Frequently Asked Questions (FAQs)

This document provides information about the study:

Lee *et al.* (2018) "Gene discovery and polygenic prediction from a 1.1-million-person GWAS of educational attainment" *Nature Genetics*, in press.

The document was prepared by several of the study's coauthors and draws from and builds on the FAQs for earlier SSGAC papers. It has the following sections:

1. **Background**
2. **Study design and results**
3. **Social and ethical implications of the study**
4. **Appendices**

For clarifications or additional questions, please contact Daniel Benjamin (djbenjam@usc.edu).

Table of Contents

1. Background

1.1 Who conducted this study? What are the group's overarching goals?

The authors of the study are members of the Social Science Genetic Association Consortium (SSGAC). The SSGAC is a multi-institutional, international research group that aims to identify statistically robust links between genetic variants and social-science-relevant traits. These include traits such as behavior, preferences, and personality that are traditionally studied by social and behavioral scientists (e.g., economists, psychologists, sociologists) but are often also of interest to health and other researchers.

The SSGAC was formed in 2011 to overcome a specific set of scientific challenges. Most traits and behaviors are associated with thousands of genetic variants. Although their collective effect can be substantial (see FAQs 1.5 & 2.3), we now know that almost every one of these genetic variants has an extremely weak effect on its own. To identify specific variants with such small effects, scientists must study at least hundreds of thousands of people (to separate weak signals from noise). One promising strategy for doing this is for many investigators to pool their data into one large study. This approach has borne considerable fruit when used by medical geneticists interested in a range of diseases and conditions (Visscher et al. 2017). Most of these advances would not have been possible without large research collaborations between multiple research groups interested in similar questions. The SSGAC was formed in an attempt by social scientists to adopt this research model.

The SSGAC is organized as a working group of the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE), a successful medical consortium. It was founded by three social scientists—Daniel Benjamin (University of Southern California), David Cesarini (New York University), and Philipp Koellinger (Vrije Universiteit Amsterdam)—who believe that studying genetic variants associated with social scientific outcomes can have substantial positive impacts across many research fields. This includes research that aims to better understand the effects of the environment (e.g., research on policy interventions, including the effects of different school environments) and interactions between genetic and environmental effects. The potential benefits also span a diverse set of research questions in the biomedical sciences, such as why and how educational attainment is linked to longevity and better overall health outcomes.

To conduct such research, the SSGAC implements genome-wide association studies (GWAS, see FAQ 1.3) of social-scientific outcomes. For example, to conduct a GWAS of educational attainment, every participating cohort uploads the (within-cohort) statistical association between educational attainment and a single-nucleotide polymorphism (SNP) in the genomes of the individuals in the cohort. A SNP is a base-pair of the genome where there is common variation in the human population (see FAQ 1.3). This statistical analysis is repeated for each SNP on the genome. The cohort-level results do not contain individual-level data – just summary statistics about these within-cohort statistical associations. The SSGAC then combines these cohort results to produce the overall GWAS results. By using existing datasets and combining cohort-level results, we can study the genetics of ~1.1 million people at very low cost. The SSGAC publicly shares the overall, aggregated results at www.thessgac.org/data so that other scientists can build on this work. These publicly available data have already catalyzed many research projects and analyses across the social and biomedical sciences (see FAQ 1.6. for examples).

The Advisory Board for the SSGAC is composed of prominent researchers representing various disciplines: Dalton Conley (Sociology, Princeton University), George Davey Smith (Epidemiology, University of Bristol), Tõnu Esko (Molecular Biology and Human Genetics, University of Tartu and Estonian Genome Center), Albert Hofman (Epidemiology, Harvard University), Robert Krueger (Psychology, University of Minnesota), David Laibson (Economics, Harvard University), James Lee (Psychology, University of Minnesota), Sarah Medland (Genetic Epidemiology, QIMR Berghofer Medical Research Institute), Michelle Meyer (Bioethics and Law, Geisinger Health System), and Peter Visscher (Statistical Genetics, University of Queensland).

The SSGAC is committed to the principles of reproducibility and transparency. Prior to conducting genetic association studies, power calculations are carried out to determine the necessary sample size for the analysis (assuming realistically small effect sizes associated with individual genetic variants). Whenever possible, we pre-register our analyses at OSF (formerly Open Science Framework). Major SSGAC publications are usually accompanied by a FAQ document (such as this one). The FAQ document is written to communicate what was found less tersely and technically than in the paper, as well as what can and cannot be concluded from the research findings more broadly. FAQ documents produced for SSGAC publications are available at https://www.thessgac.org/faqs.

In addition to educational attainment, SSGAC-affiliated papers have studied subjective well-being, reproductive behavior, and risk tolerance. The SSGAC website contains an up-to-date list of our major publications, which have been published in journals such as *Science, Nature, Nature Genetics, Proceedings of the National Academy of Sciences, Psychological Science,* and *Molecular Psychiatry*.

1.2 The current study focuses on an outcome called "educational attainment." What is educational attainment?

Educational attainment is the amount of formal education a person completes (measured as the number of years of education completed for people in our sample, all of whom are at least age 30 or older). Although educational attainment is most strongly influenced by social and other environmental factors (see FAQ 1.7), it is also influenced by thousands of genes. People vary considerably in how much education they complete. Education is recognized throughout the social and biomedical sciences as an important "predictor" (see FAQ 1.4) of many other life outcomes, such as income, occupation, health, and longevity (Ross & Wu 1995; Cutler & Lleras-Muney 2008). Educational attainment is also among the relatively few social-scientific traits for which it is feasible to conduct a large-scale genome-wide study, because educational attainment is frequently measured by a variety of cohorts, including medical cohorts, due to its robust association with health. A large-scale study is necessary (but not sufficient) to generate scientific findings that are reproducible.

1.3 What is a GWAS? Are the genetic variants identified in a GWAS "causal"?

In a genome-wide association study (GWAS), scientists look at genetic variants measured across the entire human genome to see whether any of them are, *on average*, associated with higher or lower levels of some outcome. Commonly, and in our studies, such analyses focus on the most common genetic variants—so

called single-nucleotide polymorphisms (SNPs). SNPs are sites in the genome where single DNA base pairs commonly differ across individuals. SNPs usually have two different possible base pairs, or alleles. Although there are tens of millions of sites where SNPs are located in the human genome, GWASs typically investigate only SNPs that can be measured (or imputed) with a high level of accuracy. These days, such procedures usually yield millions of SNPs that together capture most common genetic variation across people.

GWAS has been a successful research strategy for identifying genetic variants associated with many traits and diseases, including body height (Wood et al. 2014), BMI (Locke et al. 2015), Alzheimer's disease (Lambert et al. 2013), and schizophrenia (Ripke et al. 2014). It has also recently been used to identify genetic variants associated with a variety of health-relevant social science outcomes, such as the number of children a person has (Barban et al. 2016), happiness (Okbay, Baselmans, et al. 2016; Turley et al. 2018), and educational attainment (Rietveld et al. 2013; Okbay, Beauchamp, et al. 2016).

GWAS identifies genetic variants that are associated with the outcome, but an observed association with a specific variant need not imply that the variant *causes* the outcome, for a variety of reasons. First, genetic variants are often highly correlated with other, nearby variants on the same chromosome. As a result, when one or more variants in a region causally influence an outcome (in that particular environment), many non-causal variants in that region may also be identified as associated with the outcome. When GWAS results are analyzed, researchers will often tend to emphasize results for the genetic variant in a region that showed the strongest evidence of association. This variant need not be the causal variant. In fact, the causal genetic variant may not have even been measured directly. For example, GWAS that focus on common SNPs would not be able to identify rare or structural genetic variants (e.g., deletions or insertions of an entire genetic region) that are causal, but they may identify SNPs that are correlated with these unobserved variants.

Second, the frequencies of many genetic variants vary systematically across environments. If those environmental factors are not accounted for in the association analyses, some of the associations found may be spurious. To use a well-known example (Lander & Schork 1994), any genetic variants common in people of Asian ancestries will be associated statistically with chopstick use, but these variants would not *cause* chopstick use; rather, these genetic variants and the outcome of chopstick use are both distributed unevenly among people with different ancestries. This is the problem of "population stratification" discussed in Appendix 1. GWAS researchers have a number of strategies for addressing the challenges posed by population stratification (see FAQs 2.4 & 3.5 and Appendix 1).

Even in studies such as ours that attempt to address and correct for heterogeneity in genetic ancestry, allele frequencies may nonetheless vary systematically with environmental factors. For example, a genetic variant that is associated with improved educational outcomes in the parental generation may have downstream effects on parental income and other factors known to influence children's educational outcomes (such as neighborhood characteristics). This same genetic variant is likely to be inherited by the children of these parents, creating a correlation between the presence of the genetic variant in a child's genome and the extent to which the child was reared in an environment with specific characteristics. A recent study of Icelandic families showed that the parental allele that is *not* passed on to the parent's offspring is still associated with the child's educational attainment, suggesting that GWAS results for educational attainment partly

represent these intergenerational pathways (Kong et al. 2018). Our sibling analyses yield results that are consistent with this conclusion (see FAQ 2.4).

Third, variants' effects on an outcome may be indirect, so a variant that may be "causal" in one environment may have a diminished effect or no effect at all in other environments. For example, the nicotinic acetylcholine receptor gene cluster on chromosome 15 is associated with lung cancer (Thorgeirsson et al. 2008; Amos et al. 2008; Hung et al. 2008). From this observation alone we cannot conclude that these genetic variants cause lung cancer through some direct biological mechanism. In fact, it is likely that these genetic variants increase lung cancer risk through their effects on smoking behavior. In a tobacco-free environment, it is plausible that many of the associations would be substantially weaker and perhaps disappear altogether. Thus, even *if* we have credible evidence that a specific association is not spurious, it is entirely possible that the genetic variant in question influences the outcome through channels that we, in common parlance, would label environmental (e.g., smoking). Nearly forty years ago, the sociologist Christopher Jencks criticized the widespread tendency to mistakenly treat environmental and genetic sources of variation as mutually exclusive (see also Turkheimer 2000). As the example of smoking illustrates, it is often overly simplistic to assume that "genetic explanations of behavior are likely to be exclusively physical explanations while environmental explanations are likely to be social" (Jencks 1980, p.723).

In general, GWAS is just one step in a longer, often complex process of identifying causal pathways, but the results of a large-scale GWAS are a useful tool for that purpose and often lead to novel and important insights (Visscher et al. 2017). In other words, GWAS results provide important signals as to where scientists should invest future in-depth research to understand why the association exists.

1.4 In what sense do the genetic variants identified in a GWAS "predict" the outcome of interest? What do you mean by "effect size"?

When we and other scientists say that genetic variants (and other variables, such as demographics) "predict" certain outcomes, our use of the word differs in several important ways from how "predict" is used in standard language (e.g., outside of social science research papers). First, we do not mean that the presence of a genetic variant guarantees an outcome with 100% probability, or even with a high degree of likelihood. Rather, we mean that the variant is, on average across people, statistically associated with an outcome. In other words, on average, people with the genetic variant have a higher likelihood of the outcome compared to people without the genetic variant. A genetic variant is said to be statistically "predictive" of an outcome even if the presence of the genetic variant only *very weakly* increases the likelihood of that outcome—as is the case, for instance, with every SNP that we identify that is associated with educational attainment.

Second, in standard language, "prediction" usually refers to the future. In contrast, when scientists say that genetic variants "predict" an outcome, they mean that they expect to see the association in *new data*. "New data" means data that haven't been analyzed yet—regardless of whether that data will be collected in the future or has already been collected.

Finally, in standard language, a "prediction" is often an unconditional guess about what will happen. Instead of meaning it unconditionally, scientists mean that they expect to see an association in new data under

certain conditions, for example, that the environment for the new data is the same as the environment in which the variants were found in the previously studied data to be associated with the outcome. In the example given in FAQ 1.3, in which a genetic variant is associated with lung cancer due to its effect on smoking, we would *not* expect the genetic variant to be as strongly predictive of lung cancer in an environment where cigarettes are absent.

We use the term "effect size" as a concise way to refer to the magnitude of the predicted difference in the outcome resulting from having one allele of a genetic variant as opposed to the other possible allele (for example, see FAQ 2.2). The use of the word "effect" is *not* intended to imply that we believe it is generally appropriate to use the strength of the *association* between a variant and educational attainment as a measure of the variant's causal effect on educational attainment (see FAQ 1.3).

1.5 What is a polygenic score?

The results of a GWAS can be used to create a "polygenic score," an index composed of many genetic variants from across the genome. Because a polygenic score aggregates the information from many genetic variants, it can "predict" (see FAQ 1.4) far more of the variation among individuals for the GWAS outcome than any single genetic variant. Often, the polygenic scores with the most predictive power are those created using *all* the (millions of) genetic variants studied in a GWAS. The larger the GWAS sample size, the greater the predictive power (in other, independent samples) of a polygenic score constructed from the GWAS results. More precisely, the GWAS results are used to create a *formula* for how to construct a polygenic score. Using this formula, a polygenic score can then be constructed for any individual with genome-wide data. Indeed, some of the value of a GWAS is that the polygenic score it produces can be used in subsequent studies conducted in other samples.

1.6 Why conduct a GWAS of educational attainment?

We are motivated to conduct this research because we believe it can be fruitful for the social sciences and health research. In addition to the specific findings of our paper, which are discussed in Section 2 of these FAQs, the results of a GWAS of educational attainment also provide inputs for other research. For example, results from our earlier GWAS of educational attainment (Rietveld et al. 2013; Okbay, Beauchamp, et al. 2016) conducted in much smaller sample sizes (see also FAQ 1.7) have been used to:

- examine the genetic overlap between educational attainment and ADHD, schizophrenia, Alzheimer's disease, intellectual disability, cognitive decline in the elderly, brain morphology, and longevity (Pickrell et al. 2016; Warrier et al. 2016; Anderson et al. 2017; Marioni et al. 2016);
- help us better identify possible genetic subtypes of schizophrenia (Bansal et al. 2017);
- explore why educational attainment appears to be protective against coronary artery disease (Tillmann et al. 2017) and obesity (van Kippersluis & Rietveld 2017);
- control for genetic influences in order to generate more credible estimates of how changes in school policy influence health outcomes (Davies et al. 2018);

- study why specific genetic variants predict educational attainment. For example, it appears that some genetic effects on educational attainment operate through associations with cognitive performance and traits such as self-control (Belsky et al. 2016), which in turn affect educational attainment;
- study how the effects of genes on education differ across environmental contexts (Schmitz & Conley 2017; Barcellos et al. 2018); and
- develop new statistical tools that may advance our understanding of how parenting and other features of a child's rearing environment influence his or her developmental outcomes (Kong et al. 2018; Koellinger & Harden 2018).

These are just some examples of follow-up studies that previous GWASs of educational attainment have already enabled. By making the results of our analyses publicly available at https://www.thessgac.org/data, we hope to facilitate further valuable work by other researchers.

1.7 What was already known about genetic associations with educational attainment prior to this study?

Educational attainment is strongly influenced by social and other environmental factors. For example, holding all other influences equal, those who live in communities where education (at least beyond a certain level) is relatively expensive are less likely to obtain a high level of educational attainment. Even when education is free or heavily subsidized, full-time education constitutes an opportunity cost that not everyone is equally able to bear: some individuals, due to a variety of family or economic circumstances, will face more pressure than others to leave school and enter the labor force. More generally, educational outcomes are strongly influenced by environmental factors such as social norms, early-life educational experiences, and economic opportunity.

A variety of findings—from twin, family, and GWAS studies—suggest that in affluent countries, genetic factors account for some of the differences across people in their educational attainment (Branigan et al. 2013; Heath et al. 1985; Silventoinen et al. 2004). Studies have found repeatedly that identical twins raised in the same home are substantially more similar to each other in their educational attainment than fraternal twins (or other full siblings) reared together. Full siblings reared together are, in turn, more similar than half siblings reared together who, in turn, are more similar than genetically unrelated siblings (e.g., siblings who are conventionally unrelated, typically because at least one of them is adopted) reared together (Cesarini & Visscher 2017; Sacerdote 2011; Sacerdote 2007). The studies have also provided strong evidence that so-called common environment (the environmental factors shared by siblings raised in the same household) can have long-lasting effects on educational outcomes. In Sweden, the educational outcomes of adopted (i.e., genetically unrelated) brothers reared in the same households are about as similar as the educational outcomes of full siblings reared in separate homes (Cesarini & Visscher 2017). A study of Korean-American adoptees finds that adoptees assigned to households where both parents had college degrees were 16 percentage points more likely to attend college than children assigned to families in which neither parent completed college (Sacerdote 2007).

Research (like the current study) using molecular genetic data—data that measures each person's DNA and can be used to identify differences between people at the molecular level—has similarly found that common SNPs jointly predict up to 20% of variation across individuals (Rietveld et al. 2013). This predictive power may derive from many different types of mechanisms. For example, genetic variation may affect neural functions such as memory. Genetic variation may improve sleep quality (making it easier to subsequently stay awake in boring lectures). Genetic variation can affect personality traits, such as the willingness to listen politely to and follow the instructions of teachers (who aren't always right but nevertheless dictate grades and other outcomes). There may also be even more convoluted pathways. For example, genetic variation can affect one's sociability, which might draw someone into or drive someone out of the particular social environments that exist in higher education.

In prior GWAS studies, researchers have observed that some genetic variants are associated with educational attainment. In the SSGAC's first major publication (Rietveld et al. 2013), we conducted a GWAS in a sample of roughly 100,000 people and identified three genetic variants that were statistically associated with educational attainment. In 2016, the SSGAC conducted another GWAS of educational attainment, this time in a sample of around 300,000 people (Okbay, Beauchamp, et al. 2016). We found that 74 genetic variants were associated with educational attainment. These included the three genetic variants identified in our earlier study (Rietveld et al. 2013). Both of these studies involved, at the time they were conducted, the largest sample sizes ever studied for genetic associations with a social science outcome.

There were three key takeaways from the SSGAC's prior work:

(1) A GWAS approach can identify specific genetic variants statistically associated with behavioral variables if the study is conducted in large enough samples (at least 100,000 people).

(2) Genetic variants that are associated with a behavioral variable such as educational attainment are each likely to have less predictive power (i.e., a smaller effect size) than are genetic variants that are associated with a biomedical or other physical outcome (Chabris et al. 2015). For example, of the hundreds of genetics variants found to be associated with height to date (Wood et al., 2014), the genetic variant with the strongest association predicts 0.4% of the variation across individuals in height, whereas the genetic variant with the strongest association with educational attainment identified to date predicts less than one tenth (<0.04%) as much of the variation in educational attainment (Okbay, Beauchamp, et al. 2016). (The genetic variants that have not yet been identified will very likely explain less variance than those that are currently known, since statistical power is greatest for those that explain the most variance.)

(3) In the samples studied, at least 20% of the variation in educational attainment is predicted by genetic variation (Rietveld et al. 2013), implying that the genetic associations with educational attainment result from the cumulative effects of at least thousands (probably millions) of different genetic variants, not just a few.

These findings from twin, family, and GWAS studies imply that individuals who carry an allele associated with greater educational attainment will on average complete slightly more formal education than other (similarly environmentally situated) individuals who carry a different allele of the same genetic variant. Put in population terms, these findings imply that people with particular alleles will tend *on average* to complete more formal education, while people who carry other alleles will tend *on average* to complete less formal education. It is important to emphasize that these associations represent *average tendencies* in a population. Many individuals with high polygenic scores for educational attainment will not get a college degree, and vice-versa. This makes polygenic scores for educational attainment poor predictors of individual outcomes (see FAQ 3.4), but increasingly useful tools in social science research (see FAQ 2.3).

2. Study design and results

2.1 What did you do in this paper? How was the study designed? Why was the study designed in this way?

We conducted a GWAS (see FAQ 1.3) of educational attainment (see FAQ 1.2) in a sample of over 1.1 million people. The sample size we used in the current study is much larger than that used in previous GWAS of educational attainment (see FAQ 1.7). By constructing a current sample of over 1.1 million, we expected to estimate genetic effects with much greater accuracy than previous studies (with smaller samples) and, thus, to learn much more about the specific genetic variants that are associated with educational attainment.

To construct such a large sample, we combined information from our previous GWAS of roughly 300,000 research participants from 64 datasets (which we refer to as "cohorts") (Okbay, Beauchamp, et al. 2016) with data that have recently become available from seven additional cohorts. These seven new cohorts include the UK Biobank and the personal genomics company 23andMe, both of which have surveyed and genotyped hundreds of thousands of research participants.

Our study was limited to only the most common type of genetic variant: single-nucleotide polymorphisms (SNPs, see FAQ 1.3). *Unlike* most other studies, which have analyzed only the autosomes (the non-sex chromosomes), our study also included SNPs on the X chromosome (see FAQ 2.6). In total, our analyses included approximately 10 million SNPs. And, as in other GWASs, our analyses included only individuals of primarily European genetic ancestry. This restriction is needed in order to reduce statistical confounds that otherwise arise from studying populations with diverse genetic ancestries (see the discussion of population stratification in Appendix 1; see also FAQs 1.3, 2.4 & 3.5).

In the remainder of the paper, we used the findings from the GWAS for a range of additional analyses that explored (among other things):

- the extent to which siblings with different alleles end up with different amounts of formal schooling (see FAQ 2.4);
- which environmental conditions affect the size of the association between genetic variants and educational attainment (see FAQ 2.5);

- the genetic overlap between educational attainment and other outcomes, such as cognitive performance (constituting the largest GWAS of cognitive performance to date) and self-reported math ability (see FAQ 2.7);
- which other outcomes are also correlated with genetic variants that are associated with educational attainment (see FAQ 2.8); and
- the biological functions of the genetic variants identified (see FAQ 2.9).

2.2 What did you find in the GWAS of educational attainment?

In our sample of roughly 1.1 million people, we found 1,271 genetic variants that were associated with educational attainment (using the standard statistical threshold in GWAS, which adjusts for multiple hypothesis testing). This is a substantial increase from the 74 variants identified in our last GWAS of around 300,000 individuals (Okbay, Beauchamp, et al. 2016), confirming the importance of large sample size for identifying specific genetic variants associated with behavioral traits.

The current study further confirmed the finding from our earlier work that the effects of individual genetic variants on educational attainment are extremely small. The average effect size across the 1,271 genetic variants was just 1.8 weeks of schooling per allele; even the SNPs with the strongest associations only predicted around 3 weeks of additional schooling per allele. Taken together, these 1,271 SNPs accounted for just 3.9% of the variation across individuals in years of education completed.

Here is another way to think about this result. Imagine that we used the results for these 1,271 genetic variants (not the ~1 million SNPs across entire genome we discuss in FAQ 2.3) to predict the educational attainment for a new group of people (separate from our discovery sample). We could then compare each individual's *predicted* educational attainment to their *actual* educational attainment. If we did so, our results suggest that we would find that the predictions and actual outcomes correlate only very modestly (at about $r = 0.20$). That, in turn, means that if someone were predicted to complete an above average number of years of schooling (i.e., to be in the top half of educational attainment), that person would have about a 58% chance of actually being in the top half of educational attainment. Fifty-eight percent is better than chance (i.e., 50%), suggesting that a prediction based on these 1,271 SNPs has more power to predict educational attainment than a coin flip—but only a bit more power. By contrast, a prediction based on a polygenic score that combines ~1 million SNPs that we studied (see FAQs 1.5 & 2.3) has more predictive power: $r = 0.33$, corresponding to 11% of the variation across individuals.

The contrast between the 3.9% of the variation predicted by the 1,271 SNPs and the 20% known to be explained by common SNPs (see FAQ 1.7) implies that there are many other SNPs that have not yet been identified. Even larger sample sizes will be needed to identify them.

It is also important to keep in mind that educational attainment is a complex phenomenon, and our study focuses on only a tiny piece of the bigger picture. In this paper, we only examine one type of genetic variant (SNPs). Further, we conduct only preliminary analyses of how the effects of genetic variants on educational attainment differ depending on environmental conditions (see FAQ 2.5). These other genetic effects, environmental effects, and their interactions are important topics of active research and of future work by

the SSGAC. Such work includes further studies of associations between educational attainment and epigenetic marks (Linnér et al. 2017).

2.3 How predictive is the polygenic score developed in this study?

As discussed in FAQ 1.5, we can create an index using the GWAS results from around ~1 million genetic variants. Such an index is called a "polygenic score."

The polygenic score we constructed "predicts" (see FAQ 1.4) around 11% of the variation in education across individuals (when tested in independent data that was not included in the GWAS). This ~1 million SNP polygenic score predicts much more of the variation than does the genetic predictor described in FAQ 2.2, which was based on only 1,271 SNPs. Including all ~1 million SNPs tends to add predictive power because the threshold for significance/inclusion that is used to identify the 1,271 SNPs is very conservative (i.e., many of the other ~1 million SNPs are also associated with educational attainment but are not identified by our study, and on net, it turns out empirically that more signal than noise is added by including them). This study's polygenic score has much more predictive power than polygenic scores constructed from our earlier two GWAS of educational attainment, because both of those studies had much smaller sample sizes (~100,000 and ~300,000 individuals, respectively, compared with ~1.1 million individuals of the current study).

Individuals with high polygenic scores have, *on average*, higher levels of education than those with lower polygenic scores. In the present study, we found that in a U.S. sample of young adults (the National Longitudinal Study of Adolescent to Adult Health), 12% of those with the lowest 20% of polygenic scores graduated from college, compared with 57% of those with the highest 20% of polygenic scores. These results show both that polygenic scores have some predictive power but also that polygenic scores do not determine or pin down individual outcomes: even when polygenic scores are based on GWAS of many more people and therefore have even greater predictive power than ours, there will always be many people whose polygenic scores "predict" lower educational attainment who in fact attain relatively high amounts of education and vice-versa.

As we discuss in FAQ 3.4, an individual's polygenic score for education (even a polygenic score based on ~1 million SNPs) is still *not* a very accurate prediction of that individual's actual level of education attained. However, polygenic scores are useful for *scientific studies* (including social science, health research, etc.). Such studies are concerned with aggregate population trends and averages rather than with individual outcomes. In particular, because the polygenic score predicts 11% of the variation across individuals, studies of its association with other variables can be well powered in sample sizes as small as 75 individuals (but not as small as 1 individual!).

Through this lens, the fact that the current study's polygenic score for educational attainment predicts 11% of the variation across individuals in education attained is quite meaningful and rivals or exceeds the predictive power of other variables commonly used in research—none of which, taken alone, predicts a large amount of variation in a behavioral outcome. For example, using our sample in order to maximize the comparability with the polygenic score, we estimated that household income predicts ~7% of variation in educational attainment and mother's education predicts ~15%. Thus, our score has approached the

predictive power of important demographic variables and can be used in similar ways (e.g., to control for genetics as an additional confound when evaluating the effects of environmental differences or interventions).

With a relatively high level of predictive power, the polygenic score we constructed enables other research that is of value to social scientists and health researchers. Such studies are already being conducted with the (much less powerful) polygenic scores from earlier GWAS of educational attainment (see FAQ 1.6). Our new results will enable many additional applications, such as studies that use the polygenic score in relatively small samples that contain rich health and behavioral data that is expensive to collect (e.g., a randomized controlled trial that studies the effects of subsidizing higher education and uses the polygenic score as a control variable).

2.4 What did you find in the analysis of siblings?

In a sample of ~44,000 siblings (~22,000 pairs), we examined the genetic variants identified in our GWAS. Specifically, we tested whether having more alleles of particular genetic variants than one's sibling is associated with having greater educational attainment than that sibling. One purpose of this analysis was to assess to what extent GWAS results are biased by factors such as unaccounted-for "population stratification" (see Appendix 1 and FAQs 1.3 & 3.5). We found strong evidence that the genetic variants identified in the GWAS are associated with educational attainment in our sibling analysis.

However, we also found that the associations with educational attainment were substantially smaller in the sibling analysis than in the GWAS (when we conducted an analogous study of height, we did not observe any quantitative discrepancies between the GWAS and the within-family estimates after a technical correction for assortative mating). We examined a number of possible explanations for the difference. Ultimately, we believe that the GWAS estimates are larger because they partly reflect the kinds of intergenerational mechanisms discussed in FAQ 1.3 and studied by Kong et al. (2018): an individual with a genetic variant associated with greater educational attainment is more likely to have a parent with that variant. Such a parent is likely to have attributes and behaviors (such as higher income or a greater likelihood of reading to a child) that contribute to increasing the child's educational attainment. These intergenerational mechanisms are not measured in the sibling analysis (since the siblings share the same parents).

If our conjecture about the source of the discrepancy is correct, it reinforces the importance of interpreting genetic associations with caution. Behavior geneticists sometimes criticize social scientists for failing to consider a role for genetic factors when interpreting correlations between relatives (e.g., parent-child correlations in educational attainment). We believe this criticism has merit (see FAQ 1.7 for a summary of the evidence) but it goes both ways. Since the variants identified in our GWAS also show evidence of association in our sibling analyses, we can be confident that our main results are not fully explained by factors that siblings share, such as parental genotypes and many features of rearing environment. But since the associations are weaker in the sibling analyses, it is plausible that some of the predictive power – perhaps a quarter – of our GWAS-identified variants arises because the variants *are* correlated with environmental factors that siblings share. For this and other reasons, we believe it is misleading to use phrases such as

"innate ability" or "genetic endowments" to describe what is measured by polygenic scores based on our GWAS estimates.

2.5 What did you find in the analysis of environmental heterogeneity?

We expect the associations of particular genetic variants with educational attainment to depend on environmental context (such as a country's school system and the quality of an individual's schools). That is partly because the meaning of a specific educational qualification varies across time and place. It is also because genetic variants don't affect educational attainment directly. Instead, they are likely to operate through a myriad of complex pathways. For example, they may affect psychological characteristics such as cognitive abilities and personality traits that ultimately influence educational attainment (see FAQ 1.7). To take one example, one would expect genetic variants associated with educational attainment to play a smaller role in countries whose laws make education compulsory for a relatively long period of time, because this environmental factor (education laws) constrains the range of outcomes to begin with. The genetic variants we have identified as associated with educational attainment in the current environments in which they were studied would play a lesser role still in a zombie apocalypse where many schools have been overrun by walkers (if any schools at all remain, different genetic variants might be associated with educational attainment—say, those associated with muscle twitch speed or immunity to the zombie virus). Finally, because genes influence educational attainment through other traits and behaviors, different pedagogies might make some of these traits more important to educational attainment than others, which would in turn likely modify the effect sizes—and even identities—of genetic variants associated with educational attainment.

In this study, we found some evidence that the effects of genetic variation on educational attainment differed across the 71 cohorts that contributed data. Characteristics such as cognitive abilities and personality traits are likely to matter differently in different places and time periods since educational systems also vary, and the 71 cohorts contributing data come from 15 different countries and enroll people born in a wide range of years. Documenting heterogeneity across cohorts in the associations of individual SNPs is a contribution of this paper because much previous work did not have sufficient statistical power to do so. Although most researchers expected such heterogeneity, the sample size of our study made it possible to measure the existence of these differences. We performed an exploratory analysis to investigate which observable environmental factors predicted differences in genetic effects across cohorts, but we were not sufficiently well powered to identify robust results. As GWAS sample sizes continue to grow, researchers will be able to understand in greater detail how environments shape genetic effects. This is one example of how adequately-powered GWAS can help establish the limits and nuances of genetic explanations of behaviors (see FAQ 3.7).

2.6 What did you find in the analysis of the X chromosome?

In contrast to most previous GWAS (including the previous GWAS of educational attainment), this study also examined variants on the X chromosome. In addition to the 1,271 variants identified on the autosomes (the non-sex chromosomes), we identified 10 variants associated with educational attainment on the X chromosome. Part of the reason we found so few variants on the X chromosome is because we only had X chromosome data in a smaller sample size (~700,000 individuals, compared with 1.1 million for the

autosomes). But even adjusting for sample size, we found fewer variants on the X chromosome than on other chromosomes of similar length. Moreover, the variants on the X chromosome as a whole "predicted" (see FAQ 1.4) less of the variation in educational attainment than the variants on other chromosomes of similar size. These results are of interest for human geneticists, as they are some of the first GWAS evidence about the effects of SNPs on the X chromosome (on any outcome, not just educational attainment).

Finally, in separate GWAS of men and women, we found that variants on the X chromosome predict similar amounts of variation in educational attainment in men and in women. Some researchers had hypothesized that genetic influences on the X chromosome are an important source of differences in the variance in cognitive performance across men and women. While there were compelling scientific reasons to view such claims skeptically even prior to the publication of our study, our results provide further evidence against the hypothesis.

2.7 What did you find in the analysis of cognitive performance and math abilities?

In supplementary analyses, we estimated GWAS of cognitive performance (as measured by scores on cognitive tests), self-reported math ability, and self-reported highest math class completed. Each of these GWAS was estimated in a substantially smaller sample than the GWAS of educational attainment, which contained information from roughly 1.1 million individuals. This difference in sample size reflects the fact that education is simple and standard to collect in large surveys, while cognitive performance, for example, is assessed less often because it requires respondents to answer time-consuming questions.

Still, with a sample size of around 250,000, our GWAS of cognitive performance is the largest published to date. A previous GWAS of cognitive performance was based on a sample of roughly 35,000 individuals (Trampush et al. 2017). We combined the results of that study with data from over 200,000 UK Biobank respondents who completed a test of verbal and numerical reasoning. Our GWAS identified 225 genetic variants associated with cognitive performance (using a standard threshold for genome-wide significance). A polygenic score constructed from all the genetic variants "predicts" (see FAQ 1.4) 7-10% of the variation in cognitive performance across individuals. A study of cognitive performance based on an even larger sample than ours (e.g., Savage et al. 2017) is presently being conducted under the auspices of the Psychiatric Genetics Consortium (PGC). The PGC study, in turn, is a follow-up to a previously published GWAS (Sniekers et al. 2017).

Self-reported math ability and highest math class completed have not been studied with GWAS before. Our GWAS of self-reported math ability used data from around 550,000 research participants of the personal genomics company 23andMe, and our GWAS of highest math class used data from over 400,000 research participants. We identified 618 and 365 genetic variants associated with self-reported math ability and highest math class completed, respectively.

We also found that many of the genetic variants that affect educational attainment also affect cognitive performance and math abilities. Exploiting this overlap in genetic effects, we applied a recently developed method, called Multi-Trait Analysis of GWAS (MTAG) (Turley et al. 2018). By doing so, we leveraged information from our large GWAS of educational attainment to identify additional genetic variants associated with cognitive performance, self-reported math ability, and highest math class completed. For

all three outcomes, more genetic variants were identified after incorporating information about genetic correlates with educational attainment.

2.8 Are the genetic variants associated with higher educational attainment in your study also associated with other outcomes?

Yes. We found that genetic variants associated with increased educational attainment are negatively associated with grade retention (i.e., having to repeat a grade) and positively associated with grade point average (GPA), cognitive performance, and self-reported math ability. This suggests that genetic variants "predict" (see FAQ 1.4) educational attainment at least in part through their correlation with cognitive development and academic performance.

In our previous GWAS of educational attainment, we also found that the genetic variants that predict educational attainment overlap with those that predict health outcomes, including Alzheimer's disease, bipolar disorder, and schizophrenia (Okbay, Beauchamp, et al. 2016). Using the results of our previous GWAS, other researchers have identified genetic overlap between educational attainment and other outcomes, including ADHD, intellectual disability, cognitive decline in the elderly, brain morphology, and longevity (Pickrell et al. 2016; Warrier et al. 2016; Anderson et al. 2017; Marioni et al. 2016). Future research is needed to understand *why* the genetic variants linked to education overlap with those associated with these other traits.

2.9 What do your results tell us about human biology and brain development?

We can draw inferences about biological pathways using computational methods that examine whether genes known to be involved in particular biological systems are especially likely to be associated with educational attainment.

In our earlier GWAS of educational attainment (Okbay, Beauchamp, et al. 2016), we found that the identified genes tended to be strongly active in the brain, especially prenatally, and were especially likely to be involved in neural development. The additional genes identified in the current study are also strongly active in the brain and involved in neural development. However, these additional genes are active both pre- and post-natally, at virtually all stages of brain development. Moreover, many of the newly identified genes are involved in neuron-to-neuron communication in the brain.

It is not surprising that genes may influence educational attainment in part because of their effects on brain development and communication within the brain. Cognitive abilities and personality traits (such as conscientiousness and resilience) that matter for school performance may be partially reflected in how the brain is organized. It is perhaps more surprising that our study of educational attainment generates a biological picture of brain development that is clearer than those generated by previous GWAS that focused directly on brain structures. We believe that the greater clarity of the biological picture we observe is due to the relatively large sample size of our study, which afforded us greater statistical power than previous GWAS. Since it will remain much easier to measure educational attainment than to conduct brain scans in large samples of individuals, we believe that GWAS of educational attainment will continue to play a useful

role in understanding the biology of brain development and constitutes one of the benefits of this research (see FAQ 3.7).

3. Ethical and social implications of the study

3.1 Did you find "the gene for" educational attainment?

No.

We did not find "the gene for" educational attainment or anything else. We identified many genetic variants that are associated with educational attainment. Although it was once believed that scientists would discover numerous one-to-one associations between genes and outcomes, we have known for a number of years that the vast majority of human traits and other outcomes are complex and are influenced by many (thousands or even millions of) genes, each of which alone tends to have a small influence on the relevant outcome.

3.2 Well, then, did you find "the genes for" educational attainment?

Although we did find several genes that are associated with educational attainment, we believe that characterizing these as "genes for educational attainment" is still likely to mislead, for many reasons.

First, most of the variation in people's educational attainment is accounted for by social and other environmental factors, not by additive genetic effects (See FAQ 1.7). "Genes for educational attainment" might be read to imply, incorrectly, that genes are the strongest predictor of variation in educational attainment.

Second, the genetic variants that are associated with educational attainment are also associated with many other things (only some of which we identify in this study, see FAQ 2.8). These variants are no more "for" educational attainment than for the other outcomes with which they are associated.

Third, the "predictive" power (see FAQ 1.4) of each individual genetic variant that we identify is very small. Our results show that genetic associations with educational attainment are comprised of thousands, or even millions, of genetic variants, each of which has a tiny effect size. Each variant is therefore weakly associated with, rather than a strong influence on, educational attainment. "Genes for educational attainment" might misleadingly imply the latter.

Fourth, environmental factors can increase or decrease the impact of specific genetic variants. Put differently, even if a genetic variant is associated with higher or lower levels of educational attainment *on average*, it may have a much larger or smaller effect depending on environmental conditions. Indeed, in the current paper and elsewhere, we report exploratory analyses that provide evidence of such gene-environment interactions (see FAQ 2.5). Educational attainment couldn't even exist as a meaningful object of measurement if we didn't have schools, and having schools introduces societal mechanisms that influence who goes to them. Accordingly, genetic associations with educational attainment necessarily will be mediated by societal systems and therefore genetic variation should often be expected to interact with environmental factors when it influences social phenomena, such as educational attainment. "Genes for

educational attainment" suggests a stability in the relationship between these genes and the outcome of educational attainment that does not exist.

Finally, genes do not affect educational attainment directly (see FAQ 2.5). As described in FAQ 2.9, the genes identified as associated with educational attainment tend to be especially active in the brain and involved in neural development and neuron-to-neuron communication. The "predictive" power (see FAQ 1.4) of genes on educational attainment may therefore be the result of a long process starting with brain development, followed by the emergence of particular psychological traits (e.g., cognitive abilities and personality). These traits may then lead to behavioral tendencies as well as experiences and treatment by parents, peers, and teachers. All of these factors may additionally interact with the environment in which a person lives. Eventually these traits, behaviors, and experiences may influence (but not completely determine) educational attainment.

3.3 Does this study show that an individual's level of educational attainment is determined, or fixed, at conception?

No.

Social and other environmental factors account for most variation in educational attainment. But even if it were true that genetic factors accounted for *all* of the differences among individuals in educational attainment, it would *still* not follow that an individual's number of years of formal schooling is "determined" at conception. There are at least three reasons for this:

First, some genetic effects may operate through environmental channels (Jencks 1980). As an illustrative example, suppose—hypothetically—that the genetic variants we identified help students to memorize and, as a result, to become better at taking tests that rely on memorization. In this example, changes to the intermediate environmental channels—the type of tests administered in schools—could have drastic effects on individuals' educational attainment, even though individuals' genetic variants would not have changed. A genetic association with educational attainment might not be found *at all* if schools did not use tests that rely on memorization. More generally, the genetic associations that we found might not apply as strongly if the education system were organized differently than it is at present (see also FAQ 1.3).

Second, even if the genetic associations with educational attainment operated entirely through non-environmental mechanisms that are difficult to modify (such as direct influences on the formation of neurons in the brain and the biochemical interactions among them), there could still exist powerful environmental interventions that could change the genetic relationships. In a famous example suggested by the economist Arthur Goldberger, even if all variation in unaided eyesight were due to genes, there would still be enormous benefits from introducing eyeglasses (Goldberger 1979). Similarly, policies such as a required minimum number of years of education and dedicated resources for individuals with learning disabilities can increase educational attainment in the entire population and/or reduce differences among individuals.

Third, even if the genetic effects on educational attainment were not influenced by changes in the environment, those environmental changes themselves could still have a major impact on the educational

attainment of the population as a whole. For example, if young children were given more nutritious diets, then everyone's school performance might improve, and college graduation rates might increase. By analogy, 80%-90% of the variation across individuals in height is due to genetic factors. Yet the current generation of people is much taller than past generations due to changes in the environment such as improved nutrition.

3.4 Can the polygenic score from this paper be used to accurately predict a particular person's educational attainment?

No. While the "predictive" power (see FAQ 1.4) of our polygenic score is substantial—it predicts 11% of variation in educational attainment across individuals—and useful for some purposes (see FAQ 1.6), it is important to keep in mind that the score *fails to predict* the vast majority (89%) of variation in years of education across individuals. Many of those with low polygenic scores go on to achieve high levels of education, and a large proportion of those with high polygenic scores do not complete college.

Thus, an important message of this paper and our earlier papers is that DNA does *not* "determine" an individual's level of education, for multiple reasons: First, it is estimated that, at least in the environments in which we have been measuring it, the additive effects of common genetic variants will only ever predict about 20% of the variance in educational attainment across individuals. Second, *today's* polygenic score is only able to predict a little more than half of that 20% (11 percentage points). Third, since genetic variants matter more or less depending on environmental context (see FAQ 2.5), a polygenic score might be less (or more) predictive for individuals in some environments than for individuals in others. Finally, polygenic predictions only hold for as long the environment in which they were developed remains substantially the same: if the laws or pedagogy underlying a population's educational system changes substantially, then so, too, might the polygenic score. Just as eyeglasses allow those genetically predisposed to poor vision to have nearly perfect vision, innovations in education (say, an innovation that makes education irresistibly engaging, thus mitigating the risk to those with genetic variants associated with lower ability to pay attention or maintain self-control) might result in those with lower polygenic scores now achieving just as much education, on average, as those with higher polygenic scores (see also FAQs 3.2 and 3.3).

As sample sizes for GWAS continue to grow, it will likely be possible to construct a polygenic score for educational attainment whose predictive power comes closer to 20% of the variance in educational attainment across individuals (Rietveld et al. 2013). Even this level of predictive power would pale in comparison to some other scientific predictors. For example, professional weather forecasts correctly predict about 95% of the variation in day-to-day temperatures. Weather forecasters are therefore vastly more accurate forecasters than social science geneticists will ever be.

Note: The results of SSGAC studies have sometimes been used in other projects to predict individual traits. We recognize that returning individual genomic "results" can be a fun way to engage people in research and other projects and to stoke their interest in, and educate them about, genomics. But it is important that participants/users understand that these individual results are not *meaningful* predictions and should be regarded essentially as entertainment. Failure to make this point clear risks sowing confusion and undermining trust in genetics research.

3.5 Can your polygenic score be used for research studies in non-European-ancestry populations?

Only in a limited way. As a practical matter, it is possible to calculate a polygenic score for any individual for whom genome-wide data is available, but the polygenic score will be much less "predictive" (see FAQ 1.4) in non-European-ancestry populations.

Our study was conducted only using samples of individuals of European ancestries (see Appendix 1). The set of SNPs that are associated with educational attainment in people of European ancestries is unlikely to overlap perfectly with the set of SNPs associated with EA in people of non-European ancestries. And even if a given SNP is associated in both ancestry groups, the effect size—in other words, the strength of the association—will almost surely differ. This is primarily because linkage disequilibrium (LD) patterns (i.e., the correlation structure of the genome) vary by ancestry. This means that some variant may be associated with educational attainment because the variant is in LD (i.e., correlated) with a variant elsewhere in the genome that causally affects education (see FAQ 1.3). If the strength of the correlation is greater in one ancestry group than in another, then the size of the association will be larger in that ancestry group. Moreover, even if LD patterns were similar in each ancestry group, the association may differ in different groups because environmental conditions differ (see FAQ 2.5). The fact that there are differences across ancestry groups in the set of associated SNPs and their effect sizes has two important implications.

First, it means that *polygenic scores of individuals from different ancestry groups* cannot *be meaningfully compared*. A recent paper (Martin et al. 2017) illustrated this point in the context of polygenic scores for predicting height; in the sample analyzed in that paper, polygenic scores for height for individuals of European ancestries are on average larger than those of South Asian ancestries which in turn are larger than those of African ancestries. In actuality, however, populations of African ancestries represented by the sample have similar height to populations of European ancestries, and both African and European populations tend to be taller than South Asian populations.

Second, while polygenic scores can be used to predict differences across individuals *within* a sample of people of non-European-ancestries, *the amount of predictive power will be much smaller than in a sample of people of European ancestries*. Such an attenuation of predictive power has been repeatedly found in prior work (Domingue et al. 2015; Vassos et al. 2017; Domingue et al. 2017; Belsky et al. 2013). Unfortunately, this attenuation means that for non-European-ancestry populations, many of the benefits of having a polygenic score available will have to wait until large GWAS studies are conducted using samples from these populations. (Currently, most large genotyped samples are of European ancestries.)

For a more extensive, excellent discussion of these and related issues, see Graham Coop's blog post "Polygenic scores and tea drinking": https://gcbias.org/2018/03/14/polygenic-scores-and-tea-drinking/.

For more on population stratification, see FAQs 1.3 & 2.4 and Appendix 1.

3.6 What policy lessons do you draw from this study?

None whatsoever. *Any* practical response—individual or policy-level—to this or similar research would be

extremely premature and unsupported by the science. Much more research is still needed to understand *why* the genetic variants we identified are associated with educational attainment. In this respect, our study is no different from GWAS of complex medical outcomes. In medical GWAS research, it is well understood that identifying genetic variants that "predict" (see FAQ 1.4) disease risk is merely a first step toward understanding the underlying biology. It is not sufficient to assess risk for any specific individual. It is not appropriate to base policies and practices on such assessments. However, the results of our study may be useful to social scientists (e.g., by allowing them to construct polygenic scores that can be used as control variables in randomized controlled trials or in studies of gene-by-environment interactions, see FAQ 1.6).

3.7 Could this kind of research lead to discrimination against, or stigmatization of, people with the relevant genetic variants? If so, why conduct this research?

Unfortunately, like a great deal of research—including, for instance, research identifying genomic variants associated with increased cancer risk—the results can be misunderstood and misapplied. This includes being used to discriminate against those with the variants in question (e.g., in insurance markets). Nevertheless, for a variety of reasons, in this instance, we do not think that the best response to the possibility that useful knowledge might be misused is to refrain from producing the knowledge. Here, we briefly discuss some of the broad potential benefits of this research. We then describe what we take to be our ethical obligation as researchers conducting this work.

First, one benefit of conducting social science genetics research in ever larger samples is that doing so allows us to correct the scientific record. An important theme in our earlier work has been to point out that most existing studies in social-science genetics that report genetic associations with behavioral traits have serious methodological limitations, fail to replicate, and are likely to be false-positive findings (Benjamin et al. 2012; Chabris et al. 2012; Chabris et al. 2015). This same point was made in an editorial in *Behavior Genetics* (the leading journal for the genetics of behavioral traits), which stated that "it now seems likely that many of the published [behavior genetics] findings of the last decade are wrong or misleading and have not contributed to real advances in knowledge" (Hewitt 2012). One of the most important reasons why earlier work has generated unreliable results is that the sample sizes were far too small, given that the true effects of individual genetic variants on behavioral traits are tiny. Pre-existing claims of genetic associations with complex social-science outcomes have reported widely varying effect sizes, many of them purporting to "predict" (see FAQ 1.4) ten to one hundred times as much of the variation across individuals as did the genetic variants we found in this study and in our other studies.

Second, behavioral genetics research also has the potential to correct the *social* record and thereby to help *combat* discrimination and stigmatization. For instance, at various times and places throughout human history (unfortunately, including the present day), girls and women have been discouraged or even prevented from pursuing as much education as their male counterparts. There are of course many reasons why that argument has been made and sometimes prevailed, but to the extent that it is rooted in a belief in genetically-based variance differences between males and females, our study's analysis of the X chromosome finds no such evidence (see FAQ 2.6). Similarly, overestimating the role of genetics can be damaging, and the present work can help debunk this myth, too. Of the 20% of the variance in educational attainment that is related to the additive effects common genetic variants, we have found that the relationship to educational attainment depends importantly on environmental factors (see FAQ 2.5). By

clarifying the *limits* of deterministic views of complex traits, recent behavioral genetics research—if communicated responsibly—could make appeals to genetic justifications for discrimination and stigmatization *less* persuasive to the public in the future.

Third, behavioral genetics research has the potential to yield many other benefits, especially as sample sizes continue to increase—as briefly summarized in FAQ 1.6. Foregoing this research necessarily entails foregoing these and any other possible benefits, some of which will likely be the result of serendipity rather than being foreseeable. For instance, as explained in FAQ 2.9, because educational attainment is measured in far larger genotyped samples than brain function, large-scale GWAS of educational attainment have provided better insights into brain function than GWASs to date that directly examine brain function, since the latter have necessarily been conducted in much smaller samples.

In sum, we agree with the U.K. Nuffield Council on Bioethics, which concluded in a report (Nuffield Council on Bioethics 2002, p.114) that "research in behavioural genetics has the potential to advance our understanding of human behaviour and that the research can therefore be justified," but that "researchers and those who report research have a duty to communicate findings in a responsible manner." In our view, responsible behavioral genetics research includes sound methodology and analysis of data; a commitment to publish all results, including any negative results; and transparent, complete reporting of methodology and findings in publications, presentations, and communications with the media and the public, including particular vigilance regarding what the results do—and do not—show (hence, this FAQ document).

4. Appendices

Appendix 1: Quality control measures

There are many potential pitfalls that can lead to spurious results in genome-wide association studies (GWAS). We took many precautions to guard against these pitfalls.

One potential source of spurious results is incomplete "quality control (QC)" of the genetic data. To avoid this problem, we used state-of-the-art QC protocols from medical genetics research (Winkler et al. 2014). We supplemented these protocols by developing and applying additional, more stringent QC filters.

Another potential source of spurious results is a confound known as "population stratification." To give a well-known illustration, suppose we were conducting a GWAS on the use of chopsticks (Lander & Schork 1994). People of Asian ancestries are far more likely to use chopsticks than people of European ancestries. If we combined samples of Chinese and European ancestries and performed a GWAS that ignores ancestry, then we would find genetic associations for these variants. However, those associations would simply reflect the fact that allele frequencies vary across ancestry groups.

In our study we were extremely careful to correct for population stratification as much as possible. At the outset, we restricted the study to individuals of European ancestries. As is standard in GWAS, we also controlled for "principal components" of the genetic data in the analysis; these principal components capture the small genetic differences across ancestry groups within European populations, so controlling for them largely removes the spurious associations arising solely from these small differences.

After taking these steps to minimize bias stemming from population stratification, we conducted a number of analyses to assess how much population stratification still remained in our data. The results of these tests indicate that there is some, but not much.

For one such analysis, we used a subset of the individuals in our data, ~22,000 sibling pairs (from five of the datasets that contributed to our study). The key idea underlying our tests is to examine if *differences* in genetic variants across siblings are associated with *differences* in the siblings' educational attainment. If so, then these associations cannot easily be attributed to bias in the estimates of the original studies, which compared individuals from different families. When comparing individuals who have different parents, genetic differences across individuals may be confounded with environmental differences associated with the parents' genetic variants (including the parents' ancestries, as discussed above). By contrast, full siblings share the same genetic parents, and genetic differences between siblings are random. Unfortunately, because our sample of siblings (~44,000 individuals) is much smaller than our overall GWAS sample (~1.1 million individuals), our estimates of the effects of the genetic variants within the sibling pairs are much noisier than in the GWAS. However, we *can* test whether the GWAS results are entirely due to population stratification, because if they were, then the sibling estimates would not line up with the GWAS estimates. In fact, we find that the within-family estimates are more similar to the GWAS estimates in both sign and magnitude than would be expected by chance. These results imply that our GWAS results are not solely due to population stratification. However, we also found that within-family estimates are substantially smaller than the GWAS estimates, as we discuss in FAQ 2.4.

As another analysis to assess how much population stratification still remained in our data after our efforts to minimize it, we applied a state-of-the-method from statistical genetics called LD Score regression (Bulik-Sullivan et al. 2015). The results of this analysis indicated that the biases in our results due to population stratification are small.

Appendix 2: Additional reading and references

Amos, C.I. et al., 2008. Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. *Nature Genetics*, 40, pp.616–622.

Anderson, E.L. et al., 2017. The causal effect of educational attainment on Alzheimer's disease: A two-sample Mendelian randomization study. *bioRxiv [https://doi.org/10.1101/127993]*.

Bansal, V. et al., 2017. Genetics of educational attainment aid in identifying biological subcategories of schizophrenia. *bioRxiv [https://doi.org/10.1101/114405]*.

Barban, N. et al., 2016. Genome-wide analysis identifies 12 loci influencing human reproductive behavior. *Nature Genetics*, 48(12), pp.1462–1472.

Barcellos, S.H., Carvalho, L.S. & Turley, P., 2018. Education can Reduce Health Disparities Related to Genetic Risk of Obesity: Evidence from a British Reform. *bioRxiv [https://doi.org/10.1101/260463]*.

Belsky, D.W. et al., 2013. Development and evaluation of a genetic risk score for obesity. *Biodemography and Social Biology*, 59(1), pp.85–100.

Belsky, D.W. et al., 2016. The Genetics of Success. *Psychological Science*, 27(7), pp.957–972.

Benjamin, D.J. et al., 2012. The Promises and Pitfalls of Genoeconomics. *Annual Review Of Economics*, 1(4), pp.627–662.

Branigan, A.R. et al., 2013. Variation in the Heritability of Educational Attainment: An International Meta-Analysis. *Social Forces*, 92(1), pp.109–140.

Bulik-Sullivan, B.K. et al., 2015. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics*, 47(3), pp.291–295.

Cesarini, D. & Visscher, P.M., 2017. Genetics and educational attainment. *npj Science of Learning*, 2(1), p.4.

Chabris, C.F. et al., 2012. Most reported genetic associations with general intelligence are probably false positives. *Psychological Science*, 23(11), pp.1314–1323.

Chabris, C.F. et al., 2015. The fourth law of behavior genetics. *Current Directions in Psychological Science*, 24(4), pp.304–312.

Cutler, D.M. & Lleras-Muney, A., 2008. Education and Health: Evaluating Theories and Evidence. In J. House et al., eds. *Making Americans Healthier: Social and Economic Policy as Health Policy*. New York: Russell Sage Foundation.

Davies, N.M. et al., 2018. The causal effects of education on health outcomes in the UK Biobank. *Nature Human Behaviour*.

Domingue, B.W. et al., 2017. Mortality selection in a genetic sample and implications for association studies. *International Journal of Epidemiology*, 46(4), pp.1285–1294.

Domingue, B.W. et al., 2015. Polygenic Influence on Educational Attainment: New evidence from The National Longitudinal Study of Adolescent to Adult Health. *AERA Open*, 1(3), pp.1–13.

Editors, N., 2013. Dangerous work. *Nature*, 502(7469), pp.5–6.

Goldberger, A.S.A., 1979. Heritability. *Economica*, 46(184), pp.327–347.

Heath, A.C. et al., 1985. Education policy and the heritability of educational attainment. *Nature*, 314(6013), pp.734–736.

Hewitt, J.K., 2012. Editorial policy on candidate gene association and candidate gene-by-environment interaction studies of complex traits. *Behavior Genetics*, 42(1), pp.1–2.

Hung, R.J. et al., 2008. A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. *Nature*.

Jencks, C., 1980. Heredity, environment, and public policy reconsidered. *American Sociological Review*, 45(5), pp.723–736.

van Kippersluis, H. & Rietveld, C.A., 2017. Pleiotropy-robust Mendelian randomization. *International Journal of Epidemiology*, pp.1–10.

Koellinger, P.D. & Harden, K.P., 2018. Using nature to understand nurture: Genetic associations show

how parenting matters for children's education. *Science*, 359(6374), pp.386–387.
Kong, A. et al., 2018. The nature of nurture: Effects of parental genotypes. *Science*, 359(6374), pp.424–428.
Lambert, J.-C. et al., 2013. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nature Genetics*, 45(12), pp.1452–1458.
Lander, E.S. & Schork, N.J., 1994. Genetic dissection of complex traits. *Science*, 265, pp.2037–48.
Linnér, R.K. et al., 2017. An epigenome-wide association study meta-analysis of educational attainment. *Nature Publishing Group*.
Locke, A.E.A. et al., 2015. Genetic studies of body mass index yield new insights for obesity biology. *Nature*, 518(7538), pp.197–206.
Marioni, R.E. et al., 2016. Genetic variants linked to education predict longevity. *Proceedings of the National Academy of Sciences*, 113(47), pp.13366–13371.
Martin, A.R. et al., 2017. Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. *American Journal of Human Genetics*, 100(4), pp.635–649.
Nuffield Council on Bioethics, 2002. *Genetics and human behaviour: the ethical context*, London: Nuffield Council on Bioethics [http://nuffieldbioethics.org/wp-content/uploads/2014/07/Genetics-and-human-behaviour.pdf].
Okbay, A., Baselmans, B.M.L., et al., 2016. Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses. *Nature Genetics*, 48(6), pp.624–633.
Okbay, A., Beauchamp, J.P., et al., 2016. Genome-wide association study identifies 74 loci associated with educational attainment. *Nature*, 533(7604), pp.539–542.
Parens, E. & Appelbaum, P.S., 2015. An introduction to thinking about trustworthy research into the genetics of intelligence. *Hastings Center Report*, 45(S1), pp.S2–S8.
Pickrell, J.K. et al., 2016. Detection and interpretation of shared genetic influences on 42 human traits. *Nature Genetics*, 48(7), pp.709–717.
Rietveld, C.A. et al., 2013. GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. *Science*, 340(6139), pp.1467–1471.
Ripke, S. et al., 2014. Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, 511(7510), pp.421–427.
Ross, C.E. & Wu, C., 1995. The links between education and health. *American Sociological Review*, 60(5), pp.719–745.
Sacerdote, B., 2007. How Large are the Effects from Changes in Family Environment? A Study of Korean American Adoptees. *The Quarterly Journal of Economics*, 122(1), pp.119–157.
Sacerdote, B., 2011. Nature and Nurture Effects On Children's Outcomes: What Have We Learned From Studies of Twins And Adoptees? In J. Benhabib, A. Bisin, & M. O. Jackson, eds. *Handbook of Social Economics*. Elsevier/North-Holland, pp. 1–29.
Savage, J.E. et al., 2017. GWAS meta-analysis (N=279,930) identifies new genes and functional links to intelligence. *bioRxiv [https://doi.org/10.1101/184853]*.
Schmitz, L.L. & Conley, D., 2017. The effect of Vietnam-era conscription and genetic potential for educational attainment on schooling outcomes. *Economics of Education Review*, 61, pp.85–97.
Silventoinen, K. et al., 2004. Heritability of body height and educational attainment in an international context: comparison of adult twins in Minnesota and Finland. *American Journal of Human Biology*, 16(5), pp.544–555.
Sniekers, S. et al., 2017. Genome-wide association meta-analysis of 78,308 individuals identifies new loci and genes influencing human intelligence. *Nature Genetics*, 49(7), pp.1107–1112.
Thorgeirsson, T.E. et al., 2008. A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. *Nature*, 452(7187), pp.638–642.
Tillmann, T. et al., 2017. Education and coronary heart disease: Mendelian randomisation study. *BMJ (Online)*.
Trampush, J.W. et al., 2017. GWAS meta-analysis reveals novel loci and genetic correlates for general

cognitive function: A report from the COGENT consortium. *Molecular Psychiatry*, 22(3), pp.336–345.
Turkheimer, E., 2000. Three laws of behavior genetics and what they mean. *Current Directions in Psychological Science*, 9(5), pp.160–164.
Turley, P. et al., 2018. Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nature Genetics*, 50(2), pp.229–237.
Vassos, E. et al., 2017. An Examination of Polygenic Score Risk Prediction in Individuals With First-Episode Psychosis. *Biological Psychiatry*, 81(6), pp.470–477.
Visscher, P.M. et al., 2017. 10 Years of GWAS Discovery: Biology, Function, and Translation. *The American Journal of Human Genetics*, 101(1), pp.5–22.
Warrier, V. et al., 2016. Genetic overlap between educational attainment, schizophrenia and autism. *bioRxiv [https://doi.org/10.1101/093575]*.
Winkler, T.W. et al., 2014. Quality control and conduct of genome-wide association meta-analyses. *Nature Protocols*, 9(5), pp.1192–1212.
Wood, A.R. et al., 2014. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature Genetics*, 46(11), pp.1173–1186.