The impact of tracking by attainment on pupil self-confidence over time: demonstrating the accumulative impact of self-fulfilling prophecy.

Becky Francis[a], Nicole Craig[b], Jeremy Hodgen[a], Becky Taylor[a], Antonina Tereschchenko[a], Paul Connolly[c] and Louise Archer[a]

[a]UCL Institute of Education, London, UK

[b]Queen's University Belfast, Belfast, UK

[c]Lancaster University, Lancaster, UK

Abstract

The impact of self-fulfilling prophecy in education, and that of attainment grouping on pupil self-perception, remain topics of longstanding debate, with important consequences for social in/justice. Focusing on self-confidence, this article draws on 9,059 survey responses from 12-13 year olds who had experienced two years of tracking by subject ('setting'), and had provided survey responses shortly after having been placed in 'ability' sets at the start of their secondary schooling, and again a year later; enabling analysis of impact over time. After controlling for prior attainment, the gap in general self-confidence between students in the top and bottom sets for mathematics widened over time, although surprisingly this was not the case for self-confidence in mathematics and nor for the gap in general self-confidence between students in the top and bottom sets for English. These implications of these findings for interventions directed at addressing educational disadvantage are discussed.

Key words: tracking, social justice, social inequality, self fulfilling prophecy, social background, attainment grouping

The impact of tracking by attainment on pupil self-confidence over time: demonstrating the accumulative impact of self-fulfilling prophecy.

Introduction

The impact of tracking by attainment on pupil experiences and outcomes has been long debated by sociologists. There are different forms of tracking – by which we mean, division of students according to levels of educational attainment[i] – including *between-school tracking* and various manifestations of *within-school tracking*. To add to complexity in analysis of this issue, terminology tends to differ in different countries. So, for example, where attainment grouping is referred to as 'tracking' in the United States and some other countries, in the United Kingdom and many of its former colonies, the term 'tracking' is largely unknown. Instead, the specific practice is named – for example, streaming (within-school, cross-subject tracking), or setting ('tracking by subject', as it is known in the US)[ii]. Nevertheless, in spite of this complexity, the international literature on attainment grouping has been clear that these practices have implications for social in/justice in education, in relation to student experiences and outcomes.

Nancy Fraser (1997) distinguishes between social injustices of recognition and distribution, and it is arguable that both elements are manifest in attainment grouping (Author 1 et al, 2019a). Recognitive injustice is evidenced by the long-established social inequality in *allocation* to attainment groups, with pupils from low socio-economic groups (especially boys), and from certain minority ethnic groups, over-represented in low tracks (Jackson, 1964; Muijs & Dunne, 2010; Moller & Stearnes, 2012; Author 7 et al, 2018 ; Boaler, 1997), even after prior attainment has been controlled for (Dunne et al, 2007; Author 6 et al, 2019). Distributive injustice is manifest in the differences in resources and expectations found to be channelled at different tracks (Gamoran 1986; McGillicuddy & Devine, 2018; Mazenod et al., 2018; Author 1 et al, 2019b); and the difference in educational progress and outcomes wherein pupils in low attainment groups make poorer progress than their peers in higher groups (Kulik & Kulik, 1982; Slavin, 1990; Ireson et al, 2005; Kutnick et al, 2005; Steenbergen-Hu et al, 2016; EEF, 2018).

A further element of recognitive injustice instigated by attainment grouping that has interested sociologists is the labelling associated with placement in a particular 'track' or attainment group (Oakes, 1986; Boaler et al, 2000; Marks, 2016; Author 1 et al, 2017b; Mazenod et al, 2018). A variety of scholars have drawn productively on Lemert's (1951) and Becker (1963)'s theory of labelling to analyse the ways in which perceptions and labels generate self-fulfilling prophecy (Merton 1948) in

educational contexts. A striking contribution was made by Rosenthal and Jacobson's (1968) 'Pygmalion in the Classroom' study; the title capturing with self-explanatory power their findings on the effects of labelling on teacher expectations of pupils and the impact on subsequent outcomes. Rosenthal and Jacobson showed how teachers' expectations of a randomly-selected group of pupils labelled 'late bloomers' resulted in a greater increase in IQ points for these pupils over the following year than for the control group (the other pupils in the class). The findings have been widely debated, and Jussim and Harber's (2005) wide-ranging review of the evidence finds that discrepancies in pupil outcomes as a result of teacher expectations have been exaggerated. The debates provoked by 'Pygmalion in the Classroom' among psychologists have tended to focus somewhat narrowly on teacher expectations (see Jussim & Harber, 2005), rather than a more expansive view of self-fulfilling prophecy and the multiple actors that can be involved in the interactive development of understandings and behaviours precipitated by a label. Jussim and Harber (2005) maintain that this literature demonstrates that self-fulfilling prophecy due to teacher expectations *is* supported by the research evidence, but that effects are small, dissipate over time, and are often explained away by pupil 'ability'[iii]. They recognise, however, the evidence that social variables have an impact, with pupils from low socio-economic backgrounds and African American pupils more significantly affected by teacher expectations than other pupil groups (Jussim & Harber, 2005). UK research supports this finding: for example some Black ethnic groups in England are systematically under-represented in entry the higher tiers in assessment at age 14 (Strand, 2012). Jussim et al (1996) themselves found evidence of moderation of self-fulfilling prophecy effects for mathematics achievement for socioeconomic status and for ethnicity. However, when prior attainment was taken into account, these expectation effects were greatly reduced and Jussim and colleagues argue that teacher expectations were largely accurate and self-fulfilling prophecy effects much smaller than originally appeared (see also Jussim & Harber, 2005).

It is worth remarking that these pupil groups – those from low socio-economic backgrounds, and from particular minority ethnic groups - are disproportionately likely to be allocated to low attainment groups/tracks (see above), albeit Jussim and Harber (2005) do not make this connection. Their brief attention to the issue of tracking is captured in the short section 'tracking by ability level', wherein they cite a study by Smith et al (1998) which found no evidence that tracking generated stronger self-fulfilling prophecies than mixed attainment classrooms. We return to these various assertions in our Discussion section.

Meanwhile, while the above studies focused primarily on student attainment outcomes, the impact or otherwise of tracking on pupils' sense of self is similarly contested. Here again, the evidence is

complicated by a plethora of different theoretical constructs and terminology in different disciplinary and international studies. These have included attention to the constructs of self-confidence, self-esteem, self-concept, and/or self-efficacy - all of which are somewhat distinct, and reflect different disciplinary perspectives (albeit these epistemologies are not always articulated). See Authors (2017b) for a discussion of these distinctions. Sociological work tends to explore notions of self-confidence and esteem, and this is where we situate our research. We adopt the TIMMS definition of self-confidence in learning as "student's positive/negative beliefs about his/her ability to learn […] with respect to himself/herself, other students, and the teacher." (Mullis et al, 2016). Albeit in practice our survey is largely based on academic self concept measures, and its definition, which bears close conceptual affinity (see Ireson & Hallam, 2009, for detail).

As we shall explain below, there have been a range of studies on the impact of track allocation on self-confidence and/or self-concept, and previously the thrust of findings has been somewhat unclear. However, a prior paper from our [NAME] study has made a major contribution to clarity in this field, drawing on a large-scale sample of secondary school students recently set[iv] in mathematics and English. The quantitative analysis demonstrated that set placement correlated with pupil self-confidence, not only in the subject in which they were set (which of course might be expected for a range of potential reasons), but also with general self-confidence in learning (Author 1 et al, 2017b); a strong indicator of self-fulfilling prophecy precipitated the labelling integral to setting. What we did not know was whether these effects were long-lasting, and how self-confidence according to track level develops over time. This paper reports data at post-test with the same large-scale group of pupils across England, two years into the study, to make a contribution of new knowledge on this issue of significant import for social justice in education.

The existing literature on tracking and pupil self-perception

The earlier literature on student self-perception in relation to tracking/attainment grouping was outlined in detail in our prior article (Author 1 et al, 2017b), but we summarise it very briefly here. Some studies have shown a relationship between student self-concept or self-confidence, and tracking, with those in higher/academic tracks showing higher self-concept or self-esteem than those in lower/'vocational track' students (Ireson & Hallam; 2009; Chmielewski et al. 2013; Van Houtte et al. 2012; Liu et al, 2005).

Conversely, others found negligible or no relationship between tracking and self-concept (Liem et al, 2015; Kulik and Kulik, 1982). And Belfi et al's (2012) literature review surprisingly concluded that 'ability' grouping is beneficial for the academic self-concept of lower attaining students.

Likewise, the correlation between self-concept and attainment group established in Ireson & Hallam's (2009) study did not extend to *general* self-concept. Marsh and Parker's (1984) 'Big-fish-little-pond effect' concept highlights the relativistic nature of self-concept. They assert that self-concept depends on a frame of reference, observing that 'ability' grouping is likely to have "substantial effects on self-concepts within different ability groupings" (p. 799). Hence Marsh (2008) later showed that equally 'able' students have lower academic self-concept when attending schools where average attainment levels are high, than when attending schools where peer attainment is low. This would indeed, then, seem likely to have a bearing on pupils' self-perception within different tracks.

Yet contrasted with this conceptual frame and resulting hypothesis is labelling theory (see Lemert, 1951; Becker, 1963). As explained above, applied to attainment grouping, this theory suggests that the act of labelling a pupil a 'low' or 'high' attainer (or even as low or high 'ability') manifest in the allocation to a particular track, can be anticipated to precipitate a self-fulfilling prophecy (see Merton, 1948; Jackson, 1964). Here it is predicted that the different resources and expectations applied to pupils in different attainment groups, coupled with the impact of the label on students' own self-perception, leads to the prophecy being realised – in other words, impacts pupil self-confidence (and outcomes).

Our study sought to contribute to this debate, exploring the hypothesis that attainment grouping impacts pupil self-confidence, precipitating a self-fulfilling prophecy. We found a significant correlation between perceived set placement and self-confidence in the set subject. More importantly, we also found a correlation between set placement and general self-confidence in learning (Author 1 et al, 2017b). Pupils in low sets had lowest self-confidence in mathematics, English, and general learning; whereas for top sets the reverse was the case and these pupils consistently had the highest self-confidence. Application of psychosocial analysis (Hollway & Jefferson, 2013) to qualitative data from the study revealed the effects of labelling on pupil self-perception, internalisation of 'ability' labels among pupils, and the interactive processes via which tracking manifests a self-fulfilling prophecy (see Author 1 et al, 2017b).

Nevertheless, the quantitative data underpinning this prior publication was collected shortly after pupils had been placed into sets, early in their first year of secondary schooling. As such, apart from the initial impact of the label on students, the impact of setting in terms of application of resources and teacher expectations would not have had long to impact. Over time the effects could be hypothesized to be exacerbated, or conversely to be dissipated. For example, proponents of tracking often argue that pupils in low groups would be daunted by working with higher attainers (and/or

high attainers frustrated by working with lower attaining peers); and the 'big-fish-little-pond' effect (Marsh & Parker, 1984) would suggest a hypothesis that pupils within groups of similar-attaining students might grow in self-confidence over time. Moreover, although focused on impacts on IQ/attainment rather than self-confidence, Jussim & Harber (2005)'s review finds a tendency for self-fulfilling prophecy derived from teacher expectations to dissipate somewhat over time (see also Raudenbush, 1984).

Hence it appeared important to capitalise on the opportunity provided by the longitudinal nature of our research project to test the impact on self-confidence once students had experienced two academic years in set groups. As such, this article seeks to make a major contribution to the literature by building on our prior findings and the longitudinal aspect of our project to report on the development of self-confidence over time in relation to set placement.

Methodology

The wider study

The data discussed in this article draws from a large scale mixed-methods project '[NAME]', funded by the Education Endowment Foundation. The project sought to address prior gaps in the literature, by exploring: whether practice in setting[v] that remediates some of the problematic practices identified in the literature as affecting those in low groups might improve young people's progress; what comprises good practice in mixed attainment pedagogy; and the experiences and outcomes of pupils subject to attainment and mixed attainment grouping. It included the following methods:

- Two two-year interventions, one tested by a fully-powered RCT ('Best Practice in Setting') and one constituted as a randomised feasibility study ('Best Practice in Mixed Attainment Grouping')[vi], examining impact or otherwise of practice in grouping students in Year 7 and Year 8 based on research evidence.
- Surveys of pupils and teachers involved in the study
- Individual and focus group interviews with 245 pupils, and 56 teachers

The interventions and research were undertaken in 139 secondary schools (divided into intervention or control groups), and involved instigating work with and monitoring student cohorts from the beginning of Year 7 (11-12 years old) to the end of Year 8 (12-13 years old), focusing on their experiences and outcomes in English and Mathematics. English and mathematics were selected as the foci because: a) they are two subjects given longstanding priority in the national curriculum and within school performance indicators; and b) they represent diversity in content and pedagogy.

The findings of the cluster RCT have been previously reported in the evaluation report (Roy et al, 2018), showing a lack of significant impact from the intervention on the outcome measures of pupil self-confidence and attainment, in comparison to the control group. We have discussed elsewhere the reasons for this lack of effect (Francis et al, 2019), which include low fidelity (expressive of difficulties schools had in implementing the intervention effectively, see Taylor et al 2018).

Outline of the sample and data analysed in this article

The quantitative data reported here are generated by surveys, and drawn exclusively from the 'Best Practice in Setting' cluster randomised controlled trial (RCT) study of the effectiveness of schools adopting 'best practice' in attainment setting, as against 'business as usual setting' (for detail see Roy et al, 2018). Hence, all schools from which the data is collected are practising setting (either in the intervention group or the 'business as usual' control group[vii]), appropriate to our focus here on relative levels of self-confidence for pupils placed in different set levels.

The total of 126 schools were recruited to the trial through a mixture of volunteer and direct 'cold call' approach sampling, then randomised to the intervention and control groups of the RCT. Volunteer-sampled schools were recruited through a traditional and social media campaign by the authors. Direct approach-sampled schools were identified through a stratified random sample then approached by the National Foundation for Educational Research (see Styles, 2017). In total 1006 schools were approached to generate the sample. Schools were distributed across England, and the sample is broadly reflective of the national sample of schools. Schools were eligible for the study if they were state funded, non-selective by attainment and already setting in mathematics in Years 7 and 8.

The survey (including self confidence measures) was offered to all schools in the broader study. Our analysis focuses on responses from those pupils experiencing setting: 9,059 Year 8 pupil responses from control and intervention schools within the 'Best Practice in Setting' trial, including 6,167 students from 60 schools participating in the trial for mathematics, and 2,892 students in 30 schools participating in the trial for English (we refer below to the *mathematics trial*, and the *English trial*). Summary characteristics and further details of the sample are provided in Table 1. Social class background was analysed via questions concerning parental/carer occupation, with categorisation according to the highest status occupation between parents. Following this analysis, the tiered occupations were further categorised as 'professional/managerial, intermediate, and semi/unskilled'.

This analysis is based on the sample of schools where students completed the surveys at both the beginning of Year 7 and at the end of Year 8 enabling an investigation of how the change in self-confidence over time is associated with set placement. To enable comparability with our earlier

analysis of self-confidence at the start of secondary school, the sample combines the intervention and control groups, both of which allocated students to sets. Table 1 presents the sample participating in the different trials according to identity subgroups. While many of the groups contain large numbers, there are small numbers of students in some groups, particularly in relation to ethnicity.

Table 1 here

The first survey was administered in Autumn 2015, soon after pupils had arrived at secondary school and had been placed in attainment groups. The second survey was administered as pupils were reaching the end of Year 8, to explore the impact or otherwise of two school years of experience of this grouping on their self-confidence.

The questionnaire completion process was administered by school teachers, following instructions on administration protocols. Classes completed the questionnaires online. The questionnaires took approximately half an hour to complete, and included questions on perceptions of mathematics and English, liking for school, and perceptions of attainment grouping. They included various self-confidence measures constructed of a range of items. Questionnaire items were partly drawn from Ireson and Hallam (2009) with additional of our own, and had been extensively piloted with students in the pilot year of our project.

The main foci for the quantitative analysis to follow are three measures of self-confidence: self-confidence in English, self-confidence in mathematics and general self-confidence in learning. The items used for each are detailed in Table 2. These measures have been adapted from the self-confidence scales used in the international TIMSS and PIRLS studies (Martin & Mullis, 2012). All three measures were found to be unidimensional and thus valid and also, as shown, were found to be reliable. As would be expected from TIMSS, overall self-confidence decreases over Y7 and Y8 (see Table 3).

Table 2 here

Table 3 here

Schools in our sample varied in relation to the number of set levels they applied, from two to ten, with most falling between three and five (intervention schools in the setting trial had been specifically asked to cap set level number at maximum four). For the purposes of this current analysis, students were coded into three groups for English and mathematics respectively in each school: those in the very top set; those in the middle set(s); and those in the very bottom set[viii]. Thus, for a school with four sets, the top set was coded ‘1’, the middle two sets coded ‘2’ and the bottom

set coded ‘3’. Similarly, for a school with five sets, the top set was coded ‘1’, the middle three coded ‘2’ and the bottom set coded ‘3’. The breakdowns of the sample by these three categories for English and mathematics are also shown in Table 1, above.

In order to calculate how many students moved sets throughout the two years of the trial, data on students set levels were collected at the end of Year 8. The new set level data was then converted to the top, middle and bottom set allocation as was described above, and the students’ movement between sets was then calculated by giving a child a 1 if they moved in either direction between sets levels. The level of movement between sets was a little higher than expected at around 20%. This may be because schools in the intervention group were encouraged to move students between sets regularly on the basis of school assessments. However, this is not a threat to our findings, because our setting variable, set placement at the beginning of Y7, captures the effect of school’s actual setting practices on these students.

The data were analysed by fitting a series of multilevel models with students (level 1) clustered within individual subject sets (level 2) and then within schools (level 3). In each model, dummy variables representing the three categories of set level (top, middle and bottom) were included along with a series of other covariates representing gender, family occupation, ethnicity and total number of sets within the school. The models were then used to estimate the adjusted mean self-confidence scores for students in the three set levels, controlling for these covariates. Practically, this was done by adding in a series of values to the model. These values consisted of either: the relevant values of the dummy variables for the set levels (i.e. either ‘0’ or ‘1’); or the mean scores for each of the other covariates included in the model; or ‘1’ for the constant. The mean self-confidence score was then calculated by adding together the products of each of the coefficients in the model with its associated value. The standard deviations for each of the mean scores estimated were calculated using the standard error of the associated null model multiplied by the square root of the sample size to account for the clustered nature of the data and the size of each sub-sample represented the total number in each category for whom there were full data (and thus whose data were included in the model). When controlling for prior attainment, decimalised Key Stage 2 scores were used.

Findings

The three level hierarchical linear regression models were carried out and showed that there were significant differences between the self-confidence of students in top and bottom sets at post-test

when compared with an average student in the middle set and controlling for their pre-test self-confidence scores, household occupation, ethnicity and gender.

Table 4 here

Figure 1 here

It can be seen from Figure 1 that, when compared with an average student in the middle set, there is a trend for relative self-confidence in mathematics where students in the top set show significantly higher levels of self-confidence after two years (ES=.115, p<.001), and in comparison students in the bottom set show significantly lower self-confidence over time (ES=-.142, p<.001). It needs to be borne in mind that this growing gap exacerbates an unequal starting point wherein student self-confidence was shown in our prior analysis to correlate with set level shortly after placement in set groups at the beginning of secondary schooling (Author 1 et al, 2017b). As such, this is a deeply concerning finding. The trend is also shown for students' results when they report on their general self-confidence in math scores, showing that this impact on self-confidence extends beyond set subject. This indicates a strong relationship between labelling from setting and self-confidence outcomes, in the absence of other clear explanatory factors as to why general self-confidence in learning has risen overall for the highest sets and decreased overall for the lowest sets (we would otherwise expect the same rates of development for all pupils). We discuss this point further below.

The trend is slightly less clear for students in the English trial, but shows a similar trajectory. Students in the top set similarly have significantly higher relative scores on self-confidence in English when compared with the middle set after two years (ES=.073, p=.006) and general self-confidence in English (ES=.081, p=.001). However, no statistically significant differences from middle set pupils were found for students in the bottom set in English. And although the effects for top set students in the English trial were found to be significant, they were relatively small. Nevertheless, as can be seen from Figure 1, the overall results are striking: over the two years, students in the top set tend to have higher self-confidence compared to those in the middle set, whereas for students placed in the bottom sets the opposite is the case.

The above analysis was repeated with the added control for students' prior attainment at Key Stage 2[ix] in either mathematics or English (depending on which trial they were in). In other words, we sought to ensure that the trends identified above could not simply be related to prior attainment that might indicate low attainers reduce in confidence at secondary school, and/or high attainers increase in confidence; but rather trends can be attributed to the impact of setting. This was especially crucial given Jussim and Harber's (2005) conclusion that self fulfilling prophecy effects are actually minimal once 'ability' is controlled for. Seeing teacher expectations as a proxy for self-

fulfilling prophecy, their analysis of the literature leads them to assert that "teacher expectations predict student achievement primarily because they are accurate" (p. 141). Their focus is on attainment outcomes rather than self-confidence, but clearly attainment could be expected to have a bearing on self-confidence. However, we found that after controlling for prior attainment, some differences do remain in students' self-confidence two years after pre-test for those in the mathematics trial, although the effects are smaller. In the mathematics trial, students in the top set had significantly higher general self-confidence compared to the middle set after two years (ES=-.057, p=.004) whilst the bottom set had significantly lower general self-confidence (ES=-.561, p=.040). In the English trial, after controlling for prior attainment, students in the top set had significantly higher self-confidence in English (ES=.060, p=.035) after two years when compared with students in the middle set, although the effect for general self-confidence is no longer significant. All these results are presented graphically in Figure 2.

Table 5 here

Figure 2 here

As the models are three level multi-level models, the variance explained by each model was calculated by adding together the variance at the levels of the school, set level and student found in the multilevel models and subtracting the total from the total variance explained by the null model. The variance explained by the mathematics self-confidence model increased from 20.17% to 22.93% when prior attainment was controlled for. It increased further to 23.54% when the prior attainment variable used was the KS2 decimalised level in the sensitivity analysis. This was also the case for the models assessing the general self-confidence of the students involved in the mathematics trial where the variance increased from 22.40% to 24.09% when prior attainment was controlled for and 24.26% when the decimalised KS2 scores were used and in the English trial where the variance increased from 21.46% to 22.20% when prior attainment was controlled for and 23.69% when the KS2 decimalised score was used. This pattern suggests that the models controlling for prior attainment provide a better explanation of the dependent variable than those which do not. However, this does not translate to the English self-confidence models which decreased from 12.30% to 12.03% when controlling for prior attainment and even further to 12.02% when the decimalised KS2 score was used at a covariate. Although the decrease was small, it is still worth noting that controlling for prior attainment does not explain more of the variance for the student's self-confidence in English.

Overall, the analysis shows that when compared with two years previously, there was a general trend that students had higher self-confidence in the subject area of mathematics or English if they

were placed in the top set and a significantly lower self-confidence when placed in the bottom set in mathematics when compared with an average student in the middle set. This trend in self-confidence remained for those in the bottom set in mathematics and those in the top set in English after controlling for attainment level. As such, this provides novel and significant evidence on the relationship between setting and pupil self-confidence, and its development over time. It also demonstrates conclusively that, in contrast to Jussim and Harber's (2005) analysis in relation to attainment outcomes, self-fulfilling prophecies in relation to pupil self-confidence - precipitated by labelling through tracking - accumulate over time.

Discussion

The findings are important for three reasons. Firstly, they provide original evidence from a large-scale study to support the longstanding suggestions from the existing research literature that tracking – in this case setting ('tracking by subject') is inequitable, with some negative impacts on low attaining pupils that accumulate over time. As we have seen, this is has potentially important implications for social justice, both in the implications that low attainers are being ill-served in schools that apply tracking, and additionally because low attainment groups are shown to be disproportionately populated by pupils from low socio-economic backgrounds and from particular ethnic groups. (This was also shown to be the case in our study, see Author 6 et al, 2019). The new evidence that differentials in general self-confidence in learning for students placed in sets for mathematics – identified following set allocation at the beginning of secondary schooling (Author 1 et al, 2017b) – develop further over time, further advantaging high attainers in comparison to low attainers, is worrying. It is also somewhat surprising that there was no statistically significant change in the gap for self-confidence in mathematics, because, on the basis of Ireson & Hallam's (2009) findings, it would be expected that the effect of setting to be greater in the subject itself. The findings for students placed in sets for English were more equivocal. The increase in the English self-confidence of students placed in top sets for the subject in comparison to those placed in the middle set was statistically significant, although there was no statistically significant effect on general self-confidence.

Secondly, these results have important implications for interventions directed at addressing disadvantage in education. The effect size on general self-confidence differential for the students in mathematics sets is generally thought of as small ($d \approx 0.12$) in terms of Cohen's (1988) rules of thumb. However, we consider this effect to be practically significant since few educational interventions achieve an effect larger than this in trials at scale. (See Education Endowment Foundation, 2020, for a discussion of effects of this size in education.) Macleod et al. (2015) found that more than a third

of schools in England adopted setting as a strategy for addressing educational disadvantage. With respect to setting in mathematics, our results suggest that this strategy may negate the potential benefits of other more effective strategies adopted by schools, at least in terms of general self-confidence.

Thirdly, the trend provides evidence of a relationship between set placement and self-confidence. Whereas it might have been possible to argue that the relationship we have previously established between self-confidence in a subject/learning and set placement might relate more to pupils' awareness of their relative 'abilities' in the set subject than to set placement per se, this hypothesis would anticipate levels of self-confidence to hold constant, rather than for the disparity between set groups to grow. To check that trends could not be attributable to prior attainment we subjected our data to control for this factor, and show that the trends for general self-confidence for mathematics sets and for English self-confidence remain significant thereafter. Whilst these effects are relatively small, we note the concern that they appear to accumulate over time starting from an existing disparity at the beginning of secondary school. This suggests that it is the act of attainment grouping rather than other factors that precipitates these trends. We recognise that there may be other issues associated with bottom set groups that might also impede the development of self-confidence over time, such as absenteeism or exclusion – albeit it is worth noting that these may also be precipitated by designation to a bottom set group and the disassociation with schooling entailed (Author 6 et al, 2018).

This challenge to disaggregate relevant factors, and the intersectional and centrifugal nature of many of these, remains a difficulty for research. As the psychology literature suggests, there may be a range of different psychological factors and processes which mediate the affects between the receipt of an 'ability label' via tracking, and self confidence in learning. Factors potentially indicating disaffection (noted above) may also be consequential to self-confidence over time; either as expressions *of* lack of self-confidence, or possibly as causes of lack of self confidence (if, say, substantial amounts of schooling are missed). Furthermore, it may be questioned as whether these self-confidence outcomes can be attributed to the labelling precipitated by setting (and subsequent self-fulfilling prophecy), or by practices associated with setting (such as differential pedagogy provided by teachers due to their expectations of pupils; Mazenod et al, 2018; Oakes, 1985; McGillicuddy and Devine, 2018). However, firstly, many of these elements are arguably a fundamental part of the prophecy generated by the label of set level. In other words, that the institutional label of 'low set/high set', and/or 'low ability/high ability' influences the interaction and behaviours of a range of stakeholders around and including the individual pupil, is precisely what Becker (1963) would have anticipated as integral to the interactive process of realisation of the

prophecy. Moreover, given that the focus of our data here was specifically on pupil self-confidence, we consider this provides strong evidence for the impact of the labels inherent in setting on *pupil* self perception in relation to their learning, subject identification, and feelings about themselves, as learners, and about their place in school. We do not think it unreasonable to hypothesise that these trends in self-confidence likely impact on pupils' dis/associations with schooling, and in turn on pupils' perceptions of their futures. More research would be required to test this.

Indeed, this finding of the accumulation of impact of tracking on pupil self-confidence over time suggests a re-evaluation of the 'self-fulfilling prophecy' (Merton, 1941) explanation which we posited to explain our prior findings. We suggest our findings illustrate how tracking constitutes a *snowball prophecy*. The process of labelling inflicted by tracking (in this case, setting) appears to be a cumulative one that is reinforced and thus exacerbated by maintenance of these categories in tracking over time. Its impact on pupil self-confidence appeared evident shortly after placement into sets, but this fulfilment of the prophecy appears to build further over time. Hence, the original prophecy interpolated by the 'ability track' label *snowballs* as it builds momentum and impact via the various practices, understandings and behaviours on the part of the individual concerned (pupil), inter-actors (teachers, parents, peers), and organisational structures (the school and its practices). This conceptual contribution may have explanatory power in relation to the factors discussed above, and hence contributes an important clarification on the conceptualisation of self-fulfilling prophecy: that the outcome ('fulfilment') is not fixed, but rather is dynamic, and can indeed be cumulative.

In terms of social in/justice, our findings suggest that tracking is indeed promoting both distributional and recognitive injustice (Fraser, 1997). It is worth considering that Rosenthal and Jacobson's study only manipulated *positive* expectations: to simulate negative expectations would have been ethically problematic. But it is arguable that we may be doing something very similar in a routine way in subjecting pupils to labelling by tracking – and that this labelling includes negative as well as positive labels. Our study suggests a growing gap for self-confidence between bottom and top set pupils, which risks cementing existing inequalities rather than dissipating them. These findings indicate a challenge for educators, showing the importance of improving equity in practices of pupil grouping in schools.

References

Author 1 et al. 2017a

Author 1 et al. 2017b

Author 1 et al. 2019a

Author 1 et al. 2019b

Author 6 et al. forthcoming

Author 7 et al. 2018.

Becker H. 1963. *Outsiders*. New York: Free Press.

Belfi, B., Goos, M., De Fraine, B., & Van Damme, J. 2012. "The effect of class composition by gender and ability on secondary school students' school well-being and academic self-concept: A literature review". *Educational Research Review* 7: 62–74.

Boaler J. 1997. "Setting, Social Class and Survival of the Quickest". *British Educational Research Journal* 23: 575-595.

Boaler J., Wiliam D., Brown M. (2000). "Students' Experiences of Ability Grouping—disaffection, polarisation and the construction of failure". *British Educational Research Journal* 26: 631–648.

Chmielewski, A. K., et al. 2013. "Tracking Effects Depend on Tracking Type: An International Comparison of Students' Mathematics Self-Concept." *American Educational Research Journal* 50(5): 925-957.

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences* (2nd ed.). New York: Academic Press.

Education Endowment Foundation. 2018. *Education Endowment Foundation Toolkit*, London: EEF. https://educationendowmentfoundation.org.uk/evidence-summaries/teaching-learning-toolkit/setting-or-streaming/

Education Endowment Foundation. (2020). *Statement on statistical significance and uncertainty of impact estimates for EEF evaluations*. London: EEF. Downloaded from: educationendowmentfoundation.org.uk 10th March 2020.

Fraser, N. 1997. *Justice interruptus: critical reflections on the "postsocialist" condition*. New York: Routledge.

Gamoran, A. 1992. "Synthesis of research: Is ability grouping equitable?" *Educational Leadership* 50 (2): 11-17.

Hollway, W. & Jefferson, T. 2013. (Second edition) *Doing Qualitative Research Differently: A Psychosocial Approach.* London: Sage.

Ireson J., Hallam S., Hurley C. 2005. "What are the effects of ability grouping on GCSE attainment?" *British Educational Research Journal* 31: 443-458.

Ireson J., Hallam S. 2009. "Academic self-concepts in adolescence: Relations with achievement and ability grouping in schools". *Learning and Instruction* 19: 201-213.

Jackson B. 1964. *Streaming: An Education system in miniature.* London: Routledge & Kegan Paul.

Jussim, L., Eccles, J. & Madon, S.J. 1996. “Social perception, social stereotypes, and teacher expectations: Accuracy and the quest for the powerful self-fulfilling prophecy”. *Advances in Experimental Social Psychology* 28: 281-288.

Jussim, L. & Harber, K. 2005. “Teacher Expectations and Self-Fulfilling Prophecies: Knowns and Unknowns, Resolved and Unresolved Controversies”. *Personality and Social Psychology Review* 9 (2): 131-155.

Kulik C.-L. C., Kulik J.A. 1982. “Effects of Ability Grouping on Secondary School Students: A Meta-Analysis of Evaluation Findings”. *American Educational Research Journal* 19: 415-428.

Kutnick P., Sebba J., Blatchford P., Galton M., Thorpe J., with MacIntyre H., Berdondini L. 2005. *The Effects of Pupil Grouping: Literature Review. Research Report 688*: London: DfES.

Lemert E. M. 1951. *Social Pathology.* New York: McGraw-Hill.

Liem G. A. D., McInerney D., Leung A.S. 2015. “Academic Self-Concepts in Ability Streams: Considering Domain Specificity and Same-Stream Peers”. *The Journal of Experimental Education* 83: 83-109.

Liu W. C., Wang CK., Parkins, E.J. 2005. “A longitudinal study of students’ academic self-concept in a streamed setting: The Singapore context”. *British Journal of Educational Psychology*, 75: 567-586.

Macleod, S., Sharp, C., Bernardinelli, D., Skipp, A., & Higgins, S. (2015). *Supporting the attainment of disadvantaged pupils: articulating success and good practice*. London: Department for Education.

Marks, R. 2016. *Ability Grouping in Primary Schools*. London: Critical Publishing.

Mazenod, A., Francis, B., Archer, L., Hodgen, J., Taylor, B., Tereshchenko, T. & Pepper, D. 2018). “Nurturing learning or encouraging dependency? Teacher constructions of students in lower attainment groups in English secondary schools”. *Cambridge Journal of Education* 47 (1).

Merton R.K. 1948. “The Self Fulfilling Prophecy”. *Antioch Review*, 8: 195.

Madon, S.J., Jussim, L., & Eccles, J. 1997. “In search of the powerful self-fulfilling prophecy”. *Journal of Personality and Social Psychology* 72: 791-809.

Marsh H. W. 2008. The Big-fish-little-pond-effect stands up to critical scrutiny: Implications for theory, methodology and further research, Educational Psychology Review, 20, 319-350.

Marsh, H., & Parker, J.W. (1984). "Determinants of student self-concept: Is it better to be a relatively large fish in a small pond even if you don't learn to swim as well?". *Journal of Personality and Social Psychology* 47 (1): 213–231. doi:10.1037/0022-3514.47.1.213

McGillicuddy, D. & D. Devine 2018. ‘“Turned off” or “ready to fly” – Ability grouping as an act of symbolic violence in primary school’. *Teaching and Teacher Education* 70: 88-99.

Martin, M. O., & Mullis, I. V. S. (Eds.). 2012. *Methods and procedures in TIMSS and PIRLS 2011*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Muijs, D., Dunne M. 2010. “Setting by ability–or is it? A quantitative study of determinants of set placement in English secondary schools”. *Educational Research* 52: 391-407.

Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2016). TIMSS 2015 International Results in Mathematics. Retrieved from Boston College, TIMSS & PIRLS International Study Center website:http://timssandpirls.bc.edu/timss2015/international-results/

Oakes, J. 1985. *How schools structure inequality*. New Haven: Yale University Press.

Oakes, J. 1986. "Tracking, inequality, and the rhetoric of reform: Why schools don't change". *Journal of Education* 168: 60-80.

Oxford Reference (2019)
https://www.oxfordreference.com/view/10.1093/oi/authority.20110803100453192

Raudenbush, S. W. 1984. "Magnitude of teacher expectancy effects on pupil IQ as a function of the credibility of expectancy inductions: A synthesis of findings from 18 experiments". *Journal of Educational Psychology* 76: 85-97.

Rosenthal, R., & Jacobson, L. 1968. *Pygmalion in the classroom: Teacher expectations and student intellectual development*. New York: Holt.

Roy, P., Styles, B., Walker, M., Morrison, J., Nelson, J., & Kettlewell, K. (2018). *Best Practice in Grouping Students Intervention A: Best Practice in Setting Evaluation report and executive summary*. London: Education Endowment Foundation.

Slavin R. 1990. "Achievement effects of ability grouping in secondary schools: a best evidence synthesis". *Review of Educational Research* 60: 471-499.

Smith, A., Jussim, L., Eccles, J., Van Noy, M., Madon, S. J., & Palumbo, P. 1998. "Self-fulfilling prophecies, perceptual biases, and accuracy at the individual and group level". *Journal of Experimental Social Psychology* 34: 530-561.

Steenbergen-Hu S., Makel M.C., Olszewski-Kubilius P. 2016. "What One Hundred Years of Research Says About the Effects of Ability Grouping and Acceleration on K–12 Students' Academic Achievement: Findings of Two Second-Order Meta-Analyses". *Review of Educational Research* 86: 849–899.

Strand, S. (2012). The White British–Black Caribbean achievement gap: tests, tiers and teacher expectations. *British Educational Research Journal*, 38(1), 75-101. doi:10.1080/01411926.2010.526702

Styles & Roy. 2017. *Protocol for the Evaluation of Best Practice in Grouping Students Intervention A – Best Practice in Setting.* London: Education Endowment Foundation.

Van Houtte M., Demanet J., Stevens P. 2012. "Self-esteem of academic and vocational students: Does within-school tracking sharpen the difference?" *Acta Sociologica* 55: 73-89.

Table 1: Sample Characteristics for the Mathematics and English sample

		Mathematics Trial		**English Trial**	
		No.	Valid %*	No.	Valid %*
		Total		Total	
Gender	Boy	3221	52.23	1553	53.70
	Girl	2946	47.77	1339	46.30
Household Socio-Economic Background	Higher	2852	46.25	1287	44.50
	Intermediate	1909	30.96	915	31.64
	Lower	811	13.15	397	13.73
	Missing	595	9.65	293	10.13
Ever Eligible For Free School Meals	No	4699	76.20	2142	74.07
	Yes	1370	22.22	718	24.83
	Missing	98	1.59	32	1.11
Ethnicity	White	4837	78.43	2374	82.09
	Black African	148	2.40	53	1.83
	Black Caribbean	43	0.70	9	0.31
	Black Mixed	261	4.23	115	3.98
	Pakistani	165	2.68	71	2.46
	Bangladeshi	62	1.01	18	0.62
	Indian	74	1.20	27	0.93
	Chinese	20	0.32	7	0.24
	Asian Mixed	110	1.78	44	1.52
	Other	419	6.79	161	5.57
	Missing	28	0.45	13	0.45
Sets	Top	2057	33.35	916	31.67
	Middle	3091	50.12	1444	49.93
	Bottom	753	12.21	311	10.75
	Missing	266	4.31	221	7.64
Total		6167	100	2892	100

**Column percentages may not sum to 100.0% due to rounding.*

Table 2: Scales for Self-Confidence in English/Mathematics and General Self-Confidence

Scale and Items	Reliability and Summary Statistics of Scales*
Self-Confidence in English/Mathematics: • "Work in English/maths is easy for me" • "I am not very good at English/maths" • "English/maths is one of my best subjects" • "I hate English/maths" • "I do well at English/maths" • "I get good marks in English/maths" • "I learn things quickly in English/maths lessons"	Maths Scale: Alpha = 0.88 Mean = 27.22 (SD = 5.70) English Scale: Alpha = 0.86 Mean = 26.54 (SD = 5.88)
General Self-Confidence in Learning: • "I learn quickly" • "Most things I do, I do well" • "I am proud of my achievements at school" • "I can do things as well as most people" • "If I really try I can do almost anything I want to" • "I am confident in my abilities" • "I am generally high achieving in my studies"	Alpha = 0.84 Mean = 25.22 (SD = 3.94)

*Maximum in scales is 35

Table 3: Overall mean self-confidence levels at the beginning of Year 7 (pre-test) and the end of Year 8 (post-test)

	N	Year 7 (Pre-test)		Year 8 (Post-test)	
		Mean	SD	Mean	SD
Mathematics self-confidence	5119	3.94	0.85	3.63	0.94
English self-confidence	2363	3.82	0.81	3.52	0.86
General self-confidence (mathematics sets)	5268	4.25	0.59	3.93	0.78
General self-confidence (English sets)	2377	4.21	0.61	3.89	0.78

Table 4: Multilevel models used to compare post-test mean scores in self-confidence by set level, controlling for pre-test score, number of sets in school, family occupation, ethnicity and gender.

Independent variables in the Model	Dependent variable = Self-Confidence in Maths or English		Dependent Variable = General Self-Confidence	
	Maths (Model A†)	English (Model B†)	Maths (Model C†)	English (Model D†)
Number of Observations	5119	2363	5268	2377
Pre-test self-confidence (standardised) score	.457 (.013)	.344 (.017)	.375 (.011)	.355 (.015)

Set Allocation				
Top	.199 ***	.134 **	.167 ***	.116 ***
	(.033)	(.048)	(.024)	(.035)
Middle (Ref Cat)				
Bottom	-.225 ***	-.061	-.1521 ***	-.090
	(.044)	(.067)	(.034)	(.052)
No. of Sets in School	.059	-.044	.050	-.028
	(.034)	(.049)	(.023)	(.016)
Family Occupation				
Higher	.027	.025	.014	.019
	(.016)	(.023)	(.014)	(.021)
Intermediate	.0127	-.001	-.009	-.006
	(.016)	(.023)	(.014)	(.020)
Lower (Ref Cat)				
Ethnicity				
White	-.006	-.025	.002	-.018
	(.016)	(.024)	(.014)	(.022)
Asian	.026	.004	.001	-.018
	(.015)	(.023)	(.013)	(.021)
Black	.022	.034	.022	.025
	(.016)	(.026)	(.014)	(.024)
Other or Mixed (Ref Cat)				
Gender	.			
Male	080	-.058	.029	.019
	(.011)	(.016)	(.010)	(.014)
Female (Ref Cat)				
Constant	3.539	3.497	3.858	3.976
	(.029)	(.046)	(.020)	(.078)
Variance				
School Level	.138	.172	.090	.102
	(.023)	(.034)	(.017)	(.027)
Set Level	.168	.180	.082	.078
	(.019)	(.024)	(.019)	(.032)
Student Level	.756	.728	.667	.672
	(.008)	(.011)	(.007)	(.010)
-2LL	-5947.127	-2668.460	-5400.188	-2455.816

** $p<0.01$; *** $p\leq0.001$

† Estimated coefficients with associated standard errors in parentheses.

Table 5: Multilevel models used to compare post-test mean scores in self-confidence by set level, controlling for pre-test score, number of sets in school, family occupation, ethnicity, gender and prior attainment using alternative Key Stage Scores.

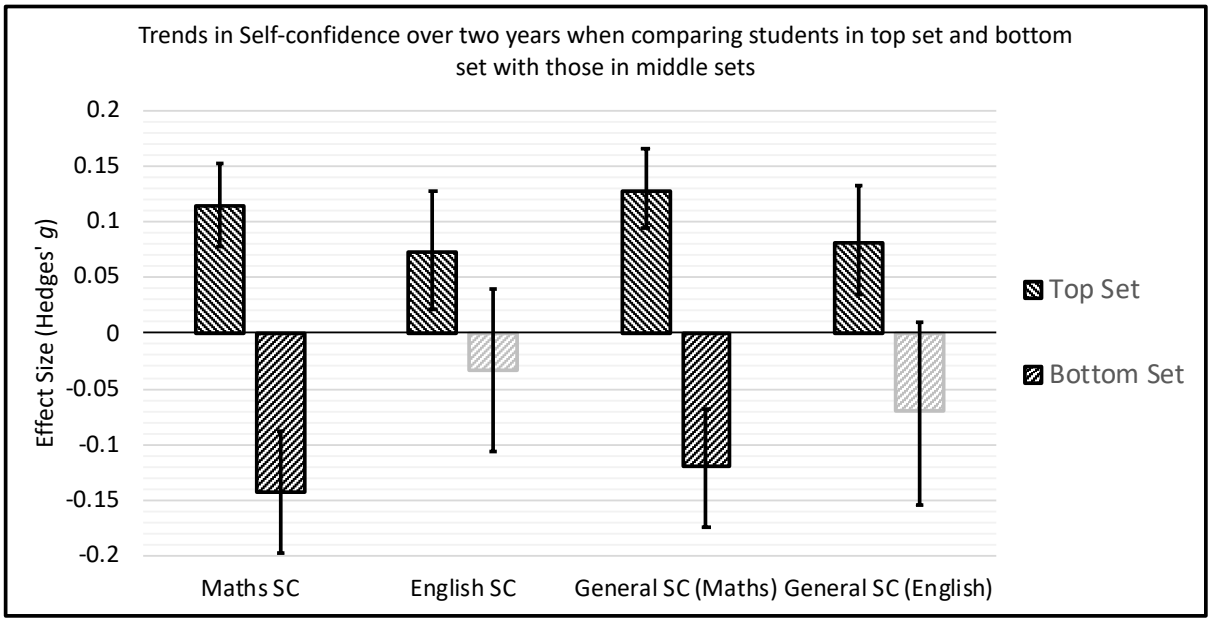
Independent variables in the Model	Dependent variable = Self-Confidence in Maths or English		Dependent Variable = General Self-Confidence	
	Maths (Model A†)	English (Model B†)	Maths (Model C†)	English (Model D†)
Number of Observations	5078	2348	5229	2362
Pre-test self-confidence (standardised) score	.424 (.013)	.339 (.017)	.360 (.011)	.345 (.016)
Set Allocation				
Top	.023 (.035)	.108 * (.051)	.073 * (.026)	.046 (.037)
Middle (Ref Cat)				
Bottom	-.081 (.044)	-.014 (.073)	-.070 * (.034)	-.006 (.054)
No. of Sets in School	.041 (.031)	-.040 (.050)	.028 (.016)	-.028 (.015)
Family Occupation				
Higher	.013 (.016)	.023 (.023)	.002 (.014)	.009 (.021)
Intermediate	.002 (.016)	-.002 (.023)	-.015 (.014)	-.012 (.021)
Lower (Ref Cat)				
Ethnicity				
White	-.012 (.016)	-.024 (.024)	-.003 (.014)	-.017 (.022)
Asian	.026 (.015)	.005 (.023)	.001 (.013)	-.018 (.021)
Black	.027 (.016)	.035 (.026)	.023 (.014)	.030 (.024)
Other or Mixed (Ref Cat)				
Gender				
Male	.073 (.011)	-.055 (.016)	.032 (.010)	.016 (.014)

Female (Ref Cat)				
Key Stage 2 score	.188	.049	.114	.102
	(.018)	(.028)	(.015)	(.022)
Constant	3.551	3.492	3.729	3.968
	(.027)	(.046)	(.070)	(.073)
Variance				
School Level	.125	.176	.088	.093
	(.021)	(.034)	(.017)	(.027)
Set Level	.141	.178	.065	.065
	(.019)	(.024)	(.022)	(.036)
Student Level	.750	.728	.665	.671
	(.008)	(.011)	(.007)	(.010)
-2LL	-5842.165	-2653.904	-5335.910	-2431.935

* $p<0.05$

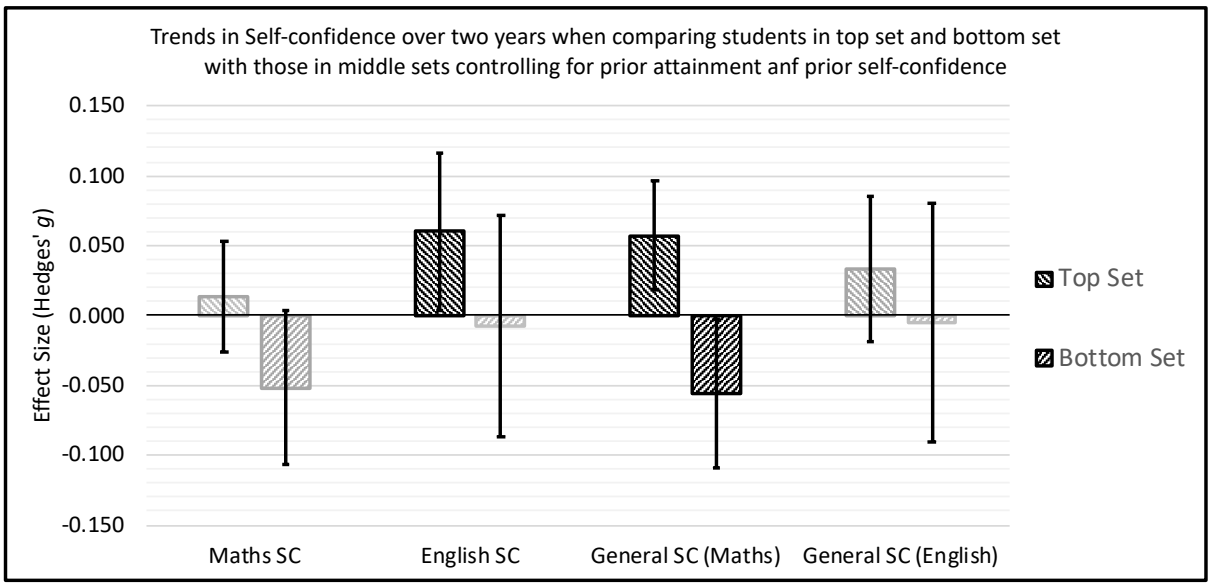
† Estimated coefficients with associated standard errors in parentheses.

Figure 1: Overview of the effect sizes and 95% confidence intervals for the three level models comparing post-test mean gains in self-confidence by set level, controlling for number of sets in school, family occupation, ethnicity and gender.



Statistically significant results presented in **bold**.

Figure 2: Overview of the effect sizes and 95% confidence intervals for the three level models to compare post-test mean gains in self-confidence by set level, controlling for pre-test self-confidence, number of sets in school, family occupation, ethnicity, gender and prior attainment.



Statistically significant results presented in **bold**.

[i] Or 'ability', as it is commonly conceived.

[ii] See Author 1 et al (2019a) for elaborated discussion.

[iii] We do not ascribe to a view of 'ability' as fixed, hence our adoption of inverted commas.

[iv] Setting is a form of attainment grouping whereby pupils are grouped together by prior attainment in the study of particular subjects. It is sometimes referred to as 'tracking by subject' in the US. It is more flexible than tracking (streaming) wherein students are banded into the same 'ability' groups for most or all subjects: in the case of setting, a pupil might be in a high set for one subject and a low set for another. However, often in practice the approaches are blurred – for example setting can take place in addition to streaming, and/or there can be clustering of set applications across a number of subjects. Setting is prevalent in English secondary (high school) education, and increasingly in primary schooling (Author 1 et al, 2019a; Hallam & Ireson 2007).

[v] See endnote iv

[vi] NFER are commissioned by the Education Endowment Foundation to perform the post-testing and to evaluate our key intervention outcomes. See https://educationendowmentfoundation.org.uk/our-work/projects/best-practice-in-grouping-students/ for information on the wider study, and the published RCT protocols.

[vii] Those schools in the intervention group had been instructed not to additionally apply other forms of tracking; but some of the schools from the control group applied streaming as well as setting

[viii] It is worth noting here that there are more pupils in the top sets than bottom sets. This is because schools frequently have larger top set groups, for example two parallel top sets (and middle set tiers) and a single – sometimes deliberately small – bottom group (Dunne et al, 2007; Author et al, 2019).

[ix] Key Stage 2 assessments are completed by pupils in England in Year 6 (age 10-11), the final year of primary school education.