Stereotyped at seven? Biases in teachers' judgements of pupils' ability and attainment

This is the *Accepted Manuscript* version of:

Campbell, T. (2015) *Stereotyped at Seven? Biases in Teacher Judgement of Pupils' Ability and Attainment.* Journal of Social Policy, 44:3, pp 517-547.
http://dx.doi.org/10.1017/S0047279415000227

Above final version Copyright © Cambridge University Press 2015.

Abstract

There is evidence that teacher judgements and assessments of primary school pupils can be systematically biased. This paper tests the proposal that stereotyping plays a part in creating these judgement inequalities and is instrumental in achievement variation according to income-level, gender, special educational needs status, ethnicity, and spoken language. Using 2008 data for almost 5000 pupils from the Millennium Cohort Study, it demonstrates biases in teachers' average ratings of sample pupils' reading and maths 'ability and attainment' which correspond to every one of these key characteristics. Findings go on to suggest that stereotypes according to each of income-level, gender, special educational needs status, and ethnicity all play some part in forming these biases. The paper strengthens the evidence that stereotyping of pupils may contribute to assessment and thereby attainment inequalities, and concludes that an increased focus on tackling this process may lead to greater parity and a narrowing of gaps.

Key words: Teacher assessment, stereotyping, attainment gaps, bias, inequalities, primary education.

Acknowledgements

This study uses data from the Millennium Cohort Study. I am grateful to the Centre for Longitudinal Studies, Institute of Education for the use of these data, and to the UK Data Archive and Economic and Social Data Service for making them available. However, these organisations bear no responsibility for the analysis or interpretation in this paper. Thanks also to the Economic and Social Research Council for the PhD funding which enabled this analysis.

Many thanks to Lucinda Platt, Lorraine Dearden, Ellen Greaves, and Meghan Rainsberry for comments, suggestions, and critique.

Introduction

Teacher assessment and pupil attainment

Since the introduction of the National Curriculum in 1988, the time dedicated to standardised assessment of English pupils has increased considerably, alongside a growing requirement that much of this assessment be performed by class teachers. Teacher judgements currently dominate children's designated attainment levels within primary education. At the time of writing, the Foundation Stage Profile (FSP; covering the years up to age five) is entirely teacher-assessed, along with the newly-introduced phonics screening test (taken at ages six and seven), and Key Stage One (KS1) attainment (age seven). Primary education culminates at age 11 with the awarding of Key Stage Two (KS2) grades, which in part comprise the results of external examinations, but which also incorporate ratings by teachers (Bew, 2011a; 2011b; Wyse *et al.,* 2008).

This approach to assessment, with its reliance on an understanding of each child built over time rather than based simply on one-off performance in a set test, has several arguable advantages. It avoids the lack of nuance of the one-shot test, and also the test's time-and place-dependency, which might result in an inaccurate picture of a child's abilities should they underperform on a given day, in the given situation, or in response to the limited test stimuli (Harlen, 2007). Some evidence indicates moreover that formalised testing can be stressful and demotivating for pupils (Harlen, 2004; 2007), and it has also been suggested that exams may be counterproductive to meaningful knowledge acquisition insofar as they encourage 'teaching to the test' at the expense of deeper, sustainable learning and wider exploration (Harlen, 2007; Wyse *et al.,* 2008). However, despite its potential advantages over more formalised and 'objective' measures, teacher assessment is not, in itself, entirely unproblematic.

The past decade's national statistics on the performance of English pupils have consistently indicated that certain groups achieve at lower levels than others throughout their early education. Low-income pupils in receipt of free school meals (FSM), pupils with any diagnosis of special educational needs (SEN), Pakistani, Bangladeshi, Black African, and Black Caribbean pupils, and pupils speaking English as an additional language are regularly reported as under-attaining in the primary phase. In addition, boys score generally at a lower level than girls at the foundation stage, though they attain higher levels at maths (and girls at English) at KS1 and KS2 (Department for Education 2012a; 2011; 2012b).

That attainment indicators depend so heavily on teacher assessment invites the question of whether these apparent achievement gaps may to some extent be an artefact of the measurement method used. There is an enduring body of evidence which indicates that teacher assessments are subject consistently to a large and significant level of error (Brookhart, 2013; Eckert *et al.*, 2006; Harlen, 2005), and, more importantly, research also indicates that some of this error may be systematic (Harlen, 2005; Robinson and Lubienski, 2011), and that there may be regular patterns of inequality in teacher judgements of English primary school pupils (Burgess and Greaves, 2009; Reaves *et al.*, 2001; Thomas *et al.* 1998).

Bias in teacher assessment

For example, examining national KS2 data, Burgess and Greaves (2009) exploit the distinction between the teacher-assessed and externally-examined components of the test, comparing marks awarded to pupils according to the two measures. They demonstrate disparities in teacher assessment which are in line with nationally-reported attainment gaps: seeming under-assessment of pupils in receipt of FSM, of pupils with SEN, and of Black Caribbean and Black African pupils. This suggests that teacher-level bias may serve to inflate and deflate the overall KS2 scores allocated to each pupil.

Analysing the English sub-sample of the Millennium Cohort Study (MCS), Hansen and Jones (2011) indicate that teachers may also be biased in their assessments of pupils at the beginning of primary school. They compare children's FSP scores to self-completed cognitive tests taken outside of school, and find greater disparities according to gender in the teacher-assessed FSP measure than in the child-completed tests. Teacher assessments pronouncedly favour girls to a greater extent than cognitive test performance, indicating that gender disproportionality at the foundation stage may, like inequalities at KS2, be attributable in part to biased judgements.

Qualitative research, some of it government-commissioned, has moreover begun to suggest mechanisms that might underpin these apparent biases in assessment and resultant attainment, particularly with regard to ethnic disparities. Evidence that perceptions and behaviours among teaching staff may play a part in creating variation has been provided by Maylor *et al.*'s (2009) evaluation of the Black Children's Achievement Programme, which concludes that, 'Institutional factors / processes including negative teacher attitudes / expectations' and 'stereotypical thinking about the ability of Black children serve to undermine teacher ability to raise Black children's attainment at an individual and group level' (p 2).

Similarly, Strand *et al.*'s (2010) investigation into *Drivers and Challenges in Raising the Achievement of Pupils from Bangladeshi, Somali and Turkish Backgrounds* reports that: 'Racism and structural inequalities may be important influences on the attainment of many Bangladeshi and Somali students' (p 18). As also suggested by Burgess and Greaves' large-scale quantitative work, these studies indicate that stereotyping at the teacher-level may provide some explanation for the ostensible attainment differentials among primary school pupils.

Biased assessments through stereotyping

There are a number of theories of what stereotypes *are*, and of behaviours associated with their presence. Many are grounded in the premises that stereotypes comprise invariant, homogenous, evaluative judgements of a given group (e.g. income, gender or ethnic group), and that stereotypes enable judgements of group members to be made quickly and with cognitive ease (Hilton and von Hipple, 1996; McGarty *et al.*, 2002.) By stereotyping, therefore, teacher judgements of pupils can be made quickly and with cognitive efficiency (though with compromised *accuracy*) based, in part, on a preconceived 'template' of the ability and attainment of low-income pupils, pupils with SEN, White pupils, Black Caribbean pupils, and so on. Stereotyping is not assumed to take place on a conscious or deliberate level: the process's efficiency is thought to be engendered by its automaticity.

Theorists argue furthermore that stereotypes must be held at the group or institutional level: '…stereotypes should be formed in line with the accepted views or norms of social groups that the perceiver belongs to' (McGarty *et al.*, 2002, p 2). The possibility, therefore, is that among the English teaching profession there exist normalised notional templates of pupil attainment, which are premised on pupil characteristics, inform judgements of each child, and skew assessments in line with these characteristics.

Building upon previous evidence to test the stereotype model

To date, little credence or focus appears to have been afforded in the policy arena to the possibility that bias and stereotyping might provide some explanation for systematic variation in children's achievement, particularly in primary school. Despite the growing body of evidence that this may be the case, policy has tended to look instead to the family-level for first causes of inequalities, often citing socio-economic differences as the primary driver, and directing resources accordingly (Department for Children, Schools and Families, 2008; Department for Education, 2010a; Department for Education 2010b; Department for Education and Skills, 2005). Yet if the process of stereotyping can definitively be implicated as instrumental in biases in teacher assessment (and consequentially as contributing to attainment disparities), this will clearly indicate a point at which intervention to mitigate these inequalities might be deployed.

However, existing research does not yet unequivocally support the theory that pupils are being stereotyped by their teachers. For example, though they show clear patterns of disparity, and though they propose and support a stereotype model, Burgess and Greaves (2009) also acknowledge an alternative explanation for their findings. Because their analysis uses comparators from within the same overall system (the teacher who assesses the pupil at KS2 also teaches them for the externally-marked KS2 test), there is a danger of causal explanatory relationships within the system. Burgess and Greaves suggest, for instance, that the notable difference between the teacher-assessed and externally-marked elements of SEN pupils' results, in particular, may be due to: '…an extreme form of "teaching to the test" for pupils with SEN…the teacher's more in-depth knowledge of the student's ability may result in a lower [teacher assessment]' (p 12). That is, teachers might explicitly train and focus on certain pupils, whom they see as less able, so that they learn to attain desirable KS2 levels in the test situation. As a result, these levels may not reflect the teacher's day-to-day perception of the pupil's ability – and this, rather than stereotyping, may be what underpins apparent biases.

Hansen and Jones' (2011) analysis partially circumvents this issue and avoids interrelatedness of measures by utilising tests of pupil 'ability' which are not explicitly associated with their schooling, and not directly influenced or reported by their teacher. Cognitive tests independently administered in children's homes as part of the MCS are compared to school-based, teacher-assessed FSP scores, arguably providing an enhanced indication that teacher judgements are biased away from manifest pupil performance.

However, while Hansen and Jones' study strengthens the evidence that recorded teacher assessments are systematically skewed, a danger remains that FSP scores do not in fact comprise direct portrayals of the mental representation – the potentially stereotype-based 'evaluative judgement' – that each assessing teacher holds of their (groups of) pupils. Because schools themselves, at the institutional level, are judged by the attainment of their pupils, and because teachers' own performance is assessed according to the attainment of their class, it is highly likely that FSP scores serve not only to describe the teacher-perceived attainment or progress of each individual child, but to inform additional purposes (Bradbury, 2011a, Harlen, 2007).

A recent report by Ofqual (2012) noted, for example, a tendency within teacher assessment to manipulate 'marks so that candidates [are] placed within certain perceived grade boundaries' (p 82), and recent reporting of national scores for the teacher-assessed phonics screening test clearly illustrates this phenomenon (Department for Education, 2012c, p4).[i] One response to a 2009 Ofsted consultation stated that: 'Schools can manipulate…scores in ways that Ofsted would be unlikely to support,'[ii] while Bradbury (2011b) describes findings from case studies where 'assessment results may be influenced by pressure from external advisors, who only recognise certain patterns of results as intelligible,' and where this moderation brings about amendments to pupils' test scores in line with established normative expectations (p 655). Recorded FSP results may, therefore, provide a somewhat inaccurate representation of teacher perceptions of a given individual or group, due to their complicity with, and the incentives of, a system where the attainment levels awarded to pupils have implications far beyond measuring and assessing each child's ability, progress or performance.

The current study

Therefore, in order to investigate more unambiguously and explicitly whether teacher-level stereotyping of pupils may relate to biased assessment according to pupil characteristics, the analysis presented in this paper uses a measure of teacher judgement which is not part of nor required by the education and assessment system, which is removed from its context, and which will not inform evaluations of performance of a teacher or their school. Confidential responses provided by teachers participating in the Millennium Cohort Study (MCS) to questions about their pupils' 'ability and attainment' (at age seven) provide a proxy for the teachers' mental representations of each pupil. These survey responses should lack the agenda inherent to the formal in-school assessments used in previous research. In addition, like Hansen and Jones' paper, the current study uses independent MCS-administered cognitive test scores (also collected at age seven) as comparators that indicate each child's contemporaneous manifest performance.

Analysis explores whether there are biases in teacher judgements of pupils which correspond to each of the key pupil characteristics underpinning recorded primary-age attainment gaps (family income-level, gender, SEN, ethnicity, EAL) and which may, as proposed, account to some extent for these gaps. Additionally, it begins to explore which of these characteristics appear to dominate and drive any apparent biases, in order further to inform potential interventions which may tackle stereotyping.

Methodology

Sample

The Millennium Cohort Study (MCS) included 11,695 English children at its first sweep in 2001, and four additional waves have taken place to date: in 2004, 2006, 2008, and 2012. This paper uses data from wave four, when the pupils were seven years old, and in year two at primary school (Plewis, 2007; data source: University of London 2011; 2012). Analysis is restricted to state school children in England, in order to allow comparison with and interpretation in the context of Department for Education (DfE) statistics on pupil attainment. Twins and triplets are excluded, because teacher bias and stereotyping may follow a different process for these pupils.

Responses to the survey of teachers at wave four which provide the data used here were received for only a sub-sample of pupils. Data on cognitive test scores is missing for a small minority, and there is also some non-response to individual questions. The base samples thus comprise those 4997 / 4985 (reading / maths) MCS children who continued to participate at wave four, whose teachers responded, and for whom there are all necessary data. Any relationships found in the current analysis can therefore be attributed with absolute certainty only to the children included – but this large sample can be used to theory-build and to explore the hypothesis that stereotyping by teachers takes place.

Estimates are weighted for the MCS's design features and for attrition to the level of the main wave four sample, as per Mostapha (2013) (attrition weights specifically for the teacher sample have not yet been developed). Annex 1, Table 11, compares key characteristics of the English singleton MCS sample at wave one to three samples at wave four: that with teacher survey response, that with all data necessary for analysis in this paper, and that without teacher response. It also contrasts estimates with and without attrition weights. It suggests some relatively minor differences between samples: that the sample used in this paper are from families slightly less often low-income than those without teacher response at wave four, who are more likely to speak only English at home; that the pupils are more often of White ethnicity, score marginally higher in the cognitive tests, and are slightly more often girls. Where comparison across waves is possible, estimates weighted for design and attrition are similar for the wave one sample and for the sample used in this paper.

Teacher judgements

Teacher-reported judgements of whether each pupil is 'well above average / above average / average / below average / well below average' at both reading and maths, respectively, form the crux of analysis. These evaluations are in response to a survey question asking the teacher to 'rate [the given] aspect of the study child's ability and attainment [reading / maths]…in relation to all children of this age…'[iv] For modelling, responses are recoded into binary variables representing a rating of 'above' or 'below' average, which indicate whether each child is judged as relatively more or less able, compared to their peers. Responses of *well above average* and *above average* are combined to form the 'above average' category, where all else is categorised 'not above average;' similarly, responses of *well below average* and *below average* are combined to one 'below average' category. While it necessitates a coarser analysis of biases, this merging of responses allows use of an easily interpretable linear probability model, and ensures robust cell sizes in logistic modelling. Four outcome variables are thereby created:

- teacher judgement of reading 'above average' / not;
- teacher judgement of reading 'below average' / not;
- teacher judgement of maths 'above average' / not;
- teacher judgement of maths 'below average' / not.

Pupil characteristics

In addition, the following measures of each of the pupil characteristics identified by DfE statistics as underpinning attainment variation are used (all are taken at wave four):

- a derived variable from parent-reported data which indicates whether the family's income is above / below an OEDC 60% of median UK income poverty indicator;
- parent-reported pupil gender;
- teacher report of any recognised SEN (yes / no);
- a derived variable from parent report denoting pupil ethnic group (White / Indian / Pakistani / Bangladeshi / Black Caribbean / Black African);
- a derived variable from parent-reported information on language(s) spoken in the pupil's household (coded to represent English only / additional languages).

Only sub-sets of breakdowns by ethnicity are reported in this paper, in order to aid meaningful interpretation and comparison with DfE statistics. The census-based eight-category ethnicity categorisation is used throughout analysis, and includes 'other' and 'mixed' classifications – but results for these groups are not presented. Descriptive statistics according to ethnicity may therefore not sum to 100%, while in modelling, noted sample sizes are for the whole sample with ethnicity data – as all are included in analysis – although only results for selected groups are outlined.

Teacher judgements and pupil characteristics

Table 1 shows the percentage of MCS pupils with each characteristic who are evaluated by their teacher as relatively more or less able than their peers, according to the definitions described above. It indicates a lower chance of being evaluated as 'above average' at reading for low-income pupils, boys, pupils with SEN, pupils of all ethnicities except White and Indian, and pupils speaking languages in addition to English. The same pattern holds for judgements of maths ability, save for a reversal according to gender, with boys more highly rated here.

****

Table 1 about here

****

Cognitive test scores

At age seven, the MCS children completed a number of cognitive tests during a home visit from a survey administrator. They included the British Ability Scale Word Reading test, and a shortened version of the Progress in Mathematics test. The Word Reading test is designed to assess children's English reading skills (see http://www.glassessment.co.uk/products/bas3). The ability score (a scaled but not otherwise standardised score) is used in analysis (see Hansen, 2012). The Progress in Mathematics test is designed to measure pupils' mathematical ability across use of numbers, shapes, and proficiency in data handling. It is intended to provide an indication of performance in maths at the given developmental stage (see http://www.gl-assessment.co.uk/products/progress-maths). A shortened version was used in the survey and entailed routing to sections of varying difficulty levels. Rasch scaling was used to convert the raw scores to a count score equivalent to that which would be attained were the full test completed (see Hansen, 2012). This scaled score is used in analysis here.

Performance on the two cognitive tests provide respective points of comparison to the teacher assessments of pupil reading and maths 'ability and attainment.' Completion of the cognitive tests shortly preceded teacher completion of their survey: the mean average time lag between cognitive test and teacher survey was 3.8 months, the median 3 months, and the mode 2 months. Comparisons using the two measures necessitate assumptions: a) that the lag between pupil test completion and teacher survey completion does not vary systematically across the pupil characteristics of interest; and b) that children delineated by each of the characteristics of interest develop at equivalent rates in their reading and maths ability and performance, at age seven (so that any apparent bias in teacher assessments cannot be attributed to slower progress during the time lag from pupil survey to teacher survey in some groups). The second of these assumptions cannot explicitly be tested using the MCS data, so remains a supposition (though as the modal time lag was short, at two months, it seems reasonably unproblematic); that the first hold is checked through additional analysis, reported later.

Test scores and pupil characteristics

Table 2, below, shows the mean Word Reading scores and Progress in Maths scores for the samples of pupils who took the tests and who also have responses to the teacher-completed question on reading / maths ability (respectively), according to each characteristic of interest.

****

Table 2 about here

****

On average, sample girls' scores on the Word Reading test are higher than boys', pupils with SEN have lower scores than those with no recognised SEN, and mean scores for low-income and Black Caribbean pupils are also relatively low. Pupils speaking languages in addition to English have higher reading scores, on average, than pupils speaking only English, and Indian, Bangladeshi and Black African pupils also have comparatively high scores.

Though measured on different scales and not, therefore, directly comparable, these descriptive statistics begin to indicate incongruities between children's cognitive test scores and judgements by their teachers. Sample pupils speaking languages in addition to English appear more likely to score relatively well on the BAS Word Reading test – but are less likely than pupils speaking only English to be rated highly at reading by their teacher. Similarly, Black African and Bangladeshi pupils score relatively highly on the Word Reading test – but are again less likely to be judged 'above average' and more likely to be judged 'below average' by their teacher

As with Word Reading scores, Table 2 indicates that sample pupils with SEN and low-income pupils are more likely to attain relatively low scores on the Progress in Mathematics test. In contrast to Word Reading, however, pupils speaking languages in addition to English score lower, on average, than pupils speaking English only, and pupils of all reported ethnicities except for White and Indian are relatively more likely to attain a lower score on this test. Mean scores for boys and girls are very similar, which again indicates some discrepancy between scores and teacher judgements of pupils' maths ability, which showed a tendency to favour boys (Table 1).

Modelling: Are some groups of pupils systematically over / under-rated by their teachers?

That there are apparent incongruities between average scores of pupils with varying characteristics for the Word Reading test and teacher judgements of reading 'ability and attainment' begins to support the possibility that there may be biases in teacher perceptions of pupils according to the pupils' characteristics. In order explicitly to investigate this, regression modelling compares teacher judgements of pupils who differ according to a given characteristic but who score at the same level on the relevant cognitive test.

The methodology here relies on a general overall relationship, across the sample, between performance on each cognitive test and teacher assessment of pupil 'ability and attainment' in the relevant domain. This relationship is strong. Within the whole sample, a naïve regression of BAS Word Reading test score on whether a pupil's teacher perceives their reading as 'above average' indicates that each additional point scored on the Word Reading test (range 10-214) is related to a likelihood of being judged 'above average' increased by 1.1 percentage point (p <. 001). For teacher judgements of reading 'below average', the relationship is inverted and there is a decrease of -.8 of a percentage point (p < .001). The relationship between point increase in Progress in Maths score (range 0 – 28) and judgement of 'above average' in maths is 4.2 percentage points (p < .001). For judgements below average it is -3.4 percentage points (p < .001).

Figure 1 presents the means and distributions of BAS Word Reading test scores for pupils judged to be at each level of reading 'ability and attainment' by their teacher, and Figure 2 presents the equivalent information for maths scores and judgements. These figures again illustrate, across all sample pupils, overall linear associations between test scores and teacher judgements. Pupils with a higher cognitive test score tend to be judged to have a higher level of 'ability and attainment' by their teacher, though this is not a prefect relationship, and there are also overlaps.

****

Figure 1 about here

****

****

Figure 2 about here

****

If there are no biases in teacher judgements according to the pupil characteristics of interest, these associations should not vary, nor the imperfection of the relationship be explained, by income-level, gender, SEN status, language, or ethnicity. Girls and boys, for example, who score at the same level on the Word Reading test, should have equal probabilities of being judged 'above average' at reading by their teacher.

A linear probability model is used to test whether this is the case. The outcome (for example) is whether a child is judged 'above average' at reading, and the predictors: pupil gender, and ability score on the reading test. The likelihood of boys being judged 'above average' at reading by their teacher is thereby compared to the likelihood of girls who score at the same level. Analysis takes the following form:

$$\begin{aligned} &Probability\ of\ being\ judged\ `above\ average'\ at\ reading\ by\ teacher_{0-1} \\ &\qquad = Constant + \ \beta 1 Boy_{0/1}\ +\ \beta 2 BAS\ word\ reading\ score + error \end{aligned}$$

The coefficient for boys represents the percentage point difference in likelihood, compared to girls who score equivalently on the Word Reading test, of being judged 'above average.' A coefficient of *0* would therefore indicate that there is no bias according to gender in teacher assessments of reading ability. A positive coefficient indicates a positive bias for boys, and a negative coefficient a negative bias.

Analysis is repeated separately for each pupil characteristic and outcome, resulting in the following basic models (Table 3). All analyses use Stata (versions 12 and 13), and include “svy” commands for weighting, unless otherwise stated.

****

Table 3 about here

****

Results

Biases in teacher judgements of pupils’ reading ability

Table 4 indicates variation in the average likelihood of MCS pupils who differ according to each characteristic (income-level, gender, SEN status, ethnicity and language) being rated relatively highly at reading, compared to peers who score equivalently on the Word Reading test. As described in Table 3, separate models were estimated for each characteristic, and findings from each discrete model are presented.

Children from low-income families, boys, pupils with any recognised diagnosis of SEN, and children who speak other languages in addition to English appear less likely to be judged ‘above average’ at reading by their teacher – despite scoring equivalently to their comparison counterparts in the reading test. All these differences are significant at p < .05 at a minimum. MCS pupils of all non-White ethnicities also appear less likely to be judged ‘above average’ at reading (compared to White pupils), and differences from the White reference group are, again, highly significant for most.

****

Table 4 about here

****

Separate models estimate the likelihood of each pupil group being judged ‘below average’ at reading, and these result are presented in Table 5. They are entirely in line with findings ‘above average,’ inverting the direction of effect. As well as being 11 percentage points less likely to be rated ’above average’ by their teachers, for example, low-income pupils are 8.3 percentage points more likely to be judged ‘below average,’ and again, this is highly significant.

****

Table 5 about here

****

Biases in teacher judgements of pupils' maths ability

In line with the lesser incongruity within the descriptive statistics, slightly fewer disparities emerge for maths (Table 6). No significant difference in teacher perceptions is found between MCS pupils speaking only English / speaking an additional language, and pupils of most ethnicities are as likely as White pupils scoring at the same level on the Progress in Maths test to be evaluated as 'above average.'

However, inverting the relationship indicated for judgements of reading 'above average,' boys are *more* likely than girls to be judged relatively highly at maths. Sample Black Caribbean pupils are significantly less likely than their equivalently performing White counterparts to be judged 'above average' – along with children from low-income families, and those with any recognised SEN.

****

Table 6 about here

****

Again, separate models estimate the likelihood of each pupil group of being judged 'below average' at maths (Table 7), and though more results again are non-significant here, those significant at $p < .05$ are entirely in line with findings 'above average.' Pupils with any diagnosis of SEN are *more* likely to be judged as 'below average' at maths compared to those without a diagnosis, low-income pupils are more likely than higher-income pupils, and Black Caribbean pupils more likely than White pupils.

****

Table 7 about here

****

Which characteristics underpin biases in judgements of reading and maths ability?

In order to begin to assess which characteristics might be important in driving these apparent biases and which stereotypes might be pertinent, analysis now incorporates each predictor variable simultaneously in a comprehensive model, and is repeated separately for teacher judgements of reading and maths 'above average.' The sample is then split between boys and girls to investigate any variation in patterns according to gender. Table 8 presents reading results for the whole sample, followed by findings for boys and girls, respectively.

****

Table 8 about here

****

Though there is a general lessening in the magnitude of biases for each characteristic, all remain significantly related at the 5 percent level to teacher judgements of sample children's reading, even when covariates are accounted for - though disparities by ethnic group appear to be moderated by the other factors, and language spoken is significant only for boys. Biases according to income-level and ethnicity appear generally to be stronger for girls, while, overall, boys remain assessed at a relatively lower level.

Table 9 presents results for teacher judgements of maths 'above average.' It suggests that gender may be key to teacher judgements of the maths ability and attainment of sample pupils (given the larger significant coefficient here than when gender is considered alone, without covariates [Table 6]). SEN status and income-level also remain significant predictors here, but biases for Black Caribbean boys and Black African pupils seem to be moderated by the covariates, and are non-significant. Accounting for confounders also renders the relationship between spoken language and teacher ratings non-significant, and in contrast to analysis for reading, there is some suggestion that biases in judgements for maths according to SEN status may be stronger for boys – though, overall, boys are more likely to be judged 'above average' at maths.

Across these analyses for reading and for maths there therefore appears to be some degree of bias according to each of four factors: income-level, gender, SEN status, and ethnicity – even accounting for every other factor, and for language spoken. Some differences in magnitude and significance are revealed according to gender among the MCS children, and relationships vary by academic domain. It seems, therefore, that stereotyping according to each of these four characteristics might underpin biases in teacher judgements of pupils, but that it may follow different trends according to subject area and gender.

****

Table 9 about here

****

Robustness checks

Four discreet robustness checks have been carried out to ensure that choices in modelling, weighting and sample selection have not influenced overall findings. Firstly, analyses are repeated using binary logistic rather than linear probability models. Results are equivalent (1). Secondly, analyses are repeated incorporated controls for age at cognitive test and time lag between test and teacher survey. This reduces sample sizes due to missing data on the timing variables, but makes little difference to the direction, significance or magnitude of findings (2).

As the pupils in the teacher sample are unevenly distributed across schools, and because some schools have several pupils and others only a single child, an additional check is carried out to examine whether extreme groups of teachers in more populous schools may be driving results. The pool for analysis is restricted to one pupil per teacher, and few differences are made to overall findings (3). Lastly, analysis is carried out *without* the wave four main sample weights, but *with* clustering at the school level, Again, the overall findings hold.

As an example, Table 10 shows key coefficients across these different analyses (1-4) for teacher judgements of reading 'above average,' according to the first model (Table 4).

Further checks and model specifications may be found in **Author's name** (** publication date** [working paper]). These include interactions of quintile versions of the cognitive test scores with each characteristic of interest, and reduced samples comprising pupils exclusively from MCS sample strata with relatively high numbers of minority ethnic and low-income families (findings hold for children in these areas).

****

Table 10 about here

****

Summary and discussion

Analysis set out to explore the possibility that biases in teacher judgements of pupils may result from stereotyping within the teaching profession and that these biases might contribute to variation in recorded attainment among primary school children. It finds that, in this MCS English sample of seven-year olds, there are inequalities in teacher perceptions of pupils' reading and maths 'ability and attainment' which correspond to the characteristics delineating these achievement gaps. On average, low-income pupils seem to be under-rated by their teachers, along with pupils with any SEN diagnosis, non-White pupils, pupils speaking languages in addition to English, and boys (reading) / girls (maths). Because both independent measures of pupil test performance and indicators of teacher perceptions of pupils which are not required by or implicit with formal in-school assessments are used in this paper, findings support the possibility that the socio-cognitive process of stereotyping may indeed be instrumental in systematic attainment differentials.

Results here are congruent with previous research indicating over- and under-assessment of pupils according to their characteristics. They provide enhanced support for the possibility that mechanisms beyond the level of the individual pupil and their family, and outside of the control of the child or their parents, appear to be at work determining assessment levels awarded. Unless these tendencies are addressed, they may continue to play some part in creating and perpetuating inequalities.

Analysis here also began tentatively to unpick the constitution of the stereotypes proposed to explain biases. It finds that income-level, gender, SEN status and ethnicity all appear to play a part in accounting for disparities in judgement of sample pupils, and that there is some variation by gender and by subject domain. This suggest that any intervention aimed at alleviating stereotyping and its effects on teacher perceptions and assessments may need to take account of the complex nature of the process and of its components, rather than simply targeting biases associated with one characteristic in isolation.

It should be noted that findings and conclusions in this paper do not serve as any condemnation of teachers – as a profession or as individuals – as enacting the process of stereotyping to any unusual (or to any deliberate) degree. As outlined in the introduction, stereotyping is conceived to be a universal, non-conscious, automatic cognitive function which enables speed and efficiency in thought and behaviour. According to theory, all individuals have a propensity to enact this function to some degree: there is no reason that teachers should be exempt, nor unusually prone. Bias in judgements of pupils is just one manifestation of the human tendency to stereotype.

Where might stereotypes of pupils originate?

Analyses here cannot indicate what may be creating and forming the stereotypes that seem to provide a template for biased teacher perceptions, and there are a number of possible explanations. Firstly, it is feasible that the expectations of different groups of children that are made pertinent to teachers through explicit characteristic-based regulation of pupil, teacher and school performance levels (Bradbury, 2011b) might reify and reinforce differentiated notions of potential and ability which become embedded and self-fulfilling.

Secondly, the messages conveyed by the various policy initiatives which require schools and teachers to focus on selected pupil groups might perpetuate an assumption that these groups are fundamentally lacking. For example, the current concentration on low-income families through the pupil premium may inadvertently imply and contribute to a stereotype that poorer pupils across the board are deficient in ability and potential. Similarly, recent initiatives targeting certain ethnic groups (Maylor *et al*., 2009; Tikley *et al.*, 2006) might build a sense that these groups are essentially less capable, and feed into differentiated expectations.

Thirdly, as suggested by Burgess and Greaves (2009), direct personal experience might inform the process of stereotyping. Teachers may form generalised templates through their everyday experiences and interactions with pupils, and if a proportion of children from a given group are observed to perform in a certain way, a teacher may form a stereotype and overgeneralise to all children in this group.

Lastly, of course, teachers function not only within schools and the education system but also within wider society. Media and other discourses regarding the societal positioning and features of different social groups may create stereotypes of these groups, potentially seeping into and influencing teachers' perceptions of the children in their classroom.

Unfortunately, the data used in this paper do not offer the possibility of testing the extent to which any or all of these potential mechanisms play a part in developing the stereotypes which appear to be held by teachers, and the interrelationships between teachers and the systems and structures within which they function cannot be established here. There may conceivably be a number of points and means of intervention through which stereotyping of pupils could be mitigated, and findings from this paper initially support one in particular: addressing and confronting the process at the teacher-level.

Tackling stereotyping

It has long been argued that self-awareness of perceptions and expectations, and self-reflectiveness, are crucial to effective teaching:

> …for teachers to optimise learning they need to have a greater awareness of the complexities of individual differences [and] the importance of perceptions and expectations of pupils on learning outcomes…(Hallam and Ireson, 1999).

Earp (2010) reviews the cognitive-psychological literature on stereotype activation and consequential behaviours and also argues (here, in relation to stereotyping according to ethnicity) for mindfulness:

> A teacher who is unaware of the basis for her judgments may conclude that they stem from the realities of her student's performance, rather than (directly or indirectly) from the activation of stereotypes about that student's [ethnic] group.

Discussing the research on ways in which teachers may thwart the stereotyping process, Earp suggests that, ‘Teachers are just the sort of people who are in a position to automate egalitarian motives,’ and describes how recent cross-disciplinary studies have indicated that it is feasible that teachers may, with time and effort, ‘train’ and tame the stereotyping mechanism. Potentially this might involve actively learning to draw on alternative stereotypes of pupils, to presume motivation and ability in each student, and / or consciously to be balanced and constructive in feedback to and interactions with all pupils. Earp concludes that, ‘it is essential that schools of education include in their curricula state-of-the-science resources on the unconscious nature of prejudice and the corresponding implications for [the] classroom.’

Though it provides the beginnings of suggestions for change, this existing literature is limited regarding the exact means by which teachers, managers and policy-makers may effectively intervene to alleviate the stereotyping process. The current paper suggests, however, that this is an area very much worthy of further investigation and trial. Increased credibility and importance should be given to the accumulating evidence that biased judgements and stereotyping might be impacting upon and shaping pupil experiences and attainment, and national resources and efforts concentrated upon addressing this possibility. Extending the current study to further explore, unpick and test the drivers of the patterns it has found should play a part in this. Analysis in this paper uses just a sample of children (albeit a relatively large one), so tendencies found particularly in the results regarding the characteristics appearing to underpin stereotypes should be explored further, in enhanced and alternative datasets. The data used in this paper is moreover extremely limited in the extent to which it can examine any role of differential school-level tendencies in creating or mitigating the biases suggested; this should also be an area for further research.

At the policy-level, consideration should be given and investigation instigated into the ways in which initiatives and communications might create or reinforce stereotypes and result in unintended consequences. If, as speculated, characteristics-based targeting and monitoring inadvertently perpetuate attainment differentials based upon these characteristics, this would be a point for intervention and reformulation. Similarly, if ostentatious implementation of targeted policies such as the pupil premium proves detrimental to the treatment of its recipients, this again suggests reconsideration and revision of methods. Finally, the recent encouragement of work-based initial teacher training (in contrast to the university-based model) may be considered in light of the findings in this paper. If a trainee learns predominantly from the practices and norms in their placement school, with less time devoted to critical pedagogical theory, might this serve only to reinforce active stereotypes and expectations, with less scope for new ideas and the challenging of normative templates?

Conclusion

This paper finds that that historical efforts to ensure parity, equality and meritocracy in the education system have not yet resulted in a parity of perception, judgement and assessment. Resources might usefully be directed as suggested here: towards building the evidence base on stereotyping; towards developing relevant interventions and strategies within teacher training and professional development; and towards avoiding the reinforcement of stereotypes and the worsening of their effects during policy intervention and associated publicity. . By recognising and challenging the existence and effects of stereotyping in these ways, it is possible that some of the longstanding and widespread inequalities among primary school children may come to be alleviated.

Notes

[i] https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/219208/main_20text_20_20sfr21-2012.pdf (p 4).
[ii] http://ofstednews.ofsted.gov.uk/article/346
[iii] See http://www.esds.ac.uk/doc/6848/mrdoc/pdf/mcs4_teacher_england.pdf for full survey documentation.

References

Bew, P. (2011a), 'Review of Key Stage 2 testing, assessment and accountability: Progress Report.' London: Department for Education, downloadable from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/180401/DFE-00035-2011.pdf (accessed 13.10.10)

Bew, P. (2011b), 'Review of Key Stage 2 testing, assessment and accountability: Final Report.' London: Department for Education, downloadable from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/176180/Review-KS2-Testing_final-report.pdf (accessed 13.10.10)

Bradbury, A. (2011a), 'Equity, ethnicity and the hidden dangers of 'contextual' measures of school performance.' *Race Ethnicity and Education*, 14:3, 277-291.

Bradbury, A. (2011b), 'Rethinking assessment and inequality: The production of disparities in attainment in early years education.' *Journal of Education Policy*, 26:5, 655-676.

Brookhart, S. M. (2013), 'The use of teacher judgement for summative assessment in the USA.' *Assessment in Education: Principles, Policy & Practice*, 20:1, 69-90.

Accepted Manuscript version of: Campbell, T. (2015) *Stereotyped at Seven? Biases in Teacher Judgement of Pupils' Ability and Attainment.* Journal of Social Policy, 44:3, pp 517-547.
http://dx.doi.org/10.1017/S0047279415000227

Burgess, S. and Greaves, E. (2009), 'Test Scores, Subjective Assessment and Stereotyping of Ethnic Minorities.' University of Bristol: Centre for Market and Public Organisation, downloadable from http://www.bris.ac.uk/cmpo/publications/papers/2009/wp221.pdf (accessed 13.10.13).

Campbell, T. (2013), 'In-school ability grouping and the month of birth effect: Preliminary evidence from the Millennium Cohort Study.' London: Centre for Longitudinal Studies, downloadable from http://www.cls.ioe.ac.uk/shared/getfile.ashx?itemtype=document&id=1618 (accessed 13.10.13).

Department for Children, Schools and Families (2008), '21st Century Schools: A World Class Education for Every Child,' downloadable from https://www.education.gov.uk/publications/eOrderingDownload/DCSF-01044-2008.pdf (accessed 13.10.13).

Department for Education (2010a), 'The Importance of Teaching: The Schools White Paper 2010,' downloadable from https://www.education.gov.uk/publications/eOrderingDownload/CM-7980.pdf (accessed 13.10.13).

Department for Education (2010b), 'The Importance of Teaching: White Paper Equalities Impact Assessment,' downloadable from https://www.education.gov.uk/publications/eOrderingDownload/CM-7980-Impact_equalities.pdf (accessed 13.10.13).

Department for Education (2011), 'National Curriculum Assessments at Key Stage 2 in England, 2010/2011 (revised),' downloadable from http://www.education.gov.uk/rsgateway/DB/SFR/s001047/sfr31-2011.pdf (accessed 13.10.13).

Department for Education (2012a), 'Early Years Foundation Stage Profile Attainment by Pupil Characteristics, England 2011/12,' downloadable from http://www.education.gov.uk/rsgateway/DB/SFR/s001098/sfr30-2012.pdf (accessed 13.10.13).

Department for Education (2012b), 'National Curriculum Assessments at Key Stage 2 in England, 2011/2012 (revised),' downloadable from http://www.education.gov.uk/rsgateway/DB/SFR/s001104/sfr33-2012v2.pdf (accessed 13.10.13).

Department for Education (2012c), 'Phonics Screening Check and National Curriculum Assessments at Key Stage 1 in England: 2012,' downloadable from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/219208/main_20text_20_20sfr21-2012.pdf (accessed 13.10.13).

Department for Education and Skills (2005), 'Higher Standards, Better Schools for all: More Choice for Parents and Pupils,' downloadable from https://www.education.gov.uk/publications/eOrderingDownload/Cm%206677.pdf.pdf (accessed 13.10.13)

Earp, B. D. (2010), ‘Automaticity in the classroom: Unconscious mental processes and the racial achievement gap.’ *Journal of Multiculturalism in Education*, 6:1, 1-22.

Eckert, T. L., Dunn, E. K., Codding, R. S., Begeny, J. C., Kleinmann, A. E. (2006), ‘Assessment of mathematics and reading performance: An examination of the correspondence between direct assessment of student performance and teacher report,’ *Psychology in the Schools*, 43:3, 247-265.

Hallam, S. and Ireson, J (1999), *Pedagogy in the Secondary School.* In Mortimore (Ed) *Understanding Pedagogy*. London: Chapman.

Hansen, K. (Ed.) (2012), ‘Millennium Cohort Study: First, Second, Third and Fourth Surveys. A Guide to the Datasets (Seventh Edition).’ London: Centre for Longitudinal Studies, downloadable from http://www.cls.ioe.ac.uk/shared/get file.ashx?id=598&itemtype=document (accessed 13.10.13)

Hansen, K. and Jones, E. (2011), ‘Ethnicity and gender gaps in early childhood.’ *British Educational Research Journal*, 37:6, 973-991.

Harlen, W. (2004), ‘A systematic review of the evidence of the impact on students, teachers and the curriculum of the process of using assessment by teachers for summative purposes.’ London: EPPI-Centre, downloadable from https://eppi.ioe.ac.uk/cms/LinkClick.aspx?fileticket=Pbyl1CdsDJU%3D&tabid=108&mid=1003 (accessed 13.10.13)

Harlen, W. (2005), ‘Trusting teachers’ judgement: research evidence of the reliability and validity of teachers’ assessment used for summative purposes,’ *Research Papers in Education*, 20:3, 245-270.

Harlen, W. (2007), ‘The quality of learning: assessment alternatives for primary education. Primary Review Research Survey 3/4.’ Cambridge: University of Cambridge Faculty of Education.

Hilton, J. L. and von Hipple. W. (1996), ‘Stereotypes.’ *Annual Review of Psychology,* 47, 237–71

McGarty, C., Yzerbyt, V. Y. and Spears, R. (2002), ‘Stereotypes as Explanations: The Formation of Meaningful Beliefs about Social Groups.’ Cambridge: Cambridge University Press.

Mostapha, T. (2013), ‘Technical report on response in the teacher survey in MCS 4 (age 7).’ London: Centre for Longitudinal Studies, downloadable from http://www.cls.ioe.ac.uk/shared/get-file.ashx?id=1749&itemtype=document (accessed 01.12.13)

Ofqual (2012), 'GCSE English 2012,' downloadable from http://www.ofqual.gov.uk/files/2012-11-02-gcse-english-final-report-and-appendices.pdf (accessed 13.10.13).

Plewis, I. (2007) (Ed.). 'The Millennium Cohort Study: Technical Report on Sampling: Fourth Edition,' London: Centre for Longitudinal Studies, downloadable from http://www.cls.ioe.ac.uk/shared/get-file.ashx?id=409&itemtype=document (accessed 13.10.13).

Reeves, D. J., Boyle, W. F., and Christie, T. (2001), 'The Relationship between Teacher Assessments and Pupil Attainments in Standard Test Tasks at Key Stage 2, 1996-98,' *British Educational Research Journal* 27:2, 141-160.

Robinson, J. P. and Lubienski, S. T. (2011), 'The Development of Gender Achievement Gaps in Mathematics and Reading During Elementary and Middle School: Examining Direct Cognitive Assessments and Teacher Ratings,' *American Educational Research Journal,* 48:2, 268–302.

Strand, S., de Coulon, A., Meschi, E., Vorhouse, J., Frumkin, L., Ivins, C., Small, L. Sood, A., Gervais, M. C., Rehman, H. (2010), 'Drivers and Challenges in Raising the Achievement of Pupils from Bangladeshi, Somali and Turkish Backgrounds.' London: Department for Children, Schools and Families, downloadable from https://www.education.gov.uk/publications/eOrderingDownload/DCSF-RR226.pdf (accessed 13.10.13).

Tikly, L., Haynes, J., Caballero, C., Hill, J., Gillborn, D. (2008), 'Evaluation of Aiming High: African Caribbean Achievement Project,' London: Department for Education and Skills, downloadable at http://webarchive.nationalarchives.gov.uk/20130401151715/https://www.education.gov.uk/publications/eorderingdownload/rr801.pdf (accessed 13.10.13).

Thomas, S., Smees, R., Madaus, G. F., and Raczek, A. E. (1998), 'Comparing Teacher Assessment and Standard Task Results in England: the relationship between pupil characteristics and attainment.' *Assessment in Education: Principles, Policy & Practice*, 5:2, 213-246.

University of London. Institute of Education. Centre for Longitudinal Studies, Millennium Cohort Study: Fourth Survey, 2008 [computer file]. 3rd Edition. Colchester, Essex: UK Data Archive [distributor], August 2012. SN: 6411, http://dx.doi.org/10.5255/UKDA-SN-6411-2.

University of London. Institute of Education. Centre for Longitudinal Studies, Millennium Cohort Study: Fourth Survey, Teacher Survey, 2008 [computer file]. Colchester, Essex: UK Data Archive [distributor], August 2011. SN: 6848 http://dx.doi.org/10.5255/UKDA-SN-6848-1.

Wyse, D., McCreery, E., Torrance, H. (2008), 'The trajectory and impact of national reform: curriculum and assessment in English primary education. Primary Review Research Survey 3/2.' Cambridge: University of Cambridge Faculty of Education.

Annex 1

****

Table 11 about here

****

Annex 2

****

Table 12 about here

****

Table 1: Percentage of pupils with each characteristic judged at each level by their teacher*

	Percentage judged ‘above average’ at reading	Percentage judged ‘below average’ at reading	Percentage judged ‘above average’ at maths	Percentage judged ‘below average’ at maths
Whole sample (n = 4997 / 4985)	45.3	22.2	39.8	20.5
Above 60% median income (n = 3593 / 3585)	52.3	16.6	45.6	16.1
Below 60% median income (n = 1404 / 1400)	26.6	37.3	24.2	32.1
Boys (n = 2494 / 2491)	40.5	27.1	42.4	21.4
Girls (n = 2503 / 2494)	50.1	17.4	37.1	19.5
No SEN diagnosis (n = 3879 / 3864)	55.7	9.3	48.5	9.2
Any SEN diagnosis (n = 1118 / 1121)	11.1	64.7	11.2	57.1
White (n = 4047 / 4032)	46.2	21.7	40.6	19.8
Indian (n = 150 / 150)	46.9	18.1	46.1	14.6
Pakistani (n = 274 / 274)	30.4	29.4	23.8	30.9
Bangladeshi (n = 85 / 86)	38.5	28.3	36.7	24.2
Black Caribbean (n = 68 / 68)	28.6	37.0	20.7	36.7
Black African (n = 112 / 112)	42.8	26.0	25.0	23.4
Speaks English only (n = 4317 / 4305)	46.0	21.9	40.5	20.1
Speaks additional languages (n = 680 / 680)	38.0	25.3	31.8	23.6

*All estimates weighted for survey design and for attrition to the main wave four survey. Ns are unweighted.

Table 2: Mean scores by characteristic on Word Reading and Progress in Maths tests*

	Mean Word Reading score (range: 10-214)	Mean Progress in Maths score (range: 0-28)
Whole sample with teacher reading / maths judgement (n = 4997 / 4985)	108.54	18.41
Above 60% median income (n = 3593 / 3585)	112.48	19.17
Below 60% median income (n = 1404 / 1400)	98.06	16.40
Boys (n = 2494 / 2491)	105.85	18.43
Girls (n = 2503 / 2494)	111.24	18.40
No SEN diagnosis (n = 3879 / 3864)	116.49	19.65
Any SEN diagnosis (n = 1118 / 1121)	82.50	14.39
White (n = 4047 / 4032)	108.00	18.61
Indian (n = 150 / 150)	117.05	19.61
Pakistani (n = 274 / 274)	108.93	15.32
Bangladeshi (n = 85 / 86)	114.95	15.68
Black Caribbean (n = 68 / 68)	101.43	16.77
Black African (n = 112 / 112)	117.74	16.81
Speaks English only (n = 4317 / 4305)	108.17	18.58
Speaks additional languages (n = 680 / 680)	112.28	16.75

*All estimates weighted for survey design and for attrition to the main wave four survey. Ns are unweighted.

Accepted Manuscript version of: Campbell, T. (2015) *Stereotyped at Seven? Biases in Teacher Judgement of Pupils' Ability and Attainment.* Journal of Social Policy, 44:3, pp 517-547.
http://dx.doi.org/10.1017/S0047279415000227

Table 3: Variables used in and structure of linear probability models

Model	Outcome	Predictors	
1	Teacher judgement of reading *above* average / not	BAS Word Reading test ability score	+ above / below 60% income
2			+ boy / girl
3			+ SEN / not
4			+ White / Indian / Pakistani / Bangladeshi / Black Caribbean / Black African
5			+ English only / additional languages
6	Teacher judgement of reading *below* average / not	BAS Word Reading test ability score	+ above / below 60% income
7			+ boy / girl
8			+ SEN / not
9			+ White / Indian / Pakistani / Bangladeshi / Black Caribbean / Black African
10			+ English only / additional languages
11	Teacher judgement of maths *above* average / not	Progress in Maths score	+ above / below 60% income
12			+ boy / girl
13			+ SEN / not
14			+ White / Indian / Pakistani / Bangladeshi / Black Caribbean / Black African
15			+ English only / additional languages
16	Teacher judgement of maths *below* average / not	Progress in Maths score	+ above / below 60% income
17			+ boy / girl
18			+ SEN / not
19			+ White / Indian / Pakistani / Bangladeshi / Black Caribbean / Black African
20			+ English only / additional languages

Table 4: Difference in percentage point likelihood of pupils with each respective characteristic being judged ‘above average’ at reading by their teacher, compared to pupils with the reference characteristic, and controlling for reading cognitive test score^

Income model	
Low income (ref = higher income)	-.110 (.014)***
Word Reading score	.010 (.000)***
Intercept	-.629 (.019)***
Gender model	
Boy (ref = girl)	-.040 (.013)**
Word Reading score	.011 (.000)***
Intercept	-.750 (.026)***
SEN model	
SEN (ref = no SEN)	-.113 (.017)***
Word Reading score	.010 (.000)***
Intercept	-.586 (.026)***
Ethnicity model	
Indian (ref = White)	-.090 (.046)*
Pakistani (ref = White)	-.168 (.027)***
Bangladeshi (ref = White)	-.151 (.058)**
Black Caribbean (ref= White)	-.107 (.039)**
Black African (ref = White)	-.138 (.056)**
Word Reading score	.011 (.000)***
Intercept	-.688 (.018)***
Language model	
Other languages (ref = English only)	-.124 (.021)***
Word Reading score	.011 (.000)***
Intercept	-.691 (.018)***

N for each model = 4997 (unweighted). *** = p < .001; ** = p < .05; * = p < .10. Standard errors in brackets. ^All estimates weighted for survey design and for attrition to the main wave four survey.

Table 5: Difference in percentage point likelihood of pupils with each respective characteristic being judged 'below average' at reading by their teacher, compared to pupils with the reference characteristic, and controlling for reading cognitive test score^

Income model	
Low income (ref = higher income)	.083 (.012)***
Word Reading score	-.009 (.000)***
Intercept	1.135 (.025)***
Gender model	
Boy (ref = girl)	.050 (.009)***
Word Reading score	-.009 (.000)***
Intercept	1.253 (.027)***
SEN model	
SEN (ref = no SEN)	.329 (.017)***
Word Reading score	-.007 (.000)***
Intercept	.864 (.029)***
Ethnicity model	
Indian (ref = White)	.045 (.040)
Pakistani (ref = White)	.086 (.029)**
Bangladeshi (ref = White)	.128 (.041)**
Black Caribbean (ref= White)	.094 (.039)**
Black African (ref = White)	.130 (.028)***
Word Reading score	-.009 (.000)***
Intercept	1.181 (.024)***
Language model	
Other languages (ref = English only)	.070 (.016)***
Word Reading score	-.009 (.000)***
Intercept	1.184 (.024)***

N for each model = 4997 (unweighted). *** = p < .001; ** = p < .05; * = p < .10. Standard errors in brackets. ^All estimates weighted for survey design and for attrition to the main wave four survey.

Table 6: Difference in percentage point likelihood of pupils with each respective characteristic being judged ‘above average’ at maths by their teacher, compared to pupils with the reference characteristic, and controlling for maths cognitive test score^

Income model	
Low income (ref = higher income)	-.104 (.016)***
Maths score	.040 (.001)***
Intercept	-.311 (.018)***
Gender model	
Boy (ref = girl)	.052 (.012)***
Maths score	.042 (.001)***
Intercept	-.396 (.018)***
SEN model	
SEN (ref = no SEN)	-.181 (.019)***
Maths score	.037 (.001)***
Intercept	-.234 (.025)***
Ethnicity model	
Indian (ref = White)	.013 (.039)
Pakistani (ref = White)	-.031 (.028)
Bangladeshi (ref = White)	.083 (.043)*
Black Caribbean (ref= White)	-.123 (.035)**
Black African (ref = White)	-.081 (.053)
Maths score	.042 (.001)***
Intercept	-.367 (.017)***
Language model	
Other languages (ref = English only)	-.011 (.020)
Maths score	.042 (.001)***
Intercept	-.369 (.017)***

N for each model = 4985 (unweighted). *** = p < .001; ** = p < .05; * = p < .10. Standard errors in brackets. ^All estimates weighted for survey design and for attrition to the main wave four survey.

Table 7: Difference in percentage point likelihood of pupils with each respective characteristic being judged ‘below average’ at maths by their teacher, compared to pupils with the reference characteristic, and controlling for maths cognitive test score^

Income model	
Low income (ref = higher income)	.070 (.014)***
Maths score	-.032 (.001)***
Intercept	.782 (.026)***
Gender model	
Boy (ref = girl)	.021 (.012)*
Maths score	-.034 (.001)***
Intercept	.812 (.026)***
SEN model	
SEN (ref = no SEN)	.356 (.020)***
Maths score	-.024 (.001)***
Intercept	.554 (.026)***
Ethnicity model	
Indian (ref = White)	-.019 (.031)
Pakistani (ref = White)	.001 (.036)
Bangladeshi (ref = White)	-.054 (.048)
Black Caribbean (ref= White)	.107 (.054)**
Black African (ref = White)	-.024 (.062)
Maths score	-.034 (.001)***
Intercept	.824 (.026)***
Language model	
Other languages (ref = English only)	-.027 (.021)
Maths score	-.034 (.001)***
Intercept	.827 (.026)***

N for each model = 4985 (unweighted). *** = p < .001; ** = p < .05; * = p < .10. Standard errors in brackets. ^All estimates weighted for survey design and for attrition to the main wave four survey.

Table 8: Difference in percentage point likelihood of pupils with each respective characteristic being judged ‘above average’ at reading by their teacher, controlling for each other factor and reading cognitive test score^

	All (n = 4997)	Boys (n = 2494)	Girls (n = 2503)
Low income (ref = higher income)	-.085 (.015)***	-.054 (.018)**	-.114 (.022)***
Boy (ref = girl)	-.033 (.013)**		
SEN (ref = no SEN)	-.100 (.017)***	-.105 (.021)***	-.103 (.021)***
Indian (ref = White)	-.052 (.047)	.033 (.072)	-.152 (.071)**
Pakistani (ref = White)	-.089 (.036)**	.035 (.053)	-.187 (.060)**
Bangladeshi (ref = White)	-.072 (.060)	-.084 (.091)	-.060 (.072)
Black Caribbean (ref= White)	-.057 (.040)	.024 (.047)	-.159 (.065)**
Black African (ref = White)	-.074 (.057)	-.020 (.077)	-.130 (.076)*
Other languages (ref = English only)	-.039 (.030)	-.092 (.042)**	.007 (.048)
Word Reading score	.010 (.000)***	.009 (.000)***	.011 (.000)***
Intercept	-.517 (.025)***	-.494 (.034)***	-.605 (.037)***

*** = p < .001; ** = p < .05; * = p < .10. Standard errors in brackets. ^All estimates weighted for survey design and for attrition to the main wave four survey. Ns are unweighted.

Table 9: Difference in percentage point likelihood of pupils with each respective characteristic being judged ‘above average’ at maths by their teacher, controlling for each other factor and maths cognitive test score^

	All (n = 4985)	**Boys (n = 2491)**	**Girls (n = 2493)**
Low income (ref = higher income)	-.091 (.017)***	-.086 (.022)***	-.100 (.024)***
Boy (ref = girl)	.074 (.013)***		
SEN (ref = no SEN)	-.185 (.020)***	-.232 (.025)***	-.125 (.025)***
Indian (ref = White)	.014 (.045)	.003 (.073)	.023 (.059)
Pakistani (ref = White)	.008 (.039)	.030 (.063)	-.002 (.054)
Bangladeshi (ref = White)	.108 (.048)**	.181 (.076)**	.043 (.085)
Black Caribbean (ref= White)	-.068 (.040)*	-.007 (.048)	-.132 (.055)**
Black African (ref = White)	-.063 (.051)	-.010 (.071)	-.113 (.077)
Other languages (ref = English only)	-.007 (.029)	-.043 (.046)	.022 (.046)
Maths score	.035 (.001)***	.034 (.002)***	.035 (.002)***
Intercept	-.217 (.025)***	-.118 (.038)**	-.238 (.035)***

*** = p < .001; ** = p < .05; * = p < .10. Standard errors in brackets. ^All estimates weighted for survey design and for attrition to the main wave four survey. Ns are unweighted.

Table 10: Difference in likelihood of pupils with each respective characteristic being judged 'above average' at reading by their teacher, compared to pupils with the reference characteristic, controlling for reading cognitive test score: robustness checks^

	Original results (*B*)	**Check 1: Logistic model (*Difference and p value for difference in model-predicted probabilities*^^)**	**Check 2: Age and timing controls (*B*)^^^**	**Check 3: One pupil per teacher (*B*)**	**Check 4: Unweighted; clustered by school (*B*)^^^^**
Low income (ref = higher)	-.11 (.014)***	-.13 (.015)***	-.11 (.015)***	-.10 (.018)***	-.11 (.013)***
Boy (ref = girl)	-.04 (.013)**	-.05 (.013)***	-.04 (.013)**	-.02 (.015)	-.04 (.011)***
SEN (ref = no SEN)	-.12 (.015)***	-.18 (.019)***	-.11 (.017)***	-.12 (.021)***	-.12 (.014)***
Indian (ref = White)	-.09 (.046)*	-.07 (.040)*	-.11 (.046)**	-.07 (.034)*	-.08 (.031)**
Pakistani (ref = White)	-.17 (.027)***	-.16 (.026)***	-.18 (.027)**	-.15 (.034)***	-.18 (.023)***
Bangladeshi (ref = White)	-.15 (.058)**	-.17 (.066)**	-.17 (.057)**	-.18 (.071)**	-.15 (.051)**
Black Caribbean (ref = White)	-.11 (.039)**	-.12 (.042)**	-.12 (.043)**	-.05 (.047)	-.10 (.041)**
Black African (ref = White)	-.14 (.056)**	-.13 (.052)**	-.13 (.059)**	-.17 (.063)**	-.19 (.039)***
Other languages (ref = English only)	-.12 (.021)***	-.12 (.020)***	-.13 (.020)***	-.11 (.021)***	-.13 (.017)***
All n.s =	4997	4997	4641	2995	4997

*** = p < .001; ** = p < .05; * = p < .10. Standard errors in brackets.

^All estimates bar Check 4 weighted for survey design and for attrition to the main wave four survey. Ns are unweighted.

^^Calculated using "margins, pwcomp (eff)" in Stata 13.

^^^See Annex 2: Table 12 for coefficients of age and timing controls

^^^^Robust SEs estimated using "vce (cluster)" in Stata 13.

Table 11: Key descriptive statistics for respective Millennium Cohort Study English single-cohort baby household samples, for comparison*

	Wave one: whole English sample, design weights only	Wave one: whole English sample, design weights plus non-response weights	Wave four: English sample with teacher survey response, design weights only	Wave four: English sample without teacher survey response, design weights only	Wave four sample used in paper for reading analysis, design weights only	Wave four: English sample with teacher survey response, design weights plus attrition weights	Wave four: English sample without teacher survey response, design weights plus attrition weights	Wave four sample used in paper for reading analysis, design weights plus attrition weights
Percent low income (OECD indicator) – at wave one	28.2%	29.5%	23.6%	25.4%	23.4%	30.3%	33.4%	30.0%
Percent low income (OECD indicator) – at wave four	-	-	21.3%	24.0%	21.8%	27.0%	30.6%	27.4%
Percent girls	48.8%	48.9%	50.1%	48.5%	50.3%	49.6%	47.5%	50.0%
Percent White	85.3%	84.6%	90.9%	85.3%	89.6%	88.0%	80.8%	86.4%
Percent Indian	2.1%	2.1%	1.7%	2.7%	1.6%	2.0%	3.0%	1.8%
Percent Pakistani	3.4%	3.6%	2.6%	4.0%	2.5%	3.5%	5.4%	3.4%
Percent Bangladeshi	1.1%	1.2%	0.7%	1.5%	0.6%	1.0%	2.2%	0.9%
Percent Black Caribbean	1.1%	1.2%	0.8%	1.6%	0.8%	1.2%	2.1%	1.1%
Percent Black African	1.7%	1.8%	1.1%	2.1%	1.1%	1.7%	2.8%	1.7%
Percent speaking English only	88.9%	88.4%	91.8%	87.6%	93.1%	89.8%	84.0%	91.1%

Accepted Manuscript version of: Campbell, T. (2015) *Stereotyped at Seven? Biases in Teacher Judgement of Pupils' Ability and Attainment.* Journal of Social Policy, 44:3, pp 517-547. http://dx.doi.org/10.1017/S0047279415000227

Mean Word Reading Test score	-	-	109.2	108.2	110.2	107.3	105.4	108.5
Mean Progress in Maths test score	**-**	**-**	18.8	18.4	18.8	18.5	17.9	18.4
n =	**11374**	**11374**	**5184**	**3107**	**4997**	**5184**	**3107**	**4997**

*Figures are presented firstly with design weights only, which account simply for known unequal selection probabilities into the initial sample (children in areas with higher number of minority ethnic and low-income families were oversampled so had a higher probability of inclusion). Secondly, adjustments for non-response (at wave one) / attrition (at wave four) are presented – these weight the sample according to differential tendencies to participation according to selected measured characteristics. Ns are unweighted.

Table 12: Difference in likelihood of pupils with each respective characteristic being judged 'above average' at reading by their teacher, compared to pupils with the reference characteristic, controlling for reading cognitive test score - Robustness check 2: Age and timing controls, with full coefficients^

	Original results (*B*)	Check 2: Age and timing controls (*B*)^^^
Low income (ref = higher)	-.11 (.014)***	-.11 (.015)***
Age in months^^		.00 (.002)*
2 month lag (ref = 0-1 month)		.01 (.030)
3 month lag (ref = 0-1 month)		.04 (.033)
4 month lag (ref = 0-1 month)		.02 (.033)
5 month lag (ref = 0-1 month)		-.04 (.036)
6 month lag (ref = 0-1 month)		-.04 (.036)
7 month lag (ref = 0-1 month)		-.05 (.039)
8-20 month lag (ref = 0-1 month)		-.06 (.045)
Boy (ref = girl)	-.04 (.013)**	-.04 (.013)**
Age in months^^		.00 (.002)
2 month lag (ref = 0-1 month)		.00 (.030)
3 month lag (ref = 0-1 month)		.04 (.034)
4 month lag (ref = 0-1 month)		.01 (.033)
5 month lag (ref = 0-1 month)		-.04 (.036)
6 month lag (ref = 0-1 month)		-.05 (.037)
7 month lag (ref = 0-1 month)		-.05 (.038)
8-20 month lag (ref = 0-1 month)		-.06 (.045)
SEN (ref = no SEN)	-.12 (.015)***	-.11 (.017)***
Age in months^^		.00 (.002)
2 month lag (ref = 0-1 month)		.01 (.030)
3 month lag (ref = 0-1 month)		.04 (.033)
4 month lag (ref = 0-1 month)		.01 (.033)
5 month lag (ref = 0-1 month)		-.04 (.035)
6 month lag (ref = 0-1 month)		-.05 (.037)
7 month lag (ref = 0-1 month)		-.05 (.038)

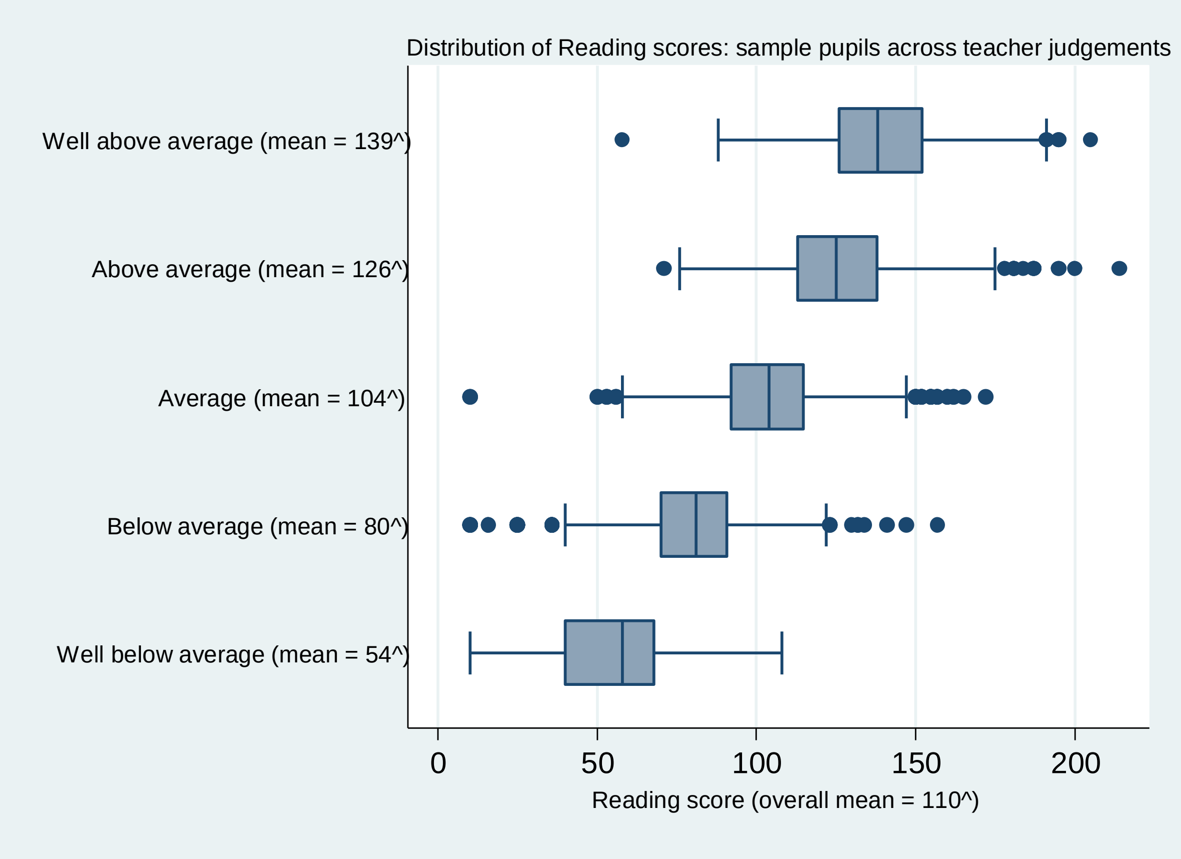
8-20 month lag (ref = 0-1 month)		-.06 (.044)
Indian (ref = White)	-.09 (.046)*	-.11 (.046)**
Pakistani (ref = White)	-.17 (.027)***	-.18 (.027)***
Bangladeshi (ref = White)	-.15 (.058)**	-.17 (.057)**
Black Caribbean (ref = White)	-.11 (.039)**	-.12 (.043)**
Black African (ref = White)	-.14 (.056)**	-.13 (.059)**
Age in months^^		.00 (.002)*
2 month lag (ref = 0-1 month)		.00 (.031)
3 month lag (ref = 0-1 month)		.04 (.033)
4 month lag (ref = 0-1 month)		.02 (.033)
5 month lag (ref = 0-1 month)		-.04 (.036)
6 month lag (ref = 0-1 month)		-.05 (.036)
7 month lag (ref = 0-1 month)		-.05 (.039)
8-20 month lag (ref = 0-1 month)		-.05 (.045)
Other languages (ref = English only)	-.12 (.021)***	-.13 (.020)***
Age in months^^		.00 (.002)
2 month lag (ref = 0-1 month)		.00 (.030)
3 month lag (ref = 0-1 month)		.04 (.033)
4 month lag (ref = 0-1 month)		.02 (.033)
5 month lag (ref = 0-1 month)		-.04 (.036)
6 month lag (ref = 0-1 month)		-.05 (.036)
7 month lag (ref = 0-1 month)		-.05 (.039)
8-20 month lag (ref = 0-1 month)		-.05 (.044)
All n.s =	4997	4641

*** = p < .001; ** = p < .05; * = p < .10. Standard errors in brackets.

^All estimates weighted for survey design and for attrition to the main wave four survey. Ns are unweighted.

^^Range = 76-97

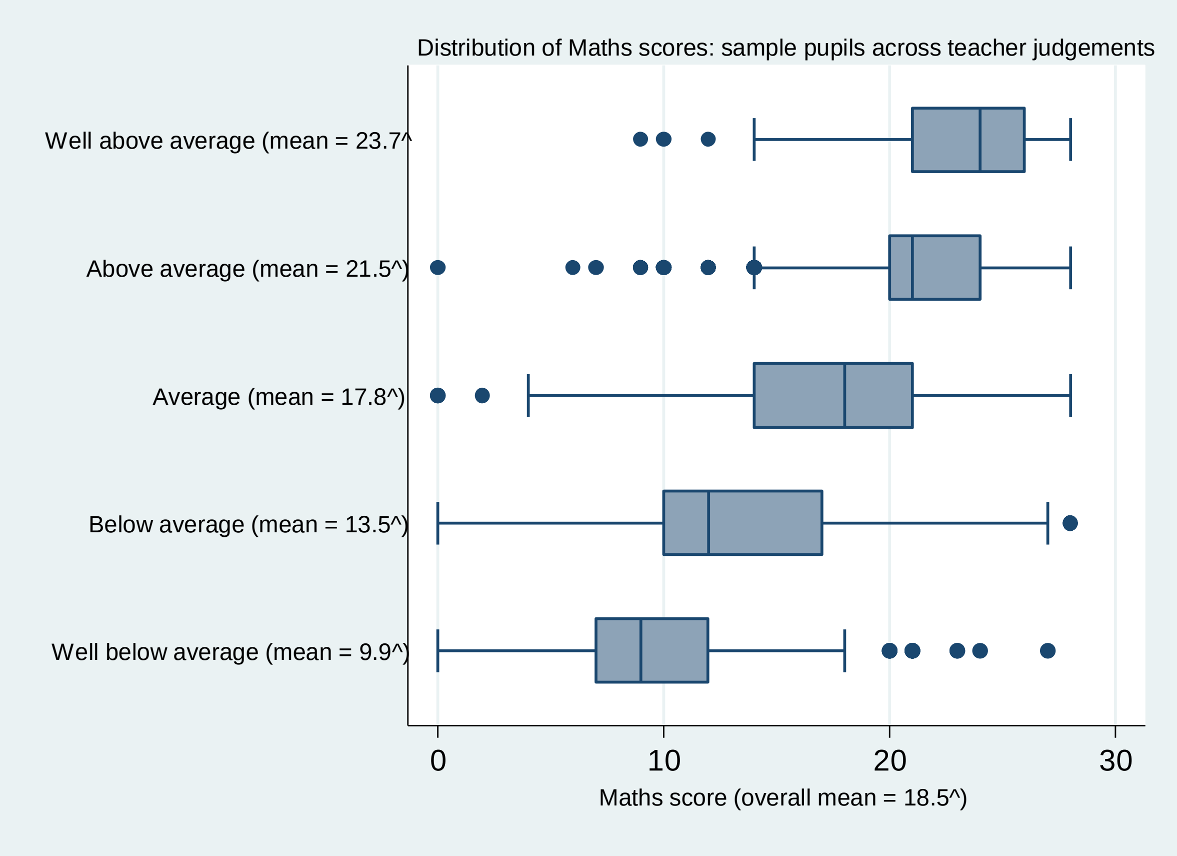
Figure 1: Distribution of and mean BAS Word Reading scores of pupils with each teacher judgement of reading 'ability and attainment'*



N = 4997 (unweighted). ^Means are unweighted; weighted estimates: overall mean = 109; well above average = 139; above average = 126; average = 104; below average = 79; well below average = 54. Line represents median, box represents 25th and 75th percentiles (Q1 and Q3, respectively), whiskers represent Q3+1.5(Q3-Q1) / Q1-1.5*(Q3-Q1).

Accepted Manuscript version of: Campbell, T. (2015) *Stereotyped at Seven? Biases in Teacher Judgement of Pupils' Ability and Attainment.* Journal of Social Policy, 44:3, pp 517-547.
http://dx.doi.org/10.1017/S0047279415000227

Figure 2: Distribution of and mean Progress in Maths scores of pupils with each teacher judgement of maths 'ability and attainment'*



N = 4985 (unweighted). ^Means are unweighted; weighted estimates: overall mean = 18.4; well above average = 23.7; above average = 21.4; average = 17.7; below average = 13.6; well below average = 10.3.
Line represents median, box represents 25th and 75th percentiles (Q1 and Q3, respectively), whiskers represent Q3+1.5(Q3-Q1) / Q1-1.5*(Q3-Q1).

Accepted Manuscript version of: Campbell, T. (2015) *Stereotyped at Seven? Biases in Teacher Judgement of Pupils' Ability and Attainment.* Journal of Social Policy, 44:3, pp 517-547.
http://dx.doi.org/10.1017/S0047279415000227