Measurement Error Correction of Genome-Wide Polygenic Scores in Prediction Samples

Elliot M. Tucker-Drob

Department of Psychology and Population Research Center

University of Texas at Austin

tuckerdrob@utexas.edu

Abstract/Introduction

DiPrete, Burik, & Koellinger (2017; http://dx.doi.org/10.1101/134197) propose using an instrumental variable (IV) framework to correct genome-wide polygenic scores (GPSs) for error, thereby producing disattenuated estimates of SNP heritability in predictions samples. They demonstrate their approach by producing two independent GPSs for Educational Attainment ("multiple indicators") in a prediction sample (Health and Retirement Study; HRS) from independent sets of SNP regression weights, each computed from a different half of the discovery sample (EA2; Okbay et al. 2016), i.e. "by randomly splitting the GWAS sample that was used for [the GPS] construction."

Here, I elucidate how a structural equation modeling (SEM) framework that specifies true score variance in GPSs as a latent variable can be used to derive an equivalent correction to the IV approach proposed by DiPrete et al. (2017). This approach, which is rooted in a psychometric modeling tradition, has a number of advantages: (1) it formalizes the assumed data-generating model, (2) it estimates all parameters of interest in a single step, (3) is can be flexibly incorporated into a larger multivariate analysis (such as the "Genetic Instrumental Variable" approach proposed by DiPrete et al., 2017), (4) it can easily be adapted to relax assumptions (e.g. that the GPS indicators equally represent the true genetic factor score), and (5) it can easily be extended to include more than two GPS indicators. After describing how the multiple indicator approach to GPS correction can specified as a structural equation model, I demonstrate how a structural equation modeling approach can be used to correct GPSs for error using SNP heritability obtained using GREML or LD score regression to produce a correction that is equivalent to an approach recently proposed by Daniel Benjamin and colleagues. Finally, I briefly discuss what I view as some conceptual limitations surrounding the error correction approaches described, regardless of the estimation method implemented.

Instrumental Variable Analysis for Causal Inference

For the reader unfamiliar with IV analysis, I begin with a brief overview of the basics of the IV analysis. Under an IV approach, variation in an instrument induces variation in a factor that is hypothesized to have a causal effect on an outcome of interest. A key assumption (termed the *exclusion* restriction) is that, beyond the mediating role of the hypothesized causal factor, the instrument is uncorrelated with the outcome. If this assumption is correct, then an IV approach can be leveraged to produce a consistent estimate the casual effect of the hypothesized causal factor on the outcome.

An example of IV analysis is a Raising of the School Leaving Age (ROSLA) or minimum compulsory schooling policy that is implemented (or rolled out at different times) pseudo randomly across municipalities (e.g. Brinch and Galloway, 2012). The key identifying assumption (the exclusion restriction) is that the policy is not directly correlated with IQ above and beyond its effect via years of

education (EduYears). A traditional two stage least squares (2SLS) approach to estimating the causal effect of EduYears on IQ involves saving the predicted values from a regression of EduYears on regional and/or temporal variation in the ROSLA policy (Stage 1). These values are then used to predict IQ (Stage 2). Covariates (e.g. time) should be included in both stages if the exclusion restriction is only plausible conditional on them. The regression coefficient estimated in the second stage represents the local average treatment effect of a year of education on IQ.

A less commonly used method of estimating an IV analysis is in a single step as a structural equation model. Figure 1 contains a path diagram representation of this structural equation model using Reticular Action Model (RAM) notation (Boker, McArdle, & Neale, 2002). This model makes the exclusion restriction explicit, as ROSLA is not allowed to correlate with the IQ residual. Note that structural equation models can be estimated using several different methods (Muthén & Muthén, 2017), perhaps the most popular of which is maximum likelihood. For instance, bootstrapping may be used when there is concern that parameters may have non-normal sampling distributions.

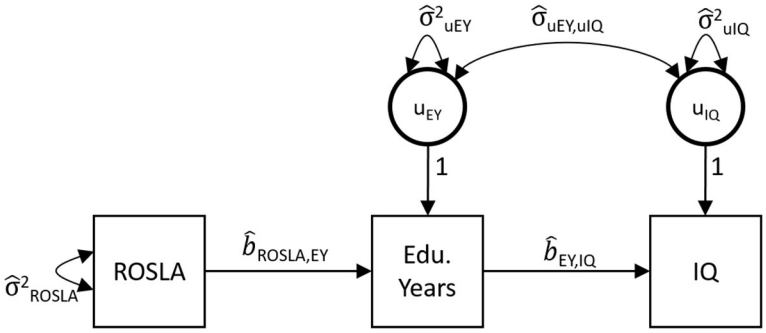

Figure 1. Note that all nonnumeric parameters in this figure are model estimates and are denoted by a hat (^). Numeric parameters are fixed values. Basic statistics that are calculated directly from the sample data have no hats.

The regression structure of this model, excluding means and intercepts, can be written as:

$$EduYears = \hat{b}_{ROSLA,EY} \times ROSLA \quad _{EY},$$

$$IQ \quad \hat{b}_{EY,IQ} \times EduYears + u_{IQ},$$

And the covariance structure of the independent variables in this model can be written as:

$$\text{Cov}\begin{bmatrix} ROSLA \\ u_{EY} \\ u_{IQ} \end{bmatrix} = \begin{bmatrix} \hat{\sigma}^2_{ROSLA} & & \\ 0 & \hat{\sigma}^2_{uEY} & \\ 0 & \hat{\sigma}_{uEY,uIQ} & \hat{\sigma}^2_{uIQ} \end{bmatrix},$$

where model-estimated parameters are denoted by a hat (^), and basic statistics calculated directly from the sample have no hats. Variances are denoted by σ, covariances are denoted by σ^2, and regression coefficients are denoted by *b*.

This model implies that

$$\hat{b}_{ROSLA,EY} \times \hat{b}_{EY,IQ} = b_{ROSLA,IQ} \quad ,$$

and

$\hat{b}_{ROSLA,EY} = b_{ROSLA,EY}$.

Therefore the key parameter $\hat{b}_{EY,IQ}$ can be calculated as the ratio of two univariate regression coefficients from the sample data, which is equivalent to the ratio of the two corresponding covariances:

$$\hat{b}_{EY,IQ} = \frac{b_{ROSLA,IQ}}{b_{ROSLA,EY}} = \frac{\sigma_{ROSLA,IQ}}{\sigma_{ROSLA,EY}}$$

Under this approach, if the ROSLA policy raises average EduYears by .5 years and the average IQ by 1.5 points, then the local average treatment effect of 1 year of education is a raise in IQ of 3 points (1.5/.5 = 3). Thus, the IV approach can be conceptualized as a form of rescaling the "treatment effect" of the instrument on the dependent variable (IQ) in units of the explanatory factor (EduYears). The $\hat{\sigma}_{uEY,uIQ}$ parameter represents unmeasured confounding between EduYear and IQ, which is independent of the instrument and causally ambiguous.

We can arrive at this same result using the 2SLS, in which the regressions are estimated separately. The expected value of EduYears from its regression on ROSLA (so called "Stage 1") is substituted into the regression Equation for IQ (so called "Stage 2") yielding what is often referred to as the reduced form equation:

$IQ = \hat{b}_{EY,IQ} \times (\hat{b}_{ROSLA,EY} \times ROSLA) + u_{IQ}$.

Given that the regressions of IQ on ROSLA and EduYears on ROSLA can be directly computed from the sample data as

$IQ = b_{ROSLA,IQ} \times ROSLA + u_{IQ}$,

and

$EduYears = b_{ROSLA,EY} \times ROSLA + u_{EY}$,

it follows that

$b_{ROSLA,IQ} = \hat{b}_{EY,IQ} \times \hat{b}_{ROSLA,EY}$,

such that

$\hat{b}_{EY,IQ} = \frac{b_{ROSLA,IQ}}{b_{ROSLA,EY}}$,

just as is implied by the structural equation model. Moving forward, I use the structural equation modeling framework, rather than the 2SLS framework.

Instrumental Variable Analysis for Error Correction

DiPrete et al. (2017) note that when multiple indicators with independent sources of error are available, these indicators can be used as instruments for one another other to correct for measurement error. Figure 2 illustrates how such an approach works in the context of the association between educational attainment (EduYears) and IQ, with EduYears measured imperfectly by two independent measures (e.g. mother and father reports of offspring EduYears).

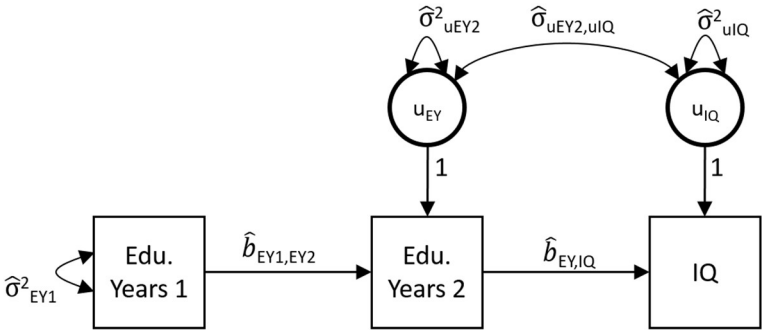

Figure 2. Note that all nonnumeric parameters in this figure are model estimates and are denoted by a hat (^). Numeric parameters are fixed values. Basic statistics that are calculated directly from the sample data have no hats.

In Figure 2, EduYears1 is used as an instrument for EduYears2 to remove measurement error from the unstandardized estimate of the EduYears→IQ association. The regression coefficient $b_{EY,IQ}$ represents the causal effect of one additional year of education on IQ in IQ points per year. The justification for the exclusion restriction is that EduYears1 is only independent of EduYears2 because of random measurement error. Therefore, EduYears1 is not expected to be associated with the IQ independent of the disattenuated effect of EduYears2.

Interestingly, in this model, the term $\sigma_{uEY2,uIQ}$ is expected to be negative, as it represents the downward bias (attenuation) of the association between EduYears2 and IQ that would have resulted from measurement error had it not been corrected for.

Under this model:

$$\hat{b}_{EY1,EY2} \times \hat{b}_{EY,IQ} = b_{EY1,IQ} ,$$

and

$$\hat{b}_{EY1,EY2} = b_{EY\ ,EY2} ,$$

Or alternatively put

$$\sigma^2_{EY1} \times \hat{b}_{EY1,EY} \times \hat{b}_{EY,IQ} = \sigma_{EY1,IQ} ,$$

which reduces to

$$\sigma_{EY1,EY} \times \hat{b}_{EY,IQ} = \sigma_{EY\ ,IQ} .$$

Therefore,

$$\hat{b}_{EY,IQ} = \frac{b_{EY\ ,IQ}}{b_{EY\ ,EY}} = \frac{\sigma_{EY\ ,IQ}}{\sigma_{EY1,EY}}$$

For example, if imperfectly measured EduYears is associated with IQ at 1 point per year and imperfectly measured EduYears predicts another imperfect measure of Eduyears at .8 points/year, then 1 year of education raises IQ by 1.25 points (1/.8 = 1.25). Importantly, in this multiple indicator context, the $\hat{b}_{EY,IQ}$ term is corrected for unreliability, but does not necessarily represent a causal effect of EduYears on IQ.

An Explicit Psychometric Model for Error Correction

Figure 3 represents an alternative modelling strategy for the same multiple indicator data. As was the case earlier, EduYears1 and EduYears2 are two imperfect measures of EduYears (e.g. mother and father reports of offspring EduYears). We create a latent variable representing true score variance in EduYears, anchoring the metric of the variable to the metric of manifest variable EduYears1 via a fixed unstandardized loading of 1. As EduYears1 and EduYears2 are expected to contain equal amounts of true score variance, we make the simplifying (but unnecessary) assumption that EduYears1 and EduYears2 load equally on the latent variable. Under this model,

$\hat{\sigma}^2_{EY} = \sigma_{EY1,EY2}$,

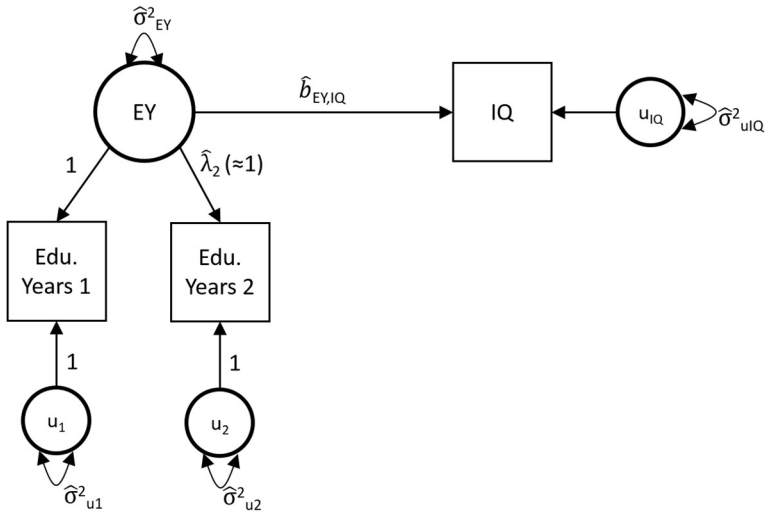

Figure 3. Note that all nonnumeric parameters in this figure are model estimates and are denoted by a hat (^). Numeric parameters are fixed values. Basic statistics that are calculated directly from the sample data have no hats.

and

$\sigma_{EY\ ,IQ} = \sigma_{EY2,IQ} = \hat{\sigma}^2_{EY} \times \hat{b}_{EY,IQ}$.

Thus

$$\hat{b}_{EY,IQ} = \frac{\sigma_{EY1,IQ}}{\sigma_{EY\ ,EY2}} = \frac{\sigma^2_{EY1} \times b_{EY1,IQ}}{\sigma^2_{EY1} \times b_{EY\ ,EY2}} = \frac{b_{EY1,IQ}}{b_{EY\ ,EY}} .$$

This estimate of $b_{EY,IQ}$ is exactly the same as that obtained from the IV error correction approach. Note here that $b_{EY,IQ}$ represents the error corrected *unstandardized regression effect* of 1 year of education on

IQ. The variance in IQ accounted for by latent EY is $\hat{\sigma}^2_{EY} \times \hat{b}^2_{EY,IQ}$, which must in turn be divided by the total variance in IQ in order to compute an R^2 value. As the variance of a latent EduYears variable is not modelled in the IV approach, this R^2 value cannot be directly computed in the IV approach.[1]

Instrumental Variable Analysis for Error Correction of GPSs

DiPrete et al. (2017) propose using the same IV error correction approach described above to obtain an estimated of SNP heritability of EduYears that is disattenuated for measurement error by producing independent genome-wide polygenic scores (GPSs) for Educational Attainment ("multiple indicators") in a prediction sample (HRS) computed from two sets of SNP regression weights, each computed from a different half of the EA2 meta-analytic sample ("by randomly splitting the GWAS sample that was used for [the GPS] construction").

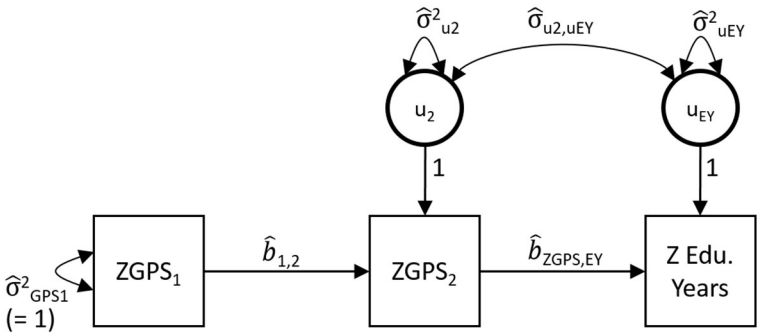

Figure 4. Note that all nonnumeric parameters in this figure are model estimates and are denoted by a hat (^). Numeric parameters are fixed values. Basic statistics that are calculated directly from the sample data have no hats.

DiPrete et al. (2017) propose using one set of GPSs as the instrument for the second set of GPSs. Crucially, because the GPSs are calculated from different halves of the parent meta-analytic sample, their errors can be plausibly assumed to be uncorrelated, allowing one GPS to serve as an instrument for the other. This approach is represented in Figure 4, which is structurally equivalent to Figure 2. However, key here is that GPS scores do not have intrinsic units, and are therefore standardized with respect to their own sample distributions. For simplicity, EduYears are also standardized in this model, such that the coefficient can be interpreted in terms of SD units of EY. Error correction works as before, such that

$$\hat{b}_{ZGPS,EY} = \frac{b_{ZGPS1,EY}}{b_{ZGPS\ ,ZGPS}} = \frac{\sigma_{ZGPS1,EY}}{\sigma_{ZGP\ \ ,ZGPS2}} = \frac{r_{ZGPS1,EY}}{r_{ZGPS1,ZGPS2}}$$

Key here is that $\hat{b}_{ZGPS,EY}$ represents the error-corrected effect of one unit of ZGPS2 on ZEduYears. As $ZGPS_2$ has been standardized with respect to the variance of $ZGPS_2$, which is itself a mixture of true genetic variance and error variance, the regression effect has been corrected for measurement error but it is still in units of a score that has been standardized with respect to observed, as opposed to latent factor, variance. Thus the square of the 2SLS regression coefficient (i.e., $\hat{b}_{ZGPS,EY}$) is not a correct estimate of the SNP heritability (see p. 9). Most generally, the square of a regression coefficient only represents R^2 (i.e.

[1] It is of further note that the IV approach requires choosing one of the two alternate forms (EY1 or EY2) as the IV. As this decision is arbitrary, it is prudent to run the IV approach both ways, and perhaps average the estimates. This is not necessary in the psychometric approach, although relaxation of the parallel alternative form assumption would allow for running the model twice, switching the anchor indicator.

SNP heritability) if it is a **standardized coefficient** from a univariate regression. ZEduYears has been appropriately standardized, but ZGPS is standardized with respect the variance of the observed indicator variable (GPS2), not the unobserved latent variable as it would need to be. As DiPret et al. (2017) explain, the corrected coefficient ($\hat{b}_{ZGPS,EY}$) must both be squared and multiplied by an estimate of the variance of the unobserved latent genetic factor to rescale it to standardized units in order to obtain an appropriate estimate of the SNP heritability. This is made explicit in the psychometric parameterization presented next.

An Explicit Psychometric Model for Error Correction of GPSs

As articulated earlier, in psychometrics multiple indicators are often leveraged as a means of error correction. In Figure 5, ZGPS1 and ZGPS2 are explicitly represented as imperfect indicators of a latent genetic factor, G, with uncorrelated errors. As in the earlier psychometric model, the metric of the latent variable (G) is linked to the metric of manifest variable EduYears1 via a fixed unstandardized loading of 1. If the two GPSs are computed from a randomly split discovery sample, we can make the simplifying (but unnecessary) assumption that the loadings are equivalent for ZGPS1 and ZGPS2. We can solve for the parameter $\hat{b}_{G,EY}$ as follows:

$$\hat{\sigma}_G^2 = \sigma_{ZGPS1,ZGP} = r_{ZGPS1,ZGPS2},$$

and

$$\sigma_{ZGPS1,EY} = \sigma_{GPS2,EY} = \hat{\sigma}_G^2 \times \hat{b}_{G,EY}.$$

Thus

$$\hat{b}_{G,EY} = \frac{\sigma_{ZGPS1,EY}}{\sigma_{ZGPS1,GPS2}} = \frac{r_{ZGPS1,EY}}{r_{ZGPS1,GPS2}},$$

which is exactly the same as the corrected $\hat{b}_{ZGPS,EY}$ estimate obtained from the IV error correction approach.

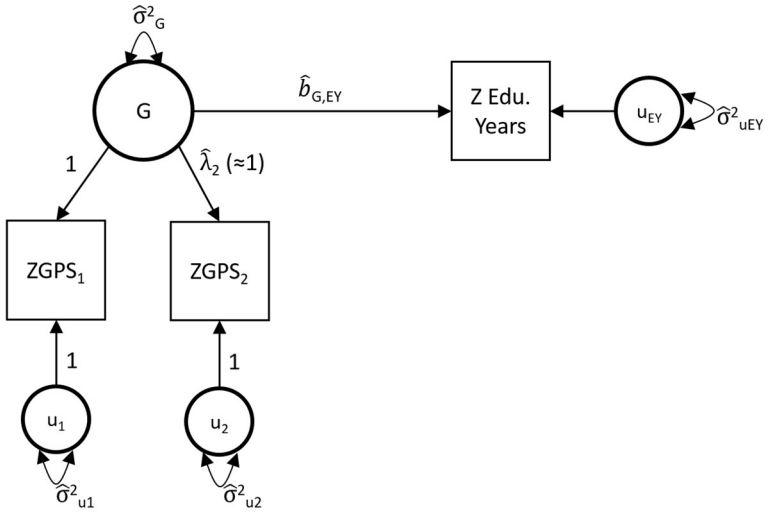

Figure 5. Note that all nonnumeric parameters in this figure are model estimates and are denoted by a hat (^). Numeric parameters are fixed values. Basic statistics that are calculated directly from the sample data have no hats.

Importantly, $\hat{b}_{G,EY}$ is an unstandardized estimate. To obtain an appropriate estimate of R^2 (i.e. SNP heritability), the variance of the independent variable (the latent variable G) must be considered. As ZEduYears is already standardized, R^2 is estimated as:

$$\hat{h}^2_{SNP} = R^2 = \hat{\sigma}^2_G \times \hat{b}^2_{G,EY} = r_{ZGPS\ ,ZGPS} \times \left(\frac{r_{ZGPS1,EY}}{r_{ZGPS1,GPS2}}\right)^2 = \frac{r^2_{ZGPS1,EY}}{r_{ZGPS1,GPS2}} .$$

Note that the above equation, in which the estimate of $\hat{\sigma}^2_G$ is equal to $r_{ZGPS1,ZGPS2}$ only holds when the loadings of $ZGPS_1$ and $ZGPS_2$ are both equal to 1. If the discovery sample has been split by convenience, rather than randomly split, the loadings of $ZGPS_1$ and $ZGPS_2$ are less likely to be equivalent. This can easily be accommodated in the psychometric model presented in Figure 5, by allowing $\hat{\lambda}_2$ to be freely estimated.

Respecifying the Psychometric Model to Standardized Units

The psychometric approach to measurement error correction can be easily re-specified in order to provide a more direct estimate of the standardized regression effect $\hat{b}_{G,EY}$ of G on the phenotype (i.e. the square root of the SNP heritability). Rather than defining the metric of G by linking it to the metric of ZGPS, we set its variance to 1, such that it is in standardized units. Again, we make the simplifying but unnecessary assumption that the loadings of ZGPS1 and ZGPS2 are equivalent. Under this specification:

$\hat{\lambda}^2 = r_{ZGPS1,ZGPS2}$,

and

$\hat{\lambda} \times \hat{b}_{G,EY} = r_{ZGPS1,EY}$,

such that

$$\hat{b}_{G,EY} = \hat{\beta}_{G,EY} = \frac{r_{ZGPS1,EY}}{\sqrt{r_{ZGPS1,ZGPS2}}},$$

which squared produces the SNP heritability:

$$\hat{h}^2_{SNP} = \hat{b}^2_{G,EY} = \frac{r^2_{ZGP\ \ ,EY}}{r_{ZGP\ \ ,ZGPS2}}.$$

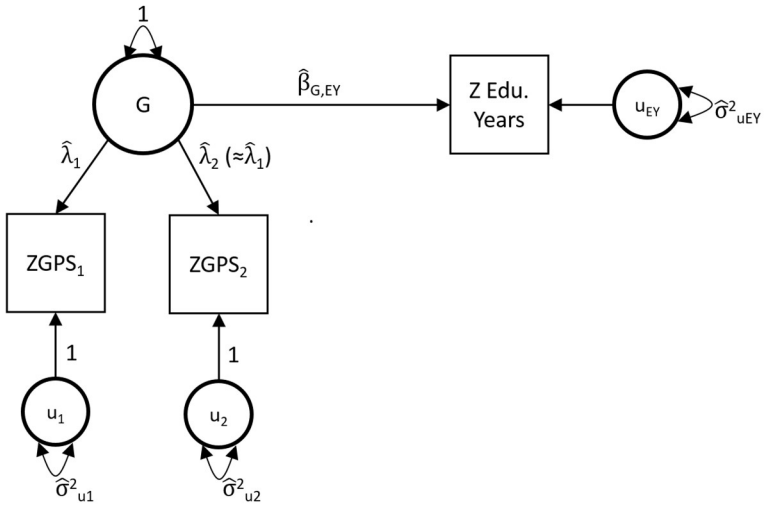

Figure 6. Note that all nonnumeric parameters in this figure are model estimates and are denoted by a hat (^). Numeric parameters are fixed values. Basic statistics that are calculated directly from the sample data have no hats.

Interestingly, this result maps on directly to the well-known psychometric correction for measurement error attenuation (Spearman, 1904; https://en.wikipedia.org/wiki/Correction_for_attenuation), where $r_{ZGPS1,ZGPS2}$ is a consistent estimate of the reliability of the GPSs, just as $r_{odd,even}$ is a consistent estimate of the reliability of the two halves of a psychometric instrument using the split-half method (https://en.wikipedia.org/wiki/Reliability_(statistics). The attenuation correction formula allows for correction due to unreliability in both variables (i.e. GPS and EY), whereas the above formula only corrects for unreliability of the GPS (the reliability of EY is implicitly assumed to be 1.0, thus dropping out of the calculation):

$$Corrected\ (r_{G,EY}) = \frac{r_{ZGPS1,EY}}{\sqrt{r_{ZGPS1,ZGPS2}} \times \sqrt{1}}$$

Broader Application of the Psychometric Approach

The psychometric approach described above provides a framework for obtaining disattenuated estimates of associations between genetic risk, G, for the discovery phenotype (e.g. EduYears) and a novel phenotype, Y (e.g. Income). Thus, it can be leveraged for much more than obtaining an estimate of the SNP heritability of the discovery phenotype (for which an estimate can already be obtained using GREML or LD score regression in the parent meta-analytic sample). The novel phenotype is simply replaced as the dependent variable. The estimated regression effect of the latent genetic factor, G, on Y ($\hat{b}_{G,Y}$) represents the disattenuated standardized effect of polygenic score for the discovery phenotype on the novel phenotype. Its square represents a disattenuated estimate of the proportion of variance explained in the novel phenotype (Income) by the polygenic score for the discovery phenotype (EduYears).

Correcting GPSs for Error Using an External Estimate of SNP Heritability

Finally, the psychometric approach provides an instructive framework for conceptualizing how information about the SNP heritability of the discovery phenotype (EduYears) that is imported from LD Score regression or GREML in the parent meta-analytic sample (e.g. EA2) can be used to disattenuate a GPSs association with a novel dependent variable (e.g. income) in a prediction sample. Key here is that SNP heritability is assumed to be equivalent in magnitude across the parent discovery dataset and prediction dataset.

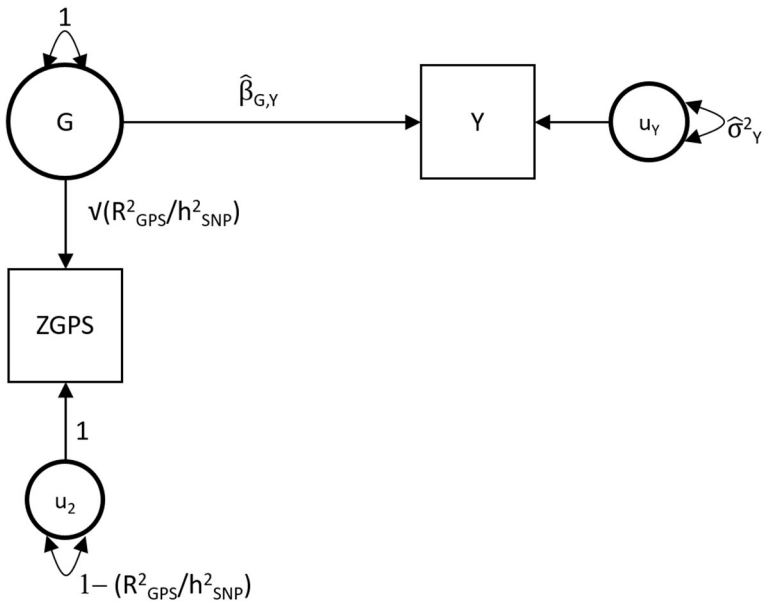

Figure 7. Note that all nonnumeric parameters with hats in this figure are model estimates and are denoted by a hat (^). The loading on G and the variance of u_2 are fixed values that are computed outside of the model.

We no longer split the parent sample in half to estimate two GPSs and instead estimate a single GPS. We compute the reliability of the GPS as the ratio of the % of variance explained in the discovery phenotype in the prediction sample (e.g. the % of variance in EduYears explained by the GPS in HRS) to the SNP heritability estimated in the parent meta-analytic sample (e.g. the SNP heritability as estimated in EA2).

$\text{Reliability}_{GPS} = R^2_{GPS\ (prediction\ sample)} / h^2_{SNP\ (discovery\ sample)}$.

This estimate is imported into a structural equation model in the prediction dataset. ZGPS is specified to load on unit variance scaled latent factor (G) at a fixed value equal to the square root of Reliability_{GPS}, and the residual variance in ZGPS is fixed to 1 - Reliability_{GPS}. The estimated regression effect ($b_{G,Y}$) of the latent genetic factor, G, on the novel phenotype Y represents the disattenuated standardized effect of polygenic score on the novel phenotype. Its square represents an estimate of the proportion of variance explained in the novel phenotype Y (e.g. Income) by the tagged SNPs relevant to the discovery phenotype (EduYears).

This model implies that

$$r_{ZGPS,Y} = \hat{\beta}_{G,Y} \times \sqrt{\frac{R^2_{GPS}}{h^2_{SNP}}}.$$

In other words, the observed association between the GPS and Y is attenuated by the square root of the GPS reliability. Solving for $\beta_{G,Y}$ yields:

$$\hat{\beta}_{G,Y} = \hat{r}_{G,Y} = \frac{r_{ZGPS,Y}}{\sqrt{\frac{R^2_{GPS}}{h^2_{SNP}}}} = \frac{r_{ZGPS,Y}}{\sqrt{Reliability(ZGPS)}} ,$$

which is equivalent the well-known psychometric correction for measurement error attenuation in the independent variable (Spearman, 1904; https://en.wikipedia.org/wiki/Correction_for_attenuation).

Although not presented in this way, this is equivalent to what Daniel Benjamin, David Cesarini, and Patrick Turley (recently implemented by Beauchamp, 2016; *Online Supplement: "Directional Selection Differentials"*) have described as scaling of the regression coefficient by the SD of the true genetic score, rather than the SD of the observed GPS. This becomes clearer as follows:

$$\hat{\beta}_{G,Y} = \hat{r}_{G,Y} = \frac{r_{ZGPS,Y}}{\sqrt{\frac{R^2_{GPS}}{h^2_{SNP}}}} = r_{ZGPS,Y} \times \sqrt{\frac{h^2_{SNP}}{R^2_{GPS}}}$$

It is important to bear in mind that the both h^2_{SNP} and R^2_{GPS} are treated as fixed terms under this approach. However, each of these terms is actually estimated with error. Therefore, the standard error of disattenuated estimate $\hat{\beta}_{G,Y}$ will be downwardly biased under this approach.

A Conceptual Note

In order to be accurate, the above formulas to correct for disattenuation of GPS associations in predictions samples rely on the assumption that the individual SNP effects composing the GPS are uniform in their degree of error, or at least that the magnitude of imprecision of the individual SNP effects is unrelated to their effect sizes. I can envision circumstances in which this may not be the case. Consider, for example, a situation in which EduYears is affected by both heritable cognitive (e.g. IQ) and noncognitive factors (e.g. conscientiousness). Further, imagine a scenario in which the SNPs relevant to IQ have higher average minor allele frequencies (MAFs), and hence their coefficients have smaller standard errors, and the SNPs relevant to conscientiousness have smaller MAFs, and hence their coefficients have larger standard errors (cf. Penke & Jokela, 2016). Thus, the GPS for EduYears will more faithfully tag the SNPs relevant to IQ than it will for conscientiousness. We might be interested in then going on use the GPS for a different prediction phenotype than EduYears, such as IQ. We would calculate GPS-IQ association in a prediction sample and disattenuate it on the basis of reliability of the GPS. However, our correction may overestimate the association by compensating for the unreliability of the conscientiousness SNPs that were relevant for EduYears even though such SNPs are not as relevant for IQ. Daniel Benjamin has suggested that this limitation may be overcome by stratifying the correction by MAF bins.

References

Beauchamp, J. P. (2016). Genetic evidence for natural selection in humans in the contemporary United States. *Proceedings of the National Academy of Sciences, 113*, 7774-7779.

Brinch, C. N., & Galloway, T. A. (2012). Schooling in adolescence raises IQ scores. *Proceedings of the National Academy of Sciences*, *109*(2), 425-430.

Boker, S. M., McArdle, J. J., & Neale, M. (2002). An algorithm for the hierarchical organization of path diagrams and calculation of components of expected covariance. *Structural Equation Modeling*, *9*(2), 174-194.

DiPrete, T. A., Burki, C., & Koellinger (2017). Genetic Instrumental Variable (GIV) Regression: Explaining Socioeconomic And Health Outcomes In Non-Experimental Data. bioRxiv 134197; doi: https://doi.org/10.1101/134197

Muthén, L.K. and Muthén, B.O. (2017). Mplus User's Guide. Eighth Edition. Los Angeles, CA: Muthén & Muthén.

Okbay, A., Beauchamp, J. P., Fontana, M. A., Lee, J. J., Pers, T. H., Rietveld, C. A., ... & Oskarsson, S. (2016). Genome-wide association study identifies 74 loci associated with educational attainment. *Nature, 533,* 539-542.

Penke, L., & Jokela, M. (2016). The evolutionary genetics of personality revisited. *Current Opinion in Psychology, 7,* 104-109.

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, *15*(1), 72-101.