OXFORD BULLETIN OF ECONOMICS AND STATISTICS, 82, 5 (2020) 0305–9049
doi: 10.1111/obes.12370

Short, Heavy and Underrated? Teacher Assessment Biases by Children's Body Size*

Nicole Black† and Sonja C. de New†,‡,§

†*Centre for Health Economics, Monash Business School, Monash University, Caulfield, Victoria, Australia. (e-mail: Nicole.Black@monash.edu)*
‡*IZA Institute of Labor Economics, Bonn, Germany. (e-mail: Sonja.deNew@monash.edu)*
§*RWI Research Network, Essen, Germany*

Abstract

We compare non-blind teacher assessments with blind national test scores in maths to examine teacher-test score disparities by children's height and weight. Relative to test scores, shorter and heavier children are rated less favourably by teachers. This teacher-test score discrepancy cannot be explained by the child's behaviours, motivation to learn or cognitive ability. Unobserved student fixed effects across subjects explain the teacher-test score discrepancy by height, but not weight. Our analysis points to biased teacher assessments as the most plausible explanation for the remaining teacher-test score gap by weight. We find harsher teacher assessments are associated with a reduction in both the child's future test performance and liking for maths 4 years later.

I. Introduction

Appearance matters in today's society. Differences in labour and marriage market outcomes by physical appearance, typically measured by stature, body mass index (BMI) and attractiveness are well documented (e.g. Hamermesh and Biddle, 1994; Sargent and Blanchflower, 1994; Averett and Korenman, 1996; Harper, 2000; Oreffice and Quintana-Domeque, 2010; Chiappori, Oreffice and Quintana-Domeque, 2012, 2016; Oreffice and Quintana-Domeque, 2016). It is increasingly recognized that the origins of these socioeconomic disparities by physical appearance may manifest in childhood, for example through influences on the formation of cognitive and socioemotional skills (Mobius and Rosenblat, 2006). Persico, Postlewaite and Silverman (2004) show that the height premium for wages can be traced back to being tall in adolescence, and suggest that teen social experiences

JEL Classification numbers: I24; I14; J24; J71.
*This paper uses unit record data from Growing Up in Australia, the Longitudinal Study of Australian Children. The study is conducted in partnership between the Department of Families, Housing, Community Services and Indigenous Affairs (FaHCSIA), the Australian Institute of Family Studies (AIFS) and the Australian Bureau of Statistics (ABS). The findings and views reported in this paper are those of the authors and should not be attributed to FaHCSIA, AIFS or the ABS. This research was supported by an Australian National Preventive Health Agency Fellowship (2AUN2013F).

are a likely channel, while Case and Paxson (2008) and Case, Paxson and Islam (2009) suggest that cognitive achievement during childhood is a likely explanation for the wage height premium. Mocan and Tekin (2010) show that the beauty premium for wages may partly operate through a greater accumulation of cognitive skills during high school.

The important role of early human capital development in explaining later life disparities by physical appearance is supported by recent studies showing a negative effect of being short (Cinnirella, Piopiunik and Winter, 2011; Scholder *et al.*, 2013; Schick and Steckel, 2015) or overweight (Ding *et al.*, 2009; Black, Johnston and Peeters, 2015; Sabia and Rees, 2015) on cognitive skill development and academic achievement during childhood and adolescence. Despite a number of hypotheses, the underlying mechanism through which weight and height (or body size) influence academic achievement remains unclear (Scholder *et al.*, 2013; Black *et al.*, 2015). This paper focuses on one potential mechanism – biased teacher assessments.

We propose that stereotyping and discrimination may lead to biased teacher assessments. It is easy for stereotypes to affect judgements made by all individuals, including teachers. Stereotypes are 'beliefs about the characteristics, attributes, and behaviours of members of certain groups' (Hilton and von Hippel, 1996, p.240). Such beliefs may be held unconsciously, and they may have positive or negative connotations. Stereotyping can lead to discriminatory behaviours among teachers by altering the lens through which they perceive their students' ability (Ferguson, 2003; Jussim and Harber, 2005). Rather than through prejudice, stereotyping can also influence teacher assessments through a process referred to as 'statistical discrimination' (Burgess and Greaves, 2013), whereby assessments are informed by a weighted average of information particular to the individual student and their group.

Certain traits (such as being tall or obese) can elicit expectations from others, including teachers, based on stereotypes. A teacher's perceptions and expectations of a child's ability may affect the child's academic development by influencing the achievement goals that teachers set, the teaching strategies employed, and the energy and effort used in the classroom (Ferguson, 2003). This can affect the opportunities that children have to learn and the grades they are given. At the same time, through what is known as 'self-fulfilling prophecies', a teacher's expectations can impact the actual performance of students by influencing their self-belief and own learning goals, strategies and effort (Ferguson, 2003; Jussim and Harber, 2005).

A large literature in social psychology has shown that physically attractive people are perceived to be more intelligent, socially skilled, warm and sociable (Feingold, 1992). The literature also suggests that height is generally linked with perceptions of positive attributes, such as attractiveness, competency and social status, especially for men (Jackson and Ervin, 1992). However, negative perceptions of heavier individuals are widespread in Western societies; overweight and obesity are often associated with characteristics such as laziness, lacking self-discipline, sloppiness, low intelligence and incompetence for both males and females (Puhl and Heuer, 2009). Such stereotypes have been shown to exist in many settings, including in schools, and appear to be upheld by educators, peers and parents of children (Puhl and Latner, 2007).

Stereotypes may evolve from true differences in productivity and performance. For example, Cipriani and Zago (2011) found that physically attractive economics students

at the University of Verona performed better than less attractive students in both written exams (where the assessor was blind to the student's appearance) and oral exams (where appearance was visible to the assessor). They concluded that payoffs to beauty reflected productivity rather than discrimination.

However, several recent studies point towards differences between objective test scores and teacher assessments by body size. Zavodny (2013) and MacCann and Roberts (2013) found significant achievement gaps by obesity in the United States when teacher assessments were used, but no (or little) difference in achievement when objective test scores were used. Building on this work Kenney *et al.* (2015) modelled the difference in teacher assessments and test scores by BMI in the United States and found that significant discrepancies exist. Furthermore, Queally *et al.* (2017) found that teacher-test score gaps in Ireland also exist by mother's BMI status.

Biased teacher assessments may be one possible explanation for the teacher-test score discrepancy found in previous studies; however, it is not possible to make this conclusion without accounting for alternative explanations. For example, unobserved characteristics of the child (such as self-confidence) may be correlated with body size and may be incorporated by teachers in their assessments, more so than is reflected in objective test scores. Alternatively, a teacher-test score gap might be due to systematic differences by body size in how students perform in externally marked exams or in high-pressure environments relative to in-class performance. To date, such explanations have not been accounted for in empirical studies. Our study aims to address this gap.

The consequences of teacher-test score discrepancies are potentially serious. Previous studies on teacher-test score discrepancies by gender, for example, have identified implications for high school and university course enrolment, occupational choices and earnings in adulthood (Lavy, 2008; Mechtenberg, 2009; Lavy and Sand, 2018). However, little is known about the future consequences of teacher-test score gaps by body size.

This study is the first, to our knowledge, that investigates whether biased teacher assessments are a plausible explanation for teacher-test score discrepancies by body size, and that explores the potential future academic consequences of such biased assessments. We use data on primary school children from the nationally representative Longitudinal Study of Australian Children (LSAC) and employ a quasi-experimental identification strategy which compares teacher assessments in maths (non-blind with respect to body size) with externally marked standardized tests in maths (blind with respect to body size). This allows us to control for unobserved factors that are fixed across assessment methods, in particular students' academic ability. The same identification strategy has been used to investigate biased grading of students by gender (Lindahl, 2007; Lavy, 2008; Hinnerich, Hoeglin and Johannesson, 2011; Cornwell *et al.*, 2013; Di Liberto and Casula, 2016; Terrier, 2016; Lavy and Sand, 2018) and ethnicity (Burgess and Greaves, 2013; Botelho, Madeira and Rangel, 2015; Campbell, 2015).

We investigate teacher-test score discrepancies using two different teacher scores which differ in their level of objectivity, allowing us to compare biases under different types of assessment methods. One teacher assessment measures the overall maths skills of a child using a single question, while the other teacher score is calculated from a multi-item survey of the student's proficiency in specific maths skills. We find that significant teacher-test score discrepancies exist by height, weight, BMI and obesity status for both types of

teacher assessments. We show that these discrepancies are not due to observed measures of socioeconomic background, various measures of cognitive ability, student's motivation or behaviours. Using a within-individual fixed effects (FE) model across subjects that controls for unobserved fixed traits of the student and teacher, we find that unobserved fixed traits explain a large proportion of the favourable teacher assessment for taller children. However, the harsher teacher assessments by weight, BMI and obesity status persist. This finding is consistent with a bounds analysis, which shows that the teacher-test score discrepancy by BMI is highly robust to selection on unobserved characteristics (Oster, 2019). We reason that biased teacher assessments are the most likely explanation for the lower teacher assessments given to heavier children.

We further demonstrate that harsher teacher assessments relative to test scores (in grade 5) are linked with poorer performance in grade 9 exams and less liking of maths in grade 9, and that this holds even after accounting for unobserved fixed traits across subjects and subject-specific characteristics of the child. These findings suggest that the consequences of biased teacher assessments are likely to be harmful to the human capital development of children.

The remainder of this paper is structured as follows: section II outlines the data and study design; section III details the empirical approach used to measure assessment bias; section IV presents our main results; section V examines the impact of teacher-test score gaps on human capital development; and in section VI, we discuss our findings.

II. Study design

The data

This study utilizes data from LSAC, a biennial representative panel survey of Australian children (see Soloff, Lawrence and Johnstone, 2005 for a detailed description of the study design). The main data is collected through face-to-face interviews with the child's primary parent at their home. Additional information is collected through interviews with the child and self-completion questionnaires from parents and the child's teacher. LSAC contains teacher assessments on the child's academic skills, which are linked to administrative records of the child's national test scores (details of these assessments are provided below). Additionally, the survey data provides rich information on the child's cognitive ability, classroom behaviours and participation in class to allow further investigation into the likely mechanisms of teacher-test score discrepancies.

We focus on Cohort K data from 2010 (wave 4) when the children were 10–11 years old. This year is chosen because it is the only year that has complete information on the national test scores and teacher assessments as well as information on the child's effort and participation in class. Of the Wave 1 original sample, the response rate for wave 4 was 84%, which is similar to other longitudinal studies of children (such as the Millennium Cohort Study in the UK). The eligible sample for our study is 2,945 children who were in grade 5 at the time of the 2010 survey. This ensures that teacher assessments and national test scores (which are administered every 2 years) occurred in the same year. Of the eligible sample, we lose 622 children due to missing information on teacher assessments, and a further 184 children due to missing information on national test scores. The primary estimation

sample consists of 1,930 participants who have non-missing information on all variables, including covariates.

We explored whether children in our estimation sample differ systematically from all eligible children in our data by estimating a probit model of the likelihood of being in the estimation sample or having a completed teacher assessment, conditioning on the child's gender, ethnicity, body mass index, family socioeconomic position and school characteristics (see Online Appendix Table A1). Characteristics associated with being in the estimation sample are: speaking English at home, living in a two-parent household and attending a school with a higher maths score. In other words, some socioeconomic indicators are positively associated with being in the estimation sample. However, the likelihood of being in the sample does not appear to be correlated with other measures of socioeconomic status, including neighbourhood SES, mother's education or father's education, nor is it correlated with the child's body size. Therefore, while our sample may not be representative of all children, sample selection does not appear to be substantial. We are able to control for observed characteristics of the child, parent and school that may be associated with selection into our sample, and in fixed-effects analyses, we are able to further control for any unobserved fixed traits of the child, parent, school, class and teacher.

Body size

The two main measures of body size are the child's height and weight, which are measured in 2010 (age 10/11) by trained interviewers using a stadiometer and digital scales. These are standardized (mean $= 0$, SD $= 1$) across the cohort. In supplementary models, we include squared terms for height and weight to allow for nonlinearities. In alternative specifications, instead of height and weight, we use z-scores for body mass index (BMI); or, categories of obese, overweight, normal and underweight, measured according to age- and gender-specific international BMI cut-off points (Cole *et al.*, 2000, 2007). In our sample, 5.2% are obese, 19.2% overweight, 70.0% normal and 5.6% underweight.[1] Online Appendix Table A2 shows the differences in characteristics between children with obesity and those of normal weight. Children with obesity attend schools that perform worse in maths and are also more likely to perform worse themselves in maths than children of normal weight. They are also more likely to be of European descent, less likely to go to an independent school, and more likely to be from families of lower socioeconomic status as measured by parent's education, their involvement in school and the socioeconomic indicator of the region (SEIFA index). Children with obesity perform worse on the problem-solving cognitive test and are more likely to have emotional, peer and conduct problems.

Non-blind teacher assessments and blind test scores

In 2008, Australia introduced the National Assessment Program in Literacy and Numeracy (NAPLAN) to provide nationally comparable measures of student achievement in order

[1] The relative body size of a child compared with their particular classmates may also be a relevant determinant of teacher-test score discrepancies. Unfortunately, we are not able to investigate this with our data as the LSAC sample is not school or class based and therefore we are unable to get information on where the child stands in the respective distribution in their class or school. However, this might be an interesting extension to our analysis for future research.

to identify strengths and weaknesses in teaching programmes and inform policy development. NAPLAN tests are administered to all students in Australian schools in grades 3, 5, 7 and 9 in one week in May each year. NAPLAN is designed to test skills in literacy and numeracy (maths) that are developed through the school curriculum. The tests are marked by a combination of computers and external expert assessors, who are blind to any identifying characteristics of the student (i.e. names and school identifiers are hidden). Therefore, NAPLAN scores provide standardized objective measures of student academic performance. The scores from NAPLAN, which are typically released in September each year, are primarily used to measure the performance of individual students compared to established standards. They are also used to report on the relative performance of schools and used to identify achievement gaps by key student characteristics, including gender, language background (other than English), Indigenous status and parental education level. This information is published online and in public annual reports. Individual results are released to students, parents and teachers.

Our study focuses on NAPLAN scores at grade 5 (age 10–11) to measure academic performance (hereafter referred to as *blind test scores*). Scores are continuous, and range from approximately 300 to 800, but for ease of interpretation, we standardize test scores (mean=0, SD=1).

There is one numeracy score, which we refer to as the maths score. We derive an overall literacy score by taking the average of the four standardized scores from reading, writing, spelling and grammar.

An advantage of LSAC is that it has two different teacher assessment scores (non-blind test scores), both of which were obtained through the LSAC Teacher Survey, completed by the child's grade 5 classroom teacher, but which differ in their level of subjectivity. The first teacher score was derived from the Academic Rating Scale (ARS), which asks teachers to rate on a 5-point scale whether the child is proficient at key skills (from 'Not yet' to 'Proficient'). There are 10 skills for maths, such as 'subtracting numbers', 'reducing fractions to lowest denominator' and 'uses strategies to multiply and divide', and 9 questions for literacy, such as 'understands and interprets a story', and 'reads fluently'. Specific examples of each skill were provided in the survey to aid consistency in skills being assessed (see Online Appendix Figure A1 for details). The ARS score is calculated by taking the average of the responses across all skills and then standardized (mean=0, SD=1).

The second teacher score is the overall assessment. Directly after completing the ARS, the teacher was asked 'overall, how would you rate this child's academic skills, compared to other children of the same grade level?' The teacher rated (i) English language & literacy skills, and (ii) Mathematical skills each on a 5-point scale (1 far below average; 2 below average; 3 average; 4 above average; and, 5 far above average). The most common assessment is 'average' with 45% of students receiving this grade for maths. For comparability, this score is standardized (mean=0, SD=1). Unlike the ARS, no specific skill or example is provided to teachers for the overall assessment; therefore, we hypothesize that teachers are more likely to consider and take into account other characteristics of the child (which may be linked to body size) in the overall assessment than under the ARS assessment.

The overall teacher assessment is very similar to the 5-point rating (typically A–E grades) that Australian students receive in their school report card, and therefore under-

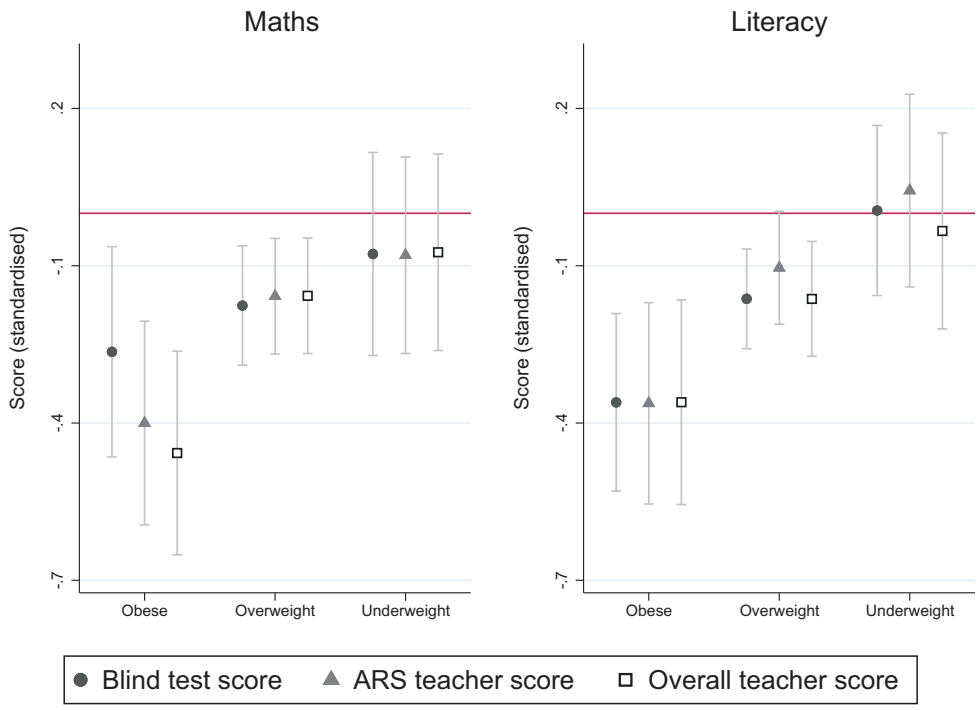

Figure 1. Raw differences in assessments by body size from test and teacher scores
Notes: Assessments undertaken in Grade 5. Scores are all standardized (mean=0, SD=0). Comparison group is 'normal weight' Grey lines represent 95% confidence intervals.

standing potential bias in this form of assessment is particularly meaningful.[2] By comparing teacher-test score discrepancies for two types of teacher assessments, we can inform the debate on which form of teacher assessment is less prone to systematic bias. Both teacher scores were administered by the LSAC survey, and therefore not used to measure the performance of the teacher or the school. This means that there was no incentive to inflate student skills and no reason for the teachers to give anything but honest assessments.

Online Appendix Tables A3–A10 explore in detail the relationship between the different teacher assessments and the blind test score. As expected, all scores are highly and significantly correlated with each other. The highest correlations are observed between the two subjective scores (the ARS score and the overall assessment) for both maths and literacy (correlation coefficients are 0.79 and 0.80 respectively). The correlations of the teacher assessments with the blind test scores are higher for literacy (ARS: 0.71; overall score: 0.75) than for maths (ARS: 0.64; overall score: 0.67). Contingency tables confirm these overall patterns.

Figure 1 shows the raw difference in mean scores for maths and literacy by body size (or weight-for-height status) using the blind test score (black circle), ARS teacher score (grey triangle) and overall teacher score (white square). The comparison group in each figure is

[2]The Australian Education Act 2013 requires schools to provide student reports at least twice a year. The reports must assess the student's progress against national standards, relative to the performance of the student's peer group and be reported on a 5-point scale (Australian Government, 2013).

children of normal weight. It shows that for both maths and literacy, there are achievement gaps by body size – children who are overweight and even more so, obese, tend to perform worse than children of normal weight. Children who are underweight perform similarly to those with a normal weight. For literacy, there is a high degree of consistency in the achievement gap associated with body weight across teacher and test scores. However, the maths achievement gap for children with obesity tends to be much larger under the two teacher scores than under the blind test score. Taken at face value, these figures could indicate a systematic underassessment in maths by teachers against children with obesity. The larger achievement gap using the overall teacher score is in accordance with the hypothesis that the single-question assessment allows for more subjectivity compared with the multi-item skill-based ARS assessment.

The observed discrepancy between test scores and teacher assessments for children with obesity in maths, but not literacy, is an empirical fact for Australia that has not been shown previously. There is little evidence to compare this finding to. While very limited evidence from the United States found body weight influenced teacher assessments but not test scores in reading (Zavodny, 2013 and Kenney *et al.*, 2015), we find it reasonable to expect differences in the teacher-test score gap between maths and literacy in Australia. Borrowing from the larger literature on teacher-test score gaps by gender, there is empirical evidence of a teacher-test score gap in maths but not literacy (e.g. in France, Terrier, 2016). It has been suggested that teachers hold a belief that maths (but not literacy) is more difficult for girls than boys due to an assumed inferior ability in this area (e.g. Riegle-Crumb and Humphries, 2012 and the literature therein). A similar belief of inferior ability in general intelligence and reasoning skills for children with obesity has been shown to be held by peers and teachers (Puhl and Latner, 2007). Given maths ability is highly correlated with general intelligence and problem-solving skills (Aiken, 1973; Navas-Sánchez *et al.*, 2014), it is conceivable that there could be biased assessments against children with obesity in maths, but not literacy.

Since the teacher-test score differences in maths achievement by obesity are particularly striking, the remainder of the paper focuses on investigating whether these raw differences in maths scores could be due to biased teacher assessments. To do this, we need to rule out alternative explanations, such as other child characteristics (e.g. classroom behaviour) which may be jointly correlated with the child's body size and teacher assessments. We also need to rule out the possibility that such characteristics are systematically correlated with the child's own performance in tests compared with in-class assessments. The following section details the methods we use to formally understand the teacher-test score differences by body size.

III. Empirical approach

Measuring discrepancies in teacher-test scores by body size

Our objective is to determine whether non-blind teacher assessments S_{it}^{nb} are systematically influenced by a child's body size, over and above the child's true ability A_{it}. To begin with, let us assume that the teacher assessment S_{it}^{nb} in maths is only determined by the child's true underlying ability A_{it} in maths and child's body size $\boldsymbol{Z}_{it}$ (measured by height and weight or alternatively BMI), with some idiosyncratic noise in teacher assessments ε_{it}^{nb}:

$$S_{it}^{nb} = a_{nb} + \beta_{nb} A_{it} + \tau \boldsymbol{Z}_{it} + \varepsilon_{it}^{nb} \tag{1}$$

where i refers to the child and t to the time. A non-zero τ would suggest a bias in teacher assessments by student's body size. However, ability A_{it} in equation (1) is unobserved and may potentially be correlated with $\boldsymbol{Z}_{it}$. Estimating equation (1) without A_{it} might lead to the false conclusion of a non-zero τ. We therefore exploit the availability of blind test scores in maths. We assume that the blind test score S_{it}^{b} is also determined by the child's true ability A_{it} and that ability is the only relevant factor in determining test scores:

$$S_{it}^{b} = a_b + \beta_b A_{it} + \varepsilon_{it}^{b}. \tag{2}$$

Ideally, S_{it}^{b} will equal A_{it}, but there will be testing noise ε_{it}^{b}, some students may perform below their capabilities on the testing day, while others may get lucky and perform above their normal capabilities.

We can combine equations (1) and (2) by subtracting the test score from the teacher score such that ability A_{it} is differenced out. Naturally, any other unobserved factors (such as child, family and school characteristics) that independently influence teacher scores would also be differenced out, provided they influence test scores and teacher scores equally. Although both test and teacher scores are standardized, they are measured on different scales. To enable comparability, we follow Cornwell *et al.* (2013) and allow S_{it}^{b} to relate to S_{it}^{nb} by a scaling factor of $\rho = \frac{\beta_{nb}}{\beta_b}$:

$$S_{it}^{nb} = \alpha_1 + \rho S_{it}^{b} + \tau \boldsymbol{Z}_{it} + \varepsilon_{1it} \tag{3}$$

where $\alpha_1 = (a_{nb} - \rho a_b)$ and $\varepsilon_{1it} = (\varepsilon_{it}^{nb} - \rho \varepsilon_{it}^{b})$. Coefficient vector τ in equation (3) measures the discrepancy between teacher assessments and test scores by characteristics $\boldsymbol{Z}_{it}$ under the assumption that $E(\varepsilon_{1it} | \boldsymbol{Z}_{it}, S_{it}^{b}) = E(\varepsilon_{1it} | S_{it}^{b})$, meaning that conditional on S_{it}^{b}, $\boldsymbol{Z}_{it}$ is as if randomly assigned and hence uncorrelated with ε_{1it}. One might be concerned if, for example, heavier students self-selected into schools or classes within schools where teachers were particularly lenient (or harsh) in their assessments or where teachers were particularly capable (or incapable) in preparing students for the exams. To control for potential non-random selection of students to schools, in our base model, we include controls $\boldsymbol{X}_{it}$ for the type as well as performance of the school. In particular, we control for the three main types of schools in Australia; public, Catholic and independent, with independent schools being the most expensive. We also include a variable that captures each school's average performance in blind maths tests (relative to other schools).

We control for potential selection into schools by socioeconomic background as a socioeconomic gradient has been demonstrated in obesity (Shrewsbury and Wardle, 2008) and height (Finch and Beck, 2011). We add controls for mother and father's education, a single-parent household indicator and an index of neighbourhood socioeconomic disadvantage (SEIFA). We also include an index of parental involvement with the school to capture the parents' engagement in their child's education.

Although selection into schools is likely to be non-random, assignment to teachers within schools should be as good as random in Australia. Using the 2011 Trends in Mathematics and Science Study (TIMSS) sample of Australian Year 8 students, Ryan (2017) found little evidence of non-random assignment of students to teachers. Nevertheless, in additional specifications, we also control for the number of years of teaching experience and the gender of the teacher (see section A16) and find that these characteristics do not

affect our results. Furthermore, our within-child FE models (see section below on sensitivity of the teacher-test score gap to fixed unobserved characteristics) control for teacher and school FE. We are therefore confident that non-random selection to teachers by student's body size is not a concern in our models.

Previous studies have shown that a student's gender and ethnicity are associated with teacher-test score discrepancies (e.g. Lavy, 2008; Burgess and Greaves, 2013). As these characteristics may also be correlated with body size, we include the gender and ethnicity of the child in our basic set of control variables $\boldsymbol{X}_{it}$. We include month FE of the LSAC teacher assessment date to control for influences of the timing of teacher assessments. This controls for possible differences in teacher assessments after NAPLAN results are released in September. Our main estimation model is:

$$S_{it}^{nb} = \alpha_2 + \rho_2 S_{it}^{b} + \tau_2 \boldsymbol{Z}_{it} + \gamma \boldsymbol{X}_{it} + \varepsilon_{2it}. \tag{4}$$

We allow test scores to relate nonlinearly to teacher scores by including the square of the test score in the model. We estimate our models using OLS, however, in the Online Appendix, we show the robustness of our results for the overall teacher assessment (which is measured on a 5-point scale) to estimating the underlying linear model as an ordered logit model in which the dependent variable is ordered with five categories (m =far below average, below average, average, above average and far above average).

Explaining teacher-test score discrepancies by body size

After establishing teacher-test score discrepancies by height and weight controlling for the basic set of controls in equation (4) outlined above, we examine possible explanations for these discrepancies. If we can rule out observed and unobserved factors correlated with body size $\boldsymbol{Z}_{it}$ that influence the teacher and test scores differently and still find a teacher-test score discrepancy by body size, then we might reasonably attribute this discrepancy to a teacher bias. We investigate this in three parts (detailed below) and subsequently draw conclusions based on the combined evidence. First, we sequentially add sets of control variables that are potentially correlated with both body size and a teacher-test score gap. Second, we estimate bounds that show the sensitivity of the conditional teacher-test score gap by body size to unobservable characteristics. Finally, we use a linear FE estimator (within individuals across maths and literacy assessments) to control for fixed unobserved characteristics of the student, school and teacher.

Sensitivity of the teacher-test score gap to observed characteristics

As shown in Online Appendix Table A2, children with obesity tend to perform worse on the problem-solving cognitive test and are more likely to have emotional, peer and conduct problems. When assessing the child's academic ability in a specific subject, teachers may be influenced by the child's general cognitive ability or intelligence and not accounting for them might lead to an overestimate of the teacher-test score gap. We test for this possible explanation by including additional covariates in equation (4) that measure the child's general cognitive ability: standardized scores from the Matrix Reasoning test (taken in

grade 5) and the Peabody Picture Vocabulary Test (taken in grade 3).[3] We also control for: whether the child has repeated a grade; whether the child is a student with special needs; and, whether the child frequently misses school days due to illness (reported by the parent).

While in principle, teachers should not let student classroom behaviours influence the assessment of academic skill level, it is difficult to achieve this degree of objectivity in practice. We examine the influence of student behaviours by controlling for scores of the child's behaviours on five dimensions: emotional problems, peer problems, hyperactivity-inattention, conduct problems and prosocial behaviour. These scores are taken from the Strength and Difficulties Questionnaire (SDQ), a widely used measure of psychological adjustment that is reported by the same teacher that assesses the academic ability of the child. It is important to note that the teacher-rated SDQ scores could also suffer from teacher bias and therefore, controlling for SDQ scores could lead to an underestimation of the teacher-test score gap by body size. We also include a score of the child's level of motivation and participation in class. This is measured by taking the average of responses by the child to eight questions about their schooling: 'I like learning', 'I get enjoyment from being at school', 'School work is interesting', 'I ask questions in class', 'I do extra work', 'I enjoy what we do in class', 'I always do my best' and 'I get excited about the work'.

Sensitivity of the teacher-test score gap to unobserved characteristics

It is possible that even after controlling for a wide range of school and sociodemographic characteristics, cognitive skills and classroom behaviours, there may be other unobserved omitted variables. For example, our set of behavioural control variables may not fully capture the behavioural traits that teachers take into account when conducting maths skill assessments. It is therefore useful to gauge the sensitivity of the teacher-test score gap to selection on unobserved characteristics (unobservables). For all of our specifications that use BMI z-scores as a single measure of body size, we include bounds that show the sensitivity of the conditional teacher-test score gap by body size to unobservables under a plausible degree of correlation between unobservables with observables. Following Altonji, Elder and Taber (2005) and Oster (2019), for the upper bound, we calculate $|\tau_2|$ under the scenario that selection on unobservables is 0 ($\delta = 0$) (i.e. unobserved characteristics do not bias τ_2). For the lower bound, we calculate the estimate for $|\tau_2|$ under the scenario that selection on unobservables equals selection on observables ($\delta = 1$). This is a reasonable lower bound of $|\tau_2|$ as the set of unobservables should not be more influential than the observed covariates, especially in more extensive specifications that have a relatively high R^2 (McConnell and Rasul, 2018). We also calculate the degree of selection on unobservables relative to selection on observables necessary to eliminate the estimated effect of body size (i.e, for τ_2 to equal 0). Bounds are calculated using the R-squared from a hypothetical regression of the teacher score on BMI and both observed and unobserved controls, R_{max}, which is assumed to be $1.3(R^2)$, where R^2 is the R-squared from a regression with controls (Oster, 2019). In

[3] The Matrix Reasoning test is a subset of the Wechsler Intelligence Scale for Children, which measures nonverbal problem-solving and reasoning skills. The Peabody Picture Vocabulary Test (PPVT) aims to measure knowledge of spoken words and receptive vocabulary. Both the Matrix Reasoning test and PPVT were administered by LSAC interviewers in the child's home. We standardize all test scores (mean=0, SD=1). To maintain sample size, children with missing scores were included as having a score of 0 and indicators for missing a score were included.

robustness specifications, we calculate the bounds assuming a more conservative R_{max} of 1 (see Altonji, Elder and Taber, 2005)

Sensitivity of the teacher-test score gap to fixed unobserved characteristics

We further explore the possible role of selection on unobservables by controlling for fixed unobserved traits across subjects that could be associated with body size (e.g. self-confidence) and which may either systematically affect the child's performance in tests differently to in-class performance, or may be valued differently by teachers to what is captured in test scores.

One recognized phenomenon that might systematically affect test or in-class performance by body size is 'stereotype threat'. Stereotype threat occurs when someone is aware of a negative stereotype about how they are likely to perform in some situation (e.g. a child with obesity may be aware of stereotypes around obesity such as being 'lazy' or 'stupid', (Puhl and Latner, 2007)), and this fear of confirming the stereotype leads to poorer performances. This fear or anxiety may be an unobserved trait that explains the teacher-test score gap.

Related to stereotype threat is the possibility that children's performance in tests is influenced by the in-class feedback and comments they receive from their teachers. To illustrate, assume two children of differing body size display equal in-class maths skills. Both are given the same critical feedback to improve their maths skills, but the heavier child is more sensitive to the feedback than the lighter child (perhaps due to lower self-esteem). In line with self-fulfilling prophecies, the performance of the heavier child in the national test scores may be adversely affected compared with their lighter peer. This would imply that there is an unobserved differential response to a teacher's feedback by the child's body size.[4]

To determine whether such fixed unobserved characteristics of the child or indeed the teacher, school or class could explain any remaining teacher-test score discrepancy, we estimate a within-individual FE model across subjects. The idea behind this approach is that if, for example, heavier children are less self-confident than leaner children, then it can be assumed that this is a fixed trait of the child, which holds constant across subjects. This is particularly likely in the current setting where primary school children typically take all their core subjects, including maths and English, in the same classroom with the same teacher.

It is possible to difference out fixed student characteristics if one has information on the student–teacher pair across multiple subjects (see e.g. Dee, 2005, 2007; Clotfelter, Ladd and Vigdor, 2010; Altinok and Kingdon, 2012; Burgess and Greaves, 2013).[5] We use a linear FE estimator (across maths and literacy assessments) to look for variation in teacher assessments across subjects within a student–teacher pair by the child's body size:

$$S_{ij}^{nb} = \alpha_3 + \rho_3 S_{ij}^{b} + \tau_3 \boldsymbol{Z}_i + \varphi M_{ij} + \vartheta \left(\boldsymbol{Z}_i \times M_{ij}\right) + \omega \boldsymbol{X}_i + \mu_i + \psi_{ij} \quad (5)$$

[4] Unobserved traits associated with stereotype threat or differential responses to teacher feedback would only lead to an overestimate of harsher teacher assessments relative to test scores by weight (or short stature) if the unobserved trait has a more *positive* influence on the test performance of heavier (shorter) children than lighter (taller) children.

[5] A more traditional within-individual FE model *over time* is not preferable for two main reasons. First, because the teacher changes each year, it is not possible to estimate the FE model within a student–teacher pair over time. Second, traits and behaviours (which may explain the teacher-test score discrepancy) can vary considerably over time among children (e.g. Haney and Durlak, 1998), and as such, will not be adequately controlled for in a FE model over time.

where S_{ij}^{nb} is the teacher assessment, and S_{ij}^{b} is the test score for child i for subject j (either maths or literacy), M_{ij} is an indicator that equals 1 if the subject is maths, and 0 if it is literacy. All observed factors that are fixed across the two scores (i.e. characteristics $\boldsymbol{X}_i$, body size $\boldsymbol{Z}_i$) are removed through differencing between maths and literacy. Similarly, any unobserved factors that affect maths and literacy performance in the same way (e.g. self-confidence, value placed on tests, behaviour or motivation, other unobserved child, family, school, teacher and neighbourhood FE) are implicitly controlled for.[6] To identify variation in teacher scores by body size, we interact body size variables with the subject M_{ij} as body size is invariant between subjects and therefore differenced out.

It is important to note that the size of the estimates are not comparable to the estimates of equation (4), but they are also not of primary interest. We are mainly interested in whether the interaction term is significantly different from zero. This would indicate that fixed unobserved traits and behaviours of the child cannot explain the teacher-test score gap by body size, and implies that biased maths teacher assessments is a possible explanation. As Burgess and Greaves (2013) note 'this (fixed effects estimation) is a powerful test since such a large degree of variation is being controlled for with the student fixed effects' (p.559). It is also important to keep in mind that an interaction effect of zero does not rule out biased teacher assessments. A coefficient of ϑ close to zero only indicates that teacher-test score gaps of equal size exist in both subjects; this gap may not necessarily be zero. Because the interpretation of the interaction term relies on the relative size of the teacher-test score gap in maths to the one in literacy, we show the literacy results in the Online Appendix and refer to them in the corresponding results section.

If the combined evidence from all three types of analyses (that control for observed characteristics, test the sensitivity to unobservable characteristics and control for unobserved fixed traits) consistently reveal that systematic and significant differences between teacher and test scores by body size remain, a reasonable conclusion is that this is due to biased teacher assessments.

IV. Main results

Identifying and explaining teacher-test score discrepancies

Table 1 shows the coefficient estimates for teacher assessments in maths for equation (4). Columns 1–3 show the estimates for the ARS teacher score, while columns 4–6 show the estimates for the overall teacher score. Column (1), Panel A shows that after controlling for blind test scores (and its square) and a set of basic controls including socioeconomic background, students who are one standard deviation (SD) taller than the mean (0.31) are given a significantly more favourable maths rating by the teacher under the ARS by about 0.08 SD ($P < 0.01$), while students who are 1 SD heavier than the mean (0.33) are rated significantly lower by about 0.06 SD ($P < 0.05$).[7] We find similar results when alternative measures of body size are used instead of height and weight. Panel B shows the

[6] Equation (5) is essentially equivalent to a 'triple-difference' approach (Breda and Ly, 2015) as it combines the comparisons of non-blind teacher assessments with blind tests with within-student comparisons across subjects.

[7] Online Appendix Table A11 shows results from regressions that include the basic covariates in a stepwise manner. Overall, the raw teacher-test score gap does not change much as each of the basic control variables are included.

TABLE 1

Explaining the observed teacher-test score gap with observed characteristics

	ARS teacher score			*Overall teacher score*		
	(1)	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*
A.						
Height z-score	0.083***	0.072***	0.073***	0.099***	0.092***	0.092***
	(0.025)	(0.024)	(0.023)	(0.024)	(0.023)	(0.023)
Weight z-score	−0.059**	−0.063***	−0.064***	−0.074***	−0.078***	−0.076***
	(0.023)	(0.022)	(0.022)	(0.023)	(0.022)	(0.022)
B.						
Obese	−0.188***	−0.197***	−0.217***	−0.270***	−0.277***	−0.279***
	(0.071)	(0.070)	(0.068)	(0.077)	(0.074)	(0.073)
Overweight	−0.030	−0.042	−0.049	−0.022	−0.030	−0.032
	(0.046)	(0.044)	(0.042)	(0.043)	(0.041)	(0.040)
Underweight	−0.027	−0.009	−0.010	−0.037	−0.030	−0.023
	(0.070)	(0.068)	(0.066)	(0.068)	(0.069)	(0.069)
C.						
BMI z-score	−0.029*	−0.037**	−0.036**	−0.036**	−0.041**	−0.040**
	(0.017)	(0.016)	(0.016)	(0.017)	(0.016)	(0.016)
[Bounds: τ_0', τ_1']	[−0.029,	[−0.037,	[−0.036,	[−0.036,	[−0.041,	[−0.040,
	−0.016]	−0.025]	−0.025]	−0.022]	−0.028]	−0.026]
δ required for coefficient of 0	2.119	3.063	2.950	2.500	3.129	2.854
R_{max}	0.582	0.640	0.687	0.633	0.675	0.709
Adjusted R^2	0.438	0.482	0.517	0.478	0.509	0.534
Test score	✓	✓	✓	✓	✓	✓
Basic controls	✓	✓	✓	✓	✓	✓
Cognitive ability	×	✓	✓	×	✓	✓
Student behaviour	×	×	✓	×	×	✓
N	1930	1930	1930	1930	1930	1930

Notes: *$P < 0.1$, **$P < 0.05$, ***$P < 0.01$. Robust standard errors in parentheses. All models estimated by OLS. Test score is standardized NAPLAN (blind) score for maths and its square. Basic controls include child's sex, ethnic background, school type, school's average NAPLAN score in maths, mother and father's education, single-parent household indicator, index of parental involvement with the school, index of neighbourhood socioeconomic disadvantage (SEIFA) and survey month dummies. Cognitive ability includes Matrix Reasoning score, Peabody Picture Vocabulary Test score, whether repeated a grade previously, whether recorded as a special needs student, and whether frequently missing school days due to illness. Student behaviour includes five scales from the teacher reported Strength and Difficulties Questionnaire (prosocial, problems with hyperactivity, emotions, peers and conduct) and an index of motivation to learn and attend school (scale of 0–4). The reported bounds show the sensitivity of the BMI estimates to selection on unobservables based on selection on observables. τ_0' assumes that the proportional degree of selection on unobservables to selection on observables is 0 ($\delta = 0$) and is therefore equivalent to our estimate for τ', while τ_1' assumes $\delta = 1$. $R_{max} = 1.3$ (R^2).

estimates for BMI categories; students with obesity are rated 0.19 SD lower than children of normal weight ($P < 0.01$); however, there is no teacher-test score gap associated with being overweight (but not obese) or underweight. Panel C shows that when a continuous measure of BMI is used, students who have a 1 SD higher BMI than the mean (0.33) are rated 0.03 SD lower ($P < 0.10$).

Column (2), which adds controls for the child's cognitive ability, shows that cognitive ability reduces the estimated score discrepancy by height (by 13%), suggesting that

teachers may take the general cognitive ability of taller children into account (over and above performance in maths tests) when assessing their maths skills. The teacher-test score discrepancy by weight increases slightly (by 7%) when cognitive ability controls are added. Of the measures of cognitive ability, being classified as a 'special needs' student and achieving a lower score on the Matrix Reasoning and PPVT tests are all independently associated with a lower teacher score relative to the blind test score.

Adding student behaviour variables in column (3) has little influence on the estimated teacher-test score discrepancy by height and weight (although children with hyperactivity and emotional problems are rated significantly less favourably by their teacher, relative to their blind test score). Although not shown, the addition of student behaviours has a considerable impact on teacher-test scores by gender. Boys and girls are rated similarly under the ARS score, but after controlling for hyperactivity, which is highly correlated with boys, there is a teacher-test score gap in favour of boys (by about 0.20 SD). A similar relationship between student behaviours and teacher-test score gaps by gender were found by Cornwell *et al.* (2013).

Even after controlling for basic controls, cognitive ability and student behaviours, we find that statistically significant teacher-test score discrepancies by height, weight, BMI and obesity exist. In the most complete specification, a 1 SD increase in height, weight and BMI relative to the mean is associated with an ARS teacher rating that is 0.07 SD higher, 0.06 SD lower and 0.04 SD lower respectively. A child with obesity is rated 0.22 SD lower than a normal weight child who performs equally well in maths blind tests.

The results in Table 1, columns 4–6 show that when the overall teacher score is examined instead of the ARS score, a very similar pattern emerges, but the magnitude of the discrepancy in scores by height and weight is even greater. In the most complete specification (column 6), a 1 SD increase in height, weight and BMI relative to the mean is associated with an overall teacher rating that is 0.09 SD higher, 0.08 SD lower and 0.04 SD lower respectively. Relative to a child of normal weight, a child with obesity is rated 0.28 SD lower.

Supplementary models, which include squared terms for height and weight or BMI confirm the main results. Estimates are very similar and, in all models, the squared terms are small and statistically insignificant.[8] The findings are also very similar if we estimate the models in Table 1 including the blind literacy test score and its square as additional control variables. Online Appendix Table A12 shows these results, which suggest that the teacher's impressions of the student's performance in literacy cannot explain our results. Estimating the models in Table 1 using an ordered probit instead of OLS also yields similar results. Online Appendix Table A13 shows the marginal effects for height, weight, BMI and BMI categories when all additional covariates are included. It shows, for example that a child with obesity is 8.3% points less likely to be rated 'above average', and 3.6 % points less likely to be rated 'far above average' by their teacher compared with a child of normal weight of equal maths performance in test scores.

Additionally, we investigated whether heterogeneity by the child's gender, teacher's gender or teacher's experience exist. Our results (see Online Appendix Table A14) show

[8] When squared terms for height and weight, or BMI, are included in the ARS teacher score equation (with all additional covariates), the estimated marginal effects (at the mean) for height, weight and BMI for the ARS teacher score are 0.07, −0.07 and −0.04 respectively, and for the overall teacher score are 0.09, −0.08 and -0.05 respectively.

TABLE 2

Heterogeneity in the teacher-test score gap by information from national test results

	ARS teacher score		*Overall teacher score*	
	(1)	*(2)*	*(3)*	*(4)*
Height	0.092***		0.086***	
	(0.033)		(0.031)	
Height × post-Sep	−0.054		−0.007	
	(0.049)		(0.050)	
Weight	−0.087***		−0.084***	
	(0.030)		(0.030)	
Weight × post-Sep	0.059		0.025	
	(0.046)		(0.048)	
BMI z-score		−0.053**		−0.050**
		(0.022)		(0.023)
BMI z-score × post-Sep		0.044		0.026
		(0.034)		(0.036)
N	1604	1604	1604	1604

Notes: *$P < 0.1$, **$P < 0.05$, ***$P < 0.01$. Robust standard errors in parentheses. Post-Sep is an indicator that equals one if the teacher assessment was conducted in October to December and equals 0 if conducted in May to August. Children whose teacher assessments were conducted during September were excluded. Includes full set of covariates, replicating columns 3 and 6 of Table 1. See Table 1 notes for description of covariates.

that there are no differences by child's gender, teacher's gender or experience in the teacher-test score discrepancy by height, weight, BMI or obesity under either form of teacher assessment.

Does information on test scores reduce the teacher-test score gap?

Results from the externally marked national tests, which are revealed in September each year, provide an opportunity for teachers to gain (new) information on the child's academic ability and potentially update their assessment of the child. If teachers took this new information into account, it is possible for teacher assessments conducted after September to be systematically less 'biased' than assessments conducted prior to September.

To examine whether information gained from national test scores reduces the teacher-test score gaps shown in Table 1, we exploit the variation in the month that the teacher assessments were conducted and re-estimate the most complete specification (i.e. columns 3 and 6 of Table 1) with an interaction term for whether or not the teacher assessment was conducted after September. We exclude all students whose teacher assessments were completed during September (because results were released part way through the month). The sample spans the teacher assessment months of May to December and we retain the survey month dummies in the model to control for individual month effects. A school year typically starts at the beginning of February and ends in December. Table 2 shows that the interaction term is negative for height and positive for both weight and BMI, which implies that there is less 'bias' in assessments conducted after teachers are exposed to the national exam results. However, the interaction terms are imprecisely estimated and not statistically significant. We therefore conclude that while our results are consistent with

new information reducing the teacher-test score gap, this deserves further investigation in future research using a larger sample of students.

Can unobserved characteristics explain the teacher-test score gap?

The teacher-test score discrepancy could not be explained by any of the observed characteristics included in our model. In the lower half of Panel C in Table 1, we report the bounds of the BMI estimate, which show the sensitivity of the estimates to selection on unobservables based on Oster (2019). The lower bound, τ_0' assumes that the proportional degree of selection on unobservables to selection on observables is 0 and is therefore equivalent to our estimate for τ', while the upper bound, τ_0' assumes the selection on unobservables equals the selection on observables. It is notable that our estimated bounds change very little as more observed covariates are included. Even in our most complete specification (columns 3 and 6), the bounds do not contain zero, which indicates that BMI is associated with significantly lower teacher scores even when selection on unobservables is as large as selection on observables.

Moreover, the results suggest that the degree of selection on unobservables (δ) needed for the teacher-test score gap by BMI to be zero is very high (about 3 times as large as the selection on observables). This gives us a high degree of confidence that our OLS estimates for BMI are unlikely to be driven by unobserved characteristics that might otherwise explain the teacher-test score discrepancy. Even when we take the conservative assumption that $R_{max} = 1$, we find $\delta = 1$ for both the ARS and overall teacher score when all covariates are included. This suggests the selection on unobservables needs to be at least as large as the selection on observables in order to eliminate the estimated teacher-test score gap by BMI.

Unobserved fixed characteristics results

To formally test whether the teacher-test score discrepancy remains after controlling for any unobserved traits that are associated with body size, we estimate the within-individual FE model across subjects from equation (5). As the coefficient of interest is an interaction term (between body size and a maths subject indicator), it needs to be interpreted in relation to the estimated teacher-test score gap by body size for literacy. Controlling for the full set of covariates, the results for literacy (see Online Appendix Table A15) indicate the coefficient estimates are positive for height and negative for weight but they are small and not statistically significant. Similarly, there is no significant association when BMI or BMI categories are used to measure body size. These results confirm the findings in Figure 1 and suggest there is no discernible teacher-test score discrepancy by body size for literacy.

The results from the FE model are shown in Table 3. Columns (1) and (4) show the OLS estimates for the estimation sample, and Columns (2) and (5) show the FE estimates for the ARS and overall teacher scores respectively. The FE estimates for height and weight are attenuated as expected, and height is no longer statistically significant (see Panel A) One possible explanation for the reduced and insignificant teacher-test score discrepancy by height is that unobserved fixed traits are correlated with height and teacher maths

TABLE 3

Fixed effects estimates of the variation in teacher assessments across subjects

	ARS teacher score			*Overall teacher score*		
	(1)	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*
	OLS	*FE*	*FE*	*OLS*	*FE*	*FE*
A.						
Height × maths	0.099***	0.023	0.023	0.102***	0.035	0.036
	(0.024)	(0.020)	(0.020)	(0.023)	(0.022)	(0.022)
Weight × maths	−0.081***	−0.054***	−0.053***	−0.087***	−0.046**	−0.045**
	(0.022)	(0.019)	(0.019)	(0.022)	(0.020)	(0.020)
B.						
Obese × maths	−0.255***	−0.092	−0.090	−0.311***	−0.179***	−0.177***
	(0.073)	(0.062)	(0.062)	(0.073)	(0.068)	(0.067)
Overweight × maths	−0.061	−0.072**	−0.070**	−0.049	−0.029	−0.027
	(0.044)	(0.035)	(0.035)	(0.042)	(0.038)	(0.037)
Underweight × maths	0.006	0.018	0.017	−0.010	0.041	0.040
	(0.067)	(0.058)	(0.058)	(0.064)	(0.064)	(0.063)
C.						
BMI × maths	−0.046***	−0.040***	−0.039***	−0.047***	−0.030**	−0.029**
	(0.016)	(0.014)	(0.014)	(0.016)	(0.015)	(0.015)
I like the subject	×	×	✓	×	×	✓
N	4166	4166	4166	4166	4166	4166

Notes: The dependent variable is the teacher score in maths and literacy. Variation is measured within students across subjects (maths and literacy). Height, weight and BMI are all measured in z-scores. Body size variables must be interacted with the maths subject indicator otherwise they would be differenced out of the model. Normal weight is the reference category in Panel B. 'I like the subject' = 1 if the student reported they like the subject. Models also include basic controls from Table 1 (child's sex, ethnic background, school type and school's average NAPLAN score in maths) interacted with maths, the blind test score and an indicator for whether the score is for maths (or literacy). The number of observations increases from 1,930 (Table 1) to 4,166 (Table 3) for two reasons: (a) we add the literacy scores for each child by stacking the maths and literacy scores in our dataset; and, (b) the fixed effects estimation automatically controls for fixed unobserved characteristics which allows us to drop fixed observable variables, resulting in fewer missing observations. Robust standard errors in parentheses. *$P < 0.1$, **$P < 0.05$, ***$P < 0.01$.

assessments[9] Self-esteem, for example, is linked with greater confidence and subjective perceptions of performance by others (Judge and Cable, 2004), which may explain the more favourable maths teacher assessment by height.

Although low self-esteem may also be correlated with obesity and lower teacher assessments, our results suggest that self-esteem (or any other fixed trait or behaviour) does not explain the teacher-test score discrepancy against heavier children. In contrast to the estimates for height, weight remains significantly associated with lower teacher assessments in maths (for both the ARS and overall teacher score). This is supported by significant estimates for BMI z-scores and BMI categories under the FE model (see Panels B and C). Under the ARS assessment, we see highly significant estimates for children in the over-

[9] Since we see no teacher-test score discrepancy in literacy scores, we conclude that the insignificant coefficient for the interaction between height and maths subject indicator in the FE model suggests an absence of a teacher-test score discrepancy by height after accounting for fixed traits, rather than the alternative explanation, which is that the teacher-test score gap by height is positive and of about equal size for maths and literacy.

weight category. The size of the estimate is even larger for children with obesity, though this is less precisely estimated. Overall, we see a striking pattern of significant and large effects for weight, BMI and heavier BMI categories, even after controlling for fixed unobserved traits.

While the significant FE estimates are by themselves a strong result, it remains possible that subject-specific traits that are associated body size, explain the teacher-test score discrepancy. It is difficult to think of examples of this, but if heavier children systematically displayed a particular enthusiasm for literacy, but not for maths, then this might explain the significant FE estimates for weight. Although we do not have subject-specific measures of enthusiasm or confidence, LSAC asks students to report whether they like maths and literacy. It is likely that this response is correlated with other subject-specific differences in student behaviours. Therefore, in an additional specification, we also control for a binary measure of whether the child likes the subject (no or sometimes=0, and yes=1). We show in columns (3) and (6) that controlling for this subject-specific child characteristic makes little difference to the FE estimates; weight (and BMI and heavier BMI categories) are still associated with a harsher teacher assessment.[10]

Taking the main FE results together with the robustness of the OLS estimates for BMI to selection on unobservables, we conclude that a teacher bias is the most likely explanation for the remaining teacher-test score discrepancy by weight BMI or obesity. For height, however, it seems that unobserved fixed characteristics of the child could explain the teacher-test score gap.

V. Impact of teacher assessments on human capital development

Biased teacher assessments may be considered an undesirable outcome in themselves, but they are arguably harmful only if they have negative future consequences. It is not clear *a priori* that harsher teacher assessments are bad and more lenient assessments are not. According to the theories on self-fulfilling prophecies, students have a tendency to live up to the expectations and assessments of their teachers, which implies that under-assessments would have damaging consequences, while over-assessments would have positive consequences (Rosenthal and Jacobson, 1968; Ferguson, 2003). However, the experimental evidence remains inconclusive (Jussim and Harber, 2005). Counter to the self-fulling prophecy hypothesis, it is feasible that students who are harshly assessed try harder to win the recognition from their teachers, which could have positive outcomes. Leniently assessed students could feel overconfident and slacken their efforts, which could have adverse outcomes.

We test for whether teacher assessments are independently associated with the future academic achievement of children, over and above their academic abilities as measured by blind test scores and unobserved fixed student characteristics. We essentially estimate a value-added FE model.

[10] As noted in footnote 6, a within-student FE model over time is not preferable to a within-student FE model across subjects in this study. Nevertheless, we show in Online Appendix Table A16 estimates using a FE model over time using grade 3 and grade 5 teacher and test scores. A significant teacher-test score discrepancy is found for weight but not height for the overall teacher assessment, which is very similar to our main FE estimates. However, for the ARS teacher assessment, there is no effect for either height or weight. It is important to note that the teacher is not the same across these two time points and therefore the student-teacher pair is not fixed in this model. Changes in child traits over time also mean that important behaviours are unlikely to be controlled for here.

Value-added fixed effects approach

Value-added models that use past academic performance to control for initial child conditions and past inputs into the production of human capital have been widely used in previous related studies on academic achievement (e.g. Todd and Wolpin, 2003, 2007). An advantage of the value-added model is that the lagged dependent variable controls for unobserved input histories and endowments, which may be correlated with both the teacher assessment and future academic achievement, thereby reducing confounding bias. However causal inferences are limited if relevant unobserved factors are not captured by the lagged dependent variable.

In our context, it might be possible that teachers observe a trait or behaviour of the student that is valued in current teacher assessments because it is viewed as an important determinant of future academic success, for example grit or perseverance. This trait may have little impact on current academic performance in grade 5 (and is therefore not captured by current test scores), but the teacher sees this as an important part of maths ability that will lead to better future maths performance when tests are more demanding in grade 9.[11]

We therefore use an extension of the value-added model to account for unobserved traits of the student that are fixed across subjects. Specifically, we use test scores and teacher assessments in maths and literacy to estimate a within-child FE value-added model across subjects. As with the FE used above, this approach allows us to control for unobserved traits of the child (and teacher, school and class) that are fixed across subjects and may be correlated with both the teacher assessment and future academic performance of the child. The approach has been used previously to measure teacher effectiveness (Slater, Davies and Burgess, 2012) and to examine the persistence of student achievement (Nicoletti and Rabe, 2018).

An alternative method used to account for unobserved heterogeneity in cognitive achievement models is the dynamic panel data estimator (Andrabi *et al.*, 2011), which removes the unobserved child fixed traits (or endowment) by comparing differences in gains in test scores measured at two different grades. As noted by Nicoletti and Rabe (2018), an advantage of our approach over the dynamic panel data estimator is that it relaxes the restrictive assumption that the unobserved child fixed traits (and other inputs for achievement) are invariant across the child's grades (see also Sass, Semykina and Harris, 2014).

Our model regresses future academic performance in blind test scores for maths in t+1 (grade 9) on blind test scores in t (grade 5) and teacher assessments:

$$S^{b}_{ijt+1} = \alpha_4 + \rho_4 S^{b}_{ijt} + \pi S^{nb}_{ijt} + \tau_4 \boldsymbol{Z}_{it} + \varphi_4 M_{ijt} + \omega_4 \boldsymbol{X}_{it} + \vartheta_{it} + \epsilon_{ijt} \tag{6}$$

where S^{b}_{ijt} is the test score for child i for subject j (either maths or literacy) and S^{nb}_{ijt} is the teacher assessment. M_{ijt} is an indicator that equals 1 if the subject is maths, and 0 if it is literacy. All observed factors $\boldsymbol{X}_{it}$ and body size $\boldsymbol{Z}_{it}$ which do not vary between subjects as well as unobserved child, family, teacher and school FE are differenced out by the between subject FE estimation.

The value-added FE model can identify the effect of teacher assessments on future academic achievement under the assumption that any subject-specific unobserved hetero-

[11] A separate normative issue is whether or not such personality traits or non-cognitive skills should be included by teachers in academic assessments of maths ability.

geneity is uncorrelated with teacher assessments. This assumption would not hold if, for example students displayed a trait (such as confidence or perseverance) in only maths (not literacy), and that trait was perceived as important in teacher assessments for maths and determined future maths test scores, but was somehow not captured by current test scores This is an unlikely, but theoretically possible scenario. To capture subject-specific student characteristics, we estimate additional models that control for whether or not the child likes the subject (as we did for the FE estimator).

We also re-estimate equation (6) using the child's future liking of the subject (in grade 9) as the dependent variable, controlling for liking of the subject test scores and teacher scores (in grade 5). The results can shed light on whether student perceptions and beliefs are a potential pathway through which teacher assessments influence future academic performance.

We are interested in π, which measures the association between the teacher assessment and future academic test performance (or future liking for the subject) over and above the influence of the student's current (blind) test performance, unobserved fixed child factors as well as child's current liking for the subject.

Results: impact of teacher assessments on future test scores

The estimates from the value-added FE model (equation (6)) for the ARS and overall teacher scores are presented in Table 4. Columns (1) and (4) show that both the ARS and overall teacher assessments (in grade 5) have a large and significant effect on the grade 9 test performance of students; a one SD increase in the teacher score is associated with a 0.18–0.19 SD in the grade 9 test score. Columns (2) and (5) show that even when we control for whether the child likes the subject, as a proxy for subject-specific child characteristics, there is still a large positive association between teacher assessments and future test scores. This has potentially harmful implications for children with obesity or who are heavier than average, as it suggests that harsher teacher scores may harm the child's future human capital development.[12]

We further show in columns (3) and (6) that teacher scores are independently associated with the child liking that subject in the future (in grade 9). This is consistent with the notion of self-fulfilling prophecies, where teacher assessments influence the student's engagement and enjoyment of a subject, which in turn influences future performances. An alternative explanation for our results is that teachers have a unique ability to predict the future performance of their students. If this were the case, teachers would have to be able to observe

[12] Following a suggestion by an anonymous reviewer, we also investigated the extent to which the relationship between obesity (grade 5) and future test scores (grade 9) can be explained by the teacher score (in grade 5) using value-added OLS and FE models that condition on grade 5 test scores (see Online Appendix Table A17). This is a strong specification because any effect that obesity has on academic performance in the same year is controlled for by the grade 5 test score. The OLS estimates show that after controlling for basic child, school, SES characteristics and blind test scores in grade 5, obesity is positively, but insignificantly, related to year 9 test scores (column 1). The FE model (our preferred specification) shows that after additionally accounting for fixed unobserved characteristics, obesity (interacted with maths) is negatively but insignificantly associated with grade 9 blind test scores (column 4). The addition of teacher scores (either ARS or overall) as a covariate in the FE specification reduces this negative association by 88%, which suggests that the teacher assessment might play a role in explaining the long-term negative (though statistically insignificant) association between obesity and future test scores.

TABLE 4

Fixed effect estimates of the effect of teacher assessments (in grade 5) on child's future test scores (in grade 9)

	Effect of ARS teacher score			*Effect of overall teacher score*		
	Test score grade 9	*Test score grade 9*	*I like the subject grade 9*	*Test score grade 9*	*Test score grade 9*	*I like the subject grade 9*
	(1)	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*
Teacher score (ARS)	0.178***	0.152***	0.073**			
	(0.026)	(0.025)	(0.028)			
Teacher score (Overall)				0.187***	0.161***	0.095***
				(0.022)	(0.022)	(0.025)
Test score (blind)	0.511***	0.458***	0.200***	0.488***	0.440***	0.185***
	(0.024)	(0.025)	(0.028)	(0.024)	(0.025)	(0.028)
Test score (blind) Sq.	0.012	0.015	0.006	0.009	0.013	0.005
	(0.011)	(0.011)	(0.012)	(0.011)	(0.011)	(0.012)
I like the subject		0.161***	0.246***		0.154***	0.240***
		(0.022)	(0.025)		(0.022)	(0.025)
Maths	0.006	0.012	0.009	0.004	0.009	0.008
	(0.016)	(0.015)	(0.017)	(0.015)	(0.015)	(0.017)
N	3061	3061	2968	3061	3061	2968

Notes: The dependent variable is the standardized NAPLAN grade 9 blind test score in maths and literacy, except in columns (3) and (6) where it is the child's self-reported liking of the subject (1=Yes; 0=Sometimes or No). Variation is measured within individuals across two subjects (maths and literacy). The teacher score (ARS or overall), blind test score, child's liking of the subject and subject indicator are included as covariates and measured in grade 5. Robust standard errors in parentheses. $*P < 0.1$, $**P < 0.05$, $***P < 0.01$.

a subject-specific trait that strongly influences future performance and is not captured by test scores or the student's enjoyment of the subject.

To put our results in more tangible terms we use our main estimates of the size of the teacher-test score discrepancy by obesity status which controls for all main covariates (from Table 1, column 6) to calculate the potential future academic consequences associated with teacher-test score gaps for a child with obesity. We focus on the estimates from the overall teacher score, because of the similarity of the format of this assessment to what students receive in their school reports.

An average grade 5 teacher-test score discrepancy of 0.28 SD against children with obesity is associated with grade 9 maths test scores that are about 4 points lower on the NAPLAN assessment scale.[13] The Australian Curriculum, Assessment and Reporting Authority (ACARA) have determined that 25 NAPLAN points is equivalent to the progress expected in one year of schooling (ACARA, 2010).[14] Given this, our results suggest that a 4-point lower assessment corresponds to a set-back of about 15% of one year of school (or 8 weeks).

[13] To calculate this, we multiplied the teacher assessment score (in SD) associated with obesity (=0.28) with the future NAPLAN scores (in SD) associated with teacher assessments (=0.18), and multiplied this with the standard deviation of NAPLAN scores (=74).

[14] Based on the difference in NAPLAN minimum standard scores across grades.

VI. Discussion and conclusion

Using data on primary school children and their teachers from LSAC, which are linked to externally graded national standardized test scores, this study identifies and examines the causes and potential consequences of teacher-test score discrepancies by student body size. Our quasi-experimental identification strategy combines teacher assessments in maths (non-blind with respect to child's body size) with externally marked national standardized test scores (blind with respect to child's body size), allowing us to control for the student's academic ability and other unobserved factors that are fixed across assessment methods.

Our results show there are significant teacher-test score discrepancies by a child's body size after accounting for basic controls. Relative to test scores, children who are shorter or heavier are rated less favourably by their teachers. These results are in line with Zavodny (2013) and MacCann and Roberts (2013) who showed that in the United States a child's weight is more negatively related to teacher assessments than test scores when analysing test scores and teacher scores separately. They also support results by Kenney *et al.* (2015) who showed that for boys in the United States increases in BMI from fifth to eighth grade were associated with decreases in teacher ratings controlling for changes in standardized test scores and age of the child.

We extend this small literature by building a model to identify bias in teacher assessments, rigorously examining alternative explanations for teacher-test score discrepancies by body size and examining potential future consequences of teacher-test score discrepancies. Our findings indicate that the teacher-test score discrepancy cannot be explained by the child's socioeconomic background, cognitive ability or classroom behaviours, including motivation to learn. We find that unobserved fixed traits potentially explain a large proportion of the favourable teacher assessment for taller children. However, the harsher teacher assessments by weight, BMI and obesity status persist. This finding is consistent with bounds analyses that show the teacher-test score discrepancy by BMI is highly robust to selection on unobservables (Oster, 2019). Taken together our analyses suggest that bias in teacher assessments is a plausible explanation for the harsher teacher assessments for heavier or obese children. We demonstrate that the teacher-test score gap arises in both the single overall assessment and the more objective multi-item ARS score, and occurs regardless of the teacher's experience and gender and the child's gender.

We test whether teacher assessments are independently associated with the future test performance of children, over and above their current test scores and other unobserved fixed traits. We show that harsher teacher assessments relative to test scores in grade 5 are linked with poorer performance in grade 9 exams, and that this holds even after accounting for subject-specific characteristics of the child such as whether they like the subject. Harsher teacher assessments are also linked with a much lower probability in the child liking maths 4 years later. These results are consistent with those of previous studies that show the significant effect of teachers on academic performance (e.g. Chetty, Friedman and Rockoff, 2014). There are no comparable studies that examine the academic impact of biased assessments by obesity. However, these results are in line with previous studies examining the relationship between gender stereotypes and academic performance (Alan, Ertac and Mumcu, 2018; Lavy and Sand, 2018; Carlana, 2019). For example, Carlana (2019) showed that teachers with stronger implicit gender stereotypes negatively impact

the maths performance of girls, resulting in an increase in the gender achievement gap over time.

The seriousness of maths achievement gaps by gender and race are widely recognized (Lavy, 2008; Burgess and Greaves, 2013; Cornwell *et al.*, 2013; Campbell, 2015; Terrier, 2016), and considerable progress has been made towards understanding the impact of teacher expectations and bias in high school and university course enrolment, occupational choices and earnings in adulthood (Lavy, 2008; Mechtenberg, 2009; Lavy and Sand, 2018). We find that accounting for socioeconomic background, cognitive ability and student behaviours, the teacher-test score gap for a child with obesity is of similar magnitude (0.22 to 0.28 SD) to the size of the teacher-test score gap against girls in our data (0.20 to 0.22 SD). With increasing rates of childhood and adolescent obesity globally (Ng *et al.*, 2014), and considerable labour market penalties due to obesity (Cawley, 2004), the consequences of teacher bias by body size needs further attention and investigation.

If teacher biases are indeed the reason for the discrepancies, a first simple step might be to make teachers aware that biases can occur given that biases often happen unconsciously (Stanley *et al.*, 2011). Bias awareness and 'bias literacy' have been shown to have potential in changing behaviours and reducing biases (Carnes *et al.*, 2015). Our results suggest that a policy response of moving from a single overall teacher assessment for maths (as is current practice in Australia), to an average of specific skill-based teacher assessments (like the 10-item ARS maths assessment) may achieve little, as our results showed that both ARS and overall teacher assessments are prone to bias. While an emphasis on (blind) scores from national tests such as NAPLAN in Australia would be preferable for a more objective and unbiased assessment of a child's maths performance, it is recognized that such exam scores may also imperfectly measure a child's true abilities and that teacher evaluations provide important complementary information. More research is needed into teacher assessment methods that can minimize bias, while still being informative.

Final Manuscript Received: March 2019

References

ACARA. (2010). *NAPLAN Score Equivalence Tables. Sydney*, Australian Curriculum, Assessment and Reporting Authority.

Aiken, L. R. (1973). 'Ability and creativity in mathematics', *Review of Educational Research*, Vol. 43, pp. 405–432.

Alan, S., Ertac, S. and Mumcu, I. (2018). 'Gender stereotypes in the classroom and effects on achievement', *Review of Economics and Statistics*, Vol. 100, pp. 876–890.

Altinok, N. and Kingdon, G. (2012). 'New evidence on class size effects: a pupil fixed effects approach', *Oxford Bulletin of Economics and Statistics*, Vol. 74, pp. 203–234.

Altonji, J. G., Elder, T. E. and Taber, C. R. (2005). 'Selection on observed and unobserved variables: assessing the effectiveness of Catholic schools', *Journal of Political Economy*, Vol. 113, pp. 151–184.

Andrabi, T., Das, J. Khwaja, A. I. and Zajonc, T. (2011). 'Do value-added estimates add value? Accounting for learning dynamics', *American Economic Journal: Applied Economics*, Vol. 3, pp. 29–54.

Australian Government. (2013). '*Guide to the Australian Education Act 2013*'. Retrieved 18 May 2017, from https://aeaguide.education.gov.au/content/d23-student-reports.

Averett, S. and Korenman, S., (1996). 'The economic reality of the beauty myth', *The Journal of Human Resources*, Vol. 31, pp. 304–330.

Black, N., Johnston, D. W. and Peeters, A. (2015). 'Childhood obesity and cognitive achievement', *Health Economics*, Vol. 24, pp. 1082–1100.
Botelho, F., Madeira, R. A. and Rangel, M. A. (2015). 'Racial discrimination in grading: evidence from Brazil', *American Economic Journal: Applied Economics*, Vol. 7, pp. 37–52.
Breda, T. and Ly, S. T. (2015). 'Professors in core science fields are not always biased against women: evidence from France', *American Economic Journal: Applied Economics*, Vol. 7, pp. 53–75.
Burgess, S. and Greaves, E. (2013). 'Test scores, subjective assessment, and stereotyping of ethnic minorities', *Journal of Labor Economics*, Vol. 31, pp. 535–576.
Campbell, T. (2015). 'Stereotyped at seven? Biases in teacher judgement of pupils' ability and attainment', *Journal of Social Policy*, Vol. 44, pp. 517–547.
Carlana, M. (2019). 'Implicit stereotypes: evidence from teachers' gender bias', *The Quarterly Journal of Economics*, Vol. 134, pp. 1163–1224.
Carnes, M., P. G. Devine, L. B. Manwell, *et al.* (2015). 'Effect of an intervention to break the gender bias habit for faculty at one institution: a cluster randomized, controlled trial', *Academic Medicine: Journal of the Association of American Medical Colleges*, Vol. 90, pp. 221–230.
Case, A. and Paxson, C. (2008). 'Stature and status: height, ability, and labor market outcomes', *Journal of Political Economy*, Vol. 116, pp. 499–532.
Case, A., Paxson, C. and Islam, M. (2009). 'Making sense of the labor market height premium: evidence from the British Household Panel Survey', *Economics Letters*, Vol. 102, pp. 174–176.
Cawley, J. (2004). 'The impact of obesity on wages', *Journal of Human Resources*, Vol. 39, pp. 451–474.
Chetty, R., Friedman, J. N. and Rockoff, J. E. (2014). 'Measuring the impacts of teachers I: evaluating bias in teacher value-added estimates', *American Economic Review*, Vol. 104, pp. 2593–2632.
Chiappori, P.-A., Oreffice, S. and Quintana-Domeque, C. (2012). 'Fatter attraction: anthropometric and socioeconomic matching on the marriage market', *Journal of Political Economy*, Vol. 120, pp. 659–695.
Chiappori, P.-A., Oreffice, S. and Quintana-Domeque, C. (2016). 'Black–White marital matching: race, anthropometrics, and socioeconomics', *Journal of Demographic Economics*, Vol. 82, pp. 399–421.
Cinnirella, F., Piopiunik, M. and Winter, J. (2011). 'Why does height matter for educational attainment? evidence from German children', *Economics and Human Biology*, Vol. 9, pp. 407–418.
Cipriani, G. P. and Zago, A. (2011). 'Productivity or discrimination? Beauty and the exams', *Oxford Bulletin of Economics and Statistics*, Vol. 73, pp. 428–447.
Clotfelter, C. T., Ladd, H. F. and Vigdor, J. L. (2010). 'Teacher credentials and student achievement in high school a cross-subject analysis with student fixed effects', *Journal of Human Resources*, Vol. 45, pp. 655–681.
Cole, T., Flegal, K., Nicholls, D. and Jackson, A. (2007). 'Body mass index cut offs to define thinness in children and adolescents: international survey', *British Medical Journal*, Vol. 335, pp. 194–197.
Cole, T. J., Bellizzi, M. C., Flegal, K. M. and Dietz, W. H. (2000). 'Establishing a standard definition for child overweight and obesity worldwide: international survey', *British Medical Journal*, Vol. 320, pp. 1240–1243.
Cornwell, C., Mustard, D. B. and Van Parys, J. (2013). 'Noncognitive skills and the gender disparities in test scores and teacher assessments: evidence from primary school', *Journal of Human Resources*, Vol. 48, pp. 236–264.
Dee, T. S. (2007). 'Teachers and the gender gaps in student achievement', *Journal of Human Resources*, Vol. 42, pp. 528–554.
Dee, T. S. (2005). 'A teacher like me: does race, ethnicity, or gender matter?' *The American Economic Review*, Vol. 95, pp. 158–165.
Di Liberto, A. and Casula, L. (2016). *Teacher Assessments Versus Standardized Tests: Is Acting "Girly" an Advantage*? IZA Discussion Paper Series No 10458. IZA. Bonn, Germany.
Ding, W., Lehrer, S. F., Rosenquist, J. N. and Audrain-McGovern, J. (2009). 'The impact of poor health on academic performance: new evidence using genetic markers', *Journal of Health Economics*, Vol. 28, pp. 578–597.
Feingold, A. (1992). 'Good-looking people are not what we think', *Psychological Bulletin*, Vol. 111, pp. 304–341.
Ferguson, R. F. (2003). 'Teachers' perceptions and expectations and the Black-White test score gap', *Urban Education*, Vol. 38, pp. 460–507.

Finch, B. K. and Beck, A. N. (2011). 'Socio-economic status and z-score standardized height-for-age of US-born children (ages 2–6)', *Economics & Human Biology*, Vol. 9, pp. 272–276.
Gershenson, S., Holt, S. B. and Papageorge, N. (2016). 'Who believes in me? The effect of student-teacher demographic match on teacher expectations', *Economics of Education Review*, Vol. 52, pp. 209–224.
Hamermesh, D. S. and Biddle, J. E. (1994). 'Beauty and the labor market', *The American Economic Review*, Vol. 84, pp. 1174–1194.
Haney, P. and Durlak, J. A. (1998). 'Changing self-esteem in children and adolescents: a meta-analytical review', *Journal of Clinical Child Psychology*, Vol. 27, pp. 423–433.
Harper, B. (2000). 'Beauty, stature and the labour market: a British cohort study', *Oxford Bulletin of Economics and Statistics*, Vol. 62, pp. 771–800.
Hilton, J. L. and von Hippel, W. (1996). 'Stereotypes', *Annual Review of Psychology*, Vol. 47, pp. 237–271.
Hinnerich, B. T., Höglin, E. and Johannesson, M. (2011). 'Are boys discriminated in Swedish high schools?', *Economics of Education Review* , Vol. 30, 682–690.
Jackson, L. A. and Ervin, K. S. (1992). 'Height stereotypes of women and men: the liabilities of shortness for both sexes', *Journal of Social Psychology*, Vol. 132, pp. 433–445.
Judge, T. A. and Cable, D. M. (2004). 'The effect of physical height on workplace success and income: preliminary test of a theoretical model', *Journal of Applied Psychology*, Vol. 89, pp. 428–441.
Jussim, L. and Harber, K. D. (2005). 'Teacher expectations and self-fulfilling prophecies: knowns and unknowns, resolved and unresolved controversies', *Personality and Social Psychology Review*, Vol. 9, pp. 131–155.
Kenney, E. L., Gortmaker, S. L., Davison, K. K. and Austin, S. B. (2015). 'The academic penalty for gaining weight: a longitudinal, change-in-change analysis of BMI and perceived academic ability in middle school students', *International Journal of Obesity*, Vol. 39, pp. 1408–1413.
Lavy, V. (2008). 'Do gender stereotypes reduce girls' or boys' human capital outcomes? Evidence from a natural experiment', *Journal of Public Economics*, Vol. 92, pp. 2083–2105.
Lavy, V. and Sand, E. (2018). 'On the origins of gender gaps in human capital: short-and long-term consequences of teachers' biases', *Journal of Public Economics*, Vol. 167, pp. 263–279.
Lindahl, E. (2007). *Comparing Teachers' Assessments and National Test Results: Evidence from Sweden*. IFAU - Institute for Labour Market Policy Evaluation Working Paper 2007:24. Uppsala, Sweden.
MacCann, C. and Roberts, R. D. (2013). 'Just as smart but not as successful: obese students obtain lower school grades but equivalent test scores to nonobese students', *International Journal of Obesity*, Vol. 37, pp. 40–46.
McConnell, B. and Rasul, I. (2018). 'Racial and ethnic sentencing differentials in the federal criminal justice system', *AEA Papers and Proceedings*, Vol. 108, pp. 241–245.
Mechtenberg, L. (2009). 'Cheap talk in the classroom: how biased grading at school explains gender differences in achievements, career choices and wages', *Review of Economic Studies*, Vol. 76, pp. 1431–1459.
Mobius, M. M. and Rosenblat, T. S. (2006). 'Why beauty matters', *American Economic Review*, Vol. 96, pp. 222–235.
Mocan, N. and Tekin, E. (2010). 'Ugly criminals', *The Review of Economics and Statistics*, Vol. 92, pp. 15–30.
Navas-Sánchez, F. J., Alemán-Gómez, Y. Sánchez-Gonzalez, J., Guzmán-De-Villoria, J. A., Franco, C., Robles, O., Arango, C. and Desco, M. (2014). 'White matter microstructure correlates of mathematical giftedness and intelligence quotient', *Human Brain Mapping*, Vol. 35, pp. 2619–2631.
Ng, M., Fleming, T., Robinson, M., Thomson, B., Graetz, N., Margono, C., Mullany, E. C., Biryukov, S., Abbafati, C., Abera, S. F., Abraham, J. P., *et al.* (2014). 'Global, regional, and national prevalence of overweight and obesity in children and adults during 1980–2013 A systematic analysis for the Global Burden of Disease Study 2013', *The Lancet*, Vol. 384, pp. 766–781.
Nicoletti, C. and Rabe, B. (2018). 'The effect of school spending on student achievement: addressing biases in value-added models', *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, Vol. 181, pp. 487–515.
Oreffice, S. and Quintana-Domeque, C. (2010). 'Anthropometry and socioeconomics among couples: evidence in the United States', *Economics & Human Biology*, Vol. 8, pp. 373–384.
Oreffice, S. and Quintana-Domeque, C. (2016). 'Beauty, body size and wages: evidence from a unique data set', *Economics & Human Biology*, Vol. 22, pp. 24–34.

Oster, E. (2019). 'Unobservable selection and coefficient stability: theory and evidence', *Journal of Business & Economic Statistics*, Vol. 37, pp. 187–204.
Persico, N., Postlewaite, A. and Silverman, D. (2004). 'The effect of adolescent experience on labor market outcomes: the case of height', *Journal of Political Economy*, Vol. 112, pp. 1019–1053.
Puhl, R. and Latner, J. (2007). 'Stigma, obesity, and the health of the nation's children', *Psychological Bulletin*, Vol. 133, pp. 557–580.
Puhl, R. M. and Heuer, C. A. (2009). 'The stigma of obesity: a review and update', *Obesity*, Vol. 17, pp. 941–964.
Queally, M., Doherty, E., Finucane, F. M. and O'Neill, C. (2017). 'Low expectations: do teachers underestimate the ability of overweight children or the children of overweight mothers?, *Economics & Human Biology*, Vol. 27, pp. 26–32.
Riegle-Crumb, C. and Humphries, M. (2012). 'Exploring bias in math teachers' perceptions of students' ability by gender and race/ethnicity', *Gender & Society*, Vol. 26, pp. 290–322.
Rooth, D. (2009). 'Obesity, attractiveness, and differential treatment in hiring: a field experiment', *The Journal of Human Resources*, Vol. 44, pp. 710–735.
Rosenthal, R. and Jacobson, L. (1968). 'Pygmalion in the classroom', *The Urban Review*, Vol. 3, pp. 16–20.
Ryan, C. (2017). '*Secondary School Teacher Effects on Student Achievement in Australian Schools*'. Melbourne Institute Working Paper No. 11/17. Melbourne. Australia.
Sabia, J. J. and Rees, D. I. (2015). 'Body weight, mental health capital, and academic achievement', *Review of Economics of the Household*, Vol. 13, pp. 653–684.
Sargent, J. D. and Blanchflower, D. G. (1994). 'Obesity and stature in adolescence and earnings in young adulthood: analysis of a British birth cohort', *Archives of Pediatrics & Adolescent Medicine*, Vol. 148, pp. 681–687.
Sass, T. R., Semykina, A. and Harris, D. N. (2014). 'Value-added models and the measurement of teacher productivity', *Economics of Education Review*, Vol. 38, pp. 9–23.
Schick, A. and Steckel, R. H. (2015). 'Height, human capital, and earnings: the contributions of cognitive and noncognitive ability', *Journal of Human Capital*, Vol. 9, pp. 94–115.
Scholder, S. V. H. K., Smith, G. D., Lawlor, D. A., Propper, C. and Windmeijer, F. (2013). 'Child height, health and human capital: evidence using genetic markers', *European Economic Review*, Vol. 57, pp. 1–22.
Shrewsbury, V. and Wardle, J. (2008). 'Socioeconomic status and adiposity in childhood: a systematic review of cross-sectional studies 1990–2005', *Obesity*, Vol. 16, pp. 275–284.
Slater, H., Davies, N. M. and Burgess, S. (2012). 'Do teachers matter? Measuring the variation in teacher effectiveness in England', *Oxford Bulletin of Economics and Statistics*, Vol. 74, pp. 629–645.
Soloff, C., Lawrence, D. and Johnstone, R. (2005). *LSAC Technical Paper Number 1: Sample Design*, Melbourne, Australian Institute of Family Studies.
Stanley, D. A., Sokol-Hessner, P., Banaji, M. R. and Phelps, E. A. (2011). 'Implicit race attitudes predict trustworthiness judgments and economic trust decisions', *Proceedings of the National Academy of Sciences*, Vol. 108, pp. 7710–7715.
Terrier, C. (2016). *Boys Lag Behind: How Teachers' Gender Biases Affect Student Achievement*. IZA Discussion Paper Series No 10343. IZA. Bonn, Germany.
Todd, P. E. and Wolpin, K. I. (2003). 'On the specification and estimation of the production function for cognitive achievement', *The Economic Journal*, Vol. 113, pp. F3–F33.
Todd, P. E. and Wolpin, K. I. (2007). 'The production of cognitive achievement in children: home, school, and racial test score gaps', *Journal of Human capital*, Vol. 1, pp. 91–136.
Zavodny, M. (2013). 'Does weight affect children's test scores and teacher assessments differently?' *Economics of Education Review*, Vol. 34, pp. 135–145.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Online Appendix.