Frequently asked questions (FAQs) to accompany the paper: Predicting Educational Achievement from DNA published in Molecular Psychiatry.

This document was prepared by the authors and the press team at King’s College, London. It aims to address common questions about the background, methods and implications of the study. For further information about the study or for additional questions, please contact the corresponding author, Robert Plomin (robert.plomin@kcl.ac.uk)

**Background to the study**

*What was the aim of the study?*

The aim of this study was to predict educational achievement and educationally relevant traits directly from DNA and to examine gene-environment interplay, by using a method called *Polygenic Scoring.*

*What do we already know about the genetic influence on educational achievement?*

We know from twin studies, which compare genetic similarity between twins to estimate the influences of genetic and environmental effects, that about 60% of the differences between individuals in educational achievement are due to genetic differences1.

Knowing that educational achievement is largely influenced by genetics helps us to better understand differences in students’ performance in the classroom. It shows us that these differences are not just due to environmental factors such as school or teaching quality, but also due to genetic factors.

The twin method provides an upper-bound estimate of genetic influence on a trait measured in a specific population, however, it does not give us any information about the actual genetic variants that contribute to a trait, nor does it tell us about an individual’s genetic risk and resilience for the trait.

One of the ways to look at the specific genetic variants acting on a trait is to perform genome-wide association studies (GWAS). From this we can then calculate genetic scores, so called *polygenic scores*, for each individual that estimate genetic propensities for a specific trait.

**Methods mentioned and used in the study**

*What is a genome-wide association study (GWAS)?*

A GWAS aims to find specific genetic variants that are linked to a particular trait. These genetic variants are also called ‘single nucleotide polymorphisms’ (SNPs). These are the most common type of genetic variation across individuals, in which a single base pair or ‘nucleotide’ differs; for example, some people may have a ‘G’ (guanine) at a specific position in the genome, others may instead have an ‘A’ (adenine).

Results from different GWAS on many traits have shown that there are no SNPs of large effect. Therefore, we can conclude that the overall genetic influence of a trait is made up by the joint effects of thousands of SNPs each with very small influence.

Our study was based on a recent GWAS that examined almost 10 million SNPs and identified 74 SNPs that were significantly associated with years of completed education2. ‘Years of education’ can be seen as a proxy measure for educational achievement. For example, in the UK, having spent 20 years in education probably indicates a doctoral degree, whereas having spent 10 years in education indicates that the individual did not continue past compulsory education.

As genetic effects are very small, findings from this GWAS showed that each education-associated SNP only explained a maximum of 0.04% of the variation in years of education across individuals – meaning that a single SNP could explain less than 0.04 % in the differences between people’s education duration. However, when all SNPs were used together, they predicted up to 4% of the variance across individuals in independent samples, highlighting the accumulative effects of SNPs. However, 4% leaves most of the genetic variance of 60% unexplained.

Because human traits are so complex and influenced by thousands of SNPs of very small effect, it is useful to consider the joint effects of all these trait-associated SNP and not just those that are found to be “significant” through GWAS. This idea forms the basis of the polygenic score method.

*What is a polygenic score?*

Recent advances in genetic research have made it possible to calculate *polygenic scores* for genotyped individuals (those who have DNA information available). This genetic score is an aggregate of thousands of SNPs that are associated with a heightened risk or resilience for a particular trait. As we know from GWAS, individual SNPs are not a good predictor themselves, but as a composite (or polygenic score) the joint effect of SNPs produce a reliable predictor of the trait.

To calculate these polygenic scores, information about the trait-SNP association is taken from GWAS analysis, usually carried out in very large samples of tens, sometime hundreds, of thousands of individuals. This information is then used to score individuals in a different sample. This is done by inspecting each person’s genetic information to identify how many of these trait-associated alleles for a SNP this person carries in their DNA. Depending on the individual’s genotype, they can carry 0, 1 or 2 of the alleles of the SNP that is associated with the trait. Some of the SNPs are more strongly associated with the trait, so they are given more weight than other SNPs. The number of trait-associated SNPs (0,1 or 2) is then summed to form the polygenic score. Therefore, individuals with high polygenic scores carry many of the trait-associated SNPs, and individuals with low polygenic scores carry few of the trait-associated SNPs.

These scores can then be used to explore links with other traits that have been measured in the sample.

**Sample measures and analysis:**

*Who did you test?*

We tested participants from the Twins Early Development Study (TEDS), which recruited over 15,000 twin pairs born in the UK between 1994-963. The sample is representative of British families in ethnicity, family socioeconomic status and parental occupation3. From birth throughout early adulthood, TEDS has assessed individuals on many different cognitive and behavioural traits. In this study, we used information from the subsample of 5,825 unrelated individuals (just one member of a twin pair), for whom we had genotype information available.

*What measures did you use?*

Educational achievement at age 7 and 12 was measured through National Curriculum based teacher ratings. Teacher ratings assessed two main abilities: English (including ‘speaking and listening’, ‘reading’ and ‘writing’) and mathematics (including ‘using and applying mathematics’, ‘numbers’ and ‘shapes, space and measures’).

Educational achievement at age 16 was measured through mathematics and English grades obtained in UK standardized examinations (GCSEs), taken at the end of compulsory education.

To measure general cognitive ability, we formed a composite of various cognitive tests administered online at age 7, 12 and 16, which measured verbal as well as non-verbal abilities.

Our family socioeconomic status (SES) measure was a composite of parental education and occupation and age of mother at the birth of the first child.

*What did you do in this study?*

In this study, we used information from the education GWAS2 described above to inform our decision about which SNPs to include in our polygenic score. We then calculated education associated polygenic scores for genotyped individuals in our sample, the Twins Early Development Study.

We then used the polygenic scores in our sample to predict educational achievement at age 7, 12 and 16, and to predict other related measures, such as intelligence and family socioeconomic status (SES).

We also tested whether in our sample genetic influences linked to educational achievement differ between individuals in high versus low SES groups. This sort of relationship is called ‘gene-environment interaction’, which means that genetic factors have a greater impact in certain environments. For example, having a high polygenic score could be a protective factor against the adverse effects of low family SES.

**Results from the current study**

*What did you find?*

We found that the polygenic score, which explained 4% of the variance in years of education for adults, explained more than twice as much variance in actual educational achievement on standardised test scores at age 16 (GCSEs). The polygenic score explained 9.1% of the variance, which is the most variance any polygenic score has explained so far for any behavioural measure.

When we compared individuals with the highest polygenic scores to the individuals with the lowest polygenic scores, we found that their educational achievement differed by a whole school grade at age 16. Participants with the highest polygenic scores had a grade between an A and B on average, those participants with the lowest polygenic scores had a grade between B and C. This makes a real difference for life chances – twice as many of individuals with the highest polygenic scores go on to university as compared to those with the lowest polygenic scores.

How early does the polygenic score predict educational achievement? We show that it predicts significantly early in school at age 7 as well as at age 12 and age 16. The polygenic score explained increasing amounts of variance in educational achievement across the school years: 2.8%, 4.6% and 9.1% at ages 7, 12 and 16, respectively.

We also found that the polygenic score for education was linked to measures related to educational achievement. The polygenic score explained 3.6% of the variance in general cognitive ability, and 7.5% of the variance in family socioeconomic status.

We tested for gene-environment interaction, but we found that genetic influences on educational achievement did not differ for individuals in different SES groups. That is, the effect of the polygenic score was as great in low as in high family SES, which could be viewed as good news in terms of educational mobility.

**What are the implications of these findings?**

No necessary policy implications follow our findings.

However, we believe that we are at a tipping point for predicting genetic strengths and weaknesses for individuals. Through polygenic scoring, we have been able to explain almost 10% of the differences between children’s achievement using DNA alone. 10% is a long way from 100% but it is a lot better than we usually do in predicting behaviour. For example, when we think about differences between boys and girls in maths, gender explains about 1% of the variance4 Motivational variables such as ‘grit’, which describes the perseverance of an individual, together with other personality measures explain only about 5% of the variance in educational achievement5.

We believe that very soon, polygenic scores will be used to identify individuals who are at particular genetic risk. Having this information will be useful as it provides the opportunity to develop prevention and intervention strategies. For example, providing additional support that is tailored to a child’s needs may prevent later developmental difficulties. However, before polygenic scores can be used in educational settings, it will be important to create dialogues between researchers and policymakers.

As for now, implications of our findings mostly relate to research in the social and behavioural sciences. Our study shows that a polygenic score based on the GWAS for years of education2 reliably and substantially predicts educational achievement, and we demonstrate that by using polygenic scores, it is possible to address important research questions. Through polygenic scores, relationships between an individual’s genetic propensity for a specific trait and any other measures, such as cognitive, behavioural and clinical traits, can be tested.

There is also converging evidence from previous research that most human traits are influenced by a dynamic interplay between genetic and environmental effects. As we have shown in our study, polygenic scores provide an excellent method to test for gene-environment interplay. For example, by testing the link between a polygenic score and an environmental measure, it is possible to examine whether the experience of a particular environment is genetically influenced (this is also called ‘gene-environment correlation’). Another type of gene-environment interplay, which can be tested with polygenic scores, is ‘gene-environment interaction’. As demonstrated in our study, it is possible to test whether the effect of the polygenic score on an outcome is dependent on the experience of a particular environment.

Most importantly, these and many more research questions can be answered without the need for extremely large sample sizes that are usually needed for most genetics research. To sufficiently detect a polygenic score based on the education GWAS2, only 250 individuals with genotype data are required. Therefore, polygenic scores make genetics research much more affordable and accessible than previously, which will transform the way research is conducted.

**References**

1 de Zeeuw EL, de Geus EJC, Boomsma DI. Meta-analysis of twin studies highlights the importance of genetic variation in primary school educational achievement. *Trends Neurosci Educ* 2015; **4**: 69–76.

2 Okbay A, Beauchamp JP, Fontana MA, Lee JJ, Pers TH, Rietveld CA *et al.* Genome-wide association study identifies 74 loci associated with educational attainment. *Nature* 2016; **advance online publication**. doi:10.1038/nature17671.

3 Haworth C, Davis OS, Plomin R. Twins Early Development Study (TEDS): A genetically sensitive investigation of cognitive and behavioral development from childhood to young adulthood. *Twin Res Hum Genet* 2013; **16**: 117–125.

4 Hyde JS. Gender similarities and differences. *Annu Rev Psychol* 2014; **65**: 373–398.

5 Rimfeld K, Kovas Y, Dale PS, Plomin R. True Grit and Genetics: Predicting Academic Achievement From Personality. *J Pers Soc Psychol* 2016; : No Pagination Specified.