Supplemental Material: Additional Methods

The causal influence of brain size on human intelligence: Evidence from within-family phenotypic associations and GWAS modeling

James J. Lee, Matt McGue, William G. Iacono,
Andrew M. Michael, Christopher F. Chabris

Study 1: Within-Family Association Between IQ and Measures of Brain Size

Sex, Age, Height, and Weight as Covariates

Earlier studies of the within-family correlation between IQ and external measures of head size have been criticized for lack of clarity over the appropriate choice of covariates (Nisbett et al., 2012). Because poorly chosen covariates can increase the bias in the estimate of a causal effect (Pearl, 2009; Lee, 2012), such criticism must be taken seriously. Here we explain why we think our particular choice of covariates is appropriate.

In the HCP data, some families include full siblings of different sexes and born at different times. There is a large sex difference in mean brain volume (Ritchie et al., 2018) and possibly a small one in performance on Raven's Standard Progressive Matrices (Lynn & Irwing, 2004; Savage-McGlynn, 2012). The relevance of the sex difference in brain volume to behavioral phenotypes is currently uncertain, and thus we would like to see any relationship between brain volume and g observed in each sex. We chose to standardize both brain volume and IQ within each sex separately, as this eliminates not only eliminates the large sex difference in mean brain volume but also the large difference in variance. Including sex as a covariate in an analysis of non-standardized variables led to larger apparent effects (results not shown).

Brain volume does vary with age, and Raven's Standard Progressive Matrices has not been age-normed in the HCP data. Since the older sibling may then differ from the younger in both brain volume and IQ solely because by virtue of being older, it is reasonable to treat age as a confounder. In all analyses where age varies within families, we used the first three powers mean-centered age as covariates.

Jensen and Johnson (1994) used the first three powers of height and weight as additional covariates. These investigators did recognize, however, that such a procedure may be overly conservative. Here is an additional reason for potential downward bias that they did not consider. Lower intelligence may in fact be a cause of being overweight, although O'Connor and Price (2018) did not find statistically significant evidence of such a relationship with the LCV method. If this causal relationship does hold, then the use of weight as a covariate may be equivalent to the unblocking of a collider, inducing negative correlations between brain size and other causes of g that may bias the estimate of the causal effect. In the primary analyses whose results are given in Table 2, we did not use height and weight as covariates.

As a robustness check, we did include the first three powers of mean-centered height and weight as additional covariates and obtained similar results (Table S3). This methodological choice thus does not appear to be especially consequential.

Study 2: Causal Inference Based on GWAS Data

Bivariate LD Score Regression

Respective Properties of LDSC and GREML

We here compare the assumptions underlying bivariate LD Score regression (LDSC; the method that we used to compute genetic correlations) to those underlying the genomic relatedness-matrix restricted maximum-likelihood (GREML) method (Lee, Yang, Goddard, Visscher, & Wray, 2012). The latter can also be used to calculate genetic correlations, although it requires individual-level data and thus will not work with GWAS summary statistics. GREML has been successfully used in many applications, including a study showing that both twin analyses and GREML produce similarly large estimates of the genetic correlations between different mental tests (Trzaskowski et al., 2013).

We proceed through the numbered list of summary points at the end of Ni, Moser, Schizophrenia Working Group of the Psychiatric Genomics Consortium, Wray, and Lee (2018), a recent work comparing the two methods.

1. If LDSC and GREML estimates are dissimilar, then Ni et al. (2018) recom-

mend that the estimate with the lower standard error (SE) should be regarded as the primary output. Since we lack an individual-level dataset with both genetic data and measures of the relevant phenotypes, we unfortunately cannot follow this suggestion. The points below, however, convince us that our LDSC estimates are valid.

2. If the number of SNPs surviving quality-control filters and employed in the analysis is small, then the standard error of LDSC estimates can become much larger than those of GREML. In all of our calculations using LDSC, the number of SNPs ranged from 1,044,377 to 1,167,435. This was always at least 79 percent of all HapMap3 SNPs with LD Scores precomputed by the developers. Ni et al. (2018) simulated no more than 800,000 SNPs in their comparison of the two methods and found that at this point the ratio of LDSC to GREML standard errors had already approached an asymptote of roughly 1.5.

   Since it is unlikely that any individual-level genetic dataset with the relevant phenotypes and a sample size in the several hundreds of thousands currently exists, we do not regard it as possible at the current time to realize this theoretical reduction in the SE.

3. When used to estimate SNP-based heritability, both GREML and LDSC can be biased as for a number of reasons, including the existence of some relationship between causal effect size and LD Score over SNPs (Speed, Hemani, Johnson, & Balding, 2012; Lee & Chow, 2014; Lee, McGue, Iacono, & Chow, 2018a). However, when used to calculate the genetic correlation (the genetic covariance divided by the square root of the two heritabilities), both methods are quite robust as a result of biases cancelling from numerator and denominator, as acknowledged by Ni et al. (2018).

4. LDSC requires that the reference sample used to calculate the LD Scores closely match the sample providing the GWAS statistics in ancestral background. If the match is insufficiently close, LDSC can be biased enough as a result of differing LD structure to warrant the use of GREML in a smaller sample of individual-level data where both phenotypes are available.

   In our application, we used the precomputed LD Scores provided by the developers, based on the European samples in 1000 Genomes (Bulik-Sullivan et al., 2015). To get some notion of how well this reference sample matches the meta-analytic sample in the GWAS of *EduYears*, we looked to the one cohort (the UK Biobank) contributing a majority of the individuals in the GWAS summary statistics of *EduYears*. The developer-provided LD Scores from 1000 Genomes

show a correlation of 0.946 with LD Scores from the UK Biobank (Ni et al., 2018). With this level of match in LD between reference and GWAS samples, LDSC was shown to provide an unbiased estimate of the genetic correlation in simulations and real-data results similar to those of GREML.

To further ensure the robustness of our own results, we reran the LCV implementation of LDSC, using *EduYears* summary statistics based on the UK Biobank and produced by the BOLT-LMM software (Loh, Kichaev, Gazal, Schoech, & Price, 2018). That is, we based the summary statistics of *EduYears* exclusively on a sample whose LD Scores are known to be highly similar to the precomputed LD Scores that we employed in our analyses. We continued to use the same GWAS summary statistics for intracranial volume (ICV). We obtained very similar results ($r_g = 0.43$, $SE = 0.08$; GCP = 0.78, $SE = 0.15$); the assumption of reasonable similarity in LD between reference and GWAS samples seems to be satisfied well enough.

5. Ni et al. (2018) point out that "the distribution of causal variants and pleiotropic effects may different across heterogeneous sources such that the estimates can be biased (capturing only common effects between heterogeneous sources)" (p. 1192). We do not regard any such bias in our case as a serious limitation. If we are examining the effects on brain size and cognition that are most consistent across settings, then the resulting estimate may be in fact quite appealing to investigators seeking reason for further biological and evolutionary inquiry.

6. Ni et al. (2018) argue that an advantage of an individual-level method such as GREML is greater precision in the partitioning of heritability and genetic covariance. We have raised the partitioning of the genetic covariance between brain size and intelligence as a future research direction (Lu et al., 2017); not having undertaken this task in the current investigation, we do not consider further the relative merits of individual-level and summary-statistic approaches to this matter.

Population Stratification

One strength of bivariate LD Score regression as a method for estimating the genetic correlation is that it can absorb the contribution of confounding (and other biases such as sample overlap) into the intercepts of the relevant regressions. This is an elegant and powerful use of the method, although GREML can be extended in this direction as well (Yang et al., 2011; de Vlaming, Johannesson, Magnusson, Ikram, & Visscher, 2017). The key assumption underlying this aspect of LD Score regression

is that the extent to which a single-nucleotide polymorphism (SNP) is confounded with environmental factors affecting one or the other phenotype is unrelated to its LD Score.

Bulik-Sullivan et al. (2015) found that a SNP's LD Score is essentially uncorrelated with its F_{ST}, a measure of population divergence in allele frequency. If generalizable, this finding would indeed provide strong evidence that LD Score regression provides robust estimates even in the face of *population stratification*—the sampling of the study cohort from groups differing in both allele frequency and exposure to environmental factors affecting the trait. This study, however, only examined populations from Northern Europe (see its Supplementary Table 2). More recent work suggests that a SNP's LD Score is indeed predictive of F_{ST} defined with respect to populations from Northern and Southern Europe respectively; this correlation seems to have led to some biases in follow-up analyses based on the GWAS of height (Berg et al., 2018). The GWAS of *EduYears* used in our investigation drew upon European populations of varying latitudes, ranging from Iceland to Sardina (Okbay et al., 2016; Lee et al., 2018b), and thus one may reasonably worry that LD Score regression will not properly remove any impact of population stratification from the summary statistics of this GWAS.

To address the possibility of uncorrected population stratification, we turned to our robustness check replacing the full *EduYears* summary statistics with those based on the UK Biobank and produced by BOLT-LMM (Loh et al., 2018). (Recall that we already used this check to address the possibility of a poor match in LD between GWAS and reference samples.) It has been shown that these GWAS statistics are largely free of the bias that renders alleles more common in Northern Europe spuriously associated with height (Sohail et al., 2018). As stated earlier, with respect to the genetic correlation between ICV and *EduYears*, we obtained very similar results ($r_g = 0.43$, $SE = 0.08$; GCP $= 0.78$, $SE = 0.15$). This suggests that population stratification in the GWAS of *EduYears* is not a serious concern in our analyses.

Recently it has been reported that GWAS of *EduYears*, our proxy for intelligence, suffer from a peculiar form of confounding: an individual derives his genetic material from his parents, whose *EduYears* appears to then influence the individual's own *EduYears* through an environmental mechanism (Sacerdote, 2007; Kong et al., 2018; Bates et al., 2018; Belsky et al., 2018). Intuitively, however, confounding of this kind should merely inflate the coefficients of SNPs that are truly associated with the trait by amounts proportional to their true coefficients (Lee, 2012) and thus leave the results of bivariate LD Score regression unaffected. We have recently carried out a more detailed analysis to confirm this intuition (Lee et al., 2018a).

The GWAS of ICV used in our investigation did not explicitly report the country-

level ancestry of its European-ancestry cohorts (Hibar et al., 2015). Nevertheless any latitudinal gradient in ICV across Europe is very unlikely to explain the moderate ICV-*EduYears* genetic correlation that we calculated. If such a gradient coincided with that of height, then the genetic correlation between ICV and height would be inflated. We used bivariate LD Score regression to calculate this genetic correlation, using the height GWAS summary statistics that are known to suffer from a bias rendering alleles more common in Northern Europe spuriously associated with height (Wood et al., 2014). We obtained an estimate of 0.24 with a standard error of 0.06, in good agreement with a more recent and larger GWAS of ICV (Adams et al., 2016). Genetic correlations are on average slightly larger than their corresponding phenotypic correlations (Sodini, Kemper, Wray, & Trzaskowski, 2018), and the sex-averaged phenotypic correlation between height and brain volume in our HCP dataset is in fact 0.24. The sex-averaged phenotypic correlation between height and head circumference among the 17-year-olds in our MCTFR dataset is 0.29. In summary, we found no evidence of population stratification in the ICV GWAS leading to severe upward biases in estimates of genetic correlations with other traits.

Latent Causal Variable

Intuition and Background

We first provide an extremely informal explanation of the Latent Causal Variable (LCV) method (O'Connor & Price, 2018). We then provide a somewhat more formal explanation and additional methodological details.

The domino model of causality that underlies frameworks such as that of Pearl (2009) provides excellent intuition here. If the falling of domino 1 knocks down domino 2, then we expect any domino (e.g., domino 0) that knocks down 1 to invariably knock down 2 as well. But suppose that there are other chains of dominos leading to 2, chains that do not pass through 1. Knocking down the dominos in such a chain will lead to the falling of 2, but not of 1.

This simple model provides exactly the intuition behind LCV. Suppose that brain size (domino 1) does indeed affect intelligence (domino 2). Then a given SNP in the genome (domino 0) with an effect on brain size must show some downstream effect on intelligence as well. But there are certainly many causes of intelligence or years of education that do not act through brain size; perhaps one of these is individual differences in the time course of synaptic plasticity. There should thus be many SNPs affecting intelligence that will show no evidence of having affected brain size.

The LCV method implements these notions by producing a numerical summary of whole-genome statistics that is positive if the statistics support brain size affecting

intelligence and negative if the statistics support the reverse causal direction. A value close to zero indicates that neither causal direction is well supported; we can analogize this to domino 0 setting off a forking chain that knocks down both 1 and 2, without 1 directly falling on 2 or *vice versa*.

More formally, the LCV method assumes that the existence of a heritable latent variable L that sends causal arrows to the traits X and Y. This is not necessarily a latent variable in the sense taken by some psychometricians (McDonald, 2003; Lee, 2012) but rather perhaps a concrete anatomical or physiological variable that happens not to have been measured. A given individual's breeding (genetic value) with respect to L is a linear combination of genotypes, the weights being the average effects of gene substitution on this L (Fisher, 1941; Lee & Chow, 2013). Let π denote the vector of these average effects. This gives us the causal fork

$$\begin{aligned} \text{SNP}_j &\xrightarrow{\pi_j} L, \\ X &\xleftarrow{q_X} L \xrightarrow{q_Y} Y. \end{aligned}$$

We now adopt the convention that L, X, and Y in the system above stand not for the actual variables but rather their breeding values (i.e., heritable components). This is justified because if SNP j has the effect π_j on the latent phenotype L, then it will have that same effect on the breeding value of L (i.e., the latent phenotype without its environmental component). Taking away the environmental perturbation of a trait does not change the fact that an alteration of genotype has a certain average effect on that trait. Similarly, SNP j has the effect $\pi_j q_X$ ($\pi_j q_Y$) on both the trait X (Y) itself and also on the trait's breeding value. We suppose that the breeding values L, X, and Y have been standardized.

The LCV method estimates the path coefficients q_X and q_Y. Note that the genetic correlation $r_g = q_X q_Y$. If $q_X = 1$ and $q_Y = r_g$, then the model becomes equivalent to the causal chain $\text{SNP}_j \to X \to Y$ for each j affecting X. To the extent that the path coefficient q_X approaches one (i.e., as residual genetic influences on X that are not propagated to Y begin to vanish), the *genetic causality proportion*

$$\text{GCP} := \frac{\log|q_Y| - \log|q_X|}{\log|q_Y| + \log|q_X|}$$

also approaches one. The developers consider GCP > 0.6 to indicate a good approximation of a causal relationship; in this case X is a reasonably good proxy for L, which has a causal effect on Y roughly in line with the genetic correlation between X and Y reflecting a wholly causal relationship.

We now give some rough intuition for how LCV estimates the quantities above. Let β_X be the effect of a given SNP on trait X and β_Y the effect of the same SNP on

trait Y. LCV makes use of the mixed fourth moments $E(\beta_X^3\beta_Y) = E(\beta_X^2 \times \beta_X\beta_Y)$ and $E(\beta_X\beta_Y^3) = E(\beta_Y^2 \times \beta_X\beta_Y)$, where E denotes the operator yielding the average of the quantity over all SNPs. If $G \rightarrow X \overset{b}{\rightarrow} Y$ is true, then SNPs with relatively large values of β_X^2 will also have relatively large values of $\beta_X\beta_Y$. Many SNPs with relatively large values of β_Y^2, on the other hand, may well show $\beta_X = 0$ and hence $\beta_X\beta_Y = 0$. The finding $E(\beta_X^3\beta_Y) > E(\beta_X\beta_Y^3)$ then indicates stronger support for $X \rightarrow Y$ over $Y \rightarrow X$; the two mixed fourth moments being close in value indicates pleiotropic confounding. (In scatterplots of the kind displayed in Fig. 1, closeness of the two mixed fourth moments has the same interpretation as both panels having the same qualitative appearance.)

Another intuitive manner of understanding the mixed fourth moments $E(\beta_X^3\beta_Y)$ and $E(\beta_Y^3\beta_X)$ is to regard them as weighted correlations over SNPs between β_X and β_Y, the weight of a given SNP being β_X^2 in one moment and β_Y^2 in the other. $E(\beta_X^3\beta_Y) > E(\beta_Y^3\beta_X)$ thus means that the correlation $\beta_X\beta_Y$ is particularly strong when β_X^2 is large but not when β_Y^2 is large. We can thus see how this comparison of mixed fourth moments generalizes the method of Pickrell et al. (2016) to genome-wide summary statistics.

With the LCV model assumptions, one can write each mixed fourth moment as a function of the path coefficients q_X and q_Y. As defined above, the GCP is in turn a function of these path coefficients. Note that the absolute value of the GCP ranges from zero (pleiotropic confounding) to one (perfect causal chain).

The LCV developers used college completion as one of the variables in their real-data analyses. The genetic correlation between college completion and *EduYears* is indistinguishable from one (Okbay et al., 2016). With the exception of low-density lipoprotein (LDL), no trait showed a potential causal effect on college completion stronger than that of height ($\widehat{\text{GCP}} = 0.33$). LDL showed $\widehat{\text{GCP}} = 0.68$ but with a large standard error of 0.3; in any case the genetic correlation between LDL and college completion was estimated to be small (-0.13). We regard these results as satisfactory outcomes of a negative-control analysis.

Respective Properties of LCV and MR

We now discuss the respective data requirements, outputs, and assumptions of LCV and a better-known family of methods for causal inference from GWAS data called *Mendelian randomization* (MR).

Data Requirements As stated earlier, LCV requires GWAS summary statistics from the entire genome; SNPs are thus included in LCV regardless of statistical

significance. In contrast, MR methods typically recommend a restriction to SNPs clearing the threshold of genome-wide significance. If there are only a handful of such SNPs, then testing frameworks accompanying MR methods may lack the statistical power to reject the null model. In this respect LCV can be considered an advance over MR in that it draws upon genome-wide summary statistics rather than a subset of SNPs reaching statistical significance.

Outputs MR will provide an estimate of the coefficient b in the causal chain $X \xrightarrow{b} Y$. However, as discussed in more detail below, the circumstances under which this coefficient is unbiased and thus validly interpretable are relatively narrow. LCV will provide an estimate of the GCP parameter (defined above). Besides its estimate being unbiased according to the criterion of Goddard, Wray, Verbyla, and Visscher (2009) in cases where the MR estimate of b is not, we think that the LCV GCP has the advantage of being interpretable in cases where the truth lies somewhere between no causality and a perfect causal chain. For instance, suppose that X is not actually a cause of Y but rather a very good proxy for some unmeasured variable that does affect Y. This will typically result in a GCP that is less than one but greater than zero; the developers recommend GCP > 0.6 as a cutoff for a sufficiently close approximation to a causal relationship between X and Y.

Assumptions It is rather difficult to compare the respective assumptions of MR and LCV, because the former uses ascertained SNPs and the latter does not. There are also different versions of MR (e.g., Bowden, Davey Smith, & Burgess, 2015); one set of assumptions does not suffice to describe all versions. To simplify matters, we assume that the MR instrument is a polygenic score for X based on all SNPs where both alleles are common. Then the assumptions of MR can be stated as follows:

(1) Those SNPs that affect Y do so through through the causal chain $X \rightarrow Y$.

(2) SNPs have no effect on Y when X is held constant. This is the so-called "exclusion restriction": any SNPs with joint effects on X and Y must reflect the hypothesized $G \rightarrow X \rightarrow Y$ causal relation exclusively rather than the pleiotropic fork $X \leftarrow G \rightarrow Y$.

On the other hand, the assumptions of LCV can be stated thusly:

(1') There exists a subset of SNPs where each SNP j makes the contributions $\pi_j q_X$ and $\pi_j q_Y$ to its total effects on traits X and Y respectively.

(2') Other effects on the two traits have a bivariate density that is mirror symmetric across both axes.

It turns out that (1) is a special case of (1') and (2) a special case of (2'). We might thus suspect that LCV should produce an output that can be validly interpreted in every case where MR does so and in additional cases as well.

The wider applicability of LCV was convincingly borne out in the extensive simulations conducted by the developers (see their Table 1 and Supplementary Tables S2–S9), which showed that LCV produces near-zero estimates of the GCP and reasonably well-calibrated false-positive rates in many null situations where MR methods (including MR-Egger) are strongly biased, including a genetic correlation reflecting pure pleiotropy, unequal numbers of genetic sites affecting the two traits (i.e., polygenicity), and unequal GWAS sample sizes. That is, conditions such as uncorrelated pleiotropic effects, correlated pleiotropic effects, unequal polygenicity, and unequal GWAS sample size were typically found to elevate the Type I error slightly, if at all, in the case of the true GCP equaling zero. Here we describe the exceptions to this trend.

1. Given a candidate causal trait with a small GWAS sample size and hence a LD Score regression slope Z-statistic of 1.4, LCV erroneously returned positive mean values of $\widehat{\text{GCP}}$ (see the developers' Supplementary Table S6). The biased estimates averaged 0.11 when the (non-causal) genetic correlation was set to 0.2; this increased to 0.27 when the genetic correlation was set to zero. The standard deviation of $\widehat{\text{GCP}}$ dramatically increased as well. Note that low statistical power to resolve effects on X also increased the false-positive rate of MR methods (see their Supplementary Table S2), so poor performance in this case should not be regarded as a unique fault of LCV.

   We do not think the relatively small sample size of the ICV GWAS is cause for concern in our real-data analysis. First, if the genetic correlation is non-causal, then higher values seem to alleviate the bias induced by small sample size. Our estimate of the ICV-*EduYears* genetic correlation exceeding 0.4 is more than twice as large as the 0.2 simulated by the developers. Second, ICV showed an LD Score regression slope Z-statistic greater than 4, closer to the value of 5 found by the developers to be adequate.

2. LCV assumes in essence that only a single latent trait is responsible for the genetic correlation. If there are multiple latent traits of varying effect and unequal polygenicity, then the equality of the mixed fourth components following from the LCV assumptions (1') and (2') in null cases may no longer hold.

In the example simulated by the developers (see their Supplementary Table S5), there was one latent trait accounting for 15 percent of X's heritability and 35 percent of Y's heritability. A second latent trait accounted for 35 percent of X's heritability and 15 percent of Y's heritability. Shared genetic influences thus account altogether for equal proportions (50 percent) of each trait's heritability, and in this sense the true GCP is equal to zero. Making the first latent trait less polygenic than the second led to upward biases in the GCP estimates; according to the developers, "[w]e expected that LCV would produce false positives, as the intermediary with lower polygenicity would disproportionately affect the mixed fourth moments" (O'Connor & Price, 2018, p. 6). Estimates of the GCP as large as $\sim$0.50 were produced by making the second latent trait 32 times as polygenic as the first. Note that multiple latent traits of varying effect and polygenicity also increased the false-positive rate of MR methods (see the developers' Supplementary Table S5).

We do not think that the possibility of multiple latent traits with unequal effect and polygenicity gives credence to a near-zero value of the GCP in our real data. Note that the case simulated by the developers qualifies as null only as a result of fine tuning. If there is one latent trait that makes a dominant contribution to the GCP, or if all latent traits make contributions to the GCP of the same sign, then the GCP estimate may not be particularly misleading. The developers themselves note that the problem of multiple latent traits requires rather extreme parameter settings (e.g., a 32-fold difference in polygenicity) to lead to a bias from zero as large as 0.5. "Thus, proportionality violations of LCV model assumptions can cause LCV (and other methods) to produce false positives, but genetic causality remains the most parsimonious explanation for high [i.e., > 0.6] GCP estimates" (O'Connor & Price, 2018, p. 7).

We think that the effect-size distributions of infant head circumference, IQ, and *EduYears* estimated by Zhang, Qi, Park, and Chatterjee (2018) provide evidence against such extreme genetic architectures, but do not pursue this point further here.

A Recent MR Analysis of Intracranial Volume and Intelligence We consider a recent application of an MR technique (Zhu et al., 2018) to the relationship between ICV and intelligence (Savage et al., 2018). This GWAS of intelligence reported significant inferred causal effects of not only ICV on IQ but also a stronger reverse effect of IQ on ICV. We performed this analysis with the same software tool (GSMR v1.0.6), employing our *EduYears* in the place of this study's IQ. We also

augmented the 7 significant ICV SNPs with the 4 additional SNPs becoming significant upon meta-analysis with HC rather than lowering the significance threshold to 10^{-5}. We obtained numerically very similar estimates (results not shown).

Since this analysis makes uses of the same data displayed in Fig. 1, the inference that intelligence has a stronger effect on ICV than the other way around is difficult to credit.

MR is in essence a two-stage regression. If $G \xrightarrow{a} X \xrightarrow{b} Y$ is the case, then the estimate of the causal effect b is the coefficient in the regression of Y on G divided by the coefficient in the regression of X on G. If the other MR assumptions hold, it is clear that this procedure yields the desired quantity, $ab/a = b$. Conversely, an MR mistakenly treating Y as the causal variable is estimating the quantity $p[a/(ab)] = p/b$, where p is the fraction of SNPs ascertained for significant association with Y that affect Y through X. If p is about 0.1—which seems plausible from the degree of sign concordance in the right panel of Fig. 1—then $b \approx 0.3$ produces a value reasonably close to what was obtained by Savage et al. (2018). It is not necessarily the case that a significant forward result will inevitably lead to a significant reverse result, particularly if SNPs affecting Y through X are not likely to be among the first to become genome-wide significant in a GWAS of Y.

Mistakenly treating Y as the causal variable can also be regarded as a violation of the MR assumptions. The exclusion restriction (2) states that the SNPs should not affect the treatment when the exposure is held constant. At the ascertained SNPs that in reality affect Y (exposure) through X (treatment), however, manipulations of genotype will show an effect on X even if Y is experimentally clamped to a constant value. The estimate of the effect of Y on X, which should be zero in expectation, will then converge to some nonzero value as a result of the assumption violation.

The GSMR method calls a procedure called HEIDI-outlier that removes SNPs from the analysis where the ratio of regression coefficients estimating the reverse causal effect is significantly atypical. One might hope that this procedure would remove all SNPs violating the exclusion restriction—that is, all SNPs affecting ICV or both traits pleiotropically. But the default setting of HEIDI-outlier ($p < .01$) removed only 7 of the 1,044 SNPs in the right panel of Fig. 1. If at least 100 SNPs in this panel affect *EduYears* through brain size and have a correctly estimated sign of its effect on ICV—as suggested by the concordance of 57 percent—then the removal of 7 SNPs cannot greatly alter the estimate of the reverse causal effect. Indeed, the estimated effect of *EduYears* on ICV hardly changes upon turning off HEIDI-outlier (results not shown). Ours is apparently not a case where HEIDI-outlier accurately identifies and removes SNPs that violate the exclusion restriction; this may be due to the noisiness of the coefficient ratios, as a result of the relatively small GWAS

sample size for ICV.

It is also worth pointing out that the GSMR method produces estimates similar to those of MR-Egger (Zhu et al., 2018), one of the methods to which LCV was compared in the simulations of O'Connor and Price (2018). MR-Egger produced an elevated false-positive rate in a number of situations where the GCP was unbiased, including a genetic correlation reflecting pure pleiotropy rather than a causal relation. Note that LCV does not require specifying a causal direction, because the GCP takes on a negative value if it is really Y affecting X.

References

Adams, H. H. H., Hibar, D. P., Chouraki, V., Stein, J. L., Nyquist, P. A., Rentería, M. E., … Thompson, P. M. (2016). Novel genetic loci underlying human intracranial volume identified through genome-wide association. *Nature Neuroscience*, *19*(12), 1569–1582. doi:10.1038/nn.4398

Bates, T. C., Maher, B. S., Medland, S. E., McAloney, K., Wright, M. J., Hansell, N. K., … Gillespie, N. A. (2018). The nature of nurture: Using a virtual-parent design to test parenting effects on children's educational attainment in genotyped families. *Twin Research and Human Genetics*, *21*(2), 73–83. doi:10.1017/thg.2018.11

Belsky, D. W., Domingue, B. W., Wedow, R., Arseneault, L., Boardman, J. D., Caspi, A., … Harris, K. M. (2018). Genetic analysis of social-class mobility in five longitudinal studies. *Proceedings of the National Academy of Sciences USA*, *115*(31), E7275–E7284. doi:10.1073/pnas.1801238115

Berg, J. J., Harpak, A., Sinnott-Armstrong, N., Joergensen, A. M., Mostafavi, H., Field, Y., … Coop, G. (2018). Reduced signal for polygenic adaptation of height in UK Biobank. *bioRxiv*. doi:10.1101/354951

Bowden, J., Davey Smith, G., & Burgess, S. (2015). Mendelian randomization with invalid instruments: Effect estimation and bias detection through Egger regression. *International Journal of Epidemiology*, *44*(2), 512–525. doi:10.1093/ije/dyv080

Bulik-Sullivan, B., Loh, P.-R., Finucane, H. K., Ripke, S., Yang, J., Schizophrenia Working Group of the Psychiatric Genomics Consortium, … Neale, B. M. (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics*, *47*(3), 291–295. doi:10.1038/ng.3211

de Vlaming, R., Johannesson, M., Magnusson, P. K. E., Ikram, M. A., & Visscher, P. M. (2017). Equivalence of LD-score regression and individual-level-data methods. *bioRxiv*. doi:10.1101/211821

Fisher, R. A. (1941). Average excess and average effect of a gene substitution. *Annals of Eugenics*, *11*, 53–63.

Goddard, M. E., Wray, N. R., Verbyla, K., & Visscher, P. M. (2009). Estimating effects and making predictions from genome-wide marker data. *Statistical Science*, *24*(4), 517–529. doi:10.1214/09-STS306

Hibar, D. P., Stein, J. L., Renteria, M. E., Arias Vasquez, A., Desrivières, S., Jahanshad, N., … Medland, S. E. (2015). Common genetic variants influence human subcortical brain structures. *Nature*, *520*(7546), 224–229. doi:10.1038/nature14101

Jensen, A. R. & Johnson, F. W. (1994). Race and sex differences in head size and IQ. *Intelligence*, *18*(3), 309–333. doi:10.1016/0160-2896(94)90032-9

Kong, A., Thorleifsson, G., Frigge, M. L., Vilhjálmsson, B. J., Young, A. I., Thorgeirsson, T. E., … Stefansson, K. (2018). The nature of nurture: Effects of parental genotypes. *Science*, *359*(6374), 424–428. doi:10.1126/science.aan6877

Lee, J. J. (2012). Correlation and causation in the study of personality (with discussion). *European Journal of Personality*, *26*(4), 372–412. doi:10.1002/per.1863

Lee, J. J. & Chow, C. C. (2013). The causal meaning of Fisher's average effect. *Genetics Research*, *95*(2–3), 89–109. doi:10.1017/S0016672313000074

Lee, J. J. & Chow, C. C. (2014). Conditions for the validity of SNP-based heritability estimation. *Human Genetics*, *133*(8), 1011–1022. doi:10.1007/s00439-014-1441-5

Lee, J. J., McGue, M., Iacono, W. G., & Chow, C. C. (2018a). The accuracy of LD Score regression as an estimator of confounding and genetic correlations in genome-wide association studies. *Genetic Epidemiology*, *42*(8), 783–795. doi:10.1002/gepi.22161

Lee, J. J., Wedow, R., Okbay, A., Kong, E., Maghzian, O., Zacher, M., … Cesarini, D. (2018b). Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nature Genetics*, *50*(8), 1112–1121. doi:10.1038/s41588-018-0147-3

Lee, S. H., Yang, J., Goddard, M. E., Visscher, P. M., & Wray, N. R. (2012). Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. *Bioinformatics*, *28*(19), 2540–2542. doi:10.1093/bioinformatics/bts474

Loh, P.-R., Kichaev, G., Gazal, S., Schoech, A. P., & Price, A. L. (2018). Mixed-model association for biobank-scale datasets. *Nature Genetics*, *50*(7), 906–908. doi:10.1038/s41588-018-0144-6

Lu, Q., Li, B., Ou, D., Erlendsdottir, M., Powles, R. L., Jiang, T., … Zhao, H. (2017). A powerful approach to estimating annotation-stratified genetic covariance via GWAS

summary statistics. *American Journal of Human Genetics*, *101*(6), 939–964. doi:10.1016/j.ajhg.2017.11.001

Lynn, R. & Irwing, P. (2004). Sex differences on the progressive matrices: A meta-analysis. *Intelligence*, *32*(5), 481–498. doi:10.1016/j.intell.2004.06.008

McDonald, R. P. (2003). Behavior domains in theory and in practice. *Alberta Journal of Educational Research*, *49*(3), 212–230.

Ni, G., Moser, G., Schizophrenia Working Group of the Psychiatric Genomics Consortium, Wray, N. R., & Lee, S. H. (2018). Estimation of genetic correlation via linkage disequilibrium score regression and genomic restricted maximum likelihood. *American Journal of Human Genetics*, *102*(6), 1185–1194. doi:10.1016/j.ajhg.2018.03.021

Nisbett, R. E., Aronson, J., Blair, C., Dickens, W., Flynn, J. R., Halpern, D. F., & Turkheimer, E. (2012). Intelligence: New findings and theoretical developments. *American Psychologist*, *67*(2), 130–159. doi:10.1037/a0026699

O'Connor, L. J. & Price, A. L. (2018). Distinguishing genetic correlation from causation across 52 diseases and complex traits. *Nature Genetics*, *50*(12), 1728–1734. doi:10.1038/s41588-018-0255-0

Okbay, A., Beauchamp, J. P., Fontana, M. A., Lee, J. J., Pers, T. H., Rietveld, C. A., … Benjamin, D. J. (2016). Genome-wide association study identifies 74 loci associated with educational attainment. *Nature*, *533*(7604), 539–542. doi:10.1038/nature17671

Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2nd ed.). New York, NY: Cambridge University Press.

Pickrell, J. K., Berisa, T., Liu, J. Z., Ségurel, L., Tung, J. Y., & Hinds, D. A. (2016). Detection and interpretation of shared genetic influences on 42 human traits. *Nature Genetics*, *48*(7), 709–717. doi:10.1038/ng.3570

Ritchie, S. J., Cox, S. R., Shen, X., Lombardo, M. V., Reus, L. M., Alloza, C., … Deary, I. J. (2018). Sex differences in the adult human brain: Evidence from 5216 UK Biobank participants. *Cerebral Cortex*, *28*(8), 2959–2975. doi:10.1093/cercor/bhy109

Sacerdote, B. (2007). How large are the effects from changes in family environment? A study of Korean American adoptees. *Quarterly Journal of Economics*, *122*(1), 119–157. doi:10.1162/qjec.122.1.119

Savage, J. E., Jansen, P. R., Stringer, S., Watanabe, K., Bryois, J., de Leeuw, C. A., … Posthuma, D. (2018). Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence. *Nature Genetics*, *50*(7), 912–919. doi:10.1038/s41588-018-0152-6

Savage-McGlynn, E. (2012). Sex differences in intelligence in younger and older participants of the Raven's Standard Progressive Matrices Plus. *Personality and Individual Differences*, *53*(2), 137–141. doi:10.1016/j.paid.2011.06.013

Sodini, S. M., Kemper, K. E., Wray, N. R., & Trzaskowski, M. (2018). Comparison of genotypic and phenotypic correlations: Cheverud's conjecture in humans. *Genetics*, *209*(3), 941–948. doi:10.1534/genetics.117.300630

Sohail, M., Maier, R. M., Ganna, A., Bloemendal, A., Martin, A. R., Turchin, M. C., … Sunyaev, S. R. (2018). Signals of polygenic adaptation on height have been overestimated due to uncorrected population structure in genome-wide association studies. *bioRxiv*. doi:10.1101/355057

Speed, D., Hemani, G., Johnson, M. R., & Balding, D. J. (2012). Improved heritability estimation from genome-wide SNPs. *American Journal of Human Genetics*, *91*(6), 1011–1021. doi:10.1016/j.ajhg.2012.10.010

Trzaskowski, M., Davis, O. S. P., DeFries, J. C., Yang, J., Visscher, P. M., & Plomin, R. (2013). DNA evidence for strong genome-wide pleiotropy of cognitive and learning abilities. *Behavior Genetics*, *43*(4), 267–273. doi:10.1007/s10519-013-9594-x

Wood, A. R., Esko, T., Yang, J., Vedantam, S., Pers, T. H., Gustafsson, S., … Frayling, T. M. (2014). Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature Genetics*, *46*(11), 1173–1186. doi:10.1038/ng.3097

Yang, J., Manolio, T. A., Pasquale, L. R., Boerwinkle, E., Caporaso, N., Cunningham, J. M., … Visscher, P. M. (2011). Genome partitioning of genetic variation for complex traits using common SNPs. *Nature Genetics*, *43*(6), 519–525. doi:10.1038/ng.823

Zhang, Y., Qi, G., Park, J.-H., & Chatterjee, N. (2018). Estimation of complex effect-size distributions using summary-level statistics from genome-wide association studies across 32 complex traits. *Nature Genetics*, *50*(9), 1318–1326. doi:10.1038/s41588-018-0193-x

Zhu, Z., Zheng, Z., Zhang, F., Wu, Y., Trzaskowski, M., Maier, R., … Yang, J. (2018). Causal associations between risk factors and common diseases inferred from GWAS summary data. *Nature Communications*, *9*(1), 224. doi:10.1038/s41467-017-02317-2

Table S1: Descriptive statistics of the sibling samples.

	HCP	MCTFR 11 yrs	MCTFR 17 yrs
Number of individuals	1,022	2,010	688
Percentage female	54.7	62.7	48.0
Age	28.9 (3.7)	11.8 (0.4)	17.4 (0.4)
Verbal IQ	109.7 (15.2)	99.1 (13.7)	95.2 (13.5)
Non-Verbal IQ	17.0 (4.8)	107.6 (15.6)	113.9 (20.8)
Brain/head size (male)	1,268 (101)	541 (17.1)	569 (16.0)
Brain/head size (female)	1,110 (88)	541 (17.9)	552 (16.1)
MZ families	138	642	223
DZ families	79	363	121

Standard deviations are given in parentheses when appropriate. See the main text for details of the intelligence testing. The unit of Non-Verbal IQ in the Human Connectome Project (HCP) is number correct out of 24, whereas the unit of all other tests is IQ point (population $SD = 15$). The unit of brain volume is cm^3 in the HCP, whereas the unit of head circumference is cm in the Minnesota Center for Twin and Family Research (MCTFR). A "family" means a complete twin pair; there are additional twins in the HCP dataset whose co-twins are absent but whose non-twin full siblings are present.

Table S2: Within-family associations between brain/head size and IQ subtests.

Dataset	Brain/head measure	IQ subtest	$\beta \pm$ SE	p-value
HCP				
All sibs	Brain volume	Vocabulary	0.120 ± 0.039	0.002
All sibs	Brain volume	Matrices	0.125 ± 0.049	0.01
MCTFR				
All sibs	Head circumference	Verbal	0.138 ± 0.028	1×10^{-6}
All sibs	Head circumference	Performance	0.155 ± 0.044	4×10^{-4}

β is the estimated partial regression coefficient of brain/head size in a model predicting the test score with family fixed effects. Both size and IQ were standardized.

Table S3: Within-family associations between brain size and IQ controlling for body size.

Dataset	Measure	$\beta \pm$ SE	ρ	p-value
HCP				
All sibs	Brain volume	0.124 ± 0.041	0.157	0.003
`MZ`	Brain volume	0.194 ± 0.189	0.246	0.31
`DZ`	Brain volume	0.047 ± 0.111	0.060	0.67
`Non-twins`	Brain volume	0.144 ± 0.066	0.183	0.03
MCTFR				
All sibs	Head circumference	0.148 ± 0.037	0.165	7×10^{-5}
`MZ 11 yr`	Head circumference	0.106 ± 0.058	0.117	0.07
`DZ 11 yr`	Head circumference	0.225 ± 0.065	0.248	6×10^{-4}
`MZ 17 yr`	Head circumference	-0.110 ± 0.107	-0.121	0.31
`DZ 17 yr`	Head circumference	0.180 ± 0.124	0.199	0.15

β is the estimated partial regression coefficient of brain/head size in a model predicting IQ with family fixed effects. Both size and IQ were standardized. In contrast to the analyses reported in Table 2 of the main text, here we included the first three powers of height and weight as additional covariates. In the HCP dataset, height and weight were standardized separately within each sex; in the MCTFR dataset, they were standardized separately within each combination of sex and age cohort. ρ, the partial regression coefficient (β) divided by the square root of the IQ test's internal-consistency reliability.