Assessment, Achievement and Participation

Stephen Gibbons*, Arnaud Chevalier**

December 2007

*Department of Geography and Environment and Centre for Economic Performance, London School of Economics.
** Department of Economics, Royal Holloway University London; Geary Institute, University College Dublin; Centre for the Economics of Education, London School of Economics and IZA, Bonn

Acknowledgements:
The project was financed by ESRC under the TLRP programme at the Centre for the Economics of Education, London School of Economics.

Abstract

Systematic divergence between face-to-face, teacher-based assessment and blind, test-based assessment can be indicative of biases in an assessment system. Such discrepancies, along pupil demographic strata, have been used in the past as evidence of statistical discrimination or social stereotyping by assessors. Moreover, teacher perceptions of pupils' abilities could influence pupils' subsequent educational outcomes and schooling participation through a number of channels, so errors in perceptions could have important consequences. We consider these issues in relation to the teacher and test-based assessments in England's National Curriculum system. Although we find evidence that teacher and test assessments diverge slightly along lines of ethnicity, gender and, to a greater extent, ability, it does not appear that this discrepancy arises through stereotyping. Moreover, the divergence between test and teacher assessments has almost no bearing on subsequent pupil outcomes.

Keywords: statistical discrimination, stereotyping, assessment, education

JEL Classifications: I2

1. Introduction

Pupil assessment plays a central role in modern schooling systems, informing teaching and learning, and facilitating school leadership and governance. It is therefore essential that assessment systems are fair, valid, reliable and suited to purpose. Ongoing formative assessment that provides teachers, parents and pupils with information about pupil skills and weaknesses is part of any education process and is rarely considered controversial. But the efficacy of summative, 'high-stakes', snapshot testing is often called into question, because of the pressure it places on teachers and pupils, and because of questions about its reliability and validity, particularly when there are incentives for schools to teach to the test. The system of national curriculum testing in England is frequently criticised along these lines, often alongside arguments for a move away from blind marked short tests towards the use of continuous teacher-based assessment as the basis for summative assessment (Brooks and Tough 2006).

Teacher assessment and test-based assessment each has its own advantages and disadvantages. Blind marked tests can be considered 'objective' in that they are marked by examiners who do not know the candidates or their demographic characteristics. However, these tests can only evaluate performance on a very limited range of questions on a specific day and can favour technique over underlying skill. Moreover, test-based assessments could mis-measure ability because the specific test was not well designed to the pupil's gender, ethnic or cultural background. Conversely, teacher assessments are usually based on observation of ability on a wider range of tasks over a longer time horizon, but may be sensitive to personal and subjective preferences on the part of the teacher, or the specific relationship and interaction between pupil and teacher. In particular, any non-blind assessment could be subject to some form of statistical discrimination or stereotyping whereby judgements are made on the basis of what is expected of pupils of similar type, rather than a pupil's personal abilities. Importantly, previous research has suggested that teachers' assessments of pupil's academic abilities do tend to differ from pupils' achievements in tests and exams in ways that are systematically related to ability, demographic and socioeconomic background (see for example Gipps and Murphy 1994, Reeves et al 2001 for England, Lavy 2004 for Israel). This is a worrying finding: since it implies that some groups can be educationally disadvantaged simply by the type of assessment to which they are exposed.

Divergence in assessments gives particular cause for concern if the tests are unbiased, and the divergence arises because face-to-face interaction influences teacher perceptions of pupil ability. Indeed, a growing empirical international literature suggests that the same teachers do not always judge pupils of different backgrounds to the same standards (Ouazad 2007, Dee 2005a). The issue is important, even if teacher-based assessments are not used for high-stakes summative assessment, because teachers guide pupils in curriculum choices and exam entries and because pupils' (and their parents') motivation may respond in more subtle ways to views of their own ability. Consequently, any divergence between teacher perceptions and test-based measures of achievement along lines of gender, ethnicity and social class, could offer at least a partial explanation for attainment gaps and differences in higher education participation patterns between these pupil groups (e.g. DfES 2003, Conner et al 2003, 2004 for England).

As well as looking again at the issue of teacher-test discrepancies in assessment, this paper provides the first systematic attempt to measure whether such discrepancies really have any influence on pupils' subsequent educational decisions and achievements. The research uses large-scale administrative datasets on England's population of school pupils in various cohorts aged 11-16 from 1997-2004, linked to information on post-16 educational participation. This linked database details the academic records and background of around two million pupils, with information on the location and characteristics of their schools and place or residence and their post-16 educational decisions.

In common with previous research, our empirical work finds evidence that teacher assessments and test scores estimate 'ability' in ways that diverge according to pupil background. Most strikingly, teacher assessments tend to favour boys over girls in English, relative to their test scores at age 14, but favour girls over boys in maths and science at age 14, and also in maths at age 11. Non-white ethnic groups and English-additional language pupils also tend to do better in age-14 English tests than would be predicted from teacher assessments. In maths and Science, there are very few systematic and stable differences between teacher assessments and test scores along lines of ethnicity. However, in all subjects, teacher assessments tend to have lower variance than actual test results which means that high achievers tend to do better, and low-achievers worse in tests than would have been expected from teacher assessments. Our results suggest that this is not because the tests are relatively noisy but because teachers tend to avoid extremes in their assessments and genuinely overestimate low-achievers and underestimate high-achievers. We find very little evidence that the divergence between teacher and test based scores are the result of any form of 'rational' statistical discrimination on the part of teachers, which would imply systematic over-assessment of high achieving groups and under-assessment of low-achieving pupil groups.

The second key question addressed by the paper is whether divergence between teacher and test-based assessment could have any bearing on pupils' subsequent academic achievement or staying on decisions. We find that teacher assessment and test-scores do provide unique information about pupils' subsequent outcomes and achievement and both appear to be noisy measures of underlying ability and propensity for success in education. However, in itself, divergence between test-based and teacher-based assessment has no systematic adverse consequences for subsequent educational outcomes. In fact, pupils who do better in tests at age 14 than teachers expect, tend to do better in their GCSEs and are more likely to stay on in education, probably because test-based measures provide a marginally stronger predictor of success along these educational lines.

In the next section (2), we consider briefly the existing literature on teacher-based, and test-based assessments and their biases in relation to gender, ethnicity and socioeconomic group. Following that, in Section 3, we describe the methods we use when analysing the relationship between pupil characteristics and the disparity in assessments, and in measuring the links between these disparities and subsequent educational attainment. Next, Section 4 explains the institutional context and the data used in the analysis and in Section 5 we discuss the empirical results. Section 6 provides conclusions.

2. Explaining assessment gaps and their consequences

All forms of assessment are imperfect and measure the object of the assessment with a degree of error. Face-to-face assessment and blind, test-based assessment can elicit different answers for a number of reasons. Moreover, the duration of testing may differ between the two assessments formats from a high stake short exam to a year long judgement that involves different types of evaluation. For example, the two types of assessment could be designed to capture different dimensions of pupil ability. Secondly, and relatedly, the assessments could be designed to measure the same thing, but a pupil is better at doing one type of assessment than the other. These differences are related to the design of the test or particular skills of the pupil. Another class of explanations are based on the characteristics, or preferences or beliefs of the person making the face-to-face assessment (e.g. a teacher) or to the interactions between that person and the individual that they are assessing. Typical explanations involve some form of ‘stereotyping’ or statistical discrimination by the assessor along ethnic or gender lines: an assessor treats an individual as an exemplar of their group (e.g. ethnic or gender) and makes judgements based on what they believe to be true about the group. An alternative story involves a pupil response to the assessor, giving an impression of poor achievement, but scoring well when tested. This may be because of ‘stereotype threat’ whereby the pupil behaves differently to the assessor because they believe the assessor is likely to engage in stereotyping. This behaviour could generate ethnic, gender or socioeconomic group differences if such tendencies are specific to particular pupil groups or of they are dependent on the match between pupils and teachers[1]. The stereotyping could also be on the student side. Steele and Aronson (1995) for example, show that students from visible minorities perform less well at tests in which their minority is not expected to do well but no difference is observed when the minority is not known to be a poor performer at this task. Yet another possibility is that the degree of 'teaching to the test' differs across pupil groups. A fairly large literature in the fields of education, psychology and educational psychology theorises on these kinds of issues and offers some empirical evidence, usually based on relatively small scale survey data (see Wright and Taylor 2007 for a way into this literature).

One thing that is fairly clear from this discussion is that it is probably impossible to separate out the precise channels of causality in any empirical analysis of divergence between test and face-to-face assessments, without some strong theoretical assumptions. Nevertheless, researchers in the economics of education have, with this hope in mind, recently started to approach the empirical questions surrounding stereotyping, statistical discrimination, and ethnic and gender biases using econometric tools on relatively large scale datasets. Lavy (2004), for example, investigates divergence between blind and non-blind assessments in high schools in Israel and points to this divergence as evidence of gender stereotyping against boys. It is difficult, however, to distinguish this claim from one in which the design of the blind tests favours boys over girls. A number of other econometric studies have considered the effect of pairing pupils with teachers of the same gender or same ethnicity. This approach focuses on

[1] For instance, if boys are more likely to act foolishly with women teachers, then male teachers may appear to rate boy pupils favourably when compared to women teachers and women teachers may appear to rate boys unfavourably compared to tests. Equally, if most teachers are women, then boys may appear to be unfairly rated by teachers when compared to girls in the population.

answering the question of whether, say, black children receive more favourable scores when assessed by black teachers than they do when assessed by white teachers. The studies by Dee (2005a, 2005b) and Ouazad (2007) use panel data to eliminate fixed-over-time individual (and even teacher) effects by observing how evaluations vary for a given pupil as he or she moves between teachers of different genders and ethnic groups (and observing how the scores a teacher awards vary according to the gender or ethnic group of each pupil they teach). Dee (2005a, 2005b) uncovers evidence that pupils have lower achievement and are more likely to be rated as inattentive, disruptive and less likely to complete homework if the teacher is of a different gender or ethnic group. Ouazad (2007) improves on this work by comparing teacher and test assessments and finds evidence effects arising from interaction between pupil and teacher ethnicity, but no gender effects. In many cases, it is difficult to disentangle whether it is the teacher's perception of the pupil or the actual pupil behaviour that varies according to the pupil-teacher match in ethnicity or gender. Indeed, Ammermueller and Dolton (2006) find similar evidence that pupil's test scores are higher when they are taught (but not assessed) by teachers of the same gender.

An important omission in considering only different-sex and different gender effects on teacher assessments is that it runs the risk of missing out on other sources of assessment bias which may not be linked to the gender or ethnic pairing of teacher and pupil. For instance, standard theories of statistical discrimination and stereotyping do not argue that the judgements are made because of *differences* in gender or ethnicity between assessor and person being assessed. According to these theories (Phelps 1972, Tajfel 1959) misjudgements are made because the assessor treats the individual as a representative of the group and bases their judgements on what they expect of individuals of a given type, rather than on the individual's personal qualities and aptitude.

All the prior literature has been concerned with evaluating and explaining systematic divergence between blind and non-blind assessment or finding evidence of gender or ethnic bias in teacher assessments. The main motivation for being concerned about this assessment divergence is, presumably, that it may have real consequences for children's subsequent outcomes. However, as far as we know, no previous research has attempted to answer this rather important question. In the next section we explain our approach to modelling teacher-test assessment gaps, and to investigating the consequences for pupil attainment and education participation.

3. Methodological discussion

3.1. Pupil characteristics and assessment disparities

Our goal is to explore, empirically, how gaps between teacher-based and test-based assessment of pupils' levels of achievement differ along ability and demographic lines, in the context of England's secondary education. The empirical approach we take is based on the assumption that both teacher-based and test-based assessments generate scores that try to measure the 'ability'of a pupil. We define 'ability' here as an unobservable attribute mapping a pupil's aptitude in a fairly specific set of tasks in a specific set of academic subject areas (literary, mathematical, scientific etc.). Our interest is, firstly, whether assessments of ability generated by tests and by teacher observation diverge systematically according to the ethnic and socioeconomic characteristics of pupils, in ways that are unrelated to their actual ability. Secondly we want to know if any such divergence impacts on pupils' subsequent achievement or propensity to continue in education.

There are many serious challenges to both of these ambitions when true ability is unobserved. Without further information, the outcome of an under-rating by a teacher and the outcome of a lucky strike on a test are observationally equivalent: a pupil does better on the test than they do in the teacher's assessment. We cannot tell which of these measures is right. Similarly, we cannot judge whether any systematic difference in teacher-based and test-based scores – for example by gender – arises because of systematic under-assessment on the part of teachers or systematic gender bias in the tests. These judgements can only be made by *a priori* assumptions about which of the assessments – test or teacher based – is more accurate. Since we do not wish to make such assumptions, we will focus in the first part of our empirical work on exploring the association between pupil characteristics and the gap between teacher and test based scores without a view as to which score is the most accurate. This exploration will be based on simple linear models, estimated by ordinary least squares, in which we regress the difference between teacher and test based assessments on a set of observable pupil and school characteristics[2]. This specification is equivalent to that set out in Lavy (2004) for estimation of gender biases in assessment systems in Israel. It eliminates fixed pupil characteristics that have identical effects on both tests and teacher-based scores, and highlights characteristics that have differential effects on these scores. Here, we consider the assessments of pupils' English, science and maths ability age 14 in England – referred to as the Key Stage 3 tests.

A key additional question is to what extent the teacher-test gap varies across the distribution of achievement levels, or 'ability'. This question is important in its own right, because a systematic trend in the gap between high ability and low ability children could suggest some structural problems with the assessment system. It is also important because groups differ in terms of their average achievement, so it is easy to confuse a systematic divergence between test and teacher assessment for particular group (low income, for example) with a systematic

[2] Unfortunately our data (discussed in detail in the next section) does not provide us with information on teacher characteristics, so this rules out exploration of the effects of being matched to a teacher of the same sex or ethnicity. The results are therefore only informative about expected outcomes for pupils of different types, conditional on the distribution of teacher characteristics in the population in England.

divergence along lines of average ability. But if ability is unobservable, how can we distinguish between these two cases empirically? The approach we take is to impose a normalising assumption that teacher and test-based assessments are ‘symmetrically’ biased either side of the true level of ability, given that we cannot prejudge which, if any, of the assessment modes is unbiased. Under this assumption, the average or sum of teacher and test-based assessment scores provide an unbiased (though noisy, in the sense that it contains random error) estimate of a pupil’s unobserved underlying ability. Hence, by including the *sum* (or mean) of teacher and test-based assessments as an explanatory variable in our models of the test-teacher *gap* in assessment, we can examine how the gap varies with both observable pupil characteristics and with levels of achievement.

However, if we use this strategy to look at the relationship between the gap and the sum of a pupil’s assessment based on the same pair of teacher and test assessments, the interpretation is not clear-cut. A negative relationship in a regression of teacher-test gaps on the sum of teacher and test scores implies only that the variance of teacher assessment scores is lower than that of test scores. This relationship could occur if teachers tend to underrate high performers and overrate low performers relative to the tests, but could also mean that tests contain more random error (noise). Conversely, a positive relationship between the gap and the sum of assessments could arise because the tests systematically under-rate high achievers or because the teacher assessments introduce more random error than the tests.

These distinctions are subtle, but important: A relationship between gaps and achievement levels attributable to random noise is an artefact of the fact that below-average pupils have a higher probability of getting a low score on the high-variance assessment, and above average pupils have a higher probability of getting a high score on the high variance assessment. This fact does not imply that either the design of the tests or the behaviour of teachers is systematically biased in favour of either high or low ability pupils. A potential solution to this problem is to estimate unobserved pupil ability based on the sum of teacher and test based assessments *during some earlier phase of assessment*, using different tests and different teachers. Any relationship between teacher-test gap and estimated ability based on past assessments is more likely than not to be linked to systematic mis-assessment of high or low ability children rather than random assessment error. This claim will be true under the assumption that the random errors in assessment introduced in the past teacher and test assessments are uncorrelated with the random errors introduced in current teacher and test assessments. Note that the data we bring to bear on this question records test and teacher based assessment scores for age-14 pupils in secondary school, and test and teacher based assessment scores for the same pupils aged-11 in primary schools. Since the assessments are made by different teachers, and at a different phase of education, the claim that the random errors in the assessments are uncorrelated across phases is fairly plausible. In any case, we can also include the teacher-test gap in the past assessment as an explanatory variable in our regressions, to control for differences in errors (bias or noise) between past teacher and test assessments on which the estimate of ability is based. These issues are set out more formally in the Technical Appendix.

3.2. Links between assessment disparities and later outcomes

The second objective of this paper is to consider whether the gaps between scores produced by different assessment methods influence pupils' subsequent educational attainment and the decision to stay on after compulsory schooling age. As discussed in the introductory sections, the underlying hypothesis behind any such relationship is that differences between teacher expectations of achievement (as reflected in teacher assessments) and the styles of assessment that forms the basis for school-leaving qualifications, could a) influence subsequent academic qualification and prospects of continuing in education because of teacher influences on the number, type and mix of qualifications attempted b) feed through to pupils' or parents personal expectations of achievement and the decision to continue in education. It will not, using the national administrative data that forms the basis of this study, prove possible to uncover the precise channels through which teacher-test assessment gaps act on subsequent outcomes. The less ambitious objective of the empirical work presented here is just to uncover evidence for the existence of such a link

Our approach to this task is to use least squares regression models to estimate the relationship between the teacher–test gaps on a pupil's age-14 assessments and various academic outcomes in the next phase of pupils' academic careers. These outcomes relate to qualifications (GCSE/NVQs) taken at minimum school leaving age (age 16) and to the decision to stay on at school or participate in other forms of education in the age 16-18 period. We also consider the mix of subjects in which pupils sit exams at age 16, in order to explore whether disparities in assessment in particular subjects could discourage further study in maths, science or English. All these outcomes are important factors in the subsequent decision to participate in higher education, and the type of higher education undertaken.

In the next section we describe in more detail the institutional context and data used in our empirical analysis.

4. Data and context

Compulsory education in state schools in England is organised into five "Key Stages". The Primary phase, from ages 4-11 spans the Foundation Stage to Key Stage 2. At the end of Key Stage 2, when pupils are 10-11, children leave the Primary phase and go on to Secondary school where they progress through to Key Stage 3 at age 14, and to Key Stage 4 at age 16. At the end of each Key Stage, prior to age-16, pupils are assessed on the basis of standard national tests and at age 16 pupils sit GCSEs (academic) and/or NVQ (vocational) tests in a range of subjects. After compulsory schooling ends at age 16, pupils can continue their education in school or at college, or sometimes in the workplace, studying for academic and/or vocational qualifications.

Those who gain suitable qualifications can, at age 18-19, enrol in Higher Education, usually at a university.

The UK's Department for Children, Schools and Families (DCSF[3]) collects a variety of data on state-school pupils centrally, because the pupil assessment system is used to publish school performance tables and because information on pupil numbers and characteristics is necessary for administrative purposes – in particular to determine funding. A National Pupil Database exists since1996 holding information on each pupil's assessment record in the Key Stage Assessments throughout their school career. Assessments at Key Stages 2 and 3 (ages 11 and 14) include a test-based component and teacher assessment component for three core curriculum areas: maths, science and English[4]. As set out in the statutory information and guidance on Key Stage 3 assessment: "The tests give a standard snapshot of attainment in English, mathematics and science at the end of the key stage. Teacher assessment covers the full range and scope of the programmes of study. It takes into account evidence of achievement in a variety of contexts, including discussion and observation" (QCA 2004). Importantly, the tests and teacher assessments are intended to measure ability, knowledge and skills along the same dimensions in the same subject areas. Since the teacher assessment is based on several measurements we may expect the variance in teacher assessment to be lower than at key stage examination.

For each subject, the teacher assessments and tests award the pupil an achievement Level on a discrete scale ranging from Below Level 1 up to Level 5 at Key Stage 2, and up to Level 7 (8 in maths) at Key Stage 3. These levels are converted into Points-based system which assigns 6 points to each Level and we work with these Points in our empirical analysis. In particular, our

[3] Until 2007, the Department for Education and Skills (DfES)

[4] We work with the overall assessment in these subjects, which is derived from various component tests.

definition of the teacher–test assessment gap is the difference in points awarded by the teacher in their assessment and the points awarded by examiners.

Since 2002, a Pupil Level Annual Census (PLASC) records information on pupil's school, gender, age, ethnicity, language skills, any special educational needs or disabilities, entitlement to free school meals and various other pieces of information including postcode of residence (a postcode is typically 10-12 neighbouring addresses)[5]. PLASC is integrated with the pupil's assessment record (described above) in the National Pupil Database (NPD), giving a large and detailed dataset on pupils along with their test histories. Tracking of pupils continues after age 16 in an integrated database of age-16-18 education that is derived from PLASC, a database called the Independent Learner Record, and from other sources.

From these sources we derive two extracts for use in our estimation. The first follows four cohorts of children from their Key Stage 2 assessment at age 11, to their Key Stage 3 assessment at age 14 in 2002-2005. The second follows the academic careers of three older cohorts of children from age-11 through to age 16 in 2002-2004, and then on to the point where they have made their post-age-16 educational choices. The first of these two extracts draws on pupil characteristics at age 14 as a basis for analysis of any systematic divergence between test and teacher based assessment. The second extract, recording pupil characteristics at age 16, allows us to explore if past teacher–test assessment gaps (at age-14, Key Stage 3) influence subsequent education decisions and outcomes. Various other data sources can be merged in at school level, including institutional characteristics (from the DCSF). In both data extracts we exclude the 12% of pupils with recognised disabilities and learning difficulties who are registered as having Special Educational Needs, whether in Special schools or mainstream schools[6]. We also focus

[5] Prior to 2002 this information was collected only at school level.

[6] This restriction is intended to exclude children with disabilities or learning difficulties and to homogenise the estimation sample. In many cases the classification of Special Educational Needs can be based on exceptionally low

solely on state Comprehensive schools, that is schools that do not choose pupils on the basis of academic ability, and we do not have data on pupils attending private schools[7]. This large and complex combined data set provides us with information on around 1.4 million children aged 14 in 2002-2005, plus just over 1 million children aged 16 in 2002-2004, with those aged 14 in 2002 represented in both datasets.

5. Results and discussion

5.1. Descriptive statistics

We begin the empirical analysis with a look at the descriptive statistics for the data set, presented in Table 1. As explained in Section 4, we have two core datasets, one based on cohorts of children age 14 in 2002-2005 and another on cohorts aged 16 in 2002-2004. The first dataset is summarised in the top panel Table 1 and is used in our analysis of the associations between pupil characteristics and the gap between teacher and test assessment scores. The second dataset is used to analyse how these assessment gaps affect subsequent outcomes, and is summarised in the lower panel. The table presents means and standard deviations for the full sample, and for various sub-samples.

The first three rows of the top panel give mean aggregated teacher and test scores in each subject, and the group differences in mean achievement can be seen by reading across the columns of the table. As is well known, Asian and black pupils, and pupils eligible for free meals score below the mean in the population in all core subject areas; boys score below girls in

(or occasionally exceptionally high) assessments of ability, without any diagnosed physiological condition, so the sample suffers from some potential selection issues, since children with the lowest teacher assessments may be excluded. In practice, inclusion or exclusion of these special needs pupils does not affect the overall findings.

[7] We exclude selective Grammar schools because most grammar school students will be participarting beyond age-16, so there would be no variation in outcome within schools when we go on to explore post-16 participation. Private schools educate around 6-7% of pupils in England as a whole.

English but slightly higher in maths and science. The bottom three rows of the top panel show the gap between teacher assessment points and the test-based points. A look down column (1) in the top panel shows that, on average over the 2002-2005, the point scores based on teacher assessments were slightly lower than those based on tests, by up to one third of a point in mathematics and English. Looking across the columns provides insights into how these gaps vary according to our socioeconomic, ethnic and demographic groups of interest. Notable features are that ethnic minorities and those with English as an additional language score even lower on teacher assessments in English than the population as a whole, with a gap as high as 0.6 points for Asian pupils. On the other hand, the gap between teacher and test assessments in science and maths is generally larger and more negative for the population (i.e. teachers assessments lower than test assessments) than it is for the ethnic and socioeconomic sub-groups. Boys seem to fare relatively badly in teacher assessments in maths and science and relatively well in English.

The bottom panel shows a range of age 16 and post-16 outcomes, again split by pupil subgroups. Pupils enter 9.8 exams on average at age 16, and whilst there is some variation across groups the differences are not dramatic. There is a lot more variation across groups in terms of their relative position in the distribution of scores from these age-16 exams, and free meal entitled pupils, black pupils and boys have relatively low attainments: the average free meal entitled pupil is at the 37th percentile in the distribution of age-16 qualifications. On the other hand Asian and, interestingly, English additional language pupils gain better qualifications than average. Post-16 participation rates follow a similar pattern, with high post-16 participation and staying on rates for Asians and those with English as an additional language. A high proportion (85.4%) of black pupils participate in post-16 academic education, but only 30.7% do so in school. Boys score below girls in their GCSEs, and are less likely to continue in academic education, either in school or elsewhere. The subject shares do not differ widely between demographic groups, but there is considerable within group variance.

These descriptive statistics reveal some interesting features in the data. The top panel in particular suggests that there are systematic differences between teacher and test-based assessments, and that these differences vary along ethnic, socioeconomic and gender lines. In Section 5.3 we extend this analysis using a regression models to explore the separate contribution of each of these pupil characteristics, and to control for pupils' achievement levels. First we look for evidence for stability in the assessment gaps over time and across different teachers, exploiting the fact that we have two periods (age 11 and age 14) and multiple tests (maths, science and English) in which we can compare tests and teacher assessments for a given pupil.

5.2. Correlation patterns in teacher and test based assessments

If the divergence between assessment methods, teacher and test-based, is generally and systematically related to fixed pupil characteristics then we can expect to see quite strong correlations between the divergence in scores for the same pupil in different subjects, and between the divergence in scores for the same pupil in different time periods. To provide a first insight into this question, Table 2 presents the Pearson's correlation coefficients between the teacher-test gap in points assigned to a pupil in each core subject, at ages 14 (the main focus of our analysis) and at age 11. At age 14, different subjects are usually taught by different teachers. Hence, any correlation across subjects at this age will reflect a tendency for similar teacher-test gaps for a given pupil, regardless of which teacher is making the assessment. As it turns out, the correlations between the assessment gaps at age 14 are very small. The highest correlation at age-14 is between science and maths assessment gaps, with a correlation coefficient of 0.081. These low figures immediately suggest that there is a relatively weak tendency for the gap in assessment at age 14 to vary according to *any* pupil characteristic at that age, and much of the gap may be pure noise; but the fact that there is any association at all is interesting. Example explanations are that: a) a given pupil exhibits characteristics or behaviours that lead teachers in different subjects to under-assess (e.g. if disruptive) or over-assess (e.g. if a convincing oral communicator) the pupil relative to his or her actual ability, and this ability is better measured in a test-based setting; b) that a given pupil is accurately assessed by teachers, but systematically gets test results that do not reflect his or her ability – for example if prone to make mistakes under stress, or if dedicated to practising example test papers to achieve good scores without real understanding.

Further insights can be had by looking at the pattern at age 11, and at the association between age-11 and age 14 scores. Firstly, the correlations between discrepancies in different subjects at age 11 are much higher, up to 0.172 in maths and science. The most obvious explanation for this finding is that a pupil is assessed by the same teacher in all the core subjects at age 11, so the correlation across subjects is strongly influenced a particular teacher's view of the pupil. In other words, it would appear from comparison of the age 11 and age 14 correlations that a large part of the divergence between teacher and test scores has something to do with the interaction between teacher and pupil, and not to do with differences in test skills. No other explanation easily accounts for the correlation between assessment gaps corresponding to different teachers being lower than the correlation between assessment gaps corresponding to the same teacher. This view is reinforced by considering the tiny – a maximum of 0.019 – correlations found between the subject gaps at ages 11 and 14, at which ages different teachers and different schools would be involved. At face value, this very low correlation for a given subject across ages suggests that fixed characteristics of the pupil that were the same at age 11 and age 14 – gender, ethnicity, social class, for example – are unlikely, on their own, to provide an important explanation for discrepancies between test and teacher scores at both ages.

5.3. Regression estimates of group divergence in teacher and test based assessments

These simple correlation patterns are, however, uninformative about the types of pupil characteristic that lead to divergence between test and teacher scores in these core subjects. For this analysis, we turn to the regression approach outlined in Section 3.1. Firstly Table 3 presents the coefficients and standard errors from regression models of teacher and test based assessment scores. The explanatory variables are pupil ethnic, demographic and socioeconomic characteristics (plus other control variables as described in the table notes). In columns (1), (3) and (5) we show the association between our basic set of pupil characteristics and pupil achievement, as measured by the sum of teacher and test assessment points – an unbiased measure of pupil abilities under our modelling assumptions. The coefficients in this column show how achievement at age 14 is related to these characteristics, and are provided as a benchmark with which to compare the gaps reported in the remaining columns. In column (2), (4) and (6) the dependent variable is the gap in points between teacher and test assessments, so the coefficients indicate the relative bias in the different modes of assessment for each pupil group.

Looking at Table 3 columns (1), we can see the a familiar story of relatively low achievement in English amongst those on free meals, black ethnic minorities, boys and those with English as an additional language. Older pupils, and pupils of mixed ethnicity and other ethnic groups (non white, black, asian or mixed) achieve relatively well. Now, column (2) illustrates how the gap between teacher scores and test scores varies according to these pupil characteristics. Note, the absolute gaps for any group must be calculated by adding up the constant term and the appropriate coefficients. More generally, the table can be interpreted by reading a minus sign on the coefficient as showing that the corresponding group tends to do relatively poorly in the teacher assessments and relatively well in the tests, referenced to the gap for the baseline group of non-free meals, white, girls with English as their first language. It is evident from the negative coefficients on the ethnic minority variables that ethnic minorities tend to have lower teacher assessments and higher test scores as compared to white pupils. Conversely, pupils on free school meals score higher on teacher assessments relative to tests compared to non-free meals pupils. Also boys, compared to girls, do relatively well on teacher assessments. Pupils with English as an additional language also do relatively poorly on teacher assessments, as do older children. The gaps are not particularly large in absolute terms, amounting to plus or minus one-quarter of a point at age-14. This gap is around 4% of one standard deviation in pupil scores. Note, however, that the differences are not trivial when viewed in the context of the gaps in achievement in column (1). Columns (3)-(6) repeat this analysis for science and maths scores at age 14. As for English there are strong differences in achievement along socioeconomic, ethnic and gender lines. However, the ethnic differences in the gap between teacher and test assessments are less marked than for English with the only significant difference indicating that blacks pupils do relatively well in teacher assessments in maths and poorly in tests in maths when compared to white pupils[8]. Low income pupils on free meals also fare relatively well in the teacher assessments, when compared to non-free meal pupils. Boys on the other hand do score relatively favourably on the tests in both maths and science. However, all the results so far on the teacher-test gap take no account of how the gap in

[8] The effects in the maths and science assessments could be attenuated because the Key Stage 3 tests in these subjects are organised into “tiers” and pupils are assigned by teachers to sit different tests according to their ability. This assignment caps the potential divergence between teacher and test assessment (e.g. a pupil assigned to a test tier covering Level 5-7 has a maximum absolute divergence of 2 levels (6 points) from the teacher’s assessment – assuming the teacher’s assessment is matched to the test tier in which the pupil is placed. Looking at Table 1 it appears that the variance of the gap in maths and science is indeed lower than in English (where there is one test for the full range) but the variance is still substantial. Given the relatively low probability of a divergence of more than two levels in English, it seems unlikely that this issue raises serious issues for our analysis in science and maths.

assessment methods might respond to differences in levels of achievement. In Table 4 we extend the regression analysis to control for pupil achievement levels.

Table 4 column (1) introduces controls for the level of pupil achievement into the model for the teacher-test gap in English. Achievement is first measured by a set of dummy variables categorising the sum of the teacher and test based assessment scores at age 14 (i.e. the dependent variable in Table 3, columns (1), (3) and (5)). Controlling for mean achievement levels in this way, makes little difference to the results on ethnicity, but now free-meals pupils appear no different from non-free-meals pupils, and older pupils no different from younger pupils in terms of the gap between teacher and test assessments. The positive coefficient for males is also reduced substantially once we control for levels of achievement. Apparently, all of the difference for low income pupils and part of the difference for boys is linked to their lower levels of attainment in English combined with a tendency for teacher assessments to exceed test scores at lower levels of achievement. Indeed, the results on the relationship between the teacher-test gap and achievement levels are now the most striking feature: pupils scoring towards bottom of the distribution do much better on the teacher assessments than the tests relative to their peers at the top of the achievement distribution. As an example, pupils scoring 48 points or less on aggregate in test and teacher assessments end up with teacher assessments that are around 3.6 points higher on the teacher assessments than the tests– a very large gap – whilst those at the other end of the distribution with 84 points score 0.75 points lower in teacher assessments than the tests.

In Section 3.1 we explained how such a negative relationship between the teacher-test gap and contemporaneous levels of achievement might arise from these modes of assessment suffering from different error variances. So, to examine whether differential assessment error drives the findings, we replace the dummy variables for current achievement levels with dummy variables for past achievement (at age 11) and an additional control for the gap between teacher assessments and tests at age 11 (column (2)). The general picture is similar, though the variation across achievement levels is less extreme. Taking into account the baseline gap between teacher and test assessments (the constant), it appears that children in the lowest age-11 achievement category score around 0.62 points higher in the teacher assessments than the tests, whilst children in the highest category score about 0.54 points lower in teacher assessments than tests. The implication of these results is that there is a general tendency for teachers to be fairly conservative in their ratings of pupil relative to the tests, rating lower ability pupils above their test scores, and rating higher ability pupils below their test scores. Switching the controls from current to past achievement levels makes little difference to the findings on ethnicity, language or free meal entitlement, though in this preferred specification older pupils score relatively high on the teacher assessments.

The next column (3) allows for school fixed effects i.e. the regressions are based on within-school variation only (the results are similar if control for residential neighbourhood fixed effects using the Census 2001 output area). The relationship between ethnicity and teacher-test gaps is reduced, with lower coefficients and reduced t-statistics. Only the "Asian" coefficient remains strongly significantly different, and the "black" coefficient is driven to zero and statistical insignificance. These changes imply that much, though not all, of the variation in teacher-test gaps across ethnic groups is linked to differences between schools (and hence also between teachers) rather than differences for pupils of different ethnicity in the same school (or taught by the same teacher). This result is important: the implication is that some schools generate wider gaps between teacher and test-based assessments than others, and ethnic minorities are more likely to be in schools that generate wide gaps. In particular, black pupils do not score any lower on teacher assessments relative to tests than do white pupils in the same school; the significant effects found in other columns are driven by comparison of black pupils with white pupils in different schools taught by different teachers. Note though, that gender differences, differences across achievement categories, and the difference for Asian pupils relative to whites remain significant in these estimates, implying that there are systematic differences between these pupil groups in the same school.

The remaining columns of Table 4 show the corresponding figures for age 14 science and maths scores respectively. In both cases it is only along gender, age and achievement lines that we see stable systematic differences in the gap between teacher and test based assessment, with boys doing better in the tests than in teacher assessments, and older pupils performing relatively poorly in the tests and higher in the teacher assessments. As for English, the strongest finding is that lower achieving children tend to be over-assessed by teachers and higher achieving children under assessed by teachers, relative to their performance in the age 14 tests. Evidence of important ethnic differences is quite weak and the coefficients are not robust to different specifications. If anything, in column (6) and (9), black and Asian pupils seem to score better on teacher assessments and less well on the tests when compared to their white school mates.

Table 5 repeats some of the analysis for the age 11 test scores of this same group of pupils. At this age there are few systematic differences across ethnic groups. Only a few features suggest any stable tendency for divergence in assessment methods across groups: black pupils do relatively well in teacher assessments in maths and science, pupils with English as an additional language do relatively badly in teacher assessments in English and boys come out badly in teacher assessments in maths. The last two findings echo those in Reeves et al (2001) for a much smaller sample from the same assessment system in 1998[9]. The differences across achievement levels are also strong, with some tendency to lower teacher assessments and higher test scores in higher achievement groups, but the pattern is not nearly as clear cut as at age 14.

On balance, looking back at the tables so far, the biggest systematic differences are found across achievement groups rather than ethnic or socioeconomic group, although there are some

[9] The cohorts in our age-14 data took their age-11 assessments in 1999-2002

strong ethnic differences in English assessment. This feature is illustrated quite clearly in Figure 1, where we chart the predicted average gaps between teacher assessment and test points for various illustrative demographic groups. The chart is based on regression models similar to Table 4, column (2), but estimated separately for each year, and extending the analysis to two older cohorts (aged 14 in 2000 and 2001). High achieving children scored lower on teacher assessments than in the tests in all years after 2000, whilst low-achieving children scored higher in teacher assessments than the tests over much of the same period. In comparison with differences by achievement level, differences across ethnic group, language or gender are quite slight. The patterns are not completely stable over time: systematic differences were very small in 2000 and again by 2005, although became substantial in the intervening years.

All in all, this picture suggests quite small and not particularly stable differences across pupil demographic and socioeconomic groups in terms of the scores assigned by teachers and the scores assigned from tests. However, on average over the period, teachers have been conservative in their assessments of high ability children and over-generous in their assessment of low ability children, when compared to the scores these pupils actually achieve in their tests. Otherwise the only important differences seem to be across ethnic and language groups in English, and gender more generally. We have also checked for interactions between pupil achievement groups and ethnic, socioeconomic and gender groups, but this analysis revealed few interesting patterns. The gender gap in English seems to be concentrated in lower achievement groups, but the gender gaps in science and maths, and all the ethnic group differences are similar for high and low achieving pupils. It is important to note too that the results on ethnic and gender differences are rarely consistent with a story of statistical discrimination, or gender or ethnic stereotyping arising from face-to-face assessments. In most cases, ethnic and gender groups are assessed higher in teacher assessments than the tests (compared to the baseline group) when their expected achievement is low, and assessed lower in the teacher assessments than the tests when their expected achievement is high. Clearly, this is not a pattern we would expect to see if expected group achievement is being used to rank individual pupils.

5.4. Group characteristics and divergence in assessment.

In Table 6, we extend the analysis consider whether the divergence between teacher and test assessments for an individual is related to the characteristics of the school group of which they are a member. We focus on the results in English, where we found strongest evidence of systematic differences across groups. The table repeats the analysis of Table 4 column (3), but with additional variables measuring the proportion of each pupil's school year group who are in each ethnic, socioeconomic and ability category. These proportions are interacted with individual pupil characteristics. Column (1) reports the individual effect, column (2) the group effect and column (3) the interaction between the two. To interpret the coefficients, note that a significant coefficient in column (3) implies that the divergence in teacher and test scores associated with a specific pupil characteristics – e.g. being black – depends on the proportion of the pupil's school mates who have this characteristics – e.g the proportion black. Significant coefficients in this column could imply that teachers' assessments of individuals depend in some way on the group to which the individual belongs, or that individual test results are in someway affected by group composition.

We will comment on some of the more interesting features of this table. First note in row 1, that although a pupil's personal free meal status does not influence the test-teacher gap in test scores, there is a tendency for teachers to rate pupils higher than the tests if the school year group as a whole has a high proportion of free meal pupils, although the effect is small in magnitude. In Table 4 we saw that Asian children were rated relatively low in teacher assessments, but in row 2 of Table 6 there is little evidence of such a direct linkage: pupils in schools with a high proportion of Asian pupils are rated lower, and Asian pupils in such classes are rated lower still, but individually none of the coefficients is statistically significant.

The effects for black pupils are also interesting, with a strong tendency for black pupils to do relatively poorly in teacher assessments (column 1), and a strong tendency for white pupils to come out less well in teacher assessments if the school has a high proportion of black pupils (column 2). However, teacher assessment of black pupils is much more favourable relative to their test scores when the proportion of black pupils in the school is high (column 3). There thus appear to be quite strong interactions between black status and the composition of the school group in determining the divergence between teacher and test based assessment in English. One candidate explanation is that the teachers are drawn to make more favourable assessments of black pupils when a high proportion of the pupils in the school are black. Another is that peer group effects are at work such that black pupils score particularly badly on formal tests when in the company of other black pupils. However, the precise reasons for these interactions must remain conjecture given the data to hand. Other interactions between pupil and group socioeconomic and demographic are not particularly striking, although it is worth noting that boys appear to be assessed favourably by teachers, relative to girls (column 1), but that the teacher ranking of girls improves relative to tests as the proportion of boys in the school increases (column 2).

As we saw in the main results in Table 4, the strongest influence on the gap between teacher and test based assessment is the level of ability or prior achievement. On this topic, the interactions in Table 6 reveal some interesting patterns. As before, pupils at the bottom of the achievement ladder, with total points less below 42 at age 11, tend to do better at age 14 on teacher assessments relative to tests when compared to mid ranking pupils (column 1). If these pupils are in schools with high proportions of low scoring pupils then the teacher assessments become increasingly favourable relative to the tests (column 3). At the other end of the ability spectrum, pupils scoring highly in their age 11 assessments in English do well in their age 14 tests relative to their teacher assessments (column 1), although the gap seems to be mitigated if there is a high proportion of high-achievers in the school. So teachers appear to grade relative to the school population.

In summary, it is difficult to gauge what precise mechanisms drive these findings. What is important about the results is that they highlight that teacher and test assessments in many cases divergence systematically according to individual characteristics *and* group composition. This finding is quite worrying, since it is not what would be expected from an unbiased assessment system. In particular, the discrepancy between teacher and test assessments at the top and bottom of the achievement distribution gives cause for concern. In the next section we go on to consider whether we should be especially concerned about divergence between teacher and test based assessment in so far as these impact in future educational decisions and opportunities.

5.5. Impacts on qualifications and post-compulsory education

Our core results concerning the influence of divergence in assessment on qualifications and subsequent outcomes appear in Table 7. We consider seven different educational outcomes at age 16 and beyond, as reported in the columns of Table 7: 1) a pupil's total number of GCSE/NVQ entries and 2) their percentile in the national distribution of GCSE/NVQ points (awarded on the basis of the number and grade of test result); 3) whether or not the pupil stays on at school into year 12[10]; or 4) whether the pupils is recorded studying for any non-vocational post-16 qualification in the Independent Learner Record data set; 5) the share of English-related subjects taken at age 16; 6) the share of Science-related subjects taken at age 16; 7) the share of maths-related subjects taken at age 16. For each outcome, we look at the association of these outcomes with the teacher-test assessment gap in maths science and English, at age 14 and 11.

[10] This is determined by whether or not a pupil appears in the Pupil Level Annual Census in year 12. If a pupil leaves the school for a year and returns, they will not be counted, but as far as we can tell this almost never occurs. Pupils who take a year out will be recorded in the Independent Learner Record.

All the results are presented for specifications that include controls for basic pupil characteristics (free meal entitlement, ethnicity, language, age and gender) plus dummy variables for prior achievement levels, as estimated by the sum of the teacher and test assessment point scores at age 14 and age 11. The specifications also allow for school-specific fixed effects, but the results are insensitive to the inclusion or otherwise of these fixed effects.

We do not tabulate the results here, but it should first be noted that we can re-cast this analysis by regressing these age-16 and post-compulsory school outcome measures on teacher assessment points and test assessment points in each of the core subjects at ages 11 and 14 separately. This approach to specifying the link between assessment scores and outcomes reveals that both teacher assessments and test assessments are positively correlated with academic outcomes at age 16 and beyond. Apparently then, higher age-14 teacher assessments, conditional on age-14 test scores, are associated with better subsequent academic outcomes. Similarly, better test scores, conditional on teacher assessments, are associated with better subsequent academic outcomes. But these results are not, on their own, informative about the effects of divergence between teacher and test scores and simply indicate that that both teacher and test assessments contain unique information about pupil ability, which is in turn correlated with the likelihood of academic success in subsequent stages of a pupil's education[11]. A better way to see the direct influence of a wider gap between teacher and test scores is to parameterise as in Table 7, which shows how outcomes change as the gap widens, holding constant the average of the teacher and test-based assessments.

[11] As it turns out, age 14 test scores are slightly better predictors of most age-16 and post-compulsory outcomes than teacher assessments in these regression models. The F-statistics for the joint tests of the statistical significance of the teacher assessment scores are in most cases much less than the F-statistics for the joint test of the significance of the age-14 test scores. This is true for all outcomes except for the probability of staying in school to year 12 and the share of English subjects, for which teacher assessments at age 14 provide better predictions.

Consider then the results for GCSE entries first, in column (1). The coefficients are multiplied × 10 for readability. The coefficients on the gap variables imply that for all subjects except English, the number of GCSE entries is increasing in the favourability of the teacher assessments relative to the tests. This is what we might expect at age 14, since teacher expectations in secondary school could be directly influential in terms of the number of papers for which a pupil is entered. This possible direct linkage cannot, however, explain the association between the divergence in assessment in primary school at age 11 and the number of GCSE/NVQ entries. An alternative explanation is that positive teacher evaluations relative to test scores encourage pupils' academic ambitions through more subtle psychological channels. However, it needs to be recognised that although highly significant (we have nearly 1 million pupils in the study), the effects are minute in terms of their magnitude. The scale of the coefficients implies that a one Level (6 point) positive gap between teacher and test based assessment scores in *every* core subject at age 11 and age 14 is linked to a seven percentage point increase in the expected number of GCSE/NVQ entries, that is an increase equivalent to seven additional GCSE/NVQ entries for every 100 pupils being "over" evaluated by a full one Level by teachers in every core subject at ages 11 and 14. Of course, that fact that the regressions include control variables for age-14 might attenuate the influence of age-11 teacher-test gaps on age-16 achievement However, removing the age-14 gap and achievement variables makes little difference to the coefficients on the age11 gap variables, and removing the age-11 achievement and gap variables makes little difference to the coefficients on the age-14 gap variables. In either case, the order of magnitude does not indicate that divergence in teacher-test assessment is a major factor behind the GCSE/NVQ entry decisions.

Although column (1) could be read as providing some suggestion that a more favourable teacher assessment engenders a positive academic attitude in pupils, this view is partially at odds with the findings in column (2). Here we show that, whilst a positive teacher-test assessment gap at age 11 is linked to marginally higher performance overall in GCSE/NVQs (even conditional on age 14 achievement), the opposite is true for divergence in assessment at age 14: at this age, it is a positive test-teacher gap that is associated with better GCSE/NVQ performance. One reading of these somewhat contradictory results is that whilst the favourable teacher assessments at the end of primary school may encourage a positive pupil response, it is the pupil qualities that generate good test results at age 14 that are most closely linked to success in formal GCSE/NVQ exams at age 16. Whatever the explanation, the magnitudes are again very small: a one-Level excess in test based assessment over teacher assessment in all core subjects at age 14 is associated with an increase in GCSE/NVQ performance that is equivalent to a mere 1.2 percentiles of the pupil distribution of GCSE/NVQ point scores. This is mirrored by an almost identical effect of a one level excess of teacher assessments over test scores in all core subjects at age 11.

The findings on the association of assessment divergence with GCSE/NVQ scores is, broadly speaking, played out further in the results on the decision to stay on at school, or to pursue post-compulsory education more generally. A relatively good teacher assessment at age 11 is linked to higher probabilities of participation in post-compulsory education, but so too is a relatively good test performance at age 14. As before, the implied effects on the probability of post-school participation (and hence Higher Education participation in subsequent years) are very small indeed. According to these models, a pupil who received a full one-Level excess teacher assessment age 11 in *all* core subjects has a 1.13 percentage point higher probability of staying on at school relative to another pupil in the same school, receiving the same mean teacher and test assessments and the same observable characteristics (and increase of 3.24% relative to the mean staying on rate of 34.92%). Although this effect is not negligible, a divergence of assessment on this scale is way outside anything observed in the actual data.

The remaining columns of Table 7 show the associations between the teacher-test gaps in the core subjects and the mix of subjects taken at age 16. There is no suggestion here of any very meaningful linkage between the divergence in assessment and the choice of subjects. In general it appears that doing relatively well in maths science and English tests at age 11 and 14 (relative to teacher assessments) is linked to a higher share of maths, science and English subjects in age 16 qualifications, but all the coefficients are so small that they are effectively zero, even when statistically significant. For example, our benchmark extreme case of a 1 level divergence of test scores over teacher assessments at age 11 and 14 would yield a 0.22 percentage point increase (1.1% relative increase) in the percentage of science related subjects taken at age 16.

We have also looked further to see if we are masking important subtleties in response by imposing a linear restriction on the effects of assessment divergence. Perhaps, for example, pupils that are "over-assessed" by teachers relative to tests experience more positive outcomes as the degree of over-assessment increases, whereas those who are under-assessed by teachers relative to tests are adversely affected or unaffected. We find occasional evidence of such non-linearities. For example, pupils who are over-assessed in the maths teacher assessment at age 11 and 14 tend to have more GCSE/NVQ entries at age 16 as the degree of over assessment increases, whilst the number of entries for those who are under assessed tends to decrease as the teacher and test assessments converge. However, for the most part there are few significant differences of this type and no general evidence of non-linearities in response. We have also considered whether teacher-test assessment gaps have bigger influences on outcomes for low achieving pupils or for high achieving pupils, but the patterns for both high and low achievers are similar and broadly in line with Table 7.

In summary, although we have found some statistically significant effects, the results in Table 7 do not appear to tell a convincing story about divergence in teacher and test-based assessments having any real impact on qualifications or post-school participation decisions.

6. Conclusions

Our empirical analysis finds evidence of systematic differences between test and teacher based assessments in national curriculum assessment at secondary school in England, using data on the population of age-14 state school pupils from 2002-2005. The biggest differences are between pupil achievement groups, with higher achieving pupils more likely to be under-assessed by teachers relative to tests, and low achieving pupils more likely to be under-assessed by the tests relative to the teachers. There are also differences by gender: boys do relatively well in teacher assessments in English, but girls do relatively well in the teacher assessments in science and maths. The gender gap between test and teacher assessments is comparable in magnitude to the gender gap in the mean test and teacher scores, though of opposite sign. There are some smaller differences by ethnic group in English assessment The reasons for these divergences between teacher and test based assessment scores are not revealed by our analysis, but statistical discrimination or stereotyping seems an unlikely explanation since any upward 'bias' in teacher assessments relative to the tests works in favour of low-achieving groups.

It is of course unlikely that any two different assessment methods will give directly comparable measures of pupil achievement and skills for every pupil, especially when there are differences in breadth of skills which are being assessed. However, mean differences across pupil groups do raise serious concerns about placing too much trust on any one form of assessment. Clearly, the current policy and pedagogical emphasis on the use of tests alone is problematic, as is any suggestion that the system is shifted to very heavy reliance on the teacher assessments (Brooks and Tough 2006). The findings in this study support the idea that a balanced use of teacher assessment and tests provides a better foundation for pupil and school evaluation (Stobart 2001).

Even so, we find little evidence that divergence between teacher assessment and actual test scores really matters much for pupil outcomes. Favourable teacher assessments are linked to marginally more GCSE/NVQ entries at age 16, suggesting a possible direct route by which teacher perceptions could influence subsequent pupil outcomes. However, the effects are very small in magnitude and we find no strong evidence here that discrepancies in assessment have any influence on qualifications or post-compulsory schooling decisions. Hence, it seems unlikely from this evidence that pupils are heavily influenced by teacher perceptions of their abilities or that teacher perceptions could be a major influence on post-16 or higher education participation rates.

7. References

Ammermueler, A. and P. Dolton (2006) Pupil-teacher gender interaction effects on scholastic outcomes in England and the USA, Discussion Paper 06-060, ZEW Mannheim

Connor H., C. Tyers, T. Modood and J. Hillage (2004) Why the Difference? A Closer Look at Higher Education Minority Ethnic Students and Graduates, Department for Education and Skills, RR 552, London

Connor H., S. Dewson, C. Tyers, J. Eccles, J. Regan and J. Aston (2001) Social Class and Higher Education: Issues Affecting Decisions on Participation by Lower Social Class Groups, Department for Education and Skills, RR 267, London

Dee (2005a) A teacher like me: does race, ethnicity or gender matter? American Economic Review 95 (2) 159-165

Dee (2005b) Teachers and the gender gaps in student achievement, National Bureau of Economic Research Working Paper w11660

Department for Education and Skills (2003) The Future of Higher Education, London: The Stationery Office

Gipps, C and P. Murphy (1994) A Fair Test? Assessment, Achievement and Equity, Buckingham: Open University Press

Brooks, R. and S. Tough (2006) Assessment and Testing: Making Space for Teaching and Learning, London: Institute for Public Policy Research

Lavy V. (2004) Do gender stereotypes reduce girls' human capital outcomes? Evidence from a natural experiment NBER Working Paper w10678

Ouazad, A. (2007) Assessed by a teacher like me: race gender and subjective evaluations, Centre for Economic Performance, London School of Economics, mimeo

Phelps, E. (1972) A statistical theory of racism and sexism, The American Economic Review 62 (4) 659-661

QCA (2004) Assessment and Reporting Arrangements Years 1-9, London: Qualifications and Curriculum Authority

Reeves, D.J., W. F. Boyle and T. Christie (2001) The relationship between teacher assessments and pupil attainments in standard tasks at Key Stage 2, 196-1998, British Journal of Educational Research 27 (2) 141-160

Steele C. and J. Aronson (1995) "Stereotype threat and the intellectual test performance of African Americans", *Journal of Personality and Social Psychology*, 69, 797-811

Stobart (2001) The validity of national curriculum assessment, British Journal of Educational Studies 49 (1) 26-39

Tajfel, H. (1959) Quantitative judgement in social perception, British Journal of Psychology, 50 16-29

Todd, P. E. and K. I. Wolpin (2003) On the Specification and Estimation of the Production Function for Cognitive Achievement, Economic Journal, 113(485), F3-33.

Wright, S.C., and D. M. Taylor (2007) The social psychology of cultural diversity: social stereotyping, prejudice and discrimination, Ch. 16 in M. A. Hogg and J. Cooper (eds.) The SAGE Handbook of Social Psychology, London: SAGE publications

8. Technical Appendix

8.1. Estimation of test-teacher assessment gaps

Our goal is to explore, empirically, how gaps between teacher-based and test-based assessment of pupils' levels of achievement differ by pupil ability and by pupil characteristics, and to seek explanations for these differences. The approach taken is based on the assumption that both teacher-based and test-based assessments are generating scores s_i^t, s_i^j that try to measure some unobservable 'ability' (a_i) of the pupil. This unobservable attribute is his or her aptitude in a fairly specific set of tasks in a specific set of academic subject areas (literary, mathematical, scientific etc.). The model set out below explores the difficulties faced in modelling the gap between teacher and test-based assessments, when actual ability is unobserved.

All forms of assessment are imperfect and measure the object of the assessment, a_i, with a degree of error. For instance, tests can be considered 'objective' in that they are marked by examiners who do not know the candidates or their socioeconomic characteristics. However, these tests can evaluate performance on a very limited range of questions on a specific day or days only and can systematically over or underestimate ability. Moreover, such test-based assessments generate a pupil score that can deviate from a_i perhaps because the specific test was not well designed to the pupil's gender, ethnic or cultural background, or because they just had a bad day. With these issues in mind, the test assessment score can be written $s_i^t = (1-\rho)a_i + u_i^t$ where u_i^t is an error term that is pupil-test and test-day specific, and ρ is a parameter that captures the general tendency of a test to inflate or deflate a pupil's true 'ability' irrespective of other pupil characteristics.

Conversely, teacher assessments are usually based on observation of ability on a wider range tasks over a longer time horizon, but may be sensitive to personal and subjective judgements on the part of the teacher, or poor relationships between pupil and teacher. In particular, teacher assessment could be subject to some form of statistical discrimination or stereotyping whereby judgements are made on the basis of what is expected of pupils of similar socioeconomic background, gender or ethnic group, rather than a pupil's personal abilities. So teacher j's assessment of pupil i can be written $s_i^j = (1+\rho)a_i + v_i^j$ where v_i^j is an error term that is pupil-teacher specific[12]. The 'errors' v_i^j and u_i^t in teacher and test-based assessments are uncorrelated with a_i, the assumption being that any correlation between the assessment error and ability is represented by ρ. Our interest is whether, in the population, assessments of ability generated by tests and by teacher observation diverge systematically according to the ethnic and socioeconomic characteristics of pupils, in ways that are unrelated to their actual ability. Formally, we can model the expected attainment of pupils conditional on observable characteristics of pupils and their context (family background, school characteristics and peer characteristics) x_i as:

$$E\left[s_i^j \mid a_i, x_i\right] = (1+\rho)a_i + x_i'\beta$$

$$E\left[s_i^j \mid a_i, x_i\right] = (1-\rho)a_i - x_i'\beta \qquad (1)$$

If teacher and test based assessments measured ability a_i perfectly, without bias in either assessment method, then the population parameters β , and ρ would be identically zero.

[12] Note that the parameterisation of the assessment inflator and deflators $(1+\rho),(1-\rho)$ ensures that any weighted average of the teacher and test assessments is a better measure of true ability than either assessment on its own. We impose symmetry in terms of the deviation of teacher and test scores from the mean, because only the symmetric deviation around the mean is identified without prior information on which assessment is right.

Divergence of β from zero indicates that, for some reason, the error created by one of the assessment methods at test-time t, or by teacher j, is correlated with pupil characteristics x_i. Divergence of ρ from zero shows that one or more of the assessments tends to under or overstate underlying ability. Note, however, that a_i is defined such that the expectation of the overall, combined assessment $E\left[s_i^j + s_i^t \mid a_i, x_i\right] = a_i$ is an unbiased estimate of the characteristic that is the target of the assessment process[13]. The aim of the empirical work in this paper is to try to estimate β and ρ by simple regression methods. The model suggested below emphasises the empirical challenges and sets out a simple framework for estimation.

Note that assessment scores can be written:

$$s_i^j = (1+\rho)a_i + x_i'\beta + \varepsilon_i^j$$

$$s_i^t = (1-\rho)a_i - x_i'\beta + \varepsilon_i^t \qquad (2)$$

where $\varepsilon_i^j, \varepsilon_i^t$ are zero mean, random error terms that are uncorrelated with pupil background or ability and uncorrelated with each other. Differencing these equations provides a starting point for estimation:

$$s_i^j - s_i^t = \rho 2a_i + 2x_i'\beta + \varepsilon_i^j - \varepsilon_i^t \qquad (3)$$

However, estimation of this equation directly is not possible given that a_i is unobserved. Aggregated test and teacher assessments provide unbiased estimate of ability if $E\left[\varepsilon_i^j\right] = E\left[\varepsilon_i^t\right] = 0$ because:

$$s_i^j + s_i^t = 2a_i + \varepsilon_i^j + \varepsilon_i^t$$

$$E\left[s_i^j + s_i^t\right] = 2a_i \qquad (4)$$

[13] As a consequence, any tendency for a group to do better in both types of assessment is classed as a difference in 'ability' a_i, rather than an overall bias in the assessment.

Superficially, these aggregated test scores (across teacher and test-based assessments) may seem attractive as proxies for the overall level of ability, for example by substituting for ability in (3) to give the empirical model

$$s_i^j - s_i^t = \rho\left(s_i^j + s_i^t\right) + 2x_i'\beta + (1-\rho)\varepsilon_i^j - (1+\rho)\varepsilon_i^t \qquad (5)$$

However, this model cannot, in general, be estimated consistently by ordinary least squares methods because the error term is correlated with the aggregated assessment $s_i^j + s_i^t$. Indeed, the parameter ρ cannot be separately identified from differences in the variance of the assessment errors $\sigma_j^2 - \sigma_t^2$. To see, this, note that the coefficient in an ordinary least squares regression of $s_i^j - s_i^t$ on $s_i^j + s_i^t$ would yield a consistent estimate of:

$$\hat{\rho} = \frac{Cov\left(s_i^j - s_i^t, s_i^j + s_i^t\right)}{Var\left(s_i^j + s_i^t\right)} = \frac{4\rho\sigma_a^2 + \sigma_j^2 - \sigma_t^2}{4\sigma_a^2 + \sigma_j^2 + \sigma_t^2} \qquad (6)$$

This is only equal to ρ if $\sigma_j^2 = \sigma_t^2 = 0$. If $\sigma_j^2 = \sigma_t^2 > 0$ then estimates of ρ are attenuated by noise in the assessment ($\sigma_j^2 + \sigma_t^2$), a problem analogous to classical measurement error, or as in 'value-added' educational models – see Todd and Wolpin (2003). In practice, it seems likely that the variance in 'ability' (σ_a^2) across pupils dominates both types of assessment error in both the covariance and variance terms in (6), implying that the bias arising from this source may be of little concern (for instance if the noise to signal ratio $\left(\sigma_j^2 + \sigma_t^2\right)/\sigma_a^2$ is 10% , but $\sigma_j^2 = \sigma_t^2$ then $\hat{\rho} = 4\rho/4.1 = 0.976\rho$.

More importantly, if $\sigma_j^2 \neq \sigma_t^2$, the difference in the variance in the assessment errors will show up in least squares estimates as tendency of one form of assessment to exaggerate ability relative to the other, that is the effect of ability on the teacher-test assessment gap operating via ρ and the effect of differences in the teacher-test variance are observationally equivalent if the aggregated assessment is used as proxy for ability. One possible solution is to use aggregated test and teacher assessments from past periods and different teachers as an estimate of ability, on the assumption that the sum of errors in tests and assessment by different teachers in previous years ($\tilde{\varepsilon}_i^j + \tilde{\varepsilon}_i^t$) is uncorrelated with difference between current test and teacher errors $\varepsilon_i^j - \varepsilon_i^t$, conditional on ability and observable characteristics. Note in addition that it is possible to control directly for $\tilde{\varepsilon}_i^j - \tilde{\varepsilon}_i^t$ in regression estimates of (5), which eliminates one channel through which the current deviation between teacher and test-based assessments might be correlated with the current aggregate of teacher and test-based assessments.

Clearly this does not fix the problem that the overall past assessment is, presumably, an even more noisy measure of current ability than the current assessments (implying attenuation of parameters). For instance, suppose the aggregated past assessment is:

$$\tilde{s}_i^j + \tilde{s}_i^t = 2\gamma a_i + \tilde{\varepsilon}_i^j + \tilde{\varepsilon}_i^t \qquad (7)$$

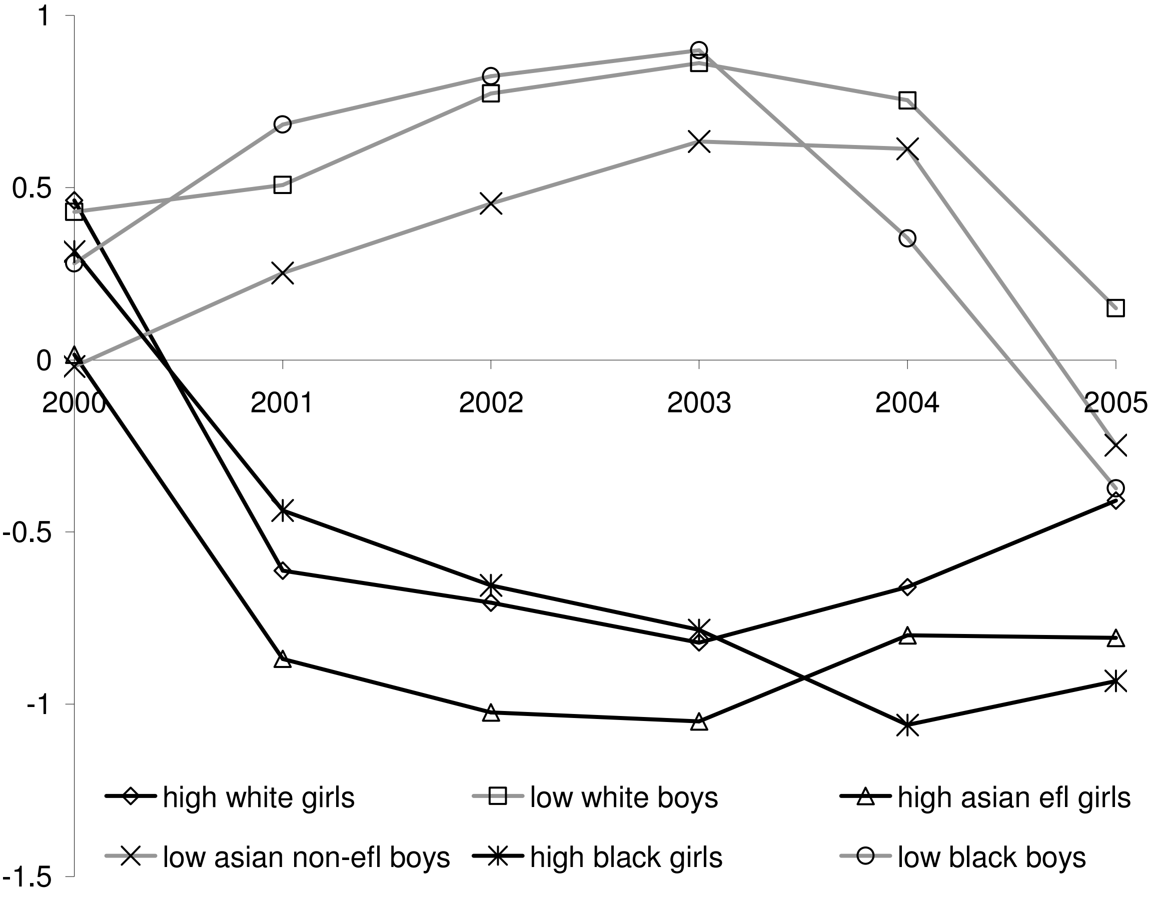
where a_i is the object of current assessment, γ is a scaling parameter $\in [0,1]$ relating past to current ability, and $\tilde{\varepsilon}_i^j, \tilde{\varepsilon}_i^t$ are the past assessment errors (uncorrelated with $\varepsilon_i^j, \varepsilon_i^t$). Then, ordinary least squares regression of $s_i^j - s_i^t$ on $\tilde{s}_i^j + \tilde{s}_i^t$ would yield a consistent estimate of:

$$\tilde{\rho} = \frac{Cov\left(s_i^j - s_i^t, \tilde{s}_i^j + \tilde{s}_i^t\right)}{Var\left(\tilde{s}_i^j + \tilde{s}_i^t\right)} = \frac{4\rho\gamma\sigma_a^2}{4\gamma^2\sigma_a^2 + \tilde{\sigma}_j^2 + \tilde{\sigma}_t^2} \qquad (8)$$

To the extent that $\gamma < 1$, this estimate provides an inflated estimate of ρ, interpreted as the effect of current ability on the teacher-test gap. As before, noise in the overall assessment of ability present through $\tilde{\sigma}_j^2 + \tilde{\sigma}_t^2$ attenuates the coefficient estimate.

9. Figures

Figure 1: Trends in estimated teacher-test gap in English Key Stage 3 points for various demographic groups, 2000-2005



High refers to high achievement at age 11, scoring 66 points in teacher and test assessments. Low refers to low achievement at age 11, scoring 42 points or less in teacher and test assessments.

10. Tables

Table 1: Descriptive statistics of the Age 14 and Age 16 samples.

	(1)	(2)	(3)	(4)	(5)	(6)
	Full sample	Free meals	Asian	Black	English additional	Male
Age-14 sample						
Age 14 English teacher + test points	69.73	64.38	67.80	67.44	67.852	67.87
	(10.22)	(9.916)	(9.94)	(9.814)	(10.030)	(10.24)
Age 14 science teacher +test points	69.93	63.78	66.50	65.65	66.81	70.54
	(11.00)	(10.62)	(11.38)	(10.67)	(11.45)	(10.97)
Age 14 maths teacher +test points	74.46	67.66	72.31	69.47	72.40	75.27
	(13.26)	(12.88)	(13.72)	(13.08)	(13.74)	(13.25)
Age 14 English teacher-test gap	-0.294	-0.212	-0.599	-0.478	-0.578	-0.214
	(4.477)	(4.706)	(4.648)	(4.581)	(4.650)	(4.551)
Age 14 science teacher-test gap	-0.082	0.122	0.021	0.045	-0.012	-0.267
	(4.014)	(4.223)	(4.268)	(4.262)	(4.276)	(3.998)
Age 14 maths teacher-test gap	-0.336	-0.160	-0.280	-0.150	-0.304	-0.552
	(3.550)	(3.716)	(3.773)	(3.756)	(3.761)	(3.540)
Observations	1439409	172352	81231	37882	107979	683945
Age 16 sample						
Total GCSE NVQ entries x 10	9.808	9.363	10.08	9.786	10.07	9.767
	(1.585)	(2.019)	(1.480)	(1.600)	(1.503)	(1.647)
GCSE NVQ percentile	52.48	37.41	55.11	45.93	54.91	49.36
	(27.43)	(26.25)	(26.74)	(26.31)	(26.94)	(27.30)
Stay on at school %	34.94	22.37	43.48	30.71	42.74	34.08
	(47.67)	(41.67)	(49.57)	(46.13)	(49.47)	(47.40)
Any non-vocational post 16 %	74.24	65.02	86.22	82.85	85.42	72.28
	(43.73)	(47.69)	(34.46)	(37.69)	(35.29)	(44.76)
Share of English-related subjects %	19.85	20.50	19.75	20.13	19.71	19.78
	(4.528)	(6.645)	(3.693)	(4.519)	(3.814)	(4.673)
Share of Science-related subjects %	19.99	20.06	19.68	19.60	19.59	20.32
	(5.436)	(7.486)	(4.739)	(5.725)	(4.875)	(5.535)
Share of maths-related subjects %	10.64	11.55	10.49	10.77	10.48	10.77
	(3.805)	(5.803)	(2.767)	(3.535)	(2.964)	(4.053)
Observations	1015446	105231	58642	24089	75271	485245

Notes: Standard deviations in parentheses

Table 2: Discrepancies between teacher and test based assessments: Correlations between various subjects and ages.

	Teacher-test English age-14	Teacher-test science age-14	Teacher-test maths age-14	Teacher-test English age-11	Teacher-test science age-11
Teacher-test English age-14	1.000	-	-	-	-
Teacher-test science age-14	0.054	1.000	-	-	-
Teacher-test maths age-14	0.042	0.081	1.000	-	-
Teacher-test English age-11	0.005	0.013	0.004	1.000	-
Teacher-test science age-11	-0.001†	0.012	0.007	0.160	1.000
Teacher-test maths age-11	-0.004	0.013	0.019	0.112	0.172

Notes: Table reports Pearson's correlation coefficients. All significant at less than 0.001% level except † p-value 0.184. Age-14 subject assessments generally made by different teachers for any one pupil. Age-11 subject assessments generally made by the same primary school teacher for any one pupil.

Table 3: Teacher and test scores: age 14 assessments

	English		Science		Maths	
	(1)	(2)	(3)	(4)	(5)	(6)
	Teacher + test points	Teacher – test points	Teacher + test points	Teacher –test points	Teacher + test points	Teacher –test points
Free meals	-5.186	0.168	-5.346	0.215	-6.732	0.191
	(0.052)	(0.023)	(0.054)	(0.019)	(0.066)	(0.015)
Asian	-0.080	-0.241	-1.415	0.091	-0.683	0.106
	(0.124)	(0.067)	(0.146)	(0.058)	(0.182)	(0.042)
Black	-1.328	-0.176	-2.711	0.102	-3.823	0.180
	(0.116)	(0.053)	(0.127)	(0.051)	(0.159)	(0.041)
Mixed	0.475	-0.220	-0.212	-0.072	-0.467	-0.004
	(0.087)	(0.045)	(0.096)	(0.045)	(0.116)	(0.031)
Other	1.124	-0.218	1.072	-0.121	2.072	0.006
	(0.149)	(0.062)	(0.172)	(0.062)	(0.216)	(0.050)
English additional	-0.483	-0.162	-0.716	-0.028	-0.180	-0.082
	(0.093)	(0.057)	(0.108)	(0.057)	(0.131)	(0.032)
Male	-3.114	0.175	0.827	-0.381	1.166	-0.424
	(0.034)	(0.015)	(0.035)	(0.013)	(0.043)	(0.010)
Age in Sept	0.181	-0.006	0.148	0.004	0.240	0.010
	(0.002)	(0.001)	(0.002)	(0.001)	(0.003)	(0.001)
Constant	68.87	-0.166	67.21	0.032	71.06	-0.090
	(0.076)	(0.043)	(0.080)	(0.033)	(0.100)	(0.025)
R-squared	0.077	0.002	0.052	0.009	0.052	0.005
Ethnicity p-value[A]	0.000	0.000	0.000	0.000	0.000	0.000
School type p-value[A]	0.000	0.030	0.000	0.007	0.000	0.829

Notes: Table reports regression coefficients. Standard errors, in parentheses. Dependent variable is sum of Key Stage 3 (Age-14) Teacher Assessed points and Test points. Standard errors are clustered on 2853 secondary schools. Underline significant at 1% or better. Unreported controls are year dummies, unknown ethnic group, school type. Baseline group is non-poor, white British girl with English first language aged 14 years and 0 months in September. Sample is the population of non-SEN pupils in Comprehensive secondary schools in England where there is complete information on age-11 and age-14 assessments and other characteristics, in Year 9 in 2002, 2003,2004, 2005. Pupils scoring maximum or minimum in both tests and teacher assessments are excluded. Sample size 1330381-1398950.

[A] p-values on a F-test of joint significance for the given set of dummy variables.

Table 4: The teacher-test assessment gap: Age 14 assessment

	Age 14 English teacher - test points			Age 14 science teacher - test points			Age 14 maths teacher - test points		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Free meals	-0.029	0.036	0.003	0.036	0.050	0.010	-0.003	0.029	-0.007
	(0.020)	(0.021)	(0.014)	(0.018)	(0.019)	(0.012)	(0.014	(0.014)	(0.010)
Asian	-0.231	-0.266	-0.118	0.046	-0.014	0.118	0.083	0.070	0.111
	(0.066)	(0.067)	(0.042)	(0.058)	(0.057)	(0.035)	(0.041)	(0.042)	(0.026)
Black	-0.197	-0.222	0.003	0.018	0.003	0.116	0.084	0.076	0.132
	(0.053)	(0.054)	(0.032)	(0.051)	(0.051)	(0.027)	(0.040)	(0.041)	(0.024)
Mixed	-0.200	-0.205	-0.088	-0.074	-0.062	0.019	-0.007	-0.012	0.042
	(0.044)	(0.045)	(0.035)	(0.044)	(0.037)	(0.030)	(0.029)	(0.030)	(0.025)
Other	-0.173	-0.219	-0.102	-0.083	-0.118	-0.047	0.052	0.046	0.094
	(0.061)	(0.062)	(0.050)	(0.061)	(0.062)	(0.043)	(0.050)	(0.049)	(0.037)
English additional	-0.168	-0.212	-0.080	-0.058	-0.094	0.043	-0.076	-0.115	-0.049
	(0.057)	(0.058)	(0.046)	(0.051)	(0.050)	(0.046)	(0.030)	(0.031)	(0.023)
Male	0.072	0.119	0.089	-0.355	-0.340	-0.350	-0.381	-0.362	-0.369
	(0.015)	(0.015)	(0.011)	(0.013)	(0.013)	(0.011)	(0.010)	(0.010)	(0.008)
Age in Sept	-0.001	0.003	0.002	0.009	0.017	0.018	0.015	0.022	0.022
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Age 11 gap	-	0.005	0.003	-	0.024	0.023	-	0.015	0.014
		(0.002)	(0.002)		(0.002)	(0.001)		(0.001)	(0.001)
Teacher+test score <=42	-	0.869	0.846	-	0.871	0.882	-	1.107	1.118
		(0.024)	(0.023)		(0.025)	(0.023)		(0.017)	(0.017)
48	3.584	0.143	0.149	1.568	0.496	0.501	1.370	0.504	0.514
	(0.069)	(0.017)	(0.016)	(0.080)	(0.019)	(0.018)	(0.046)	(0.014)	(0.014)
54	0.068	Baseline	Baseline	0.029	Baseline	Baseline	0.196	Baseline	Baseline
	(0.015)			(0.011)			(0.011)		
60	-0.974	-0.097	-0.083	0.199	-0.351	-0.340	0.548	-0.162	-0.156
	(0.040)	(0.016)	(0.015)	(0.045)	(0.014)	(0.013)	(0.048)	(0.013)	(0.012)
66	Baseline	-0.288	-0.271	Baseline	-0.900	-0.882	-0.019	-0.289	-0.271
		(0.022)	(0.020)		(0.019)	(0.018)	(0.006)	(0.016)	(0.014)
72	-0.361	-	-	-0.343	-	-	-2.051	-	-
	(0.040)			(0.042)			(0.040)		
78	-0.109	-	-	0.050	-	-	Baseline	-	-
	(0.014)			(0.007)					
84	-0.773	-	-	-1.035	-	-	-1.616	-	-
	(0.049)			(0.057)			(0.047)		
90	-	-	-	-	-	-	0.058	-	-
							(0.005)		
96		-	-	-	-	-	-1.147	-	-
							(0.074)		
Fixed effects	-	-	School	-	-	School	-	-	School
Constant	0.022	-0.251	-0.291	0.044	0.057	0.009	0.139	-0.313	-0.313
	(0.038)	(0.081)	(0.034)	(0.029)	(0.033)	(0.025)	(0.022)	(0.025)	(0.018)
R-squared	0.028	0.006	0.056	0.021	0.022	0.069	0.061	0.022	0.054
Ethnicity^A	0.000	0.000	0.005	0.094	0.134	0.000	0.142	0.217	0.000
School type^A	0.299	0.197	-	0.051	0.021	-	0.500	0.489	-

Notes: Table reports regression coefficients. Standard errors, in parentheses. Dependent variable is Key Stage 3 (Age-14) Teacher Assessed Level and Test Level converted to National Curriculum Points. Standard errors are clustered on 2853 secondary schools. Underline significant at 1% or better. Unreported controls are year dummies, unknown ethnic group, school type. Baseline group is non-poor, white British girls with English first language aged 14 years and 0 months in September. Sample is the population of non-SEN pupils in Comprehensive secondary schools in England where there is complete information on age-11 and age-14 assessments and other characteristics, in Year 9 in 2002, 2003,2004, 2005. Pupils scoring maximum or minimum in both tests and teacher assessments are excluded. Sample size 1330381-1398950.

^{A} p-values on a F-test of joint significance for the given set of dummy variables

Table 5: The teacher-test-assessment gap: age 11 assessment

	Age 11 English teacher - test points		Age 11 science teacher - test points		Age 11 maths teacher - test points	
Free meals	0.111 (0.011)	-0.006 (0.007).	-0.036 (0.012)	-0.088 (0.011)	-0.015 (0.010)	-0.070 (0.010)
Asian	-0.018 (0.030)	-0.023 (0.027)	0.104 (0.033)	0.100 (0.030)	0.037 (0.029)	-0.049 (0.029)
Black	0.030 (0.027)	0.010 (0.024)	0.139 (0.030)	0.122 (0.027)	0.126 (0.026)	0.089 (0.025)
Mixed	-0.089 (0.028)	-0.039 (0.026)	-0.061 (0.031)	-0.023 (0.028)	0.073 (0.026)	0.069 (0.026)
Other	-0.125 (0.035)	-0.087 (0.032)	0.044 (0.038)	0.060 (0.035)	-0.044 (0.034)	-0.024 (0.034)
English additional	-0.104 (0.024)	-0.118 (0.022)	-0.007 (0.025)	0.014 (0.023)	-0.065 (0.022)	-0.073 (0.022)
Male	0.010 (0.007)	-0.050 (0.006)	0.005 (0.007)	0.044 (0.007)	-0.274 (0.006)	-0.251 (0.006)
Age in Sept	-0.005 (0.001)	0.005 (0.001	0.010 (0.001)	0.016 (0.001)	0.010 (0.001)	0.015 (0.001)
Point score <=42	-	-0.076 (0.007)	-	-0.317 (0.018)	-	0.055 (0.008)
48	-	-2.267 (0.028)	-	-3.003 (0.029)	-	0.831 (0.034)
54	-	Baseline	-	Baseline	-	Baseline
60	-	-2.653 (0.025)	-	-2.529 (0.026)	-	-0.582 (0.033)
Constant	-0.455 (0.051)	0.167 (0.012)	-0.746 (0.017)	0.035 (0.013)	-0.128 (0.014)	-0.182 (0.012)
R-squared	0.005	0.004	0.126	0.137	0.006	0.019
Ethnicity^A	0.000	0.000	0.020	0.000	0.000	0.000
School type^A	0.001	0.000	0.047	0.000	0.000	0.000

Notes: Table reports regression coefficients. Standard errors, in parentheses. Dependent variable is Key Stage 2 (Age-11) Teacher Assessed Level and Test Level converted to National Curriculum Points. Standard errors are clustered on 14791 primary schools. Underline significant at 1% or better. Unreported controls are year dummies, unknown ethnic group, school type. Baseline group is non-poor, white British girls with English first language aged 14 years and 0 months in September. Sample is the population of non-SEN pupils in Comprehensive secondary schools in England where there is complete information on age-11 assessments and other characteristics for Year 9 pupils (age-14) in 2002, 2003,2004, 2005. Sample size approx 1.1million.

^{A} p-values on a F-test of joint significance for the given set of dummy variables

Table 6: Group and interaction effects in the test-teacher-assessment gap in English

	Pupil effect	School group effect	Interaction	Row F-test
Free meals	0.030	0.682	-0.085	0.042
	(0.035)	(0.247)	(0.171)	
Asian	-0.025	-0.252	-0.335	0.044
	(0.068)	(0.410)	(0.253)	
Black	-0.184	-0.887	0.974	0.008
	(0.066)	(0.346)	(0.311)	
Mixed	-0.172	-1.052	2.105	0.021
	(0.067)	(0.795)	(1.618)	
Other	-0.023	-0.381	-1.323	0.160
	(0.068)	(0.834)	(0.973)	
English additional	-0.115	-0.294	0.079	0.146
	(0.066)	(0.389)	(0.217)	
Male	0.341	0.571	-0.517	0.000
	(0.131)	(0.179)	(0.270)	
Age in Sept	-0.028	0.010	0.006	0.100
	(0.021)	(0.042)	(0.004)	
Prior assessment <=42	0.446	-0.222	2.907	0.000
	(0.042)	(0.339)	(0.288)	
Prior assessment = 48	0.051	-1.048	0.954	0.000
	(0.041)	(0.441)	(0.038)	
Prior assessment = 54	Baseline	Baseline	Baseline	
Prior assessment = 60	-0.250	-0.931	1.121	0.000
	(0.055)	(0.447)	0.371	
Prior assessment = 66	-0.584	-0.034	1.243	0.000
	(0.061)	(0.319)	(0.255)	
Column F-test (excluding prior assessment)	0.000	0.000	0.000	

Notes: Table reports regression coefficients and standard errors. Dependent variable is Key Stage 3 (Age-14) Teacher Assessed Level and Test Level converted to National Curriculum Points. Standard errors are clustered on 2852 secondary schools. Underline significant at 1% or better. Unreported controls are year dummies, unknown ethnic group, school type and prior test-teacher gap. School type dummies insignificant (p-value 0.375). Baseline group is non-poor, white British girls with English first language aged 14 years and 0 months in September. Sample is the population of pupils in Comprehensive secondary schools in England where there is complete information on age-11 and age-14 assessments and other characteristics, in Year 9 in 2002, 2003,2004, 2005. Sample size 1362696

Table 7: GCSEs, staying on and teacher assessments.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Total GCSE NVQ entries x 10	GCSE NVQ percentile	Stay on at school %	Any non-vocational post 16 %	Share of English-related subjects %	Share of Science-related subjects %	Share of maths-related subjects %	Mean (s.d.)
Teacher test gap in English age-14	-0.016	-0.043	-0.003	-0.049	-0.003	0.004	0.004	-0.062
	(0.005)	(0.005)	(0.011)	(0.010)	(0.002)	(0.002)	(0.001)	(4.508)
Teacher test gap in science age-14	0.016	-0.111	-0.044	-0.073	0.000	-0.015	-0.004	-0.056
	(0.005)	(0.056)	(0.013)	(0.012)	(0.001)	(0.002)	(0.001)	(3.964)
Teacher test gap in maths age-14	0.026	-0.048	0.006	-0.048	0.000	-0.006	-0.007	-0.221
	(0.052)	(0.006)	(0.013)	(0.013)	(0.001)	(0.002)	(0.001)	(3.587)
Teacher test gap in English age-11	0.017	0.040	0.030	0.016	-0.004	-0.004	-0.002	-0.166
	(0.005)	(0.006)	(0.014)	(0.015)	(0.002)	(0.002)	(0.001)	(2.849)
Teacher test gap in science age-11	0.038	0.074	0.100	0.047	-0.004	-0.010	-0.007	-0.116
	(0.005)	(0.005)	(0.014)	(0.014)	(0.002)	(0.002)	(0.001)	(2.991)
Teacher test gap in maths age-11	0.036	0.094	0.058	0.059	-0.005	-0.006	-0.004	0.176
	(0.005)	(0.006)	(0.015)	(0.016)	(0.002)	(0.002)	(0.001)	(2.757)
R-squared	0.356	0.726	0.397	0.159	0.167	0.173	0.127	
Mean (sd) of dependent variable	97.892	52.392	34.923	74.237	19.848	20.000	10.639	
	(16.404)	(27.502)	(47.672)	(43.733)	(4.528)	(5.436)	(3.805)	

Notes: Table reports regression coefficients and standard errors. Teacher-test gap variable is Teacher Assessed Level minus Test Level converted to National Curriculum Points. Standard errors are clustered on 2852 secondary schools. Underline significant at 1% or better. Controls variable are prior achievement (dummy variables for sum of test and teacher points at age 11 and age 14), ethnic group, free meals, language, gender, age in months, school fixed effects. Sample is the population of non-SEN pupils in Comprehensive secondary schools in England where there is complete information on assessments and other characteristics, in Year 9 in 2000, 2001, 2002. Sample size 1010191 to 1015527. Means and standard deviations of teacher-test gaps reported for maximum sample.