Delaney, Judith M.; Devereux, Paul J.

Working Paper

Gender Differences in Teacher Judgement of Comparative Advantage

IZA Discussion Papers, No. 16635

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: Delaney, Judith M.; Devereux, Paul J. (2023) : Gender Differences in Teacher Judgement of Comparative Advantage, IZA Discussion Papers, No. 16635, Institute of Labor Economics (IZA), Bonn

This Version is available at:
https://hdl.handle.net/10419/282762

I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 16635

Gender Differences in Teacher Judgement of Comparative Advantage

Judith M. Delaney
Paul J. Devereux

NOVEMBER 2023

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 16635

Gender Differences in Teacher Judgement of Comparative Advantage

Judith M. Delaney
University of Bath, University College London and IZA

Paul J. Devereux
School of Economics and Geary Institute, University College Dublin CEPR, IZA, and NHH

NOVEMBER 2023

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Gender Differences in Teacher Judgement of Comparative Advantage

Much research shows that students take account of their perceived comparative advantage in mathematics relative to verbal skills when choosing college majors and career tracks. There is also evidence for an important role for comparative advantage in explaining the gender gap in college STEM major choice. For these reasons, it is important to understand why student perceptions of comparative advantage may differ from true comparative advantage as determined by actual abilities. One plausible pathway is through teachers. We study gender differences in teacher evaluations of student comparative advantage relative to comparative advantage as measured by test scores. We show that findings are very sensitive to the methods used; commonly used methods are not equivalent and can give different results as they target different estimands. Using two recent UK cohort surveys, we show that these conceptual issues matter in practice when we evaluate whether teachers are likely to over-estimate female comparative advantage in English relative to mathematics. Our preferred estimates provide no evidence that teachers exaggerate the female advantage in English relative to mathematics and generally suggest the opposite. We conclude that differences in teacher judgement by gender do not provide another reason for the gender gap in STEM.

JEL Classification: J16, I21, I23

Keywords: teacher bias, gender gaps, STEM, comparative advantage, math and English skills

Corresponding author:
Judith M. Delaney
University of Bath
Claverton Down
Bath
BA2 7AY
Great Britain
E-mail: judithmdelaney@gmail.com

1. Introduction and Motivation

Women are underrepresented in Science, Technology, Engineering, and Mathematics (STEM) jobs and occupations. This gender gap in STEM has important implications for society and the economy as it is widely agreed that having an adequate supply of STEM graduates is important for both innovation and economic productivity (Peri et al. 2015). In addition, given the typically larger earnings of STEM workers, having more females working in STEM jobs may help to decrease the gender gap in earnings (Card and Payne, 2021).

Much research has shown that several factors may explain the gender gap in STEM at tertiary level including personal attributes such as self-confidence, competitiveness, and preferences as well as external factors such as peers and role models (McNally, 2020). There is also evidence of an important role for comparative advantage in explaining this gap.[1] Standardized international tests tend to find a gender difference in mathematics favouring boys (often small, and non-existent in some countries), while boys tend to score significantly worse than girls in reading (Borgonovi et al., 2018).[2] However, decisions will likely depend on how students perceive their comparative advantage and teachers may play an important role in affecting how students think about their own abilities.[3] Additionally, teacher views of student comparative advantage may matter if teachers are involved in the college admissions process as letter writers or if teacher grade evaluations or predicted grades are used to determine college

[1] Papers that explore this issue include Aucejo and James (2021), Delaney and Devereux (2019; 2021), Goulas et al. (2022), Saltiel (2022), Shi (2018) and Speer (2017; 2023).
[2] See Cavaglia et al. (2020) and Delaney and Devereux (2021) for a review of the literature on gender gaps in educational achievement.
[3] There is much evidence that teacher opinions and assessments affect student beliefs about their abilities. Terrier (2020) shows that girls who benefit from teacher favouritism are more likely to select a science track in high school. Carlana (2019) shows that teacher gender stereotypes can cause girls to underperform in mathematics. Lavy and Sand (2018) show that grading biases can have long lasting impacts on academic achievement and course taking in high school. Also, there is evidence that students' own beliefs about their capabilities are influenced in various ways by their teachers (Dee 2015; Gershenson, Holt, and Papageorge 2016).

admission.[4] We study whether, relative to test scores, teachers tend to overestimate male comparative advantage in mathematics and female comparative advantage in verbal skills.

While much literature is concerned with estimating various types of teacher biases, we show conceptually that there is no single "teacher bias" effect. When researchers have access to teacher assessments (TA) as well as test scores (TS), there are two broad approaches taken to study gender bias. The first approach is to regress TA – TS on gender. The second approach is to regress TA on TS and on gender. While these methods often appear to be used interchangeably, we show that they are not equivalent and can give different results as they target different estimands. Even if teachers have no gender bias, teacher evaluation differences by gender may systematically deviate from test score differences if the distribution of test scores differs between boys and girls. Therefore, we conclude that it is important for researchers to be clear about the target estimand.

Using two recent UK cohort surveys, we show that these conceptual issues matter in practice when we evaluate whether teachers are likely to over-estimate female comparative advantage in English relative to mathematics. In our application, the gender effects differ substantially depending on the method used and the signs even switch across specifications. Our preferred estimates provide no evidence that teachers exaggerate the female advantage in English relative to mathematics and generally suggest the opposite. An implication is that it is unlikely that teacher misperceptions of comparative advantage by gender are an important cause of the gender gap in STEM.

The remainder of the paper proceeds as follows: In section 2 we discuss the two main specifications that have been used in the literature and how they differ conceptually. Section 3 outlines the empirical strategy, section 4 discusses the institutional background, and, in section

[4] It has also been shown that teachers have a substantial impact on students both in the short-term while in school but also affecting outcomes later in life. (Chetty et al. 2014a, 2014b).

5, we describe the two surveys we use in the analysis. Our main empirical results are presented in section 6 with robustness checks in section 7. Section 8 concludes.

2. The Two Approaches

As mentioned earlier, researchers tend to take one of two broad approaches to measuring teacher bias, either regressing TA – TS on female or regressing TA on TS and female. The first approach has been taken by Reeves et al. (2001), Lindahl (2006), Cornwell et al. (2013), Falch and Naper (2013), and Gibbons and Chevalier (2008). Reeves et al. (2001) used UK Key Stage Two (KS2) data from 1996-1998 and found small effects with teachers tending to underrate (relative to the test score) males in mathematics and females in English. Gibbons and Chevalier (2008) use UK administrative data to regress the difference between Key Stage Three (KS3) test scores and teacher evaluations on a range of student characteristics including gender. They find that "Boys, compared to girls, do relatively well on teacher assessments in English, but relatively poorly in mathematics and science". Lindahl (2007) uses Swedish data to compare national tests with teacher assessments at age 16 and finds that females do much better on teacher assessments in Swedish, English, and math. Using Norwegian data at the end of compulsory education, Falch and Naper (2013) find that teacher grades favour girls in both languages and mathematics when compared to anonymously evaluated central exit exams. Cornwell et al. (2013) show using US data that, amongst students with the same test scores, boys are assessed lower by teachers in reading, math, and science and this disparity can be largely explained by non-cognitive skills.[5]

[5] Cornwell et al. (2013) implement this by separately regressing TA and TS on gender. They have continuous measures and standardize both TA and TS to have mean 0 and standard deviation of 1.

The second approach of regressing TA on TS and female has been taken by Cornwell et al. (2013), Campbell (2015), and McCoy et al. (2021).[6] As mentioned above, Cornwell et al. regress TA – TS on gender but they take the approach of regressing TA on TS and gender when they also add a control for non-cognitive skills. Using MCS data at age 7, Campbell (2015) studies the probability of being judged above average at maths and reading, conditional on the test score, and finds a negative effect for girls in mathematics and a positive effect for girls in reading. McCoy et al. (2021) focus on mathematics and, using a logit model for teacher assessments, also find a negative effect for girls.[7]

Conceptual Framework

In this section, we set some key ideas based on the general case where teachers are providing information about a characteristic of the student, which we label as ability. Define the true ability of the student as *A*, let *TA* be the teacher assessment of the ability of the student, and *F* denote a binary indicator for whether the student is female. We assume that teachers receive an unbiased estimate of *A*, which we call *T* so that $T = A + v$, where $E(v) = 0$, $cov(A, v) = 0$, and $E(v|F = 0) = E(v|F = 1)$.[8] We assume that teachers make their assessment as

[6] This type of approach has also been used to study teacher bias in dimensions other than gender. See Burgess and Greaves (2013) for differences by ethnic group in the UK, Botelho et al. (2015) for differences by race in Brazil, and Alesina et al. (2018) for differences between immigrants and other students.

[7] More broadly, there are several studies of the effects of anonymous grading of assessments. For example, Burgess et al. (2022) provide evidence from Denmark that, while girls are largely unaffected, boys are more likely to get higher Danish scores and lower math scores in oral exams when the examiner is an external person rather than their teacher. Similarly, Breda and Ly (2015) use non-anonymous oral and anonymous written tests in an entrance exam to an elite higher education institution in France and find males are biased against in male-dominated subjects such as math and females are biased against in female-dominated subjects such as literature and biology. Lavy (2008) compares scores on blind and non-blind exams for students in their senior year in high school in Israel and finds that boys are discriminated against in mathematics and English as well as several other subjects in the sciences and humanities. A disadvantage of using blind and non-blind scores is that the type of exam may vary across assessments and that students may perform differently across such assessments. Hinnerich et al. (2011) overcome this issue by using a field experiment that randomly assigns blind and non-blind scores to the same compulsory high school Swedish exam. They find no evidence of grading biases by gender.

[8] The signal received by the teacher could be gender-biased but, for simplicity, we ignore this possibility as, in practice, it is indistinguishable from $\beta \neq 0$ in equation (1).

$$TA = \alpha + \delta T + \beta F + \epsilon, \quad (1)$$

where *F* denotes a binary indicator for whether the student is female, *T* is the teacher perception of the true ability of the student, and ϵ is an idiosyncratic error term.

We can evaluate whether teacher behaviour leads to systematic disparities between *TA* and *A* that differ by gender:

$$E(TA - A|F = 0) = E(\alpha + \delta T + \beta F + \epsilon - A|F = 0)$$

$$= E(\alpha + (\delta - 1)A + \delta v + \beta F + \epsilon|F = 0)$$

$$= \alpha + (\delta - 1)E(A|F = 0)$$

Likewise, $E(TA - A|F = 1) = \alpha + (\delta - 1)E(A|F = 1) + \beta$. Therefore,

$$E(TA - A|F = 1) - E(TA - A|F = 0) = (\delta - 1)\{E(A|F = 1) - E(A|F = 0)\} + \beta$$

This makes clear that there are two types of teacher behaviour that can lead to systematic disparities between *TA* and *A* by gender. First, there is explicit gender bias if $\beta \neq 0$. Second, even if $\beta = 0$, teachers may choose a value of δ less than 1. This implies that teacher assessments regress towards the mean compared to the signal of ability they receive; this will tend to disadvantage students towards the top of the ability distribution compared to those towards the bottom.[9] If girls have higher ability on average than boys, then a value of δ less than 1 leads $TA - A$ to be smaller for girls than for boys. To give an extreme example of this, if teachers give all students the same *TA*, then there will trivially be no gender gap in *TA* at any particular value of *A* or *T*. However, if girls typically have higher values of *A* than boys,

[9] Regression towards the mean has been found in teacher assessments with students with higher test scores being more likely to have teacher assessments that are under-rated relative to the test score than students with lower test scores (Burgess and Greaves, 2013; Gibbons and Chevalier, 2008).

$E(TA - A|F = 1) - E(TA - A|F = 0)$ will be negative. This issue would not arise if the distribution of ability was the same for boys and girls.

This simple analysis makes clear the importance of defining the estimand of choice, whether it is $E(TA - A|F = 1) - E(TA - A|F = 0)$ or whether it is the structural gender bias parameter, β. If the focus is on whether teachers are gender-biased, then obtaining a consistent estimate of β should be the priority. For the more general question about whether there are systematic gender gaps in teacher assessment relative to ability then $E(TA - A|F = 1) - E(TA - A|F = 0)$ would be the more desired choice.[10]

Estimation

The discussion so far has been about the relationship between *TA*, the teacher assessment, and *A*, the true level of ability. However, *A* is unobserved, and, in its place, we have a test score measure. Define *TS* as the score of the student in the test. We assume that *TS* is an unbiased estimator of *A* so that $TS = A + u$, where $E(u) = 0$, $cov(A, u) = 0$, and $E(u|F = 0) = E(u|F = 1)$.

Measuring Gender Disparities

We can estimate $E(TA - A|F = 1) - E(TA - A|F = 0)$ using the observed *TS* rather than the unobserved *A*.

$$E(TA - TS|F = 1) - E(TA - TS|F = 0)$$

$$= E(TA - A|F = 1) - E(TA - A|F = 0) - \{E(u|F = 1) - E(u|F = 0)\}$$

$$= E(TA - A|F = 1) - E(TA - A|F = 0).$$

[10] The supplementary material to Breda and Hillion (2016) considers similar issues in the context of an oral and written exam in which the skills required for one may differ from those required for the other. Unlike us, they treat the structural gender bias parameter, β as the only estimand of interest.

Note that the assumptions on u are not innocuous here. If the test score is gender-biased in some unknown fashion, there is no way of evaluating whether the teacher assessment is gender-biased. Likewise, if $cov(A, u) = 0$ does not hold so that measurement error in *TS* is different at different parts of the distribution, there is no way of evaluating whether the teacher assessment is gender-biased if girls and boys are located at different parts of the ability distribution.

In practice, we implement this approach by regressing *TA* – *TS* on an indicator for female. This approach determines whether, on average, the gender gap in teacher assessments matches up with the gender gap in test scores. Importantly, this does not imply anything about the value of β – regressing TA – TS on female is not informative about whether teachers are actually biased (in a direct sense) as the estimand is affected by mean regression of teacher judgements and the distribution of test scores of boys and girls. For this approach to be reliable, it is important that *TA* and *TS* are measured in comparable units so that a one unit increase in *TA* is equivalent to a one unit increase in *TS*. We use various standardization approaches to ensure this.

Estimating Structural Teacher Bias (β)

Consider now, regressing *TA* on *TS* and female in order to estimate β. Using this approach, we consider that there is no teacher bias if, on average, for each particular value of the test score, the average teacher assessment is the same for girls and boys.

Measurement error in *TS* causes real problems as given $TS = A + u$ and $T = A + v$, $T = TS - u + v$. When we replace *T* with *TS* in equation (1), the positive covariance between *TS* and u implies that the OLS estimator of $\delta, \hat{\delta}$, is inconsistent and biased towards zero. If *TS* is correlated with gender, then $\hat{\beta}$ will also be inconsistent with an upward bias if girls have higher average test scores than boys and a downward bias if they have lower average test scores

than boys. One approach to this problem is to use instrumental variables and researchers have used lagged values of *TS* as instruments for *TS*.[11] If the instrument is valid, this provides consistent estimates of δ and β.[12]

Note that, even with a consistent estimator, if δ is less than 1 and $E(TS|F=1) \neq E(TS|F=0)$, $\hat{\beta} \neq E(TA-TS|F=1) - E(TA-TS|F=0)$. Therefore, measurement error is not the only reason why the two approaches (regressing *TA – TS* on gender versus regressing *TA* on *TS* and gender) may give different findings for the effect of gender – the estimands are different across the two approaches. However, the two methods give the same gender estimate if $\hat{\delta} = 1$.

3. Empirical Methodology

Using percentile ranks, we create measures of comparative advantage (CA) for each student in each year based on their exam scores and teacher evaluations at a given age. Ranks are elements of [0,1] and are calculated based on the entire sample. We break ties using the mean rank of all students who have the same test score or evaluation.[13] We define the test score measure of CA (the "*TS*") as rank of English test score minus rank of mathematics test score; equivalently, we define the teacher evaluation measure of CA (the "*TA*") as rank of English teacher evaluation minus rank of mathematics teacher evaluation.

While we are specifically interested in comparative advantage, there is also a methodological advantage to studying CA rather than the individual English and mathematics grades and evaluations that underly it. Students may have fixed characteristics that influence

[11] Cornwall et al. (2013), Botelho et al. (2015), Terrier (2020), and Ferman and Fontes (2022) have used lagged test scores to instrument for current test scores.
[12] The IV approach also may help to reduce bias from reverse causality if teacher evaluations proxy for teacher attitudes towards the child and these teacher attitudes affect test scores.
[13] In our data, teacher evaluations generally range from 1 to 8 and so there are many ties. We find that our estimates are similar if we assign ties either the highest rank or the lowest rank.

their test scores in both subjects; for example, a student may be a poor exam taker in standardized exams. Likewise, non-academic student-specific factors such as behaviour in class may influence teacher evaluations. So long as these student characteristics are common to both English and mathematics, calculating the difference between the English and the mathematics scores eliminates these sources of bias. At primary school level, in which students have the same teacher for English and mathematics, this approach also accounts for fixed teacher characteristics such as if some teachers place more weight on non-cognitive skills in teacher evaluations.

Specification 1: Estimating the Structural Teacher Bias Parameter

This specification regresses the teacher measure on gender, controlling for the exam-based measure. The basic specification has the form:

$$TA = \beta_0 + \beta_1 Female + \delta TS + u. \qquad (2)$$

Female is a binary variable denoting whether the student is female or not.[14] In addition to allowing *TS* to enter linearly, we also show estimates where it enters much more flexibly using polynomials, fixed effects, and having English and mathematics test scores enter the equation separately. We also show specifications where, to account for measurement error in *TS*, we use lagged test scores as an instrument for current student test scores.

Specification 2: Estimating Differences in the Gap between Teacher Assessments and Ability

In this specification we are interested in whether the gap between *TA* and *TS* differs systematically by gender and so we regress the difference between the teacher measure and the exam measure on gender. This basic specification has the form:

[14] We have also estimated specifications whereby we include controls for whether the student is white, and quintiles of the family income distribution and the results were very similar, but the sample size falls considerably due to missing values in these variables. Since we do not expect these characteristics to vary by gender and to increase sample size, we have decided to omit them.

$$TA - TS = \alpha_0 + \alpha_1 Female + e. \quad (3)$$

4. Institutional Setting

The national curriculum in England and Wales has 5 Key Stages (KS), approximately corresponding to ages 7, 11, 14, 16 and 18. The National Curriculum sets levels of achievement that students are expected to meet at various stages of their educational trajectory. Nationally set exams enable teachers, parents, and students to understand how well they are performing relative to what is expected at their age. At the end of KS 1, there are exams in English and math.[15] At the end of KS 2 there are exams in English and math and at the end of KS 3 there are exams in English, math and science.[16] At the end of KS 4 students take the General Certificate of Secondary Education (GCSE) exams and at the end of KS 5 students take their A-level exams. The key stage exams are standardised and taken at the end of the school year and the exams are graded externally by an anonymous marker. We use the KS 2 and KS 3 exams that are taken at approximately ages 11 and 14 as our measures of students test scores. One potential issue is that the KS exam scripts include the name of the student. In theory, graders could be influenced by whether it is a male or female name, but Baird (1998), Hanna and Linden (2012), and Chowdhury et al. (2020) find no evidence for this type of grader behaviour.

In addition to the key stage exams, at the end of KS 1, 2 and 3, teachers are asked to provide an assessment of whether each student is meeting the learning objectives as set out in the national curriculum. There is a lot of emphasis placed on teacher assessments as they are

[15] The KS1 exams are marked internally by the school and schools are not required to report the test results for KS 1 so we do not have information on the test scores at age 7; schools are required to report teacher assessments at age 7 and these are contained in the dataset. As of the academic year 2023, the KS 1 tests are no longer compulsory.
[16] The KS 3 exams were abolished in 2008.

statutory and need to be reported to the Department for Education. Teachers are given guidance and a framework for how to best assess the students' performance and are asked to inform their assessment based on interactions with students throughout the school year including performance on in-school tests. Teachers must provide demonstrable and reliable evidence to support their evaluation to ensure that the assessments are objective and comparable across schools. There are 8 response options ranging from "National curriculum Level 1 achieved" up to "National curriculum level 8 achieved". It is expected that the typical 11-year-old student would achieve level 4. We use the answers to the teacher assessments at KS 2 and 3 that are evaluating students at ages 11 and 14 respectively, as our measure of teacher evaluations.

5. Data

We use two different cohort studies that contain the requisite information on test scores and teacher evaluations. These allow us to study students at different ages and in different years as it is important to show whether the findings hold across multiple cohorts and how any gender gaps evolve with age. In each survey (Millennium Cohort Study (MCS) and Next Steps (NS)), we study children at the ages for which we have both exam grades in English and mathematics as well as teacher evaluations of their abilities in these subjects. The ages that fulfil these requirements are ages 7 and 11 in the MCS and ages 11 and 14 in Next Steps.[17]

Millennium Cohort Study

The Millennium Cohort Study (MCS) follows the lives of approximately 19,000 young people born in the UK between 2000 and 2002. To date, participants have been surveyed at various stages across the life cycle including at 9 months, and at ages 3, 5, 7, 11, 14, 17 and 22. The age 11 survey was carried out in 2012/13 and includes information on around 13,500

[17] For more information about the UK cohort studies and the linked administrative data, see University College London et al. (2020, 2021).

cohort members. The survey contains rich information on teacher evaluations of student ability in various subjects and information on teacher characteristics. In addition, the MCS dataset has been linked to the National Pupil Database (NPD) which contains information on KS 2 exam test scores and KS 2 teacher evaluations at age 11.

The NPD data are only available for students in English state schools and so our sample is limited to students in English state schools with non-missing data on KS exams and teacher evaluations. The MCS also includes cognitive tests in math and English at age 7 that were administered to the children during a home visit as part of the survey – we will later use these to create an instrumental variable for comparative advantage at age 11.[18]

Next Steps Cohort Dataset

Next Steps (NS), previously known as the Longitudinal Study of Young People in England, follows the lives of around 16,000 people in England born in 1989-90. Similar to the MCS, these data have been linked to the NPD records and so we have information on Key Stage exams at ages 11 and 14 in mathematics and English as well as Key Stage teacher evaluations at these ages.[19]

Descriptive Findings

Table 1 displays the summary statistics for both datasets. It is clear that, on average, females outperform males in English while males score higher in math and this pattern holds across all ages and cohorts. Similarly, teachers, on average, assign a higher evaluation to females in English and to males in math. Our *TS* measure defined as rank of the student KS

[18] The English test used was the Word Reading test which is the verbal skills subscale of the British Ability Scales. This test is used to elicit childrens' skills in reading. The math test comprised a shortened version of the Progress in Mathematics (PiM) test. The PiM test assesses the child's skills in math content based on the UK National Curricula.

[19] The KS3 exams were abolished in 2008 and so we do not have KS exam and teacher assessments at age 14 for the MCS cohort.

test score in English minus rank of student KS test score in math is positive for females and negative for males showing that, on average, girls do better in English KS tests relative to math and conversely for boys. We also see that the teacher assessed comparative advantage (*TA*) defined as rank of the student's English teacher evaluation minus rank of the student's mathematics teacher evaluation is higher for females than males. In addition, the average difference in *TA – TS* is slightly larger for males than females.

Table 1: Descriptive Analysis for MCS and NS Datasets

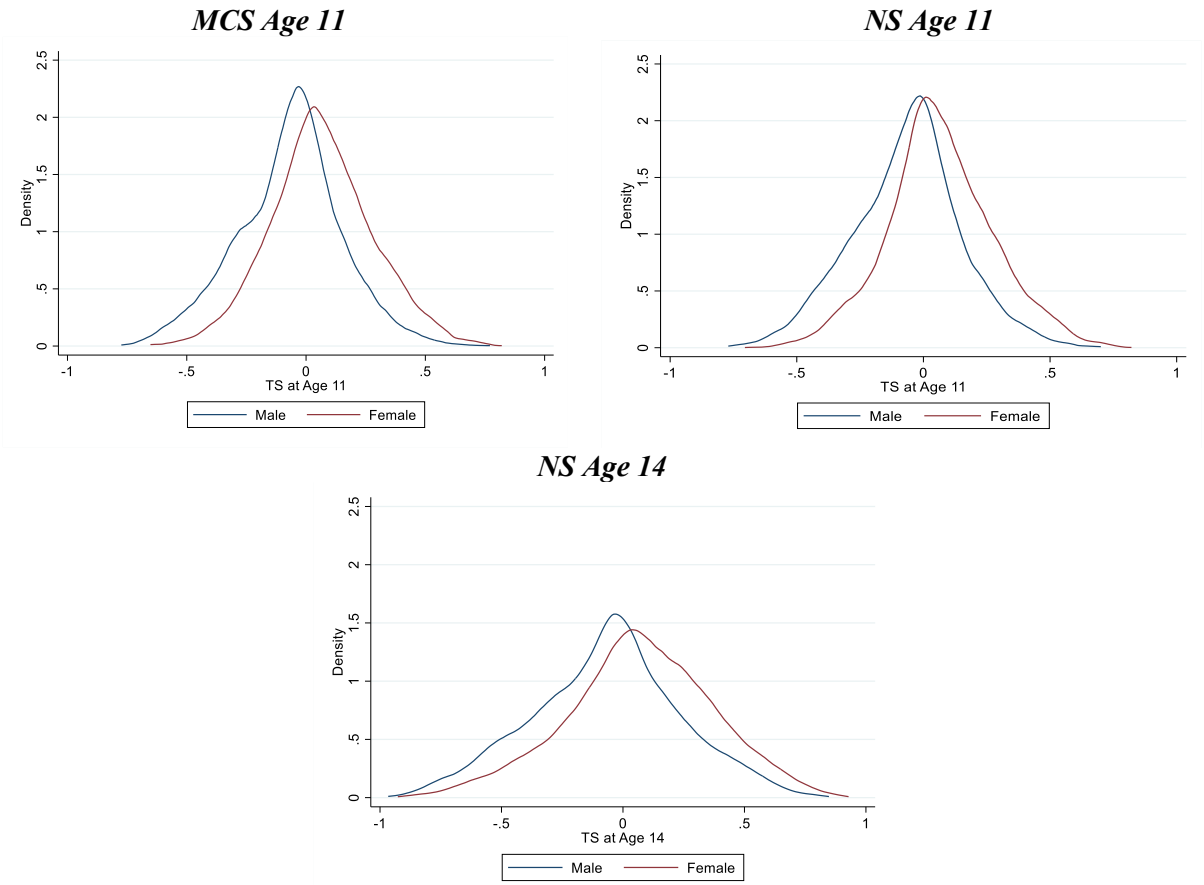
	Male		Female		Difference in Means
	Mean	SD	Mean	SD	
	Millennium Cohort Study Age 11				
Teacher Female Age 11	0.75	(0.43)	0.76	(0.43)	-0.01
English Test Score Age 11	73.79	(14.31)	77.70	(13.65)	-3.91***
Math Test Score Age 11	75.43	(18.77)	72.31	(18.55)	3.12***
Rank English Test Score Age 11	0.46	(0.29)	0.54	(0.29)	-0.08***
Rank Math Test Score Age 11	0.53	(0.29)	0.47	(0.28)	0.06***
TS at Age 11	-0.07	(0.22)	0.07	(0.22)	-0.13***
English Teacher Evaluation Age 11	4.32	(0.63)	4.47	(0.62)	-0.15***
Math Teacher Evaluation Age 11	4.49	(0.74)	4.40	(0.69)	0.10***
Rank English Teacher Evaluation Age 11	0.47	(0.26)	0.53	(0.26)	-0.06***
Rank Math Teacher Evaluation Age 11	0.52	(0.27)	0.48	(0.26)	0.04***
TA at Age 11	-0.05	(0.22)	0.05	(0.23)	-0.10***
TA-TS Age 11	0.02	(0.22)	-0.02	(0.21)	0.04***
Observations	*2,252*		*2,391*		
	Next Steps Age 11 and 14				
English Test Score Age 11	58.39	(13.53)	62.42	(13.00)	-4.03***
Math Test Score Age 11	64.70	(20.39)	61.50	(19.90)	3.20***
Rank English Test Score Age 11	0.46	(0.28)	0.54	(0.29)	-0.09***
Rank Math Test Score Age 11	0.52	(0.29)	0.48	(0.28)	0.05***
TS at Age 11	-0.07	(0.22)	0.07	(0.21)	-0.13***
English Teacher Evaluation Age 11	3.90	(0.73)	4.07	(0.71)	-0.17***
Math Teacher Evaluation Age 11	4.08	(0.76)	4.00	(0.73)	0.08***
Rank English Teacher Evaluation Age 11	0.47	(0.26)	0.53	(0.26)	-0.06***
Rank Math Teacher Evaluation Age 11	0.51	(0.27)	0.49	(0.26)	0.03***
TA Age 11	-0.04	(0.21)	0.04	(0.20)	-0.09***
(TA-TS) Age 11	0.02	(0.22)	-0.02	(0.22)	0.05***
English Test Score Age 14	42.38	(16.47)	47.96	(15.89)	-5.58***
Math Test Score Age 14	79.65	(20.69)	76.68	(20.63)	2.97***
Rank English Test Score Age 14	0.45	(0.29)	0.55	(0.28)	-0.10***
Rank Math Test Score Age 14	0.52	(0.29)	0.48	(0.29)	0.04***
TS at Age 14	-0.07	(0.31)	0.07	(0.30)	-0.14***
English Teacher Evaluation Age 14	5.01	(0.99)	5.33	(0.94)	-0.32***
Math Teacher Evaluation Age 14	5.62	(1.26)	5.57	(1.19)	0.05*
Rank English Teacher Evaluation Age 14	0.46	(0.28)	0.54	(0.27)	-0.09***
Rank Math Teacher Evaluation Age 14	0.51	(0.29)	0.49	(0.27)	0.01*
TA Age 14	-0.05	(0.23)	0.05	(0.22)	-0.10***
(TA-TS) Age 14	0.02	(0.36)	-0.02	(0.36)	0.03***
Observations	*6,097*		*6,076*		

Note: TS is measured for each student as the percentile rank of their English test score minus the rank of their mathematics test score. TA is measured for each student as the percentile rank of their teacher evaluation in English minus the rank of their teacher evaluation in mathematics. The last column shows the t-statistic for the difference in means. *** p<0.01, * p<0.1.

Figure 1 shows the distribution of *TS* by gender for all cohorts and ages. It is clear that at every age and for each cohort, the *TS* distribution is shifted to the right for females. This implies that, on average, females are doing relatively better at English than maths as compared

to boys who tend to score higher in math than English. Given that the distribution of *TS* is not the same across gender, the two approaches may lead to different estimates of teacher biases. We discuss this more in the results section.

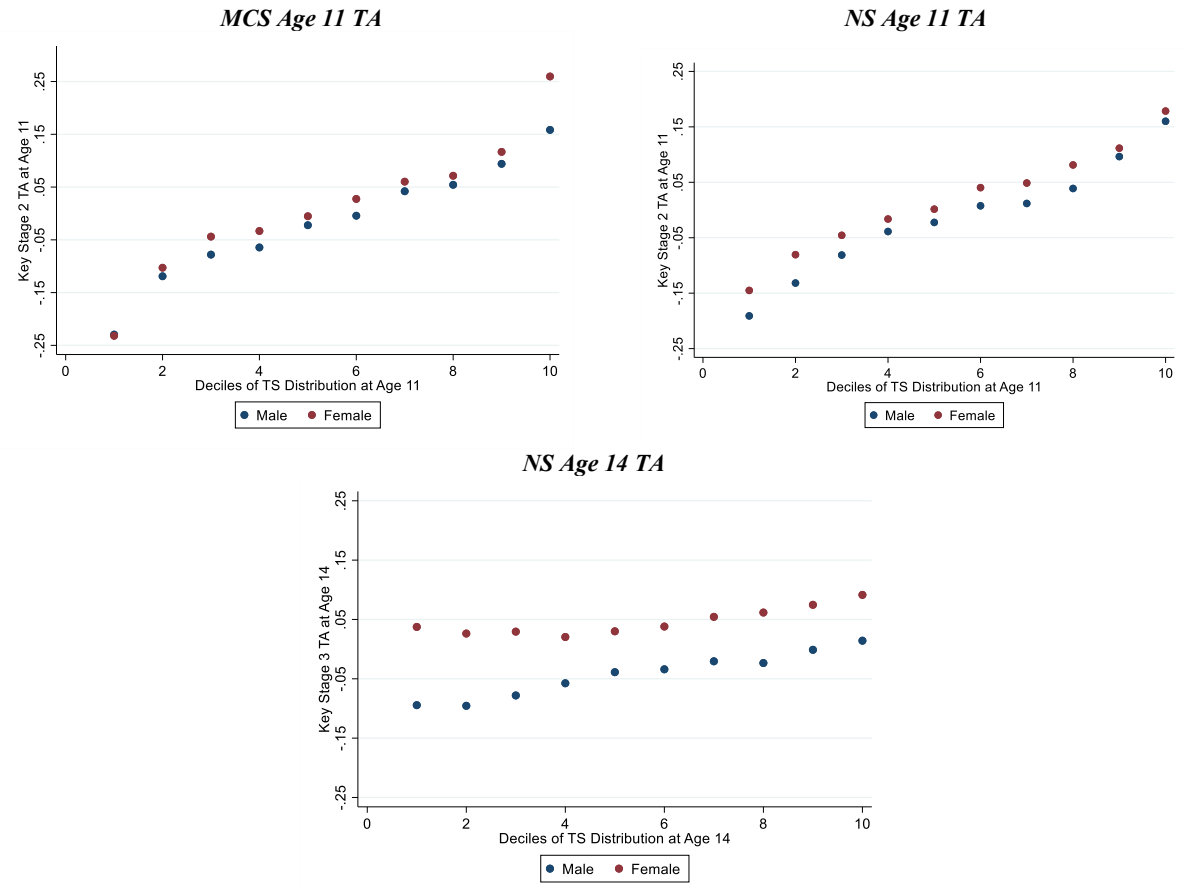
Figure 1: Distribution of TS by Gender



Note: *TS* is measured for each student as the percentile rank of their English test score minus the rank of their mathematics test score.

Figure 2 shows how *TA* varies across the *TS* distribution. As expected, there is a positive relationship such that students with a higher *TS* have a higher *TA*. However, across all cohorts and ages, at each decile of the *TS* distribution, *TA* is larger for females than males. The gender gap in *TA* is largest for the Next Steps cohort at age 14; the gender difference in TA conditional on *TS* decile is much smaller for the other cohorts and ages. This suggests that any specification that includes *TS* as a control in the regression may lead to results that imply that teachers' judgement of *CA* is in favour of females.

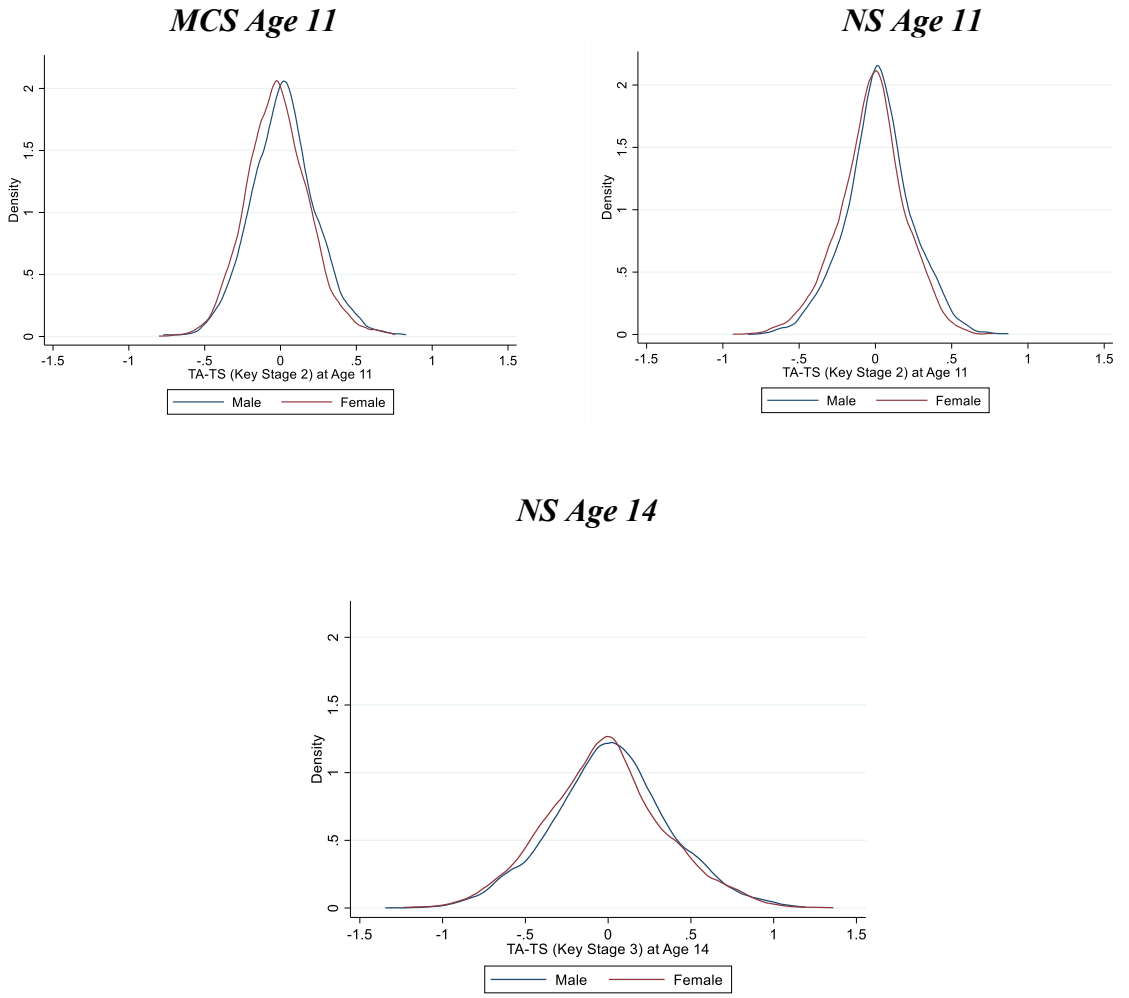
Figure 2: Distribution of TA across Deciles of TS Distribution by Gender



Note: *TS* is measured for each student as the percentile rank of their English test score minus the rank of their mathematics test score. *TA* is measured for each student as the percentile rank of their teacher evaluation in English minus the rank of their teacher evaluation in mathematics.

Figure 3 shows the distribution of *TA-TS* by gender for all cohorts and ages. The *TA-TS* distribution is very similar for both genders. However, at age 11, the *TA-TS* distribution is slightly shifted to the right for males suggesting that the specification that has *TA-TS* as a dependent variable may imply teacher judgement of *CA* to be biased in favour of males.

Figure 3: Distribution of TA-TS by Gender



Note: *TS* is measured for each student as the percentile rank of their English test score minus the rank of their mathematics test score. *TA* is measured for each student as the percentile rank of their teacher evaluation in English minus the rank of their teacher evaluation in mathematics.

6. Results

In this section, we discuss the estimates from the two main specifications that have been used in the literature.

Specification 1

The first panel of Table 1 shows the estimates where the dependent variable is teacher assessed comparative advantage, *TA*, and the student test score comparative advantage, *TS*, is

controlled for in the regression (equation (2)). As discussed earlier, if consistently estimated, the coefficient on *FEMALE* from this specification will capture explicit teacher bias. The results show that, conditional on student test score comparative advantage, teachers are more likely to systematically judge females as having a comparative advantage in English relative to math compared to boys. This finding is evident across both surveys and across different ages. The magnitude of the estimates ranges from 11% of a standard deviation in teacher assessed comparative advantage at age 11 in the MCS, 14% of a standard deviation at age 11 in NS, to almost 40% of a standard deviation in teacher assessed comparative advantage at age 14 in NS.

Instrumental Variable Analysis

The administrative KS tests at age 11 and 14 provide rich measures of student achievement on a given day. Nonetheless, there may be measurement error in the scores and, as discussed above, this could lead to inconsistent estimates of both the female and *TS* parameters, and therefore, in this section, we instrument our current test scores with lagged test scores. Due to data limitations, we can only do this for the age 11 survey in the MCS and age 14 survey in NS. We use the cognitive tests in math and English administered at age 7 in the MCS and form a measure of *TS* at age 7 that is equal to the rank in the English test score at age 7 minus the rank in the math test score at age 7 and use this measure to instrument for *TS* at age 11. Similarly, we use *TS* at age 11 in NS as an instrument for *TS* at age 14. The first stage estimates reported in Table A1 in the appendix show that lagged *TS* is very strongly predictive of current *TS* and the F-statistic is very large across all specifications.

The second panel of Table 2 shows the estimates when we use lagged *TS* as an instrument for current *TS*. As expected, the coefficient on *TS* is now much larger across all ages and cohorts and is just above 1 for the age 14 estimates. Similarly, we expected that the coefficient on female in the analysis that did not account for measurement error would be

upward biased if females on average had a higher *TS* than boys and again this is what we find. These female estimates are very different to the OLS estimates from Specification 1 which were all positive and statistically significant. The estimates obtained by instrumenting the current TS measure suggest that teachers do not misjudge student comparative advantage by gender at age 11 in the MCS study. However, at age 14 in the NS cohort, the coefficient is now -0.046 which implies that teachers are more likely to judge boys as being relatively better at English than math as compared to girls.

Specification 2

The last panel of Table 2 shows the estimates for the regression described in equation (3) where the dependent variable is (*TA–TS*) and *TS* is not entered as a control variable. Across all ages and cohorts, the coefficient on female is negative implying that teachers are more likely to judge males as being relatively better at English than math as compared to girls. The coefficients range from -0.035 to -0.045. The magnitude translates to 18% of a SD of the (*TA-TS*) distribution at age 11 in the MCS, 20% of a SD of the (*TA-TS*) distribution at age 11 in NS, and 10% of a SD of the (*TA-TS*) distribution at age 14 in NS. Overall, the estimates for specification 2 are all the opposite sign to the OLS estimates from specification 1 but are more similar to the estimates that use instrumental variable analysis to account for measurement error in test scores.

Table 2: Teacher Judgement of Student Comparative Advantage by Cohort and Age

VARIABLES	(1) **MCS Age 11**	(2) **NS Age 11**	(3) **NS Age 14**
		Dependent Variable is TA	
Female	0.024***	0.030***	0.089***
	(0.006)	(0.003)	(0.004)
TS	0.538***	0.431***	0.092***
	(0.014)	(0.009)	(0.007)
Observations	4,643	12,173	12,173
R-squared	0.314	0.238	0.062
SD of Dependent Variable	0.226	0.210	0.233
	IV Analysis: Dependent Variable is TA and Lagged TS used as an IV for Current TS		
Female	0.002		-0.046***
	(0.009)		(0.009)
TS	0.704***		1.077***
	(0.052)		(0.044)
Observations	4,643		12,173
R-squared	0.288		.
SD of Dependent Variable	0.226		0.233
		Specification 2: Dependent Variable is TA - TS	
Female	-0.038***	-0.045***	-0.035***
	(0.006)	(0.004)	(0.006)
Observations	4,643	12,173	12,173
R-squared	0.008	0.010	0.002
SD of Dependent Variable	0.214	0.221	0.358

Note: Robust standard errors in parentheses. *** p<0.01. *TA* is defined as rank in English teacher evaluation minus rank in math teacher evaluation. *TS* is defined as rank in English test score minus rank in math test score. Tied test scores and teacher evaluations are assigned the average rank.

How to Reconcile these Estimates?

As discussed in Section 2, even with consistent estimates, if the test score distribution differs by gender and if the coefficient on *TS*, $\hat{\delta}$, is not equal to 1 then the two specifications will give different estimates of the gender coefficient. Figure 1 showed that there was indeed a large difference in *TS* by gender with females doing better in English relative to math compared to males. However, even with different distributions, the two methods would yield equivalent estimates if the coefficient on *TS* was equal to 1. Interestingly, we find that the female estimate

obtained using IV at age 14 in NS was -0.046 which is very similar to the estimate obtained using Specification 2 of -0.035. When we look at the coefficient on *TS* in this IV specification, we find that the coefficient is 1.08, very close to 1. Indeed, we find that the IV estimates that yield consistent estimates all give female estimates much closer to the estimates of specification 2 and all have a coefficient on *TS* that gets closer to 1.

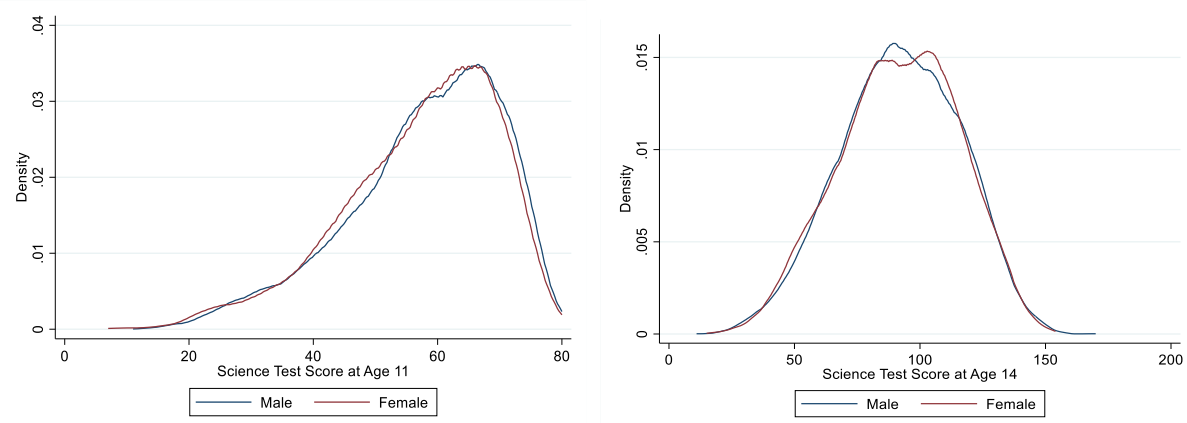
So, what is the takeaway? If the estimand of choice is the structural teacher bias parameter, then this would suggest focusing on the estimates from the IV regression as these estimates account for measurement error and other potential endogeneities associated with using current *TS* in the regression. These findings imply that at age 11 there is no gender bias in teacher judgement of student comparative advantage by gender, while at age 14, teacher judgement is actually in favour of boys, with boys being more likely to have higher *TA* even conditional on *TS*. However, if researchers are interested in whether the gap between *TA* and *TS* differs by gender, then they should consider the estimates from Specification 2. These estimates suggest that teacher evaluations of student comparative advantage tend to be in favour of boys rather than girls, i.e., that the gap between *TA* and *TS* is larger for boys than girls. Interestingly, this would suggest that females are not more likely to get higher teacher evaluations in English relative to math and so it is unlikely that teacher evaluations form a reason for the lower likelihood of girls to pursue STEM subjects in high school or college.

To understand further how teacher evaluations may be related to the gender gap in STEM, we make use of the KS science exams and science teacher assessments that are available in NS at age 11 and 14. While comparative advantage in English relative to math may be an important factor in the decision to continue with studying STEM, science ability may be just as important. Table A2 in the appendix shows the effect of gender on teacher evaluations in science using methods analogous to those used thus far. We find that the female estimates are very small and almost the same across Specifications 1 and 2, consistent with our earlier

findings that teachers do not mistakenly believe that girls have a comparative advantage in English relative to mathematics. The similarity across Specifications 1 and 2 occurs because, as shown in Figure 4 below, the science test score distribution is almost identical for males and females so that average Female *TS* is approximately equal to average Male *TS* and so the right-hand side in the equation below is approximately equal to β.

$$E(TA - A|F = 1) - E(TA - A|F = 0) = (\delta - 1)\{E(A|F = 1) - E(A|F = 0)\} + \beta$$

Figure 4: Distribution of Science Test Score by Gender in NS



7. Robustness Checks

In this section, we examine how robust our estimates are to various changes in specification. Tables 3, 4 and 5 show the estimates for robustness checks for the age 11 MCS sample, age 11 NS sample, and age 14 NS sample, respectively.[20]

[20] In addition to these reported robustness checks, we have verified that the results are very similar if we assign tied scores the highest or lowest rank instead of assigning ties the average rank. Also, we have information on schools and so can include school fixed effects. The results with school fixed effects are very similar to our baseline estimates. We have chosen not to include school fixed effects as the sample size is much smaller and may be less representative as at least 2 students per school are required in the survey.

Control Flexibly for Test Scores

Our Specification 1 controls linearly for student *TS*. However, it may be that English and mathematics test scores have separate effects on teacher judgement of comparative advantage and so we replace the comparative advantage (*TS*) measure with separate linear controls for mathematics and English test scores. In addition, we look at whether the estimates change when we enter *TS* more flexibly by including a quartic polynomial in *TS* in place of the linear control for *TS*. The second and third panels of Tables 3, 4 and 5 show that the estimates barely change when we include English and math controls separately or include a quartic polynomial.[21]

Standardise Test Score and Teacher Evaluation Distributions

Even though they are both measured as percentile ranks, the measures of *TA* and *TS* may not be directly comparable as the two distributions may have different variances. This is a particular issue for Specification 2 as it involves subtracting *TS* from *TA*. Therefore, we estimate specifications where we standardise *TA* and *TS* so that they are more comparable. To do this, we standardise separately the math test score, English test score, math teacher assessment and English teacher assessment so that each has the same standard deviation.[22] The estimated gender effects become slightly smaller in absolute terms but are quite similar to the unstandardized estimates. This is not surprising given that Table 1 shows that the standard deviation of each of these distributions is quite similar.

[21] We also tried including *TS* fixed effects in place of the linear *TS* control and found that this had negligible effects on the estimates. In the interest of space, we decided not to show the estimates with *TS* fixed effects.
[22] In standardizing the 4 measures to have the same standard deviation, we also ensure that the resulting standard deviation of the new standardized (*TA-TS*) variable is the same as the standard deviation of the unstandardised measure of (*TA-TS*). Therefore, the gender estimates are directly comparable with baseline.

Aggregate the Test Score Distribution

The student test scores take on many more values than the teacher evaluations due to the discrete nature of the teacher evaluations. The limited nature of the teacher evaluations means that teachers may be constrained from finely distinguishing between different students and this may vary by gender. Therefore, as a robustness check we aggregate the student test score distribution (*TS*) to the level of the teacher evaluation distribution (*TA*). To do this, we aggregate the math test score distribution to the level of the teacher math evaluation and similarly, we aggregate the English test score distribution to the level of the teacher English evaluation. Then we calculate *TS* as the rank in the aggregated English distribution minus the rank in the aggregated math distribution and we use this as our measure of *TS* in the regressions. Using this new aggregated *TS* distribution makes our Specification 2 gender estimates smaller in absolute terms but they remain negative and statistically significant, suggesting that the difference in the level of aggregation between the teacher and test score distributions does not drive our main findings.

Adding Controls for Non-Cognitive Measures

Ferman and Fontes (2022) and Cornwall et al. (2013) show that teachers are more likely to give better behaved students higher evaluations. At the age 11 survey in the MCS, we have information on teacher reports of student obedience, student self-reports of misbehaviour and also the responses to the Strength and Difficulties Questionnaire (SDQ). The SDQ is a behavioural questionnaire that is given to individuals aged between 2 and 17 and has 25 questions related to emotional symptoms, conduct problems, hyperactivity/inattention, pro-social behaviour, and peer relationship problems. We add controls for the scores in each of the 25 questions as well as controls for teacher and student reports of disobedience into all our specifications. We find that the addition of these non-cognitive measures has very little effect

on our estimates. This is unsurprising as, although boys are more likely to have behavioural issues than girls, there is little reason to believe that there would be a differential effect between English and mathematics.[23]

Teacher Evaluations Reported in the Survey

The MCS study asked teachers at age 11 to fill out a questionnaire related to the child. One of the questions asked how teachers viewed the child's ability relative to the average student. In particular, the survey asked, "In so far as your professional experience will allow, please rate this child in relation to all children of this age (i.e., not just their present class or, even, school)." and there were 5 responses, either "Well above average" "Above average" "Average", "Below average" or "Well below average". This question was asked for both English and mathematics.[24]

The last panel of Table 3 shows that, while there are some quantitative differences, the estimates using this alternative measure of *TA* are qualitatively similar to the baseline estimates. The estimate for Specification 1 increases from 0.024 to 0.052 while that for Specification 2 falls (in absolute terms) from -0.038 to -0.029.

[23] Similarly, although boys are much more likely to have a special education need (SEN), adding a control for this to the regression has little effect on the estimates.
[24] Campbell (2015) uses teacher survey reports about pupils and argues that, because they are confidential, they may be less influenced by factors that may affect KS teacher evaluations.

Table 3: Robustness Checks of Main Estimates for MCS Age 11 Sample

VARIABLES	(1) Specification 1: Dependent variable is *TA* and *TS* is a control	(2) IV Specification: Dependent variable is *TA* and *TS* is a control	(3) Specification 2: Dependent variable is (*TA-TS*)
		Baseline Estimates	
Female	0.024***	0.002	-0.038***
	(0.006)	(0.009)	(0.006)
Observations	4,643	4,643	4,643
R-squared	0.314	0.288	0.008
		Control for Math Score and English Score (rather than TS)	
Female	0.025***	0.001	N/A
	(0.006)	(0.009)	
Observations	4,643	4,643	
R-squared	0.317	0.289	
		Control for Quartic in TS	
Female	0.025***	0.004	N/A
	(0.006)	(0.016)	
Observations	4,643	4,643	
R-Squared	0.317	N/A	
		Standardise TS and TA Distributions	
Female	0.025***	0.001	-0.026***
	(0.006)	(0.009)	(0.006)
Observations	4,643	4,643	4,643
R-Squared	0.314	0.287	0.004
		Aggregate the TS Distribution	
Female	0.044***	-0.018	-0.012*
	(0.006)	(0.012)	(0.007)
Observations	4,643	4,643	4,643
R-Squared	0.301	N/A	0.001
		Control for Behavioural and Non-Cognitive Measures	
Female	0.021***	-0.002	-0.041***
	(0.006)	(0.009)	(0.007)
Observations	4,517	4,517	4,517
R-Squared	0.328	0.301	0.028
		Use Survey Reported Teacher Evaluations	
Female	0.052***	0.011	-0.029***
	(0.005)	(0.008)	(0.006)
Observations	4,643	4,643	4,643
R-Squared	0.272	0.160	0.005

Note: Robust standard errors in parentheses. *** p<0.01, * p<0.1 *TS* is measured for each student as the percentile rank of their English test score minus the rank of their mathematics test score. *TA* is measured for each student as the percentile rank of their teacher evaluation in English minus the rank of their teacher evaluation in mathematics. The standardised *TS* at age 7 is used to instrument the standardised *TS* at age 11 and the aggregated *TS* at age 7 (that was aggregated to match *TA* at age 7) is used to instrument the aggregated *TS* measure at age 11.

Table 4: Robustness Checks of Main Estimates for NS Age 11 Sample

VARIABLES	(1) Specification 1: Dependent variable is TA and TS is a control	(2) Specification 2: Dependent variable is (TA-TS)
	Baseline Estimates	
Female	0.030***	-0.045***
	(0.003)	(0.004)
Observations	12,173	12,173
R-squared	0.238	0.010
	Control for Math Score and English Score (rather than TS)	
Female	0.031***	N/A
	(0.003)	
Observations	12,173	
R-squared	0.240	
	Control for Quartic in TS	
Female	0.030***	N/A
	(0.003)	
Observations	12,173	
R-Squared	0.238	
	Standardise TS and TA Distributions	
Female	0.025***	-0.035***
	(0.003)	(0.004)
Observations	12,173	12,173
R-Squared	0.237	0.006
	Aggregate the TS Distribution	
Female	0.046***	-0.020***
	(0.003)	(0.004)
Observations	12,173	12,173
R-Squared	0.226	0.002

Note: Robust standard errors in parentheses. *** p<0.01. TS is measured for each student as the percentile rank of their English test score minus the rank of their mathematics test score. TA is measured for each student as the percentile rank of their teacher evaluation in English minus the rank of their teacher evaluation in mathematics. The standardised TS at age 7 is used to instrument the standardised TS at age 11 and the aggregated TS at age 7 (that was aggregated to match TA at age 7) is used to instrument the aggregated TS measure at age 11.

Table 5: Robustness Checks of Main Estimates for NS Age 14 Sample

VARIABLES	(1) Specification 1: Dependent variable is TA and TS is a control	(2) IV Specification: Dependent variable is TA and TS is a control	(3) Specification 2: Dependent variable is (TA-TS)
		Baseline Estimates	
Female	0.089***	-0.046***	-0.035***
	(0.004)	(0.009)	(0.006)
Observations	12,173	12,173	12,173
R-squared	0.062	.	0.002
		Control for Math Score and English Score (rather than TS)	
Female	0.091***	-0.029***	N/A
	(0.004)	(0.010)	
Observations	12,173	12,173	
R-squared	0.067	.	
		Control for Quartic in TS	
Female	0.088***	-0.052***	N/A
	(0.004)	(0.014)	
Observations	12,173	12,173	
R-Squared	0.062	.	
		Standardise TS and TA Distributions	
Female	0.087***	-0.043***	-0.030***
	(0.004)	(0.009)	(0.006)
Observations	12,173	12,173	12,173
R-Squared	0.062	NA	0.002
		Aggregate the TS Distribution	
Female	0.090***	-0.049***	-0.024***
	(0.004)	(0.011)	(0.006)
Observations	12,173	12,173	12,173
R-Squared	0.061	.	0.001

Note: Robust standard errors in parentheses. *** p<0.01. TS is measured for each student as the percentile rank of their English test score minus the rank of their mathematics test score. TA is measured for each student as the percentile rank of their teacher evaluation in English minus the rank of their teacher evaluation in mathematics. The standardised TS at age 7 is used to instrument the standardised TS at age 11 and the aggregated TS at age 7 (that was aggregated to match TA at age 7) is used to instrument the aggregated TS measure at age 11.

Other Specifications

While it is natural to use the difference in English and math test scores and teacher evaluations to measure comparative advantage, there are other approaches that could be used to measure comparative advantage. One approach that is similar to Specification 1 is to create

a binary dependent variable denoting whether the rank in the teacher evaluation (TE) in English is greater than the rank in the teacher evaluation in math and use this as the dependent variable. Then to regress this outcome on female and an indicator variable denoting whether the rank in English test score is greater than the rank in math test score:

$$I(English\ \ TE > Math\ TE) = \kappa_o + \kappa_1 Female + \gamma \mathrm{I}(English\ Score > Math\ Score) + \epsilon \qquad (4)$$

We also estimate a version of equation (4) that is in the spirit of Specification 2 whereby we incorporate the test score measures into the dependent variable and exclude controls for test scores from the regression. Specifically, the dependent variable is calculated as the difference between an indicator variable denoting whether the rank of the English teacher evaluation is greater than the rank of the math teacher evaluation, and an indicator variable denoting whether the rank of the English test score is greater than the rank of the math test score:

$$I(English\ TE > Math\ TE)\ -\ \mathrm{I}(English\ Score > Math\ Score)\ = \eta_0 + \eta_1 Female + \xi \qquad (5)$$

Table 6 shows the results from estimating equations (4) and (5). Interestingly, across all specifications, the same pattern emerges as in the baseline specifications. Regressions in the top panel (similar to Specification 1) give a positive coefficient on female suggesting that teachers are more likely to judge females as having a comparative advantage in English relative to maths as compared to boys with equivalent test scores. When considered relative to the standard deviation of the dependent variable, the magnitudes of the coefficients on female are similar to those in Table 2 with the estimates implying an effect of 17% of a standard deviation for the MCS age 11 study, 23% of a standard deviation for the NS age 11 study, and 40% of a standard deviation for the NS age 14 study.

To account for measurement error in the test score measures, we use lagged test scores and create a binary variable denoting whether the rank in the lagged English test score is greater than the rank in the lagged math test score as an instrumental variable. We find that using this new binary variable as an instrument leads to no statistically significant effects of gender at age 11 in the MCS but to a negative estimate on female for the age 14 NS cohort. This is the same pattern that we saw in our baseline specifications in Table 2. Finally, we see that the specification that includes both test scores and teacher evaluations in the dependent variable (and omits test scores from the control set) results in a negative coefficient on female, similar to Specification 2 in Table 2. The magnitudes translate to teachers judging boys as having a higher CA in English relative to maths compared to girls that is equivalent to 20% of a standard deviation of the dependent variable at age 11 in the MCS, and 18% of a standard deviation at age 11 in NS and at age 14 in the NS.

Lastly, we estimate a specification similar to Burgess and Greaves (2013) whereby the dependent variable is an indicator variable denoting whether *TA* is greater than *TS*:

$$I(TA > TS) = \pi_0 + \pi_1 Female + \lambda TS + \zeta. \quad (6)$$

Like Burgess and Greaves, we include the *TS* measure as a control variable. However, we also estimate a specification whereby we exclude *TS* from the set of controls; this is similar in spirit to Specification 2. Once again, the findings are consistent with the other specifications with positive coefficients on female when we control for *TS* and negative coefficients when we do not. Overall, we conclude that our findings are robust to the functional form of the comparative advantage measures used.

Table 6: Robustness Checks for Other Specifications

VARIABLES	(1) **MCS Age 11**	(2) **NS Age 11**	(3) **NS Age 14**
	Dependent Variable is I(English Teacher Evaluation > Math Teacher Evaluation)		
Female	0.073***	0.091***	0.092***
	(0.012)	(0.007)	(0.004)
I(English Test Score > Math Test Score)	0.270***	0.213***	0.048***
	(0.012)	(0.007)	(0.004)
Observations	4,643	12,173	12,173
R-squared	0.126	0.105	0.057
SD of Dependent Variable	0.419	0.389	0.233
	IV Analysis: Dependent Variable is I(English Teacher Evaluation > Math Teacher Evaluation) and Lagged I(English Test Score > Math Test Score) used as an IV		
Female	-0.013	N/A	-0.037***
	(0.020)		(0.011)
I(English Test Score > Math Test Score)	0.615***		0.695***
	(0.064)		(0.040)
Observations	4,643		12,173
R-squared	.		.
SD	0.419		0.233
	Dependent Variable is 1(English Teacher Evaluation > Math Teacher Evaluation) – 1(English Test Score > Math Test Score)		
Female	-0.108***	-0.098***	-0.097***
	(0.016)	(0.010)	(0.009)
Observations	4,643	12,173	12,173
R-squared	0.010	0.009	0.009
SD of Dependent Variable	0.529	0.532	0.521

Note: Robust standard errors in parentheses. *** p<0.01. In the IV analysis at age 11, the instrument used is an indicator for whether age 7 rank in math score is larger than age 7 rank in English score, I(English Score 7 > Math Score 7). Similarly, in the NS data, the instrument used at age 14 is I(English Score 11 > Math Score 11). We cannot estimate an IV model for NS at age 11 as there are no test scores available in the dataset prior to age 11.

Table 7: Robustness Checks for Other Specifications

VARIABLES	(1) MCS Age 11	(2) NS Age 11	(3) NS Age 14
	Dependent Variable is I(TA>TS)		
Female	0.029**	0.054***	0.086***
	(0.014)	(0.009)	(0.007)
TS	-0.849***	-0.980***	-0.988***
	(0.029)	(0.017)	(0.008)
Observations	4,643	12,173	12,173
R-squared	0.144	0.183	0.367
SD of Dependent Variable	0.492	0.50	0.50
	Dependent Variable is I(TA>TS) (Lagged TS used as IV for current TS)		
Female	-0.006	N/A	-0.057***
	(0.022)		(0.012)
TS	-0.591***		0.060
	(0.123)		(0.060)
Observations	4,643		12,173
R-squared	0.132		0.002
SD of Dependent Variable	0.492		0.499
	Dependent Variable is I(TA>TS)		
Female	-0.086***	-0.076***	-0.049***
	(0.015)	(0.009)	(0.009)
Observations	4,643	12,173	12,173
R-squared	0.007	0.006	0.002
SD of Dependent Variable	0.492	0.499	0.499

Note: Robust standard errors in parentheses. *** p<0.01, ** p<0.05. *TS* is measured for each student as the percentile rank of their English test score minus the rank of their mathematics test score. *TA* is measured for each student as the percentile rank of their teacher evaluation in English minus the rank of their teacher evaluation in mathematics.

8. Conclusion

We conclude by summarizing our main findings. While much literature is concerned with estimating various types of teacher biases, we show conceptually that there is no single "teacher bias" effect. Even if teachers have no gender bias, teacher evaluation differences by gender may systematically deviate from test score differences if the distribution of test scores

differs between boys and girls. Therefore, we conclude that it is important for researchers to be clear about the target estimand.

Using two recent UK cohort surveys, we show that these conceptual issues matter in practice when we evaluate whether teachers are likely to over-estimate female comparative advantage in English relative to mathematics. In our application, the gender effects differ substantially depending on the method used and the signs can even switch across specifications. Our preferred estimates provide no evidence that teachers exaggerate the female advantage in English relative to mathematics and generally suggest the opposite. One implication is that it is unlikely that teacher misperceptions of comparative advantage by gender are an important cause of the gender gap in STEM.

References

Alesina, A., Carlana, M., Ferrara, E.L., Pinotti, P., (2018). Revealing stereotypes: Evidence from immigrants in schools. Technical report, *National Bureau of Economic Research*.

Aucejo, E. and James, J. (2021). 'The path to college education: The role of math and verbal skills', *Journal of Political Economy*, vol. 129(10), pp. 2905–46.

Baird, Jo-Anne. (1998). What's in a name? Experiments with blind marking in A-level examinations. *Educational Research* 40, no. 2:191–202.

Borgonovi, F., Choi, Á., & Paccagnella, M. (2018). "The evolution of gender gaps in numeracy and literacy between childhood and adulthood", *OECD Education Working Papers*, No. 184, OECD Publishing, Paris.

Botelho, Fernando, Ricardo A. Madeira, and Marcos A. Rangel (2015). "Racial Discrimination in Grading: Evidence from Brazil." *American Economic Journal: Applied Economics* 7, no. 4 (2015): 37–52. http://www.jstor.org/stable/24739058.

Breda, T. & Hillion, M. (2016) Teaching accreditation exams reveal grading biases favor women in male-dominated disciplines in France. *Science* 353, 474–478.

Breda, T., Ly, S.T., (2015). "Professors in core science fields are not always biased against women: Evidence from France." *American Economic Journal: Applied Economics* 7 (4), 53–75

Burgess, Simon, Daniel Sloth Hauberg, Beatrice Schindler Rangvid, and Hans Henrik Sievertsen. (2022). "The importance of external assessments: High school math and gender gaps in STEM degrees." *Economics of Education Review*, 88: 102267.

Burgess, S., & Greaves, E. (2013). Test scores, subjective assessment, and stereotyping of ethnic minorities. *Journal of Labor Economics, 31*(3), 535–576.

Campbell, T. (2015). Stereotyped at seven? Biases in teacher judgement of pupils' ability and attainment. Journal of Social Policy, 44(3), 517–547.

Card, D. and A.A. Payne. (2021). High School Choices and the Gender Gap in STEM. *Economic Inquiry*. 59: 9-28.

Carlana, Michela. (2019). "Implicit Stereotypes: Evidence from Teachers' Gender Bias." *The Quarterly Journal of Economics* 134(3): 1163–1224.

Cavaglia, C., Machin, S., McNally, S., & Ruiz-Valenzuela, J. (2020). "Gender, Achievement and Subject Choice in English Education, Paper" Prepared for *Oxford Review of Economic Policy* issue on Gender

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff (2014a), "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates," American Economic Review, September 2014, 104 (9), 2593–2632.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff (2014b), "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood," American Economic Review, September 2014, 104 (9), 2633–2679.

Chowdhury, Shyamal and Klauzner, Ilya and Slonim, Robert. (2020). "What's in a Name? Does Racial or Gender Discrimination in Marking Exist?" *IZA Discussion Paper* No. 13890.

Cornwell, C. M., Mustard, D. B., & van Parys, J. (2013). "Noncognitive Skills and the Gender Disparities in Test Scores and Teacher Assessments: Evidence from Primary School." *Journal of Human Resources*, 48(1), 236–264.

Dee, Thomas S. (2015). "Social Identity and Achievement Gaps: Evidence from an Affirmation Intervention." *Journal of Research on Educational Effectiveness* 8 (2): 149–68.

Delaney J.M. and P.J. Devereux (2019). "Understanding Gender Differences in STEM: Evidence from College Applications." *Economics of Education Review*, Volume 72, October, Pages 219-238.

Delaney J.M. and P.J. Devereux (2021). "High School Rank in Math and English and the Gender Gap in STEM." *Labour Economics,* Volume 69, April, Pages 101969.

Delaney J.M. and P.J. Devereux (2021). "The Economics of Gender and Educational Achievement: Stylized Facts and Causal Evidence." *Oxford Research Encyclopedia of Economics and Finance*, https://doi.org/10.1093/acrefore/9780190625979.013.663

Falch, Torberg, and Linn Renée Naper. (2013). "Educational Evaluation Schemes and Gender Gaps in Student Achievement." *Economics of Education Review* 36: 12–25.

Ferman, Bruno, and Luiz Felipe Fontes (2022). "Assessing knowledge or classroom behavior? Evidence of teachers' grading bias" *Journal of Public Economics.* Volume 216, 104773

Gershenson, Seth, Stephen B Holt, and Nicholas W Papageorge (2016). "Who Believes in Me? The Effect of Student–Teacher Demographic Match on Teacher Expectations." *Economics of Education Review* 52: 209–24.

Gibbons, Stephen and Arnaud Chevalier (2008). "Assessment and age 16+ education participation.", *Research Papers in Education*, 23:2, 113-123, DOI: 10.1080/02671520802048638

Goulas, S., Griselda, S., and Megalokonomou, R. (2022). Comparative advantage and gender gap in STEM. *Journal of Human Resources*.

Hanna, Rema N., and Leigh L. Linden. (2012). "Discrimination in Grading." *American Economic Journal: Economic Policy*, 4 (4): 146-68**.**

Hinnerich, B.T., Höglin, E., Johannesson, M., (2011). Are boys discriminated in Swedish high schools? Economics of Education Review 30 (4), 682–690.

Lavy, V. (2008). Do gender stereotypes reduce girls' or boys' human capital outcomes? Evidence from a natural experiment. *Journal of Public Economics*, 92(10):2083–2105.

Lavy, V. and Sand, E. (2018). On the origins of gender human capital gaps: Short and long term consequences of teachers' stereotypical biases. *Journal of Public Economics*.

Lindahl, E. (2007). Comparing teachers' assessments and national test results: evidence from sweden. Technical report, Working Paper.

McCoy, S., Byrne, D., and O'Connor, P (2021), Gender stereotyping in mothers' and teachers' perceptions of boys' and girls' mathematics performance in Ireland, *Oxford Review of Education*, Available online: https://doi.org/10.1080/03054985.2021.1987208.

McNally, S. (2020). "Gender Differences in Tertiary Education: What explains STEM Participation?", IZA Policy Paper, No. 165, Institute of Labor Economics (IZA), Bonn.

Peri, G., Shih, K. and Sparber, C. (2015). Stem workers, h-1b visas, and productivity in us cities, *Journal of Labor Economics* 33(S1), S225–S255. 1, 2.1, 4.

Reeves, D.J., Boyle, W.F., & Christie, T. (2001). The relationship between teacher assessments and pupil attainments in standard test tasks at Key Stage 2, 1996-98. *British Educational Research Journal*, *27(2)*, 141–160. https://doi.org/10.1080/01411920120037l0

Saltiel, F. (2022). "Multidimensional Skills and Gender Differences in STEM Majors." *Economic Journal* 133 (651): 1217– 1247.

Shi, Y. (2018). 'The puzzle of missing female engineers: Academic preparation, ability beliefs, and preferences', *Economics of Education Review*, vol. 64, pp. 129–43.

Speer, J. D. (2017). "The Gender Gap in College Major: Revisiting the Role of Pre-college Factors." *Labour Economics, 44*, 69–88.

Speer, J. D. (2023). "Bye Bye Ms. American Sci: Women and the Leaky STEM Pipeline." *Economics of Education Review,* 93, April 2023.

Terrier, Camille. (2020). "Boys Lag Behind: How Teachers' Gender Biases Affect Student Achievement." *Economics of Education Review* 77: 101981.

University College London, UCL Institute of Education, Centre for Longitudinal Studies, Department for Education. (2021). *Millennium Cohort Study: Linked Education Administrative Datasets (National Pupil Database), England: Secure Access*. [data collection]. *2nd Edition.* UK Data Service. SN: 8481, http://doi.org/10.5255/UKDA-SN-8481-2

University College London, UCL Institute of Education, Centre for Longitudinal Studies. (2020). *Next Steps: Linked Education Administrative Datasets (National Pupil Database), England, 2005-2009: Secure Access*. [data collection]. *6th Edition.* UK Data Service. SN: 7104, http://doi.org/10.5255/UKDA-SN-7104-6

Table A1 : Instrumental Variable Estimation First Stage Regressions

VARIABLES	(1) Rank TS Age 11 in MCS	(2) Rank TS Age 14 in NS
Female	0.123***	0.088***
	(0.006)	(0.006)
Rank TS Age 7 MCS	0.194***	
	(0.011)	
Rank TS in Age 11 NS		0.358***
		(0.013)
F-Statistic	372.14	753.62
Observations	4,643	12,173
R-squared	0.147	0.107

Note: Robust standard errors in parentheses. *** p<0.01. TS is measured for each student as the percentile rank of their English test score minus the rank of their mathematics test score. TS at age 7 in the MCS is used to instrument for TS at age 11 in the MCS. TS at age 11 in NS is used to instrument for age 14 TS for the NS cohort. There are no lagged test scores available prior to age 11 in NS and so we cannot do the IV analysis for the age 11 NS cohort.

Table A2: Teacher Judgement of Student Science Ability at Age 11 and 14 in NS

VARIABLES	(1) **NS Age 11**	(2) **NS Age 14**
	Specification 1: Dependent Variable is Teacher Science Evaluation	
Female	-0.002	-0.007
	(0.003)	(0.005)
Science Test Score	0.652***	0.314***
	(0.005)	(0.008)
Observations	11,959	11,959
R-squared	0.536	0.107
SD of Dependent Variable	0.257	0.276
	IV Analysis: Dependent Variable is Teacher Science Evaluation and Science Test Score at Age 11 is used as an Instrument for Science Test Score at Age 14	
Female		-0.005
		(0.010)
Science Test Score		1.919***
		(0.043)
Observations		11,959
R-squared		N/A
SD of Dependent Variable		0.276
	Specification 2: Dependent Variable is Teacher Science Evaluation – Science Test Score	
Female	0.005	-0.006
	(0.004)	(0.006)
Observations	11,959	11,959
R-squared	0.000	0.000
SD of Dependent Variable	0.202	0.328

Note: Robust standard errors in parentheses. *** p<0.01. The rank of the teacher science evaluation and science test score is used in the analysis. Tied test scores and tied teacher evaluations are assigned the average rank.