Dearden, Lorraine; Micklewright, John; Vignoles, Anna

Working Paper

The effectiveness of English secondary schools for pupils of different ability levels

IZA Discussion Papers, No. 5839

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: Dearden, Lorraine; Micklewright, John; Vignoles, Anna (2011) : The effectiveness of English secondary schools for pupils of different ability levels, IZA Discussion Papers, No. 5839, Institute for the Study of Labor (IZA), Bonn, http://nbn-resolving.de/urn:nbn:de:101:1-201107283627

This Version is available at:
http://hdl.handle.net/10419/52120

DISCUSSION PAPER SERIES

IZA DP No. 5839

The Effectiveness of English Secondary Schools for Pupils of Different Ability Levels

Lorraine Dearden
John Micklewright
Anna Vignoles

July 2011

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

The Effectiveness of English Secondary Schools for Pupils of Different Ability Levels

Lorraine Dearden
IoE, University of London,
IFS and IZA

John Micklewright
IoE, University of London
and IZA

Anna Vignoles
IoE, University of London
and IZA

Discussion Paper No. 5839
July 2011

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
E-mail: iza@iza.org

IZA Discussion Paper No. 5839
July 2011

ABSTRACT

The Effectiveness of English Secondary Schools for Pupils of Different Ability Levels*

'League table' information on school effectiveness in England generally relies on either a comparison of the average outcomes of pupils by school, e.g. mean exam scores, or on estimates of the average value added by each school. These approaches assume that the information parents and policy-makers need most to judge school effectiveness is the average achievement level or gain in a particular school. Yet schools can be differentially effective for children with differing levels of prior attainment. We present evidence on the extent of differential effectiveness in English secondary schools, and find that even the most conservative estimate suggests that around one quarter of schools in England are differentially effective for students of differing prior ability levels. This affects an even larger proportion of children as larger schools are more likely to be differentially effective.

JEL Classification: I2

Keywords: school effectiveness, school choice, value added, England

Corresponding author:

John Micklewright
Department of Quantitative Social Science
Institute of Education
University of London
20 Bedford Way
London WC1H 0AL
United Kingdom
E-mail: J.Micklewright@ioe.ac.uk

* This research was supported by ESRC grant RES-576-25-0014 to the ADMIN node of the National Centre for Research Methods at the Institute of Education. We thank Claire Crawford, Luke Sibieta and participants at the 2010 ESRC Research Methods Festival for comments. We are very grateful to James Brown for his advice and help on tests of correlations reported in the Appendix.

1. Introduction

The information currently provided in England to measure the effectiveness of a school is based on the mean performance of the school as a whole, i.e. for the average child. We show here that schools are often differentially effective for students of differing levels of prior attainment. Differential effectiveness is not necessarily a positive or negative phenomenon. A high performing school that does even better for its high ability pupils is better than a low performing school that is consistently ineffective for all its pupils. However, if differential effectiveness is a widespread phenomenon it does suggest that from an accountability perspective, policy-makers need to have information that enables them to determine the extent to which a school is being equally effective for all its pupils. The main points we make are relevant to the increasing number of accountability schemes based on test scores that are emerging in education systems in various countries.

The measures of school performance that have been published to date by the UK Department for Education are all based on the mean achievement level or gain in the school. For example, the modeling of 'Contextualised Value-Added' (CVA) for the performance measures that are currently published, leads to a single figure for the estimated CVA of each school, which 'indicates the value the school has added *on average* for its pupils, given what is known about the circumstances of its pupil intake' (DCSF 2010, our emphasis). The academic literature has long recognised however, that schools may be differentially effective (Jesson and Gray, 1991; Teddlie and Reynolds, 2000; Thomas et al. 1997; Wilson and Piebalga 2008). A school that benefits a high attaining pupil may not do so well for a child of modest or low attainment, or vice versa. Equally a school may do best for pupils in the middle of the distribution.

Differential effectiveness might occur due to schools' strategic choices and response to incentives provided by 'league table' measures of school performance. In England there has been an historical emphasis on a particular threshold - achieving 5 A*-C grades at age 16 exams (General Certificates of Secondary Education or GCSEs). If schools are judged on the proportion of pupil achieving this threshold, they may put additional effort into improving the grades of students near to the threshold and less effort into students with little possibility of achieving it or indeed those who will achieve the threshold come what may. Differential effectiveness may also occur as a result of choices on student grouping. In mixed ability classes teachers may focus on the average student and others may get less attention. If there is

streaming by ability together with positive peer effects, greater progress will be made in the higher ability classes.[1] If the positive impact of the peer effect is outweighed by the ability of teachers to better target their teaching in streamed classes (Duflo, Dupas and Kremer, 2008), schools that do group their students by ability may in fact be more effective for students from right across the prior ability distribution. Another potential cause of differential effectiveness may be a teacher allocation mechanism that assigns more able teachers or indeed unqualified teachers to particular types of student, such as the most or least able.

Some of these potential explanations can be tested in our data. For example, we can assess whether schools that have higher or lower mean levels of prior attainment appear to be more or less differentially effective. We can observe whether schools that have more even pupil intakes, i.e. students from across the full prior attainment distribution, are less differentially effective than schools that have a more skewed intake. Unfortunately we do not have data on the ability grouping policy of each school or on teacher allocation mechanisms, and hence we cannot establish the importance of these possible causes.

Section 2 presents the measures of pupil attainment in primary and secondary school that are central to the analysis together with our data, which refer to the population of children in state secondary schools in England. Section 3 describes how we use these attainment measures to measure differential effectiveness. Section 4 uses our measures to show the extent of differential effectiveness. We find that even the most conservative estimate suggests that around one quarter of schools in England are differentially effective for students of differing prior ability levels. This affects an even larger proportion of students since we find that larger schools are more likely to be differentially effective, partly because with larger sample sizes it is easier to detect differential effectiveness. Section 5 concludes.

2. Measures of attainment at primary and secondary level

Our measures of differential effectiveness and the government's estimates of average school performance both make use of two sets of information on pupil achievement. The first is the pupil's 'Key Stage 2' (KS2) attainment, derived from externally marked national tests in mathematics and English taken at age 10/11 in the final year of primary schooling. This

[1] The evidence on peer effects is mixed e.g. Angrist and Lang (2004); Figlio and Page (2002); Lavy, Paserman and Schlosser (2008).

provides a measure of attainment prior to entry to secondary school which is exogenous to the secondary schools' efforts to improve learning.[2]

Pupils obtain a raw fine-grade score in the KS2 tests. The scores are grouped into 5 levels for each subject to aid understanding by parents and teachers. We collapse the 5 x 5 combinations of mathematics and English levels into 8 groups, reflecting the pupil's level of ability in these two subjects: below level 3-3 (which we label 2-2); level 3-3; level 4-3; level 3-4; level 4-4; level 4-5; level 5-4 and level 5-5 in mathematics and English respectively.[3] The 'expected' level of achievement at KS2, i.e. the level that the government indicates represents adequate progress, is level 4. We use data on these KS2 levels rather than the underlying scores because parents are provided with information on their child's level in each subject area but are not told the raw fine-grade scores. For the same reason, we do not use the average point score at KS2 that features in the Department of Education's modeling of CVA.

The distribution of state school secondary school pupils across prior attainment groups are shown in Table 1. Our data come from the integrated National Pupil Database (NPD)/Pupil Level Annual School Census (PLASC) data set for two cohorts of pupils in year 11 (age 15/16) in 2006/7 and 2007/8. We use two cohorts to boost sample sizes for the different achievement groups within each school. We restrict attention to state school pupils for whom we have KS2 test scores (or teacher assessments[4]), which is the great majority.[5] This represents 1,116,982 children who are aged 16 by the end of the academic year in question. The dataset provides us with nationally comparable information on attainment for children at the end of both primary schooling and compulsory secondary schooling.

Table 1 shows group 4-4 to contain 30 percent of the sample with another 40 percent in a higher group. Hence some 30 percent of children did not reach the 'expected' level 4 in at least one subject.

---

[2] There are also Science tests at KS2, which were also externally marked prior to 2010. We make use only of the mathematics and English scores to increase the manageability of the analysis and because science is no longer a mandatory test at KS2.

[3] Below level 3-3 includes not just those that achieve level 1 or 2 but also those working towards 1 but who have not achieved it. This below level 3 group is much more heterogeneous than the other groups. There were a very small minority of students with highly variable scores across subjects. To make group sizes tractable those with 4-2 or 2-4 are allocated to 4-3 or 3-4 and 5-2 or 5-3 are allocated to 5-4 and if 2-5 or 3-5 allocated to 4-5.

[4] Teacher assessments are used where actual test scores are missing. Teacher assessments are given in terms of level and have a highly correlation with actual KS2 results (Gibbons and Chevalier, 2008). In the data we use, the teacher assessments and the level actually achieved by the child have correlations of 0.81 for English and 0.83 for mathematics.

[5] See Table 1 for details of missing KS2 scores. We drop absentees, those disallowed from the KS2 test (generally this happens due to special educational needs limiting the child's ability to access the curriculum) and those missing due to not being in the country the previous year. Since our focus is on accountability systems we restrict our analysis to pupils in public funded state schools. The sample in Table 1 refers to the sample we use after these exclusions.

The second set of attainment information we use is the child's test scores from the 'Key Stage 4' (KS4) GCSE examinations taken at age 15/16 during the last year of compulsory schooling. Again these examinations are externally set and marked, and the results are therefore nationally consistent. We use a univariate summary of achievement in GCSEs that combines information on the number of subjects examined and the grades achieved. We start with the 'capped' average point score that takes into account the pupil's 8 highest grades at GCSE or equivalent.[6] We then add the points the student achieved in English and mathematics GCSE exams to the total capped score[7]. This ensures that these essential academic skills are included in our attainment measure, an important point given that schools have an incentive to maximize the points a student achieves at GCSE and may do this by encouraging students to take less demanding subjects rather than more rigorous subjects such as English or maths. If already present in the 'capped 8' score, mathematics and English enter our measure twice. This augmented capped score is used by the Department of Education in the official CVA model.[8]

3. Measuring school effectiveness

To provide a robust estimate of the extent of differential effectiveness across schools, we calculate various measures of GCSE performance for pupils in each of the 8 KS2 prior attainment combinations. For each school we average across the values of these measures for its pupils in each group to obtain 8 summary statistics of pupil performance. If these group averages vary significantly within a school (with measures that are appropriately scaled), we classify that school as being differentially effective.

Our simplest measures are shown in the first row of Table 2. For Measure 1, we take the individual's KS4 score and subtract the mean score of other individuals in the KS2 ability group (Table 1 shows these means).[9] For example, for a pupil in the modal 4-4 group, we take her KS4 score and subtract the mean KS4 score for all other pupils in the country in the

[6] There are vocational qualifications that are deemed to be equivalent to GCSE and these are included in this attainment measure.
[7] If a student achieves no GCSEs they are awarded a score of zero.
[8] According to the new scoring system introduced between 2002–03 and 2003–04, 58 points were awarded for a grade A* at GCSE, 52 for an A, 46 for a B, 40 for a C, 34 for a D, 28 for a E, 22 for F, and 16 for a G. Marks are allocated for standard GCSEs, but also for all qualifications approved for use pre-16, such as entry-level qualifications, vocational qualifications, and AS levels taken early.
[9] The dataset on which these simple measures are based exceeds one million students and hence standard errors are very small.

4-4 group. We therefore obtain a measure of the number of KS4 points by which someone differs from their group average.

Measure 1 is a value added measure, similar to the first stage of the calculations of the value-added (VA) measure that the DfE[10] published in 2002-5 prior to the development of the current CVA model. The Department defined 10 ordered groups on the basis of average KS2 fine grade score across English, mathematics and science. They then calculated the difference between each individual's KS4 score and the median KS4 score of individuals in the relevant group (DCSF 2010a, Table A). We use the mean group score rather than the median, our groups are somewhat different (not being based on the univariate fine grade score average and not taking account of science), and our KS4 measure is the 'augmented' capped 8 score rather than the simple capped 8 score. What differs sharply however, is how the figures are then used: the Department summarised school performance by taking the average of these individual-level differences across *all* pupils in the school. We calculate 8 separate averages for each school, one for each group defined by KS2 performance (below 3-3; 3-4, etc).

Measure 1 is an 'absolute' measure of KS4 GCSE points. It does not recognise that a difference of a given number of points may not mean the same thing in terms of school performance at different places in the KS4 distribution. It may be harder for schools to push pupils to gain 6 points who enter the school at the top of the KS2 distribution than at the bottom. One reason for this is the use of a capped score, the other is that grades are right censored with significant (and increasing) numbers of pupils achieving A*, the top grade, at GCSE. This problem can be seen in Table 1 which shows the distinct fall in the dispersion of KS4 scores as one moves down the table: group 5-5 has a standard deviation of scores that is some 40 percent lower than that in the first three groups.

This concern leads to Measure 2: the GCSE score for each individual is transformed into a standardised score, i.e. a z-score. The metric is now the group standard deviation. We label this a 'relative' measure. As before, we then calculate 8 group averages for Measure 2 for each school.

Our other measures are more sophisticated versions of these two first measures. Measures 3 and 4 take account of the account of the individual's exact level of prior KS2 attainment. For each of the 8 groups, we regress pupils' KS4 scores on their average KS2

[10] Department of Children Schools and Families as it was then.

fine-grade scores (averaging across English and mathematics only). Each regression is estimated for all pupils in the group nationally, using the samples shown in Table 1.[11]

$$KS4_{ig} = a_g + b_g.KS2_{ig} + u_{ig} \qquad g = 1..8 \text{ groups} \qquad (1)$$

For Measure 3 we take the residuals from the regression as our estimates of individual-level value-added – the difference between actual KS4 score and that predicted on the basis of KS2 fine-grade score – and then average these figures for each group in each school. For Measure 4, we standardise the individual level figures by dividing by the group standard deviation. (Measures 1 and 2 are the special case of $b_g = 0$, all g, since in that case the OLS estimate of a_g is the mean group score.)

Factors other than prior attainment also influence pupil performance e.g. socio-economic background. We augment the regression model in equation (1) to include some age 11 characteristics from PLASC in addition to the KS2 test score: gender, month of birth, neighbourhood measure of deprivation based on home address (IDACI), whether eligible for Free School Meals (FSM), English as an Additional Language, whether the individual has special educational needs (distinguishing between statemented and non-statemented), and ethnicity.[12] We again estimate the regressions separately for each of the 8 groups defined by KS2 attainment. The results are used in the manner described above to give us Measure 5 (absolute) and 6 (relative). The main difference from the current DfE CVA measure is that the regressions are estimated separately for each group and that we use the results to estimate the contextualised average value added by each school for each of 8 groups separately rather than for the school as a whole.

We do not envisage that in practice all 6 measures in Table 2 would be published – a choice between the different measures would need to be made and the analysis that follows is intended to inform that choice.

4. What the proposed information reveals

4.1 An example

[11] Full regression results are available from the authors.

[12] If they have no KS2 score we use KS2 teacher assessment instead. If their KS 2 score is missing and they have no KS2 teacher assessment they are excluded from the sample altogether. If their ethnicity is missing, they are dropped. If other characteristics are missing, they are allocated to a missing data variable.

To illustrate, we use an example from a particular school, school X. Table 3 shows the number of pupils in the school in each prior attainment group over the two year cohorts and the percentage this represents of the total (last two columns). Some schools, including supposedly comprehensive schools, only admit pupils from the upper or lower range of the prior attainment distribution. By showing the proportion of pupils that fall into each attainment category we can indicate the selective nature of the school's intake. School X has an intake that is very similar to the national distribution.

The final row in the table gives the p-value from an F-test of the hypothesis that the 8 means for the measure in question do not differ.[13] Information on the statistical significance of any differences is needed with these proposed measures, as much as for the Department's CVA measure.[14] Many schools with different ranks in league tables of CVA are not actually significantly different from one another.

School X appears to be a 'good' school. In general, pupils' mean GCSE scores are above the national mean. In the large modal group, group 4-4, the pupils score 21 points above the national mean on average (Measure 1) or nearly a quarter of a standard deviation (Measure 2), differences that are well determined. There is a strong suggestion that school X is differentially effective. On Measures 1-4, the data comfortably reject at the 5 percent level the hypothesis of equal means across the 8 prior attainment groups. School X seems to be less effective for its highest ability group of 5-5 pupils. On average these pupils score below the national group mean by nearly 12 points (Measure 1) or nearly 0.2 of a standard deviation (Measure 2). However, using measures of performance that in addition take account of pupil characteristics other than prior attainment, the differences across the groups are less clear cut and we only just reject the null at the 5 percent level (Measure 5) or at the 10 percent level (Measure 6).

4.2. National analysis of the proposed information

To what extent is the differential effectiveness of school X evident in other schools? Table 4 shows the proportion of schools that are found to be differentially effective according

[13] We assume that each group of pupils represents a simple random sample from a super population of possible pupils. In the case of Measures 3-6, we ignore sampling error in the parameter estimates obtained from the national level regressions that enter the calculations; Sampling error is small given the large sample sizes and the simplicity of the models.

[14] The school performance tables (http://www.education.gov.uk/performancetables/) published by the Department do currently include upper and lower limits of a confidence interval for the whole-school mean CVA, although these are generally disregarded when these data are reproduced elsewhere in the media (see Leckie and Goldstein's contribution to this special issue).

to our various measures and using a 5 percent significance level for the test of difference in group means. Between 25 and 40 percent of schools are differentially effective, depending on the measure used. The figure is highest with the simple absolute measure of value added (Measure 1) and lowest figure for the relative measure which takes account of fine-grade KS2 prior attainment score and pupil characteristics (Measure 6). The other measures are fairly consistent suggesting around one third of schools are differentially effective, though this represents a higher proportion (nearly 50%) of pupils since larger schools are more likely to be differentially effective (see Table 5).

To start to explain *why* schools are differentially effective, we need to know the characteristics of the schools concerned. In the Introduction, we proposed a number of potential explanations, several of which were related to the prior attainment of pupils admitted to the school. Figure 1 explores the relationship between this prior attainment, measured by the average KS2 test scores of a school's pupils, and the likelihood of the school being differentially effective. As in Table 4, differential effectiveness is defined as the mean values of pupil performance at 16 differing significantly across prior attainment groups (at the 5 percent level). For ease of exposition we focus on one measure of performance, namely the conservative Measure 6 which takes account of both fine-grade KS2 score and pupil characteristics using standardized scores. (Similar patterns emerge regardless of the measure used.)

Figure 1 suggests a positive but non linear relationship between average KS2 intake and the likelihood of a school being differentially effective. Very high average KS2 intake – schools in the top decile of the KS2 distribution – does appear to be associated with a greater incidence of differential effectiveness. We then consider the relationship between differential effectiveness and secondary school achievement as measured by mean GCSE scores – see Figure 2. It may be that some schools maximise their mean GCSE score by focusing on certain groups of students (e.g. those near the 5A*-C GCSE threshold) at the expense of others. We do find that the likelihood of a school being differentially effective rises with mean GCSE scores, which when considered alongside Figure 1, suggests a positive relationship between the value added by a secondary school and the likelihood of being differentially effective. This may be down to strategies pursued by schools in their attempt to maximise GCSE scores that have the effect of adding more value for some students than others.

We look at this directly in Figure 3 by examining the relationship between the average value added of the school, as measured by its CVA score, and the likelihood of the school being differentially effective. We might expect that good schools that on average add more value are more consistent (less differential effectiveness) as such schools serve their pupils equally. However, Figures 1 and 2 suggest that an alternative hypothesis might be that schools that are more effective in CVA terms may be less consistent (more differential effectiveness) because they are strategic in the choices they make to maximise average CVA. Figure 3 shows that schools with high CVA are more likely to be differentially effective. This does not prove that differential effectiveness is caused by schools pursuing strategies to maximised CVA but it is suggestive.

We investigate these patterns in a multivariate (probit) model, where the dependent variable takes a value of 1 if the school is differentially effective and zero otherwise where differential effectiveness is defined as in Figures 1-3. Table 5 shows marginal effects. In column 1 we control for the average KS2 intake of the school, to determine whether schools with more able intakes are more likely to be differentially effective, and the standard deviation of the school's KS2 intake scores, to test the hypothesis that schools with a wider range of student abilities may struggle to add equal value and hence that the likelihood of being differentially effective may be greater where intake is more variable. Moving across to column 2 we include controls for school size and the number of KS2 prior attainment groups that are represented in the school. This latter measure allows for the fact that a school may be less likely to be differentially effective if there are only a few different prior attainment groups in the school: this too measures variability in the KS2 intake. In column 3 we include variables describing the school, including the gender balance and ethnic profile. In column 4 we include the mean GCSE score of the school. This enables us to test whether for a given level of KS2 score, higher GCSE achieving schools are more differentially effective.

Schools with very high mean KS2 scores are significantly more likely to be differentially effective. For example, schools with KS2 intakes in the top decile of the distribution are 15 percentage points more likely to be differentially effective than those in the bottom decile (significant at the 10% level). This confirms that the results in Figure 1 hold after controlling for the variability of KS2 intake, which is insignificant in the model. When we control for school size and prior attainment groups in column 2 and other covariates in column 3, we no longer find any significant relationship between KS2 intake and the likelihood of a school being differentially effective. In column 4 we control for GCSE

score. For a given level of GCSE performance, schools with higher KS2 intakes are less likely to be differentially effective. Put another way, holding KS2 constant, those schools with higher GCSE achievement level are more likely to be differentially effective (significant at the 1% level). This indicates that schools that add more value between KS2 and GCSE are more likely to be differentially effective as hypothesized.

Larger schools are more likely to be differentially effective, partially reflecting the fact that larger sample sizes will enable us to reject the null hypothesis of no statistical difference between groups more readily. Other observable school characteristics, such as the percentage of children from any particular ethnic group, are not always correlated with the likelihood of the school being differentially effective. The variable measuring the number of prior attainment groups represented in the school remains statistically insignificant in all the models.

We confirm from this analysis that schools that add more value, i.e. have higher GCSE scores conditional on KS2 scores, are more likely to be differentially effective. We cannot determine whether schools achieve higher value added by exhibiting strategic behavior which leads to the school being differentially effective: the direction of causality is not clear.

4.3. Robustness Checks

We undertook several sensitivity analyses to check the robustness of our findings. Cell size is an issue for some schools, with very small numbers of pupils in some prior attainment groups. We re-estimated our analyses with a minimum requirement of 10 pupils in each cell. The results did not vary substantially to those presented here. Table 1 highlighted lower variation in GCSE outcomes in the highest prior attainment group (group 5-5). As discussed earlier, we anticipate some right censoring here which could affect schools that predominantly take their students from this group. We recalculated our estimates excluding 164 selective (grammar) schools.[15] As Table 6 shows, this did not change our key findings substantially. When we exclude selective schools, 30-40 percent of schools appear to be differentially effective, depending on the measure used.

We would expect our results to be sensitive to the level of statistical significance we use when testing differences in means across prior attainment groups. Continuing to exclude

[15] The great majority of state secondary schools in England do not explicitly select their intake using a criterion of ability.

selective schools, we checked whether using a 1 percent rather than a 5 percent level of significance makes a difference to our substantive findings. Table 6 shows the proportion of schools that are differentially effective is reduced when we use the more stringent criterion. Nevertheless, between 1 in 6 and 1 in 4 schools, depending on the measure used, are still estimated to be differentially effective.

A final validation is to assess the extent to which our various alternative measures of school effectiveness provide different estimates of a particular school's effectiveness. We ranked schools according to their effectiveness on the measures described in Table 2. We then calculated the correlation in the rank of schools across these different measures. For reasons of space we only present the results which use absolute measures (Measures 1, 3, 5), although the results are similar when we include the relative measures. Table 7 shows there is a high correlation between a school's rank across different prior attainment groups when using the M1 measure (the deviation of each pupil from the mean for their prior attainment group), the M3 measure (which controls for fine-grade KS2 point score) and indeed the M5 measure (which controls for both fine-grade KS2 point score and pupil characteristics). The rank correlation of schools across the various measures is always in excess of 0.85. Hence whilst controlling for pupil characteristics and their fine-grade KS2 score does make a difference to the ranking achieved by a school, the correlation between all measures of effectiveness is very high.

Lastly we note that our analysis abstracts from some important issues. First, children sort into different types of schools and hence, as is the case with all models of this type, including the Department for Education's CVA model, we ignore the fact that some of the apparent effectiveness of a particular school may be actually attributable to unobserved characteristics of the pupils who choose to enrol in that school. Second, we ignore any mobility of students and attribute the student's gain in achievement to the school that they are in when they take their GCSEs. Again we know from other evidence that this is an important issue (e.g. Goldstein et al. 2007) but we do not attempt to deal with it here.

5. Discussion and Conclusions

A crucially important question is whether the differential school effectiveness that we observe is of policy importance. One, albeit crude, way of assessing this is to determine the extent to which schools would be ranked differently in school league tables if rankings took account of differential effectiveness. To consider this we ranked schools for each prior

attainment group, according to our simplest school effectiveness measure (M1). We then correlated these rankings across prior attainment groups (results are similar when we use other measures from Table 2). Table 8 shows that the correlation between a school's league table rank across the different prior attainment groups is not high. A school that is ranked high for low attaining children is not necessarily ranked as high for higher attaining children. For example, the correlation between the rank for the lowest group (below level 3) and the highest group (5-5) is just 0.37. The bottom row of Table 8 also shows the correlation between the average rank of the school (across all groups) and its ranking for each prior attainment group. Schools ranked high on average also tend to be higher ranked for each prior attainment group. Unsurprisingly, there is a strong correlation between the average measure and the rank for the modal prior attainment group (4-4) but for lower or higher groups the correlation is lower (e.g. 0.7 for group 2-2).[16]

Hence we find evidence of differential effectiveness that would matter in practice, in terms of parents choosing schools or policy-makers assessing the effectiveness of schools for different types of pupils. We have also found that schools that add more value between KS2 and GCSE are more likely to be differentially effective. We would like to know about the nature of this differential effectiveness, i.e. do differentially effective schools systematically add more value at the bottom or top of the prior attainment distribution. Figure 4 shows the mean value added for each prior attainment group for schools we have identified as differentially effective and for schools that are not differentially effective using the conservative Measure 6 (and a 5 percent significance level). Some schools are differentially effective because they add more value at the bottom of the prior attainment distribution whilst others are differentially effective because they add more value at the top. On average, across all differentially effective schools, we find they are generally more effective than non differentially effective schools, in that they add at or above the average for all prior attainment groups. Further, on average, differentially effective schools tend to add marginally more value for lower attainment groups. Differential effectiveness is not necessarily a negative indicator, as noted in the Introduction. Indeed, if the majority of schools are not that effective for low ability students while the others are effective for all students regardless of ability, these latter schools could be the ones that appear differentially effective: mean group

[16] In the Appendix, we consider how this conclusion is affected by sampling variation, recognizing that schools are being ranked on mean values that are calculated for what are quite often relatively modest cell sizes. We conclude that our results in Table 8 are not just driven by 'noise' resulting from small sample sizes.

performance is not an objective benchmark of the progress children should make – it is just the average for the group.

If schools' performance were presented by prior attainment group, theywould have an incentive to focus on the performance of all pupils. The government has already recognised the impact of the school accountability system in creating incentives that drive schools' behaviour. For example, it has abolished the so called 'equivalencies rule', whereby non-GCSE qualifications were equated with GCSEs even though many were arguably not as challenging, because it encouraged schools to get pupils to take less challenging qualifications (Wolf, 2011). The government is introducing more 'rigorous' measures of achievement, such as the percentage of pupils achieving a more academic set of GCSEs known as the 'English Baccalaureate'[17]. These changes will affect schools with different levels of pupil prior attainment differently. Schools with advantaged intakes are likely to be doing English Baccalaureate GCSE subjects anyway. Schools in less advantaged areas are likely to have been doing a far higher proportion of 'equivalent' vocational qualifications. Whilst this change is bedding in it will therefore be important to consider how schools are responding and to ensure that the performance of all prior attainment groups is monitored and that groups are not neglected in the rush to focus on those who can achieve the English Baccalaureate. Looking at school performance by prior attainment group is one way of doing this.

[17] A*-C GCSE grades for maths, English, two sciences, history or geography, and a language.

References

Angrist, J., and K. Lang (2004) 'Does School Integration Generate Peer Effects? Evidence from Boston's Metco Program' *American Economic Review*, 94(5):1613-34

DCSF (2010) *Key Stage 2 to Key Stage 4 (KS2-KS4) Contextual Value Added Measure (CVA)* http://www.dcsf.gov.uk/performancetables/schools_08/s3.shtml (accessed 8.1.10)

DCSF (2010a) *Secondary School Achievement and Attainment Tables 2004* http://www.dcsf.gov.uk/performancetables/schools_04.shtml (accessed 30.4.10).

Duflo, E., Dupas, P., and Kremer, M. (2008) "Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya". NBER Working Paper 14475.

Figlio, D. And Page, M. (2002). "School Choice and the Distributional Effects of Ability Tracking: Does Separation Increase Inequality?" *Journal of Urban Economics* 51: 497-514.

Gibbons, S. and Chevalier, A. (2008) 'Assessment and age 16+ education participation", *Research Papers in Education,* 23 (2): 113-123.

Goldstein, H., Burgess, S., and McConell, B. (2007) 'Modelling the effect of pupil mobility on school differences in educational achievement', *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, 170 (4): 941-954.

Jesson, D and Gray J (1991). 'Slants on Slopes: Using Multi-level Models to Investigate Differential School Effectiveness and its Impact on Pupils' Examination Results.' *School Effectiveness and School Improvement: An International Journal of Research, Policy and Practice*. 2(3):230-247.

Lavy, V., Paserman, D. and Schlosser A. (2008) 'Inside the Black Box of Ability Peer Effect: Evidence from Variation of Low Achiever in the Classroom' NBER working paper 14415.

Teddlie, C. and Reynolds, D. (2000) *The International Handbook of School Effectiveness Research*, Reynolds, Falmer Press, London and New York.

Thomas, S, Sammons, P, Mortimore, P and Smees, R, (1997) 'Differential secondary school effectiveness : examining the size, extent and consistency of school and departmental effects on GCSE outcomes for different groups of students over three years', *British Educational Research Journal,* 23 (4): 451-469.

Wilson D and Piebalga A (2008) 'Performance measures, ranking and parental choice: an analysis of the English school league tables' *International Public Management Journal*, 11: 233-66.

Wolf, A. (2011) *Review of Vocational Education: The Wolf Report,* commissioned by the Department for Education.

Table 1: KS2 prior attainment groups and KS4 results for year 11 children in state secondary schools in 2006/7 and 2007/8

KS2 group	Frequency	**%**	% (cumul.)	% missing KS2	KS4 mean	KS4 std. dev.
22	73,922	6.6	6.6	6.1	208.1	106.5
33	102,591	9.2	15.8	3.7	274.4	105.3
34	73,063	6.5	22.3	2.4	313.3	106.2
43	96,762	8.7	31.0	1.3	332.2	96.5
44	339,519	30.4	61.4	1.3	383.4	88.9
45	119,474	10.7	72.1	0.8	433.0	79.2
54	113,325	10.1	82.2	0.6	437.7	75.4
55	198,326	17.8	100.0	0.5	491.4	65.1
Total	1,116,982	100.0				

Note: the sample includes pupils for whom no KS2 score is present if there is a teacher assessment of their KS2 score available. The sample excludes private school children. See main text for further details.

T**able 2: Alternative measures of school effectiveness**

	Absolute (metric: KS4 points)	*Relative* (metric: group KS4 SDs)
Raw VA score	1. $diff_{KS4} = KS4\text{-}KS4_{mean}$	2. $Z_{KS4} = [KS4\text{-}KS4_{mean}]/KS4_{SD}$
Regression adjusted VA score	3. residual of regression of KS4 on KS2	4. residual of regression of Z_{KS4} on Z_{KS2}, where latter defined analogously [equivalent to Measure 3 divided by $KS4_{SD}$]
Context adjusted VA score	5. as for measure 3 but with controls in regression	6. as for measure 4 but with controls in regression

Table 3: Effectiveness by prior attainment group for School X: means (standard errors in brackets)

	Absolute Measures			Relative Measures				
Groups	1. *Raw VA score*	3. *Regression adjusted VA score*	5. *Context Adjusted VA*	2. *Raw VA score*	4. *Regression adjusted VA score*	6. *Context Adjusted VA*	No. Obs.	% total
Group 22	34.0	28.5	35.2	0.319	0.267	0.330	46	6.9
	[12.3]	[12.4]	[12.0]	[0.115]	[0.116]	[0.112]		
Group 33	19.5	20.1	14.8	0.185	0.191	0.140	62	9.3
	[12.7]	[12.6]	[11.9]	[0.121]	[0.120]	[0.113]		
Group 34	28.4	27.6	20.9	0.268	0.260	0.196	34	5.1
	[14.6]	[14.4]	[13.3]	[0.137]	[0.135]	[0.125]		
Group 43	33.0	31.4	25.3	0.342	0.325	0.263	48	7.2
	[13.0]	[12.0]	[11.7]	[0.134]	[0.124]	[0.121]		
Group 44	21.4	21.1	17.5	0.241	0.238	0.197	225	33.8
	[4.6]	[4.4]	[4.3]	[0.052]	[0.050]	[0.048]		
Group 45	14.0	14.2	10.6	0.177	0.179	0.133	75	11.3
	[9.1]	[8.5]	[8.2]	[0.114]	[0.108]	[0.104]		
Group 54	15.4	13.9	12.2	0.204	0.185	0.161	78	11.7
	[7.1]	[6.5]	[6.8]	[0.094]	[0.086]	[0.090]		
Group 55	-11.6	-7.8	-5.8	-0.178	-0.120	-0.090	98	14.7
	[5.8]	[5.2]	[4.9]	[0.089]	[0.080]	[0.075]		
p-value	0.005	0.020	0.039	0.003	0.018	0.061		

Note: Sample size 666. School X is an actual school from the NPD/PLASC database. The p-value in final row is from an F-test of the hypothesis that the 8 group means are the same.

Table 4: Percentage of all schools that are differentially effective by type of measure

	Absolute measures %	Relative measures %
Raw VA score	40.0	35.2
Regression adjusted VA score	37.9	31.7
Context adjusted VA score	31.7	25.0

Note: the number of schools is 3,096. The table shows the percentage of all schools that are differentially effective on each measure as indicated by a F-test of the differences in means across the 8 ability groups within a school (5 percent level of significance).

Table 5: Probit model of probability of a school being differentially effective, Measure 6 (5% significance level)

	1	2	3	4
KS2 score 2nd decile	0.021	-0.007	0.000	-0.050
	(0.037)	(0.038)	(0.039)	(0.036)
KS2 score 3rd decile	0.047	-0.005	-0.001	-0.073**
	(0.039)	(0.040)	(0.041)	(0.036)
KS2 score 4th decile	0.043	-0.020	-0.010	-0.102*
	(0.040)	(0.041)	(0.042)	(0.035)
KS2 score 5th decile	0.071+	-0.001	0.006	-0.107*
	(0.043)	(0.044)	(0.045)	(0.037)
KS2 score 6th decile	0.035	-0.042	-0.037	-0.154*
	(0.042)	(0.042)	(0.043)	(0.033)
KS2 score 7th decile	0.067	-0.020	-0.014	-0.153*
	(0.044)	(0.046)	(0.047)	(0.035)
KS2 score 8th decile	0.043	-0.050	-0.042	-0.185*
	(0.046)	(0.046)	(0.047)	(0.032)
KS2 score 9th decile	0.042	-0.051	-0.042	-0.196*
	(0.052)	(0.050)	(0.051)	(0.032)
KS2 score 10th decile (highest)	0.147+	0.023	0.034	-0.193*
	(0.078)	(0.072)	(0.074)	(0.042)
Standard deviation of KS2 score	0.000	-0.020	-0.018	-0.003
	(0.016)	(0.021)	(0.021)	(0.021)
Number of pupils		0.001*	0.001*	0.001**
		(0.000)	(0.000)	(0.000)
Number of pupils squared		0.000+	0.000	0.000
		(0.000)	(0.000)	(0.000)
Number of prior attainment groups		-0.011	-0.011	0.004
		(0.014)	(0.015)	(0.015)
Male			0.057	0.094**
			(0.042)	(0.042)
White Other			-0.312	-0.340
			(0.212)	-(0.218)
Black African			-0.198	-0.254
			(0.295)	-(0.292)
Black Caribbean			0.269	0.157
			(0.307)	(0.315)
Black Other			-0.419	-0.194
			(0.814)	(0.815)
Indian			-0.082	-0.205
			(0.146)	(0.147)
Pakistani			0.176	0.077
			(0.101)+	(0.102)
Bangladeshi			-0.469	-0.567**
			(0.255)+	(0.257)

Chinese			1.507	0.763
			(1.412)	(1.420)
Asian Other			-1.096	-1.302**
			(0.650)+	(0.643)
Mixed			0.040	0.210
			(0.389)	(0.393)
Other			-0.377	-0.457
			(0.460)	(0.433)
Unknown			-0.024	-0.229
			(1.014)	(1.014)
GCSE points score (mean)				0.002*
				(0.000)
Observations	3096	3096	3096	3096

Note: Standard errors in parentheses. + significant at 10% level; ** significant at 5% level; * significant at 1% level. The dummies for KS2 scores indicate deciles of the distribution of schools by their pupils' average KS2 marks (averaged of maths and English). The base case is a school in the bottom decile of the distribution. For ethnicity, the base case is White British. The marginal effects are calculated at the means of all variables. A school is differentially effective if its group means for Measure 6 are significantly different at the 5% level.

Table 6: Percentage of schools that are differentially effective by type of measure (absolute measures only) at different levels of statistical significance (selective schools excluded)

	% differentially effective	
	5% sig. level	1% sig. level
Raw VA score (M1)	37.0	23.4
Regression adjusted VA score (M3)	35.2	21.6
Context Adjusted VA score (M5)	29.8	17.0

Note. The figures relate to 2,932 schools that are not (explicitly) selective in their intakes.

Table 7: Rank correlations of schools by group across different measures of school effectiveness (absolute Measures 1, 3 and 5)

	M1 and M3	M1 and M5	M3 and M5
Group 22	0.99	0.92	0.93
Group 33	0.99	0.91	0.92
Group 34	0.99	0.88	0.89
Group 43	0.99	0.91	0.92
Group 44	0.99	0.87	0.89
Group 45	0.98	0.86	0.89
Group 54	0.98	0.89	0.91
Group 55	0.97	0.86	0.90

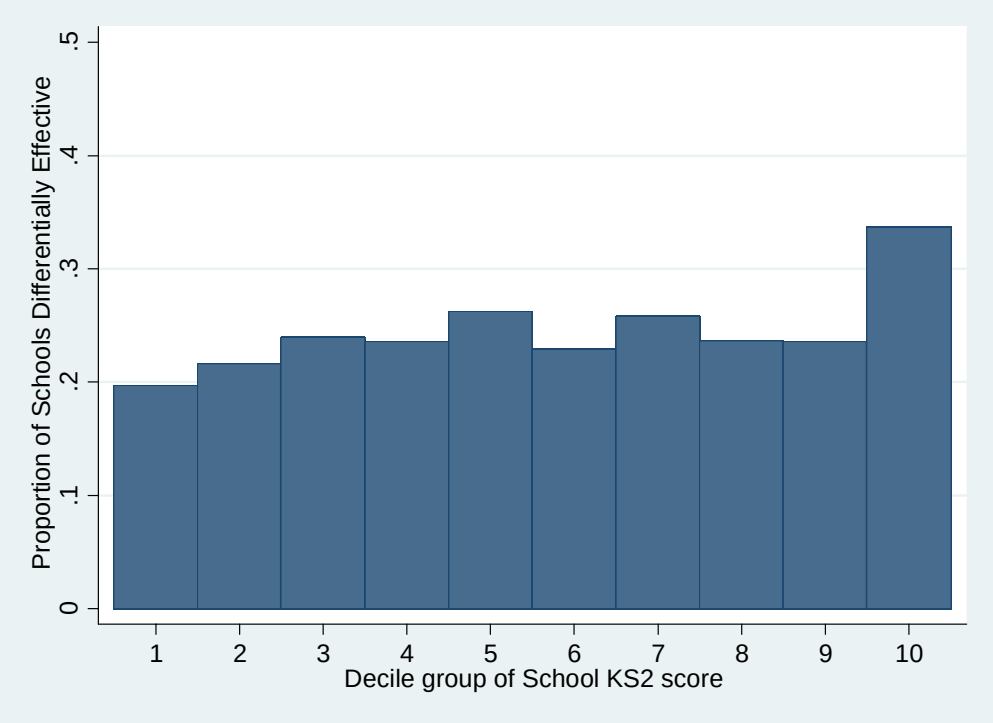
Note: The figures relate to 2,932 schools that are not (explicitly) selective in their intakes.

Table 8: Rank correlations of schools by group based on Measure 1 absolute raw VA score (selective schools excluded)

	Group 22	Group 33	Group 44	Group 55
Group 22	1.00			
Group 33	0.68	1.00		
Group 44	0.58	0.71	1.00	
Group 55	0.37	0.49	0.71	1.00
Average	0.70	0.82	0.94	0.74

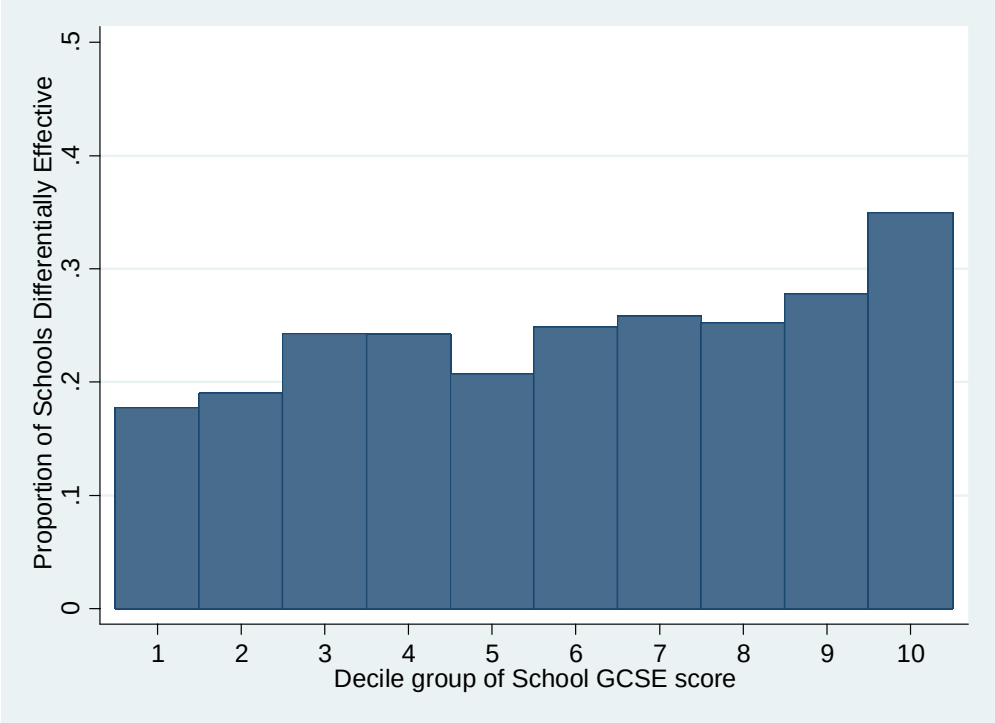
Note: The figures relate to 2,932 schools that are not (explicitly) selective in their intakes.

Figure 1: The proportion of schools that are differentially effective by decile of school mean KS2 score, Measure 6 (5% significance level)



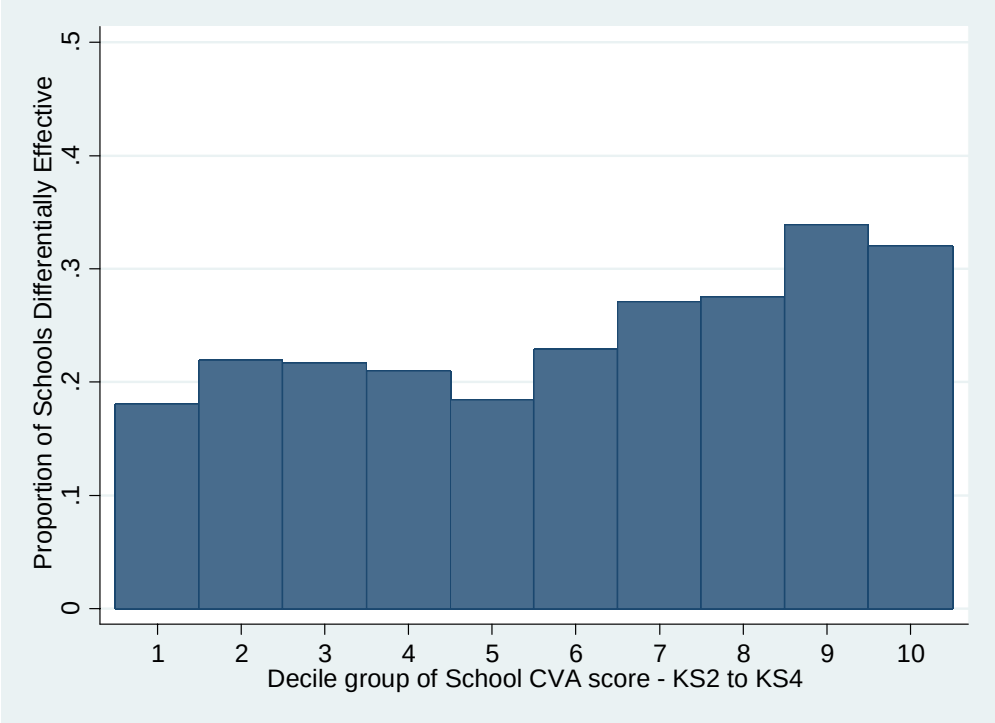
Note: sample of 3,096 schools including selective schools.

Figure 2: The proportion of schools that are differentially effective by school mean KS4 GCSE score, Measure 6 (5% significance level)



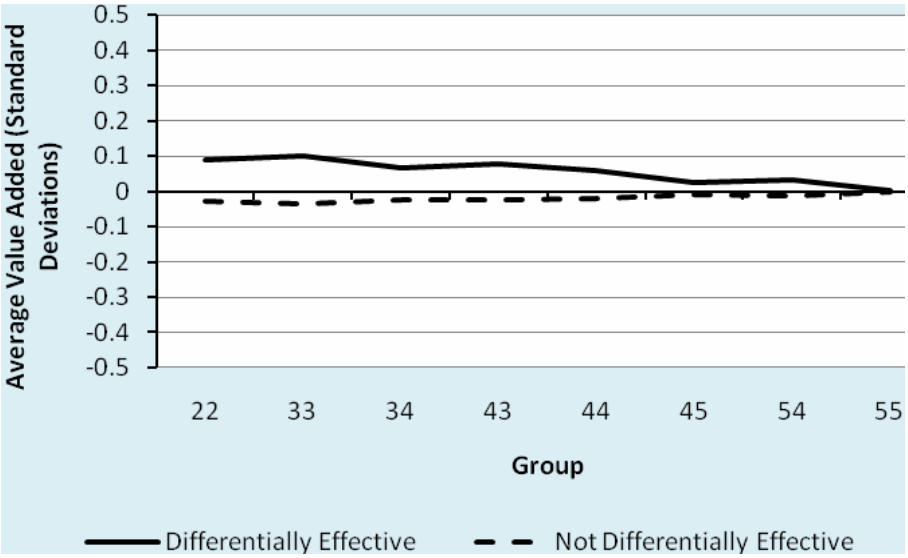
Note: sample of 3,096 schools including selective schools.

Figure 3: The proportion of schools that are differentially effective by school mean CVA score, Measure 6 (5% significance level)



Note: sample of 3,096 schools including selective schools.

Figure 4: Mean value added (in standard deviations) by group for differentially effective schools compared to non differentially effective schools, Measure 6 (5% significance level)



Note: sample of 3,096 schools including selective schools. A school is differentially effective if its group means for Measure 6 are significantly different at the 5% level. The graph shows the mean across schools in each group (weighted by the number of pupils in the school in the group concerned) of the school average value-added on Measure 6.

Appendix: Correlations in schools' rankings on effectiveness measures for different prior attainment groups: the impact of sampling error[18]

Table 8 shows the strength of the association between a ranking of schools by their average effectiveness for one prior attainment group, e.g. the low ability group 2-2, with that for another group, e.g. the high ability group 5-5. However, the number of observations used to compute the average effectiveness for any ability group within a particular school is often not that large. For example, in the case of the single school we considered in Table 3, there are less than 50 students in each of three of the different ability groups. Hence although the data we use to compute our effectiveness measures refer to the population of all children in secondary schools in two years, it is useful to think of the observations for each ability group for each school in the data as representing a sample: a sample drawn from a 'super-population' of all possible students that could attend each school. That is, sitting behind the means for each group that we observe in the data for each school is a set of means for the super-population in each group in the school. Viewed in this way, the correlations presented in Table 8 are estimates of (super-) population parameters that are subject to sampling error.

Imagine that the 'true' correlation of schools' mean scores for a particular measure (M1 to M6) for, say, groups 2-2 and 5-5 were 1.0. In other words, if our calculations of the means were based on the full super-population of students in each group in each school, then the schools would be ranked in exactly the same way for each of the two ability groups. However, the correlation would not be perfect in the actual data available for analysis because of the 'noise' resulting from sampling variation – in each school we only have a sample from the group super-population. We want to assess how important this sampling variation is likely to be with the size of the samples that we have and hence how likely it is that low values of correlations in our actual data shown in Table 8 are just the result of sampling variation.

How might we estimate the extent of the sampling error? One possibility would be to bootstrap our estimates. We adopt a different approach, simulating the sampling distribution for the correlations under assumptions about the super-population values.

[18] We are very grateful to James Brown, University of Southampton, for suggesting the approach taken in this Appendix. He carried out the simulations and has co-authored our write-up of them here.

Procedure

Our procedure is to simulate the sampling distribution under the assumption that the sample standard deviation of scores within a group for a given school is a good estimate of the super-population standard deviation of scores for that group within the school. For the purpose of this exercise, we focus on the Pearson correlation rather than the Spearman's rank correlation.

Consider the 2,372 schools that have students in both the 2-2 and 5-5 groups.

1. School 1 has, say, 30 students in group 2-2 and 40 students in group 5-5, standard deviations in our data for the measure with which we are concerned (M1 to M6) equal to $SD22_1$ and $SD55_1$.

2. Draw a random sample of size 30 students from a normally distributed infinite population with a mean=X_{A1} and standard deviation=$SD22_1$. Record the sample mean=$x22_1$

3. Draw a random sample of size 40 students from a normally distributed infinite population with a mean=X_{B1} and standard deviation=$SD55_1$. Record the sample mean=$x55_1$

4. Go to School 2. School 2 has, say, 50 students in group 2-2 and 65 students in group 5-5, with standard deviations in our data equal to $SD22_2$ and $SD55_2$.

5. Draw a random sample of size 50 students from a normally distributed infinite population with a mean=X_{A2} and standard deviation=$SD22_2$. Record the sample mean=$x22_2$

6. Draw a random sample of size 65 students from a normally distributed infinite population with a mean=X_{B2} and standard deviation=$SD55_2$. Record the sample mean=$x55_2$

7. Continue process for each of the 2,372 schools.

8. Now compute the correlation coefficient of the two sets of sample means [$x22_1$, $x22_2$, $x22_3$...$x22_{2372}$] and [$x55_1$, $x55_2$, $x55_3$...$x55_{2372}$].

9. Perform steps 1 to 8 1,000 times, giving 1,000 estimates of the correlation coefficient.

10. Inspect the distribution of these 1,000 correlation coefficients. Compare with the one correlation we have calculated with our actual data.

In performing this simulation, there are two parameters that can be varied. The first is the correlation of principal interest, that is the correlation in the super-population between schools' means scores on the measure concerned (M1 to M6) for group 2-2 and their scores for group 5-5. We call these the 'between-school' correlations (Table 8 provides estimates of these from our actual data). The second is the correlation within a school between the mean scores for its group 2-2 students and its group 5-5 students. We call these the 'within-school'

correlations and draw samples to control the correlation, i.e. to achieve an assumed value. (The assumption made here determines the relationship between X_{A1} and X_{B1} and between X_{A2} and X_{B2} at Steps 2 and 3 and Steps 5 and 6 in the simulation.) We can assume that this correlation is zero, which is the simplest case, or we can allow for a positive correlation reflecting the idea that if a school has a 'good draw' of 2-2 students in terms of the sample mean score on the measure concerned, it is more likely also to have a good draw of 5-5 students on account of some influence common to both groups that raises scores e.g. good teachers or a particularly good cohort. Arguably, the closer the prior ability groups, e.g. 3-3 and 3-4, the higher this within-school correlation will be.

Results

Figure A1 investigates the sensitivity of the results of the simulation to the choice of the within-school correlation. We fix the super-population value of the between-school correlation for groups 2-2 and 5-5 at the value actually observed in our sample data, 0.33. (Note we are considering Pearson correlations in this Appendix rather than the Spearman rank correlations considered in Table 8 – the rank correlation for group 2-2 and 5-5 is 0.37.) Each vertical bar shows the range that includes 95% of the 1000 sample estimates of the between-school correlation that were produced in the simulation. We vary the within-school correlation across the range from zero to one. As the within-school correlation rises, the range of the estimates of the between-school correlation moves up since the sample 'draws' for the two groups in each school increase in similarity in the sense described above. The results show that actual observed sample value of the between-group correlation – shown by the horizontal line – can be seen as consistent with a super-population between-group correlation of the same value, 0.33, and a within-school correlation of 0.3 to 0.5 but at the 5% significance level we would reject the hypothesis that within-school correlation was 0.2 or less of 0.6 or more.

We next repeat the exercise considering only schools with at least 30 pupils in both groups, 2-2 and 5-5, so as to strengthen the validity of the use of the central limit theorem in the simulation. This reduces the number of schools included in the simulation to only 554. The observed sample between-school correlation increases to 0.42. Restricted to the larger schools in this way, we find the same pattern as before, albeit around the higher observed

correlation, with more variation in the distributions as each correlation is based on a reduced number of schools.

We now return to the full data set of 2,372 schools and present sampling distributions in which we vary the super-population (i.e. the 'true') between-school correlation from 0.1 up to 1.0. We do this under three different assumptions about the within-school correlation, setting this to zero, 0.4 and 0.7, leading to Figures A2-A4. In each case the horizontal line shows the value of the between-group correlation that is observed in our actual sample data. Note that as the assumed 'true' between-school correlation approaches 1.0, we have to obtain a distribution that lies wholly below the 'true' value as a correlation coefficient cannot exceed this level.

An obvious feature of these graphs – and the key result for our purposes – is that the horizontal line showing the sample value of the between-school correlation intersects with very few of the vertical bars indicating the range for 95% of the values in each simulation. In other words, in most cases our observed sample value is not consistent (at the 5% significance level) with the between-school value assumed in the different simulations. For example, Figure A3 suggests that the observed sample value of 0.33 is consistent with the 'true' value being 0.3 (and the within-school correlation being 0.4) but is not consistent with values of 0.1-0.2 or values of 0.4 or more. Hence we can conclude with a reasonable degree of confidence that the relatively low observed sample value is not simply the result of sampling error pushing the sample value well below a much higher 'true' value in the super-population. In fact we can reject the hypothesis that the true value is 0.5 or more irrespective of whether the within-correlation is assumed to be zero (Figure A2), to be moderate, 0.4 (Figure A3), or whether it is assumed to be quite high, 0.7 (Figure A4).

This is an encouraging result – it suggests that our results in Table 8 are not just driven by 'noise' resulting from relatively modest numbers of students in each ability group in each school. However, we should note that we have considered only the most extreme comparison, the low ability group 2-2 and the high ability group 5-5, and a more complete analysis would also consider the sampling distributions for groups that are closer in average ability. (For example, consider groups 3-3 and 4-4, where Table 8 shows the means to have a between-school correlation of 0.71 in our data. Imagine that the within-school correlation was high, e.g. 0.7, which seems reasonable for groups of similar ability. Were our simulations in this case to result in a diagram like Figure A4, we would not be able to reject the null hypothesis of a 'true' between school correlation of 0.9 or even 1.0.)

Figure A1: Varying the within-school correlation, between-school correlation = 0.33 (groups 2-2 and 5-5)

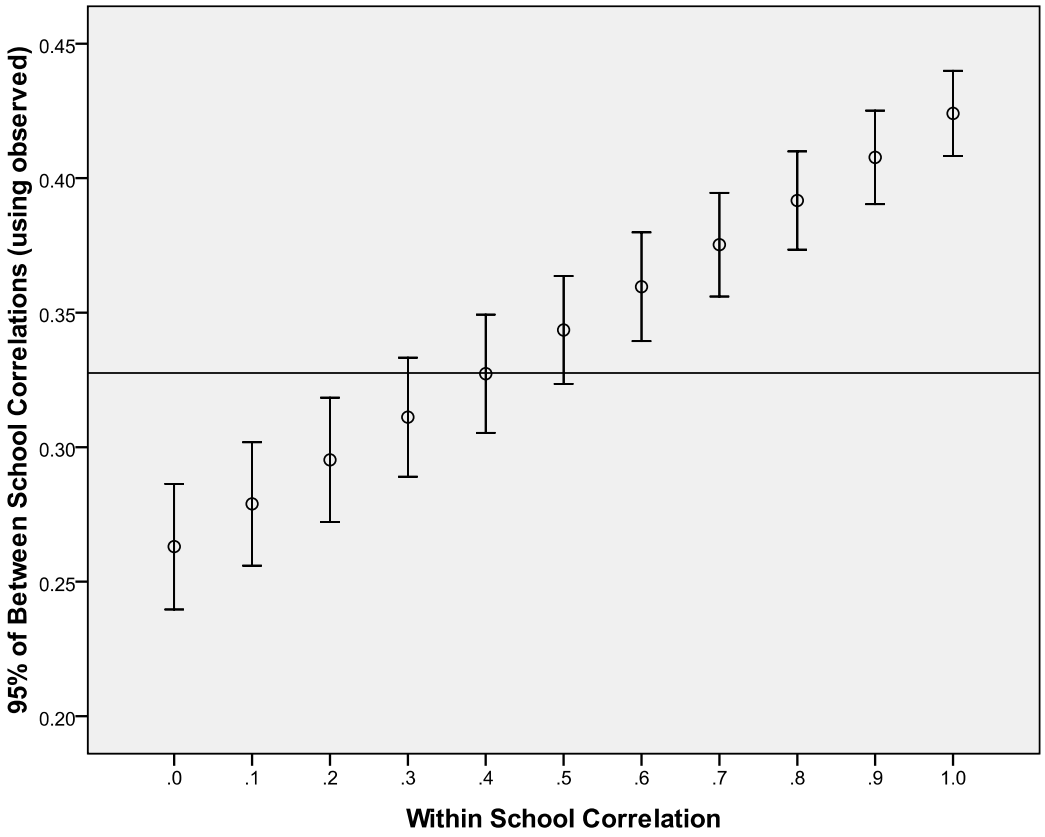


Figure A2: Varying the between-school correlation, within-school correlation = 0 (groups 2-2 and 5-5)

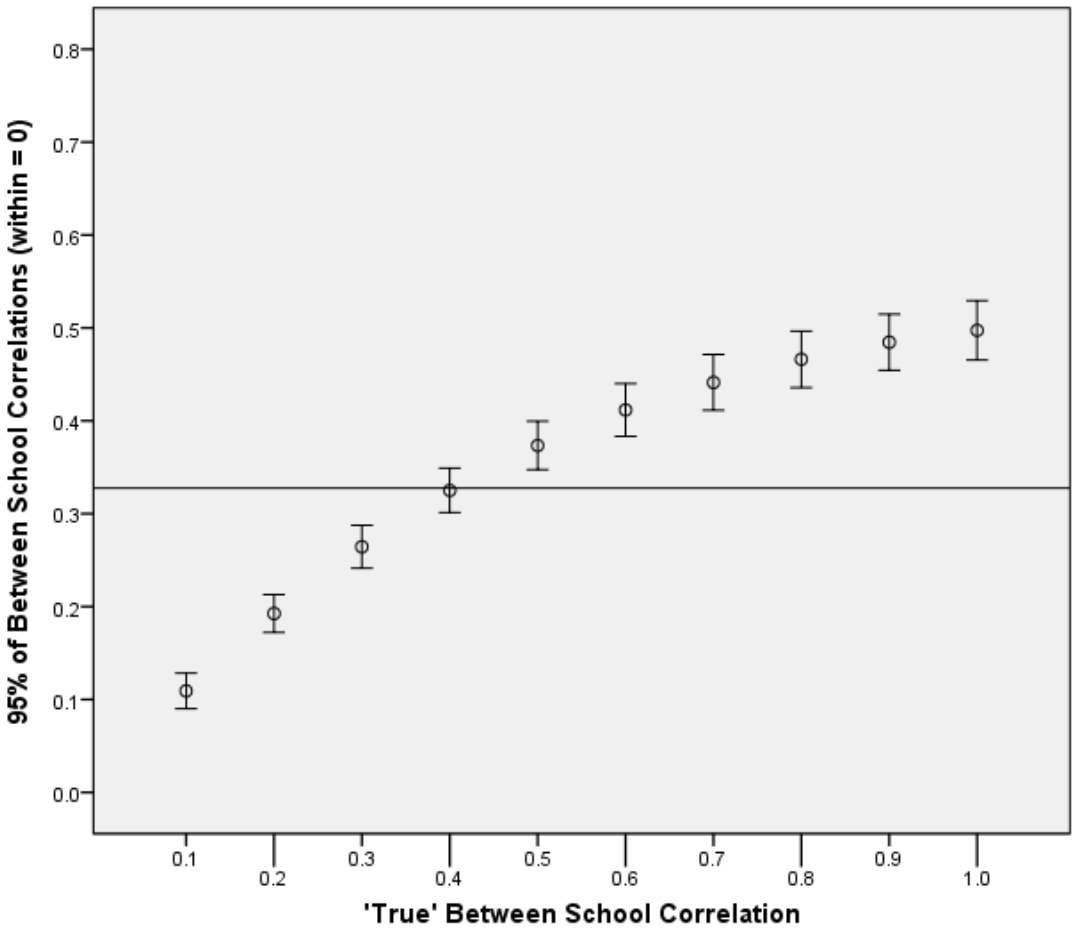


Figure A3: Varying the between-school correlation, within-school correlation = 0.4 (groups 2-2 and 5-5)

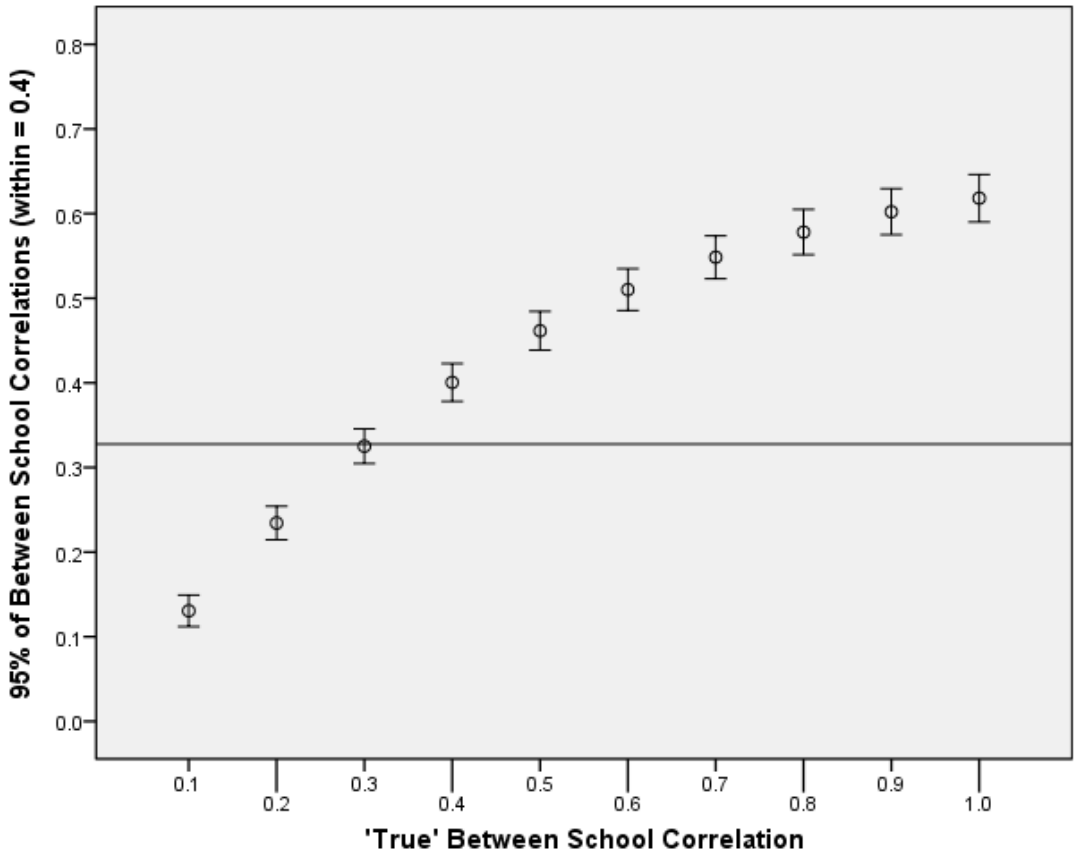


Figure A4: Varying the between-school correlation, within-school correlation = 0.7 (groups 2-2 and 5-5)