It's not which school but which set you're in that matters: the influence of ability-grouping practices on student progress in mathematics[1]

Dylan Wiliam & Hannah Bartholomew, King's College London

Abstract

The mathematics achievement of a cohort of 955 students in 42 classes in six schools in London was followed over a four-year period, until they took their GCSEs in the summer of 2000. All six schools were regarded by Ofsted as providing a good standard of education, and all were involved in teacher-training partnerships with universities. Matched data on key stage 3 test scores and GCSE grades were available for 709 students, and these data were analysed in terms of the progress from key stage 3 test scores to GCSE grades. Although there were wide differences between schools in terms of overall GCSE grades, the average progress made by students was similar in all six schools. However, within each school, the progress made during key stage 4 varied greatly from set to set. Comparing students with the same key stage 3 scores, students placed in top sets averaged nearly half a GCSE grade higher than those in the other upper sets, who in turn averaged a third of a grade higher than those in lower sets, who in turn averaged around a third of a grade higher than those students placed in bottom sets. In the four schools that used formal whole-class teaching, the difference in GCSE grades between top and bottom sets, taking key stage 3 scores into account, ranged from just over 1 grade at GCSE to nearly 3 grades. At the schools using small-group and individualised teaching, the differences in value-added between sets were not significant. In two of the schools, a significant proportion of working class students were placed into lower sets than would be indicated by their key stage 3 test scores.

Background

Reforming education is, as many politicians have discovered, a tricky business. Because the day-to-day practice of a teacher is so intimately linked to the teacher's personality, many aspects of teachers' practice are difficult to change. Furthermore, because what teachers actually do in classrooms is so weakly theorised, attempts at reform have tended to concentrate on administrative aspects of practice, such as the number of episodes into which an hour's instruction ought to be segmented, rather than addressing what, exactly, should be happening in each episode. It is hardly surprising, therefore, that the issue of how cohorts of students should be organised in schools has been a hotly debated issue for many years, and has been a key 'policy lever' that politicians have tried to use to change practice.

The physical arrangement of most schools dictates that students need to be organised into groups of between 25 and 40—most of the rooms in schools are not big enough to take more than 40 students, and there are not enough rooms to have much fewer than 25 in each group. In very small schools (ie where there are fewer than 20 students in each year-group) constraints of funding require that students are taught in mixed-age groups (ie a group in which the difference in age between the youngest and oldest students is, by design, rather than due to retention or acceleration of individual students, more than one year). However, where resourcing is sufficient to provide one teacher per year-group, this almost invariably results in one class per year-group. Although the range of achievement in a mixed-age class is often not much more than in a single age cohort (Wiliam, 1992), there appears to be a great reluctance in schools in the UK to adopt mixed-age classes unless absolutely necessary.

In this context it is important to note that, unlike many other countries, schools in the UK have made little, if any, use of grade-retention (ie requiring students to repeat a year of schooling until they have attained the necessary level of achievement for promotion to the next year). Data from TIMSS (Beaton, Martin, Mullis, Gonzalez, Smith, & Kelly, 1996; Beaton, Mullis, Martin, Gonzalez, Kelly, & Smith, 1996) suggest that less than 1% of students in England are taught 'out of age' while approximately 25% of students in France and Germany have had to repeat at least one year before the age of 14. The use of grade-retention (and acceleration) is intended to reduce the range of achievement in a year-cohort, and where such strategies are not used, other ways of dealing with the range in student achievement are often felt to be necessary (although whether they are, of course, is another matter).

Because of the reluctance to create mixed-age classes, UK schools with one or two teachers per age cohort have tended to focus on within-class grouping, while in schools with larger cohorts, between-class grouping has been more important. Since the primary aim has been to reduce the range of attainment in a class because it is believed that this makes teaching easier, both within-class and between-class groupings strategies have focused on grouping students on the basis of assumptions about ability, achievement, attainment, or, in some cases, motivation. Such grouping systems are usually referred to by schools as ‘grouping by ability’ or ‘ability grouping’ even though what is meant by ability (and in particular whether this is some fixed notion of ability, or just what a student is able to do at a particular time) is rarely made clear. For the remainder of this paper, we will continue to refer to ‘ability grouping’ and ‘mixed ability’ classes simply because that is how schools describe the practice and this provides a convenient label. However, we would wish to make it plain that we believe that such notions of ability are not in any way well-founded and are of dubious validity as predictors of potential.

The advantages and disadvantages of grouping students by ability in schools have been debated for many years. For the first half of the last century, the idea of ‘streaming’ children—allocating them to teaching groups according to some measure of general ability (usually ‘intelligence’ tests)—was so natural as to be unremarkable. Almost all secondary schools arranged students into ‘streams’, so that a student would be taught in the same ‘stream’ for most or all of their subjects, and many of the larger primary schools did the same. Beginning in the 1960s, however, associated with the interest in child-centred education, there was increasing concern about the effects of streaming in primary schools (Jackson, 1964), and the use of between-class ability grouping declined in primary schools during the 1970s and 1980s.

In secondary schools, while streaming may have reduced the range of achievement within a teaching group, the range of achievement was still very wide and in most secondary schools, subject-specific ability grouping or ‘setting’ was superimposed on streaming, especially in mathematics and modern languages.

In the 1960s and 1970s, a number of reports highlighted the problems of disaffection from school experienced by students in lower streams in secondary schools (eg Hargreaves, 1967; Lacey, 1970) and the Banbury enquiry found little evidence that streaming improved academic achievement (Newbold, 1977; Postlethwaite & Denton, 1978). The result was a general move away from between-class grouping of students on the basis of general ability. However, as streaming was abandoned, between-class grouping of students by ability for particular subjects, particularly mathematics and modern foreign languages, was retained, and although precise figures are impossible to establish, it seems likely that the proportion of secondary schools grouping students by ability for at least one subject (usually mathematics) has never dropped below 90%.

More recently, government pronouncements have proposed that “setting should be the norm in secondary schools” (Department for Education and Employment, 1997 p38), and increasing numbers of primary schools are also making use of between-class ability grouping because they believe that this will help improve scores on national curriculum tests (Office for Standards in Education, 1998).

Reviews of research on the effects of ability grouping (Hallam & Toutounji 1996, Harlen & Malcolm 1997, Sukhnandan & Lee 1998) have found little evidence that between-class ability grouping does, in fact, increase attainment. However, many studies have found aptitude-treatment interactions in the effects of ability-grouping—in other words, that the effects of ability-grouping are not the same for all students. For example, several studies have found that the use of between-class ability grouping increases the achievement of the highest attainers, at the expense of lower-attainers, particularly in mathematics (eg Kerckhoff, 1986; Hoffer, 1992; Linchevski & Kutscher, 1998) and Boaler (1997) found a third-order effect in that girls were disadvantaged by placement in top sets.

In the most comprehensive study of between-class ability-grouping conducted in the UK in recent years, Ireson and Hallam (Ireson & Hallam, 2001; Ireson, Hallam, & Hurley, 2002) have investigated the impact of ability grouping on attainment in English, mathematics and science in 45 secondary schools in the UK. They found that the amount of setting experienced by students in key stage 3 (that is, how many years they were taught in setted, as opposed to mixed-ability groups) improved performance slightly in mathematics, but not in science. However, in key stage 4, they found that the amount of setting experienced had no impact in mathematics, or English, but had a slight positive impact on achievement in science. Across both key stages combined, the amount of setting experienced by students had no overall impact in any subject but Ireson and Hallam did find that set placement (ie high, middle, or low sets) influenced progress, both in key stage 3 and in key stage 4. For example, compared with students in middle groups, students in low groups achieved on average approximately one-quarter of a grade less, and students in high sets achieved one-quarter of a grade more at GCSE.

The fragility of these effects suggests that between-class ability-grouping cannot be understood as a simple phenomenon with predictable results. Rather, the practices of ability-grouping are likely to vary from school to school, and if we are to understand how ability-grouping impacts on attainment and influences attitudes, it is necessary to look in detail at how setting is put into practice in schools.

The sample

Between 1996 and 2000 we followed a cohort of 955 students, in six secondary schools in the Greater London area, as they moved from year 8 in 1996-97 to year 11 in 1999-2000. All the students in the six schools were taught mathematics in mixed-ability groups in year 7, but by year 11, all the students were being taught in subject-specific ability-groups or 'sets'. The six schools were chosen to provide a range of learning environments and contexts. During the period of data collection, each of the schools was inspected by the Office of Her Majesty's Chief Inspector of Schools (Ofsted), and each was regarded as providing a satisfactory or good standard of education and all were partner schools with Higher Education Institutions for initial teacher training. The schools were located in five different local education authorities. Some of the school populations were mainly white, others mainly Asian, while others included students from a wide range of ethnic and cultural backgrounds. Table 1 provides information about the cohort of students followed, including the GCSE results, the number of students starting year 10, the number for whom key stage 3 scores were available, and the number in each school actually taking GCSE, together with a brief description of the intake (note that percentages have been rounded to prevent identification of schools).

Table 1 about here

We collected data on the entire cohort of students via questionnaires administered in years 8, 9, 10 and 11, conducted interviews with over 100 students, individually and in pairs, conducted over 150 lesson observations, and collected data on performance on the national curriculum tests at the end of year 9, and on the GCSE examinations at the end of year 11.

In an earlier paper in this journal (Boaler, Wiliam & Brown 2000) we showed that teachers changed their style of teaching when faced with classes of students who had been grouped by ability. In contrast to what had been found by Bennett, Desforges, Cockburn and Wilkinson (1984) for within-class grouping in primary schools, we found that teachers over-estimated the capability of students in the top set, giving them work that was often too demanding, and expecting them to be able to do it quickly, while they underestimated the capability of those in the bottom set. Subsequent papers have described the ways in which setting serves to structure and constrain opportunities for students (Bartholomew 1999, 2000, 2001). In this paper we report specifically on the impact of setting arrangements on the progress of students between KS3 and GCSE, and draw out implications for current government policy on ability-grouping in schools.

From table 1 it can be seen that the proportion of students enrolled at the beginning of year 10 who take GCSE at the school varies from well over 90% (153 out of 165) at Alder school to under two-thirds (112 out of 176) at Cedar. Although some of this appears to be the result of 'drop-out' (ie students who leave and do not go on to take GCSE elsewhere) our discussions with staff at the schools lead us to conclude that the majority of cases are the result of 'turnover' (that is students leaving for other schools). However, it has not been possible to trace these students. For the purpose of this paper the data consists of the 709 students who took both the key stage 3 tests and their GCSE examinations in the same school.

Of course, we cannot assume that the missing data are representative of the data that are present. In fact, those without key stage 3 scores average nearly a whole grade lower at GCSE than those for whom key stage 3 scores are available, and those who did not take GCSE at the six schools scored just over half a level lower at key stage 3 than those who did take GCSE at the six schools. This suggests that the students excluded from this analysis are significantly different from the students included, and so some caution must be exercised when interpreting the results. However, since the main thrust of this paper is the relationship *between* key stage 3 scores and GCSE grades in different sets, the missing data are likely to reduce the numbers in the lower sets, but this is not likely to have a significant impact on the relationship between key stage 3 scores and GCSE grades within those sets.

Results at key stage 3

In order to be able to compare the results of the six schools , and in order to compare the performance of students within the same school when they took different tiers of the key stage 3 mathematics tests, the tiers of the mathematics tests were equated using the level thresholds published by the Qualifications and Curriculum Agency. This allowed all students to be placed on a single mark scale, from 0 to 150. The results of the six schools are shown in figure 1 below[2].

Figure 1 about here

As can be seen, the results of four of the schools are broadly comparable, with Cedar's scores significantly lower, and those at Firtree significantly higher (the shaded areas around the median in each of the box and whisker plots in figure 1 represent 95% confidence intervals for pairs of plots, so that if the shaded regions do not overlap, the two medians are significantly different). Analysis of variance revealed no statistically significant sex differences in key stage 3 scores at any of the schools, although the girls at Redwood did average 12 marks (approximately half a level) higher than the boys which was close to statistical significance ($p = 0.06$).

We collected data on parental occupation from each student during the administration of the questionnaires while the students were in year 9[3]. These were coded using the 7-point OPCS classification, and for the purpose of analysis reported here, these codes were further condensed to a simple dichotomy (ie middle-class/working class).

In four of the six schools, there was no significant difference between the scores of students from working-class and middle-class backgrounds. At Hazel School, middle-class students outscored working-class students by 7 marks, but at Alder school, working-class students outscored middle-class students by 8 marks.

Setting in key stage 4

All six schools taught mathematics to mixed-ability groups when students were in year 7 (age 11). One of the schools (Alder) allocated students to 'setted' ability groups for mathematics at the beginning of year 8 (age 12), three others (Firtree, Redwood and Willow) 'set' the students at the beginning of year 9 (age 13), and the other two schools set students at the beginning of year 10.

Three of the schools, Alder, Redwood and Willow, operated 'traditional' setting arrangements with the students being grouped into five (Redwood and Willow) or six (Alder) strictly hierarchical sets. At two schools (Cedar and Hazel), the age cohort was divided into roughly parallel blocks, with setting within each block (at Cedar the two blocks were of unequal size, and were divided into 3 and 4 sets respectively; at Hazel, each block was divided into five sets). At Firtree School, there were three parallel 'top' sets (so that over 40% of the age-cohort at Firtree school was nominally in a 'top set' in year 10), two parallel set twos, two parallel set threes, and then a set four and a set five. In year 11 at Firtree, students in the three top sets were distributed into two top sets and a set two (so that the set twos became set threes and so on). Table 2 shows the distribution of students into the different sets at the six schools at the beginning of year 10.

Table 2 about here

It is important to note that setting takes different forms in different schools. In most of the schools, setting was clearly based on some measure of attainment in mathematics. However, in Willow School, the sets described as sets 4 and 5 (out of 5) on the timetable were frequently referred to by teachers as the 'behaviour' group and the 'language' group respectively. The strategy had apparently been, at one time, to concentrate students with challenging behaviour in one class, and those with language problems in another. However, many of the students with the most challenging behaviour had subsequently been permanently excluded from the school, with the result that the 'behaviour' group was a rather small group, with few challenging students. Similarly, while the 'language' group did have some students for whom English was not their mother tongue, there were also students who spoke only English, and when asked about the constitution of the group, the teacher said that he did not know the basis on which students had been allocated to his class.

In fact, in four of the schools (Cedar, Firtree, Hazel and Willow) we found that the nature of the setting arrangements were not transparent, so that students (and in some cases, teachers) were not aware of how one set related to the others. At Alder, and, to a lesser extent, Hazel, sets were referred to by their rank, so that students knew that they were in, for example, set 4. The complexity of these arrangements, and the fact that in some schools neither the teachers nor the students understood the basis of the grouping arrangements, suggest that great caution is needed in drawing conclusions about what is going on in schools from what is reported as happening.

We found no significant differences in the key stage 3 scores of boys and girls in any set at any of the six schools (although see section on 'Sex and social class differences' below). However, by using set number as a *dependent* variable in a general linear model, with key stage 3 score and social class as *independent* variables , we did find evidence that working class students were placed in lower sets than middle class students with comparable key stage 3 scores at Alder and Firtree ($p = 0.03$ at Alder, $p < 0.01$ at Firtree). This effect was particularly strong at Firtree school where approximately half the working class students were in a lower set than would be indicated by their key stage 3 scores.

Because of the complexity of the setting arrangements, the number of the set itself conveys little about the experience of the students. For example, set 3 at Cedar might be a bottom set (depending on which block it was in), while at Alder, set 3 would be a middle set. Ireson & Hallam (2000) used a three-fold classification of set into 'higher', 'middle', and 'lower'. However, given what we had learned about the particular circumstances operating in top and bottom sets, reported in Boaler, Wiliam & Brown (2000) we decided on a four-fold classification of set. The highest set (or sets if there were two parallel top sets) at each school was classified as 'top' and the lowest set at each school was classed as 'bottom'. The intervening sets were classified as either 'upper' and 'lower' and since lower sets tend to be smaller than upper sets, middle sets were classified as 'lower'. Thus, with five sets, set 2 would be classed as 'upper' and sets 3 and 4 would be classed as 'lower', while for six sets, sets 2 and 3 would be classed as 'upper' and sets 4 and 5 would be classed as 'lower'.

GCSE results

The GCSE results obtained by the schools are shown in table 3. As might be expected, there is a tendency for the schools with the highest key stage 3 scores to have better GCSE grades, although the distribution of grades is complex. For example, while Firtree has more than twice as many students achieving grades A and B as Alder, the proportion of students achieving at least a grade 'E' at the two schools is the same (86%).

Table 3 about here

A scatterplot of the GCSE score (converted to a uniform mark scale to take into account the different tiers) against the key stage 3 score shows that the relationship is linear, although the distribution of both variables is slightly platykurtic (this plot is not included here due to limitations of space). Plotting the relationship between key stage 3 scores and GCSE score separately for each school shows that the relationship between key stage 3 score and GCSE score is not the same at each school. The most successful school in terms of raw GCSE scores (Firtree) is actually no better than average in terms of value-added from key stage 3 to key stage 4 (in fact, if anything, the value-added here is worse than average, although this is not statistically significant).

Table 4 gives the relative value-added (in terms of GCSE grades) for each of the six schools from key stage 3 to GCSE, in comparison with the other five. As can be seen, for four of the six schools, the relative value-added is not significantly different from zero, but in two of the schools, this difference is significant—at Cedar, students achieve on average 0.28 grades higher than would be expected given their key stage 3 scores, and at Redwood, are achieving 0.27 grades less than would be expected. However, it is worth noting that these differences are actually quite small in terms of the achievement of the students. For example at Redwood, students achieve on average slightly over one quarter of a grade less than would be expected from their key stage 3 scores. This is equivalent to one student in four achieving one grade lower than expected in mathematics—a very small effect and almost negligible in comparison to the effect of prior attainment (ie key stage 3 score). Put another way, the progress made in mathematics in key stage 4 by a student at Redwood School is hardly any different from that made by a student at Firtree School, which is regarded as a highly successful school.

Table 4 about here

The effects of setting

In order to investigate the effects of setting, the relationship between key stage 3 score and GCSE score was calculated separately for top, upper, lower, and bottom sets in each school, and where the relationship differed from set to set, this was tested for significance using analysis of covariance. The relationship across the six schools and for individual schools is shown in table 5.

Table 5 about here

Overall, students in top sets achieve over half a grade (ie 0.58 grades) higher at GCSE than would be expected from their key stage 3 scores, while those in the bottom sets score jusst over half a grade (0.51

grades) lower than would be expected from their key stage 3 scores. Furthermore, in four of the six schools, these effects are absolutely consistent, with top sets doing better than upper sets, who in turn do better than lower sets, who do better than bottom sets. At Redwood, lower sets do slightly better than upper sets (although this is not significant) and at Cedar, the bottom sets do best of all.

Although the trend for higher sets to do better, even when prior attainment is taken into account, is consistent across five of the six schools, the size of the effect varies markedly. At Hazel School, being placed in the top set, rather than the bottom, would improve your GCSE score by about half a grade, while at Redwood, the improvement would be just over three-quarters of a grade. At Alder, it would be well over a grade, over two grades at Firtree, and nearly three grades at Willow. At Cedar, in contrast, it would actually make your GCSE grade worse! The differential performance by set explains some of the differences in overall GCSE performance found in table 3 above.

Now of course, these comparisons are merely for illustration—given the extent of curricular differentiation that we found in our earlier paper, a student moved from a bottom set to a top set would find that they had missed out a great deal of work, and would almost certainly struggle to catch up. But it does show that the set into which you are allocated—an allocation over which students have little, if any, influence at any of the six schools—makes a huge difference to how well you do, and much more of a difference than which school you go to.

What is, perhaps, most interesting, is that the schools where these differences are least marked are Cedar and Hazel, which are both above average in terms of value-added in key stage 4. These are schools which delayed the introduction of setting until the beginning of year 10 and are the schools where teachers continued to make extensive use of small-group and individualised work in key stage 4. Echoing the results of our earlier paper (Boaler et al 2000), it appears that the most pernicious effects of setting may not be necessary consequences of grouping students by ability, but appear when teachers use traditional, teacher-directed whole-class teaching.

Sex and social class differences

Overall, in the six schools, boys outperformed girls at both key stage 3 and at GCSE. At key stage 3, the boys outperformed the girls by nearly half a level (girls' average level: 4.56; boys average level 5.03) and by almost half a grade at GCSE (using the standard GCSE points scores, girls averaged 4.95 and boys 5.37). Because KS3 levels and GCSE grades are not measured on the same scale, we cannot compare them directly, but we can convert the differences between males and females to standardised effect sizes by dividing the difference between the scores of males and females by the (pooled) standard deviation of the scores (Willingham and Cole, 1997). This procedure yields a standardised difference of d=0.35 in favour of boys at key stage 3 and of d=0.24 at GCSE. Girls therefore do, in fact, 'close the gap' somewhat on boys during key stage 4. However, in looking at these data, it is important to bear in mind that the allocation of students to sets in the six schools is not representative. In particular, girls are under-represented in the top, lower and bottom sets, and over-represented in the upper sets. A general linear model of GCSE scores, with key stage 3 scores and sex as independent variables shows that while the relationship between key stage 3 scores and GCSE scores are relatively similar for boys and girls in lower, upper and top sets, there is, in bottom sets, a considerable (and statistically significant) interaction between sex and key stage 3 score ($p<0.01$). In fact higher attaining boys in bottom sets achieve up to a whole grade less at GCSE than girls with similar key stage 3 scores. The effect of this is to depress the attainment of low-attaining boys even further.

Middle class students out-performed working-class students by more than a whole grade at GCSE, but this effect is already mostly present at the end of KS3. Adding social class to the general linear model for GCSE scores showed that working-class students do make less progress in key stage 4 than middle-class students ($p<0.01$), and the size of the effect (just under one-tenth of a grade at GCSE) is consistent with it being caused by the over-representation of working-class students in lower sets referred to earlier.

Discussion

The relationship between key stage 3 scores and GCSE grades will be subject to a number of influences. For example, while all schools are, presumably, trying to maximise their GCSE scores, the same may not be true at key stage 3 (although the recent imposition of targets for schools for achievement at key stage 3 as well as for GCSE may change this). The relationship between key stage 3 scores and GCSE scores may not, therefore, be a valid measure of comparison *between* schools, although of course, it will be a better measure *within* each of the six schools.

A more serious objection to the conclusion that set placement does affect progress in key stage 4 is that schools allocate students to sets on criteria other than just the key stage 3 test results (and in this regard, it is certain that the overlap in key stage 3 scores between sets is substantial). The lower 'value-added' in key stage 4 for lower sets may not be related to set placement at all, but could be because schools place students into particular sets based on notions of 'educability'. If teachers are indeed able to identify students who are capable of getting good GCSE grades despite a modest performance in the key stage 3 tests, then we would expect to find that the apparent value-added during key stage 4 would be highest in the upper sets, and lower in the lower sets, which is what we found in most of the schools. However, if this is the reason for the effect, then we should expect key stage 3 scores to predict GCSE scores least well in schools where set-placement makes most difference, because set placement is not being based on key stage 3 scores. In fact, we find the reverse; there is a modest (and of course, with only six schools, non-significant) but positive correlation of 0.29 between the proportion of variance in GCSE scores accounted for by key stage 3 scores and the size of the difference (in GCSE grades) that set placement makes. This combined with our observations of the teaching in different sets reported in our earlier papers (Boaler, Wiliam and Brown, 2000; Bartholomew, 2000; 2001) leads us to believe that the effects we report are attributable to the process of setting, and the kinds of teaching that result. In brief, teachers teaching bottom sets were generally the least well-qualified to teach mathematics, had lower expectations of their students, frequently set work that was undemanding (often just copying off the chalkboard), used a narrower range of teaching approaches and hardly ever responded to students' frequent requests for more demanding work. In contrast, top sets tended to be allocated well-qualified teachers, who tended to go too fast for many students (particularly girls). Most importantly, teachers teaching setted classes tended to treat the whole class as being of identical 'ability' and made little or no provision for differentiation. The same teachers, when teaching mixed-ability classes, used a wider range of approaches, took greater account of individual differences, and were, in our admittedly subjective view, better teachers, even though they disliked teaching mixed-ability groups.

The data reported here provide further evidence that ability-grouping does not raise average levels of achievement, and, if anything, tends to depress achievement slightly, which is entirely consistent with results from studies conducted in the 1960s and 1970s in the UK, and with the more recent studies conducted in the USA .

More importantly, this study replicates a key finding from earlier studies (eg Hoffer, 1992; Kerchkoff, 1986; Linchevski & Kutscher, 1998) that while ability grouping in mathematics has little overall effect on achievement, it does produce gains in attainment for higher achieving students at the expense of losses for lower attaining students (see also Venkatakrishnan & Wiliam, 2003). This produces an increase in the spread of achievement within the age cohort. In this context is worth noting that every country that outperforms England in mathematics makes less use of ability grouping. Indeed, one of the key findings from international comparisons is that the greater the difference of achievement *between* classes of the same age, the worse that country's overall levels of achievement in mathematics are likely to be (Bursten, 1992)—again consistent with the pattern found here.

The research reported here suggests that, in terms of mathematics attainment, it doesn't really matter very much which school you go to. However, it matters very much which set you get put into. The irony is that current government policy is to allow parents choice as to which school their children attend, which makes little difference in terms of the results their children are likely to achieve. At the same time, by presuming that setting should be the norm in secondary schools, the government is denying parents the choice that really matters—being able to send one's children to a school that does not set for mathematics. Of course, as we know from studies of school choice (see, for example, Gewirtz, Ball and Bowe, 1995) setting is valued by middle-class parents who presumably assume that their children will be in the top sets, but given the disadvantages that setting produces for those who are not placed in the higher sets, we should question whether the parents of higher-attaining children should be allowed to secure advantages for their (already advantaged) children in this way.

However, abolishing setting overnight is not the answer. Time is needed to develop strategies for teachers to work effectively with mixed-ability groups, but the evidence, from both the UK and from abroad (see, especially, Linchevski & Kutscher, 1998), is that teachers can develop strategies for working with mixed-ability groups.

The current government claims to be interested in developing educational practice that is informed by research evidence. And yet, it continues to advocate the adoption of setting in all secondary schools despite the accumulating evidence that setting does not improve overall standards of achievement (and in fact probably lowers them), while also contributing to social exclusion by polarising achievement, and in particular by disadvantaging students from working class backgrounds. One is led, inescapably, to the conclusion that the government's support for ability-grouping is not based on evidence at all, but on political grounds. Setting is presumably believed to be popular with (some) voters. But surely a government elected to a second term, with nearly two-thirds of the seats in Parliament, could begin to think about what might actually improve achievement in our schools, rather than what is politically expedient.

Correspondence: Dylan Wiliam, King's College London, Franklin-Wilkins Building, Stamford Street, London SE1 9NN. Email dylan.wiliam@kcl.ac.uk.

Notes

References

Bartholomew, H. (1999, September) *Setting in stone? How ability-grouping practices structure and constrain achievement in mathematics.* Paper presented at British Educational Research Association 25th annual conference held at University of Sussex. London, UK: King's College London School of Education.

Bartholomew, H. (2000, September) *Negotiating identity in the community of the mathematics classroom* Paper presented at British Educational Research Association 26th annual conference held at Cardiff University. London, UK: King's College London Department of Education and Professional Studies.

Bartholomew, H. (2001, September) *Positioning students in setted mathematics groups* . Paper presented at British Educational Research Association 27th annual conference, held at University of Leeds. London, UK: King's College London Department of Education and Professional Studies.

Beaton, A. E.; Martin, M. O.; Mullis, I. V. S.; Gonzalez, E. J.; Smith, T. A. & Kelly, D. L. (1996). *Science achievement in the middle school years: IEA's third international mathematics and science study*. Chestnut Hill, MA: Boston College.

Beaton, A. E.; Mullis, I. V. S.; Martin, M. O.; Gonzalez, E. J.; Kelly, D. L. & Smith, T. A. (1996). *Mathematics achievement in the middle school years*. Boston, MA: Boston College.

Bennett, N.; Desforges, C.; Cockburn, A. & Wilkinson, B. (1984). *The quality of pupil learning experiences*. London, UK: Lawrence Erlbaum Associates.

Boaler, J. (1997). *Experiencing school mathematics: teaching styles, sex and setting*. Buckingham, UK: Open University Press.

Boaler, J.; Wiliam, D. & Brown, M. L. (2000). Students' experiences of ability grouping—disaffection, polarisation and the construction of failure. *British Educational Research Journal,* **27**(5), 631-648.

Boaler, J.; Wiliam, D. & Zevenbergen, R. (2000). The construction of identity in secondary mathematics education. In J. F. Matos & M. Santos (Eds.), Proceedings of *Mathematics Education and Society* conference, vol (pp. 192-202). Montechoro, Portugal: Centro de Investigação em Educação da Faculdade de Ciências Universidade de Lisboa.

Bursten, L. (Ed.) (1992). *The IEA study of mathematics III: student growth and classroom processes*. Oxford, UK: Pergamon.

Department for Education and Employment (1997). *Excellence in schools*. London, UK: The Stationery Office.

Gewirtz, S.; Ball, S. J. & Bowe, R. (1995). *Markets, choice and equity in education*. Buckingham, UK: Open University Press.

grouping in mathematics. *Journal for Research in Mathematics Education,* **29**(5), 533-554.

Hallam, S. & Toutounji, I. (1996). *What do we know about the grouping of pupils by ability? A research review*. London, UK: University of London Institute of Education.

Hargreaves, D. H. (1967). *Social relations in a secondary school*. London, UK: Routledge & Kegan Paul.

Harlen, W. & Malcolm, H. (1997). *Setting and streaming: a research review*. Edinburgh, UK: Scottish Council for Research in Education.

Hoffer, T. B. (1992). Middle school ability grouping and student achievement in science and mathematics. *Educational Evaluation and Policy Analysis,* **14**(3), 205-227.

Ireson, J. & Hallam, S. (2001). *Ability grouping in education*. London, UK: Paul Chapman.

Ireson, J.; Hallam, S. & Hurley, C. (2002, September) *Ability grouping in the secondary school: effects on GCSE attainment in English, mathematics and science.* Paper presented at British Educational Research Association 28th annual conference held at University of Exeter. London, UK: University of London Institute of Education.

Jackson, B. (1964). *Streaming: an education system in miniature*. London, UK: Routledge and Kegan Paul.

Kerckhoff, A. C. (1986). Effects of ability grouping in British secondary schools. *American Sociological Review,* **51**(6), 842-858.

Lacey, C. (1970). *Hightown Grammar : the school as a social system*. Manchester, UK: Manchester University Press.
Linchevski, L. & Kutscher, B. (1998). Tell me with whom you're learning and I'll tell you how much you've learned: mixed ability versus same-ability
Newbold, D. (1977). *Ability grouping: the Banbury enquiry*. Windsor, UK: NFER Publishing Company.
Office for Standards in Education (1998). *Setting in primary schools: a report from the Office of Her Majesty's Chief Inspector of Schools*. London, UK: Office for Standards in Education.
Postlethwaite, K. & Denton, C. (1978). *Streams for the future? The long-term effects of early streaming and non-streaming—the final report of the Banbury enquiry*. Banbury, UK: Pubansco.
Sukhnandan, L. & Lee, B. (1998). *Streaming, setting and grouping by ability: a review of the literature*. Slough, UK: National Foundation for Educational Research in England and Wales.
Venkatakrishnan, H. & Wiliam, D. (2003). Tracking and mixed-ability grouping in secondary school mathematics classrooms: a case study. *British Educational Research Journal,* **29**(2), 189-204.
Wiliam, D. (1992). Special needs and the distribution of attainment in the national curriculum. *British Journal of Educational Psychology,* **62**, 397-403.
Willingham, W. S. & Cole, N. S. (Eds.). (1997). *Gender and fair assessment*. Mahwah, NJ: Lawrence Erlbaum Associates.

School	Name	type	percent 5 A*-C	starting Y10	KS3 score?	taking GCSE	description of intake
A	Alder	mixed	45%	165	163	153	mainly white, middle & working class
B	Cedar	girls	30%	176	143	112	mainly Asian and working class
C	Firtree	mixed	65%	183	167	167	mainly white and middle class
D	Hazel	mixed	35%	168	161	142	ethnically diverse, middle & working class
E	Redwood	mixed	20%	153	107	116	ethnically diverse, mainly working class
F	Willow	mixed	45%	110	104	89	mainly Asian & African-Caribbean, middle and working class
	Totals			955	845	779	

Note: In an earlier paper (Boaler, Wiliam & Brown 2000) the schools were referred to by letter (A-F). Here, consistent with other related papers (eg Boaler, Wiliam & Zevenbergen, 2000), pseudonyms are used.

Table 1: details of the schools involved in the study

Figure 1: Box and whisker plots of key stage 3 scores

	bottom	lower	upper	top	total
Alder	11	71	53	30	165
Cedar	21	50	51	54	176
Firtree	6	49	49	79	183
Hazel	21	81	28	38	168
Redwood	19	70	32	32	153
Willow	8	30	41	31	110
Total	86	351	254	264	955

Table 2: Allocation of students to sets

	entries	A*	A	B	C	D	E	F	G
Alder	153	0%	7%	21%	46%	71%	86%	97%	99%
Cedar	112	1%	4%	10%	25%	45%	65%	83%	96%
Firtree	167	4%	19%	44%	63%	74%	86%	94%	99%
Hazel	142	2%	9%	25%	45%	65%	85%	95%	100%
Redwood	116	0%	1%	13%	29%	55%	73%	91%	97%
Willow	89	2%	12%	28%	54%	62%	78%	92%	97%

Table 3: Reverse cumulative frequency of GCSE grades achieved

School	relative value-added	p-value
Alder	-0.12	0.06
Cedar	0.28	<0.01
Firtree	-0.09	0.18
Hazel	0.05	0.47
Redwood	-0.27	<0.01
Willow	0.15	0.07

Table 4: Relative value-added in terms of GCSE grades

	Relative value-added (measured in GCSE grades)						
Set	Alder	Cedar	Firtree	Hazel	Redwood	Willow	Overall
top	0.64	*0.05*	1.12	*0.22*	*0.44*	1.43	0.58
upper	0.22(-)	*0.05*	0.34	*0.02*	*-0.17*	0.83	0.16 (-)
lower	-0.16	*-0.37*	-0.47	*0.02*	*0.14*	-0.72	-0.22
bottom	-0.72	*0.27*	-0.99	*-0.25*	*-0.41*	-1.54	-0.51

Notes

1. Differences in italics are not significant ($p > 0.05$)
2. (-) indicates a significant interaction between KS3 score and set in favour of lower attainers
3. Scheffe post hoc comparisons show that all the overall differences except those between lower sets and bottom sets are statistically significant ($p < 0.01$)

Table 5: Relative value-added in terms of GCSE grades by set

[1]Paper presented at the 27th annual conference of the British Educational Research Association, University of Leeds, September 2001.
[2]In a box and whisker plot, the box represents the attainment of the middle half of the data, with the line indicating the value of the median. The whiskers extend far enough to include most of the remaining data (specifically, the whiskers extend far enough to encompass 99.5% of normally distributed data).
[3]Our experience has been that it is difficult to collect reliable data on parental occupation without actually visiting classrooms and collecting the data ourselves. We asked students to provide information on the jobs done by parents or guardians, or, if they were out of work, what job they did when they last worked. An indication of the problematic nature of the data is provided by one incident when we collected information at Redwood School. A girl asked one of us (DW) for help as she didn't know what to put for her father's job. When asked, "What does your father do?", the girl replied, "He's a waiter, but when we were in Iran, he was a professor of History".