Test scores, subjective assessment and stereotyping of ethnic minorities

Simon Burgess
University of Bristol, CMPO

Ellen Greaves
CMPO

July 2009

Abstract

We assess whether ethnic minority pupils are subject to low teacher expectations. We exploit the English testing system of "quasi-blind" externally marked tests and "non-blind" internal assessment to compare differences in these assessment methods between White and ethnic minority pupils. We find evidence that some ethnic groups are systematically "under-assessed" relative to their White peers, while some are "over-assessed". We propose a stereotype model in which a teacher's local experience of an ethnic group affects assessment of current pupils; this is supported by the data.

Keywords: Subjective assessment, stereotypes, education, test score gaps, ethnic minorities
JEL Codes:

Corresponding author:
Simon Burgess
simon.burgess@bristol.ac.uk

Many thanks to the DCSF who provided the PLASC/NPD data for this paper, and to the ESRC for funding through CMPO. Thanks also for comments from seminar audiences at Carnegie Mellon University, IZA, CMPO, SoLE and ESPE, plus comments from Steve Gibbons, Carol Propper, Helen Simpson, Sarah Smith, Liz Washbrook and Deborah Wilson.

1. Introduction

There are many stories of how having a teacher who "believed in me" changed a pupil's life for the better. There are also stories of teachers having low expectations particular groups of pupils and of what they can achieve. Such low expectations may lead to pupils reducing their effort at school, and therefore to achieving lower levels of human capital. This 'self-fulfilling prophecy' was first discussed by Arrow (1972) and more recently by (Mechtenberg, 2006) in the context of a cheap talk game between teachers and pupils. More generally, (Hoff & Pandey, 2006) argue that the propagation of negative stereotypes is part of the broad pattern of persistent inequalities. (Ferguson, 2003) review of the literature on the black-white attainment gap concludes that "teachers' perceptions, expectations, and behaviours probably do help sustain, and perhaps even expand the black-white test score gap"[1].

In this paper, we test whether there are systematic differences between objective and subjective assessment measures across ethnic minority and white pupils in England, and, having found such differences, examine the form they take. This study therefore contributes to the debate around the educational performance of some ethnic groups (particularly pupils of Black Caribbean, Pakistani or Bangladeshi ethnicity), and the implications of this for their future life chances.

The analysis is based on large-scale observational data. We have access to five annual censuses of all state school pupils in England, providing a large sample for most minorities. Our empirical strategy is to first test for the existence of systematic differences between objective and subjective measures of a pupil's ability level. We do this in a very flexible way, allowing for heterogeneous responses in a number of dimensions. Secondly, we interpret the pattern of differences across ethnic groups, subjects, schools and pupils to test different theories of the source of the

[1] A number of studies have addressed whether low expectations on the behalf of teachers are detrimental to student progress, see a recent review in (Jussim & Harber, 2005). The expectancy effect, or "living down to expectations", has appeared frequently in the education literature, for example in (Weinstein, 2002), though Rosenthal and Jacobson's early (1968) study has been strongly criticised (see Snow, 1995, and Raudenbush, 1984). The stereotype threat effect suggests that students who fear that others will assess them through the lens of a negative stereotype perform badly in test situations (Thomas S Dee, 2009; Steele & Aronson, 1995). There are institutional possibilities too: students may be unjustifiably placed in lower ability classes, and thereby be entered into lower tier exam papers (Strand, 2007).

objective/subjective gap. The English National Curriculum is built around measuring concrete, well-defined levels of achievement, assessed separately for English, maths and science. At age 11 the level achieved is assessed in two ways: by a written exam, nationally set and remotely marked, and by assessment by the pupil's own teacher. We characterise these methods as objective and "quasi-blind" (the test) and subjective and "non-blind" (the teacher assessment), and it is a comparison of these two measures of the same level that we exploit[2]. We find statistically and quantitatively significant differences in the test/assessment differences across ethnic groups. The census nature of the dataset means that we observe all the pupils in a school, allowing us to control for school fixed effects. The differences we observe remain, even working only off this within-school variation.

Our paper adds to a small literature comparing "blind" and "non-blind" assessment methods in schools (Lavy, 2004), discrimination in hiring (Goldin & Rouse, 2000), or discrimination by institution in refereeing (Blank, 1991). Using data from matriculation exams in Israel, Lavy (2004) finds a negative bias in teachers' assessment for male students. The negative effect of being male on "non-blind" tests as opposed to objective "blind" tests occurs at all points in the ability distribution. Lavy suggests that the bias is not due to statistical discrimination, as the bias is present even in sub samples where males outperform females, but instead is related to teachers' own characteristics and behaviour. The finding of male bias is corroborated in evidence from the Swedish education system, where females are more generously rewarded in teacher assessed "School Leaving Certificates" than test results (Lindahl, 2007). Using the same data source as we do, (Gibbons & Chevalier, 2008) interpret differences between test scores and teacher assessments as indicative of assessment bias or uncertainty in teacher assessments, focusing in particular on discrepancies by socioeconomic status. (Hanna & Linden, 2009) run a field experiment in India, randomly assigning identifying cover pages to exam scripts and comparing the marks with and without these cover sheets. They find evidence of significant discrimination with exams assigned to lower caste children being given grades between 0.03 and 0.09 standard deviations below those assigned to higher caste children. All these results are consistent with earlier research with smaller sample sizes (Reeves, Boyle, & Christie, 2001; Thomas, Madaus, Raczek, & Smees, 1998). Qualitative work adds

[2] The test is marked outside the school by a marker the pupil has never met. However, the script does contain the pupil's name, so we describe this as quasi-blind. We discuss this further below.

to this picture: in his study of a UK multi-ethnic school (Gillborn, 1990) argues that "teacher-student interaction was fraught with conflict and suspicion" for Black Caribbean pupils.

The second component of our empirical strategy is to analyse the patterns in the test/assessment differences to test different theories of the source of the subjective assessment. The advantage of the richness of our administrative data is that it offers variation across a number of margins. First and most obviously, it covers different ethnic groups, some of which outperform white students and some of which do less well. Second, we have exactly equivalent data across three subjects. Third, we have variation across 16557 schools and 4 years. The central point about the estimated test/assessment differences is that they are not uniform across these margins. There are variations across different ethnic groups, with the gap being negative for some ethnicities and positive for others. There are also variations across subject within ethnic group. Finally, there are variations across schools within ethnic group, within subject. We show that the pattern of the test/assessment differences fits a stereotype model[3] reasonably well, and that four other possible explanations are rejected. We show that the past performance of a specific ethnic group in a specific school matters for the current teacher assessment of pupils of that group in that school. We also show that the stereotype factor is more important in schools where that group is relatively scarce.

Fryer and Jackson's (2008) model of category formation and decision-making is useful. As motivation, they quote the social psychologist Allport: "the human mind must think with the aid of categories". They model the optimal formation of such categories, particularly focussing on ethnicity, and show that optimal decision-making involves the formation of a "prototype" for each category based on some statistic of the members of the category. The properties of the prototype are used as the basis for decisions. This is related to the idea of statistical discrimination (Arrow, 1973; Phelps, 1972). Analysis of categorical ways of processing information and making decisions has a long history (Fiske, 1998 provides a review). In social psychology, the exemplar-based model of social judgement argues that individuals are categorised into groups and stored in memory as "exemplars", or representations of their group (Smith

[3] Fryer and Jackson make a distinction between prototypes, the model held by the decision-maker, and stereotypes, a model that the decision-maker believes is widely held. We use the more widely-used term stereotype here, as we cannot know the exact mechanism through which teachers form their views.

& Zarate, 1992). In the context of this paper, this approach suggests that a teacher will categorise students and create prototypes or exemplars to make conscious or unconscious judgements about future students of the same group. (Chang & Demyan, 2007) show that teachers hold these exemplars or stereotypes, and show that they differ across ethnic groups. In related work, (T. S. Dee, 2005) shows that assignment to a demographically similar teacher influences the teachers' subjective evaluations of student behaviour and performance.

Section 2 sets out the structure we use to interpret and identify the results and section 3 then explains the detail of the dataset. Section 4 reports the results and finally section 5 concludes.

2. Measuring Students' Ability

The National Curriculum in England sets standards of achievement in each subject for pupils aged five to 14[4]. These standards are defined by a set of Key Stage levels ranging from 1 to 8[5]. A pupil's ability in a subject is therefore defined by the Key Stage level they attain. These Key Stage levels are absolute, concrete measures defining a set of skills that the child has mastered, not relative marks. A rich set of descriptors are provided for the levels[6]; some examples are given in Appendix Figures 1, 2 and 3. The level achieved depends on the human capital of the pupil. The formation of human capital is of course very widely studied, but is not the key focus of this paper. The Key Stage level a child has reached is assessed in two ways, and these are at the centre of our analysis.

First, a level is assigned in English, mathematics and science from the nationally set and remotely marked Key Stage tests. They are seen as an objective "snapshot"

[4] The National Curriculum in England has been organised around four compulsory Key Stages, each rounded off by exams: Key Stage 1 with tests at age 7, Key Stage 2 tested at age 11, Key Stage 3 tested at age 14 and Key Stage 4 (also known as GCSEs) tested at age 16, the end of compulsory schooling. As of 2009, testing at age 14 was abolished.

[5] From Department for Children Schools and Families (DCSF) at http://www.dcsf.gov.uk/performancetables/primary_07/p5.shtml
Most seven-year-olds are expected to achieve level 2, most 11-year-olds are expected to achieve level 4, and most 14-year-olds are expected to achieve level 5 or 6.

[6] See:
http://www.dfes.gov.uk/rsgateway/DB/SFR/s000640/SFR09_2006.pdf,
http://www.ncaction.org.uk/subjects/maths/levels.htm, and
http://www.ncaction.org.uk/subjects/maths/judgemnt.htm

measure of a pupil's ability. Key Stage scripts still carry the pupil's name, and to a degree names can identify different ethnic groups. This may influence the level awarded by markers (see Brennan, 2008), although (Baird, 1998) finds no evidence of this in the case of gender. We characterise these tests as quasi-blind, in that the marker knows nothing about the pupil other than their name.

Second, teachers make a personal assessment of each child's level in the same three subjects. The assessment is based on the teacher's interaction with the child over the year, the child's performance in in-school tests, and a set of "probing questions" provided by the Department for Children, Schools and Families (DCSF) specifically to help assess each pupil's level. The teacher must provide some evidence of their pupils' work to justify the TA awarded, although the role of class interaction and observation is acknowledged. TA is taken seriously by schools, and the emphasis on rigorous TA has recently increased. The QCA (Qualifications and Curriculum Authority) now provides materials online to support teachers in "aligning their judgements systematically with national standards"[7].

In primary schools in England, almost all pupils have one teacher who teaches them all the subjects[8]. Any differences between the three subjects are therefore not explained by different teacher characteristics. Our analysis focuses on any systematic differences between these two measurement approaches.

a. Measurement of ability level

Denote the true underlying and unobserved level achieved by pupil i as L_i. There are two measurement functions, *KS* denoting the level returned by the Key Stage test, and *TA* being the level given in the teacher assessment. Pupil i has characteristics X_i, including their ethnicity, and attends school $s(i)$. We assume that these measurement functions work as follows:

$$KS_i = f\left(L_i, X_i, \lambda_{s(i)}\right) \qquad (1)$$

KS should reflect the true level, *L* and ideally *KS* should equal *L*, but there will be testing noise: some pupils may have a bad day and perform below potential, others might get lucky and "over-perform". In addition to random influences, we allow

[7] For example see: http://www.standards.dfes.gov.uk/secondary/keystage3/respub/englishpubs/ass_eng/optional_tasks/
[8] See http://careersadvice.direct.gov.uk/helpwithyourcareer/jobprofiles/profiles/profile820/

systematic factors through characteristics *X* to influence the measurement of *L* by *KS*. The possible bases for this are discussed below in section 5. Finally, to control for school policies regarding interpretation of the National Curriculum, "teaching to the test" and test conditions in the school, we include school effects, $\lambda_{s(i)}$.

We assume that the teacher assessment measurement function is as follows, where pupil *i* is taught by teacher *j(i)*:

$$TA_i = g\left(L_i, X_i, A_{j(i)}, \theta_{s(i)}\right) \qquad (2)$$

Again, discussion of why *X* may matter in changing the relationship between L and TA is postponed to section 5. Teacher attitudes, *A*, may influence the mapping from *L* to *TA*, and this possibility is the key difference between (1) and (2). The term $\theta_{s(i)}$ captures any potential school effects. We think of teacher attitudes as having three main components:

$$A_{j(i)} = \phi_j + \varphi_{j(i)}.X_i + \eta_{s(i)} + \varepsilon_{ijs} \qquad (3)$$

The first is a common effect, independent of pupil type. For example, some teachers may be naturally pessimistic or negative and tend to under-grade all their pupils. Second, some teachers may have differential attitudes between observable pupil types. This is the key focus here, the idea being that an interaction between *X* and *j* affects the mapping from *L* to *TA*. Third, schools may either directly influence teachers' attitudes, or may select teachers with particular attitudes through their hiring policies. Finally, we assume an element of randomness in attitudes.

The analysis in this paper is based on the difference between the two measurements:

$$d_i \equiv TA_i - KS_i = \Delta\left(L_i, X_i, A_{j(i)}, \mu_{s(i)}\right) \qquad (4)$$

where $\mu_{s(i)}$ combines $\theta_{s(i)}$ and $\lambda_{s(i)}$. We have to take a number of steps to make this operational. First, we assume that (1) and (2) and so (4) are linear in L_i. Second, while our dataset contains rich information on pupils and schools, we do not know the assignment of pupils to specific teachers. The $A_{j(i)}$ terms are therefore unobservable and are substituted out through (3):

$$d_i = \alpha L_i + (\gamma_0 + \gamma_{s(i)})X_i + \mu_{s(i)} + \xi_i \qquad (5)$$

Where $\gamma_{s(i)} = E(\varphi_{j(i)} \mid j \in s)$ is the mean value of teacher attitudes to characteristic *X* among teachers in school *s*. By simply taking the mean of $\varphi_{j(i)}$ we are assuming that the assignment of teachers to pupils within school is independent of $A_{j(i)}$ and X_i. This may not be the case: it may be that particular teachers are typically assigned to particular groups of pupils. We check for this potential source of bias in the empirical work below. The school effect $\mu_{s(i)}$ combines $\lambda_{s(i)}$, $\theta_{s(i)}$, $\eta_{s(i)}$ and $E(\phi_{j(i)} \mid j \in s)$. Finally, since L_i is unobserved, we invert *f(.)* to replace L_i by KS_i, X_i and λ [9]. This gives us our empirical model, and in the following analysis we examine the probability:

$$pr(TA_i < KS_i) = pr(d_i < 0) = \alpha.KS_i + \boldsymbol{\beta}.\mathbf{X}_i + \mu_{s(i)} + \xi_i \qquad (6)$$

We are particularly interested in the role of the ethnicity identifiers in *X*, and in any variation in β across schools and subjects. This framework gives us a basis for interpreting cases where *TA* and *KS* differ. In cases where *TA* < *KS*, either the pupil over-performs in the test or the teacher under-assesses the pupil's level, and *vice versa* for cases where $TA > KS$. A key issue for the interpretation of our results is whether any effect of characteristics *X* on the gap *d* arises through its impact on differential test-taking ability (conditional on true ability) in (1) or through differential teacher attitudes (conditional on true ability) in (2).

One reason for concern over under-assessment of some groups of students is the potential impact on their academic performance. It is straightforward to incorporate this into an extension of the model. Suppose pupils' achievements depend on both their underlying level of ability (*L* as before) and the effort they exert, denoted ε. So we rewrite (1) and (2) as $KS_i = f(L_i, \varepsilon_i, X_i, \lambda_{s(i)})$ and $TA_i = g(L_i, \varepsilon_i, X_i, A_{j(i)}, \theta_{s(i)})$. We assume that effort depends on ability, characteristics and also the teacher's attitude to the pupil, $\varepsilon_i = h(L_i, X_i, A_{j(i)})$. Substituting for ε and *A*, we reach the counterpart to (5):

$$d_i = k_1 L_i + \left(k_2 + \gamma_{s(i)}(1 - \eta(\psi - \sigma))\right)X_i + \mu_{s(i)} + \xi_i \qquad (7)$$

[9] This seems the natural way to parameterise the relationship, rather than having on TA on the right hand side. Statistically, this exploits the greater variation in KS than TA that we show below, and intuitively KS is a more objective measure than TA. Possibly for these reasons, the results are less clear cut conditioning on the TA score.

where k_1 and k_2 are constants, η is the effect of teacher attitude on pupil effort, and ψ and σ are the effect of effort on KS and TA respectively. Two points follow from this. First, the key coefficient is still $\gamma_{s(i)}$, and if this is zero – if there are no differential teacher attitudes – then the channel of impact on *d* via effort is also zero. Second, if pupil effort has roughly equal effects on TA and KS, then most of the quantitative impact of teacher attitudes is the direct one through $\gamma_{s(i)}$. In this paper we focus on establishing the existence and cause of differential teacher assessments; we leave to our future work the study of the impact of this on outcomes via student effort.

3. Data

a. PLASC/NPD

The National Pupil Database (NPD) combines information on pupil and school characteristics from the Pupil Level Annual School Census (PLASC) dataset with information on pupil attainment, as well as incorporating reference data on schools and Local Authorities (LAs). The dataset relates to England and covers state schools[10], which educate around 94% of all pupils in England.

It is a statutory requirement for all state schools in England to return this data and in consequence the NPD is highly accurate and complete. This removes problems of self-selection and attrition common in many datasets, although pupils may have missing results for other reasons such as absence on the day of the test. Whereas smaller datasets contain insufficient samples of ethnic minorities for robust estimation, the large scale of the NPD allows analysis for all but the smallest minority groups in England.

PLASC contains the following pupil level information: within–year age, gender and ethnicity, eligibility for free school meals (FSM), whether the student has English as an additional language (EAL), and whether the student has Special Educational Needs (SEN). Very specific ethnicity codes are now available in PLASC, but in some

[10] Independent schools will be present if they take KS2 and KS3 exams. As this is not compulsory for independent schools, and taking the tests is unlikely to be random, we restrict our sample to state schools.

analyses we focus on the relatively larger groups: White, Black Caribbean, Black African, Indian, Pakistani, Bangladeshi, and Chinese ethnicity pupils.

A pupil's socioeconomic status is proxied by whether or not s/he is eligible for FSM. This in turn derives from eligibility for certain kinds of welfare benefits, principally Income Support and Income-based Job Seekers Allowance. This simple dichotomous measure does not capture all aspects of socioeconomic status, and identifies only those at the bottom of the income distribution. Our measure is likely to be a good measure of true eligibility, but this is an imprecise measure of poverty (see (Hobbs & Vignoles, 2007), for some evidence on this). School data includes: type of school[11], whether the school is selective, single or mixed sex, and the location of school. We construct measures of school composition from the pupil data.

We provide some descriptive statistics on the key variables in Appendix Table 1. These show that schools over this period remained overwhelmingly white, with all ethnic minorities together making up 13.4% of our sample. Numerically the most important are Black Caribbean pupils (1.5%), Black African (1.6%), Indian (2.2%), Pakistani (2.6%), Bangladeshi (1%), mixed White-Black Caribbean (0.8%) and Chinese (0.3%). Around 17% of pupils are eligible for FSM, and 9.5% have English as an additional language.

b. Sample definition

Our analysis focuses on the relationship between KS and TA at age 11. We treat the data as a series of repeated cross-sections over the years 2002-2005, so our dataset is longitudinal across schools not pupils[12]. After dropping students in certain schools[13] and with missing observations on either KS or TA, we end up with a sample of 2,255,383 pupils over four years, each taking English, maths and science[14]. These pupils are in 16,557 primary schools, where the school cohort has a mean size of 54.8. While this is a huge sample, it does not contain large numbers of minority students, who in total make up 13.4% of the final sample. In robustness checks, we restrict the

[11] Type of School refers to whether the school is a faith school, academy or other. For definitions see: http://www.direct.gov.uk/en/Parents/Schoolslearninganddevelopment/ChoosingASchool/DG_4016312

[12] So the fixed point is a school*KS2 subject (English, maths, science), and we trace four generations of pupils as they pass through that point.

[13] We keep students in community and community special schools, academies, voluntary aided, voluntary controlled, city technology colleges, foundation and foundation special schools.

[14] This involves dropping 151,379 observations, or 6.3% of the original sample.

range of schools in our analysis and consider only those with more than 5 pupils of a particular ethnic group. This yields datasets that differ in size, but are a great deal smaller than the 2.26m pupils noted above, given that 67% of school-cohorts in our final sample comprise at least 90% white students.

4. Results

a. Teacher assessments

The distribution of scores is more compressed in TA than KS, with the variance in KS scores around 20% higher than variance in TA in English and maths, and 10% in science. This means that students of lower ability are awarded higher TA than KS on average, while students of higher ability receive a lower TA than KS. Table 1 shows the TA-KS difference for students with KS level equal to 3, 4 or 5, accounting for around 90% of all students. Students scoring level 3 in the KS on average receive a TA above 3, yielding a positive difference. At higher KS scores, TA-KS is negative; for those achieving level 5 in English KS for example, the mean TA score is 4.7. This relative compression of the TA scores may reflect centrality bias in teachers' assessments and/or larger testing noise in KS scores (see (Grund & Przemeck, 2008; Prendergast, 1999).

The lower part of Table 1 confirms that most of the distribution of TA-KS covers the values (-1, 0, +1). About three quarters of pupils have TA-KS equal to zero, consistent with previous research (Reeves et al, 2001; Thomas et al, 1998; Gibbons et al. 2008), and less than 5% of pupils have an absolute difference greater than one. There are differences between subjects: TA-KS<0 or "under-assessment" is much more common than "over-assessment" in English and science, but in maths "over-assessment" is slightly more common. This is true in all years (2002-2005), which suggests robust patterns of over and under assessment, and may be due to the nature of the subject, for example the degree of subjectivity.

b. Pupil level analysis – assessments and ethnicity

We now turn to an analysis by ethnicity. Table 2 shows the percentage of each group with TA<KS, TA=KS or TA>KS. Given the strong dependence of TA-KS on the KS level, we present results for a given level (level 4 in KS, the "expected" level of attainment at KS2). We focus particularly on TA<KS and find that 12.4% of white pupils have TA<KS in English. This compares to 17.2% of Black Caribbean students, 18.3% of Black African, 20.2% of Pakistani and 18.1% of Bangladeshi students. Indian and Chinese students are more comparable to their white peers, but still have greater proportions of TA<KS; 13.8% and 13.3% respectively.

There are differences between subjects: in maths and science, the proportion of students with TA<KS is around 23% *lower* for Chinese than for White students, while the proportion of Indian students is essentially the same as white students in science. The degree of discrepancy between white students and other ethnic groups varies between subjects, for example the TA<KS rate is about 63% higher for Pakistani than for white students in English, 50% in maths and 41% higher in science.

If TA falling below KS represents random error in assessment, then the distribution of (TA-KS) should be symmetric and the frequency of (TA>KS) should be comparable to that of (TA<KS). For some ethnic groups in some subjects however, the rates at which TA<KS are far higher than the rates at which TA>KS. In Science, the proportion of TA<KS is around 3 times higher than TA>KS for Pakistani students, and over 2 time higher for Bangladeshi students. This represents a non-random allocation of TA relative to KS for some groups, although differences in maths are less marked.

Focusing on other groups, students eligible for FSM are more likely to have TA<KS in all subjects. The largest difference between groups is between those students with SEN and those without. For pupils with SEN, the proportion of pupils with TA>KS is exceptionally small (in the range 2.1% to 4.5% in all cases), while around a third of students with SEN have TA<KS in English and science. This could be an extreme form of "teaching to the test" for pupils with SEN, whereas the teacher's more in-depth knowledge of the student's ability may result in a lower TA. It is possible however that SEN is correlated with worse behaviour, or that a label of SEN serves to reinforce teachers' low expectations. We return to this issue below.

We now turn to model these differences in a multivariate setting. We use a linear probability model for the likelihood that TA<KS. We offer four specifications for the model, which all include the students' KS score to account for the strong negative relationship and isolate differences between groups at the same KS level. In specification 1 we include ethnicity coefficients only; in specification 2 we add other personal characteristics; and in specification 3 we add school characteristics and LA fixed effects. Finally in specification 4 we include school fixed effects in preference to LA effects and school variables. We analyse each specification separately by subject; results for English are shown in Table 3, maths in Table 4 and science in Table 5. We present only the coefficients for ethnicity in these tables, but the full results for Table 3 are presented in Appendix Table 2. The coefficients are marginal effects; a positive coefficient of 0.04 corresponds to a 4 percentage point increase in the probability that TA<KS relative to white students. We include the raw mean for the proportion of TA<KS for each group in the table for comparison.

In Table 3, specification 1 there are statistically significant and quantitatively substantial positive marginal effects for the majority of ethnic groups. The largest effects are for Black Caribbean, Black African, Pakistani and Bangladeshi students, and students of other Asian ethnicity. In specification 2 all coefficients decline due to the correlation of ethnicity with poverty (FSM), SEN and EAL[15]. The largest changes are for Pakistani and Black African students, and also for Bangladeshi pupils for whom the coefficient becomes insignificant. Coefficients for Indian and Chinese students become negative, indicating a lower probability of TA<KS than white pupils, statistically significantly for Chinese students and marginally significant for Indian students. Adding school characteristics in specification 3 further reduces the coefficients for the South Asian groups. The coefficient for Indian students is now -0.018 and statistically significant; Indian students are 1.8 percentage points less likely than white students to have TA<KS. In column 4 we add school fixed effects, so marginal effects now derive from variation within schools. There remain substantial and significant positive effects for Black Caribbean and Black African students, 2.5 and 1.7 percentage points more likely than white students to have TA<KS respectively. Chinese, Indian and Mixed White Asian students now have substantially negative coefficients. Looking across specifications, the biggest unconditional effects

[15] 15% of white pupils receive FSM, compared to 31% Black Caribbean, 42% Black African, 35% Pakistani, and 50% Bangladeshi.

for Pakistani and Bangladeshi pupils have largely been explained by personal characteristics and school fixed effects. While the coefficient for Black Caribbean students was within the range of others in column 1, it is noticeably larger than others in column 4, suggesting persistent differences for this group.

In Tables 4 and 5 we present the equivalent results for maths and science. Focusing on specification 4, we see a very similar pattern to English. Coefficients are positive and significant for Black Caribbean and Black African students, negative for Indian, Chinese and Mixed White Asian students, and close to zero for Pakistani and Bangladeshi students. Again, Black Caribbean students have the largest positive coefficient, with a marginal effect of 0.014 in maths and 0.035 in science. The coefficients for Indian and Chinese students in maths in science are quantitatively substantial and negative, most notably -0.046 for Chinese students in maths, and -0.066 in science.

c. Robustness checks

We interpret these results as arising from the student-teacher interaction modelled in section 2. An obvious alternative is that it arises from student-teacher assignment. For example, following the findings of (Clotfelter, Ladd, & Vigdor, 2005), it may be that minority students are disproportionately assigned to inexperienced teachers who are less adept at forming assessments. Non-random assignment *between* schools is dealt with in specification (4) above by the inclusion of school fixed effects. We address non-random assignment *within* schools by restricting the sample to one-class-per-cohort schools, thus ensuring that all students are taught by the same teacher. The results, reported in column 1 of Appendix Table 3, show little qualitative difference from the main results above: Black Caribbean students remain above 2 percentage points more likely to have TA<KS, and Indian and Chinese students 2 percentage points less likely. The coefficient for Black African students is reduced by almost half, but this is the exception.

It is clear from the basic data in Table 2 that the relationship between TA and KS is very different for students with SEN, and this designation is correlated with some ethnic minority groups. Whilst the results above control for SEN, we repeat the

analysis omitting these students. We find that results are robust and in fact are stronger in the majority of cases, see column 2 in Appendix Table 3.

Functional form may be an issue. Since we cannot observe the underlying level *L*, we cannot directly investigate *KS(L)* and *TA(L)*. But we can restrict the range of KS scores we run the analysis over, as it is conceivable that pupils achieving outlier levels are driving the results. Restricting the sample to only students who achieve level 4 in the KS tests, we find that the absolute level of the coefficients is very similar for most ethnic groups; see column 3 of Appendix Table 3. The coefficient for Black Caribbean students is 0.026, compared with 0.025 in our main specification for example. We conclude that our results do not seem to be driven by those observations at the extremes of the distribution.

We repeat the analysis, restricting the sample to school-cohorts with at least 5 of the designated minority group and at least 5 white students. This removes all-white school-cohorts from the analysis and ensures that comparisons between groups are made in schools with sufficient numbers. This subset is no longer representative, being more urban and generally poorer[16]. Under this sample restriction the coefficients for Black Caribbean and Pakistani students remain similar to those in the main results, but coefficients for Black African and Indian students decrease in magnitude. This is reported in column 4 of Appendix Table 3.

The effects of poverty and of ethnicity can be difficult to disentangle, and FSM eligibility is rather coarse. We add to this a fine grained measure of neighbourhood disadvantage, based on the full unit postcode (zipcode) of each student. This classifies very small neighbourhoods (typically around 15 dwellings) into 61 neighbourhood types[17]. Including dummy variables for each type leaves the results for ethnic minority indicators effectively unchanged (column 5 of Appendix Table 3).

Finally we report on an exercise to consider any stereotyping effects within the white population. Clearly, the 84% of students who are white are not a simple homogeneous group and one way we can distinguish them is by their neighbourhood. This will be correlated with a variety of other factors that might influence teachers' views of these students. We divide the 61 neighbourhood types into the poorest third, middle

[16] In the full sample, 17% of pupils have FSM, but in restricted samples this rises to 30% for Black Caribbean sample, 36% for Black African, 27% for Bangladeshi and 43% for Pakistani. The percentage falls in to 16% in the Indian sample. In each restricted sample, 99% of pupils are in urban schools, compared to 82% in the full sample.

[17] This is commercial geo-demographic data, MOSAIC, kindly supplied to us by Experian.

(omitted category) and least poor thirds, and introduce indicators for these in the analysis. The results suggest the same factors at work[18]: we find a coefficient of 0.018 for whites living in poor neighbourhoods (compare 0.025 for Black Caribbean ethnicity, and 0.036 for FSM eligibility), and -0.019 for whites living in the least poor neighbourhoods (compare -0.018 for Indian ethnicity and -0.019 for Chinese ethnicity).

5. Interpretation

Having established the existence and nature of the assessment gaps, we now turn to interpreting their cause. Adopting simplified linear forms from the measurement framework above, we have for pupil *i* in school *s* that $KS_i = \beta.L_i + \delta.X_i + \varepsilon_i$ and $TA_i = \alpha.L_i + \gamma_{s(i)}.X_i + \upsilon_i$, yielding the gap as:

$$(TA - KS)_i = ((\alpha/\beta) - 1)KS_i + (\gamma_{s(i)} - (\alpha/\beta)\delta)X_i + \omega_i$$

We want to establish whether the impact of *X* on the conditional gap arises principally through its impact on the teacher's subjective assessment, or through the test score. We consider the latter first, focussing on two reasons why ethnic minority status might affect the test score.

a. Minority status and the Test Score

First, it could be that the tests are culturally biased against some groups, typically argued to be black students and poor students (Gipps, 1992; Murphy & Pardaffy, 1989). If this were a major factor, we would expect to see these groups performing less well in the tests than in their teachers' assessments. However, our results run exactly counter that view: these are precisely the groups we show to be achieving more than their teachers' assessments.

Second, it could be that some ethnic groups take school tests more seriously than other groups, and more seriously than day-to-day school work. This behaviour might arise because of a perceived differential rate of return to test results, or because of cultural differences in the importance attached to schooling. This approach would be

[18] In specification 4 for English.

reflected in higher test scores than assessments relative to the other groups. There is some *a priori* plausibility to this as it is often argued that some minority groups see education as more important than white students do ((Burgess, Wilson, & Briggs, 2009; Connor, Tyers, Modood, & Hillage, 2004)). Some of the results we find do not fit well with this hypothesis. Indian students are an example a group that place a high value on educational attainment, yet the coefficient we find in this case is negative, opposite to what the theory would predict. Nevertheless, the hypothesis is worthy of closer investigation and we test it as follows. One implication of the argument is that the differences that arise between ethnic groups should not vary systematically across subjects or across schools within ethnic group. That is: if some students give added importance to the Key stage tests, prepare more and try harder in the test, then this behaviour should apply equally to maths as to English, and in one school as in another. This is what we test.

In the top panel of Table 6, we first test for the equality of subject effects by ethnicity. Specifically, we include interactions between ethnicity and subject using specification 4 of table 3 pooled across subjects, and test whether $\beta_{english}^{[group]} = \beta_{maths}^{[group]}$, *all groups*. The results strongly reject this hypothesis. We also test for differences in the ethnic group coefficients across schools; again, we would not expect systematic variations under this hypothesis. We test the significance of ethnicity*school dummy interactions. Because of the large number of variables, we do this ethnic group by group[19], and report the results in the second panel of Table 6. The data strongly reject the presence of a single ethnic group effect, constant across schools. This is true for all ethnic groups, but particularly so for Pakistani and Black African students. In fact, in a test reported below, we show that any argument based solely on the behaviour of students can be ruled out.

b. Minority status and Teacher Assessments

We now consider two hypotheses concerning the impact of minority status on the teacher assessment. First, it may be that teachers have simple discriminatory views, believing white students to be more able than others. There are two results standing against this straightforward view. First, there are clear differences in effect among the

[19] That is, each row is a separate regression, containing students of the named ethnic group plus white students.

non-white groups, including some (for Indian and Chinese students) of the opposite sign. Second, even if we broaden the hypothesis to allow for a discriminatory view distinguishing different minority groups, we have shown that there are significant differences between subject outcomes within an ethnic group. It is difficult to see how these could be explained under this hypothesis.

A second hypothesis is that differences in pupils' behaviour drives the difference in teachers' assessments. Interpretation of this is not straightforward. In principle, a teacher ought to be able to see through behaviour in class and correctly judge the level of attainment of a pupil. Under this view, the behaviour is taken account of by the teacher and does not directly affect her/his assessment. This might be an overly optimistic view of what a hard-pressed teacher can accomplish in a class of 30 students, and we need to consider the possibility that it is student behaviour differences that are generating the conditional assessment gaps. We do this by first implementing a further test using this model, and then revisit the issue below in a supplementary dataset.

Because we have three observations for each student (English, maths and science), we can introduce student fixed effects. These will therefore control for any individual behaviour patterns that the students may display[20]. In a primary school, the same teacher is with the students all day, teaching a mix of different subjects, and so it cannot be the case that a student engages in one set of behaviours in maths and another in science. We therefore look for variation across subjects within a teacher-student match, by ethnicity. This is a powerful test since such a large degree of variation is being controlled for with the student fixed effects. The results are in the third panel of Table 6. They show despite controlling for behaviour and other fixed student characteristics, there are significant ethnicity*subject differences in the conditional assessment gap. Nevertheless, since it is an important potential component of the story, we return to the issue of behaviour in sub-section (d) below.

We have argued against the hypotheses considered above on the grounds of the presence of variation in the conditional gap across subjects and across schools that the hypotheses cannot explain. Any successful explanation therefore needs to be able to explain that variation; we set out such a model based on stereo-typing.

[20] Indeed, it also controls completely for any individual differences in behaviour, ability or effort, reinforcing the refutation above of differences in exam preparation.

c. A model of assessment formation with stereotypes

We assume that it is costly in time and effort for a teacher to form an assessment of a pupil's ability level. Given this, it is rational for a teacher to use all available information as long as it is cheap and reliable, available through categorisation of experiences. Based on the categorical model of cognition (Fryer & Jackson, 2008), we assume that the teacher combines information derived from observing and questioning the specific, individual pupil in front of them with the prototype for that pupil's group. Let *g(i)* denote *i*'s group. We assume that the teacher assessment arises from a weighted average of the specific pupil and the prototype information:

$$TA_{i(j)} = \pi.obs_{i(j)} + (1-\pi).\hat{L}_{g(i)} \qquad (8)$$

where, $\hat{L}_{g(i)}$ is the prototype, and we assume that the specific observation of the pupil depends on the factors set out in (2). For $\hat{L}_{g(i)}$ we use the past mean test score of group g in that school, $\overline{ks}_{g(i),s(i),t-1}$. This is a specific, perhaps rather narrow, assumption, implying that the prototype comes only through previous experience in that school and not from broader stereotypes. Under this approach, our empirical model is expanded as follows:

$$pr(d_i < 0) = \alpha.KS_i + \boldsymbol{\beta}.\mathbf{X}_i + \delta\overline{KS}_{g(i),s(i),t-1} + \mu_{s(i)} + \varepsilon_i \qquad (9)$$

We expect δ to be negative – a low group mean Key Stage score leads the teacher to make a low assessment, and therefore, given the pupil's actual ability, a higher chance of finding that *TA* is below *KS*. Note that the model also has school fixed effects, so the variation in past group scores is across years and across groups within-school.

This approach accords with the broad facts. The groups more likely to be "under-assessed" are those whose educational performance is widely portrayed as being poor, such as Black Caribbean students. Groups known to perform well in certain contexts are "over-assessed" on average, for example ethnic Chinese students in maths and science. Figure 1 illustrates this relationship more formally. We take groups defined by ethnicity, gender, and FSM status, and consider each subject separately. On the vertical axis we plot the mean effect on the likelihood of under-assessment as measured by that group's coefficient from specification 4 in Tables 3, 4, and 5. The

horizontal axis plots actual national performance for a given group relative to their obvious comparator group. The figure shows a clear negative relationship. Groups performing poorly relative to their comparator group are under-assessed, for example students with FSM and Black Caribbean students, which lie in the top left quadrant of the graph. Groups performing better than their comparator group tend to be over-assessed. They are in the bottom right quadrant, and include Mixed White Asian and Chinese students in all subjects (especially maths and science), and female students in English. There are very few points in the other quadrants, suggesting that the relationship holds for most groups.

We now address this more formally by estimating (9). We restrict the sample to the largest groups to ensure that the within-school group averages are robust. To be included in the sample, group *g* in a given school cohort must have 5 or more students of the same group in the previous year. The cohort must also have 5 or more students of the comparator group for comparison; we run the original regression (excluding $\overline{ks}_{g(i),s(i),t-1}$) on this restricted sample for comparison. This is a strong test for the prototype model since much of the basis for a teacher's view may come from more diffuse and general sources.

Table 7 shows that for each subject, that the stereotype variable $\overline{ks}_{g(i),s(i),t-1}$ is strongly significant, giving weight to the hypothesis that teachers' previous experience affects current assessment. The effect is quantitatively strongest in English, then science and then maths. This is in line with the idea that judgement is most subjective in English and least in maths, although the sizes relative to the ethnicity coefficients are broadly similar[21]. After inclusion of $\overline{ks}_{g(i),s(i),t-1}$, changes in the ethnic group coefficients vary in magnitude. For significant coefficients however, the coefficients fall in absolute value as expected. The percentage decreases for the coefficients are not huge, for example for Black Caribbean students there is a 12% decrease in English, 20% in maths and 11% in science. For Black African students the respective decreases are 14%, 33% and 28%. For female students in English (not shown in the tables), including the stereotype reduces the coefficient by 63% from -0.0155 to -0.00578.

[21] To account for broader influences, we have also used the national performance of the group in the previous year as the stereotype variable. This has inadequate variation between years for identification however, and is not significant for any subject: coefficients for the national mean in the previous year are -0.029 (t-stat of 1.3) in English, -0.018 (0.8) in maths and 0.009 (0.4) in science. Full results are available from the authors.

We have also experimented with a regression combining subjects which therefore brings in variation across subject as well as across group. This exploits the variation between Chinese students' performance in maths from Chinese students' performance in English for example (see Figure 1). This modification yields similar results; a slightly more significant coefficient for $\overline{ks}_{g(i),s(i),t-1}$ of -0.0133 (t statistic of 25) and an absolute decline in the group coefficients of about the same magnitude.

Finally, we are able to follow up an implication of the Fryer and Jackson (2008) framework. They show that minority experiences are more coarsely sorted than majority experiences, and also that the variance from categorisation depends on the size of the groups. We interpret this as suggesting that the prototype information will be more valuable in situations when the experience is rarer and so categorised more coarsely. Specifically, in schools where teachers regularly meet pupils from ethnic minorities, they may be categorised more finely, confident in their ability to judge an individual pupil, whereas teachers in schools with few ethnic minorities will rely more on the coarse, more inaccurate, prototype information. We test this by splitting the sample up into observations where ethnic minorities form the majority of a school-cohort and those where they form a minority. The results are in table 8. The coefficient on the stereotype variable is more than twice as large in absolute value for the latter case, and significantly different.

This school-based stereotyping behaviour is not the full explanation of the assessment gap, but the results suggest that it is an important part of the story. It may be that teachers also draw information from the national-level patterns displayed in Figure 1, and that that factor explains the remaining gaps, but this is difficult to test given the lack of variation. (Sewell, 1997) writes that teachers "cannot escape the wider perceptions" that exist about Black boys.

d. Student Behaviour

Finally we return to the potential role of differences in student behaviour in explaining the conditional assessment gap. (Bennett, Gottesman, Rock, & Cerullo, 1993) find that pupils' behaviour in class has an effect on teacher assessments. (Pedulla, Airasian, & Madaus, 1980)) also find that teachers' judgements of performance are confounded with judgement of other academically related behaviours, such as

attention and persistence. Since our main dataset is an administrative dataset, it has no detailed data on students' behaviour in class or their views. We instead turn to a survey dataset to provide richer insight into the effect of young people's behaviour. We use the first wave of the Longitudinal Study of Young People in England (LSYPE), a panel study of around 14,000 young people aged 13 and 14 in 2004[22]. This dataset provides a very rich set of questions on most aspects of school life, and includes questions regarding the pupils' attitude to and behaviour in school, as well as all the variables used in our analysis above. School identifiers are not available, however.

We first investigate whether some ethnic groups report worse behaviour or less effort in class. The results in table 9 show that Indian and Pakistani pupils are significantly less likely than white pupils to cause trouble in most of their classes, and Black Caribbean, Black African and Bangladeshi pupils are not significantly different from white students. All ethnic minority groups are more likely than whites to report working hard in classes, to spend 4 or 5 nights a week on their homework, and to like school. These regressions were run controlling for KS2 English level, but relationships remain if this variable is excluded.

We examine the role of these behaviours in potentially influencing assessment. These results can only be suggestive as the behaviour variables are self-reported and there is a timing problem in that the behaviour relates to age 13-14, while the assessment relates to a period 2-3 years previously. Nevertheless, they provide some insight. The results are in table 10. As might be expected, "reporting praise from your teachers" is significantly negatively correlated with the probability of under assessment in all subjects. Reporting working hard and liking school are also negatively correlated, but not significant. We find that pupils that report "causing trouble in more than half of their classes" are 3.5 percentage points more likely than others to be under assessed. These behavioural variables are jointly significant at 1%. These results suggest that TA is influenced by non-academic factors. It is interesting to note that conditioning on behaviour, the coefficients for some minority groups become significantly positive in some cases, but given the much smaller sample sizes of minorities this may not be robust. In summary, whilst the survey data shows that student behaviours and

[22] More detail about the dataset can be found at http://www.esds.ac.uk/longitudinal/access/lsype/L5545.asp

attitudes do have an influence on the likelihood of under-assessment, such adverse behaviours are if anything more common among white pupils.

5. Conclusions

We have shown that there are enduring and significant differences in teachers' assessments of pupils from different ethnic groups. On average, Black Caribbean and Black African pupils are under-assessed relative to white pupils, and Indian, Chinese and mixed white-Asian pupils are over-assessed. These differences remain after controlling for individual characteristics, and also for school fixed effects. For pupils of Bangladeshi or Pakistani ethnicity, a substantial average under-assessment in the unconditional analysis largely disappears with the introduction of these controls. The fact remains, however, that it is the unconditional differences that will be written on the pupils' record.

There are important differences across subjects within these ethnic groups, and differences between schools across groups and subjects. The observed patterns do not seem to reflect a straightforward discriminatory viewpoint, culturally biased tests, or student behaviour. Pupils in particular ethnic groups and subjects that typically score highly tend to be over-assessed, and vice versa, which matches a model of categorisation and stereotyping. We fit such a model to the data and show that this does explain part of the statistical role of ethnicity. When forming an assessment of a pupil's likely progress, teachers use information on the past performance of members of that group in that school from previous years. The dependence of a pupil's assessment on the performance of others of her ethnic group locally means that school composition matters. This is a form of indirect peer effect, and suggests another basis for parents and pupils selecting particular schools.

These results matter for two debates. First, if the systematic teacher under-assessment of some groups is reflected in lower teacher effort for these pupils, then this may impact on their educational outcomes. Given that the school performance of some groups, particularly Black Caribbean boys, remains a matter of concern, this finding is of some relevance. It also seems likely that pupils feeling under-valued by their

teachers are more likely to disengage from the education process altogether, and to reciprocate by under-valuing education and qualifications.

Second, one prominent discussion on education policy in England is the "problem of over-testing" (Brooks & Tough, 2006). It is argued that pupils are subjected to too many written tests, and that some should be replaced by teacher assessments. Along with the work of (Gibbons & Chevalier, 2008), the results here suggest that this might be severely detrimental to the recorded achievements of children from poor families, and for children from some ethnic minorities. For example, in English, using teacher assessment instead of the Key Stage test decreases the proportion of students achieving at least the expected level of attainment[23] by 5.6 percentage points for Black Caribbean pupils, 6.4 for Black African pupils, 4.6 for Indian, 7.0 for Pakistani, 6.9 for Bangladeshi and 4.1 for Chinese, compared with 3.3 for White pupils. This implies a larger raw attainment gap when measured through TA than KS. Given that 'setting' in secondary school classes may depend on earlier recorded attainment and that motivation may also be affected by a lower level, the use of assessment rather than testing may increase attainment gaps between ethnic groups later in academic life.

[23] The expected level of attainment at KS2 is level 4.

References

Arrow, K. (1973). The Theory of Discrimination. *In Orley Ashenfelter and Albert Rees, eds. Discrimination in Labor Markets. Princeton, N.J.: Princeton University Press*, 3-33.

Baird, J. A. (1998). What's in a name? Experiments with blind marking in A-level examinations. *Educational Research, 40*(2), 191-202.

Bennett, R. E., Gottesman, R. L., Rock, D. A., & Cerullo, F. (1993). Influence of Behavior Perceptions and Gender on Teachers Judgments of Students Academic Skill. *Journal of Educational Psychology, 85*(2), 347-356.

Blank, R. M. (1991). The effects of double blind versus single-blind reviewing - experimental-evidence from the American Economic Review. *American Economic Review, 81*(5), 1041-1067.

Brooks, R., & Tough, S. (2006). Assessment and Testing: Making space for teaching and learning.

Burgess, S., Wilson, D., & Briggs, A. (2009). The Dynamics of School Attainment of Englands Ethnic Minorities. *Forthcoming; Journal of Population Economics*.

Chang, D. F., & Demyan, A. (2007). Teachers' stereotypes of Asian, Black, and White students. *School Psychology Quarterly, 22*(2), 91-114.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. (2005). Who teaches whom? Race and the distribution of novice teachers. *Economics of Education Review, 24*(4), 377-392.

Connor, H., Tyers, C., Modood, T., & Hillage, J. (2004). Why the Difference? A Closer Look at Higher Education Minority Ethnic Students and Graduates. *Institute for Employment Studies*.

Dee, T. S. (2005). A teacher like me: Does race, ethnicity, or gender matter? *American Economic Review, 95*(2), 158-165.

Dee, T. S. (2009). Stereotype Threat and the Student-Athlete. *NBER Working Paper Series, w14705*.

Ferguson, R. F. (2003). Teachers' perceptions and expectations and the Black-White test score gap. *Urban Education, 38*(4), 460-507.

Fryer, R., & Jackson, M. O. (2008). A categorical model of cognition and biased decision making. *B E Journal of Theoretical Economics, 8*(1).

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

Gillborn, D. (1990). Race, Ethnicity and Education: Teaching and Learning in Multi-Ethnic Schools.

Gipps, C. V. (1992). National Curriculum Assessment: a research agenda. *British Educational Research Journal, 18*(3), 277 - 286.

Goldin, C., & Rouse, C. (2000). Orchestrating impartiality: The impact of "blind" auditions on female musicians. *American Economic Review, 90*(4), 715-741.

Grund, C., & Przemeck, J. (2008). Subjective Performance Evaluation and Inequality Aversion. *IZA Discussion Paper No. 3382*.

Hanna, R., & Linden, L. (2009). MEASURING DISCRIMINATION IN EDUCATION. *NBER WORKING PAPER SERIES: Working Paper 15057*.

Hobbs, G., & Vignoles, A. (2007). Is Free School Meals a Valid Proxy for Socio-Economic Status (in Schools Research)? *CEE Discussion Paper*.

Hoff, K., & Pandey, P. (2006). Discrimination, social identity, and durable inequalities. *American Economic Review, 96*(2), 206-211.
Jussim, L., & Harber, K. D. (2005). Teacher expectations and self-fulfilling prophecies: Knowns and unknowns, resolved and unresolved controversies. *Personality and Social Psychology Review, 9*(2), 131-155.
Lavy, V. (2004). *Do Gender Stereotypes Reduce Girls' Human Capital Outcomes? Evidence from a Natural Experiment*: SSRN.
Lindahl, E. (2007). Comparing teachers' assessments and national test results - evidence from Sweden.
Mechtenberg, L. (2006). Cheap Talk in the Classroom: How biased grading at school explains gender differences in achievements, career choices, and wages. *Forthcoming; Review of Economic Studies*.
Murphy, K. R., & Pardaffy, V. A. (1989). BIAS IN BEHAVIORALLY ANCHORED RATING-SCALES - GLOBAL OR SCALE-SPECIFIC. *Journal of Applied Psychology, 74*(2), 343-346.
Pedulla, J. J., Airasian, P. W., & Madaus, G. F. (1980). DO TEACHER RATINGS AND STANDARDIZED TEST-RESULTS OF STUDENTS YIELD THE SAME INFORMATION. *American Educational Research Journal, 17*(3), 303-307.
Phelps, E. S. (1972). Statistical Theory of Racism and Sexism. *American Economic Review, 62*(4), 659-661.
Prendergast. (1999). The Provsion of Incentives in Firms. *Journal of Economic Literature, 27*, 7-63.
Reeves, D. J., Boyle, W. F., & Christie, T. (2001). The relationship between teacher assessments and pupil attainments in standard test tasks at Key Stage 2, 1996-98. *British Educational Research Journal, 27*(2), 141-160.
Sewell, T. (1997). Black masculinities and schooling: How Black boys survive modern schooling. *Trentham Books Limited*.
Smith, E. R., & Zarate, M. A. (1992). Exemplar-Based Model of Social Judgment. *Psychological Review, 99*(1), 3-21.
Steele, C. M., & Aronson, J. (1995). Stereotype Threat and the Intellectual Test-Performance of African-Americans. *Journal of Personality and Social Psychology, 69*(5), 797-811.
Strand, S. (2007). Minority Ethnic Pupils in the Longitudinal Study of Young People in England (LSYPE).
Thomas, S., Madaus, G., Raczek, A., & Smees, R. (1998). Comparing Teacher Assessment and Standard Task Results in England: the relationship between pupil characteristics and attainment. *Assessment in Education: Principles, Policy & Practice, 5*(2), 213-246.
Weinstein, R. S. (2002). Reaching Higher: The Power of Expectations in Schooling *Harvard University Press*.

Table 1: Summary statistics for the difference between Teacher Assessment (TA) and Key Stage Test level (KS); TA-KS

	TA-KS, English			TA-KS, Maths			TA-KS, Science		
KS	Mean	SD	% N	Mean	SD	% N	Mean	SD	% N
3	0.12	0.50	15.95	0.18	0.51	19.41	0.14	0.55	9.75
4	-0.03	0.49	50.52	0.01	0.43	45.27	-0.06	0.49	45.59
5	-0.29	0.48	27.29	-0.18	0.42	29.52	-0.32	0.50	42.22

	English	Maths	Science
TA-KS	%	%	%
-1	15.04	9.80	19.70
0	71.23	75.99	71.86
1	9.35	10.25	6.49

Note. The sample was taken from academic years 2001/2002 to 2004/2005, and includes only those with both TA and KS results. SD stands for Standard Deviation, TA for Teacher Assessment; KS for Key Stage Test. TA-KS is the difference between TA and KS, measured in levels. The top panel shows summary statistics for TA-KS at different levels of KS, by subject. The bottom panel shows the proportion of students with a difference between TA and KS of -1, 0, or 1, by subject.

Table 2: Teacher Assessment (TA) relative to the Key Stage Test (KS), given that the pupil achieved level 4 in the Key Stage Test, by subject

	English (%)			Maths (%)			Science (%)		
Variable	TA<KS	TA=KS	TA>KS	TA<KS	TA=KS	TA>KS	TA<KS	TA=KS	TA>KS
Ethnic Group									
White	12.4	77.5	10.2	7.9	82.5	9.6	13.6	78.2	8.3
Black Caribbean	17.2	75.2	7.5	10.2	81.3	8.5	17.3	76.0	6.7
Black African	18.3	74.3	7.4	10.6	81.0	8.4	16.9	75.7	7.4
Indian	13.8	76.3	10.0	8.4	80.5	11.1	13.8	76.4	9.9
Pakistani	20.2	73.6	6.2	11.9	80.3	7.8	19.2	74.2	6.6
Bangladeshi	18.1	75.2	6.7	11.2	81.0	7.8	16.5	75.8	7.7
Chinese	13.3	76.0	10.7	6.0	81.1	12.9	10.6	76.1	13.3
Special Educational Needs (SEN)									
No SEN	9.6	79.2	11.2	6.0	83.3	10.7	8.7	81.1	10.2
SEN, without statement	32.9	65.0	2.1	19.3	77.5	3.3	29.2	68.6	2.2
SEN, with statement	33.6	62.7	3.8	22.2	73.2	4.5	39.6	58.0	2.4
Free School Meals (FSM)									
No FSM	11.8	77.5	10.6	7.5	82.4	10.1	12.5	78.5	8.9
FSM	19.1	75.1	5.8	11.9	81.5	6.6	20.2	74.7	5.1
English as an additional language (EAL)									
Not EAL	12.5	77.4	10.1	7.9	82.5	9.6	13.7	78.1	8.2
EAL	18.1	74.4	7.5	10.6	80.6	8.8	17.2	75.0	7.8
Male	14.3	76.5	9.1	9.0	81.7	9.3	15.1	76.5	8.4
Female	11.6	77.8	10.6	7.4	82.8	9.7	12.9	79.1	8.0

Note. The sample was taken from academic years 2001/2002 to 2004/2005, and includes only those with both TA and KS results. Cells give the proportion of the group with TA<KS, TA=KS and TA>KS, given that students achieved level 4 in the Key Stage test in the subject. Level 4 is the expected level of attainment at KS2 (DCSF). TA stands for Teacher Assessment; KS for Key Stage Test.

Table 3: The probability that TA<KS in English. The dependent variable is binary, equal to one if TA<KS.

Variable	Specification 1 β	t stat	Specification 2 β	t stat	Specification 3 β	t stat	Specification 4 β	t stat	Raw mean % TA<KS
KS2 score	0.066	171.65	0.105	190.13	0.108	192.03	0.110	199.10	
Ethnic Group									
Black Caribbean	0.040	12.87	0.027	8.68	0.027	9.52	0.025	11.39	0.172
Black African	0.048	14.56	0.019	5.87	0.018	6.41	0.017	7.15	0.178
Black Other	0.035	8.53	0.016	3.99	0.014	3.53	0.012	3.36	0.173
Indian	0.017	4.81	-0.006	1.65	-0.018	5.35	-0.018	7.61	0.171
Pakistani	0.059	15.46	0.027	6.47	0.015	3.70	0.010	3.66	0.181
Bangladeshi	0.046	8.95	0.004	0.84	-0.001	0.21	0.002	0.71	0.178
Other Asian ethnicity	0.044	9.36	0.028	5.87	0.019	4.17	0.017	4.22	0.186
Chinese	0.007	1.45	-0.015	3.00	-0.017	3.49	-0.019	4.15	0.171
Mixed White and Black Caribbean	0.028	9.03	0.020	6.52	0.013	4.51	0.013	4.65	0.173
Mixed White and Black African	0.013	2.21	0.004	0.77	0.003	0.59	0.004	0.74	0.162
Mixed White and Asian	-0.011	2.85	-0.014	3.64	-0.018	4.71	-0.015	4.19	0.150
Mixed Other	0.013	3.99	0.003	1.02	0.000	0.01	-0.001	0.29	0.168
Other	0.044	14.77	0.020	6.82	0.018	6.34	0.017	6.63	0.179
Missing	0.013	4.67	0.008	2.99	0.007	2.68	0.007	3.02	0.157
Reference group: White									0.150
Other personal characteristics?	No		Yes		Yes		Yes		
School characteristics?	No		No		Yes		No		
LA fixed effects?	No		No		Yes		No		
School fixed effects?	No		No		No		Yes		
R^2	0.044		0.071		0.078		0.074		
Number of Observations	2255382		2255382		2227352		2255382		
Number of Schools	16550		16550		15719		16550		

Note. The sample was taken from academic years 2001/2002 to 2004/2005, and includes only those with both TA and KS results. OLS regressions were run with standard errors clustered by school. See Appendix Table 3 for full results. Specification 1 includes Key Stage Test level and a set of ethnicity dummies only. Specification 2 also controls for observable pupil characteristics such as whether they have free school meals (an indicator for poverty status). Specification 3 also includes school characteristics, such as faith school status, and LA fixed effects. Specification 4 includes school fixed effects in place of school characteristics and LA fixed effects. Full details of all specifications are given in Appendix Table 2.

Table 4: The probability that TA<KS in Maths. The dependent variable is binary, equal to one if TA<KS.

	Specification 1		Specification 2		Specification 3		Specification 4		Raw mean
Variable	β	t stat	β	t stat	β	t stat	β	t stat	% TA<KS
KS2 score	0.044	144.24	0.067	165.14	0.069	166.51	0.070	173.80	
Ethnic Group									
Black Caribbean	0.023	9.75	0.017	7.30	0.017	8.07	0.014	7.84	0.105
Black African	0.021	8.65	0.007	2.92	0.007	3.28	0.006	2.87	0.106
Black Other	0.015	4.77	0.005	1.73	0.004	1.36	0.002	0.62	0.103
Indian	0.001	0.44	-0.009	3.50	-0.017	7.30	-0.019	9.71	0.106
Pakistani	0.033	11.31	0.018	5.68	0.007	2.25	0.002	1.23	0.114
Bangladeshi	0.030	6.94	0.008	1.92	0.003	0.66	0.000	0.04	0.117
Other Asian ethnicity	0.000	0.14	-0.010	2.87	-0.015	4.34	-0.014	4.48	0.102
Chinese	-0.030	8.18	-0.045	11.72	-0.047	12.21	-0.046	12.69	0.092
Mixed White and Black Caribbean	0.007	2.88	0.003	1.33	-0.002	0.79	-0.002	1.07	0.100
Mixed White and Black African	-0.006	1.38	-0.010	2.19	-0.010	2.34	-0.010	2.34	0.090
Mixed White and Asian	-0.011	3.42	-0.011	3.65	-0.013	4.37	-0.011	3.59	0.096
Mixed Other	-0.001	0.47	-0.005	2.12	-0.006	2.60	-0.006	2.81	0.100
Other	0.020	7.71	0.004	1.70	0.003	1.12	0.003	1.22	0.115
Missing	0.004	2.03	0.002	0.97	0.003	1.31	0.002	1.21	0.100
Reference group: White									0.100
Other personal characteristics?	No		Yes		Yes		Yes		
School characteristics?	No		No		Yes		No		
LA fixed effects?	No		No		Yes		No		
School fixed effects?	No		No		No		Yes		
R^2	0.027		0.045		0.053		0.047		
Number of Observations	2255382		2255382		2227352		2255382		
Number of Schools	16550		16550		15719		16550		

Note. The sample was taken from academic years 2001/2002 to 2004/2005, and includes only those with both TA and KS results. OLS regressions were run with standard errors clustered by school. See Appendix Table 3 for full results. Specification 1 includes Key Stage Test level and a set of ethnicity dummies only. Specification 2 also controls for observable pupil characteristics such as whether they have free school meals (an indicator for poverty status). Specification 3 also includes school characteristics, such as faith school status, and LA fixed effects. Specification 4 includes school fixed effects in place of school characteristics and LA fixed effects. Full details of all specifications are given in Appendix Table 2.

Table 5: The probability that TA<KS in Science. The dependent variable is binary, equal to one if TA<KS.

Variable	Specification 1 β	t stat	Specification 2 β	t stat	Specification 3 β	t stat	Specification 4 β	t stat	Raw mean % TA<KS
KS2 score	0.096	141.93	0.149	189.46	0.155	190.78	0.162	207.20	
Ethnic Group									
Black Caribbean	0.040	11.11	0.026	7.26	0.034	11.01	0.035	14.30	0.215
Black African	0.032	8.80	0.011	3.26	0.018	5.90	0.019	7.16	0.199
Black Other	0.029	6.36	0.014	3.18	0.018	4.11	0.017	4.52	0.211
Indian	-0.001	0.34	-0.010	2.19	-0.026	6.70	-0.028	10.92	0.198
Pakistani	0.048	11.44	0.028	6.15	0.006	1.35	0.002	0.76	0.206
Bangladeshi	0.029	4.71	-0.003	0.54	-0.008	1.48	-0.010	2.70	0.200
Other Asian ethnicity	0.012	2.39	0.004	0.78	-0.005	1.14	-0.005	1.18	0.201
Chinese	-0.052	10.82	-0.063	12.51	-0.066	13.43	-0.066	14.44	0.167
Mixed White and Black Caribbean	0.029	8.30	0.015	4.45	0.010	2.99	0.011	3.45	0.224
Mixed White and Black African	-0.001	0.24	-0.012	1.96	-0.010	1.67	-0.010	1.70	0.194
Mixed White and Asian	-0.016	3.76	-0.019	4.40	-0.023	5.63	-0.019	4.75	0.196
Mixed Other	0.000	0.11	-0.012	3.68	-0.012	3.89	-0.011	3.69	0.206
Other	0.023	7.04	0.008	2.53	0.009	3.01	0.010	3.76	0.206
Missing	0.015	4.72	0.008	2.56	0.009	2.92	0.010	4.14	0.213
Reference group: White									0.203
Other personal characteristics?	No		Yes		Yes		Yes		
School characteristics?	No		No		Yes		No		
LA fixed effects?	No		No		Yes		No		
School fixed effects?	No		No		No		Yes		
R^2	0.045		0.087		0.097		0.095		
Number of Observations	2255382		2255382		2227352		2255382		
Number of Schools	16550		16550		15719		16550		

Note. The sample was taken from academic years 2001/2002 to 2004/2005, and includes only those with both TA and KS results. OLS regressions were run with standard errors clustered by school. See Appendix Table 3 for full results. Specification 1 includes Key Stage Test level and a set of ethnicity dummies only. Specification 2 also controls for observable pupil characteristics such as whether they have free school meals (an indicator for poverty status). Specification 3 also includes school characteristics, such as faith school status, and LA fixed effects. Specification 4 includes school fixed effects in place of school characteristics and LA fixed effects. Full details of all specifications are given in Appendix Table 2.

Table 6: Hypothesis tests to help interpretation of differences between groups.

Test	Subject	Ethnic group	Controls	Sample	F stat	P value	Reject?
Equality of subject effects by ethnicity	English, Maths	All	Full set of pupil controls	Full sample, as in main regressions	189.2	0.00	Y
	English, Science	All	Full set of pupil controls	Full sample, as in main regressions	113.29	0.00	Y
	Maths, Science	All	Full set of pupil controls	Full sample, as in main regressions	120.2	0.00	Y
Equality of effects within ethnic group, across schools	English only	Black Caribbean	Full set of pupil controls	Pupils in all school years with at least 5 Black Caribbean pupils	3791.92	0.00	Y
		Black African	Full set of pupil controls	Pupils in all school years with at least 5 Black African pupils	4624.73	0.00	Y
		Indian	Full set of pupil controls	Pupils in all school years with at least 5 Indian pupils	11122.7	0.00	Y
		Pakistani	Full set of pupil controls	Pupils in all school years with at least 5 Pakistani pupils	7018.47	0.00	Y
Pupil fixed effects	All	All	Pupil fixed effects Interaction between subject*ethnic group and subject*gender	Full sample, as in main regressions	678.03	0.00	Y

Notes: 1) The first panel reports results for the test of equality of subject effects by ethnicity. Does ethnicity have the same effect on p(TA<KS) across subjects? Results were computed pairwise due to size limitations of the model. All ethnic groups were included in the model, with a full set of pupil controls (as in specification 2 in Appendix Table 2). The F statistic reports the value of the F test for whether coefficients for ethnicity are equal. In each pairwise combination the hypothesis that coefficients across subjects are equal is rejected.

2) The second panel reports results for tests of equality in P(TA<KS) within ethnic group, between schools. These tests were completed separately for each ethnic group, with a full set of pupil controls. We restrict the sample to white students and students of the ethnic group in question, in school years where there are at least 5 students of the ethnic group. In each case, we reject the null hypothesis that the effect of the respective group is the same across schools.

3) The third panel reports results for a test of equality between subject, within pupil. Looking at differences in P(TA<KS) between subject for each pupil removes any pupil specific effects on TA and KS such as classroom behaviour. We allow coefficients to vary by subject and ethnicity, and by subject and gender. We reject the null hypothesis that interaction terms for subject and ethnic group are zero.

Table 7: Statistical discrimination: The impact of local (school level) performance of your own group in the previous year on the probability that TA<KS.
The dependent variable is binary and equal to one if TA<KS

	English				Maths				Science			
	No *t*-1		*t*-1		No *t*-1		*t*-1		No *t*-1		*t*-1	
Variable	β	t stat	β	t stat	β	t stat	β	t stat	β	t stat	β	t stat
KS2 score	0.108	178.88	0.108	178.99	0.069	155.23	0.069	155.22	0.162	192.68	0.162	192.77
Ethnic Group												
Black Caribbean	0.033	10.94	0.029	9.75	0.015	6.29	0.012	5.04	0.035	11.03	0.031	9.71
Black African	0.022	6.86	0.019	6.18	0.003	1.28	0.002	0.65	0.018	5.44	0.013	3.92
Indian	-0.009	3.00	-0.003	0.90	-0.020	7.68	-0.018	6.61	-0.029	8.65	-0.028	8.30
Pakistani	0.015	4.55	0.013	3.83	0.003	1.16	0.002	0.59	0.007	1.82	0.000	0.09
Bangladeshi	0.007	1.60	0.008	1.77	-0.004	1.04	-0.004	1.17	-0.008	1.56	-0.012	2.33
Chinese	-0.032	3.87	-0.024	2.94	-0.059	9.55	-0.054	8.58	-0.084	9.83	-0.080	9.31
School mean by group (t-1)			*-0.031*	*20.02*			*-0.011*	*8.85*			*-0.026*	*12.60*
Pupil level characteristics	Yes				Yes				Yes			
School level fixed effects	Yes				Yes				Yes			
R^2	0.073		0.073		0.045		0.045		0.096		0.096	
Number of Observations	1545838		1545838		1545838		1545838		1545838		1545838	

Note. The sample was taken from academic years 2001/2002 to 2004/2005, and includes only those with both TA and KS results. The local (school level) mean is the mean KS2 score in the previous academic year, for your specific ethnic group. The column headed 't-1' gives coefficients for the regression in which the local mean score of the group in the previous year is included. The regression is otherwise the same as in specification 4 in tables 3, 4, and 5. The column headed 'No t-1' does not include the local mean score in the previous year, and therefore has an identical specification to that in tables 3, 4, and 5. The coefficients here vary slightly from tables 3, 4, and 5, as the sample for the regression is restricted to be the same as in regression including t-1.

Table 8: Statistical Discrimination: The impact of local (school level) performance of your own group in the previous year on the probability that TA<KS.
Is the impact the same where ethnic group is in the minority/majority in the school? The dependent variable is binary and equal to one if TA<KS, for English only.

	No *t*-1		*t*-1, full sample		*t*-1, white majority		*t*-1, white minority	
Variable	β	t stat	β	t stat	β	t stat	β	t stat
KS2 score	0.108	178.88	0.108	178.99	0.110	170.18	0.095	60.61
Ethnic Group								
Black Caribbean	0.033	10.94	0.029	9.75	0.029	5.90	0.029	7.19
Black African	0.022	6.86	0.019	6.18	0.018	3.33	0.020	4.95
Indian	-0.009	3.00	-0.003	0.90	-0.007	1.76	-0.006	1.20
Pakistani	0.015	4.55	0.013	3.83	0.017	3.40	0.009	1.89
Bangladeshi	0.007	1.60	0.008	1.77	0.008	1.04	0.006	0.99
Chinese	-0.032	3.87	-0.024	2.94	-0.026	2.60	-0.024	1.67
School mean by group (t-1)			*-0.031*	*20.02*	*-0.037*	*18.23*	*-0.017*	*6.96*
Pupil level characteristics	Yes		Yes		Yes		Yes	
School level fixed effects	Yes		Yes		Yes		Yes	
R^2	0.073		0.073		0.073		0.076	
Number of Observations	1545838		1545838		1387902		148426	

Note. The sample was taken from academic years 2001/2002 to 2004/2005, and includes only those with both TA and KS results. The local (school level) mean is the mean KS2 score in the previous academic year, for your specific ethnic group. Column 1 gives the coefficients for ethnicity variables in the regression where no past school mean is included. Column 2 gives the coefficients for ethnicity variables, including the ethnic group specific past school mean KS2 English score, on the full sample. The specification in column 3 is the same as in column 2, this time on the sample of school cohorts where white students are in the majority (more than half of the relevant student cohort). Column 4 is the same regression, this time on the sample of school cohorts where white students are in the minority (less than half of the relevant student cohort). In column 4 the majority is 'all other' ethnic groups.

Table 9: The correlation of attitudes towards school and ethnicity. Each dependent variable is binary.

	Independent Dummy Variables for Ethnicity										
	Black Caribbean		Black African		Indian		Pakistani		Bangladeshi		
Dependent Variable:	β	t stat	β	t stat	β	t stat	β	t stat	β	t stat	R^2
Believe school is a waste of time	-0.027	2.39	-0.046	3.26	-0.042	6.52	-0.045	5.51	-0.056	6.35	0.032
In trouble in more than 1/2 classes	-0.004	0.19	-0.019	0.80	-0.069	5.88	-0.066	4.26	-0.054	2.60	0.027
Parents report often quarrel	-0.062	2.53	-0.123	4.42	-0.141	8.71	-0.227	13.89	-0.329	23.48	0.013
Student has been suspended	0.078	3.52	-0.028	1.32	-0.067	8.59	-0.103	11.37	-0.136	13.14	0.052
Student has been expelled	0.011	1.04	-0.002	0.56	-0.004	2.81	-0.005	2.24	-0.007	4.60	0.006
Completes 4/5 nights of homework	0.026	1.18	0.235	8.24	0.203	10.59	0.152	8.01	0.171	7.64	0.043
Like school	0.035	1.37	0.145	5.04	0.127	6.99	0.162	8.49	0.185	7.57	0.009

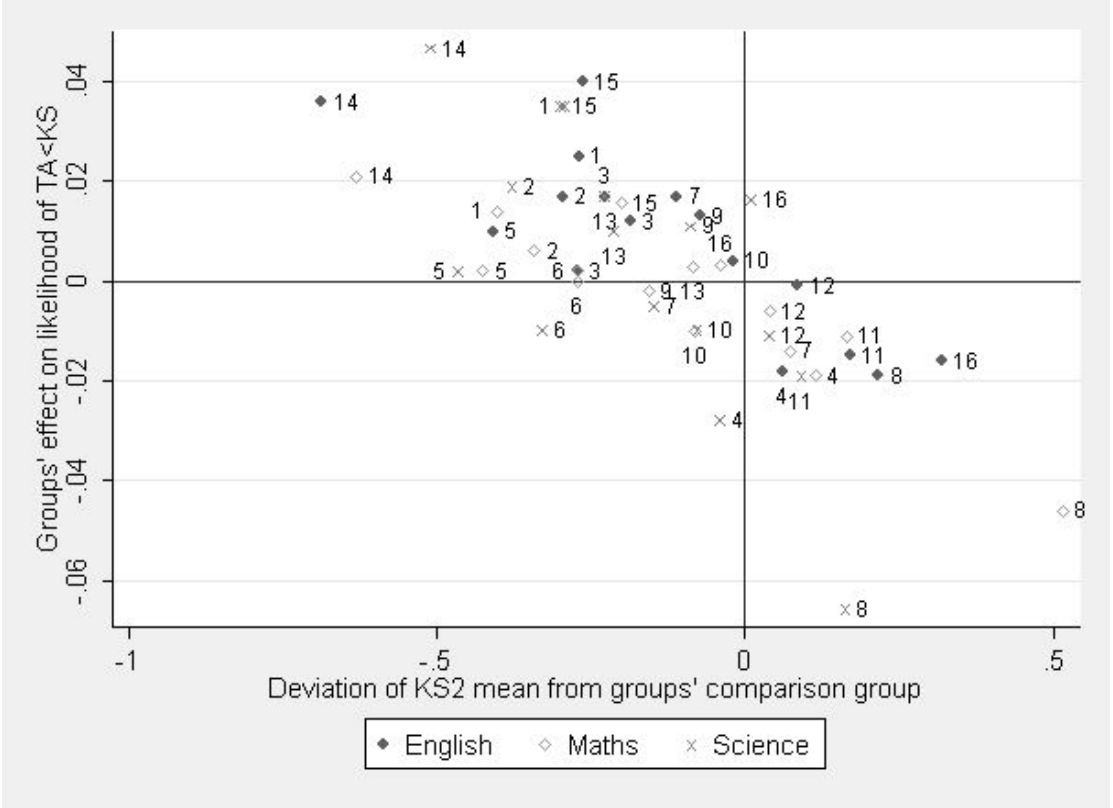
Note. The sample was taken from the Longitudinal Survey of Young People in England (LSYPE), wave 1, when students are 13-14. All dependent variables are binary, and each row represents a different regression. The sample size in each regression is 12378. The columns show the coefficients for ethnic group dummies in relation to White students. Survey weights are applied to the regressions. Robust standard errors are also applied.

Table 10: Probability of TA<KS, by subject, including behavioural variables. Dependent variable is binary, equal to one if TA<KS.

	English TA<KS		Maths TA<KS		Science TA<KS	
Variable	β	t stat	β	t stat	β	t stat
KS2 score	0.113	29.18	0.060	21.68	0.148	30.02
Selected coefficients						
Black Caribbean	0.041	2.07	0.006	0.46	0.063	2.96
Black African	0.053	2.40	0.029	1.74	0.019	0.92
Indian	0.028	1.80	0.003	0.28	0.013	0.89
Pakistani	0.038	2.67	0.030	2.29	0.058	3.78
Bangladeshi	0.073	3.34	0.035	2.21	0.066	3.12
Teachers praise student	0.003	0.36	0.009	1.61	-0.002	0.25
Pupil does 4/5 nights of homework	-0.033	4.04	-0.015	2.39	-0.037	4.44
Pupil likes school	-0.007	0.81	-0.013	2.07	-0.008	1.00
Pupil works hard in class	-0.004	0.46	0.008	1.26	0.012	1.37
Pupil thinks school a 'waste of time'	0.034	2.32				
Pupil causes trouble in >1/2 classes	0.009	0.92	-0.001	-0.16	0.004	0.38
Other personal characteristics?	Yes		Yes		Yes	
School characteristics?	No		No		No	
R^2	0.074		0.043		0.095	
Number of Observations	12378		12396		12374	

Note. The sample was taken from LSYPE, wave 1, and includes only those with both TA and KS results at KS2. The regression is as close as possible to the specification in Appendix Table 2; it includes variables month-of-birth, free school meals status, special education needs status, whether the pupils has English as an Additional Language and gender. This regression also includes behavioural variables, which are self reported from the pupil.

Figure 1: The correlation of relative group performance in KS2 tests and the likelihood of having TA<KS.



Number	Group
Comparison group – White pupils	
1	Black Caribbean
2	Black African
3	Black Other
4	Indian
5	Pakistani
6	Bangladeshi
7	Other Asian ethnicity
8	Chinese
9	Mixed White and Black Caribbean
10	Mixed White and Black African
11	Mixed White and Asian
12	Mixed Other
13	Other
Comparison group – pupils without FSM	
14	Free School Meals (FSM)
Comparison group – pupils without EAL	
15	English as Additional Language (EAL)
Comparison group – male pupils	
16	Female

Note. Mean KS2 scores for each group in each subject were calculated using the full sample used in regressions. This sample uses data from 2001/2002 to 2004/2005, for all pupils with both TA and KS scores. The groups' effect on the likelihood of TA<KS was taken from regression coefficients in specification 4 (including school fixed effects) in tables 3, 4, 5. 'Relative' group performance was calculated as the mean of the group in question, minus the mean of the appropriate reference group, for example non-FSM pupils for FSM pupils.

Appendix Figure 1: Examples of probing questions in maths at KS2.

Teaching objective	Probing questions	Examples of what pupils should know and be able to do
Algebra		
Use systematic trial and improvement methods and ICT tools to find approximate solutions to equations such as $x^3 + x = 20$. SMTP Planned assessment by: Algebra 3	How do you go about choosing a value (of x) to start? How do you use the previous outcomes to decide what to try next? How do you know when to stop? How would you improve the accuracy of your solution? Is your solution exact? Can this equation be solved using any other method? Why?	Use trial and improvement for equivalent problems, e.g. • A number plus its cube is 20. What's the number? • The length of a rectangle is 2cm longer than the width. The area is 67.89. What's the width? Pupils should have opportunities to use a spreadsheet for trial and improvement methods.
Construct and solve linear equations with integer coefficients, using an appropriate method. SMTP Planned assessment by: Algebra 3	How do you decide where to start when solving a linear equation? Given a list of linear equations, ask: Which of these are easy to solve? Which are difficult and why? What strategies are important with the difficult ones? $6 = 2p - 8$. How many solutions does this equation have? Give me other equations with the same solution? Why do they have the same solution? How do you know? How do you go about constructing equations from information given in a problem? How do you check whether it works?	Solve linear equations such as: $3c - 7 = -13$ $1.7m^2 = 10.625$ $4(z + 5) = 8$ $4(b - 1) - 5(b + 1) = 0$ $\frac{12}{(x + 1)} = \frac{21}{(x + 4)}$ Construct linear equations, e.g. The length of a rectangle is three times its width. Its perimeter is 24cm. Find its area.
Generate terms of a sequence using term-to-term and position-to-term definitions of the sequence, on paper and using ICT; write an expression to describe the nth term of an arithmetic sequence. SMTP Planned assessment by: Algebra 1/2	The term-to-term rule for a sequence is +2. What does that tell you about the position-to-term rule? Do you have enough information to find the rule for the nth term? Why? What do you look for in a sequence to help you to find the position-to-term (nth term) rule? How would you go about finding the position-to-term (nth term) rule for this information on a sequence: Position 3 5 10 Term 11 19 39?	

Source: Assessing pupils progress in mathematics at Key Stage 3, Secondary National Strategy.
http://www.standards.dcsf.gov.uk/secondary/keystage3/subjects/maths/focus/asses_maths/ma_app_ass_mats/version_b/

Appendix Figure 2: KS2 level descriptions in English (reading). Assessment guidelines for teachers

Level 5		**Across a range of reading** ▪ most relevant points clearly identified, including those selected from different places in the text ▪ comments generally supported by relevant textual reference or quotation, even when points made are not always accurate	**Across a range of reading** ▪ comments develop explanation of inferred meanings drawing on evidence across the text, *e.g. "you know her dad was lying because earlier she saw him take the letter"* ▪ comments make inferences and deductions based on textual evidence, *e.g. in drawing conclusions about a character's feelings on the basis of their speech and actions*	**Across a range of reading** ▪ comments on structural choices show some general awareness of writer's craft, *e.g. "it tells you all things burglars can do to your house and then the last section explains how the alarm protects you"* ▪ various features relating to organisation at text level, including form, are clearly identified, with some explanation, *e.g. "each section starts with a question as if he's answering the crowd"*
Level 4		**Across a range of reading** ▪ some relevant points identified ▪ comments supported by some generally relevant textual reference or quotation, *e.g. reference is made to appropriate section of text but is unselective and lacks focus*	**Across a range of reading** ▪ comments make inferences based on evidence from different points in the text, *e.g. interpreting a character's motive from their actions at different points* ▪ inferences often correct, but comments are not always rooted securely in the text or repeat narrative or content	**Across a range of reading** ▪ some structural choices identified with simple comment, *e.g. 'he describes the accident first and then goes back to tell you why the child was in the road'* ▪ some basic features of organisation at text level identified, *e.g. 'the writer uses bullet points for the main reasons'*
Level 3	**In most reading** ▪ range of strategies used mostly effectively to read with fluency, understanding and expression	**In most reading** ▪ simple, most obvious points identified though there may also be some misunderstanding, *e.g. about information from different places in the text* ▪ some comments include quotations from or references to text, but not always relevant, *e.g. often retelling or paraphrasing sections of the text rather than using it to support comment*	**In most reading** ▪ straightforward inference based on a single point of reference in the text, *e.g. 'he was upset because it says "he was crying"'* ▪ responses to text show meaning established at a literal level *e.g. "walking good" means "walking carefully"* or based on personal speculation *e.g. a response based on what they personally would be feeling rather than feelings of character in the text*	**In most reading** ▪ a few basic features of organisation at text level identified, with little or no linked comment, *e.g. 'it tells about all the different things you can do at the zoo'*

Source: DCSF, The Standards Site, Primary Framework for literacy and mathematics
http://www.standards.dcsf.gov.uk/secondary/framework/files/downloads/pdf/English_assessment_guidelines.pdf

Appendix Figure 3: KS2 level descriptions in Maths. Assessment guidelines for teachers.

Level 5	• identify and obtain necessary information to carry through a task and solve mathematical problems • check results, considering whether these are reasonable • solve word problems and investigations from a range of contexts • show understanding of situations by describing them mathematically using symbols, words and diagrams • draw simple conclusions of their own and give an explanation of their reasoning	• use a wider range of properties of 2-D and 3-D shapes and identify all the symmetries of 2-D shapes • use language associated with angle and know and use the angle sum of a triangle and that of angles at a point • reason about position and movement and transform shapes • measure and draw angles to the nearest degree, when constructing models and drawing or using shapes • read and interpret scales on a range of measuring instruments, explaining what each labelled division represents • solve problems involving the conversion of units and make sensible estimates of a range of measures in relation to everyday situations • understand and use the formula for the area of a rectangle and distinguish area from perimeter
Level 4	• develop own strategies for solving problems • use their own strategies within mathematics and in applying mathematics to practical contexts • present information and results in a clear and organised way • search for a solution by trying out ideas of their own	• use the properties of 2-D and 3-D shapes • make 3-D models by linking given faces or edges and draw common 2-D shapes in different orientations on grids • reflect simple shapes in a mirror line, translate shapes horizontally or vertically and begin to rotate a simple shape or object about its centre or a vertex • choose and use appropriate units and instruments • interpret, with appropriate accuracy, numbers on a range of measuring instruments • find perimeters of simple shapes and find areas by counting squares
Level 3	• select the mathematics they use in a wider range of classroom activities • try different approaches and find ways of overcoming difficulties that arise when they are solving problems • begin to organise their work and check results • use and interpret mathematical symbols and diagrams • understand a general statement by finding particular examples that match it • review their work and reasoning	• classify 3-D and 2-D shapes in various ways using mathematical properties such as reflective symmetry for 2-D shapes • begin to recognise nets of familiar 3-D shapes, e.g. cube, cuboid, triangular prism, square-based pyramid • recognise shapes in different orientations and reflect shapes, presented on a grid, in a vertical or horizontal mirror line • describe position and movement • use a wider range of measures including non-standard units and standard metric units of length, capacity and mass in a range of contexts • use standard units of time

Source: DCSF, The Standards Site, Primary Framework for literacy and mathematics
http://www.standards.dcsf.gov.uk/secondary/framework/files/downloads/pdf/Mathematics_assessment_guidelines.pdf

Appendix Table 1: Summary statistics for important variables

Variable	Mean	KS2 SD	N
TA English	3.911	0.88	2255382
TA Maths	3.978	0.89	2255382
TA Science	4.119	0.80	2255382
KS English	3.883	1.14	2255382
KS Maths	3.896	1.13	2255382
KS Science	4.236	0.90	2255382
Binary Variables			
White	0.841	0.37	1896966
Black Caribbean	0.015	0.12	32902
Black African	0.016	0.13	36834
Black Other	0.005	0.07	10930
Indian	0.022	0.15	48732
Pakistani	0.026	0.16	59501
Bangladeshi	0.010	0.10	22328
Other Asian ethnicity	0.004	0.07	10116
Chinese	0.003	0.06	7009
Mixed White and Black Caribbean	0.008	0.09	17536
Mixed White and Black African	0.002	0.04	4287
Mixed White and Asian	0.004	0.06	9294
Mixed Other	0.007	0.09	16590
Other Ethnic Group	0.012	0.11	26255
No Special Educational Needs	0.776	0.42	1751106
Special Educational Needs, no statement	0.195	0.40	438731
Special Educational Needs, with statement	0.029	0.17	65463
No Free School Meals	0.830	0.38	1871184
Free School Meals	0.170	0.38	384030
English as first language	0.905	0.29	2040477
English as Additional Language	0.095	0.29	214016
Male	0.509	0.50	1147608
Female	0.491	0.50	1107774

Note. The sample was taken from academic years 2001/2002 to 2004/2005, and includes only those with both TA and KS results. SD stands for Standard Deviation. TA stands for Teacher Assessmen;, KS for Key Stage Test. The expected level of attainment at KS2 is level 4.

Appendix Table 2: Probability of TA<KS, English. Dependent variable is binary, equal to one if TA<KS.

	Specification 1		Specification 2		Specification 3		Specification 4		Raw mean
Variable	β	t stat	β	t stat	β	t stat	β	t stat	% TA<KS
KS2 score	0.066	171.65	0.105	190.13	0.108	192.03	0.110	199.10	
Ethnic Group									
Black Caribbean	0.040	12.87	0.027	8.68	0.027	9.52	0.025	11.39	17.2
Black African	0.048	14.56	0.019	5.87	0.018	6.41	0.017	7.15	17.8
Black Other	0.035	8.53	0.016	3.99	0.014	3.53	0.012	3.36	17.3
Indian	0.017	4.81	-0.006	1.65	-0.018	5.35	-0.018	7.61	17.1
Pakistani	0.059	15.46	0.027	6.47	0.015	3.70	0.010	3.66	18.1
Bangladeshi	0.046	8.95	0.004	0.84	-0.001	0.21	0.002	0.71	17.8
Other Asian ethnicity	0.044	9.36	0.028	5.87	0.019	4.17	0.017	4.22	18.6
Chinese	0.007	1.45	-0.015	3.00	-0.017	3.49	-0.019	4.15	17.1
Mixed White and Black Caribbean	0.028	9.03	0.020	6.52	0.013	4.51	0.013	4.65	17.3
Mixed White and Black African	0.013	2.21	0.004	0.77	0.003	0.59	0.004	0.74	16.2
Mixed White and Asian	-0.011	2.85	-0.014	3.64	-0.018	4.71	-0.015	4.19	15.0
Mixed Other	0.013	3.99	0.003	1.02	0.000	0.01	-0.001	0.29	16.8
Other Ethnic Group	0.044	14.77	0.020	6.82	0.018	6.34	0.017	6.63	17.9
Missing	0.013	4.67	0.008	2.99	0.007	2.68	0.007	3.02	15.7
Reference group: White									15.0
Special Education Needs (SEN)									
SEN, without statement			0.144	127.82	0.146	131.69	0.150	145.56	17.1
SEN, with statement			0.235	118.97	0.230	112.72	0.233	117.41	9.1
SEN, missing			0.045	0.62	0.051	0.70	0.021	0.21	18.3
Reference Group: No SEN									15.1
Free School Meals (FSM)									
FSM			0.043	45.01	0.042	47.90	0.036	49.65	15.8
FSM, missing			0.073	2.31	0.081	2.47	0.068	1.98	13.6
Reference Group: No FSM									15.3

English as an additional language (EAL)								
EAL		0.037	16.26	0.038	18.10	0.040	23.34	17.8
EAL, missing		-0.006	0.30	-0.003	0.15	0.002	0.07	13.4
Reference Group: No EAL								15.1
Female		-0.015	27.84	-0.015	28.79	-0.016	30.54	15.3
Reference Group: Male								15.4
Pupil took KS2 in 'wrong' year		-0.007	1.32	0.022	3.47	0.023	3.72	13.5
Number of pupils in school (per 1000)				0.026	3.67			
Type of School								
Selective				-0.049	2.37			6.9
Urban				0.011	4.81			15.5
Foundation				-0.012	2.15			15.3
Voluntary Aided				-0.025	2.32			15.1
Voluntary Controlled				-0.020	1.93			14.7
Christian				0.012	1.18			14.8
Roman Catholic				0.011	0.99			15.2
Muslim				-0.084	5.52			16.5
Jewish				-0.016	0.82			14.0
Other Faith				0.032	1.42			16.7
Year fixed effects	No		Yes		Yes		No	
LA fixed effects	No		No		Yes		No	
R^2	0.044		0.071		0.078		0.074	
Number of Observations	2255382		2255382		2227352		2255382	
Number of Schools	16550		16550		15719		16550	

Note. The sample was taken from academic years 2001/2002 to 2004/2005, and includes only those with both TA and KS results. Pupils took KS2 exam in 'wrong' year if they are not in the correct school year according to their date of birth. OLS regressions were run with standard errors clustered by school.

Appendix Table 3: Results under different restrictions/robustness checks.

	One class only		Non SEN only		Level 4 only		5 plus only		MOSIAC	
Variable	β	t stat	β	t stat	β	t stat	β	t stat	β	t stat
KS2 score	0.109	140.69	0.136	162.30					0.112	198.37
Ethnic Group										
Black Caribbean	0.022	5.47	0.032	12.20	0.0259	8.32	0.027	7.91	0.024	10.76
Black African	0.009	2.17	0.025	8.75	0.0162	4.91	0.005	1.40	0.015	6.29
Black Other	0.004	0.72	0.019	4.44	0.0106	2.11			0.011	3.13
Indian	-0.023	4.77	-0.028	10.55	-0.0178	5.72	-0.008	1.64	-0.016	6.61
Pakistani	0.007	1.29	0.007	2.24	0.0209	5.72	0.011	1.74	0.011	3.99
Bangladeshi	0.001	0.15	-0.004	1.02	0.00457	0.99	-0.007	0.67	0.002	0.67
Other Asian ethnicity	0.020	2.61	0.016	3.54	0.0148	2.70			0.017	4.10
Chinese	-0.022	2.74	-0.027	5.52	-0.0134	2.11			-0.020	4.24
Mixed White and Black Caribbean	0.007	1.58	0.016	5.16	0.0105	2.84			0.011	3.80
Mixed White and Black African	0.013	1.40	0.001	0.23	0.00493	0.69			0.003	0.49
Mixed White and Asian	-0.022	3.62	-0.020	5.01	-0.00781	1.71			-0.014	3.91
Mixed Other	-0.001	0.29	-0.001	0.27	-0.00384	1.05			-0.002	0.78
Other Ethnic Group	0.010	2.04	0.022	7.55	0.0150	4.13			0.017	6.59
Other personal characteristics?										
School characteristics?										
LA fixed effects?										
School fixed effects?										
R^2	0.074		0.068		0.070				0.075	
Number of Observations	702995		1751112		1139391				2226478	
Number of Schools	11010		15808		15585				16497	

Note. The sample was taken from academic years 2001/2002 to 2004/2005, and includes only those with both TA and KS results. OLS regressions were run with standard errors clustered by school. 'One class' restricts the sample to schools with <=35 pupils. 'Non SEN' restricts the sample to those pupils with no special educational need. 'Level 4 only' restricts the sample to pupils that got a KS2 English test level of 4. KS2 is therefore no longer needed as a control. '5 plus only' runs separate regressions for each ethnic group, but are reported together here for comparison. In these regressions, the sample includes only white students and students of the ethnic group, in schools where there are >=5 of each. Sample sizes are 60533, 71337, 91362, 85779, and 21906 respectively. MOSAIC includes detailed neighbourhood characteristics in the regression in addition to the free school meals indicator.