort Study (MCS) data

We now turn to the Millennium Cohort Study (MCS) to provide an empirical example of the method set out in sections 2.3 and 3.3. This is a nationally representative sample of children born in the UK between 2000 and 2001. Information has thus far been collected at four ages - when children were approximately 1, 3, 5 and 7 years old. Other authors have investigated the progress of initially high ability children from poor homes using these data, applying the methodology that prevails in the existing literature but recognising that regression to the

mean may be a problem in their estimates (Blanden and Machin 2010). We attempt to take their work a step further by considering how results change once we try to take RTM into account. We note that other studies have investigated this issue using the British Cohort 1970 data (BCS 70). However, we have concerns regarding the very limited test information collected in this resource at the earliest time points (22 and 42 months). We feel this limits the usefulness of the BCS 70 data to study the topic at hand, and therefore analyse the MCS instead. Further discussion of the application of the methodology to BCS 70 data can be found in Appendix 2.

As with any longitudinal survey, the MCS does suffer from non-response. Although 19,488 children were included in the initial study, only 14,043 remain by wave 4. The survey organisers have produced a set of high quality response weights to take this attrition into account. We apply these weights throughout the analysis. Some children also have missing data on key variables, leaving us with a working sample of 9,449 individuals. In particular, at age 3 around 1,000 children did not complete at least one of the cognitive assessments. Cross-tabulations suggested that these children tended to come from lower socio-economic backgrounds, with below average performance on later tests. We have investigated the extent to which results change after taking into account such non-response by making an adjustment to the MCS survey weight; all substantive conclusions remain intact.

Based on this sample, we then define:

Low SES = Bottom quartile of equivalised household income

Middle SES = Second or third quartile of household income

High SES = Top quartile of equivalised household income.

We do not dwell on whether income is the appropriate measure of advantage here, but have simply checked that all substantive results hold when using alternative measures of family background, such as highest level of parental education.

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 was only administered to children who speak English; thus the sample includes English speakers only. The latter assessment (Bracken) measures concepts that parents and teachers

traditionally teach children in preparation for formal education. Each child is then categorised by the survey organisers into one of five groups, very delayed, delayed, average, advanced and very advanced, based on their total Bracken score. This measure has been validated against various other indicators of childhood intelligence, such as the WPPSI-R measure of IQ (Laughlin 1995), and is now widely used in the 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, and is hence particularly useful for studying the topic at hand. The correlation between age 3 BAS vocabulary and Bracken test scores is 0.57. Table 2 also provides a cross-tabulation of these two measures for high SES, top income quartile, and low SES, bottom income quartile, groups. Further details are provided in Appendix 3. This clearly illustrates that there is a strong association between socio-economic background and children's test scores, even when measured at this very early age.

Table 2

Having two cognitive measures at age 3 is of obvious appeal given the method we proposed in sections 2.3 and 3.3. Specifically, we take children who have been defined as delayed or very delayed on the Bracken assessment as low ability children and those classified as advanced or very advanced as an indicator of high ability. The other assessment, the vocabulary subset of BAS, will be used as the first observation point. Thus, in the terminology of section 2.3 and 3.3, the Bracken test is used to classify children into ability groups (Y_{i1}) and the age 3 BAS vocabulary test is the baseline observation (Y_{i1}^*) from which change is measured from. Substantive results remain unchanged if the tests are used the other way around – with age 3 BAS used to define ability groups and Bracken as the baseline observation. It is important to recognise, however, that as these two age 3 tests were taken by children on the same day, their errors may be correlated. Recalling the discussion in section 2.3 and Appendix 1, this may mean we are only able to reduce the problem of RTM due to test error rather than completely eradicate it, and thus potentially overstate the decline in the cognitive achievement of high ability low SES children.

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) and is therefore taken as an indicator of children's language ability at age 7.

Substantive findings can be found in Figure 5. Panel A presents results using the existing methodology, with children's age 3 BAS vocabulary test scores used to both divide respondents into ability groups and as the initial observation from which to measure change from. The pattern seen is familiar. Those with scores in the top quartile see a rapid decline between ages 3 and 5, particularly those from low income backgrounds who move, on average, from roughly the 90th to the 50th percentile. Indeed, by age 7 initially high scoring children from poor homes have been overtaken by their less able, but affluent, peers.

Figure 5

In the right hand panel, ability groups are defined using the age 3 Bracken test and the BAS test score is used as the initial observation from which change is measured from, in an attempt to correct for the problem of regression to the mean. There is now no suggestion that the cognitive skills of bright children from poor homes 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. This could, however, be the result of residual regression to the mean effects as discussed in Appendix 1. Nevertheless, the message from the analysis of the MCS should be clear. When one takes into account the problem of regression to the mean a very different conclusion is reached 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. This result should, however, be considered within the limitations of the method that we have applied and the data available to us.

These results are of course specific to the MCS cohort. A quite different pattern might be observed in other datasets, such as the BCS 1970 which has been widely used to study this topic. Children born in the 1970's may have been more dependent on home learning and pre-school support and, consequently, may well have displayed different trajectories to that seen in the MCS. Unfortunately, it is problematic to apply the methodology proposed to the BCS data due to the limited quality and quantity of tests available, particularly before children turn 5. Appendix 2 discusses this issue in detail. In essence this leads us to the conclusion that, with regards to the BCS 70 cohort, regression to the mean is only one possible explanation

for the pattern that has been found by other researchers, and that we are unable to rule out the possibility that high ability – low SES children in that particular generation did suffer a dramatic decline in their cognitive skills. We argue, however, that the methods previously used to analyse the BCS70 data are flawed and that firm conclusions for the BCS 70 cohort cannot be reached. Policymakers should thus be made aware that there is little robust evidence, from either the BCS 70 or MCS cohort, that initially able children from poor homes are overtaken by rich children who were initially lower achievers.

5. 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. This can induce substantial bias into estimates of the educational achievement trajectories of different SES and ability groups, and thus lead to the wrong conclusions being drawn from trends in the data. Simulation evidence clearly shows how, using the existing methodology, one can find a large decline in test performance for bright children from poor homes even when no real change is taking place. Statistical error can therefore potentially explain why we see bright children from poor homes falling behind their affluent high ability peers in a number of studies which apply the existing methodology to UK data. Having described the strengths and weaknesses of an alternative approach, we provide an empirical application to the MCS data. This is therefore the first paper to provide empirical evidence on the trajectories of children from different ability/SES groups based on a method that tries to take the problem of regression to the mean into account. The results, using what we argue is a more robust methodology, provide little evidence that the cognitive skills of initially able low SES children suffer a significant decline between the ages of 3 and 7.

It is therefore important to make clear what these findings imply for public policy regarding social mobility in the UK. We confirm that socio-economic gaps in children's test scores are large and apparent from a very early age. This is consistent with an array of theoretical and empirical evidence which suggests that the earliest years are critical in terms of children's cognitive development (Cunha et al 2006). We do not find evidence, however, of a significant relative decline in cognitive skills suffered by initially able children from disadvantaged homes between age 3 and 7. In other words we find no evidence to support the widely held view amongst academics and policymakers that highly able children from poor homes get overtaken by their affluent but less able peers before the end of primary school.

This implies that whilst family background has a major influence on the child's earliest level of cognitive development, it does not have a strengthening impact that would cause SES gaps in children's cognitive achievement to widen further beyond the age of 3. This finding is, in our opinion, entirely consistent with current government policy of early investment in disadvantaged groups. In fact, by dispelling the myth that bright poor children rapidly lose their talent as they develop, at least for the MCS cohort between ages 3 and 7, we provide further support for interventions that focus on the early years. Government should therefore maintain a focus on reducing the socio-economic achievement gaps that appear long before the start of secondary school.

References

- Blanden, J. and Machin, S. (2007) Recent Changes in Intergenerational Mobility in Britain. Sutton Trust, London. (Available from <http://www.suttontrust.com/public/documents/summaryintergenerationalmobility.pdf>.)
- Blanden, J. and Machin, S. (2010) Intergenerational Inequality in Early Years Assessments. In *Children of the 21st century: The first five years* (eds K. Hansen, H. Joshi, and S. Dex), pp. 153-168. Bristol: The Policy Press.
- Cabinet Office (2011) Opening Doors, Breaking Barriers: A Strategy for Social Mobility, HM Government, London. (Available from http://www.dpm.cabinetoffice.gov.uk/sites/default/files_dpm/resources/opening-doors-breaking-barriers.pdf.)
- Cunha, F. Heckman, J. and Lochner, L. (2006) Interpreting the Evidence on Life Cycle Skill Formation. In *Handbook of the Economics of Education* (eds E. Hanushek and F. Welch), pp. 697-812. Amsterdam: Holland North.
- Davis, C. (1976) The Effect of Regression to the Mean in Epidemiologic and Clinical Studies, *American Journal of Epidemiology*, 104, 493-498.
- Duncan, G. and Magnuson, K. (2011) The Nature and Impact of Early Achievement Skills, Attention Skills and Behavior Problems. In *Whither Opportunity? Rising Inequality, Schools, and Children's Life Chances* (eds G. Duncan and R. Murnane), pp. 47-69. New York: Russell Sage Foundation.
- Ederer, F. (1972) Serum Cholesterol: Effects of Diet and Regression Toward the Mean, *Journal of Chronic Disorders*, 25, 277-289.
- Feinstein, L. (2003) Inequality in the Early Cognitive Development of British Children in the 1970 Cohort, *Economica*, 70, 73-97.
- Feinstein, L. (2004) Mobility in Pupils' Cognitive Attainment During School Life, *Oxford Review of Economic Policy*, 20, 213-229.

Field, F. (2010) The Foundation Years: Preventing Poor Children Becoming Poor Adults, Independent Review on Poverty and Life Chances, Cabinet Office, London. (Available from <http://webarchive.nationalarchives.gov.uk/20110120090128/http://povertyreview.independent.gov.uk/media/20254/poverty-report.pdf>.)

Galton, F. (1886) Regression Towards Mediocrity in Hereditary Stature, *The Journal of the Anthropological Institute of Great Britain and Ireland*, 15, 246–263.

Goodman A. Sibieta L. and Washbrook E. (2009) Inequalities in Educational Outcomes Among Children Aged 3 to 16, Final report for the National Equality Panel, Institute for Fiscal Studies, London. (Available from <http://sta.geo.useconnect.co.uk/pdf/Inequalities%20in%20education%20outcomes%20among%20children.pdf>.)

Hansen, K. and Joshi, H. (2010) Millennium Cohort Study Fourth Survey: A User's Guide to Initial Findings, Centre for Longitudinal Studies, Institute of Education, University of London, London. (Available from http://eprints.ioe.ac.uk/5931/1/MCS_3_Descriptive_Report_Oct_2008.pdf.)

Jerrim, J. and Vignoles, A. (2011) 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, Department of Quantitative Social Science working paper 11/01, Institute of Education, London. (Available from <http://ideas.repec.org/p/qss/dqsswp/1101.html>.)

Laughlin, T. (1995) The School Readiness Composite of the Bracken Basic Concept Scale as on Intellectual Screening Instrument, *Journal of Psychoeducational Assessment*, 13, 294-302.

Marsh, H. and Hau, K. (2002) Multilevel Modelling of Longitudinal Growth and Change: Substantive Effects or Regression Toward the Mean Artefacts?, *Multivariate Behavioral Research*, 37, 245-282.

Marmot M. (2010) Fair Society, Healthy Lives: The Marmott Review, Marmot Review, University College London, London. (Available from <http://www.instituteofhealthequity.org/>.)

Parsons, S. Schoon, I. Rush, R. and Law, J. (2011) Long-term Outcomes for Children with Early Language Problems: Beating the Odds, *Children and Society*, 25, 202 – 214.

Reardon, S. (2011) The Widening Academic Achievement Gap Between the Rich and the Poor: New Evidence and Possible Explanations. In *Whither Opportunity? Rising Inequality, Schools, and Children's Life Chances*, (eds G. Duncan and R. Murnane), pp.91-116. New York: Russell Sage Foundation.

Schoon, I. (2006) *Risk and Resilience: Adaptations in Changing Times*, Cambridge: Cambridge University Press.

Table 1. Descriptive statistics drawn from simulated data

	Reality	Observed
Number of children <u>observed</u> as high ability		
High SES	43,957	40,014
Low SES	6,043	9,986
Average error on first test (ϵ_1) for those defined as high ability		
High SES	0	0.6
Low SES	0	1.1

Notes:

Table refers to simulated data. It illustrates: (a) the number of children defined as high ability (b) 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 simulated high and low SES groups. The first column on the left (labelled ‘reality’) refers to when children’s true ability is perfectly observed. The column to the right of this illustrates when a test (which is subject to error) is used.

Table 2. Cross-tabulation of high and low SES children's age 3 Bracken classification against their age 3 BAS vocabulary test quartile (column percentages)

(a) Low SES

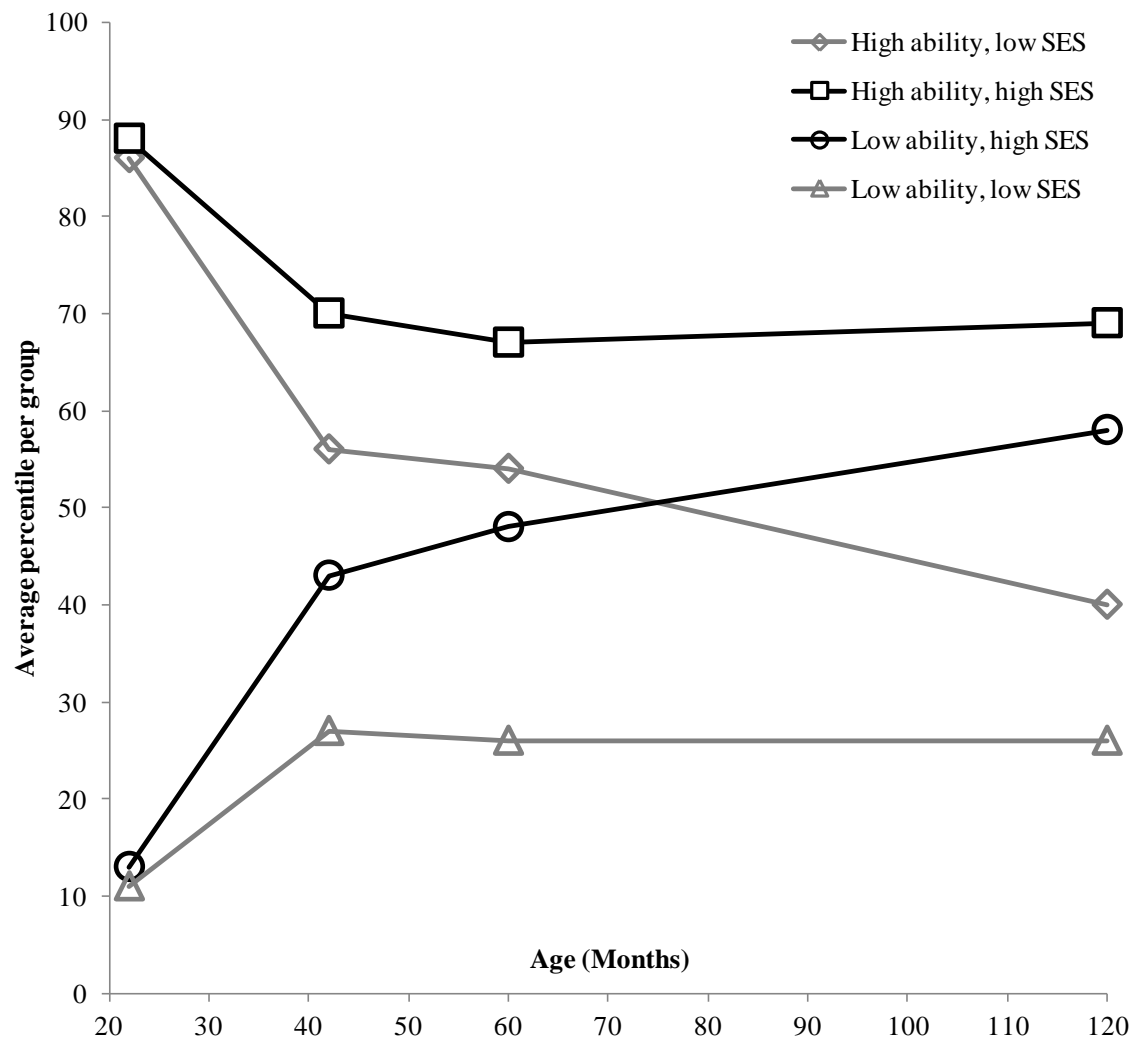
		Age 3 BAS Classification			
		Bottom Quartile	Second Quartile	Third Quartile	Top Quartile
Age 3 Bracken Classification	Very delayed	7	1	1	0
	Delayed	38	17	8	2
	Average	52	72	71	59
	Advanced	2	9	18	31
	Very advanced	1	1	2	7
	Total	100	100	100	100

(b) High SES

		Age 3 BAS Classification			
		Bottom Quartile	Second Quartile	Third Quartile	Top Quartile
Age 3 Bracken Classification	Very delayed	2	1	0	0
	Delayed	12	4	2	0
	Average	73	66	57	37
	Advanced	12	24	32	45
	Very advanced	1	6	9	18
	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. Figures refer to column percentages.

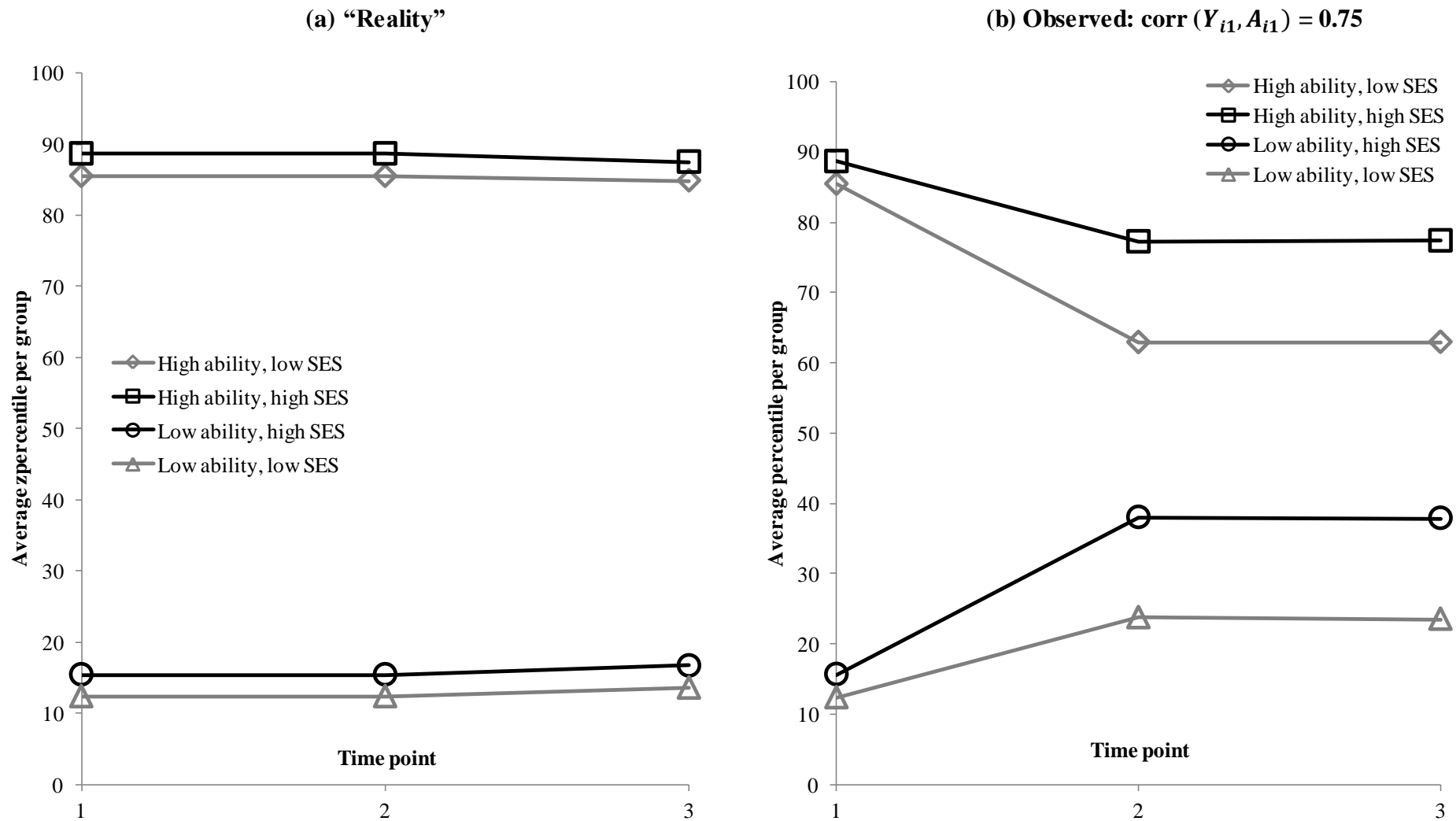
Figure 1. The development of high and low ability children by socio-economic group – evidence from the existing literature



Notes:

Figure adapted from Feinstein (2003) Figure 2. Based upon British Cohort Study 1970 data.

Figure 2. Simulation results using existing methodology, when children's true ability does not change over time



(c) Observed: $\text{corr}(Y_{i1}, A_{i1}) = 0.4$

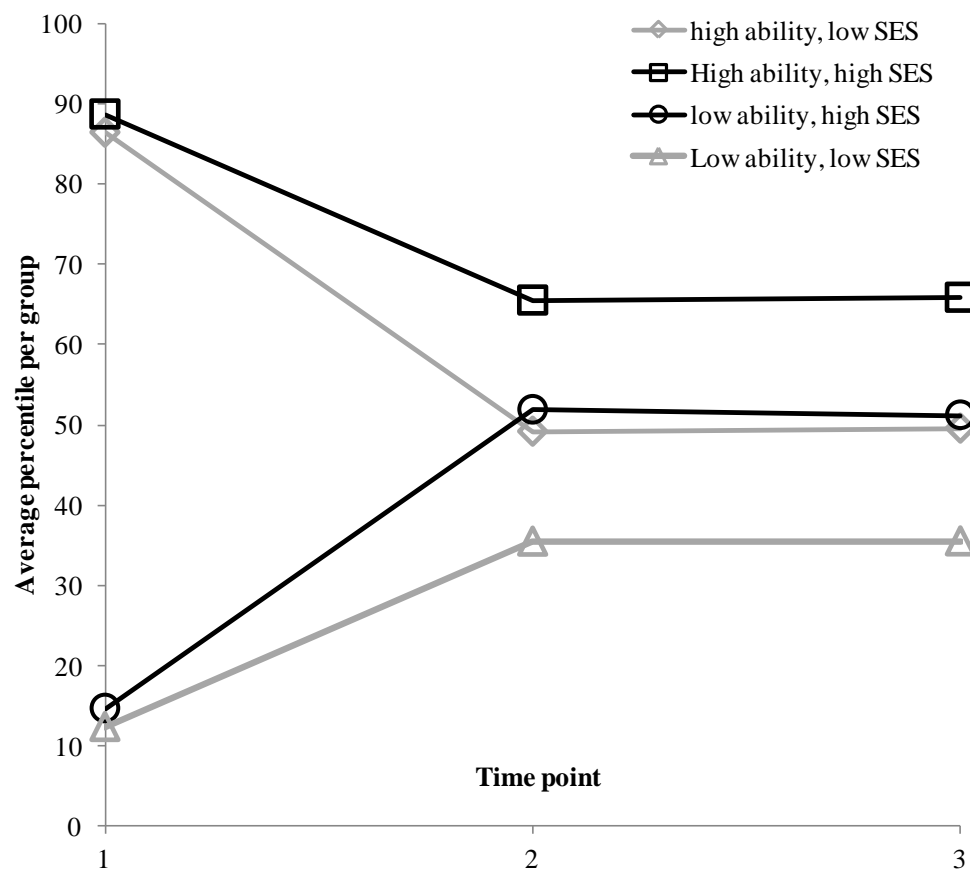
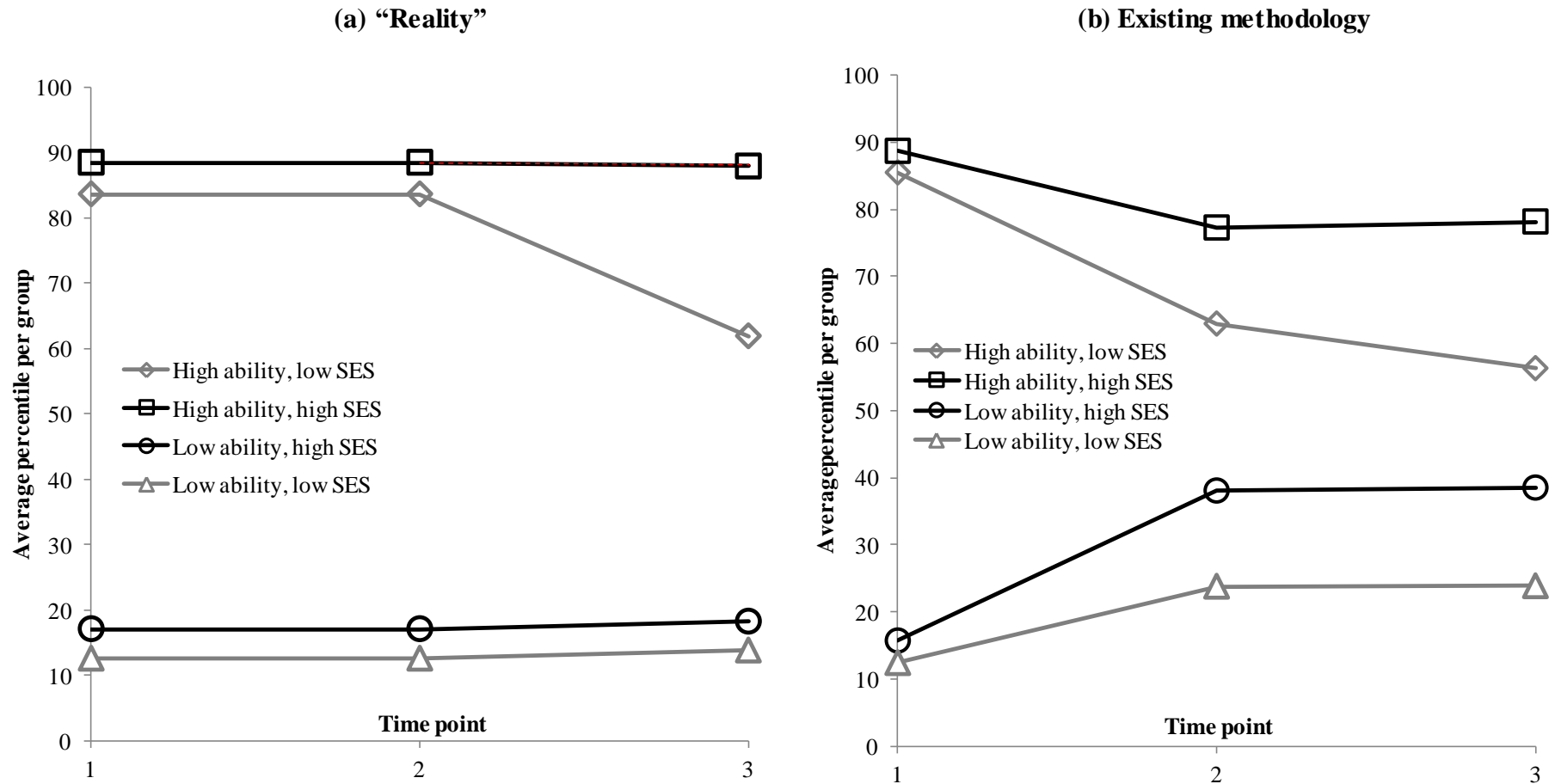
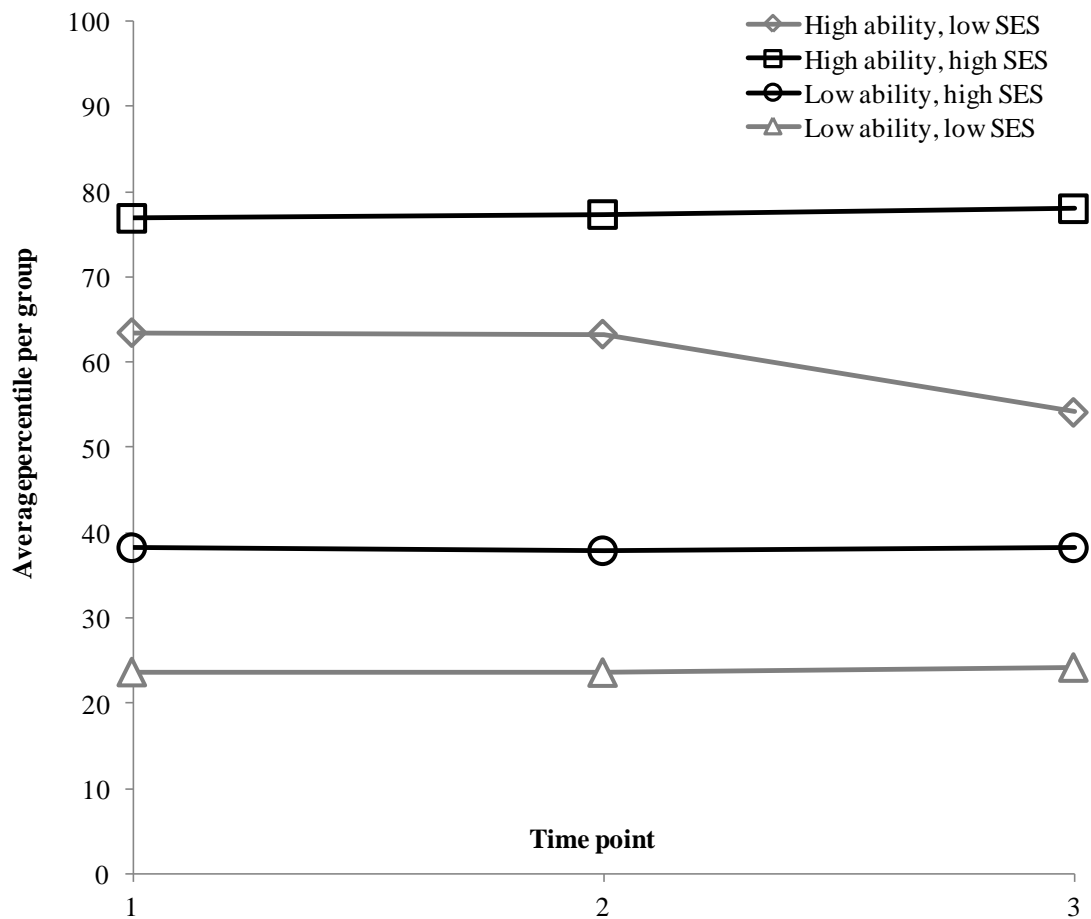


Figure 3. Simulation results using the existing methodology, when there is a sharp fall in true ability for high ability – low SES children between time points 2 and 3



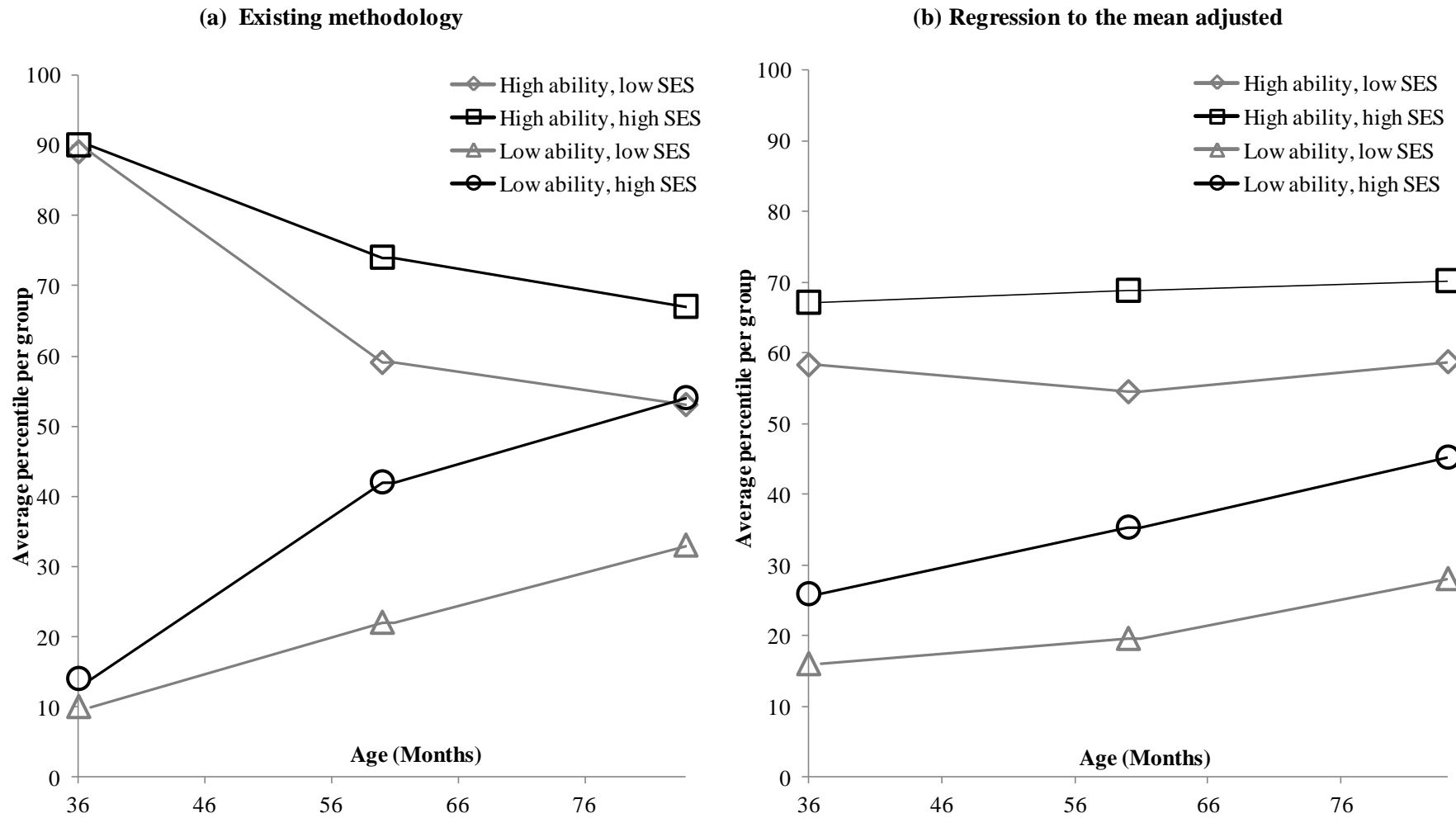
Notes: Diagram produced from simulated data, described in detail in section 3. Children's (hypothetical) age runs along the x-axis, while the average percentile rank for each group is on the y-axis. Panel A on the left refers to when children's true ability is perfectly observed (i.e. it is the actual cognitive trajectory that researchers wish to estimate). Panel B refers to what researchers observe when applying existing methodology.

Figure 4. Simulation results using the method proposed in section 2.3



Notes: Diagram produced from simulated data, described in detail in section 3. Children's (hypothetical) age runs along the x-axis, while the average percentile rank for each group is on the y-axis. The "true" cognitive trajectories are the same as those presented in Figure 3 panel A. These are the results that one observes when applying the method described in section 2.3.

Figure 5. Estimated cognitive gradients in MCS when using different methodologies



Note: Estimated cognitive trajectories based upon the MCS. The left hand panel refers to estimates using existing methodology. The right hand panel is the equivalent figures when applying the methodology proposed in section 2.3

