Author Accepted Manuscript: Morris, T., Davies, N., Dorling, D., Richmond, R. and Davey Smith, G. (2018) Examining the genetic influences on educational attainment and the validity of value-added measures of progress, British Educational Research Journal, forthcoming, accepted July, pre-publication draft version: https://www.biorxiv.org/content/early/2018/01/03/233635

Examining the genetic influences on educational attainment and the validity of value-added measures of progress.

Tim T Morris* [1], Neil M Davies [1], Danny Dorling [2], Rebecca C Richmond [1], and George Davey Smith [1]

1. MRC Integrative Epidemiology Unit, University of Bristol, BS8 2BN, United Kingdom.

2. School of Geography and the Environment, University of Oxford, Oxford, OX1 3QY, United Kingdom.

Corresponding author: Tim T Morris, MRC Integrative Epidemiology Unit, University of Bristol, Oakfield House, Oakfield Grove, BS8 2BN, UK. Email: Tim.Morris@bristol.ac.uk.

The authors declare no conflict of interest. This publication is the work of the authors and Tim Morris will serve as guarantor for the contents of this paper.

Key words: Education; value-added; genetics; ALSPAC.

Word count: 6382

Abstract: 241

Accepted for publication in the British Educational Research Journal on 10 July 2018

Abstract

Value-added measures of educational progress are frequently used by education researchers and policy makers to assess the performance of teachers and schools, influencing performance related pay and position in school league tables. They are designed to control for all underlying differences between pupils and so should provide measures of school and teacher impact upon pupil progress. However, questions have been raised about whether value-added measures are unbiased measures of performance. In this study, we exploit genetic data from a UK birth cohort to investigate how successfully value-added measures control for genetic differences between pupils. We use raw, contextual, and teacher reported value-added measures built from point score test data at ages 11, 14 and 16 and teacher rated ability at ages 11 and 14. Independent sample sizes for analyses range from 4,600 to 6,518. Our findings demonstrate that genetic differences between pupils explain little variation in raw value-added measures but explain up to 20% of the variation in contextual measures that additionally control for background characteristics (95%CI: 6.06% to 35.71%). Value-added measures built from teacher rated ability are more responsive to genetic differences between pupils, with 36.3% of their cross sectional variation being statistically accounted for by genetics (95%CI: 22.8% to 49.8%). In conclusion, our findings provide evidence that value-added measures of educational progress can be influenced by genetic differences between pupils, and therefore may provide a biased measure of school and teacher performance. We include a glossary of genetic terms for educational researchers interested in the use of genetic data in educational research.

Glossary

Allele

Alleles are the different variant forms of genetic variation found at a specific point on a chromosome. Specific alleles associate with different phenotypic traits (e.g. outcomes such as educational attainment).

Allele frequency

Allele frequency is the prevalence of a given allele at a genetic locus in the sample, expressed as a decimal, fraction, or percentage. The allele frequency is reported in terms of the proportion of alleles that are the effect allele (e.g. allele associated with higher levels of education) or in terms of the number of minor, or less frequent alleles (termed *minor allele frequency*).

Assortative mating

Assortative mating refers to the non-random manner in which people sort into partnerships with partners who have more similar social and biological characteristics such as height, education, and personality than would be expected by chance alone. This may be due to assortment based upon partner choice, convergence in characteristics due to interaction with a partner over time, or social homogamy (Robinson et al., 2017). Rates of assortative mating may vary between populations and over time.

Common genetic variation

Common genetic variation refers to all genetic variants across the genome in which the minor or rare allele occurs relatively frequently, i.e. above 1%.

Dynastic effects

Dynastic effects refer to the direct effects of parents' phenotypes on their offspring. An example of this in the education context would be highly educated parents creating a nourishing learning environment for their children via buying books and helping their children learn to read. These effects are sometimes referred to as "genetic nurture" (Kong et al., 2018).

Gene

A stretch of DNA that encodes protein, made up of exons interspersed with introns formed by a distinct sequence of nucleotides constituting a section of a chromosome. This is distinct from a genetic variant, which can occur in a gene (intragenic), or outside of a gene (intergenic).

Genetic relatedness matrix

Identical twins share all their germline genome and have a genetic relatedness of one. Siblings share half their genetic code and have a genetic relatedness of 0.5. If two unrelated individuals drawn at random from the population are more genetically similar than you would expect by chance, they will have a genetic relatedness of greater than zero. If they are less alike than you would expect by chance they will have a genetic relatedness of less than zero. The genetic relatedness matrix stores the genetic relatedness between every pair of individuals in a sample. It can be used to estimate heritability using GCTA (see below).

Genome-wide association study (GWAS)

A genome-wide association study (GWAS) tests the associations of hundreds of thousands of genetic variants and an outcome (the phenotype). Due to the number of associations being tested

simultaneously and therefore issues of multiple hypothesis testing, strict P-value thresholds (conventionally 5×10−8) are used to account for multiple testing. The combination of these strict P-value thresholds and small SNP effect sizes means that GWAS require very large samples. GWAS are typically separated into two parts; an analysis is first performed in a discovery cohort to identify nominally genome-wide significant SNPs, and secondly performed in an independent validation cohort to validate these SNPs.

Genome-wide complex trait analysis (GCTA)

Genome-wide complex trait analysis (GCTA) is a statistical programme and method used to estimate the additive genetic contribution of common SNPs to the heritability of a trait using variance components. GCTA compares the genetic similarity between unrelated individuals and compares it to their similarity on phenotypic traits; where pairs of unrelated individuals are genetically and phenotypically similar this provides evidence that phenotypic variation can be explained by genotypic variation. GCTA studies typically require sample sizes in the many thousands.

Haplotypes

Haplotypes are a specific sequence of alleles that are inherited together from a parent, leading to conserved sequences across generations.

Hardy-Weinberg Equilibrium

Hardy-Weinberg Equilibrium refers to the principle that genetic variation across a population will remain constant (in equilibrium) over generations in the absence of external disruptive or evolutionary factors. External factors that may disrupt Hardy-Weinberg Equilibrium include non-random mating, mutations, genetic drift and natural selection.

Heritability

Heritability is the proportion of total phenotypic variance in a population that can be explained by genetic variance, and therefore ranges from zero (no phenotypic variance explained) to one (all phenotypic variance explained). Broad sense heritability (H^2) is defined as the total proportion of variance in a trait that is due explained by all genetic variation, *inclusive* of additive genetic variance, dominance and epistasis (gene-gene interactions). Narrow sense heritability (h^2) is the proportion of total variance in a trait that is explained by additive genetic variance. Heritability is a population rather than an individual parameter and is specific to both the population and environment under analysis. SNP heritability (h^2_{SNP}) is a measure of narrow sense heritability calculated from a given set of genetic variants (SNPs).

Heterozygosity

Heterozygosity refers to the occurrence of two *different* alleles at a specific genetic locus.

Homozygosity

Homozygosity refers to the occurrence of two of the *same* alleles at a specific genetic locus.

Identity by descent

Where a segment of the genome shared by multiple people is due to inheritance from a common ancestor.

Linkage disequilibrium

Linkage disequilibrium refers to the combination of alleles at two or more loci occurring more frequently than would be expected by chance. This typically occurs for variants in close proximity in the genome. This is in violation of Mendel's second law of inheritance which states that the identity of an allele should provide no information about alleles at other points in the genome.

Phenotype/phenotypic trait

A phenotype is the trait or characteristic of interest, for example: educational attainment, cognition, and socioeconomic position are all phenotypes.

Polygenic trait

Polygenic trait is the term used to refer to phenotypic traits that are influenced by many SNPs, the majority of which can only explain a very small proportion of variance in a trait. Most human behavioural traits that are influenced by our DNA are polygenic, being influenced by a small degree by hundreds or thousands of SNPs.

Polygenic score

A polygenic score (PGS – sometimes referred to as a *polygenic risk* score) is a summed score of the number of alleles associated with a phenotypic trait. These scores are often weighted by the genetic variant's effect size on the phenotype as estimated from a published GWAS. Polygenic scores can use genetic variants that were associated with the phenotype at different P-value thresholds ranging from genome-wide significance ($p<5\times10^{-8}$) to liberal thresholds such as $p<0.5$. Polygenic scores therefore indicate the summed influence that all genetic variants which are identified at a given level of GWAS significance have on a phenotypic trait. Because of power restrictions on the ability of GWAS to identify SNPs with a small effect size, polygenic scores will omit many variants and therefore not provide an estimate of the total genetic impact on a trait.

Population stratification

Population stratification occurs when different subpopulations may have systematic differences in allele frequencies due to ancestral differences, such as non-random mating between sub populations. These differences can occur because of geographical separation. The association of genetic variants and phenotypes can be confounded by population stratification. To control for population stratification, studies that estimate heritability use Principal Components Analysis (PCA)applied to the genome-wide SNP data to infer population structure, and include the resultant principal components as covariates in analysis to account for population specific variations in allele distributions.

Single nucleotide polymorphism (SNP)

A single nucleotide polymorphism (SNP) is a genetic variant of a single base pair at a specific position in the genome.

Introduction

Value-added (VA) measures are frequently used by education researchers and policy makers to assess the performance of teachers and schools, and therefore impact upon performance related pay, position in school league tables, and school accountability (Leckie and Goldstein, 2009; Ray et al., 2009). Because VA measures compare a student's academic performance to their performance at an earlier stage, they are designed to control for between-individual time-invariant differences such as a child's underlying level of ability (McCaffrey et al., 2004). They are therefore considered to provide a reliable measure of educational progress independent of the selection of pupils, background characteristics and innate ability (Chetty et al., 2014a). This makes them a fairer measure of comparison for the effectiveness of teachers and schools than raw attainment scores which are confounded by earlier attainment and can unfairly assess the final grades of schools with more disadvantaged student intakes (Leckie and Goldstein, 2009). Contextual value-added (CVA) measures have also been developed which additionally account for a range of additional time-invariant background factors beyond the teacher or schools control such as gender, ethnicity, special educational needs, and month of birth. There has however been debate over which factors should be adjusted for and the extent to which CVA measures sufficiently adjust for these (Todd and Wolpin, 2003). Despite research demonstrating that children assigned to high-VA teachers outperform children assigned to low-VA teachers (Chetty et al., 2014b) and therefore that VA can correctly identify the most able teachers, there has been criticism of the extent to which they successfully control for time-invariant factors (Gorard et al., 2013; Taylor and Nguyen, 2006).

What can genetic data offer educational research?

Because VA measures are designed to control for all time-invariant factors, they should be robust to genetic differences between individuals. This is because the genetic variants that we inherit are fixed at birth and cannot be altered by teacher or school performance. Genetic data therefore offers a unique opportunity to investigate educational performance in novel ways. Over the past decade there has been a growing interest in the contribution of genetics towards a range of human behavioural outcomes including educational attainment. A number of studies using molecular genetics and twin approaches have provided strong evidence that genetics contribute towards educational attainment (Branigan et al., 2013; Davies et al., 2015; Deary et al., 2007; Krapohl et al., 2014; Selzam et al., 2016). Amongst samples of unrelated individuals, genetic contribution is estimated using a narrow sense heritability statistic, which is defined as the proportion of the total variation in attainment that can be explained by genetic variation. For an accessible review of methods for estimating heritability and the drawbacks involved in such methods see (Tenesa and Haley, 2013). Heritability is therefore just a correlational statistic and does not imply that a behaviour is immutable, nor does it provide any information on exactly *how* and *why* particular genetic variants associate with outcomes. Furthermore, heritability is dependent upon the trait in question, the population being studied, and the spatiotemporal circumstances surrounding the population (Davey Smith, 2011; Plomin et al., 2008). Heritability estimates can therefore be expected to vary to a certain degree across studies, but it remains "the most useful summary statistic for the genetic contribution to… complex [traits]" such as education (Tenesa and Haley, 2013, p. 140). Like social inequality, heritability is a useful measure even if the exact mechanisms which generate it are not fully understood. A complete understanding of education is not possible without understanding the contribution of genetics (or indeed all other factors) and encorporating genetic evidence could profoundly affect our understanding of education. Furthermore, including

genetic data in studies offers the potential to improve estimation of the influence of social factors by reducing residual variation and genetic confounding in statistical models. However, despite increasing sample sizes of genome-wide association studies (GWAS) used to identify genetic variants that associate with educational attainment, we are still only able to account for a minority of genetic effects. This approach to using data treats genetics not as an unmalleable prognosis of educational attainment, but as a tool to more accurately and reliably assess *modifiable* influences on educational attainment.

A large meta-analysis of studies estimated the heritability of educational attainment at 40% (95% CI: 35-45%) (Branigan et al., 2013). There is considerable variation in heritability estimates between twin studies and molecular genetic studies, to an extent due to the former analysing all genetic variance (i.e. including dominance and epistasis) while the latter analyse only additive genetic variance. Using data from the Twins Early Development Study (TEDS), narrow sense heritability of attainment at age 16 has been estimated at 31% (95% CI: 7.5% to 54.5%) (Krapohl and Plomin, 2016), half that of the broad sense heritability estimate in the same sample of 62% (95% CI: 58% to 67%) (Krapohl et al., 2014). While this points to an important contribution of genetics towards educational attainment, it also demonstrates that between 33% and 42% of the variation in educational attainment between individuals in TEDS is due to *non*-genetic (i.e. social) factors. When looking at international or temporal variations in educational outcomes, social factors are likely to be more visible. For example, in the UK attendance to university has increased hugely over the last couple of generations, far quicker than genetic changes across the population would be possible. Holding time and place constant, genetic variation is likely to explain a greater amount of variation in educational outcomes. This reflects a general principle regarding what generates between individual and between group differences (Davey Smith, 2011; Keyes and Galea, 2016).

Educational attainment is influenced by many thousands of genetic variants each of which have a very small effect size rather than a singular (or even small number of) genetic variants which have large effect sizes. The largest genome-wide association study of educational attainment conducted to date identified only 74 genetic variants with a discovery sample size of 293,723, with the single strongest variant estimated to account for only 0.035% of the variation in educational attainment defined by years of education (Okbay et al., 2016). These findings are consistent with a polygenic model in which it is the combination of small effects of many genetic variants that influences a trait.

It is also, of course, important to realise that educational attainment is no a singular construct that can be perfectly measured with a single score. If in a particular society use of imagination was valued above focus on strictly following a set curriculum, then what would be considered a good education would differ. There would very likely still be genetic variants that contribute to some children being more imaginative than others, but these would likely not be the same genetic variants that lead some children to more diligently do their homework and revise for exams as compared to others.

Using genetic data to investigate the validity of VA measures

Despite this evidence, educational research has generally been slow to incorporate genetic information (Plomin and Walker, 2003), though some development has been made (Jerrim et al., 2015). Harnessing genetic information offers novel opportunities to investigate educational phenomena. For example, if the goal of VA measures is to control out between-individual

differences and provide a fair assessment of teachers and schools, then they *must* account for genetic factors which directly influence attainment. Recent genetic analyses applied to VA measures have suggested that they are prone to genetic differences and therefore perform poorly at controlling for time-invariant differences between children. In TEDS the heritability of a value-added measure was estimated at 52% (95% CI: 48%–57%), similar to the heritability as raw attainment scores (Haworth et al., 2011). This implies that VA measures do not provide a fair assessment of teachers and schools but are instead genetically biased and may disadvantage schools by the ability of their intakes. The study by Haworth and colleagues used teacher reported ability instead of the directly assessed attainment scores which are used to inform school league tables and educational policy. This is important because teacher rated ability is likely to be a less accurate measure of student achievement than National Curriculum test-assessed point scores. It will be characterised by greater systematic error due to confounding by teacher-reporting variation on the bias of traits such as ethnicity or physical attractiveness (Burgess and Greaves, 2013; Gershenson et al., 2016; Hansen, 2016), and therefore may induce bias in heritability estimates. The heritability for VA measures built from reading and maths test assessments in the same study were similar though, suggesting that such bias may not be present. However, these assessments used were completed online, differed between the two measurement occasions, and did not align with the National Curriculum assessments and so are likely to contain greater measurement error than official data. Further investigation into the ability of VA measures to control for genetic differences between children is required to assess their suitability for informing policy, assessing teacher performance, or determining school accountability.

In this study we exploit molecular genetic data to first estimate the heritability of educational attainment at three time points throughout the compulsory educational lifecourse, and second investigate the ability of VA measures created from examination assessment data to control for time-invariant between individual genetic differences (one of their desirable properties). If VA measures are estimated to have non-zero heritability it suggests that they are susceptible to the influence of genetics. Throughout this paper we use the term *heritability* to refer to a measure of heritability calculated from a given set of genetic variants (SNPs, or single nucleotide polymorphisms).

Methods and materials

Study sample

Participants were children from the Avon Longitudinal Study of Parents and Children (ALSPAC). Pregnant women were eligible to enrol if they had an expected date of delivery between April 1991 and December 1992 and were resident in the (former) Avon Health Authority area in South West England (for full details of the cohort profile and study design see Boyd et al (2013) and Fraser et al (2013)). The study website contains details of all the data that is available through a fully searchable data dictionary (ALSPAC data dictionary available at http://www.bris.ac.uk/alspac/researchers/data-access/data-dictionary/). The ALSPAC cohort is largely representative of the UK population when compared with 1991 Census data; however there is under representation of some ethnic minorities, single parent families, and those living in rented accommodation. Ethical approval for the study was obtained from the ALSPAC Ethics and Law Committee and the Local Research Ethics Committees.

From the core sample of 14,775 children 14,115 have data on at least one measure of educational attainment. From these children genetic data was available for 7,988 after quality control and removal of related individuals. We use the largest available samples in each of our analyses to increase precision of estimates, regardless of whether a child contributed data to the other analyses.

Genetic data

In short, DNA of the ALSPAC children was extracted from blood, cell line and mouthwash samples, then genotyped using references panels and subjected to standard quality control approaches. In full, the 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. The resulting raw genome-wide data were subjected to standard quality control methods. Individuals were excluded on the basis of gender mismatches; minimal or excessive heterozygosity (where a genetic locus contains two different alleles); disproportionate levels of individual missingness (>3%) and insufficient sample replication (identity by descent (IBD) < 0.8). Population stratification was assessed by multidimensional scaling analysis and compared 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 of < 1%, a call rate of < 95% or evidence for violations of Hardy-Weinberg Equilibrium (HWE) (P < 5E-7) were removed. Cryptic relatedness (where two individuals in the sample are close relatives, but this is unknown) was measured as proportion of IBD (> 0.1). Related subjects that passed all other quality control thresholds were retained during subsequent phasing and imputation. 9,115 subjects and 500,527 SNPs 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 subjects due to potential ID mismatches. This resulted in a dataset of 17,842 subjects containing 6,305 duos and 465,740 SNPs (112 were removed during 'liftover' and 234 were out of HWE after combination). Haplotypes (a group of alleles inherited together) were estimated using ShapeIT (v2.r644) which uses relatedness during phasing. We obtained 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 date December 2013). Imputation of the target data was performed using Impute V2.2.2 against the 1000 genomes reference panel (Phase 1, Version 3) (all polymorphic SNPs excluding singletons), using all 2186 reference haplotypes (including non-Europeans). This gave 8,237 eligible children with available genotype data after exclusion of related subjects using cryptic relatedness measures.

Education data

Educational attainment

Our measures of educational attainment are average fine graded point scores at each of the major stages of education in the UK; Key Stage 2 assessed at age 11; Key Stage 3 assessed at age 14; and Key Stage 4 assessed at age 16. Point scores were used to obtain a richer measure of a child's attainment than level bandings, with the distributions of the raw scores presented in the supplementary material section 2. Scores for the ALSPAC cohort were obtained through data linkage

to the UK National Pupil Database (NPD), which represents the most accurate record of individual educational attainment available in the UK. We extracted all scores from the Key Stage 4 database as this includes attainment at earlier Key Stages and provides a larger sample size than the earlier databases.

Value-added measures

We use two sets of value-added (VA) measures in our analyses. First, we calculated a raw VA score as the difference between standardised point scores at different Key Stages. This VA measure can be considered the child's cohort specific VA score as it is based upon the rank ordering of the child in the cohort at each occasion. Second, contextual value-added (CVA) measures were extracted from the NPD linked to ALSPAC participants. The CVA measures - using the example of a CVA score between ages 11 and 14 - are calculated as the difference between a child's given exam score (age 14) and the score that would be predicted from that child's previous Key Stage exam score (age 11). The models used to calculate CVA measures are estimated within a multilevel framework whereby students are nested within schools and the intercept is permitted to vary across schools. The CVA models also account for gender; special educational needs; eligibility for free school meals (a proxy for low income); first language; school mobility; ethnicity; month of birth; an indicator of whether a child has been in care; and residential area level deprivation. We present results for both the CVA and VA measures throughout our analyses.

We also use a value-added measure based on teacher-assessed ability of children (TAVA). Teachers are required to grade their students at multiple time points in English, Mathematics and Science on a scale of 1 to 8. These grades reflect the level at which a teacher deems a student to be working, with higher levels reflecting students working at a more advanced stage. Because the levels run throughout a child's educational career they can be compared at different time points to assess progress. The TAVA measure we use was calculated as the difference between the mean level of English, Maths and Science at ages 11 and 14, thus representing progress during these years. Teacher-reported ability is not available in the NPD at age 16, meaning that our TAVA measure only covers the one educational period.

Educational attainment polygenic score

An educational attainment polygenic score was generated from genetic data based upon the 74 independent SNPs identified at genome-wide significance ($p = 5 \times 10^{-08}$) in the largest GWAS of education to date (Okbay et al., 2016). Each genetic variant was weighted by the effect size of the variant in the replication cohort of the meta-analysis, the UK Biobank, and these doses were summed using the allelic scoring function in PLINK (version 1.9) (Purcell et al., 2007). The resulting polygenic score provides an estimate of the summed influence that all genetic variants which are identified at GWAS significance have on educational attainment.

Statistical analysis

To estimate the proportion of variation in educational attainment and VA measures that can be attributed to common genetic variation (the SNP heritability) we run a series of univariate analyses using generalized restricted maximum likelihood (GREML) in the software package GCTA. GCTA uses measured SNP level variation to estimate the genetic similarity between every pair of unrelated individuals in the sample and compares this to their phenotypic similarity. Unrelated participants

(less related than 2nd cousins) are determined using the ALSPAC Genetic Relatedness Matrices (GRMs). Our univariate analyses are specified as follows:

$$y = X\beta + g + \epsilon$$

where y is the heritability of a phenotype, X is a series of covariates, g is a normally distributed random effect with variance σ_g^2, and ϵ is residual error with variance σ_ϵ^2. Heritability is defined as the proportion of total phenotypic variance (genetic variance plus residual variance) that can be attributed to common genetic variation:

$$h_{SNP}^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_\epsilon^2}$$

If across the sample genetically similar pairs are more phenotypically similar than genetically dissimilar pairs then heritability estimates of the phenotype will be higher. Population stratification can bias the estimates of heritability. Population stratification can occur if different subpopulations in the sample have systematic differences in allele frequencies due to ancestral differences. To control for these population specific variations in allele distributions, the first 20 principal components of inferred population structure are included as covariates in analyses. All continuous variables were inverse-normally-rank-transformed to have a normal distribution, a requirement of GCTA. Power calculations are presented in supplementary material section 3, Table S2. Briefly, we are suitably powered to detect heritability estimates greater than 0.15. All code used to generate the results in this study are available from https://github.com/timtmorris/VA-heritability.

Results

Table 1 displays the number of study children who provide information for each of the analyses and descriptive statistics for each of the variables used. Sample size is higher for the raw attainment measures but reflects some data loss in the raw VA and CVA measures due to missing attainment and background factors respectively. The CVA measures have higher standard deviations than the raw VA measures because they are measured using differences between predicted point scores and realised point scores, whereas the raw VA measures use differences in standardised score differences.

Table 1: Descriptive statistics for children included in the analyses.

	n	Mean	SD
KS2 points	6132	28.04	3.85
KS3 points	4960	35.97	6.19
KS4 points	6518	39.89	9.48
KS 2-3 VA score	4904	0.06	0.42
KS 2-4 VA score	6088	-0.03	0.63
KS 3-4 VA score	4924	-0.07	0.49

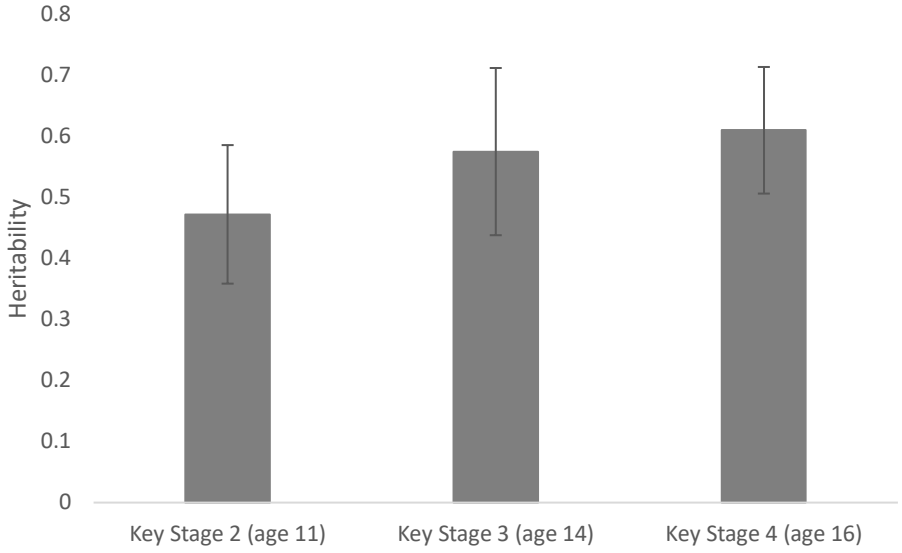
KS 2-3 CVA score	4600	0.10	2.52
KS 2-4 CVA score	6028	0.98	56.33
KS 3-4 CVA score	4914	1.08	45.21

SD, standard deviation; VA, raw value-added; CVA, contextual value-added.

Educational attainment

The SNP heritabilities of educational attainment measured at each of the Key Stages are presented in Figure 1. Heritability rises with age from 47.3% (95%CI: 35.9 to 58.7) at age 11 to 57.6% (95%CI: 43.9 to 71.3) at age 14 and 61.1% (95%CI: 50.7 to 71.5) at age 16. This suggests that in the ALSPAC sample genetic variation contributes towards around half of the total variance in educational attainment using end of Key Stage exam scores. These heritabilities are higher than would be expected given the estimated heritability of 40.0% (95% CI: 35.3-44.7) from a meta-analysis of educational attainment (Branigan et al., 2013).

Figure 1: Heritability of Key Stage average point score attainment.



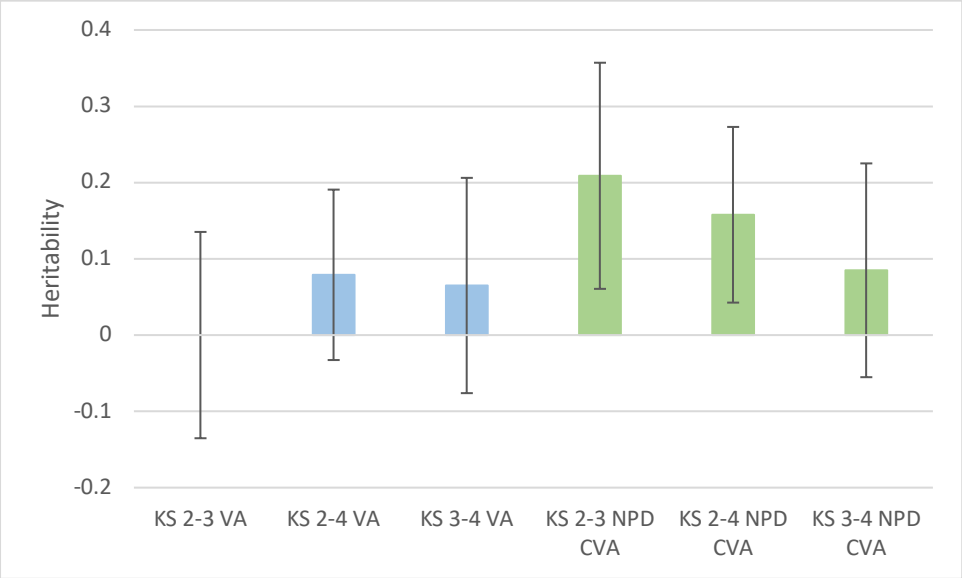
Sample sizes: KS2=6132; KS3=4960; KS4=6518. See supplementary material section 4 for full model results.

Value-added measures

Figure 2 displays the heritability of VA measures, which would be zero if the VA measures control for all time-invariant genetic differences between children. There little evidence that common genetic variation can explain the raw VA measures, far smaller than for the raw attainment scores at each Key Stage (KS). The raw VA measures from ages 11 to 14, 11 to 16, and 14 to 16 show SNP heritabilities of <0.1% (95%CI: -13.5 to 13.5), 7.9% (95%CI: -3.3 to 19.0), and 6.5% (95%CI: -7.6 to 20.6) respectively. The estimate of heritability for the KS2-3 raw VA score is constrained to zero due because the predicted heritability is -0.04, which is outside of the bounds of zero to one. The

heritability estimates for the CVA measures though are consistently higher than those for the corresponding raw VA measures. There is evidence that the CVA scores are heritable for the period of age 11 to 14 (h^2_{SNP}: 20.09%; 95%CI: 6.06 to 35.71), and the period of age 11 to 16 (h^2_{SNP}: 15.77%; 95%CI: 4.26 to 27.29). Heritability estimates for the period from age 14 to 16 are lower with little strong evidence for a heritable component (h^2_{SNP}: 8.49%; 95%CI: -5.52 to 22.51). These results consistently suggest that relatively little of the variation in raw VA measures – the unadjusted progress that students make from one Key Stage to another – can be attributed to common genetic variation. However, they also suggest that the variation in some CVA measures – those that adjust for background factors – *can* in part be attributed to common genetic variation.

Figure 2: Heritability of Value-added measures.



KS, key stage; VA, raw value-added; CVA, contextual value-added. Children are aged 11 at KS2, 14 at KS3 and 16 at KS4. Sample sizes: KS 2-3 VA=4904; KS 2-4 VA=6088; KS 3-4 VA=4924; KS 2-3 CVA=4600; KS 2-4 CVA=6028; KS 3-4 CVA=4914. See supplementary material section 4 for full model results.

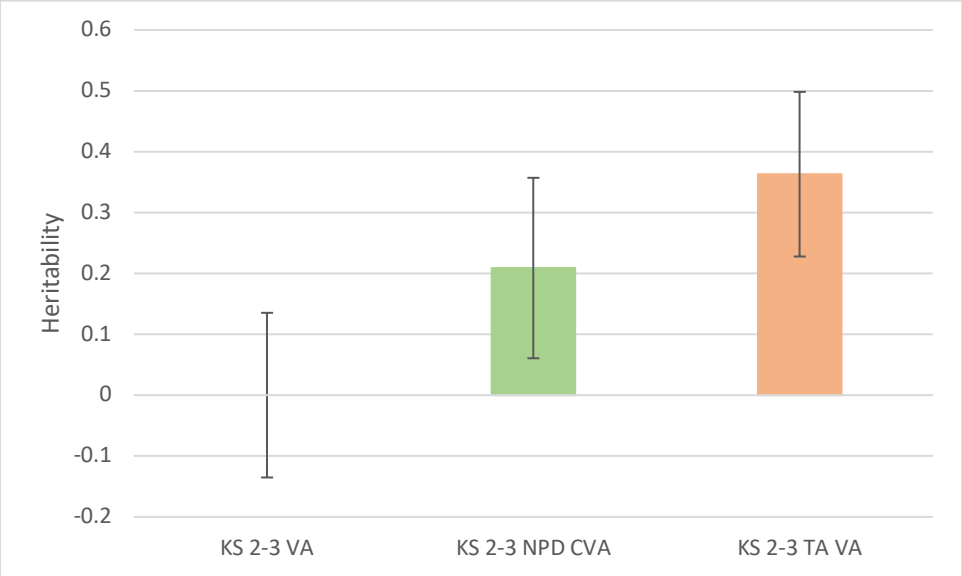
To explore this, we ran a series of simulations to determine scenarios in which CVA measures may be more genetically biased than raw VA measures (see supplementary material section 1). The simulations demonstrate that CVA measures overstate heritability more than raw VA measures where the baseline input score contains measurement error (VA h^2_{SNP}= 9.99%; CVA h^2_{SNP}=16.66% given measurement error of 0.25).

These results contrast with those from a previous study which estimated heritability of VA measures using teacher assessed ability at 50% (Haworth et al., 2011). Teacher assessed ability is a good proxy for true student ability; the correlations between teacher assessed ability and point scores at KS2 and KS3 are 0.884 and 0.921 respectively. However, teacher assessed ability may be confounded by teacher-report bias which will not be present in directly assessed examination scores. If the baseline measure of achievement is measured with error due to teacher-reported bias, then it will not fully control for prior achievement. As our simulations demonstrate, this may inflate the estimated

heritability. Furthermore, because teacher assessed ability is measured using fewer bandings than the rich point scores used by exams assessments, the variability of scores is lower as children cannot be differentiated within bands. Where the baseline measures are more constrained they will be less effective at controlling for initial differences and may further lead to upward bias in heritability estimates. It is therefore possible that the high heritability observed in the previous study may be due in part to imprecision or teacher bias in the baseline measurement compared to point scores that are traditionally used in educational research.

We investigated if the discrepancy between our results and those from the previous study may have reflected genuine differences (a sample issue) or differences caused by alternative methods of assessment (a measurement issue). A teacher assessed value-added (TAVA) measure similar to that used in the previous study was created for the period of education from age 11 to age 14 (teacher rated ability was unavailable at age 16). The ages that the teacher assessments were made were similar to TEDS (ages 10 and 12). Figure 3 presents the GCTA heritability results of our TAVA measure compared to the raw VA and CVA measures at this age. The results suggest a moderate amount of heritability in the TAVA measure of 36.3% (95%CI: 22.8 to 49.8). This exceeds the heritability point estimates presented in Figures 1 and 2 and suggests that VA measures using teacher rated ability are likely to reflect both student progress and genetic differences between students. The correlations between the TAVA measure and the VA and CVA measures are 0.333 and 0.405 respectively, far lower than the correlation between the VA and CVA measures of 0.919. We return to this in the discussion.

Figure 3: Heritability of KS2-3 Value-added measures.

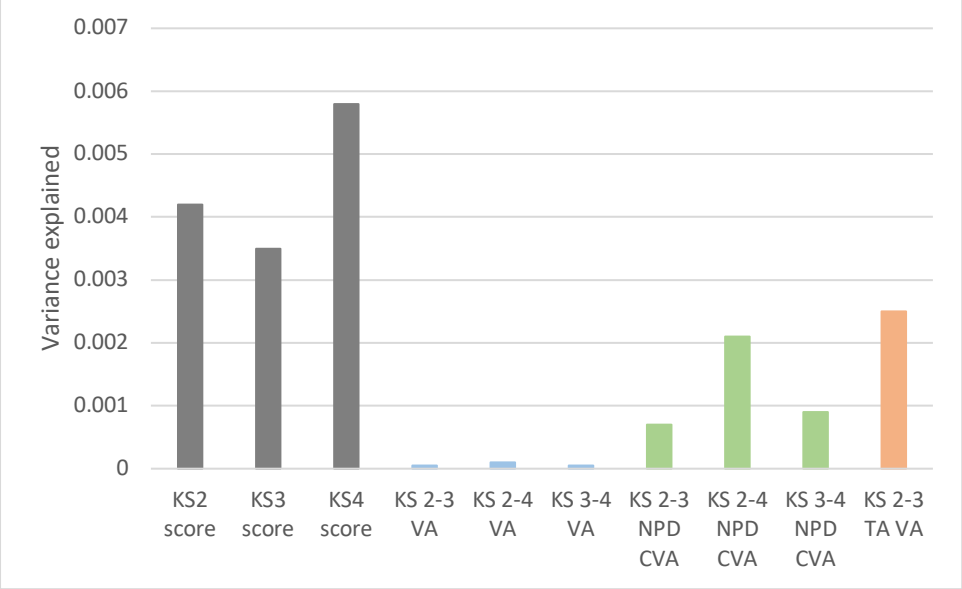


KS, key stage; VA, raw value-added; CVA, contextual value-added; TA VA, teacher assessed value-added. Children are aged 11 at KS2, 14 at KS3 and 16 at KS4. Sample sizes: KS 2-3 VA=4904; KS 2-3 CVA=4600; KS 2-3 TA VA=5070. See supplementary material section 4 for full model results.

Polygenic score results

The final stage of the analysis was to investigate the amount of variance in educational attainment and VA measures that an educational attainment polygenic score (EA PGS) predicts. The results (Figure 4) show that the EA PGS accounts for a small but detectable proportion of variance in educational attainment at each Key Stage, varying from 0.35% to 0.58%. The EA PGS accounts for negligible variation in the raw VA measures (≤0.01%), providing further evidence that these successfully account for time-invariant differences between children and are not influenced by genetic factors. Consistent with the GCTA results the EA PGS predicts a small but detectable amount of variation in the CVA measures (0.07% to 0.21%), though this is far smaller than the raw attainment scores. Again, this suggests that these adjusted CVA measures are more strongly associated with genetic differences than raw VA measures. The EA PGS explains a greater amount of variation in the TAVA measure (0.25%) than the raw VA (<0.01%) and CVA (0.07%) measures for this period, closer to the heritability of the age 11 and 14 point scores themselves. Given that the EA PGS only includes 74 variants which each have a small effect size, this provides further evidence that value-added measures based upon teacher rated ability are likely to be considerably biased by common genetic variation associated with educational attainment.

Figure 4: Proportions of variance in outcomes attributed to the EA PGS.



KS, key stage; VA, raw value-added; CVA, contextual value-added; TA VA, teacher assessed value-added. Children are aged 11 at KS2, 14 at KS3 and 16 at KS4. Sample sizes: KS2 score =6132; KS3 score =4960; KS4 score=6518; KS 2-3 VA=4904; KS 2-4 VA=6088; KS 3-4 VA=4924; KS 2-3 CVA=4600; KS 2-4 CVA=6028; KS 3-4 CVA=4914; KS 2-3 TA VA=5070. See supplementary material section 4 for full model results.

Discussion

Attainment throughout the educational lifecourse

Our genome wide heritability estimates using all common genetic variants provide molecular genetic evidence for a substantial heritable component of educational attainment in ALSPAC, conforming to previous findings from other samples. We find that the heritability of educational attainment increases from 47% at age 11 to 61% at age 16, suggesting genetics explains a greater proportion of the variation in outcomes at age 16 than earlier ages. This is important because the age 16 exams play a large role in setting a child's future chances in further education and the labour market. The differences across ages are small and could be due to various factors. For example, parents, teachers and others may work harder at later ages to make the environment more homogenous than at earlier ages. Attitudinal behaviours may also be partly responsible because older children have more freedom to choose their own educational effort with less parental influence. It is also important to note that in some subjects such as Maths, children are compelled to enter different tiers based upon ability as assessed after the age 14 exams which will have placed ceiling and floor caps on the scores that they could attain in the age 16 examinations.

Our estimated heritability is higher than the 40.0% (95% CI: 35.3 to 44.7) estimated using meta-analyses (Branigan et al., 2013). Furthermore, it is more similar to the broad than narrow sense heritability estimated in TEDS: 62% (95% CI: 58% to 67%) and 31% (95% CI: 7.48% to 54.52%) respectively (Krapohl et al., 2014; Krapohl and Plomin, 2016). It is possible that this may reflect a true difference whereby the heritability of point scores is higher than that of years of education (as used in the meta-analysis) or grades (as used in TEDS). However, it must be noted that this may also reflect the fact that the ALSPAC cohort is more spatiotemporally homogenous than other cohorts used in the meta-analysis; all the children in ALSPAC were born within three years in the same geographical location and they mostly experienced the same school system, albeit in a very socially divided largely urban area. This means there may be a smaller set of environmental factors influencing our results which could result in reduced environmental variation. This will in turn increase the residual variation in the phenotype that can be attributed to common genetic variation and the resulting heritability estimate.

Value-added measures

We found that very little of the variation in raw value-added measures built from point score data could be explained by common genetic variation. Surprisingly, we found that these raw VA measures outperformed CVA measures which additionally control for background factors. Our results imply that raw VA measures may be less prone to between-individual genomic differences than CVA measures and therefore offer a more valid measure for the value added by schools or teachers to a child's education. CVA measures appear to be susceptible to genetic differences between children and may therefore not offer fair assessments of the contribution that teachers and schools make towards a child's educational progress. Our simulations suggest that the reduced performance of CVA measures could be due to measurement error in the baseline scores. Because KS2 and KS3 point scores are determined from a smaller range of subjects than KS4 point scores, they are likely to provide less precise measures of overall academic ability which results in a baseline measure containing greater measurement error. This measurement error is then inflated when additional contextual factors are accounted for, resulting in CVA measures being more biased by genetic differences than raw VA measures.

It has been argued that adjusting for factors can increase bias where input measures are broad or crude proxies, or related to parental choice (Todd and Wolpin, 2003), and our results provide genetic evidence that supports this. Our simulations may also help to explain why VA measures built from teacher reported ability further overestimate heritability. They are likely to contain greater measurement error than VA measures built from point scores and therefore demonstrate higher heritability. Careful consideration must therefore be taken when constructing VA measures, and caution should be exercised when using them in educational research or for policy purposes. While the raw measures in our sample appear to be largely independent of genetic background and may provide an *indication* of the contribution that teachers and schools make to a child's educational progress, our findings demonstrate that contextual measures will provide an unfair reflection of teachers and school and could unfairly penalise those depending upon the intake that they receive. Our results only suggest that CVA measures are less effective at controlling for *genetic* differences between children than raw VA measures. It is possible however that they may be overall less biased and more effective at controlling for other between-individual social and demographic factors.

The VA measures created from teacher rated ability for the same children show considerably higher heritability, of around 36%, suggesting that teacher rated ability may be more prone to between-individual genomic differences than official test point scores data. However, it must be stressed that the error around the point estimates only provides strong evidence for a difference between the raw and teacher reported VA measures. Several factors may account for this higher heritability estimate. For example, there is evidence that teacher rated ability may be influenced by heritable factors such as attractiveness which could lead to confounding bias (Clifford and Walster, 1973; Talamas et al., 2016). Furthermore, teacher rated ability is likely to be a less precise measure of ability than end of key stage point scores due to its reduced richness as a measure of attainment and therefore reduced imposed variability. As with the CVA measures, if teacher rated prior attainment is measured with bias or low precision then heritability estimates may be upwardly biased. A policy implication of this is that baseline measurements need to be accurately measured to eliminate bias. Teacher assessments are unlikely to be sufficiently precise to eliminate this bias. The difference in correlations between the TAVA measure and the VA/CVA measures serves to highlight that while teachers are generally adept at assessing their pupil's overall ability, there is enough error in their assessments that the value-added measures are biased, leading to a heritable component.

Our teacher rated VA estimate contrasts with that from the only previous study examining genetic influences on VA measures which estimated heritability in TEDS at 50% (Haworth et al., 2011). It is possible this discrepancy arises because the two cohorts represent different samples. However, the two cohorts are drawn from the same country and the age of participants differs only by a maximum of 6 years. It is therefore unlikely that this leads to the observed differences (assuming teachers of the geographically concentrated ALSPAC study were as accurate at determining student ability as the general population of teachers). Second, it is possible that the teacher reported ability collected by TEDS may have been less accurate or subject to greater bias than those linked from the UK National Pupil Database (NPD) to the ALSPAC cohort. It is likely that a child's level would have been decided with more care in the official (and contractually required) National Curriculum reports used by ALSPAC than the optional survey used by TEDS. Third, the ALSPAC analytical sample is likely to be more genotypically homogenous than the TEDS sample because it is geographically concentrated rather than national. Fourth, it is also possible that the discrepancy between the TEDS and ALSPAC samples is due to the age at which teacher reported ability was measured. ALSPAC measures were

taken at ages 11 and 14 (the end of Key Stages 2 and 3) while the TEDS measures were taken at ages 10 and 12. It is important to note that in the study by Haworth et al (2011), VA measures built from test assessments had similar heritability to those built from teacher rated data, providing a suggestion that teacher rated measures may not underperform compared to test data. However, the test data used were not drawn from NPD data but instead from online assessments, used only reading and Maths ability, differed between the two measurement occasions, and the timing did not align with National Curriculum Key Stages. This is likely to have resulted in greater measurement error which our simulations demonstrated could lead to inflation in the resulting VA measures. Furthermore, twin studies may be more susceptible to teacher biases of appearance (Hansen, 2016) than studies of unrelated individuals because of the similarity of appearance between twins. It is also possible that this discrepancy in findings is at least in part due to differences between GCTA and twin models, however the discrepancy is likely too large to be fully accounted for by these model differences. Ultimately, further work replicating the GCTA approach is required on other datasets such as TEDS to resolve these differences and determine if they are due to differences in modelling approach or differences in the data measures used.

Polygenic scores

We found that the EA PGS explained between 0.35% and 0.58% of the variation in educational attainment depending on the stage of education. The EA PGS explains far less variation in exam scores than the GCTA estimates because it uses a set of only 74 genetic variants that associate with education at genome-wide levels of significance. This is a tiny proportion of the genome wide variants used by GCTA, resulting in lower explanatory power. This is demonstrated by two recent studies using TEDS data; a genome wide score using 108,737 SNPs explained 9% of the variance in age 16 grades (Selzam et al., 2016) whereas a score using only 5,733 SNPs explained 1.5% of the variance in the same trait (Krapohl and Plomin, 2016). While lowering the potential explanatory power of the score, using only the 74 genome-wide significant variants ensures that our score contains only the variants that are robustly associated with educational attainment. Regarding VA measures, the polygenic score results corroborated with the GCTA results. The EA PGS explained negligible variance in the raw VA measures (≤0.01%) but a greater amount of variance in the CVA measures (0.07% to 0.21%) and the teacher rated VA measure (0.25%). These results further demonstrate that value-added measures which adjust for background variables or those built from teacher rated measures of ability may be confounded by genetics, even when only accounting for 74 variants which associate with educational attainment.

Limitations

Our results provide only estimates of the variance in educational attainment and value-added measures across the lifecourse, and do not imply that common genetic variants determine the educational attainment of an individual. The major limitation with this work relates to the potential of GCTA to overestimate heritability, which can occur where model assumptions, particularly that of even linkage disequilibrium between SNPs, are violated (Speed et al., 2012). Furthermore heritability estimates from GCTA are sensitive to the sampling of participants, the accuracy of phenotypic measurement, and the structure of the genetic relatedness matrix underlying the data (Krishna Kumar et al., 2016). These limitations however have been strongly refuted by the authors of GCTA (Yang et al., 2017, 2016). Nevertheless, our estimates are comparable to those previously conducted using twin designs despite SNP heritability typically being lower than that derived from twin studies,

raising the possibility that our high heritability estimates may suffer from over inflation. One further issue that may lead to overestimation in GCTA heritability estimates is that of dynastic effects, where the parental phenotype/genotype directly affects the offspring's outcomes through the creation of specific types of environments. Such indirect genetic effects, termed 'genetic nurture' or 'environmental bias', have recently been demonstrated to upwardly bias GCTA heritability estimates of educational attainment because the methods used here are unable to distinguish between direct and indirect genetic effects (Kong et al., 2017; Young et al., 2017). Our heritability estimates for educational attainment point scores will also be susceptible to bias by assortative mating, whereby parents non-randomly select partners based upon level of education, as demonstrated by previous work (Robinson et al., 2017). There may also be unobserved differences between individuals (residual population structure) biasing our results; we attempted to account for this by using the first twenty principle components of population structure however, we cannot be certain that these will correct for all differences. This limitation could be overcome with genotypic data on mother-father-offspring trios and future studies should exploit the growing availability of data to investigate this hypothesis. Given that data on teacher reported ability was only available for ages 11 and 14 we were unable to examine if the bias in value-added measures based upon teacher reported ability was consistent between the ages of 11 to 16 and 14 to 16. Future studies with teacher reported ability at multiple timepoints should examine any such variation by age. As with all measures based upon educational attainment, random measurement error at the individual level will exist within the data (such as a child suffering illness at the time of examination). However, these random changes will likely provide only a minimal amount of bias at the aggregate level given our sample size. Conversely, teacher rated measures could be susceptible to longer term factors that may impact a child's educational performance such as parental illness. Finally, it is inevitable that our measure of value-added will still contain measurement error. Random measurement error will not be related to genetic (or other) underlying factors and could therefore bias heritability estimates towards the null. However, it is unlikely that our raw VA measure will suffer from heavy bias because the NPD Key Stage examinations represent the most accurate objective assessment of a child's educational ability.

Concluding remarks

In conclusion, our results demonstrate that common genetic variation contributes towards around half of the total variance in educational attainment measured by exam scores throughout the compulsory educational lifecourse in the ALSPAC sample. Our results also suggest that raw value-added measures are robust to genomic differences between children but that contextual value-added measures which further control for additional background factors and those built from teacher reported ability may be genetically biased. These value-added measures should therefore be used with caution in educational research and policy as they have the potential to provide unfair assessment and accountability of teachers or schools, and they may bias performance and position in school league tables.

References

Boyd, A., Golding, J., Macleod, J., Lawlor, D.A., Fraser, A., Henderson, J., Molloy, L., Ness, A., Ring, S., Davey Smith, G., 2013. 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. doi:10.1093/ije/dys064

Branigan, A.R., Mccallum, K.J., Freese, J., 2013. Variation in the heritability of educational attainment: An international meta-analysis. Soc. Forces 92, 109–140. doi:10.1093/sf/sot076

Burgess, S., Greaves, E., 2013. Test Scores, Subjective Assessment, and Stereotyping of Ethnic Minorities. J. Labor Econ. 31, 535–576. doi:10.1086/669340

Chetty, R., Friedman, J.N., Rockoff, J.E., 2014a. Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. Am. Econ. Rev. 104, 2593–2632. doi:10.1257/aer.104.9.2593

Chetty, R., Friedman, J.N., Rockoff, J.E., 2014b. Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood. Am. Econ. Rev. 104, 2633–2679. doi:10.1257/aer.104.9.2633

Clifford, M.M., Walster, E., 1973. The Effect of Physical Attractiveness on Teacher Expectations. Sociol. Educ. 46, 248. doi:10.2307/2112099

Davey Smith, G., 2011. Epidemiology, epigenetics and the "Gloomy Prospect": Embracing randomness in population health research and practice. Int. J. Epidemiol. 40, 537–562. doi:10.1093/ije/dyr117

Davies, N.M., Hemani, G., Timpson, N.J., Windmeijer, F., Davey Smith, G., 2015. The role of common genetic variation in educational attainment and income: evidence from the National Child Development Study. Sci. Rep. 5, 16509.

Deary, I.J., Strand, S., Smith, P., Fernandes, C., 2007. Intelligence and educational achievement. Intelligence 35, 13–21. doi:10.1016/j.intell.2006.02.001

Fraser, A., Macdonald-wallis, C., Tilling, K., Boyd, A., Golding, J., Davey smith, G., Henderson, J., Macleod, J., Molloy, L., Ness, A., Ring, S., Nelson, S.M., Lawlor, D.A., 2013. Cohort profile: The avon longitudinal study of parents and children: ALSPAC mothers cohort. Int. J. Epidemiol. 42, 97–110. doi:10.1093/ije/dys066

Gershenson, S., Holt, S.B., Papageorge, N.W., 2016. Who believes in me? The effect of student–teacher demographic match on teacher expectations. Econ. Educ. Rev. 52, 209–224. doi:10.1016/j.econedurev.2016.03.002

Gorard, S., Hordosy, R., Siddiqui, N., 2013. How unstable are "school effects" assessed by a value-added technique? Int. Educ. Stud. 6, 1–9. doi:10.5539/ies.v6n1p1

Hansen, K., 2016. The relationship between teacher perceptions of pupil attractiveness and academic ability. Br. Educ. Res. J. 42, 376–398. doi:10.1002/berj.3227

Haworth, C.M.A., Asbury, K., Dale, P.S., Plomin, R., 2011. Added value measures in education show genetic as well as environmental influence. PLoS One 6. doi:10.1371/journal.pone.0016006

Jerrim, J., Vignoles, A., Lingam, R., Friend, A., 2015. The socio-economic gradient in children's reading skills and the role of genetics. Br. Educ. Res. J. 41, 6–29. doi:10.1002/berj.3143

Keyes, K., Galea, S., 2016. Population Health Science. Oxford University Press.

Kong, A., Thorleifsson, G., Frigge, M.L., Vilhjalmsson, B., Young, A.I., Thorgeirsson, T.E.,

Benonisdottir, S., Oddsson, A., Halldorsson, B. V., Masson, G., Gudbjartsson, D.F., Helgason, A., Bjornsdottir, G., Thorsteinsdottir, U., Stefansson, K., 2017. The nature of nurture: effects of parental genotypes. doi.org 219261. doi:10.1101/219261

Kong, A., Thorleifsson, G., Frigge, M.L., Vilhjalmsson, B.J., Young, A.I., Thorgeirsson, T.E., Benonisdottir, S., Oddsson, A., Halldorsson, B. V, Masson, G., Gudbjartsson, D.F., Helgason, A., Bjornsdottir, G., Thorsteinsdottir, U., Stefansson, K., 2018. The nature of nurture: Effects of parental genotypes. Science (80-.). 359, 424–428. doi:10.1126/science.aan6877

Krapohl, E., Plomin, R., 2016. Genetic link between family socioeconomic status and children/'s educational achievement estimated from genome-wide SNPs. Mol Psychiatry.

Krapohl, E., Rimfeld, K., Shakeshaft, N.G., Trzaskowski, M., McMillan, A., Pingault, J.-B., Asbury, K., Harlaar, N., Kovas, Y., Dale, P.S., Plomin, R., 2014. The high heritability of educational achievement reflects many genetically influenced traits, not just intelligence. Proc. Natl. Acad. Sci. 111, 15273–15278. doi:10.1073/pnas.1408777111

Krishna Kumar, S., Feldman, M.W., Rehkopf, D.H., Tuljapurkar, S., 2016. Limitations of GCTA as a solution to the missing heritability problem. Proc. Natl. Acad. Sci. 113, E61–E70. doi:10.1073/pnas.1520109113

Leckie, G., Goldstein, H., 2009. The limitations of using school league tables to inform school choice. J. R. Stat. Soc. Ser. A Stat. Soc. 172, 835–851. doi:10.1111/j.1467-985X.2009.00597.x

McCaffrey, D.F., Lockwood, J.R., Koretz, D., Louis, T. a., Hamilton, L., 2004. Models for Value-Added Modeling of Teacher Effects. J. Educ. Behav. Stat. 29, 67–101. doi:10.3102/10769986029001067

Okbay, A., Beauchamp, J.P., Fontana, M.A., Lee, J.J., Pers, T.H., Rietveld, C.A., Turley, P., Chen, G.-B., Emilsson, V., Meddens, S.F.W., Oskarsson, S., Pickrell, J.K., Thom, K., Timshel, P., de Vlaming, R., Abdellaoui, A., Ahluwalia, T.S., Bacelis, J., Baumbach, C., Bjornsdottir, G., Brandsma, J.H., Pina Concas, M., Derringer, J., Furlotte, N.A., Galesloot, T.E., Girotto, G., Gupta, R., Hall, L.M., Harris, S.E., Hofer, E., Horikoshi, M., Huffman, J.E., Kaasik, K., Kalafati, I.P., Karlsson, R., Kong, A., Lahti, J., Lee, S.J. van der, DeLeeuw, C., Lind, P.A., Lindgren, K.-O., Liu, T., Mangino, M., Marten, J., Mihailov, E., Miller, M.B., van der Most, P.J., Oldmeadow, C., Payton, A., Pervjakova, N., Peyrot, W.J., Qian, Y., Raitakari, O., Rueedi, R., Salvi, E., Schmidt, B., Schraut, K.E., Shi, J., Smith, A. V., Poot, R.A., St Pourcain, B., Teumer, A., Thorleifsson, G., Verweij, N., Vuckovic, D., Wellmann, J., Westra, H.-J., Yang, J., Zhao, W., Zhu, Z., Alizadeh, B.Z., Amin, N., Bakshi, A., Baumeister, S.E., Biino, G., Bønnelykke, K., Boyle, P.A., Campbell, H., Cappuccio, F.P., Davies, G., De Neve, J.-E., Deloukas, P., Demuth, I., Ding, J., Eibich, P., Eisele, L., Eklund, N., Evans, D.M., Faul, J.D., Feitosa, M.F., Forstner, A.J., Gandin, I., Gunnarsson, B., Halldórsson, B. V., Harris, T.B., Heath, A.C., Hocking, L.J., Holliday, E.G., Homuth, G., Horan, M.A., Hottenga, J.-J., de Jager, P.L., Joshi, P.K., Jugessur, A., Kaakinen, M.A., Kähönen, M., Kanoni, S., Keltigangas-Järvinen, L., Kiemeney, L.A.L.M., Kolcic, I., Koskinen, S., Kraja, A.T., Kroh, M., Kutalik, Z., Latvala, A., Launer, L.J., Lebreton, M.P., Levinson, D.F., Lichtenstein, P., Lichtner, P., Liewald, D.C.M., Cohort Study, L., Loukola, A., Madden, P.A., Mägi, R., Mäki-Opas, T., Marioni, R.E., Marques-Vidal, P., Meddens, G.A., McMahon, G., Meisinger, C., Meitinger, T., Milaneschi, Y., Milani, L., Montgomery, G.W., Myhre, R., Nelson, C.P., Nyholt, D.R., Ollier, W.E.R., Palotie, A., Paternoster, L., Pedersen, N.L., Petrovic, K.E., Porteous, D.J., Räikkönen, K., Ring, S.M., Robino, A., Rostapshova, O., Rudan, I., Rustichini, A., Salomaa, V., Sanders, A.R., Sarin, A.-P., Schmidt, H., Scott, R.J., Smith, B.H., Smith, J.A., Staessen, J.A., Steinhagen-Thiessen, E., Strauch, K., Terracciano, A., Tobin, M.D., Ulivi, S., Vaccargiu, S., Quaye, L., van Rooij, F.J.A., Venturini, C., Vinkhuyzen, A.A.E., Völker, U., Völzke, H., Vonk, J.M., Vozzi, D., Waage, J., Ware, E.B., Willemsen, G., Attia, J.R., Bennett, D.A., Berger, K., Bertram, L., Bisgaard, H., Boomsma, D.I., Borecki, I.B., Bültmann, U., Chabris, C.F., Cucca, F.,

712 Cusi, D., Deary, I.J., Dedoussis, G. V., van Duijn, C.M., Eriksson, J.G., Franke, B., Franke, L.,
713 Gasparini, P., Gejman, P. V., Gieger, C., Grabe, H.-J., Gratten, J., Groenen, P.J.F., Gudnason, V.,
714 van der Harst, P., Hayward, C., Hinds, D.A., Hoffmann, W., Hyppönen, E., Iacono, W.G.,
715 Jacobsson, B., Järvelin, M.-R., Jöckel, K.-H., Kaprio, J., Kardia, S.L.R., Lehtimäki, T., Lehrer, S.F.,
716 Magnusson, P.K.E., Martin, N.G., McGue, M., Metspalu, A., Pendleton, N., Penninx, B.W.J.H.,
717 Perola, M., Pirastu, N., Pirastu, M., Polasek, O., Posthuma, D., Power, C., Province, M.A.,
718 Samani, N.J., Schlessinger, D., Schmidt, R., Sørensen, T.I.A., Spector, T.D., Stefansson, K.,
719 Thorsteinsdottir, U., Thurik, A.R., Timpson, N.J., Tiemeier, H., Tung, J.Y., Uitterlinden, A.G.,
720 Vitart, V., Vollenweider, P., Weir, D.R., Wilson, J.F., Wright, A.F., Conley, D.C., Krueger, R.F.,
721 Davey Smith, G., Hofman, A., Laibson, D.I., Medland, S.E., Meyer, M.N., Yang, J., Johannesson,
722 M., Visscher, P.M., Esko, T., Koellinger, P.D., Cesarini, D., Benjamin, D.J., 2016. Genome-wide
723 association study identifies 74 loci associated with educational attainment. Nature 533, 539–
724 542. doi:10.1038/nature17671

725 Plomin, R., DeFries, J., McClearn, G., McGuffin, P., 2008. Behavioral Genetics, 5th Editio. ed. Worth,
726 New York.

727 Plomin, R., Walker, S.O., 2003. Genetics and educational psychology. Br. J. Educ. Psychol. 73, 3–14.
728 doi:10.1348/000709903762869888

729 Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de
730 Bakker, P.I.W., Daly, M.J., Sham, P.C., 2007. PLINK: A Tool Set for Whole-Genome Association
731 and Population-Based Linkage Analyses. Am. J. Hum. Genet. 81, 559–575. doi:10.1086/519795

732 Ray, A., Mccormack, T., Evans, H., 2009. Value Added in English Schools. Educ. Financ. Policy 4, 415–
733 438. doi:10.1162/edfp.2009.4.4.415

734 Robinson, M.R., Kleinman, A., Graff, M., Vinkhuyzen, A.A.E., Couper, D., Miller, M.B., Peyrot, W.J.,
735 Abdellaoui, A., Zietsch, B.P., Nolte, I.M., van Vliet-Ostaptchouk, J. V., Snieder, H., Alizadeh, B.Z.,
736 Boezen, H.M., Franke, L., Harst, P. van der, Navis, G., Rots, M., Snieder, H., Swertz, M.,
737 Wolffenbuttel, B.H.R., Wijmenga, C., Abecasis, G.R., Absher, D., Alavere, H., Albrecht, E., Allen,
738 H.L., Almgren, P., Amin, N., Amouyel, P., Anderson, D., Arnold, A.M., Arveiler, D., Aspelund, T.,
739 Asselbergs, F.W., Assimes, T.L., Atalay, M., Attwood, A.P., Atwood, L.D., Bakker, S.J.L., Balkau,
740 B., Balmforth, A.J., Barlassina, C., Barroso, I., Basart, H., Bauer, S., Beckmann, J.S., Beilby, J.P.,
741 Bennett, A.J., Ben-Shlomo, Y., Bergman, R.N., Bergmann, S., Berndt, S.I., Biffar, R., Di Blasio,
742 A.M., Boehm, B.O., Boehnke, M., Boeing, H., Boerwinkle, E., Bolton, J.L., Bonnefond, A.,
743 Bonnycastle, L.L., Boomsma, D.I., Borecki, I.B., Bornstein, S.R., Bouatia-Naji, N., Boucher, G.,
744 Bragg-Gresham, J.L., Brambilla, P., Bruinenberg, M., Buchanan, T.A., Buechler, C., Cadby, G.,
745 Campbell, H., Caulfield, M.J., Cavalcanti-Proença, C., Cesana, G., Chanock, S.J., Chasman, D.I.,
746 Chen, Y.-D.I., Chines, P.S., Clegg, D.J., Coin, L., Collins, F.S., Connell, J.M., Cookson, W., Cooper,
747 M.N., Croteau-Chonka, D.C., Cupples, L.A., Cusi, D., Day, F.R., Day, I.N.M., Dedoussis, G. V., Dei,
748 M., Deloukas, P., Dermitzakis, E.T., Dimas, A.S., Dimitriou, M., Dixon, A.L., Dörr, M., van Duijn,
749 C.M., Ebrahim, S., Edkins, S., Eiriksdottir, G., Eisinger, K., Eklund, N., Elliott, P., Erbel, R.,
750 Erdmann, J., Erdos, M.R., Eriksson, J.G., Esko, T., Estrada, K., Evans, D.M., de Faire, U., Fall, T.,
751 Farrall, M., Feitosa, M.F., Ferrario, M.M., Ferreira, T., Ferrières, J., Fischer, K., Fisher, E., Fowkes,
752 G., Fox, C.S., Franke, L., Franks, P.W., Fraser, R.M., Frau, F., Frayling, T., Freimer, N.B., Froguel,
753 P., Fu, M., Gaget, S., Ganna, A., Gejman, P. V., Gentilini, D., Geus, E.J.C., Gieger, C., Gigante, B.,
754 Gjesing, A.P., Glazer, N.L., Goddard, M.E., Goel, A., Grallert, H., Gräßler, J., Grönberg, H., Groop,
755 L.C., Groves, C.J., Gudnason, V., Guiducci, C., Gustafsson, S., Gyllensten, U., Hall, A.S., Hall, P.,
756 Hallmans, G., Hamsten, A., Hansen, T., Haritunians, T., Harris, T.B., van der Harst, P.,
757 Hartikainen, A.-L., Hassanali, N., Hattersley, A.T., Havulinna, A.S., Hayward, C., Heard-Costa,
758 N.L., Heath, A.C., Hebebrand, J., Heid, I.M., den Heijer, M., Hengstenberg, C., Herzig, K.-H.,
759 Hicks, A.A., Hingorani, A., Hinney, A., Hirschhorn, J.N., Hofman, A., Holmes, C.C., Homuth, G.,

760 Hottenga, J.-J., Hovingh, K.G., Hu, F.B., Hu, Y.-J., Huffman, J.E., Hui, J., Huikuri, H., Humphries,
761 S.E., Hung, J., Hunt, S.E., Hunter, D., Hveem, K., Hyppönen, E., Igl, W., Illig, T., Ingelsson, E.,
762 Iribarren, C., Isomaa, B., Jackson, A.U., Jacobs, K.B., James, A.L., Jansson, J.-O., Jarick, I., Jarvelin,
763 M.-R., Jöckel, K.-H., Johansson, Å., Johnson, T., Jolley, J., Jørgensen, T., Jousilahti, P., Jula, A.,
764 Justice, A.E., Kaakinen, M., Kähönen, M., Kajantie, E., Kanoni, S., Kao, W.H.L., Kaplan, L.M.,
765 Kaplan, R.C., Kaprio, J., Kapur, K., Karpe, F., Kathiresan, S., Kee, F., Keinanen-Kiukaanniemi,
766 S.M., Ketkar, S., Kettunen, J., Khaw, K.-T., Kiemeney, L.A., Kilpeläinen, T.O., Kinnunen, L.,
767 Kivimaki, M., Kivmaki, M., Van der Klauw, M.M., Kleber, M.E., Knowles, J.W., Koenig, W., Kolcic,
768 I., Kolovou, G., König, I.R., Koskinen, S., Kovacs, P., Kraft, P., Kraja, A.T., Kristiansson, K.,
769 Krjutškov, K., Kroemer, H.K., Krohn, J.P., Krzelj, V., Kuh, D., Kulzer, J.R., Kumari, M., Kutalik, Z.,
770 Kuulasmaa, K., Kuusisto, J., Kvaloy, K., Laakso, M., Laitinen, J.H., Lakka, T.A., Lamina, C.,
771 Langenberg, C., Lantieri, O., Lathrop, G.M., Launer, L.J., Lawlor, D.A., Lawrence, R.W., Leach,
772 I.M., Lecoeur, C., Lee, S.H., Lehtimäki, T., Leitzmann, M.F., Lettre, G., Levinson, D.F., Li, G., Li, S.,
773 Liang, L., Lin, D.-Y., Lind, L., Lindgren, C.M., Lindström, J., Liu, J., Liuzzi, A., Locke, A.E., Lokki, M.-
774 L., Loley, C., Loos, R.J.F., Lorentzon, M., Luan, J., Luben, R.N., Ludwig, B., Madden, P.A., Mägi, R.,
775 Magnusson, P.K.E., Mangino, M., Manunta, P., Marek, D., Marre, M., Martin, N.G., März, W.,
776 Maschio, A., Mathieson, I., McArdle, W.L., McCaroll, S.A., McCarthy, A., McCarthy, M.I.,
777 McKnight, B., Medina-Gomez, C., Medland, S.E., Meitinger, T., Metspalu, A., van Meurs, J.B.J.,
778 Meyre, D., Midthjell, K., Mihailov, E., Milani, L., Min, J.L., Moebus, S., Moffatt, M.F., Mohlke,
779 K.L., Molony, C., Monda, K.L., Montgomery, G.W., Mooser, V., Morken, M.A., Morris, A.D.,
780 Morris, A.P., Mìhleisen, T.W., Mìller-Nurasyid, M., Munroe, P.B., Musk, A.W., Narisu, N., Navis,
781 G., Neale, B.M., Nelis, M., Nemesh, J., Neville, M.J., Ngwa, J.S., Nicholson, G., Nieminen, M.S.,
782 Njølstad, I., Nohr, E.A., Nolte, I.M., North, K.E., Nöthen, M.M., Nyholt, D.R., O'Connell, J.R.,
783 Ohlsson, C., Oldehinkel, A.J., van Ommen, G.-J., Ong, K.K., Oostra, B.A., Ouwehand, W.H.,
784 Palmer, C.N.A., Palmer, L.J., Palotie, A., Paré, G., Parker, A.N., Paternoster, L., Pawitan, Y.,
785 Pechlivanis, S., Peden, J.F., Pedersen, N.L., Pedersen, O., Pellikka, N., Peltonen, L., Penninx, B.,
786 Perola, M., Perry, J.R.B., Person, T., Peters, A., Peters, M.J., Pichler, I., Pietiläinen, K.H., Platou,
787 C.G.P., Polasek, O., Pouta, A., Power, C., Pramstaller, P.P., Preuss, M., Price, J.F., Prokopenko, I.,
788 Province, M.A., Psaty, B.M., Purcell, S., Pìtter, C., Qi, L., Quertermous, T., Radhakrishnan, A.,
789 Raitakari, O., Randall, J.C., Rauramaa, R., Rayner, N.W., Rehnberg, E., Rendon, A., Ridderstråle,
790 M., Ridker, P.M., Ripatti, S., Rissanen, A., Rivadeneira, F., Rivolta, C., Robertson, N.R., Rose,
791 L.M., Rudan, I., Saaristo, T.E., Sager, H., Salomaa, V., Samani, N.J., Sambrook, J.G., Sanders, A.R.,
792 Sandholt, C., Sanna, S., Saramies, J., Schadt, E.E., Scherag, A., Schipf, S., Schlessinger, D.,
793 Schreiber, S., Schunkert, H., Schwarz, P.E.H., Scott, L.J., Shi, J., Shin, S.-Y., Shuldiner, A.R.,
794 Shungin, D., Signorini, S., Silander, K., Sinisalo, J., Skrobek, B., Smit, J.H., Smith, A.V., Smith,
795 G.D., Snieder, H., Soranzo, N., Sørensen, T.I.A., Sovio, U., Spector, T.D., Speliotes, E.K.,
796 Stančáková, A., Stark, K., Stefansson, K., Steinthorsdottir, V., Stephens, J.C., Stirrups, K., Stolk,
797 R.P., Strachan, D.P., Strawbridge, R.J., Stringham, H.M., Stumvoll, M., Surakka, I., Swift, A.J.,
798 Syvanen, A.-C., Tammesoo, M.-L., Teder-Laving, M., Teslovich, T.M., Teumer, A., Theodoraki, E.
799 V., Thomson, B., Thorand, B., Thorleifsson, G., Thorsteinsdottir, U., Timpson, N.J., Tönjes, A.,
800 Tregouet, D.-A., Tremoli, E., Trip, M.D., Tuomi, T., Tuomilehto, J., Tyrer, J., Uda, M.,
801 Uitterlinden, A.G., Usala, G., Uusitupa, M., Valle, T.T., Vandenput, L., Vatin, V., Vedantam, S., de
802 Vegt, F., Vermeulen, S.H., Viikari, J., Virtamo, J., Visscher, P.M., Vitart, V., Van Vliet-
803 Ostaptchouk, J. V., Voight, B.F., Vollenweider, P., Volpato, C.B., Völzke, H., Waeber, G., Waite,
804 L.L., Wallaschofski, H., Walters, G.B., Wang, Z., Wareham, N.J., Watanabe, R.M., Watkins, H.,
805 Weedon, M.N., Welch, R., Weyant, R.J., Wheeler, E., White, C.C., Wichmann, H.-E., Widen, E.,
806 Wild, S.H., Willemsen, G., Willer, C.J., Wilsgaard, T., Wilson, J.F., van Wingerden, S.,
807 Winkelmann, B.R., Winkler, T.W., Witte, D.R., Witteman, J.C.M., Wolffenbuttel, B.H.R., Wong,
808 A., Wood, A.R., Workalemahu, T., Wright, A.F., Yang, J., Yarnell, J.W.G., Zgaga, L., Zhao, J.H.,
809 Zillikens, M.C., Zitting, P., Zondervan, K.T., Medland, S.E., Martin, N.G., Magnusson, P.K.E.,
810 Iacono, W.G., McGue, M., North, K.E., Yang, J., Visscher, P.M., 2017. Genetic evidence of

assortative mating in humans. Nat. Hum. Behav. 1, 16. doi:10.1038/s41562-016-0016

Selzam, S., Krapohl, E., von Stumm, S., O'Reilly, P.F., Rimfeld, K., Kovas, Y., Dale, P.S., Lee, J.J., Plomin, R., 2016. Predicting educational achievement from DNA. Mol Psychiatry.

Speed, D., Hemani, G., Johnson, M.R., Balding, D.J., 2012. Improved heritability estimation from genome-wide SNPs. Am. J. Hum. Genet. 91, 1011–1021. doi:10.1016/j.ajhg.2012.10.010

Talamas, S.N., Mavor, K.I., Perrett, D.I., 2016. Blinded by Beauty: Attractiveness Bias and Accurate Perceptions of Academic Performance. PLoS One 11, e0148284.

Taylor, J., Nguyen, A.N., 2006. An Analysis of the Value Added by Secondary Schools in England: Is the Value Added Indicator of Any Value?*. Oxf. Bull. Econ. Stat. 68, 203–224. doi:10.1111/j.1468-0084.2006.00159.x

Tenesa, A., Haley, C.S., 2013. The heritability of human disease: estimation, uses and abuses. Nat. Rev. Genet. 14, 139–149. doi:10.1038/nrg3377

Todd, P., Wolpin, K., 2003. On the Specification and Estimation of the Production Function for Cognitive Achievement. Econ. J. 113, F3–F33. doi:10.1111/1468-0297.00097

Yang, J., Lee, S.H., Wray, N.R., Goddard, M.E., Visscher, P.M., 2016. Commentary on "Limitations of GCTA as a solution to the missing heritability problem," bioRxiv. doi:10.1101/036574

Yang, J., Zeng, J., Goddard, M.E., Wray, N.R., Visscher, P.M., 2017. Concepts, estimation and interpretation of SNP-based heritability. Nat. Genet. doi:10.1038/ng.3941

Young, A.I., Frigge, M.L., Gudbjartsson, D.F., Thorleifsson, G., Bjornsdottir, G., Sulem, P., Masson, G., Thorsteinsdottir, U., Stefansson, K., Kong, A., 2017. Estimating heritability without environmental bias. Doi.Org 218883. doi:10.1101/218883

Acknowledgements

We are extremely grateful to all the families who took part in this study, the midwives for their help in recruiting them, and the whole ALSPAC team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists and nurses. Ethical approval for the study was obtained from the ALSPAC Ethics and Law Committee and the Local Research Ethics Committees. The UK Medical Research Council and the Wellcome Trust (Grant ref: 102215/2/13/2) and the University of Bristol provide core support for ALSPAC. The Medical Research Council (MRC) and the University of Bristol fund the MRC Integrative Epidemiology Unit [MC_UU_12013/1, MC_UU_12013/9]. NMD is supported by the ESRC via a Future Research Leaders Fellowship [ES/N000757/1]. The funders had no role in the conceptualization, design, data collection, analysis, decision to publish, or preparation of the manuscript.

Author contributions

NMD, DD and GDS conceived the study; TTM conducted the analysis; TTM drafted the manuscript. All authors contributed to the final version.

Competing interests

The authors declare no competing interests.