Assessment and age 16+ education participation

Stephen Gibbons*, Arnaud Chevalier**

January 2008

*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

This paper summarises our research into the relationship between pupil assessment at age 14 (Key Stage 3) and participation in age 16+ education. We assume, in line with previous literature, that a systematic gap between teacher-based assessment and externally-marked tests indicates some type of assessment bias or uncertainty, either in the testing procedure or as a result of teachers' perceptions of pupils' skills. We go on to explore whether these errors could have any consequence for pupils' subsequent educational attainment and participation. In common with other work, we find that teacher and test assessments diverge slightly along lines of pupil characteristics, especially prior achievement, that are clearly observable to the teacher but less so to external assessors. However, this divergence does not conform to standard notions of stereotyping by teachers. Moreover, the divergence between test and teacher assessments at age 14 has almost no bearing on pupil qualifications or participation in education after age 16, and hence seems unlikely to influence participation rates in higher education.

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. However, the validity and reliability of pupil assessment procedures has been a central question in educational research for many years, especially in terms of the relationship between assessment and equity (Gipps and Murphy 1994). The justification for this interest is, presumably, the belief that errors in assessment have consequences for pupils, teachers, school leaders or others who are evaluated on the basis of these assessments. There are of course some situations where the potential long-run consequences of mis-assessment are obvious, for example if students are awarded final qualifications that understate their skills and abilities, barring entry to higher education. But mis-assessment at earlier stages of schooling could affect educational trajectories in mores subtle ways, by misleading or making pupils uncertain about their abilities[1].

This paper is a summary of our research that considers two linked questions that are pertinent to the widening participation agenda in England. First we consider whether measures of bias or uncertainty in assessment in the National Curriculum at Key Stage 3 – specifically the difference between teacher and test-based assessment at age 14 – are linked systematically to observable pupil characteristics. Many papers have compared teacher and test-based assessments before, but most have used small samples with limited coverage, and we extend the analysis to cover the majority of the population of England's secondary school pupils over 4 years. Secondly we consider the important, but largely unexplored question of whether these measures of uncertainty and bias in assessment have any relationship with pupil's subsequent educational outcomes at age 16 and beyond. Our work is the first to provide empirical evidence on this issue. Our overall hypothesis is that some demographic

[1] We use the term 'ability' here to mean a set of skills and competencies at a particular point in time, not innate aptitude or talent.

groups may be under-represented in higher education, because they experience greater bias or uncertainty in the feedback they get about their abilities at school. This bias and uncertainty can affect participation rates because a) it means pupils have imperfect information about their abilities and/or b) because part of the bias and uncertainty arises through teacher mis-perceptions and teachers have a direct or indirect influence on pupil's subsequent choices. 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).

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 of residence plus details on their post-16 educational decisions. In common with previous research, our empirical work finds evidence that teacher assessments and test scores estimate current 'ability' in ways that diverge according to pupil background, particularly in relation to gender. Most strikingly, pupils who are high prior achievers tend to do better, and low-achievers worse in tests than would have been expected from teacher assessments at age 14.

This tendency is carried through into the links between pupil demographic characteristics and the gap between teacher and test-based assessment scores, with low-achieving groups performing relatively well in teacher assessments, and high-achieving groups performing relatively badly in teacher assessments. This evidence is not consistent with traditional forms of statistical discrimination or stereotyping, which would imply systematic over-assessment of high achieving groups and under-assessment of low-achieving pupil groups. 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.

The second question addressed by this research is whether bias and uncertainty in assessment has any bearing on pupils' subsequent academic achievement or staying on decisions. We find that, in itself, divergence between test-based and teacher-based assessment has no systematic adverse consequences for subsequent educational outcomes and participation. 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 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 3 explains the institutional context and the data used in the analysis and in Section 4 we discuss the empirical results. Section 5 provides conclusions.

2. Background and methods

Our first research goal is to explore empirically whether gaps between teacher-based and test-based assessment of pupils' levels of achievement differ along demographic lines, and along lines of prior achievement. We study these issues in the context of England's secondary education. Our second goal is to find out if divergence in assessment – interpreted as an indicator of assessment bias or uncertainty – has any link to pupils' subsequent achievement or propensity to continue in education. We consider the effects of divergence in assessment both in terms of the difference between test and teacher scores, and in terms of the absolute (unsigned) deviation between the outcomes of these modes of assessment.

A particular issue of concern in the literature has been the possibility of assessment bias arising as a result of stereotyping or statistical discrimination, whereby judgements about

individuals are made on the basis of what is expected of pupils of similar type. Such stereotyping can arise in examination contexts if examiners can deduce pupil race or gender from names or other information on test scripts, although Baird (1998) finds that A-Level marking is not highly sensitive to clues about candidates' gender. Generally, the existing evidence on test marking bias does not find strong evidence that it is pervasive (see Baird 1998 for a review).

Face-to-face assessment gives more cause for concern regarding stereotyping, and because it is sensitive to personal and subjective preferences on the part of the assessor, or to the specific relationship and interaction between pupil and assessor. With this in mind, a substantial literature in the fields of social psychology, education and economics has considered questions related to stereotyping and its implications for face-to-face assessment (for example, see Wright and Taylor 2007, Steele and Aronson 1995, Gipps and Murphy 1994, Reeves, et al 2001, Lavy 2004, Dee 2005a, 2005b, Ammermueller and Dolton 2006 and Ouazad 2007). Many of these studies are based on comparison of test-based and face-to-face assessments, and other studies have implicitly or explicitly considered similar questions in relation to teacher predictions and actual grades in tests and exams (Murphy 1981, Delap 1995, Thomas et al 1998, Dhillon 2005). Most of these studies indicate that teachers' assessments of pupils' academic abilities can differ from pupils' achievements in tests and exams in ways that are systematically related to ability, demographic and socioeconomic background. This finding is a worrying since it implies that some groups can be educationally disadvantaged simply by the type of assessment to which they are exposed. However face-to-face, teacher-based assessment clearly offer many advantages over tests, because the former are usually based on observation of ability on a wider range of tasks over a longer time horizon, whereas tests can only evaluate performance on a very limited range of questions on a specific day, can favour technique over underlying ability, and may favour one cultural background or gender. It is therefore unclear which of these modes of assessment provides the

most reliable or accurate assessment, and in the analysis that follows we give each mode equal weight.

The first part of the analysis is based on simple linear regression models, estimated by ordinary least squares, in which we regress the difference or absolute 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. In our case, we consider the assessments of pupils' English, science and maths ability at age 14 in England – referred to as the Key Stage 3 tests.

An important consideration is to what extent the teacher-test gap in scores varies across the distribution of prior achievement levels. 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 divergence along lines of average ability. Our approach to measuring prior achievement – in a way that is not biased in favour of teacher or test based assessment – is to use the average of teacher and test-based assessment scores from pupils' age 11 assessments (referred to as Key Stage 2 assessments), which children undertake at the end of primary school. By including the *mean* of teacher and test-based prior assessments as an explanatory variable in our models of the test-teacher *gap* in assessment, we can examine

[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.

how the gap varies with both observable pupil characteristics and with levels of prior achievement measured. Note that in this setting, the age 11 and age-14 assessments are made by different teachers and at a different phases of education.

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. Our approach to this task is to use least-squares regression models to estimate the relationship between the teacher–test gaps in pupils' age-14 assessment, 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.

3. Data and context

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 since 1996 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

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

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" (e.g. QCA 2004). Importantly however, the tests and teacher assessments seem intended to measure a pupil's current ability, knowledge and skills along the same dimensions in the same subject areas and the output of each mode of assessment is a comparable measure of pupil achievement scored in terms of National Curriculum "Levels". 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 definition of the teacher–test assessment gap is the difference, or absolute (unsigned) difference between the points awarded by the teacher in their assessment and the points awarded by examiners. Note that 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.

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

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

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

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 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

[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 (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.

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.

4. Results and discussion

4.1. Descriptive statistics

Table 1 presents some simple descriptive statistics for the data set used in our analysis. As explained in Section 3, 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, summarised in the top panel Table 1, is used in our analysis of the associations between pupil characteristics and the gap between teacher and test assessment scores. The second dataset, used to analyse whether these assessment gaps affect subsequent outcomes, 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 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 from a large body of research, 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 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.

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 4.2 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.

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

For this analysis, we turn to the regression approach outlined in Section 2. Our main findings are succinctly summarised in Figure 1, but the figure requires careful explanation and reading.

First, we estimate regression models with the *gap* between teacher and test based assessment scores in each core subject at Key Stage 3 as the dependent variable. The estimation sample is based on 1.4 million Year 9 pupils (aged 13-14) in 2002-2005. The key explanatory variables for which we report the coefficients are pupil ethnic indicators (black,

Asian, mixed, other, English additional language) demographic characteristics (age in months, gender) and an indicator of socioeconomic background and income (entitled to free school meals). We also show the coefficients corresponding to the variables for mean teacher and test scores received by the pupil in the subject assessment at age 11. Our regression models also include the following control variables: year dummies (0-1 indicators), an 'unknown ethnic group' indicator, and the gap between teacher and test scores in the pupil's Key Stage 2 assessments at age 11. In the results we report here we also control for fixed-over-time school specific factors (including teachers) that affect all pupils in a given school equally, that is we allow for secondary school specific "fixed effects" and estimate the regressions using the deviations of the variables from the school specific means.

Note that the regression specification implies that all the effects are measured relative to a *baseline group* of white girls with English as a first language, aged 13 and 0 months in September at the beginning of the year, not entitled to free meals, and with a mean age-11 score of 27 (corresponding to expected achievement of Level 4 in both teacher and test assessments). For this baseline group, the gap between teacher and test based assessments is about 6% of one standard deviation of the variation in achievement points in English and maths, with teacher scores below test scores, but the gap is effectively zero in science.

Next we estimate regression models with the *mean* of the teacher-based and test-based assessments as the dependent variable, using the same set of explanatory variables. Figure 1 then plots the coefficients from the three "*gap*" regressions against the coefficients from the three "*mean*" regressions, to show how the bias in the assessment varies with the expected achievement of the demographic, ethnic or socioeconomic group in question. Each data point in the graph is represented by a label that signifies the pupil group to which it corresponds: L3 represents pupils achieving Level 3 or less in both teacher and test assessments at age 11 Key Stage 2. L3+ represents pupils achieving a combination of Level 3 and Level 4 in the age 11 Key Stage 2 test and teacher assessments. L4+ represents a combination of Level 4 and Level

5, and Level 5 represents pupils achieving Level 5 in both modes of assessment at Key Stage 2. The other symbols are: F Free Meals, B Black, A Asian, X Mixed ethnicity, R Other ethnicity, L English additional language, M Male, O Old (birthday September). There are three data labels of each type, corresponding to the English, maths and science regression coefficients.

Figure 1 can be interpreted by recognising that data points in the top two quadrants of the diagram represent pupil groups who do relatively well in the teacher assessments and relatively poorly in the teacher assessments at age 14 Key Stage 3, referenced to the gap for the baseline group. Data points in the bottom two quadrants represent pupil groups who do relatively well in the tests. Thinking now about mean levels of achievement, data points in the right hand two quadrants represent pupil groups who have higher achievements, on average, in the combined test and teacher assessments at age 14 Key Stage 3, again referenced to the baseline group. Data points in the left hand quadrants represent lower achieving pupil groups.

The most striking feature of the chart is that there are some very substantial gaps between teacher and test scores at age 14 Key Stage 3 with respect to levels of prior achievement. Pupils scoring towards bottom of the distribution at age 11 Key Stage 2 (L3, L3+, top left quadrant) do relatively well on the teacher assessments at age 14 Key Stage 3, whilst their peers at the top of the achievement distribution (L4+, L5, bottom right quadrant) so relatively well in the tests. As an example, in English, the difference between the baseline group (Level 4 in both assessment) and pupils scoring Level 3 or less on the test and teacher assessments at age 11 corresponds to around 16% of one standard deviation of the achievement levels at age 14. A comparable gap of the opposite sign can be observed for pupils scoring Level 5 at age 11 in science.

In comparison with these results on prior achievement, the differences by free meal entitlement, ethnic group, and demographics are rather modest. There is no difference between low income pupils (“F”, on free meals, close to the horizontal zero line) and others in

terms of the teacher-test gap in assessment points. The coefficients are small and statistically insignificant at the 5% level. Neither is there any clear relationship in general for the ethnicity groups. Some of the coefficients are negative (lower quadrants) because the ethnic minorities tend to have significantly lower teacher assessments and higher test scores in English, when compared to white pupils. On the other hand, minority ethnic groups tend to have significantly higher teacher-test assessment gaps in maths (higher quadrants), but there are no statistically significant patterns in science. Pupils with English as an additional language do relatively poorly on teacher assessments in all subjects, but this is only statistically significant in maths (at the 5% level). One notable feature is that boys, compared to girls, do relatively well on teacher assessments in English, but relatively poorly in mathematics and science and these gender differences are always significant at the 1% level. The last two findings echo those in Reeves et al (2001) for age-11 assessments in 1998[8]. Lastly, older children seem to be rated relatively well in teacher assessments than tests, particularly in science. All these gaps are fairly modest, at most around 5% of one standard deviation in terms of achievement levels at age 14[9].

The general impression given by Figure 1 is of a strong general downward trend, with positive teacher-test assessment gaps for pupil groups who have low achievements (on the left hand side of the diagram) and negative teacher-test assessment gaps for high achievers

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

[9] 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. 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.

(on the right hand side of the diagram). We show this clearly using a trend line, fitted using a quadratic polynomial, showing that there is a general tendency for groups with higher than average achievement to have more negative gaps between teacher and test assessments (the teacher scores are lower than the tests) whilst groups with lower than average achievements score relatively well on the teacher assessments.

This pattern could suggest a general tendency for teachers to be fairly cautious, relative to the tests, in their ratings of pupils. Teachers give lower achieving groups favourable ratings relative to the tests, whereas the tests favour high ability groups. The net result is that the distribution of teacher assessments is compressed relative to the tests. Note, that although the variance of teacher assessments could also be lower than the variance of the tests because the former is based on repeated teacher observations of pupils over a long period (and hence potentially more precise), this cannot easily account for negative relationship between the gap and prior achievement seen in Figure 1. An important implication of Figure 1 is, however, that the results on ethnic and gender differences are rarely consistent with standard stories of statistical discrimination, or gender or ethnic stereotyping arising from face-to-face assessments. In our case, face-to-face assessments favour demographic groups with lower levels of achievement. Clearly, this is not a pattern we would expect to see if expected group achievement is being used to rank individual pupils[10].

We have carried out many additional analyses: checking for interactions between pupil achievement groups and ethnic, socioeconomic and gender groups; allowing for interactions between individual pupil characteristics and school composition; estimating models without

[10] The results are, however, partly consistent with the idea that external examiners are more susceptible to stereotyping and statistical discrimination than teachers, but only under the assumption that external examiners can deduce gender and ethnicity from the candidates names and test scripts. It is not at all clear how external assessors could detect "English as an additional language" status or prior achievement, so examiner stereotyping seems an unlikely explanation for all the patterns we observe.

school fixed effects, or with residential neighbourhood fixed effects; replacing the dependent variable with absolute deviations between test and teacher scores. None of these additional analyses change the general pattern observed in Figure 1. However, we do find some evidence that the teacher-test gap depends on both individual characteristics and school composition, and that differences between schools make a much bigger contribution (up to 6%) than individual pupil characteristics (less than 1%) to the overall variance in the gap between test and teacher assessment scores. Again these facts seem to reinforce the view that the patterns say more about teacher behaviour than about validity of external assessment.

Although it is difficult to gauge what precise mechanisms drive these findings, the results do highlight that teacher and test assessments in many cases divergence systematically across pupil groups, with group composition, and between schools (and hence teachers). This finding is potentially quite worrying, if it is expected that the test and teacher assessments should be arriving at broadly similar conclusions throughout the achievement distribution. However, as we have noted, a lower variance in the distribution of assessment scores from teachers, relative to tests, may be a natural outcome of being able to observe pupils over longer periods. 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 the opportunity to participate in higher education.

4.3. Impacts on qualifications and post-compulsory education

To start this part of the analysis, we first consider to what extent assessments at age 14 provide predictors of education outcomes at age 16 and beyond. To do this we estimate regression models of four age 16+ pupil outcomes. We report on four different educational outcomes at age 16 and beyond: 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); or 3) whether the pupils is recorded studying for any non-vocational post-16 qualification in the Independent Learner Record data set; and 4) whether

the pupil is recorded as staying on at school. The key explanatory variables in the first part of this investigation are the teacher and test assessment scores at age 14 Key Stage 3 in English, maths and science. We control for the full set of pupil demographic, ethnic and socioeconomic characteristics, age 11 Key Stage 2 achievement variables and other factors described in Section 4.2, plus our regressions allow for secondary school fixed effects.

The coefficients on age 14 Key Stage 3 assessments are always jointly and individually significant in these regressions, and Table 2 shows the correlations between the predictions derived from the Key Stage 3 assessment scores (using their corresponding regression coefficient estimates) and the dependent variable. As might be expected, both teacher assessments and test-based assessments are quite strongly positively correlated with pupil age 16-plus outcomes (conditional on each other, and on pupil past achievements and characteristics) indicating that both assessments contain unique information about the pupil achievements. In fact, there is not much to choose between these assessment modes as predictors of subsequent pupil outcomes. Having demonstrated that the assessments are correlated with later achievements and education decisions, we now go on to explore the central issue of whether *discrepancy* between these two types of assessment – implying bias or uncertainty in assessment of pupil achievement – has any relationship with age 16 achievement and post-16 decisions.

The best way to consider the specific impact of divergence between teacher and test assessments, and hence any influence on pupils arising from teacher perceptions, is to observe how pupil outcomes change as the *gap* between teacher and test scores widens, holding constant the average of the teacher and test-based assessments. Hence Figure 2 reports the coefficients from regressions of pupil age 16-plus outcomes on measures of the *gap* between teacher and test assessments in maths, science and English, at age 14 and 11. We do not report results for staying on at school since these are very similar to those reported for any post-16 academic participation. All the results are presented for regression specifications that include

controls for basic pupil characteristics (free meal entitlement, ethnicity, language, age and gender) plus dummy variables for prior achievement levels based on the sum of the teacher and test assessment point scores at age 14 and age 11[11]. The specifications also allow for school-specific fixed effects, but the results are insensitive to the inclusion or otherwise of these fixed effects. Solid shading indicates coefficients that are statistically significant at the 1% level. The dataset contains around 1.1 million pupils, aged 15-16 in 2002-2004.

To aid interpretation, we have standardised our coefficients in Figure 2 so that the height of the bar represents the association between a one standard deviation change in the teacher-test point gap, and the outcome measured in terms of standard deviations of the pupil distribution. Given this scaling, it is immediately clear that divergence between teacher and test assessment has very little impact on pupil age 16-plus outcomes, regardless of the fact that most of our coefficients are statistically significant.

Consider then the results for GCSE entries represented by the first of the three bars in each group. 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 re-emphasised that 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

[11] It in fact makes little difference whether or not we control for mean age 14 achievement.

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.

Although the findings on GCSE/NVQ entries might suggest that a more favourable teacher assessment engenders a positive academic attitude in pupils, this view is partially at odds with the findings in on GCSE/NVQ attainment. These results are shown by the second bar in each group of three bars in Figure 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, 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 are, broadly speaking, played out further in the results on the decision to stay on at school, or to pursue post-compulsory education more generally (the third bar in each group in Figure 2). 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 equal teacher and test assessments (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.

We have also considered effects on the share of English-related subjects, maths related subjects and science related subjects taken at age 16, and whether or not a pupil is recorded staying on at school in Year 12 but there seem to be no strong influences on these outcomes either. 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. We have also looked further to see if pupils that are assessed positively by teachers relative to tests experience more positive outcomes as the degree of over-assessment increases. We find occasional evidence of such non-linearities, but for the most part there are few significant differences of this type. 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. Lastly we looked at the effect of the absolute divergence between teacher and test scores to see if the magnitude rather than the sign of the gap matters, but find no evidence that it does, and so no evidence that uncertainty in age-14 assessment has a particular role to play in age 16+ education decisions.

In summary, although we have found some statistically significant effects, the results in Figure 2 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.

5. Conclusions

The goal of this research was to consider whether bias and uncertainty in assessment at school plays any role in generating patterns of age-16 achievement and post-school education participation, and hence whether assessment accuracy issues are relevant for policy that seeks to widen the participation of groups that are under-represented in higher education. This aim was motivated by previous research that has suggested that stereotyping in assessment has important consequences for pupils' educational outcomes.

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 smaller differences by gender and ethnic group, but these follow the general pattern that groups at risk for lower achievement tend to receive more favourable teacher assessments, whilst higher achieving groups do better in the tests. The reasons for these divergences between teacher and test based assessment scores are not revealed by our analysis. Statistical discrimination or stereotyping on the part of teachers 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).

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 our evidence that pupils are heavily influenced by teacher perceptions of their abilities or by any other form of bias in school assessment, or that these factors could be a major influence on post-16 pupil decisions or higher education participation rates.

6. 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

Baird, Jo-Anne (1998) What's in a name? Experiments with blind marking in A-Level examinations, Educational Research, 40 (1) 191-202

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

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

Delap, M. R. (1995) Teahers' estimates of candidates' performances in public examinations, Assessment in Education 2 (1) 75-92

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

Dhillon, D. (2005) Teachers' estimates of candidates grades: Curriculum 2000 Advanced Level Qualifications, British Educational Research Journal 31 (1) 69-88

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

Murphy, R. J. L (1981) O Level grades and teachers' estimates as predictors of the A-Level results of UCCA applicants, British Journal of Educational Psychology 51 () 1-9

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

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

Thomas, S., G. F. Maduas, A.E. Raczek, R Smees (1998) Comparing Teacher Assessmen and Standard Task Results in England: the relationship between pupil

characteristics and attainment, Assessment in Education, 5(2) 213-246

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

Figure 1: Relationship between teacher-test points gap and age-14 predicted achievement points by Key Stage 2 achievement level, demographic, ethnic and free-meal groups.

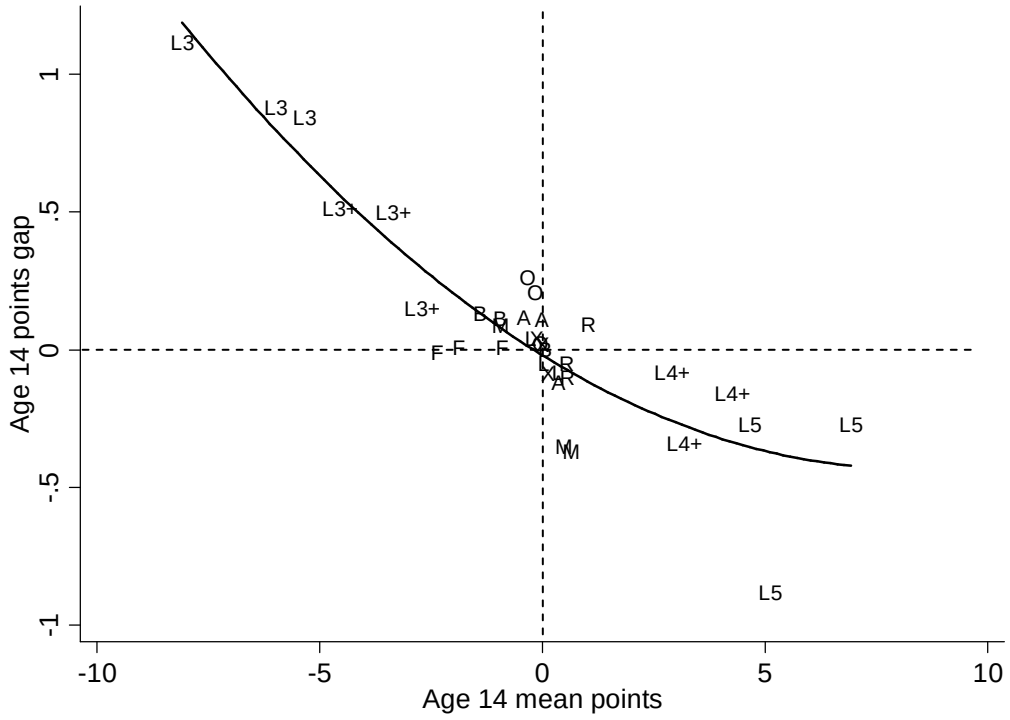
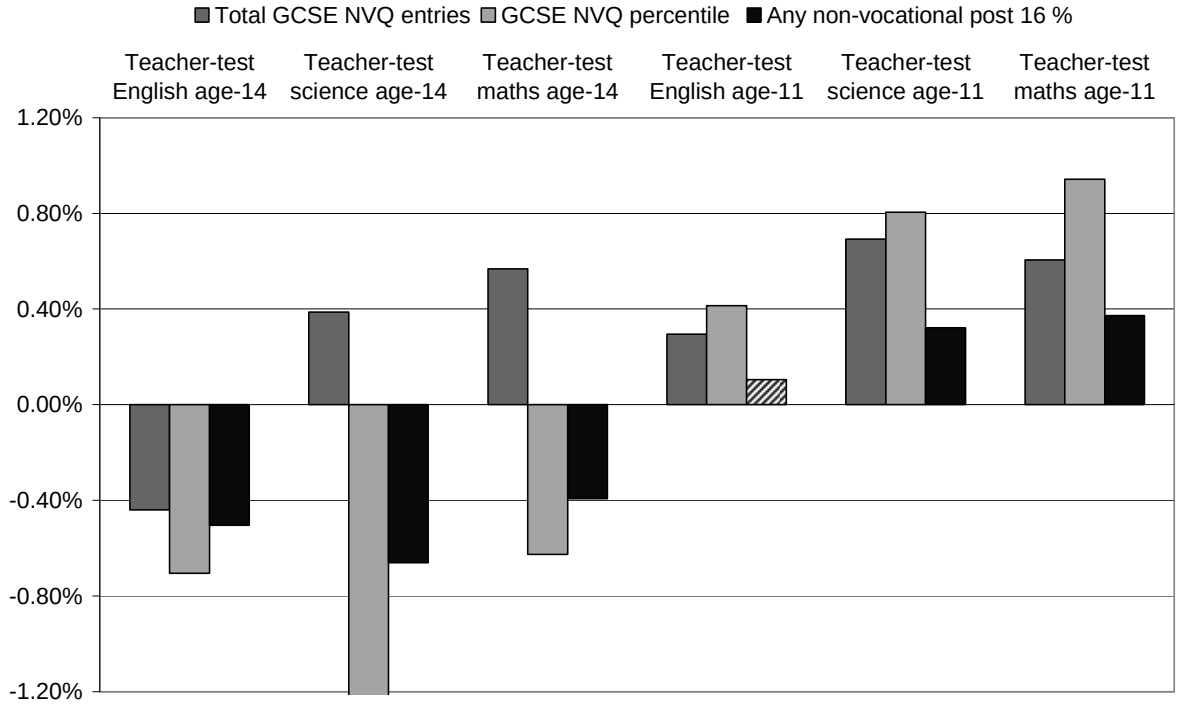


Figure plots coefficients from regression of teacher minus test points on pupil characteristics, against coefficients from regression of mean teacher and test points on pupil characteristics, as described in text. Key: F Free Meals, B Black, A Asian, X Mixed ethnicity, R Other ethnicity, L English additional language, M Male, O old (birthday September), L3 Level 3 at age 11 (both test and teacher assessments), L3+ Level 3-4 at age 11, L4+ Level 4-5 at age 11, L5 Level 5 at age 11. Baseline group is non-free meals white ethnicity, English first language, birthday in August, girls, scoring Level 4 on teacher and tests assessments at age 11.

Figure 2: Standardised association between teacher-test gaps in assessment and Age 16+ outcomes.



Notes: Figure reports regression coefficients from models of relationship between age 14 Key Stage 3 teacher-test assessment score gap and age 16+ pupil outcomes. Coefficients are based on regression models of the listed outcomes, with the gap between teacher and test Key Stage 3 assessment scores as explanatory variables. Control variables are indicators of pupil entitled to free meals, ethnic group, English additional language, gender, age in months, and year of observation. Specifications include indicators of achievement at age 11 Key Stage 2 and mean of teacher and test score at age 14, and allow for secondary school fixed effects.

7. Tables

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

	(1) Full sample	(2) Free meals	(3) Asian	(4) Black	(5) English additional	(6) Male
Age-14 sample						
Age 14 mean English teacher and test points	34.87 (5.11)	32.19 (4.96)	33.90 (4.97)	33.72 (4.91)	33.93 (5.02)	63.94 (5.12)
Age 14 mean science teacher and test points	34.97 (5.5)	31.89 (5.31)	33.25 (5.69)	32.83 (5.34)	33.41 (5.73)	35.27 (5.49)
Age 14 mean maths teacher and test points	37.23 (6.63)	33.83 (6.44)	36.16 (6.86)	34.74 (6.54)	36.2 (6.87)	37.64 (6.63)
Age 14 English teacher-test points gap	-0.294 (4.477)	-0.212 (4.706)	-0.599 (4.648)	-0.478 (4.581)	-0.578 (4.650)	-0.214 (4.551)
Age 14 science teacher-test points gap	-0.082 (4.014)	0.122 (4.223)	0.021 (4.268)	0.045 (4.262)	-0.012 (4.276)	-0.267 (3.998)
Age 14 maths teacher-test points gap	-0.336 (3.550)	-0.160 (3.716)	-0.280 (3.773)	-0.150 (3.756)	-0.304 (3.761)	-0.552 (3.540)
Observations	1439409	172352	81231	37882	107979	683945
Age 16 sample						
Total GCSE NVQ entries x 10	9.808 (1.585)	9.363 (2.019)	10.08 (1.480)	9.786 (1.600)	10.07 (1.503)	9.767 (1.647)
GCSE NVQ percentile	52.48 (27.43)	37.41 (26.25)	55.11 (26.74)	45.93 (26.31)	54.91 (26.94)	49.36 (27.30)
Any non-vocational post 16 %	74.24 (43.73)	65.02 (47.69)	86.22 (34.46)	82.85 (37.69)	85.42 (35.29)	72.28 (44.76)
Observations	1015446	105231	58642	24089	75271	485245

Notes: Table reports means. Standard deviations in parentheses

Table 2: Teacher and test based assessments at Key Stage 3 as predictors of age 16+ pupil outcomes. Partial correlations between regression predictions and outcomes.

	Total GCSE/NVQ entries	GCSE/NVQ percentile	Any non-vocational education post 16	Stays on at school
Prediction from age-14 tests	0.352	0.791	0.335	0.339
Prediction from age 14 teacher assessments	0.349	0.786	0.329	0.337

Table reports correlations between predictions from age 14 Key Stage 3 assessments and age 16+ pupil outcomes. Predictions are based on regression models of the listed outcomes, with age 14 Key Stage 3 assessment point scores in English, maths and science as explanatory variables. Control variables are indicators of pupil entitled to free meals, ethnic group, English additional language, gender, age in months, and year of observation. Specifications include indicators of achievement at age 11 Key Stage 2 and allow for secondary school fixed effects. Predictions correspond to the baseline group of non-free meals white ethnicity, English first language, birthday in August, girls, scoring Level 4 on teacher and tests assessments at age 11.