# Investigating how the accuracy of teacher expectations of pupil performance relate to socioeconomic and genetic factors

**Ciarrah-Jane Barry**
University of Bristol
**Neil Davies**
University of Bristol
**Tim Morris** ( tim.morris@bristol.ac.uk )
University of Bristol

**Research Article**

**Keywords:** education, assessment, inequality, ALSPAC, discrimination, prediction, genetic

**Posted Date:** September 27th, 2021

**DOI:** https://doi.org/10.21203/rs.3.rs-927102/v1

# Abstract

Teacher expectations of pupil ability can influence pupil's educational progression, impacting subsequent streaming and exam level entry. Systematic errors in the accuracy of teacher expectations of pupil achievement may therefore have a lasting detrimental effect on a child's education and life prospects. Associations between socioeconomic and demographic factors with teacher expectation accuracy have been previously investigated, but it is not known how expectation accuracy may relate to genetic factors. We investigated these relationships using data on nationally standardized exam results at ages 11 and 14 from a UK longitudinal cohort study. We found that teacher expectation of achievement was strongly correlated with subsequent achievement, that teacher expectation accuracy was patterned by pupil socioeconomic background but not teacher characteristics, and that teacher expectation accuracy related to pupil's genetic liability to education. We find no strong evidence for heritability in teacher reporting accuracy, suggesting that the majority of variation in teacher expectation accuracy can be attributed to non-genetic factors.


# Introduction

Teacher's expectations of ability can affect pupil's academic achievement throughout their educational career from initial enrolment through to the end of compulsory schooling [1–6]. They can influence the subjects that pupils take, whether they are entered into an advanced stream, the level of exam they are entered to, how long they remain in education and ultimately their educational attainment [7–10]. These expectations are based on a teacher's understanding and experience with pupils over an extended period and can have advantages over pupil achievement measured by test performance [11,12]. For example, expectations may avoid a misleading representation of a pupil's ability if they tested on a particularly good or bad day, they avoid incentives to "teach to the test", they may remove the stress of formalised testing, and they can ensure that ability is measured using a broader range of factors than test performance alone [13–15].

However, disadvantages also exist with teacher expectations. First, there is potential for either conscious or unconscious bias against specific pupils or groups, such as by gender, socioeconomic background, ethnicity or special educational needs status [6,10,16–20]. Teacher expectation theory posits that while teachers form inferences about their student's future academic achievement for individuals and groups of students, their expectations of pupils may be biased by students' backgrounds [21–23]. Second, variation in teacher and classroom characteristics may result in systematic differences in the accuracy of teacher expectations. Teachers with larger class sizes have less individual contact time with each pupil, meaning that their expectations may be less reliable than teachers with smaller classes [24]. Third, teachers may only have a small sample of previous students to draw upon so the accuracy of their expectations of future pupil performance may be dependent on their level of experience [25]. Given these advantages and disadvantages, teacher expectations can be used for early assessment and streaming before being replaced by formalised testing and assessment later in schooling. However, recent policy updates

highlight how this is not always the case. The Covid-19 pandemic led to the use of assessments from teacher expectations for determining academic performance in the UK, following heavy criticism of the UK Government's initial statistical model for exam results which was reported to have widespread inconsistencies.

Previous studies have shown that systematic differences exist in the accuracy of teacher expectations of subsequent achievement across groups of pupils. On average teachers underestimate outcome for students with special educational needs, those of black-African and black-Caribbean ethnic origin, and those of a lower socioeconomic position, and boys [16,17,26–28]. For example, a meta-analysis of 39 studies demonstrated that teacher's expectations of pupils was linked to pupils' ethnicity, with higher expectations being held for European-American pupils than for ethnic-minority pupils [29]. These systematic differences can be detrimental to pupils who are under or overestimated by their teachers. For example, pupils who felt undervalued by teachers may be less likely to be engaged in school and have lower achievement than expected [30]. Conversely, pupils whose ability is overestimated may be overlooked by teachers or placed into streams that are too advanced for them and they therefore may not receive adequate support to accomplish their academic potential [1,9,16].

New data and methods offer the opportunity to examine the accuracy of teacher expectations in novel ways. The use of genetic data in educational research is growing and there is now evidence that many genetic variants (single nucleotide polymorphisms, SNPs) associate with educational attainment, achievement and progress [31–33]. The largest genome-wide association study of education to date identified over 1,000 SNPs which combined into a polygenic score (PGS) explain around 12% of the variation in educational attainment. Because genetic variation is set at birth and cannot be affected by environmental factors post-conception, associations between genetic factors and individual characteristics are robust to confounding and reverse causation that pervade much educational research [34]. While genetic variation is not directly observable, its effects on education are.

If teachers overestimate the achievement of a given group (for example girls from high socioeconomic position (SEP) backgrounds) and underestimate the achievement of another group (for example boys from low SEP backgrounds) then we might expect the difference between teacher expectations and exam results to be partially explained by the educational attainment polygenic score. To investigate this, we estimated the association of teacher expectation accuracy with socioeconomic, demographic and genetic factors in the Avon Longitudinal Study of Parents and Children (ALSPAC), a UK longitudinal cohort. We answered three research questions: 1) How accurately do teacher expectations associate with realised achievement? 2) How do teacher expectations associate with pupil socioeconomic and demographic factors and teacher characteristics? 3) Can pupil's common genetic variation explain differences in the accuracy of teacher expectations?

# Results

Due to attrition and item non-response in the ALSPAC cohort, the complete case samples of ALSPAC participants available for analyses are 2,341 at Key Stage 2 (age 11) and 3,696 at Key Stage 3 (age 14). We therefore ran multiple imputation to recover missing data and increase the statistical power of our analyses. Our multiple imputation sample was 7,465, with imputations run over 100 iterations. Teacher expectation accuracy was obtained by regressing realised achievement in standardised national examinations for Mathematics, English and Science on teacher expectations of performance in these subjects.

# Association of teacher expectations and achievement

Each one standard deviation (SD) increase in teacher expectation was associated with a 0.88 (95% CI: 0.87, 0.89) and 0.92 (95% CI: 0.91, 0.93) SD increase in realised achievement at Key Stages 2 and 3 respectively in the imputed data (Table 1). Teacher expectations explained a large amount of variation in realised pupil achievement as demonstrated by the high $R^2$ values of 0.78 and 0.85 at Key Stages 2 and 3 respectively.

Table 1
Standardised association between achievement and teacher expectations.

| | Effect estimate (95% CI) | | $R^2$ | |
|---|---|---|---|---|
| | Complete case | Imputed | Complete case | Imputed |
| Key Stage 2 (age 11) | 0.76 (0.74, 0.78) | 0.88 (0.87, 0.89) | 0.76 | 0.78 |
| Key Stage 3 (age 14) | 0.83 (0.82, 0.84) | 0.92 (0.91, 0.93) | 0.85 | 0.85 |

# Phenotypic predictors of teacher expectation accuracy

There was evidence that teacher expectation accuracy associated with some demographic and socioeconomic measures at both Key Stages (Table 2). At Key Stage 2 there was strong evidence that pupils with less educated mothers underperformed their teacher's expectations relative to pupils with degree educated mothers (O-level: -0.13, 95% CI: -0.23, -0.034; Vocational: -0.29, 95% CI: -0.41, -0.16; CSE: -0.38, 95% CI: -0.49, -0.26). There was also strong evidence that children from families of lower parental social class underperformed their teacher's expectations relative to pupils whose parents were in the highest social class jobs (II: -0.11, 95% CI: -0.20, -0.015; III non-manual: -0.17, 95% CI: -0.28, -0.07; III manual: -0.22, 95% CI: -0.34, -0.10; IV: -0.34; 95% CI: -0.50, -0.18). There was strong evidence that pupils born later in the school year slightly underperformed compared to those who were born earlier in the year (-0.011, 95% CI: -0.018, -0.005). Associations were consistent at Key Stage 3 for maternal education (O-level: -0.16, 95% CI: -0.28, -0.04; Vocational: -0.21, 95% CI: -0.35, -0.08; CSE: -0.28, 95% CI: -0.41, -0.15) and month of birth (-0.013, 95% CI: -0.021, -0.005), but not for parental social class. There was also strong evidence that children from families in the lowest two income brackets underperformed teacher

expectations at Key Stage 3 compared to those in the top bracket £100–199 per week: -0.12; 95% CI: -0.23, -0.005; Less than £100 per week: -0.17; 95% CI: -0.33, -0.018). There was no strong evidence that children of differing sex or SEN status under/overperformed their teacher's expectations.

Table 2

**Associations between teacher expectation, socioeconomic and demographic variables.** Positive values reflect pupils who overperformed relative to their teacher's expectations, while negative values reflect pupils who underperformed relative to their teacher's expectations. Results for complete case analyses presented in Supplementary Table S1.

| | Teacher expectation accuracy at Key Stage 2 (age 11) | | Teacher expectation accuracy at Key Stage 3 (age 14) | |
|---|---|---|---|---|
| | **Coefficient (95% CI)** | **P value** | **Coefficient (95% CI)** | **P value** |
| **Gender** | | | | |
| *Female* | *Reference* | | *Reference* | |
| *Male* | 0.010 (-0.04, 0.06) | 0.710 | 0.024 (-0.036, 0.08) | 0.433 |
| **Month of delivery**[1] | -0.011 (-0.018, -0.005) | 0.001 | -0.013 (-0.021, -0.005) | 0.002 |
| **SEN status** | | | | |
| *Not statemented* | *Reference* | | *Reference* | |
| *Has a statement* | -0.00024 (-0.23, 0.23) | 0.998 | -0.09 (-0.35, 0.17) | 0.502 |
| **Mothers highest education** | | | | |
| *Degree* | *Reference* | | *Reference* | |
| *A level* | -0.09 (-0.18, 0.0017) | 0.054 | -0.04 (-0.16, 0.07) | 0.438 |
| *O level* | -0.13 (-0.23, -0.034) | 0.008 | -0.16 (-0.28, -0.04) | 0.007 |
| *Vocational* | -0.29 (-0.41, -0.16) | < 0.001 | -0.21 (-0.35, -0.08) | 0.002 |
| *CSE* | -0.38 (-0.49, -0.26) | < 0.001 | -0.28 (-0.41, -0.15) | < 0.001 |
| **Parental social class** | | | | |
| *I* | *Reference* | | *Reference* | |
| *II* | -0.11 (-0.20, -0.015) | 0.023 | 0.08 (-0.014, 0.18) | 0.096 |
| *III non-manual* | -0.17 (-0.28, -0.07) | 0.001 | 0.05 (-0.06, 0.17) | 0.379 |
| *III manual* | -0.22 (-0.34, -0.10) | < 0.001 | 0.021 (-0.11, 0.15) | 0.751 |
| *IV* | -0.34 (-0.50, -0.18) | < 0.001 | 0.028 (-0.15, 0.21) | 0.764 |
| *V* | -0.018 (-0.37, 0.33) | 0.919 | 0.11 (-0.31, 0.53) | 0.600 |
| **Income, £ per week** | | | | |
| *Over 400* | *Reference* | | *Reference* | |

*[1] Where September = 1, October = 2 etc. [2] Data available for KS2 only*

| | Teacher expectation accuracy at Key Stage 2 (age 11) | | Teacher expectation accuracy at Key Stage 3 (age 14) | |
|---|---|---|---|---|
| | **Coefficient (95% CI)** | **P value** | **Coefficient (95% CI)** | **P value** |
| *300–399* | -0.019 (-0.10, 0.06) | 0.637 | 0.025 (-0.07, 0.12) | 0.602 |
| *200–299* | -0.039 (-0.12, 0.04) | 0.342 | -0.025 (-0.12, 0.07) | 0.595 |
| *100–199* | -0.07 (-0.18, 0.028) | 0.153 | -0.12 (-0.23, -0.005) | 0.041 |
| *Less than 100* | -0.032 (-0.17, 0.10) | 0.643 | -0.17 (-0.33, -0.018) | 0.028 |
| **Teacher gender[2]** | | | | |
| *Female* | | | | |
| *Male* | -0.039 (-0.12, 0.040) | 0.335 | | |
| **Length of teaching time[2]** | | | | |
| *10 + years* | | | | |
| *3–9 years* | 0.038 (-0.031, 0.11) | 0.285 | | |
| *1–2 years* | 0.10 (-0.06, 0.26) | 0.230 | | |
| *Less than 1 year* | 0.07 (-0.15, 0.29) | 0.534 | | |
| **Class size[2], per additional 10 pupils** | -0.025 (-0.10, 0.05) | 0.532 | | |
| **Constant** | 0.44 (0.20, 0.68) | < 0.001 | 0.18 (0.07, 0.29) | 0.002 |

*[1] Where September = 1, October = 2 etc. [2] Data available for KS2 only*

There was little evidence that teacher gender (-0.039, 95% CI: -0.12, 0.40), teacher experience (e.g. 1–2 years vs 10 or more years: 0.07, 95% CI: -0.15, 0.29), or class size (-0.025, 95% CI: -0.10, 0.05 per additional 10 pupils) were associated with teacher expectation accuracy. Results were consistent across the imputed and complete case analyses (Supplementary Table S1).

# Genotypic predictors of teacher expectation accuracy

We assessed the association between teacher expectation accuracy and a pupil's polygenic score for educational attainment built from the largest GWAS of education to date [32]. There was strong evidence for an association between a pupil's educational attainment polygenic score and teacher expectation accuracy at both Key Stage 2 (0.13; 95% CI: 0.11, 0.16) and Key Stage 3 (0.10; 95% CI: 0.08, 0.13) (Table 3). These associations persisted after adjustment for demographic and socioeconomic factors, and the first 20 principal components of ancestry (KS2: 0.08; 95% CI: 0.06, 0.11; KS3: 0.07; 95% CI: 0.04, 0.10).

This suggests that pupils with higher polygenic scores for educational attainment were more likely to outperform their teachers' expectations than children with lower polygenic scores.

Table 3

**Associations between teacher expectation and pupil's polygenic scores (PGS).** Positive values reflect pupils who overperformed relative to their teacher's expectations, while negative values reflect pupils who underperformed relative to their teacher's expectations. n = 7,465. Full results in Supplementary Table S2. Results for complete case analyses presented in Supplementary Table S3.

| | Teacher expectation accuracy at Key Stage 2 (age 11) | | Teacher expectation accuracy at Key Stage 3 (age 14) | |
|---|---|---|---|---|
| | Coefficient (95% CI) | P-value | Coefficient (95% CI) | P-value |
| PGS only | 0.13 (0.11, 0.16) | < 0.001 | 0.10 (0.08, 0.13) | < 0.001 |
| PGS adjusted for covariates | 0.08 (0.06, 0.11) | < 0.001 | 0.07 (0.04, 0.10) | < 0.001 |

# SNP heritability of teacher expectation accuracy

We used GCTA-GREML[35] to assess SNP heritability of teacher expectation accuracy. We found weak evidence for SNP heritability of teacher expectation accuracy at both ages. SNP heritability was estimated at 6.5% (95% CI: -6.7%, 19.6%) for Key Stage 2 and 14% for Key Stage 3 (95% CI: -4.6%, 33.4%) after adjustment for principal components of ancestry. Similar to the PGS results, this suggests that pupil's under or over-performance relative to their teacher's expectations may relate to their genome-wide genetic variation (Fig. 1).

# Discussion

Our results suggested that teachers' expectation of their pupil's achievement was generally accurate at two Key Stages of UK education (ages 11 and 14). We found evidence that teacher expectation accuracy was related to some socioeconomic or demographic factors, principally pupil's maternal education and age in year at both Key Stages, and parental social class and household income at Key Stage 2 and 3 respectively. This patterning was consistently in the same direction, whereby pupil's from more disadvantaged backgrounds underperformed compared to their advantaged peers. Our findings conform to those from previous studies which have found differential teacher expectation accuracy towards certain groups of pupils such as those from lower socioeconomic position backgrounds [3,6,10,16,27,28,36–38]. For example, disparity was found between teacher assessed measures and Foundation Stage Profile assessment on socioeconomic and demographic factors including income, gender, special educational needs status and ethnicity [4]. We found little evidence that teacher gender, teacher experience or class size

were associated with the accuracy of teacher expectations. This contrasts to previous research that has observed strong associations between these factors and the accuracy of teacher expectations [16,25].

We found mixed evidence for associations between genetic factors and teacher expectation accuracy. Using a polygenic score from a large GWAS of educational attainment [32], we found strong evidence that pupils with higher polygenic scores were more likely to outperform their teachers' expectations compared to children with lower polygenic scores. Using all genomewide data within a GREML-GCTA framework we found only weak evidence for SNP heritability at both ages, though these results should be interpreted with caution due to their imprecision. These results suggest that some of the variation in teacher reporting accuracy can be explained by genetic variation at the pupil level. Conversely, they suggest that the overwhelming majority of variation in teacher expectation accuracy can be explained by non-genetic (environmental and residual) factors. Genetic liability towards educational performance could operate through a range of mediating mechanisms, such as personality characteristics or attitudes to learning and schoolwork [16,36]. In this way, 'invisible' genetic variation may become visible to teachers, influencing their expectations of a pupil's future performance. Future studies with larger sample sizes are required to verify these findings, however our results build upon previous studies demonstrating robust associations between genetic factors and achievement throughout schooling [31,39,40].

Our analyses were unable to determine whether inaccuracies in teacher expectation were due to error or detrimental bias on the teachers part, i.e. whether teachers were prejudiced against specific groups of pupils [2,41]. Prejudice on the teachers part however may have been more likely to result in pupils *over* performing their teachers (unfairly negative) expectations. Similarly, pupil behavioural change due to self-fulfilling prophecies from inaccurate teacher forecasts may be expected to result in accurate teacher expectations, even if they are prejudiced [42]. Regardless, our findings highlight that some groups of students systematically underperform their teachers' expectations. We were unable to reliably investigate how ethnicity related to teacher expectations due to the ethnic homogeneity of the ALSPAC cohort and the low numbers of ethnic minority participants. Previous studies conducted on more ethnically representative cohorts have shown that teacher expectations differ by pupil ethnicity [6,18,30].

Several limitations exist with this study. First, generalisability of these findings to the wider UK school population may be limited. Our sample was ethnically homogenous and restricted to those who were recruited from a single geographical area over a three school-year period. This tightly defined sampling frame means that there will likely be reduced environmental and genetic variation in our sample compared to the broader population of the UK. Furthermore, within ALSPAC there is greater attrition for pupils from families of lower socioeconomic position and poorer general health [43], meaning that this demographic are underrepresented and our complete case analyses may be biased. Results from our multiple imputation analyses were broadly consistent with the complete case analyses, suggesting that bias due to attrition may be limited.

Second, the age of the sample may also limit the generalisability of our findings. We examined the accuracy of teacher expectations at ages 11 and 14, but our results may not be transportable to earlier of

later stages of education. Additionally, the participants were educated between 2001 and 2006 and teacher expectations may have change since this period.

Third, the accuracy of teacher expectations may have differed across Maths, English and Science subjects as many participants will have had subject specific teachers. Our decision to combine across these subjects was taken to provide a more accurate measure of the pupil's overall academic performance and reflect any general teacher expectation bias. Furthermore, teacher expectations of pupil achievement were provided as categorical levels for each subject, meaning that there was reduced variation in this measure when compared to point scores used for assessment.

Fourth, many variables were subject to potential measurement error. For example, family socioeconomic position is a complex construct encompassing education, income, wealth and other factors, yet we were only able to proxy this using weekly household income, parental social class and the highest level of maternal education [44,45]. To improve statistical power, we leveraged the larger sample of responses from study mother's reports of their partners occupation and income, but it is likely that the mother reports will have contained greater measurement error than direct partner reports.

Finally, because of the ethnic homogeneity of the ALSPAC sample, the European-centric focus of genetic studies, and the need to exclude non-Europeans due to systematic ancestry differences arising from population stratification (which can induce spurious genotype-phenotype associations) [46–48], we were only able to perform the analyses amongst white participants of European ancestry. Previous studies have demonstrated that the direction of teacher expectation accuracy varies by ethnicity [4]. However, previous work has demonstrated that trait-associated genetic markers do not perform well across ancestral groups [49]. Larger genotyped samples of ethnic minorities are therefore required to explore this issue further.

In conclusion, this study investigated potential patterns of teacher expectation accuracy by socioeconomic, demographic and genetic factors. We found evidence of systematic socioeconomic and genetic patterning in teacher expectations. Pupils from more disadvantaged backgrounds underperformed their teachers' expectations compared to their more advantaged peers, and those with higher genetic liability for educational attainment outperformed their teachers' expectations relative to pupils with lower genetic liability for educational attainment.

# Methods

# Study participants

We used data from the Avon Longitudinal Study of Parents and Children (ALSPAC), a longitudinal birth cohort study based in Bristol, UK. ALSPAC initially recruited 14,541 pregnant women with an expected delivery date between April 1991 and December 1992. When the oldest children were approximately seven years of age, an attempt was made to recruit eligible children that were not included in the original

sample. This resulted in a total eligible sample size of 15,454 pregnancies of which 14,901 children were alive at one year of age. For full cohort details and study design see [43,50]. The ALSPAC cohort was representative of the UK population in 1991 on many criteria, but had underrepresentation of ethnic minorities, single parent families and those of lower socio-economic position. Ethical approval for the study was obtained by the ALSPAC Ethics and Law Committee and the Local Research Ethics Committees. This study was approved by the ALSPAC Executive committee under the project identifier B2193. All methods were performed in accordance with the relevant guidelines and regulations. Informed consent for the use of data collected via questionnaires and clinics was obtained from participants following the recommendations of the ALSPAC Ethics and Law Committee at the time. Consent for biological samples has been collected in accordance with the Human Tissue Act (2004). Questionnaires were completed by study mothers, the child's schoolteachers and headteachers to obtain information relating to family background and the school/classroom. The study website contains details of all the data that is available through a fully searchable data dictionary and variable search tool (see http://www.bristol.ac.uk/alspac/researchers/our-data/). Due to low numbers in the minority ethnicity groups and problems with multi-ancestry genetic analyses (below), all non-White participants were excluded.

# Educational outcomes

# Realised achievement

We used fine graded achievement scores at two of the major "Key Stages" of UK education (Key Stage 2, at ages 7–11, Key Stage 3, at ages 11–14) obtained from the UK National Pupil Database (NPD) through data linkage to the ALSPAC cohort. Achievement scores were determined from examinations at the end of each Key Stage.

# Teacher expected achievement

Teacher expected achievement was available from the NPD as categorical variables indicating the national curriculum level (i.e. 3, 4, 5) that a pupil was expected to achieve for each of Mathematics, English and Science. The average of these three measures was taken and converted to a point score to enable comparability with achievement scores (Supplementary Table S4). All achievement scores were rounded for comparability.

# Accuracy of teacher expectations

To determine the accuracy of teacher expectations we used residuals from a regression of realised achievement (the dependent variable) on teacher expectation scores (the independent variable) at Key Stages 2 and 3. A positive value therefore indicates that a pupil outperformed their teacher's expectation in their examinations. For ease of interpretation, accuracy scores were standardised to follow a normal distribution with mean zero and standard deviation one. This enabled investigation into systematic inaccuracy of teacher expectations against a variety of demographic and teacher characteristics, and exploration of heterogeneity in the teacher expectations across groups.

# Covariates

Information on participant sex at birth and month of birth was obtained from birth records. Month of birth was recoded with September as the first month to represent age in school year (the school year in the UK starts in September). Self-report questionnaires completed by the study mothers during pregnancy provided information on family socioeconomic position. Socioeconomic position was proxied by parental social class based on occupation at cohort member birth, the mother's highest education qualification at cohort member birth, and family income at cohort member age four. Study mothers reported their own and their partners' occupation, with responses coded to the Standard Occupational Classification (SOC) codes and converted to social class based on occupation, with the following seven bands: I (Professional occupations); II (Managerial and technical occupations); III-NM (Skilled non-manual occupations); III-M (Skilled manual occupations); IV (Partly-skilled occupations); V (Unskilled occupations); Armed forces. Armed forces responses to social class were recoded to II due to low number of observations. Where both maternal and paternal social class were available, the highest social class was taken. Mothers highest level of education was categorised as follows: Degree; A-level (a post-compulsory qualification at age 18); O-level (a subject based academic qualification at age 16); Vocational qualifications; Certificate of Secondary Education (CSE, a general qualification at age 16). Family income per week was reported in the following bands: less than £100; £100–199; £200–299; £300–399; £400 or more; Don't know. Responses of "don't know" were coded as missing.

Teachers provided information through self-report questionnaires at Key Stage 2 (age 11) about their gender (categorised as "Male" or "Female"), length of service (less than one year, 1–2 years, 3- years, 10 + years), and the number of pupils per class. School headteachers provided information on the Special Educational Needs (SEN) status of participants, characterised as: 1) Has a statement; 2) Currently being assessed; 3) Not statemented; 4) Has been refused. SEN status of 2, 3 and 4 were re-coded to "Not statemented".

# Genotyping, quality control and imputation

DNA of the ALSPAC children was extracted from blood, cell line and mouthwash samples, then genotyped using reference panels and subjected to standard quality control approaches. ALSPAC children were genotyped using the Illumina HumanHap550 quad chip genotyping platforms by 23andme subcontracting the Wellcome Trust Sanger Institute, Cambridge, UK and the Laboratory Corporation of America, Burlington, NC, US. Standard quality control methods were applied to the resulting genome-wise data. Individuals were excluded based on sex-mismatch, minimal or excessive heterozygosity (< 0.320 and > 0.345), individual missingness greater than 3% and insufficient sample reduction (IBD < 0.8). Multidimensional scaling was used to stratify the population, comparing with Hapmap II (release 22) European descent (CEU), Han Chinese, Japanese and Yoruba reference populations; all individuals with non-European ancestry were removed. SNPs with a minor allele frequency (MAF) of less than 1%, a call rate of less than 95% or evidence for violations of Hardy-Weinberg equilibrium (p-value $< 5 \times 10^{-7}$) were removed. Cryptic relatedness was measured as a proportion of identity-by-descent (IBD > 0.1), with related

participants passing all other quality control thresholds retained in subsequent phasing and imputation, described fully in supporting information. 8,237 children passed these quality control filters.

Children's genotypes were jointly phased and imputed with the genotypes of the ALSPAC mothers (Illumina human660W quad (mothers)), combining 477,482 SNP genotypes which were in common between the samples. SNPs with genotype missingness above 1% were removed due to poor quality (11,396 SNPs removed) and a further 321 participants due to potential ID mismatches. This resulted in a dataset of 17,842 participants containing 465,740 SNPs (112 removed during liftover and 234 were out of Hardy Weinberg Equilibrium after combination). Haplotypes were estimated using ShapeIT (v2.r644), utilizing relatedness during phasing. A phased version of the 1000 genomes reference panel (Phase 1, Version 3) from the Impute2 reference data repository (phased using ShapeIT v2.r644, haplotype release data Dec 2013) was obtained. Imputation of the target data was performed using Impute V2.2.2 against the reference panel (all polymorphic SNPs excluding singletons), using all 2186 reference haplotypes (including non-Europeans). This gave 8,237 eligible children and 8,196 eligible mothers with available genotype data after exclusion of related participants using cryptic relatedness measures described previously.

# Polygenic scores

A PGS was generated from the largest GWAS of educational attainment. ALSPAC participants were excluded from the meta-analysis used to generate the PGS to reduce bias due to overfitting alongside 23andMe participants due to data sharing agreements. The PGS was created using the software package PRSice [51] using all SNPs that were identified to associate with years of education. The scores were calculated as a weighted sum of educational attainment associated SNPs weighted by their effect size. SNPs were clumped and the SNP with the smallest P-value in each 250kb window was retained. All other SNPs in linkage disequilibrium with an $r^2 > 0.1$ were removed.

# Multiple imputation

Due to attrition and item non-response, 2,341 and 3,696 participants had missing data on at least one variable at KS2 and KS3 respectively (**Figure A1**). To increase statistical power and reduce potential selection bias within primary findings due to attrition [52], we conducted multiple imputation by chained equations (MICE) [53] under the assumption that data were missing at random (MAR) conditional upon the data included in the imputation model [54]. To help overcome this assumption, additional covariates were included as predictors in the imputed dataset to incorporate additional information and improve imputation accuracy [53,55,56]. MICE is a method based on data augmentation, iteratively estimating parameters for the distribution of each variable and using this to predict the missing values. All variables except for genotype were imputed and variables that did not follow a normal distribution were transformed. Imputation was conducted in Stata 16 using the *mim* command [57,58]. A total of 100 imputed datasets were generated, and the results were pooled for each regression analyses. We only imputed phenotypic data using this approach (see *Genotyping, quality control and imputation* for imputation of genetic data). The pooled imputed dataset contained 7,465 participants, of whom all had

genetic data. We found little evidence that distributions of the imputed and observed values differed (Table 4). While the proportion of missing data were large in some cases, previous research has demonstrated that this will not bias imputation results and is not a reliable guide for comparing the accuracy between complete case and multiple imputation analyses [59]. Table 4 displays the distributions of key variables across the complete case and multiple imputation samples.

Table 4

**Complete case and imputed summary statistics.** Imputed summary statistics calculated from 100 imputed datasets.

| Variable | Complete case dataset | | Imputed dataset [3] | |
|---|---|---|---|---|
| | N | Mean (SD) | N | Mean (SD) |
| **Gender** | 5,499 | | 7,465 | |
| *Male* | 2,780 | 50.55 | 3,810 | 51.04 |
| *Female* | 2,719 | 49.45 | 3,655 | 48.96 |
| **Month of delivery**[1] | 5,499 | 6.65 (3.72) | 7,465 | 6.63 (3.74) |
| **SEN status** | 5,499 | | 7,465 | |
| *Has a statement* | 64 | 1.16 | 7,368 | 98.70 |
| *Not statemented* | 5,435 | 98.84 | 97 | 1.30 |
| **Mothers highest education** | 5,499 | | 7,465 | |
| *Degree* | 940 | 17.09 | 1,131 | 15.14 |
| *A level* | 1,451 | 26.39 | 1,863 | 24.95 |
| *O level* | 1,977 | 35.95 | 2,663 | 35.68 |
| *Vocational* | 460 | 8.37 | 692 | 9.28 |
| *CSE* | 671 | 12.20 | 1,116 | 14.95 |
| **Parental social class** | 5,499 | | 7,465 | |
| *I* | 868 | 15.78 | 1,088 | 14.57 |
| *II* | 2,457 | 44.68 | 3,247 | 43.51 |
| *III non-manual* | 1,374 | 24.99 | 1,891 | 25.32 |
| *III manual* | 585 | 10.64 | 880 | 11.80 |
| *IV* | 191 | 3.47 | 316 | 4.22 |
| *V* | 24 | 0.44 | 43 | 0.58 |
| **Family income per week** | 5,499 | | 7,465 | |
| *Less than £100* | 270 | 4.91 | 436 | 5.88 |
| *£100-£199* | 710 | 12.91 | 1,040 | 13.92 |

*[1] Where September = 1, October = 2 etc. [2] Information available for KS2 only. [3] Summary statistics calculated from 100 imputed datasets.*

| Variable | Complete case dataset | | Imputed dataset [3] | |
|---|---|---|---|---|
| | N | Mean (SD) | N | Mean (SD) |
| *£200-£299* | 1,498 | 27.24 | 2,011 | 26.94 |
| *£300-£399* | 1,306 | 23.75 | 1,745 | 23.31 |
| *More than £400* | 1,715 | 31.19 | 2,233 | 29.95 |
| **Number of pupils on class register [2]** | 2,914 | 2.83 (0.51) | 7,465 | 2.82 (0.50) |
| **Teacher gender [2]** | 2,898 | | 7,465 | |
| *Male* | 728 | 25.12 | 1,874 | 25.12 |
| *Female* | 2,170 | 74.88 | 5,591 | 74.88 |
| **Length of service [2]** | 2,580 | | 7,465 | |
| *Less than 1 year* | 48 | 1.65 | 123 | 1.64 |
| *1–2 years* | 110 | 3.79 | 264 | 3.57 |
| *3–9 years* | 1,156 | 39.85 | 2,933 | 39.31 |
| *10 or more years* | 1,587 | 54.71 | 4,145 | 55.49 |

*[1] Where September = 1, October = 2 etc. [2] Information available for KS2 only. [3] Summary statistics calculated from 100 imputed datasets.*

# Statistical analysis

We estimated associations between realised achievement and teacher expectations, expectation accuracy and teacher characteristics, and expectation accuracy and polygenic scores using linear regression. We performed separate analyses for Key Stages 2 and 3 to assess teacher expectation accuracy at ages 11 and 14.

We estimated the SNP heritability of the teacher expectation accuracy using genome-wide complex trait analysis (GCTA) with genomic-relatedness-based restricted maximum-likelihood estimation (GREML) [35]. SNP heritability is defined as the proportion of total variation in a phenotype (the teacher expectation accuracy) that can be explained by common genetic variation in all measured SNPs [35], akin to a correlation coefficient [60]. GCTA first estimates the genetic similarity between every pair of unrelated individuals using measured variation across the genome, and compares this similarity with the phenotypic similarity of each pair. If more genetically similar pairs are more phenotypically similar than genetically dissimilar pairs, then the heritability estimate for a phenotype will be higher.

For all genotypic analyses we make two further sample restrictions. First, we restrict the sample to unrelated participants in ALSPAC (less related than 2nd cousins) as indicated by their genotypic

similarity. Second, we restrict our sample to participants of European ancestry only due to poor polygenic score performance in diverse ancestral groups. Further to these selections, we control for the first 20 principal components of population structure in all genotypic analyses to reduce population stratification bias.

## Declarations

# Acknowledgements

The Medical Research Council (MRC) and the University of Bristol support the MRC Integrative Epidemiology Unit [MC_UU_00011/1]. The Wellcome Trust support CJB via a PhD [218495/Z/19/Z]. The Economics and Social Research Council (ESRC) support NMD via a Future Research Leaders grant [ES/N000757/1], a Norwegian Research Council Grant number 295989, and TTM via a postdoctoral fellowship [ES/S011021/1]. No funding body has influenced data collection, analysis or its interpretation. This publication is the work of the authors, who serve as the guarantors for the contents of this paper.

# Author contributions

CJB analysed and cleaned the data, interpreted the results, wrote and revised the manuscript. TTM conceived the study, interpreted results, supervised the analysis and revised the manuscript. NMD supervised the analysis, interpreted the results and revised the manuscript.

# Data availability

The empirical dataset has been archived with the ALSPAC study under the project identifier B2193 and will be made available to individuals who obtain the necessary permissions from the study's executive committee.

# Code availability

All code used to process and analyse the data are available at https://github.com/Ciarrah/tchr_rprtng_ccrcy/.

# Competing interests

The authors declare no competing interests.

## References

1. Südkamp, A., Kaiser, J. & Möller, J. Teachers' Judgments of Students' Academic Achievement BT - Teachers' Professional Development: Assessment, Training, and Learning. in (eds. Krolak-Schwerdt, S., Glock, S. & Böhmer, M.) 5–25(SensePublishers, 2014). doi:10.1007/978-94-6209-536-6_2
2. Benner, A. D. & Mistry, R. S. Congruence of mother and teacher educational expectations and low-income youth's academic competence. *Journal of Educational Psychology*, **99**, 140–153 (2007).
3. Artelt, C. Teacher Judgments and their Role in the Educational Process. *Emerging Trends in the Social and Behavioral Sciences*, **1–16**, https://doi.org/10.1002/9781118900772.etrds0402 (2016).
4. Campbell, T. Stereotyped at Seven? Biases in Teacher Judgement of Pupils' Ability and Attainment. *J. Soc. Policy*, **44**, 517–547 (2015).
5. Hanna, R. & Linden, L. Measuring discrimination in education(2009).
6. Harvey, D. G. & Slatin, G. T. The relationship between child's SES and teacher expectations: A test of the middle-class bias hypothesis. *Soc. Forces*, **54**, 140–159 (1975).
7. Parsons, S. & Hallam, S. The impact of streaming on attainment at age seven: evidence from the Millennium Cohort Study. *Oxford Rev. Educ*, **40**, 567–589 (2014).
8. Meissel, K., Meyer, F., Yao, E. S. & Rubie-Davies, C. M. Subjectivity of teacher judgments: Exploring student characteristics that influence teacher judgments of student ability. *Teach. Teach. Educ*, **65**, 48–60 (2017).
9. Baker, C. N., Tichovolsky, M. H., Kupersmidt, J. B., Voegler-Lee, M. E. & Arnold, D. H. Teacher (mis) perceptions of preschoolers' academic skills: Predictors and associations with longitudinal outcomes. *J. Educ. Psychol*, **107**, 805 (2015).
10. Hinnant, J. B., O'Brien, M. & Ghazarian, S. R. The longitudinal relations of teacher expectations to achievement in the early school years. *J. Educ. Psychol*, **101**, 662 (2009).
11. Airasian, P. W. *Classroom assessment: Concepts and applications* (ERIC, 2001).
12. Goldstein, H. Using pupil performance data for judging schools and teachers: scope and limitations. *Br. Educ. Res. J*, **27**, 433–442 (2001).
13. Resnick, L. B. & Resnick, D. P.. Assessing the thinking curriculum: New tools for educational reform. in Changing assessments37–75(Springer, 1992).
14. Shepard, L. A. Why We Need Better Assessments. *Educ. Leadersh*, **46**, 4–9 (1989).
15. Bosker, R. J., Creemers, B. P. M. & Stringfield, S. *Enhancing educational excellence, equity and efficiency: evidence from evaluations of systems and schools in change* (Springer Science \& Business Media, 1999).
16. Hansen, K. The relationship between teacher perceptions of pupil attractiveness and academic ability. *Br. Educ. Res. J*, **42**, 376–398 (2016).
17. Harlen, W. Trusting teachers' judgement: Research evidence of the reliability and validity of teachers' assessment used for summative purposes. *Res. Pap. Educ*, **20**, 245–270 (2005).
18. Connolly, P. *et al.* The misallocation of students to academic sets in maths: A study of secondary schools in England. *Br. Educ. Res. J*, **45**, 873–897 (2019).

19. Ready, D. D. & Wright, D. L. Accuracy and inaccuracy in teachers' perceptions of young children's cognitive abilities: The role of child background and classroom context. *Am. Educ. Res. J*, **48**, 335–360 (2011).
20. McKown, C. & Weinstein, R. S. Modeling the Role of Child Ethnicity and Gender in Children's Differential Response to Teacher Expectations 1. *J. Appl. Soc. Psychol*, **32**, 159–184 (2002).
21. Wang, S., Rubie-Davies, C. M. & Meissel, K. A systematic review of the teacher expectation literature over the past 30 years. *Educ. Res. Eval*, **24**, 124–179 (2018).
22. Chalmers, T. C., Celano, P., Sacks, H. S. & Smith, H. Jr Bias in treatment assignment in controlled clinical trials. *N. Engl. J. Med*, **309**, 1358–1361 (1983).
23. Brophy, J. E. Research on the self-fulfilling prophecy and teacher expectations. *J. Educ. Psychol*, **75**, 631 (1983).
24. Blatchford, P., Russell, A., Bassett, P., Brown, P. & Martin, C. The effect of class size on the teaching of pupils aged 7–11 years. *Sch. Eff. Sch. Improv*, **18**, 147–172 (2007).
25. Mulholland, L. A. & Berliner, D. C. Teacher Experience and the Estimation of Student Achievement(1992).
26. Elhoweris, H., Mutua, K., Alsheikh, N. & Holloway, P. Effect of children's ethnicity on teachers' referral and recommendation decisions in gifted and talented programs. *Remedial Spec. Educ*, **26**, 25–31 (2005).
27. Bianco, M., Harris, B., Garrison-Wade, D. & Leech, N. Gifted girls: Gender bias in gifted referrals. *Roeper Rev*, **33**, 170–181 (2011).
28. Miller, C. K., McLaughlin, J. A., Haddon, J. & Chansky, N. M. Socioeconomic class and teacher bias.*Psychol. Rep.*(1968).
29. Tenenbaum, H. R. & Ruck, M. D. Are teachers' expectations different for racial minority than for European American students? A meta-analysis. *J. Educ. Psychol*, **99**, 253 (2007).
30. Burgess, S., Greaves, E. T. & Scores Subjective Assessment, and Stereotyping of Ethnic Minorities. *J. Labor Econ*, **31**, 535–576 (2013).
31. Morris, T. T., Davies, N. M., Dorling, D. & Richmond, R. C. & Davey Smith, G. Testing the validity of value-added measures of educational progress with genetic data.*Br. Educ. Res. J.***0**, (2018).
32. Lee, J. J. *et al.* Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nat. Genet*, https://doi.org/10.1038/s41588-018-0147-3 (2018).
33. Harden, K. P. *et al.* Genetic Associations with Mathematics Tracking and Persistence in Secondary School. bioRxiv(2019). doi:10.1101/598532
34. Smith, G. D. *et al.* Clustered environments and randomized genes: a fundamental distinction between conventional and genetic epidemiology. *PLoS Med*, **4**, 1985–1992 (2007).
35. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: A tool for genome-wide complex trait analysis. *Am. J. Hum. Genet*, **88**, 76–82 (2011).

36. Borghans, L., Golsteyn, B. H. H., Heckman, J. & Humphries, J. E. Identification problems in personality psychology. *Pers. Individ. Dif*, https://doi.org/10.1016/j.paid.2011.03.029 (2011).
37. Kuklinski, M. R. & Weinstein, R. S. Classroom and developmental differences in a path model of teacher expectancy effects. *Child Dev*, **72**, 1554–1578 (2001).
38. Malouff, J. M. & Thorsteinsson, E. B. Bias in grading: A meta-analysis of experimental research findings. *Aust. J. Educ*, **60**, 245–256 (2016).
39. Rimfeld, K. *et al.* The stability of educational achievement across school years is largely explained by genetic factors. *npj Sci. Learn*, **3**, 16 (2018).
40. Morris, T. T. & Davies, N. M. & Davey Smith, G. Can education be personalised using pupils' genetic data? *Elife*, **9**, e49962 (2020).
41. Sorhagen, N. S. Early teacher expectations disproportionately affect poor children's high school performance. *J. Educ. Psychol*, **105**, 465 (2013).
42. Jussim, L., Robustelli, S. L. & Cain, T. R. Teacher Expectations and Self-Fulfi lling Prophecies. in Handbook of motivation at school363–394(Routledge, 2009).
43. Boyd, A. *et al.* Cohort Profile: the 'children of the 90s'–the index offspring of the Avon Longitudinal Study of Parents and Children. *Int J Epidemiol*, **42**, 111–127 (2013).
44. Entwislea, D. R. & Astone, N. M. Some practical guidelines for measuring youth's race/ethnicity and socioeconomic status. *Child Dev*, **65**, 1521–1540 (1994).
45. Hauser, R. M. Measuring socioeconomic status in studies of child development. *Child Dev*, **65**, 1541–1545 (1994).
46. Morris, T. T., Davies, N. M., Hemani, G. & Smith, G. D. Population phenomena inflate genetic associations of complex social traits.*Sci. Adv.***6**, (2020).
47. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet*, https://doi.org/10.1038/ng1847 (2006).
48. Novembre, J. *et al.* Genes mirror geography within Europe. *Nature*, https://doi.org/10.1038/nature07331 (2008).
49. Duncan, L. *et al.* Analysis of polygenic risk score usage and performance in diverse human populations. *Nat. Commun*, https://doi.org/10.1038/s41467-019-11112-0 (2019).
50. Fraser, A. *et al.* Cohort profile: The avon longitudinal study of parents and children: ALSPAC mothers cohort. *Int. J. Epidemiol*, **42**, 97–110 (2013).
51. Euesden, J., Lewis, C. M. & O'Reilly, P. F. PRSice: Polygenic Risk Score software., **31**, 1466–1468 (2015).
52. Graham, J. W. Missing data analysis: Making it work in the real world. *Annu. Rev. Psychol*, **60**, 549–576 (2009).
53. Azur, M. J., Stuart, E. A., Frangakis, C. & Leaf, P. J. Multiple imputation by chained equations: what is it and how does it work? *Int. J. Methods Psychiatr. Res*, **20**, 40–49 (2011).

54. He, Y., Zaslavsky, A. M., Landrum, M. B., Harrington, D. P. & Catalano, P. Multiple imputation in a large-scale complex survey: a practical guide. *Stat. Methods Med. Res*, **19**, 653–670 (2010).
55. Schafer, J. L. Multiple imputation in multivariate problems when the imputation and analysis models differ. *Stat. Neerl*, **57**, 19–35 (2003).
56. Barnard, J. & Meng, X. L. Applications of multiple imputation in medical studies: from AIDS to NHANES. *Stat. Methods Med. Res*, **8**, 17–36 (1999).
57. Zhang, Z. Multiple imputation with multivariate imputation by chained equation (MICE) package.*Ann. Transl. Med.***4**, (2016).
58. Galati, J. C., Royston, P. & Carlin, J. B. MIM: Stata module to analyse and manipulate multiply imputed datasets(2013).
59. Madley-Dowd, P., Hughes, R., Tilling, K. & Heron, J. The proportion of missing data should not be used to guide decisions on multiple imputation. *J. Clin. Epidemiol*, https://doi.org/10.1016/j.jclinepi.2019.02.016 (2019).
60. Vinkhuyzen, A. A. E., Wray, N. R., Yang, J., Goddard, M. E. & Visscher, P. M. Estimation and partition of heritability in human populations using whole-genome analysis methods. *Annu. Rev. Genet*, **47**, 75–95 (2013).

# Figures

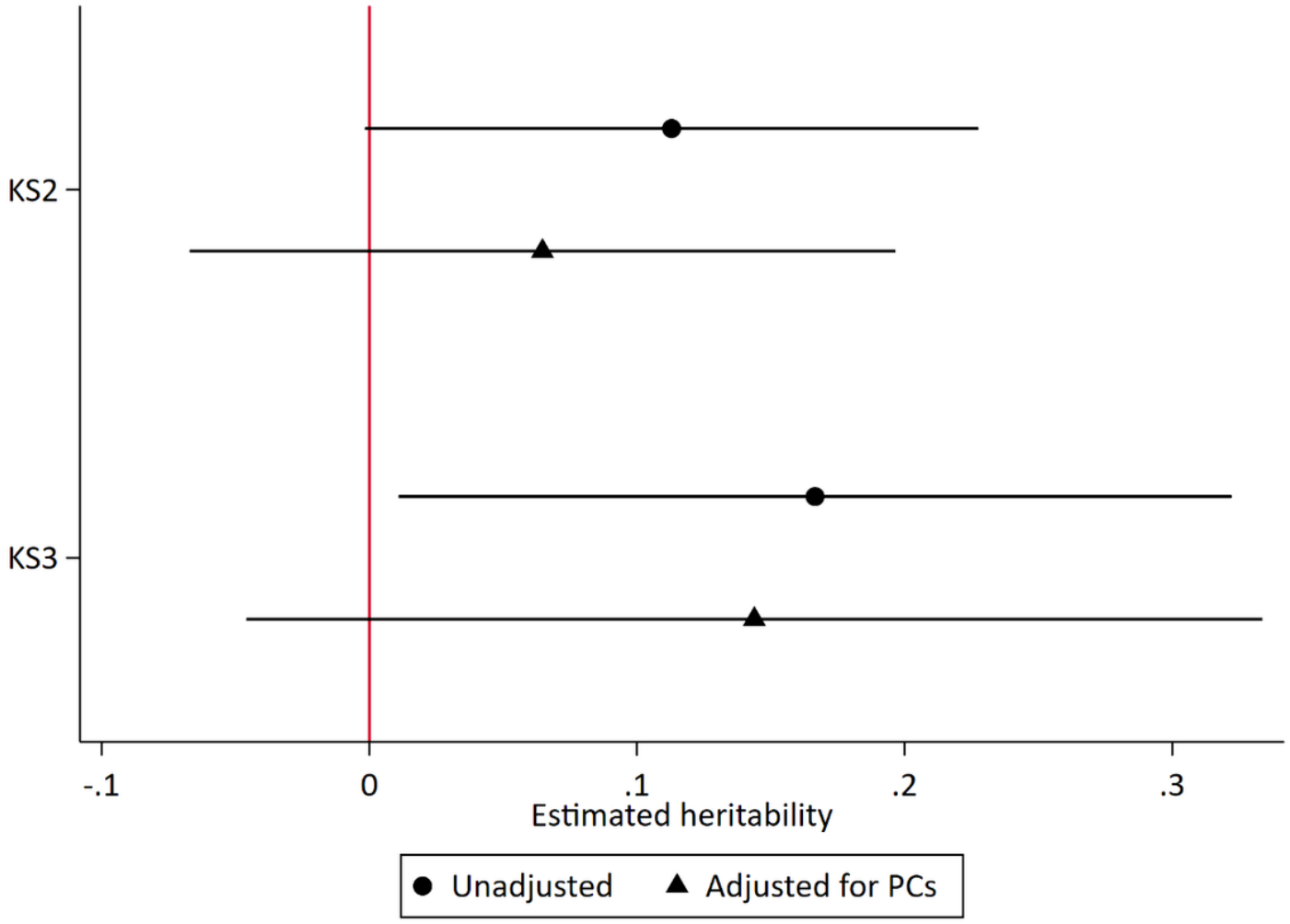


**Figure 1**

Estimated SNP heritability of teacher expectation accuracy at Key Stage 2 (age 11) and Key Stage 3 (age 14). Models control for the first 20 principal components of population structure.

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- teacherbiasalspacsupplement.docx