

1           **Polygenic Scores for Plasticity: A New Tool for**  
2           **Studying Gene-Environment Interplay**

3           Rebecca Johnson<sup>\*\*1</sup>, Ramina Sotoudeh<sup>\*2</sup>, and Dalton Conley<sup>\*, 2, 3</sup>

4                     <sup>1</sup>Program in Quantitative Social Science, Dartmouth College

5                             <sup>2</sup>Department of Sociology, Princeton University

6                             <sup>3</sup>Office of Population Research, Princeton University

7   Last edited: June 17, 2021

---

<sup>\*\*</sup>Equal first authorship. The authors would like to thank members of the Conley Biosociology Lab and the University of Wisconsin Social Genomics workshop for helpful feedback on the project. Results from this research were presented earlier at the National Institute on Aging supported 2018 Integrating Genetics and Social Science Conference (R13-AG062366) at the University of Colorado, Boulder. Corresponding author is: Dalton Conley, [dconley@princeton.edu](mailto:dconley@princeton.edu)

8

## Abstract

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

Outcomes of interest to demographers—fertility; health; education—are the product of both an individual’s genetic makeup and his or her social environment. Yet Gene  $\times$  Environment research (GxE) currently deploys a limited toolkit on the genetic side to study gene-environment interplay: polygenic scores (PGS, or what we call mPGS) that reflect the influence of genetics on *levels* of an outcome. The purpose of the present paper is to develop a genetic summary measure better suited for GxE research. We develop what we call *variance polygenic scores* (vPGS), or polygenic scores that reflect genetic contributions to plasticity in outcomes. The first part of the analysis uses the UK Biobank ( $N \sim 326,000$  in the training set) and the Health and Retirement Study (HRS) to compare four approaches for constructing polygenic scores for plasticity. The results show that widely-used methods for discovering which genetic variants affect outcome variability fail to serve as distinctive new tools for GxE. Then, using the polygenic scores that *do* capture distinctive genetic contributions to plasticity, we analyze heterogeneous effects of a UK education reform on health and educational attainment. The results show the properties of a new tool useful for population scientists studying the interplay of nature and nurture and for population-based studies that are releasing polygenic scores to applied researchers.

24

**Keywords:** Gene-environment interactions; BMI; education; UK Biobank; HRS

## 25 1 Introduction

### 26 1.1 The growth of using genome-wide measures to study genetic moderation of 27 environments

28 A wide range of research has shown how outcomes of interest to demographers—e.g, fertility; ed-  
29 ucational attainment; diseases with marked disparities such as obesity—are influenced by both an  
30 individual’s genetic makeup and his or her social environment. In turn, this research program,  
31 also called *gene × environment* (G×E) research, has undergone a large shift in how researchers  
32 summarize genetic variation.

33 Earlier research focused on how *single* or *small sets of* genetic variants moderated social en-  
34 vironments to affect outcomes. These include studies of how polymorphisms in specific genes like  
35 *MAOA*, or the promoter region of *5-HTTP*, moderate social conditions like stressful childhood ex-  
36 periences or parental abuse (e.g., Guo et al., 2008) (for a review, see Seabrook and Avison (2010)).

37 Two developments led researchers to abandon studying how small sets of genetic variants mod-  
38 erate environments. First was the failure of many single gene G×E studies to replicate (Duncan and  
39 Keller, 2011; Keller, 2014; Border et al., 2019). Second was growing evidence that most outcomes  
40 of interest to social and behavioral scientists—educational attainment; body mass index (BMI);  
41 depression—are “polygenic,” that is, the result of small contributions of many variants across the  
42 genome, rather than “monogenic” (Boyle et al., 2017). As a result, researchers have moved away  
43 from studying how single genes or small sets of genes moderate environments to using polygenic  
44 scores (PGS) that summarize genome-wide contributions. As Section 1.3 shows, PGS have become  
45 the workhorse tool that social scientists use when studying genetic moderation of environments.  
46 As a result, large social science cohort studies—the Health and Retirement Study (Ware et al.,  
47 2018); the National Longitudinal Study of Adolescent to Adult Health (Braudt and Harris, 2020);  
48 the Wisconsin Longitudinal Study; the Fragile Families and Child Wellbeing Study—have either  
49 already released or are considering releasing polygenic scores alongside their standard survey mea-  
50 sures.

51 The proliferation of polygenic scores as the workhorse tool for studying genetic moderation  
52 of social environments raises the question: what genome-wide summary *should* researchers use?  
53 Until now, researchers have developed scores that are meant to predict the conditional mean of an  
54 outcome. One problem with then using these scores to study Gene by Environment interactions  
55 (G × E) is that the PGS used in the interaction term is constructed from a meta-analysis of levels  
56 effects across multiple cohorts that differ temporally and geographically. As a result, the PGS may  
57 be particularly ill-suited for G×E analysis since it is based on the extraction of a signal for a main  
58 effect that is common across the plausible range of environments with which researchers may seek  
59 to interact it (i.e. the multiple cohorts, countries and contexts on which it is based). By contrast,  
60 by estimating a vPGS based explicitly on variation as the estimand in the training, the score may  
61 capture signals of variation across environmental contexts. More broadly, the goal of the present  
62 paper is to expand social scientists’ methodological toolkit by presenting a new summary measure:  
63 scores summarizing genetic contributions to plasticity.

64 The remainder of the introduction proceeds as follows. First, we outline two distinct forms of  
65 genetic moderation of social environments (Section 1.2). The first is when the environment’s impact  
66 on some outcome depends on that individual’s *genetic propensity to attain that same outcome*—for  
67 instance, pre-K having a larger effect on academic outcomes among children with an already-high  
68 genetic propensity towards high educational achievement. Building on the discussion in (Domingue  
69 et al., 2020), we call this form of genetic moderation *moderation through dimming or amplifying*.

70 Second is when the environment’s impact on some outcome depends on that individual’s *propen-*

71 *sity towards variability* in an outcome. We can call this form of moderation *moderation through*  
72 *plasticity*. We argue that the majority of GxE research implicitly uses the first model of genetic  
73 moderation (moderation via dimming or amplifying). By making these implicit choices explicit,  
74 we highlight that little existing research uses tools suited for measuring the second form of genetic  
75 moderation.

76 Next, we outline how variance polygenic scores (vPGS) may capture this second form of genetic  
77 moderation (Section 1.4). We show that researchers' focus on using methods for detecting genetic  
78 contributions to plasticity have thus far largely used the methods to find "top hits," or a limited set  
79 of single nucleotide polymorphisms (SNPs) significantly associated with variability. We show our  
80 paper fills a gap by using these methods to construct genome-wide summary measures useful for  
81 GxE research, complementing other recent calls for better methods to detect genetic moderation  
82 of social environments (Domingue et al., 2020) and applications of vPGS (Schmitz et al., 2021).

## 83 1.2 Implicit models of genetic moderation: outcome moderation versus vari- 84 ability

85 Past typologies of different types of gene-environment interactions have focused on differences in  
86 the *shape* of the interaction (e.g., Boardman et al., 2014; Derringer et al., 2019). For instance,  
87 Boardman et al. (2014) and Derringer et al. (2019) each summarize three shapes of interactions:  
88 (1) diathesis-stress, where those with both a risky genotype and a highly-stressful environment  
89 have adverse outcomes; (2) vantage-sensitivity, where those with a less risky genotype and a low-  
90 stress environment have particularly good outcomes; and (3) cross-over or differential susceptibility,  
91 where those with a risky genotype have adverse outcomes in high-stress environments but also have  
92 some of the best outcomes in supportive environments (Boyce and Ellis, 2005; Ellis et al., 2011).  
93 Researchers investigating genetic moderation of environments distinguish between these shapes  
94 through both theory and the form the interaction effect takes—for instance, diathesis-stress having  
95 a crossover shape.

96 Yet shape is only one dimension of how genotypes can moderate environments. The second  
97 dimension, which occurs regardless of shape, is what form of genetic variation moderates the impact  
98 of the environment on some outcome. Here, we review two types.

### 99 1.2.1 Moderation through dimming and amplifying

100 The first type of interaction, coined by Domingue et al. (2020), is moderation through an indi-  
101 vidual's genotype dimming or amplifying an environment. This occurs when a social environment  
102 either impedes or removes an impediment to the expression of a genetically-influenced outcome.  
103 For instance, people may vary in their genetic propensity to complete formal schooling (Lee et al.,  
104 2018). However, in certain societies, there may be limited access to schooling for the population or  
105 some subgroup within the population (e.g., access to higher education was limited for women for  
106 much of the twentieth century in the U.S. and elsewhere). If that constraint is removed, individ-  
107 uals' genetic propensities towards higher education that had enjoyed no avenue for expression can  
108 then become manifest. In this case, we would expect a significant interaction in a model where a  
109 person's years of schooling is regressed on (1) an indicator for the cohorts impacted by education  
110 reform and (2) a summary measure of a person's genetic propensity to complete formal schooling.  
111 The coefficient between the genetic summary measure and reform would be null or smaller in the  
112 pre-reform years; it would become significant and positive during the post-reform years.<sup>1</sup>

<sup>1</sup>Put differently, a polygenic score trained in societies where those constraints were attenuated or absent would poorly predict education before an expansion of schooling and then predict in an improved way — i.e. show increased

113 As another example, economic changes in the U.S. have removed caloric constraints for much  
114 of the population. Genetic predispositions towards higher BMI now interact with an altered food  
115 environment (Guo et al., 2015; Conley et al., 2016). Those with a genetic predisposition towards  
116 higher BMI, which stems from a genetic architecture in part related to regulation of appetite and  
117 impulse control and in part related to metabolism (Locke et al., 2015), are more highly impacted  
118 by the new food environment.

### 119 1.2.2 Moderation through plasticity

120 The case of BMI, however, also highlights a different form of genetic variation that can interact with  
121 changes to the environment. Individuals may vary not only in their propensity towards *higher* or  
122 *lower* BMI, but also vary in their propensity towards *changes in BMI* in the face of environmental  
123 changes. Some individuals have genotypes that are *less buffering* of environmental changes. When  
124 the environment changes (in either direction), their BMI is likely to exhibit large changes. Other  
125 individuals have genotypes that are *more buffering* of environmental changes. When they enter a  
126 more calorie-rich or more calorie-restricted environment, their BMI is less likely to change because  
127 they adapt to that environment in ways that minimize changes, regardless of where they were on  
128 the BMI distribution at baseline. A genetic predisposition towards higher or lower *levels* of BMI  
129 might be very different than a genetic predisposition towards *changes* in BMI in the face of shifting  
130 environmental conditions.

131 We call this form of genetic moderation *moderation through plasticity*. Plasticity can take two  
132 forms. First is variation in *within-individual plasticity*, which is relevant for dynamic outcomes like  
133 BMI and depression that change as individuals progress through the life course. As we discuss  
134 in the Conclusion, estimating genetic contributions to *within-person* variability is complicated by  
135 the lack of data with both large-scale cohorts that have been genotyped and repeated measures  
136 across genotyped individuals. Second, and more immediately tractable, is *population-level variation*  
137 *in plasticity*. To make more concrete, consider a shock that affects BMIs in a population—for  
138 instance, neighborhood violence that leads to more sedentary activity. While one form of gene by  
139 environment interaction might predict that those with genetic propensities towards high BMI are  
140 most impacted by the change, a plasticity-focused interaction would instead find individuals, that  
141 regardless of their propensity towards higher or lower BMI, are most sensitive to the environmental  
142 shock.

### 143 1.3 Gene $\times$ environment research using genome-wide polygenic scores has largely 144 focused on moderation using levels scores

145 The previous section showed that one form of genetic moderation of environments is *moderation*  
146 *through dimming or amplifying*: those with different propensities towards an outcome are differ-  
147 entially impacted by some environmental change. Yet in a particular context—changes to neigh-  
148 borhoods interacting with genotype to impact BMI; changes to education policy interacting with  
149 genotype to affect schooling—genetic predispositions towards greater variability may also play a  
150 role.

151 Yet researchers' workhorse tool for studying genetic moderation of environments—polygenic  
152 scores for levels of an outcome—has inadvertently narrowed their focus to outcome moderation.  
153 Researchers use a three-step process when they develop and use these scores:

---

genetic penetrance — once access to formal education was opened up.

154 **Step one:** estimate separate linear regressions of some outcome ( $Y$ ) in a large training  
155 sample, to develop weights that reflect each variant’s contribution to levels of that outcome

156 **Step two:** use the weights from step one to construct a polygenic score (PGS) in a separate  
157 sample

158 **Step three:** interact that polygenic score with some measure of “E” to study genetic mod-  
159 eration of environments

160 Researchers in step one have focused on genetic contributions to levels of an outcome, rather  
161 than genetic contributions to variability. Table 1, focusing on recent gene by environment studies  
162 that use polygenic scores, shows that the majority focus on how the impact of environments on some  
163 outcome vary among people with different propensities for that same outcome—e.g., the impact  
164 of neighborhood features on Type II diabetes having a larger impact on those with higher genetic  
165 propensities.

166 *Table 1 about here*

167 With the exception of Domingue et al. (2017), who examine how a genetic risk score for wellbeing  
168 buffers the impact of the loss of a spouse on depression, nearly all studies examine the dimming or  
169 amplifying mechanism. Furthermore, this focus on one form of genetic moderation is often *implicit*,  
170 with the researchers stating that they are studying gene by environment interactions, rather than  
171 stated as an *explicit* estimand (Lundberg et al., 2020), with the researchers stating that they are  
172 studying a particular type of gene by environment interaction. The failure to make the specific  
173 type of moderation explicit has led to missed opportunities to examine other forms of moderation.

#### 174 1.4 Variance polygenic scores as a tool for examining new forms of genetic 175 moderation

176 The implicit focus on one form of genetic moderation of social environments stems from the re-  
177 liance on one tool for  $G \times E$  research: polygenic scores trained to predict levels of an outcome. We  
178 follow others’ recent calls to expand social scientists’ toolbox for studying genetic moderation of  
179 environments. Recently, Domingue et al. (2020) discuss “dimmer-type” gene-environment interac-  
180 tions, which corresponds with outcome moderation, and “lens-type” gene-environment interactions,  
181 which take a different form.<sup>2</sup> They argue that while social scientists often frame  $G \times E$  research as  
182 wanting to study lens-type interactions, social scientists’ reliance on polygenic scores for levels of an  
183 outcome might impede their progress. As they put it: “The selection of PGS effects for examining  
184 lens-type  $G \times E$  may be particularly challenging in that we construct PGSs from GWASs that only  
185 include main effects of SNPs. If the environmental context of the participants in the GWAS sample  
186 used to construct the PGS is similar to that in the test sample used to estimate  $G \times E$  then it is  
187 unlikely to include SNPs that demonstrate lens-type patterns as the main effects of these SNPs will  
188 be close to zero” (p. 10). This call suggests that better tools for either variability moderation or  
189 “lens-type” moderation are genome-wide summary measures (PGS) constructed from weights that  
190 more closely mirror theory behind  $G \times E$ .

191 Here, we present one approach: constructing genome-wide summary measures from models that

---

<sup>2</sup>As they describe: “When considering lenses, the relative effect of a given genotype may be positive for a ‘low’ level of the relevant environmental exposure and negative for ‘high’ levels of the exposure, or vice versa” (p. 9).



192 measure genetic contributions to variability in outcomes.<sup>3</sup> In the language of statistical genetics,  
193 these models are called “vQTL analyses,” or models for detecting variance-affecting loci. In turn,  
194 researchers have developed a variety of approaches for detecting genetic contributions to variability.  
195 But thus far, the researchers have only used these approaches to find the “top SNPs”—a few SNPs  
196 that have the lowest p-values in regressions performed separately for each SNP. They have not yet  
197 used the weights from these models to construct genome-wide scores for plasticity.

198 [Rönnegård and Valdar \(2011\)](#) first coined the term vQTL to discuss genetic contributions to  
199 trait variability. One of the earliest attempts at vQTL analysis was [Yang \(2012\)](#), who operational-  
200 ize variability as a person’s Squared Z-score of a trait—the person deviates from the mean of an  
201 outcome in either direction. [Wang et al. \(2019\)](#) and others use the classic Levene’s test, which  
202 examines whether the error variance significantly differs across subgroups—in the genetics case,  
203 across the three subgroups (AA, AB, and BB) at a given variant. Yet these attempts can lead to  
204 false positives when trying to distinguish between variants that affect the mean of an outcome and  
205 variants that affect the variance.

206 Two methods aim to control for this mean-variance conflation. [Conley et al. \(2018\)](#) use sibling  
207 pairs to examine how variation in the sibling pair’s combined count of minor alleles at a locus con-  
208 tributes to that sibling pair’s standard deviation in the trait, controlling for the sibling pair’s mean  
209 levels of a trait. [Young et al. \(2018\)](#) decompose trait variance into two components—an “additive  
210 effect” and a “dispersion effect”—and argue that the latter provides a measure of “when a SNP has  
211 a variance effect beyond that which can be explained by a general mean-variance relationship” (p.  
212 1613).<sup>4</sup>

213 There is a significant gap in the use of these methods to study gene-environment interplay.  
214 Researchers have used each method to find “top hit” loci that contribute to variability in traits  
215 like BMI ([Yang, 2012](#); [Conley et al., 2018](#); [Young et al., 2018](#)).<sup>5</sup> Some such as [Wang et al. \(2019\)](#),  
216 [Young et al. \(2018\)](#), and [Marderstein et al. \(2020\)](#) have also interacted these highly significant  
217 single SNPs one by one with measures of social environments. No studies of which we are aware  
218 have explored whether the “variance weights” that these methods generate can be aggregated to  
219 produce what we call *variance polygenic scores*, or genome-wide summary measures of a person’s  
220 plasticity. vPGS can expand demographers’ toolbox for studying gene-environment interplay. Our  
221 study is the first to build and characterize the properties of this new tool.

## 222 1.5 Research goals/questions

- 223 1. What are best practices for building variance polygenic scores?
- 224 2. When we build these scores, do they reflect distinctive genetic contributions to  
225 variability in a trait, or are they too correlated with scores for levels of an outcome  
226 to serve as a new tool for gene-environment research?
- 227 3. Applying the scores to a real-world example (education reform in the UK), what

---

<sup>3</sup>This approach complements the approach in [Boardman et al. \(2014\)](#) of studying genetic moderation of specific environmental shocks. In particular, in their study, they use what they call a genome-wide gene-by-environment interaction (GWGEI) approach that regresses level of an outcome (BMI) on each SNP’s interaction with an environmental moderator (education). They note the promise of the approach for capturing G×E, but also challenges with statistical power.

<sup>4</sup>Other methods that we do not include in the present review because of their similarity to the four we focus on include the new deviation regression model ([Marderstein et al., 2020](#)), which models the absolute difference between an individual’s phenotype value and the phenotype medians within each genotype, and the double generalized linear model (DLGM) (?).

<sup>5</sup>These are SNPs with effects on the outcome that fall below some p-value threshold.

228 forms of moderation do we see?

## 229 2 Methods

### 230 2.1 Estimating vPGS weights in the UK Biobank

231 To build the two types of scores for comparison—a typical PGS (hereafter: mPGS) for levels of  
232 an outcome; a vPGS for variability in an outcome—we use the UK Biobank, a dataset containing  
233 about 500,000 individuals from across the United Kingdom. The sample was limited to respondents  
234 who passed quality control and were of British ancestry, using information provided by the UK  
235 Biobank, leaving us with 408,219 in our final analytic sample. Further information regarding  
236 sample construction and quality control can be found in Online Supplement Section [S.1](#).

237 This size of the UK Biobank allows us to divide the sample into training and test sets while  
238 still maintaining sufficient statistical power for fitting GWAS and vQTL. Training and test sets  
239 were produced by randomly sampling respondents. 80% of the British subsample of the UKB was  
240 included in the training set; the remaining 20% made up the test set.

241 We analyze four outcomes: height, body mass index (BMI), educational attainment, and number  
242 of children ever born, a measure of fertility. The inverse normal transformations of the outcomes  
243 were calculated. Traits were also z-scored to create a second set of dependent variables, used in  
244 the Squared Z-score analyses. Unless specified as z-scored, a trait/outcome should be assumed to  
245 be inverse normal transformed.

246 For each outcome, first a regression was run predicting the inverse normal of that outcome, such  
247 that the weights reflect the contribution of each genetic locus to the mean level of the outcome.  
248 These regressions were performed using the software PLINK (version 1.9), controlling for age, sex,  
249 array, and the first 40 PCs.<sup>6</sup> We refer to these regression weights as weights for `Levels` PGS, and  
250 they correspond to the traditional tools used in  $G \times E$  research.

251 A set of second identical regressions predict the Squared Z-score, rather than inverse normal, of  
252 the outcome, corresponding with the method for vQTL analysis discussed in [\(Yang, 2012\)](#). Again,  
253 age, sex, array, and the first 40 PCs were included as controls. Since the z-score is squared, values  
254 which are the same number of standard deviations above or below the mean will receive the same  
255 value. Thus, the regression predicts distance from the mean, rather than the mean-level itself,  
256 though, as we argue above, this will still be correlated with the mean. We refer to the weights and  
257 polygenic scores produced by these regressions as `Squared Z`.

258 Third, regressions were run for each outcome on the sibling subsample of the UK Biobank,  
259 which includes 19,294 white British sibling pairs, while controlling for the same set of covariates as  
260 above. For each sibling pair, the intra-sibling mean and SD were calculated. We then residualized  
261 the SD with the mean and used this new residualized standard deviation as our outcome variable.  
262 Since each sibling pair was represented twice in the data, we used only one member of each sibling  
263 pair in the final regressions. We refer to the weights and polygenic scores produced by this method  
264 as `Sibling SD`.

265 Fourth, a Mean-Variance vQTL analysis using Levene’s test for variance heterogeneity was run  
266 using OSCA ([www.cnsgenomics.com/software/osca](http://www.cnsgenomics.com/software/osca)) [\(Wang et al., 2019\)](#). The Levene’s test does not  
267 estimate the effect size and standard error, but rather assesses the equality of variances between  
268 sample groups (in this case, those that do and do not have a given allele). Thus, following [\(Zhang](#)  
269 [et al., 2019\)](#), OSCA re-scales the test statistics (p value) to effect size and standard error using

---

<sup>6</sup>Terms with interactions and higher order age variables ( $age^2$ ,  $age^2 * sex$ ), which have been employed in other studies, such as [\(Young et al., 2018\)](#), were excluded due to issues with multicollinearity.



270 Z-statistics. We refer to the weights and polygenic scores produced by this method as as **Levene's**.  
271 OSPA requires one to distinguish continuous from discrete controls. As such, we treated the binary  
272 variables sex and array as discrete and age and PCs as continuous.

273 Fifth, a Mean-Variance vQTL analysis using heteroskedastic linear mixed models was similarly  
274 used (Young et al., 2018). Their method produces additive (mean) and variance effects. It also  
275 allows us to derive what they term dispersion effects, which are variance effects that are independent  
276 of the mean effect. We use the weights produced by these dispersion effects in subsequent analyses,  
277 referring to them as HLMM. Age, sex, array, and the first 40 PCs were included as both mean and  
278 variance covariates.

279 Finally, to ensure that results comparing the different vPGS were due to true differences between  
280 the scores, and not due to differences that arise from the smaller sample size and lower precision in  
281 the sibling-based method, for every vQTL or GWAS analysis run on the full sample an analogous  
282 analysis was run on a randomly-chosen subsample, where the number of respondents was set to be  
283 equal to the number of sibling pairs in the UK Biobank.

## 284 2.2 Constructing vPGS in the Health and Retirement Study (HRS)

285 Using the weights from the previous step, we constructed vPGS in the Health and Retirement  
286 study. The HRS sample is restricted to (1) self-identified European Americans, who (2) pass  
287 the HRS preprocessing procedure and are within 2 standard deviations of the mean of the first  
288 two principal components of their racial/ethnic group. This leaves  $N = 10,554$  respondents in  
289 the genotyping sample. Then, we filter to respondents with at least one wave of BMI, a primary  
290 outcome, collected ( $r*bmi$ ).  $N = 5,744$  respondents remained after this exclusion.<sup>7</sup> The replication  
291 code contains details on the outcome variable construction; most notably, since the HRS is time-  
292 varying with several waves, we took the most recently observed value of the outcome for each  
293 respondent.

## 294 2.3 Analytic approach

### 295 2.3.1 Relationship between mPGS/vPGS and levels of an outcome

296 We use three tools to explore whether vPGS can capture genetic contributions to variability in an  
297 outcome, distinct from genetic contributions to levels of an outcome.

298 First, we estimate the following linear regression, where  $i$  indexes a respondent,  $PGS$  indicates  
299 the levels PGS (mPGS) or a variance PGS (vPGS), and  $Y$  is levels of the outcome trait (converted  
300 to the standard normal scale).  $X_i$  includes the first 5 principal components (PCs). Our coefficient  
301 of interest is  $\beta_1$ —we expect the levels PGS to significantly predict levels of a trait. We also conduct  
302 a robustness check where, in addition to controlling for age and sex in the *construction* of the  
303 vPGS weights using the UKB, we also control for these covariates in the regression. We find no  
304 substantive differences with these additional covariates.

305 In turn, SNPs are a mix of four types: (1) SNPs that affect neither levels of an outcome nor  
306 variance in an outcome, (2) SNPs that affect levels of an outcome but not its variance, (3) SNPs  
307 that affect variance in an outcome but not levels of an outcome, and (4) SNPs that affect both  
308 variance in and levels of an outcome. For traits that are not normally distributed, isolating SNPs of  
309 the third type is made more difficult by the fact that any SNP that affects the mean of an outcome  
310 will also affect the outcome's variance (Young et al., 2018). Here, we aim to construct plasticity

<sup>7</sup>We did this approach, rather than imputation, because the missingness was in the outcome variable rather than in a predictor.

311 scores that capture genetic contributions to variability, and that are therefore comprised of SNPs of  
312 type three (SNPs that affect variance but not the mean) purged of general mean-variance artifacts  
313 from non-normal distributions and SNPs of type four. Since the distribution of type three and type  
314 four SNPs should be constant across each of the vPGS we compare, we interpret a larger positive  
315 coefficient on the vPGS from a regression of levels of an outcome on the vPGS as evidence that the  
316 vQTL method is picking up either (1) a large share of type four SNPs relative to type three SNPs  
317 or (2) has weights that fail to adjust for the mechanical relationship between mean and variance.  
318 For these regressions, our samples are the HRS and the held-out test set of the UKB.

$$Y_i = \alpha + \beta_1 PGS_i + \gamma X_i + \epsilon_i \quad (1)$$

319 Second, we examine whether these patterns of correlation *after* constructing the vPGS in each  
320 sample are also present in the *underlying weights* that summarize each SNP's contribution. We use  
321 linkage disequilibrium score (LD) score regression (Bulik-Sullivan et al., 2015b) for two purposes.  
322 First, we use the technique to compare the heritability of levels of an outcome to the heritability of  
323 plasticity in that outcome measured using the four techniques discussed above (Squared Z; Levene's  
324 test; HLMM; sibling SD). Then, we compare the underlying genetic correlations between (1) levels  
325 and variability for each outcome, (2) across outcomes in levels and variability (Bulik-Sullivan et al.,  
326 2015b) (for a social science application of genetic correlations, see: (Wedow et al., 2018)).

327 Finally, since the analyses of heritability show very low heritability for vPGS like HLMM and  
328 sibling SD that are less confounded with levels of an outcome, we conduct two validation exercises  
329 to investigate whether the vPGS are capturing some form of plasticity and that the two scores  
330 do not just represent random noise. Using the HRS, which, unlike the UKB, has repeated mea-  
331 surements of the phenotype over time, we explore the relationship between each vPGS and two  
332 forms of plasticity. The first form of plasticity is *within-person variability*, which we measure using  
333 two versions of the within-person standard deviation in BMI: a version using raw values of BMI  
334 and a version detrended using age-specific trends. The second form of plasticity is *unexplained*  
335 *population-level variability*, which we operationalize by regressing BMI on age, sex, and the first  
336 five PCS, and then using the squared residual from that regression as the outcome.

337 Together, these analyses show that two of the polygenic scores for plasticity—one constructed  
338 using the Squared Z-score of an outcome; the other constructed using Levene's test for variance  
339 heterogeneity—fail to summarize genetic contributors unique to variability in an outcome. How-  
340 ever, two of the polygenic scores for plasticity—one summarizing “dispersion” effects; the other  
341 constructed from sibling variation—capture more distinctive genetic contributions. We focus on  
342 these two tools for the application we discuss in the next section, but include results with all four  
343 scores in the Online Supplement.

### 344 2.3.2 Comparing mPGS versus vPGS as moderators of a UK education reform

345 We use these preferred plasticity scores to study heterogeneous effects of a large-scale education  
346 reform initiated in 1972 in England, Scotland, and Wales that extended how long students were  
347 legally required to stay in school from 15 to 16 years old (Barcellos et al., 2018). Using Barcellos  
348 et al. (2018)'s regression discontinuity design, we evaluate the extent to which the two effective vPGS  
349 are able to detect different forms of genetic moderation of this educational shock than the standard  
350 levels polygenic scores. We examine two different cases using the same reform as the exogenous  
351 environmental context: one where, following Barcellos et al. (2018), we examine the moderating  
352 role that the genetic risk of obesity plays in the relationship between education on body size and  
353 another where we evaluate the influence of genetic plasticity on downstream educational outcomes.  
354 Both outcomes are affected by gene and environment interactions but differ in the kinds of GxE

355 effects they exhibit: while the effect for body size is the result of outcome moderation, the latter  
356 results from plasticity.

357 More specifically, following [Barcellos et al. \(2018\)](#), we use 2SLS to first instrument whether  
358 someone stayed in school until 16 years of age (Educ16) with whether they were younger than 16  
359 when the reform went into place (post reform), and were therefore legally required to stay the extra  
360 year. This residualized version of Educ16 is then interacted in the second step with BMI PGS to  
361 evaluate whether there is an interaction between educational attainment and genes as they affect  
362 health outcomes. We also report a reduced form regression, where PGS is directly interacted with  
363 the post reform variable.

364 The main health outcome is Body Size - a weighted combination of BMI, waist-to-hip ratio,  
365 and body fat percentage. In a separate NBER preprint, [Barcellos et al. \(2019\)](#) identified the point  
366 in the Body Size distribution at which they should have the most power to detect an effect. We use  
367 this same distributional threshold to create an Above Threshold version of Body Size, the results  
368 for which can be compared to the continuous version of the outcome.

369 In a second set of analyses, we examine the role of genes in the impact of an additional year  
370 of education on downstream educational outcomes. Here, rather than instrumenting educational  
371 attainment, we look specifically at whether the effect of one's genotype on education outcomes  
372 differs depending on whether one was born before or after the reform. This is akin to the reduced  
373 form regressions reported for BMI. We follow the previous literature ([Barcellos et al. \(2018\)](#)), which  
374 found no effect of the reform on the likelihood of attending college, and focus on the outcomes of  
375 those who left school at the age of 18 or younger. We examine the effect of the reform on four  
376 educational outcomes, previously used in the literature. First, we consider whether the respondent  
377 left school at age 16 or later, since despite the reform some students still opted out of attending  
378 college through age 16 (Left School 16 or later). Second, we consider whether respondents achieved  
379 any certifications as a result of their education (Certification). For the last two outcomes, we  
380 explore whether they achieved specific certifications: O-levels or CSE (which were equivalent and  
381 later replaced by the GCSE in 1988) and A-levels.

382 Controls, for both sets of models, include a quadratic term for the number of days that passed  
383 from when the respondent was born until the time of reform (to factor out any time trends), dummy  
384 variables for the month born, sex, age at time of assessment in days, age squared, dummy variables  
385 for country of birth, the first 15 PCs, mPGS, the interaction between those PCs and Educ 16 (or,  
386 in the reduced form, Post Reform), and the interaction between mPGS with Educ16. Triangular  
387 kernel weights were used to assign more weight to observations closer to the reform and time trends  
388 were allowed to vary before and after the reform ([Barcellos et al. \(2018\)](#)). Because we will show that  
389 the Squared Z-score and Levene's test plasticity scores fail to capture variability distinct from mean  
390 effects, we do not present them in the main text but instead present them in the Supplementary  
391 Materials (Section [S](#)).

## 392 **3 Results**

### 393 **3.1 How do the plasticity scores relate to levels of a trait?**

394 The first question that arises when using plasticity scores for GxE research is: is the plasticity  
395 score simply capturing genome-wide contributions to levels of an outcome, rather than capturing  
396 genome-wide contributions to variability in an outcome? If the plasticity score looks very similar  
397 to social scientists' standard tool for GxE research, it is less useful as a new tool for capturing  
398 distinctive forms of genetic moderation. Figure [1](#) summarizes the results of Model [1](#), or whether  
399 the plasticity score significantly predicts levels of an outcome. The left hand side shows the results

400 in the smaller sample size HRS; the right panel the results in the larger sample size UK Biobank  
401 test set. Each bar represents one of the four outcomes of interest to demographers: height; BMI;  
402 education; number of children ever born (NEB).

403 We see that, as expected, the levels PGS predict levels of an outcome. But in three out of the  
404 four traits, the Squared Z vPGS significantly predicts levels of a trait. In two of the four traits, the  
405 Levene’s test vPGS significantly predicts levels of a trait. In contrast, the sibling SD and HLMM  
406 vPGS were only significant for one out of the four traits in the HRS sample, though were significant  
407 for more traits in the UKB test set.

408 Overall, the results show that researchers hoping to use plasticity scores for gene-environment  
409 research should be careful to choose one of the tools that captures distinctive genetic contributions  
410 to plasticity apart from genetic contributions to an outcome’s mean. Online Supplement Section  
411 [S.3](#) presents additional results, which include comparing the scores’ significance when we match the  
412 sample size of the non-sibling scores to the sample size in the sibling-based analyses.<sup>8</sup>

413

*Figure 1 about here*

### 414 **3.2 What is the genetic correlation between mPGS and vPGS?**

415 The previous results show that when we aggregate weights from the different vQTL methods to  
416 produce a polygenic score for plasticity, some of the scores—most notably, the Squared Z vPGS—  
417 perform similarly to an mPGS in predicting levels of a trait. As a result, the score may be a less  
418 useful tool for examining certain forms of gene-environment interplay since they fail to capture  
419 distinctive genetic contributions to variability.

420 Here, we examine whether we can use tools aimed at using weights from mPGS to infer (1)  
421 heritability and (2) genetic correlation to examine the genetic architecture of plasticity.

422 Online Supplement Section [S.5](#) contains the results from using LD score regression to examine  
423 the *univariate heritability* of each of the four outcomes—first, levels of an outcome (replicating  
424 previous work) and second, plasticity in that outcome (extending that work). We find that the  
425 only valid estimates of heritability are for the squared Z-score, possibly due to the method requiring  
426 weights with a certain degree of precision to generate non-zero heritability. Future research should  
427 investigate better methods for estimating heritability for less well powered vQTL weights.

428 Since the squared Z-score was the only one with non-zero heritability across outcomes, we  
429 examine the genetic correlation between (1) levels of each outcome (replicating past work by [Bulik-](#)  
430 [Sullivan et al. \(2015a\)](#)), (2) plasticity in each outcome (extending that work), and (3) levels and  
431 plasticity. Notably, these genetic correlations are *prior* to estimating the scores in a sample, so  
432 reflect a shared genetic architecture between contributors to levels of an outcome and contributors  
433 to variability in an outcome.

434 The top panel of Table [2](#) shows the genetic correlation between the levels PGS and the Squared  
435 Z vPGS for each of the outcomes. It shows that the weights for the levels PGS for that trait are  
436 significantly correlated with the weights for the Squared Z vPGS.

437 The middle panel of Table [2](#) shows *between-trait* patterns of genetic correlation for (1) the levels

---

<sup>8</sup>This robustness check helps guard against us finding that the sibling SD score does not significantly predict levels of an outcome, while the non-sibling scores do, due to inadequate power for the sibling score compared to the scores estimated in a larger sample size. The fact that the patterns hold in the matched sample size supports our claim that the Squared Z and Levene’s test scores are less useful as distinctive tools.

438 PGS<sup>9</sup> and (2) the Squared Z vPGS. The analysis investigates whether there are similar patterns  
439 of cross-trait genetic correlation in variability in addition to levels. The results show that the  
440 patterns are similar except for the relationship between BMI and height. In particular, the one  
441 exception is that levels of height and BMI are negative genetically correlated (in other words, those  
442 with genetic propensities to be taller also have genetic propensities towards lower BMI, replicating  
443 the relationship in [Bulik-Sullivan et al. \(2015a\)](#)) but plasticity in height and BMI is positively  
444 correlated. As we discuss in the Conclusion, this relationship deserves more attention in future  
445 research and could reflect differential sensitivity to environmental inputs to growth.

446

*Table 2 about here*

447 The bottom panel of the table also highlights the underlying genetic correlation between an  
448 mPGS meant to purge variance effects (the additive weights from the HLMM method) and each  
449 vPGS within a trait, which shows generally negative patterns. Appendix Section [S.5](#) shows a visual  
450 summary of these correlations.

### 451 3.3 Validation that the vPGS correlates with plasticity

452 The previous sections showed that (1) especially the squared Z score vPGS was highly correlated  
453 with levels of an outcome and (2) that vPGS was the only one to have precise-enough estimates to  
454 be able to examine heritabilities and genetic correlations. Yet the noisiness of the vPGS estimates  
455 raises a question: could the results of Section [3.1](#) stem from scores that reflect random noise, rather  
456 than true contributions to variability?

457 Here, we report the results of the validation exercise discussed in Section [2.3.1](#). Figure [2](#) shows  
458 the results of relating each vPGS to within-person variability in BMI among respondents with  
459 at least three waves of BMI observations (Online Supplement Section [S.6](#) discusses details of the  
460 sample construction and shows the full regression results). We see that individuals with higher  
461 vPGS have significantly more over-time variability in BMI than individuals with lower vPGS.  
462 Online Supplement Section [S.6](#) also discusses a validation exercise where we regress the squared  
463 residual of BMI on each of the vPGS.

464

*Figure 2 about here*

### 465 3.4 Summing up thus far: which vPGS can serve as new tools for GxE?

466 Taken together, the results show that the Squared Z vPGS vPGS is less useful for social scientists  
467 looking for a new tool to examine gene-environment interplay. The vPGS significantly predicts  
468 levels of an outcome across four diverse traits (height; BMI; education; number of children ever  
469 born). The Squared Z-score also exhibits patterns of underlying genetic correlation similar to those  
470 between levels of a trait. In contrast, the sibling standard deviation method ([Conley et al., 2018](#))  
471 and dispersion weights ([Young et al., 2018](#)) show better properties in capturing distinctive genetic  
472 contributions to plasticity that appear less confounded with levels of an outcome.

473 Why might past research studying methods for vQTL have missed the ways in which certain  
474 methods fail to capture distinctive genetic effects? Section [S.7](#) in the Online Supplement begins

<sup>9</sup>These replicate results from [Bulik-Sullivan et al. \(2015a\)](#) for overlapping outcomes and also extend their analysis to look at additional outcomes like number ever born.



475 with the common way that researchers assess whether a method for detecting vQTLs overlaps with  
476 a method for detecting mean effects/normal QTLs: examining whether the two methods select  
477 similar SNPs as “top hits.”<sup>10</sup> The results show that while the top hits comparison reveals some  
478 degree of overlap—for instance, the Squared Z score and Levene’s top hits display more overlap  
479 with the levels top hits than the other methods—this comparison might be too conservative. In  
480 particular, an mPGS and vPGS might not happen to have overlap in the SNPs with p values below  
481 a threshold but the two might have overlap in SNPs with non-zero weights that contribute to the  
482 final scores. Overall, the combined results show that social scientists interested in using vPGS as a  
483 new tool should look carefully at whether the vPGS is distinctive from, or nearly identical to, the  
484 mPGS for that outcome.

## 485 4 Using the vPGS to examine heterogeneous impacts of education 486 reform

487 Having examined the properties of the different polygenic scores for plasticity, their distinctiveness  
488 from the standard tool for GxE research (mPGS), and their relationships to one another, we can  
489 use them to adjudicate between different mechanisms of genetic moderation. Specifically, we have  
490 argued that interactions between mPGS and the environment in predicting an outcome capture  
491 outcome moderation, while variance polygenic scores can capture a different form of heterogeneous  
492 effects.

493 With this in mind, we turn to using mPGS and different vPGSs in a practical example to explore  
494 which kind of genetic moderation is at play. We build upon the research of [Barcellos et al. \(2018\)](#).  
495 They investigate the impact of an educational reform that raised the required age of schooling  
496 from 15 to 16 years in England, Scotland, and Wales. Unlike measures like an individual’s own  
497 educational attainment or their parent’s educational attainment, which can lead to false-positive  
498 gene-environment interactions through confounding between the environmental shock and parent  
499 genotype (discussed in [Conley, 2016](#)), the reform’s timing is exogenous to genotype. It allows us  
500 to study the different forms of genetic moderation, as well as a chance to examine the performance  
501 of different vPGS measures in an applied example.

502 We evaluate two sets of outcomes. First, following [Barcellos et al. \(2018\)](#), we evaluate whether  
503 there was genetic moderation of the reform’s impact on health outcomes in the form of body  
504 size. If the form this moderation takes is *outcome moderation*, then the mPGS for BMI would  
505 significantly interact with the reform—the reform might have a larger impact on those with an  
506 already-low genetic propensity towards obesity (amplifying their advantage) or it might have a  
507 larger impact on those with a high genetic propensity (buffering their risk). Alternately, if the form  
508 this moderation takes is *variability moderation* (significant interaction between the vPGS and the  
509 post-reform indicator), the reform has larger impacts on those who, across many shocks, experience  
510 more swings in BMI.

511 Second, extending [Barcellos et al. \(2018\)](#), we evaluate whether there was genetic moderation  
512 of the reform’s impact on educational outcomes. Here, *outcome moderation* occurs if the reform  
513 has a larger impact on those with especially high or low genetic propensities towards educational  
514 attainment. Under the “education as the great equalizer hypothesis” ([Barcellos et al., 2020](#)), we  
515 might expect that those with the lowest educational polygenic score are the most impacted by the  
516 extra year of mandatory education. *Variability moderation* occurs if the reform has heterogeneous

---

<sup>10</sup>Researchers use this in conjunction with simulations comparing the methods, but those simulations likewise largely focus on one or two top causal SNPs.



517 effects on individuals with different underlying genetic plasticity<sup>11</sup>.

#### 518 4.1 Genetic moderation of the education reform's impact on body size

519 For the body size models, presented in Table 4, the interaction between mPGS and being exposed  
520 to the reform is the only statistically significant result. None of the polygenic scores for plasticity  
521 show a significant interaction, with the exception of Squared Z-Score vPGS (Online Supplement  
522 Section S.8, which uncovers a marginally significant result (on the Above Threshold outcome)).  
523 This result is likely due to the high correlation between the Squared Z-Score vPGS and mPGS.

524 These results suggest that outcome moderation (rather than plasticity) is the main form that  
525 genetic moderation of the education reform takes when impacting these measures of health. Put  
526 differently, and as visualized in Figure 3, the reform has larger impacts on reducing obesity-related  
527 measures among those with already-higher genetic propensities towards obesity. The results largely  
528 replicate those found in Barcellos et al. (2018), and show that the mPGS the original authors used  
529 ended up corresponding to the type of genetic moderation that unfolded.

530 *Table 3 about here*

531 *Figure 3 about here*

#### 532 4.2 Genetic moderation of the education reform's impact on educational at- 533 tainment

534 While the reform's impact on health outcomes follows the pattern of outcome moderation, the  
535 reform's impact on educational attainment might take a different form. When examining this  
536 impact, we find a significant interaction between the HLMM polygenic score for plasticity and Post  
537 Reform when predicting three of the four education outcomes: Left School 16 or later, Certification,  
538 and O-levels or CSE. The interactions between the HLMM plasticity score and Post Reform remain  
539 significant when controls are included. By contrast, we find significant interactions between mPGS  
540 and Post Reform for only one of the four outcomes: Left School 16 or later. The results are  
541 presented in Table 5 and visualized in Figure 4, which compares the predicted educational outcomes  
542 for children in the lowest, middle, and upper terciles of the HLMM vPGS distribution before and  
543 after the reform. For the significant interactions, the results show that those with higher HLMM  
544 polygenic score attain lower levels of education outcomes prior to the reform but equal levels  
545 of education outcomes after the reform, potentially because they had enhanced sensitivity to the  
546 positive effects of the reform. And because the results in Section 3.1 show that the HLMM plasticity  
547 scores capture genetic contributions to plasticity in outcomes distinct from genetic contributions  
548 to the conditional mean, we are more confident that the effect is a true positive.

---

<sup>11</sup>Here, regressions that use vPGS also control for mPGS to ensure that the observed effects do not simply reflect outcome moderation. To understand how the inclusion of mPGS affects these results, we report regressions where mPGS is not controlled for in section S.8 of the SI. There, we also report the full regression results for the GxE analyses reported in the main text.

549

Table 4 about here

550

Figure 4 about here

## 551 5 Discussion

552 Recognizing the biosocial nature of most outcomes of interest to demographers, social scientists  
553 are increasingly interested in how genetic variation moderates the impact of life-course events that  
554 range from society-wide education reforms to targeted policy interventions aimed at specific sub-  
555 groups. That is, in addition to estimating direct main effects of genotypes and environments, social  
556 and behavioral scientists often seek to model the mutual dependence of nature and nurture. Nu-  
557 merous metaphors have been offered for this causal model of human traits—e.g., genetics as a lens  
558 (Domingue et al., 2020) or genetics as a prism refracting environmental influences into heteroge-  
559 neous treatment effects (Conley and Fletcher, 2018).

560 In this paper, we argue that social scientists’ workhorse measure of  $G$  in GxE research—a ge-  
561 netic summary measure that reflects genetic contributions to levels of an outcome—commits those  
562 researchers to an implicit model of genetic moderation of environments. The model corresponds to  
563 what we call outcome moderation, and to what others have recently called “dimmer-type” moder-  
564 ation (Domingue et al., 2020). While this model may characterize some forms of gene-environment  
565 interplay, there are likely other forms of gene-environment interplay that summary measures con-  
566 structed from aggregating effects on an outcome’s mean fail to capture.

567 We propose the use of polygenic scores for plasticity as an addition to social scientists’ toolbox.  
568 We first investigate the properties of this tool before applying it. First, focused on best practices,  
569 we show how conflation between genetic effects on an outcome’s mean and effects on that outcome’s  
570 variance begin with SNP-level analyses but then appear in the constructed scores. The conflation  
571 also makes it difficult to investigate whether plasticity in outcomes like BMI displays different pat-  
572 terns of heritability than levels of those outcomes, though an initial analysis of genetic correlations  
573 shows an interest flip where levels of BMI and height are negatively genetically correlated but  
574 plasticity in the two has a positive correlation. As a whole, we argue that researchers interested in  
575 a polygenic score for plasticity as a *distinctive* summary measure of genotype should be careful to  
576 construct scores based on weights from methods that try to adjust for false positive effects on the  
577 mean.

578 Second, applying the scores to a real-world application, we show how adding an  $E \times vPGS$   
579 analysis to an  $E \times mPGS$  analysis can detect a particular type of GxE interaction that deploying  
580 only an mPGS would obscure. Building on Barcellos et al. (2018), we show that, in line with their  
581 results but contrary to our priors, *outcome moderation* best characterizes the education reform’s  
582 impact on health outcomes. But genetic plasticity might better explain the reform closing gaps  
583 in educational attainment between low and high-plasticity youth. These results show that one  
584 cannot know in advance with great certainty which form of moderation will be operative and thus  
585 researchers should test for both forms.

## 586 5.1 Limitations and directions for future research

587 The first limitation is that our application of the polygenic scores for plasticity was limited to one  $E$ :  
588 an education reform in the UK. In turn, and relevant to the theoretical discussion in Section 1.2, we  
589 might imagine that different environmental treatments are more or less likely to exhibit moderation  
590 by a person’s plasticity. Due to the issues others have raised about false positives where researchers  
591 think they are detecting GxE effects but instead are detecting unobserved confounding between  
592 genotype and environment (Conley, 2016; Domingue et al., 2020), we prioritized studying the effect  
593 of an “E” that was clearly causally identified over examining how multiple, potentially-confounded  
594 “E” interact with each of the focal vPGS. Future research leveraging other natural experiments  
595 that alter environments should incorporate plasticity scores to investigate their relevance for other  
596 contexts.

597 Second, as we outline in Section 1.2, there are at least two ways we can think about genetic con-  
598 tributions to plasticity. The first, *within-individual plasticity*, would require an estimation strategy  
599 where we train the inputs weights to the PGS on repeated measures of the same outcome within  
600 an individual (e.g., variation in BMI across many years). This form of plasticity is a promising  
601 avenue for future research. Predicting within-person (or within-family) variation over time without  
602 having to know explicitly what the fluctuating environmental factors are may prove to be a useful  
603 exploratory exercise before researchers try to hypothesize about specific factors in the environment  
604 that may be causing the fluctuation in genotypically-plastic individuals. Moreover, one could imag-  
605 ine using a within-person variability score to identify individuals who might be responsive to an  
606 intervention in advance—be that a drug trial or an educational intervention. The advantages of  
607 identifying such individuals include increased statistical power for the identification of effects in a  
608 pilot study before investing in a larger, more costly study. In terms of the feasibility of this second  
609 type of plasticity score, unfortunately, data sources like UKB that contain a large enough sample  
610 size to estimate new weights for polygenic scores lack large-scale repeated measures of the same  
611 individual. Once these data sources become available, future research can construct scores better  
612 designed for this form of plasticity.

613 Third, we might imagine two forms of plasticity. One form of plasticity is *trait specific* and  
614 occurs in response to various environmental triggers—so an individual with high BMI plasticity  
615 might have more BMI variability in response to many different environmental shocks (e.g., educa-  
616 tion reform; changes in food landscape; changes in peer group eating behavior). But another form  
617 of plasticity may be both *trait specific* and *environment specific*—so an individual may not have  
618 “generally high BMI plasticity,” but instead have high plasticity of BMI in response to a certain  
619 type of environmental trigger. The present approach to estimating variance-affecting SNPs weights  
620 and constructing *cross-environmental* vPGS captures the first type of plasticity, but fails to capture  
621 the second. For the second type of plasticity, researchers in statistical genetics are using flexible,  
622 machine learning methods to (1) focus on a specific “E” or environmental shock, (2) interact that  
623 “E” with many SNPs, (3) use regularization and other methods to find top-performing “SNP:E”  
624 interactions (for an early application noting challenges, see: Boardman et al., 2014); more recently,  
625 Frost et al. (2016) use elastic net penalized regression to zero-out many of the SNP:“E” interac-  
626 tions). The weights from those methods focused on interactions between a specific “E” and each  
627 SNP could be used to estimate vPGS specific to certain environmental triggers.

628 Fourth, the present paper focuses still on scores that can be interacted with a specific measure  
629 of environment. But twin- and other pedigree-based approaches to estimating heritability have long  
630 been deployed to estimate GxE, for example, by assessing whether the additive heritability estimate  
631 changes in the face of differing social conditions (e.g., social class background; birth cohort)(e.g.,  
632 Boardman et al., 2010; Vink and Boomsma, 2011). Newer methods such as GREML-based molec-

633 ular methods allow for similar analysis under the same general framework where a shift in the  
634 additive (SNP) heritability is evidence of GxE (e.g., [Rimfeld et al., 2018](#)). However, since these  
635 methods each take an approach of variance decomposition, these methods cannot distinguish be-  
636 tween different heritabilities due to a change in the genetic variance or a corresponding shift in the  
637 environmental variance. More importantly, by testing for changes in the total, additive heritability,  
638 as is the case for GxE studies using polygenic scores based on levels regressions, these approaches  
639 may be missing important GxE that do not result from differences in the predictive power of levels'  
640 effects. One way to think about the plasticity or variance effect as it interacts with the environment  
641 is as an “environmental (and genetic)” epistasis term. That is, it is a non-additive effect that is not  
642 captured in traditional models. The goal of using vQTL methods is to capture this non-additive  
643 effect.

644 Finally, there is growing attention to how standard polygenic scores (mPGS) do not represent  
645 “pure” genetic measures of propensities; instead, the weights reflect a combination of direct genetic  
646 effects and biases from population stratification and genetic nurture ([Kong et al., 2018](#); [Trejo and](#)  
647 [Domingue, 2019](#); [Zaidi and Mathieson, 2020](#)). Because of these issues, for generating vQTL weights,  
648 the ideal design is either having the genotypes of two or more siblings along with the parent geno-  
649 type, or having the genotype of three or more siblings and being able to add a fixed effect for the  
650 sibling pair. However, these methods require large-enough samples with one of those family-based  
651 structures. In the present paper, the sibling SD method provides one approach to addressing bias  
652 in the vQTL weights but future research should explore changes in vQTL weights when generated  
653 using a family-based design. In sum, the present article aims to equip demographers and social  
654 scientists with an additional tool for studying the interplay of genes and environment, one that  
655 captures a broader range of how these interactions play out in applied settings. As cohort stud-  
656 ies make polygenic scores available to applied researchers, our paper suggests complementing the  
657 mPGS scores they are currently releasing with scores aimed at capturing genetic contributions to  
658 plasticity.

## 659 References

- 660 Amin, V., Böckerman, P., Viinikainen, J., Smart, M. C., Bao, Y., Kumari, M., Pitkänen, N.,  
661 Lehtimäki, T., Raitakari, O., and Pehkonen, J. (2017). Gene-environment interactions between  
662 education and body mass: Evidence from the uk and finland. *Social Science & Medicine*, 195:12–  
663 16.
- 664 Barcellos, S. H., Carvalho, L. S., and Turley, P. (2018). Education can reduce health differences re-  
665 lated to genetic risk of obesity. *Proceedings of the National Academy of Sciences*, 115(42):E9765–  
666 E9772.
- 667 Barcellos, S. H., Carvalho, L. S., and Turley, P. (2019). Distributional effects of education on  
668 health. Technical report, National Bureau of Economic Research.
- 669 Barcellos, S. H., Carvalho, L. S., and Turley, P. (2020). Is education the great equalizer? *NBER*  
670 *Working Paper*.
- 671 Boardman, J. D., Blalock, C. L., and Pampel, F. C. (2010). Trends in the genetic influences on  
672 smoking. *Journal of health and social behavior*, 51(1):108–123.
- 673 Boardman, J. D., Domingue, B. W., Blalock, C. L., Haberstick, B. C., Harris, K. M., and McQueen,  
674 M. B. (2014). Is the gene-environment interaction paradigm relevant to genome-wide studies?  
675 the case of education and body mass index. *Demography*, 51(1):119–139.
- 676 Border, R., Johnson, E. C., Evans, L. M., Smolen, A., Berley, N., Sullivan, P. F., and Keller, M. C.  
677 (2019). No support for historical candidate gene or candidate gene-by-interaction hypotheses for  
678 major depression across multiple large samples. *American Journal of Psychiatry*, 176(5):376–387.
- 679 Boyce, W. T. and Ellis, B. J. (2005). Biological sensitivity to context: I. an evolutionary–  
680 developmental theory of the origins and functions of stress reactivity. *Development and psy-*  
681 *chopathology*, 17(02):271–301.
- 682 Boyle, E. A., Li, Y. I., and Pritchard, J. K. (2017). An expanded view of complex traits: from  
683 polygenic to omnigenic. *Cell*, 169(7):1177–1186.
- 684 Braudt, D. and Harris, K. M. (2020). Polygenic scores (pgss) in the national longitudinal study of  
685 adolescent to adult health (add health)–release 2.
- 686 Bulik-Sullivan, B., Finucane, H. K., Anttila, V., Gusev, A., Day, F. R., Loh, P.-R., Duncan, L.,  
687 Perry, J. R., Patterson, N., Robinson, E. B., et al. (2015a). An atlas of genetic correlations across  
688 human diseases and traits. *Nature genetics*, 47(11):1236.
- 689 Bulik-Sullivan, B. K., Loh, P.-R., Finucane, H. K., Ripke, S., Yang, J., Patterson, N., Daly, M. J.,  
690 Price, A. L., and Neale, B. M. (2015b). Ld score regression distinguishes confounding from  
691 polygenicity in genome-wide association studies. *Nature genetics*, 47(3):291–295.
- 692 Conley, D. (2016). Socio-genomic research using genome-wide molecular data. *Annual Review of*  
693 *Sociology*, 42:275–299.
- 694 Conley, D. and Fletcher, J. (2018). *The Genome Factor: What the social genomics revolution*  
695 *reveals about ourselves, our history, and the future*. Princeton University Press.

- 696 Conley, D., Johnson, R., Domingue, B., Dawes, C., Boardman, J., and Siegal, M. (2018). A sibling  
697 method for identifying vqtls. *PloS one*, 13(4).
- 698 Conley, D., Laidley, T. M., Boardman, J. D., and Domingue, B. W. (2016). Changing polygenic  
699 penetrance on phenotypes in the 20 th century among adults in the us population. *Scientific*  
700 *Reports*, 6:30348.
- 701 Derringer, J., Livengood, J., and Briley, D. (2019). Gene-by-environment interactions in human  
702 individual differences.
- 703 Domingue, B., Trejo, S., Armstrong-Carter, E., and Tucker-Drob, E. (2020). Interactions between  
704 polygenic scores and environments: Methodological and conceptual challenges.
- 705 Domingue, B. W., Liu, H., Okbay, A., and Belsky, D. W. (2017). Genetic heterogeneity in depres-  
706 sive symptoms following the death of a spouse: Polygenic score analysis of the us health and  
707 retirement study. *American Journal of Psychiatry*, 174(10):963–970.
- 708 Duncan, L. E. and Keller, M. C. (2011). A critical review of the first 10 years of candidate gene-by-  
709 environment interaction research in psychiatry. *American Journal of Psychiatry*, 168(10):1041–  
710 1049.
- 711 Ellis, B. J., Boyce, W. T., Belsky, J., Bakermans-Kranenburg, M. J., and Van IJzendoorn, M. H.  
712 (2011). Differential susceptibility to the environment: An evolutionary–neurodevelopmental the-  
713 ory. *Development and psychopathology*, 23(1):7–28.
- 714 Frost, H. R., Shen, L., Saykin, A. J., Williams, S. M., Moore, J. H., and Initiative, A. D. N. (2016).  
715 Identifying significant gene-environment interactions using a combination of screening testing  
716 and hierarchical false discovery rate control. *Genetic epidemiology*, 40(7):544–557.
- 717 Guo, G., Liu, H., Wang, L., Shen, H., and Hu, W. (2015). The genome-wide influence on human bmi  
718 depends on physical activity, life course, and historical period. *Demography*, 52(5):1651–1670.
- 719 Guo, G., Tong, Y., and Cai, T. (2008). Gene by social context interactions for number of sexual  
720 partners among white male youths: Genetics-informed sociology. *American Journal of Sociology*,  
721 114(S1):S36–S66.
- 722 Halldorsdottir, T., Piechaczek, C., Soares de Matos, A. P., Czamara, D., Pehl, V., Wagenbuechler,  
723 P., Feldmann, L., Quickenstedt-Reinhardt, P., Allgaier, A.-K., Freisleder, F. J., et al. (2019).  
724 Polygenic risk: Predicting depression outcomes in clinical and epidemiological cohorts of youths.  
725 *American Journal of Psychiatry*, pages appi–ajp.
- 726 Herd, P., Freese, J., Sicinski, K., Domingue, B. W., Mullan Harris, K., Wei, C., and Hauser, R. M.  
727 (2019). Genes, gender inequality, and educational attainment. *American Sociological Review*,  
728 84(6):1069–1098.
- 729 Keller, M. C. (2014). Gene× environment interaction studies have not properly controlled for  
730 potential confounders: the problem and the (simple) solution. *Biological psychiatry*, 75(1):18–24.
- 731 Kong, A., Thorleifsson, G., Frigge, M. L., Vilhjalmsdottir, B. J., Young, A. I., Thorgeirsson, T. E.,  
732 Benonisdottir, S., Oddsson, A., Halldorsson, B. V., Masson, G., et al. (2018). The nature of  
733 nurture: Effects of parental genotypes. *Science*, 359(6374):424–428.



- 734 Lee, J. J., Wedow, R., Okbay, A., Kong, E., Maghzian, O., Zacher, M., Nguyen-Viet, T. A., Bowers,  
735 P., Sidorenko, J., Linnér, R. K., et al. (2018). Gene discovery and polygenic prediction from a  
736 genome-wide association study of educational attainment in 1.1 million individuals. *Nature*  
737 *genetics*, 50(8):1112–1121.
- 738 Liu, H. and Guo, G. (2015). Lifetime socioeconomic status, historical context, and genetic in-  
739 heritance in shaping body mass in middle and late adulthood. *American sociological review*,  
740 80(4):705–737.
- 741 Locke, A. E., Kahali, B., Berndt, S. I., Justice, A. E., Pers, T. H., Day, F. R., Powell, C., Vedantam,  
742 S., Buchkovich, M. L., Yang, J., et al. (2015). Genetic studies of body mass index yield new  
743 insights for obesity biology. *Nature*, 518(7538):197–206.
- 744 Lundberg, I., Johnson, R., and Stewart, B. M. (2020). Setting the target: Precise estimands and  
745 the gap between theory and empirics.
- 746 Marderstein, A. R., Davenport, E., Kulm, S., Van Hout, C. V., Elemento, O., and Clark, A. G.  
747 (2020). Leveraging phenotypic variability to identify genetic interactions in human phenotypes.  
748 *bioRxiv*.
- 749 Mullins, N., Power, R., Fisher, H., Hanscombe, K., Euesden, J., Iniesta, R., Levinson, D., Weiss-  
750 man, M., Potash, J. B., Shi, J., et al. (2016). Polygenic interactions with environmental adversity  
751 in the aetiology of major depressive disorder. *Psychological medicine*, 46(4):759–770.
- 752 Papageorge, N. W. and Thom, K. (2020). Genes, education, and labor market outcomes: evidence  
753 from the health and retirement study. *Journal of the European Economic Association*, 18(3):1351–  
754 1399.
- 755 Rimfeld, K., Krapohl, E., Trzaskowski, M., Coleman, J. R., Selzam, S., Dale, P. S., Esko, T.,  
756 Metspalu, A., and Plomin, R. (2018). Genetic influence on social outcomes during and after the  
757 soviet era in estonia. *Nature human behaviour*, 2(4):269–275.
- 758 Robinette, J. W., Boardman, J. D., and Crimmins, E. M. (2019). Differential vulnerability to  
759 neighbourhood disorder: a gene  $\times$  environment interaction study. *J Epidemiol Community Health*,  
760 73(5):388–392.
- 761 Rönnegård, L. and Valdar, W. (2011). Detecting major genetic loci controlling phenotypic vari-  
762 ability in experimental crosses. *Genetics*, 188(2):435–447.
- 763 Schmitz, L. L. and Conley, D. (2017). The effect of vietnam-era conscription and genetic potential  
764 for educational attainment on schooling outcomes. *Economics of education review*, 61:85–97.
- 765 Schmitz, L. L., Goodwin, J., Miao, J., Lu, Q., and Conley, D. (2021). The impact of late-career  
766 job loss and genetic risk on body mass index: Evidence from variance polygenic scores. *Scientific*  
767 *reports*, 11(1):1–15.
- 768 Seabrook, J. A. and Avison, W. R. (2010). Genotype–environment interaction and sociology:  
769 Contributions and complexities. *Social Science & Medicine*, 70(9):1277–1284.
- 770 Trejo, S., Belsky, D. W., Boardman, J. D., Freese, J., Harris, K. M., Herd, P., Sicinski, K., and  
771 Domingue, B. W. (2018). Schools as moderators of genetic associations with life course attain-  
772 ments: evidence from the wls and add heath. *Sociological science*, 5:513–540.

- 773 Trejo, S. and Domingue, B. W. (2019). Genetic nature or genetic nurture? quantifying bias in  
774 analyses using polygenic scores. *BioRxiv*, page 524850.
- 775 Vink, J. M. and Boomsma, D. I. (2011). Interplay between heritability of smoking and environ-  
776 mental conditions? a comparison of two birth cohorts. *BMC public health*, 11(1):1–7.
- 777 Wang, H., Zhang, F., Zeng, J., Wu, Y., Kemper, K. E., Xue, A., Zhang, M., Powell, J. E., Goddard,  
778 M. E., Wray, N. R., et al. (2019). Genotype-by-environment interactions inferred from genetic  
779 effects on phenotypic variability in the uk biobank. *Science advances*, 5(8):eaaw3538.
- 780 Ware, E., Schmitz, L., Gard, A., and Faul, J. (2018). Hrs polygenic scores—release 2: 2006–2012  
781 genetic data. *Ann Arbor: Survey Research Center, University of Michigan*.
- 782 Ware, E. B., Schmitz, L. L., Faul, J., Gard, A., Mitchell, C., Smith, J. A., Zhao, W., Weir, D., and  
783 Kardia, S. L. (2017). Heterogeneity in polygenic scores for common human traits. *BioRxiv*, page  
784 106062.
- 785 Wedow, R., Zacher, M., Huibregtse, B. M., Mullan Harris, K., Domingue, B. W., and Boardman,  
786 J. D. (2018). Education, smoking, and cohort change: Forwarding a multidimensional theory of  
787 the environmental moderation of genetic effects. *American Sociological Review*, 83(4):802–832.
- 788 Yang, J. e. a. (2012). FTO genotype is associated with phenotypic variability of body mass index.  
789 *Nature*, 490(7419):267–272.
- 790 Young, A. I., Wauthier, F. L., and Donnelly, P. (2018). Identifying loci affecting trait variability and  
791 detecting interactions in genome-wide association studies. *Nature genetics*, 50(11):1608–1614.
- 792 Zaidi, A. A. and Mathieson, I. (2020). Demographic history mediates the effect of stratification on  
793 polygenic scores. *Elife*, 9:e61548.
- 794 Zhang, F., Chen, W., Zhu, Z., Zhang, Q., Nabais, M. F., Qi, T., Deary, I. J., Wray, N. R., Visscher,  
795 P. M., McRae, A. F., et al. (2019). Osa: a tool for omic-data-based complex trait analysis.  
796 *Genome biology*, 20(1):1–13.

797 **Tables**

**Table 1: What type of moderation do recent gene by environment studies examine?**

The table presents a (non-exhaustive) list of recent gene by environment studies that use polygenic scores as the measure of genotype. With the exception of the study in gray, all study outcome moderation, or how the impact of an environmental trigger or buffer on an outcome varies by an individual’s genetic propensity towards that same outcome.

<b>Study</b>	<b>Outcome</b>	<b>Environment</b>	<b>PGS used to examine moderation</b>
Barcellos et al. (2018)	BMI	Education reform	BMI
Liu and Guo (2015)	BMI	Childhood and adult SES	BMI
Amin et al. (2017)	Educational attainment	Educational attainment	BMI
Trejo et al. (2018)	Educational attainment; job status	School SES; school stratification; environment-agnostic heterogeneity in school-level random slopes on PGS	Educational attainment
Schmitz and Conley (2017)	Educational attainment	Veteran status (instrumented with Vietnam draft lottery)	Educational attainment
Herd et al. (2019)	Educational attainment	Gender/cohort	Educational attainment
Robinette et al. (2019)	Type II diabetes	Neighborhood disorder	Type II diabetes
Domingue et al. (2017)	Depressive symptoms	Spousal loss	Subjective wellbeing; depression
Halldorsdottir et al. (2019)	Depression	Childhood abuse	Depression
Mullins et al. (2016)	Depression	Childhood stressful life events and trauma	Depression
Papageorge and Thom (2020)	Educational attainment	Family SES	Educational attainment

**Table 2: Genetic correlation results** The top panel shows the between-trait correlation in the standard levels weights. The middle panel shows the between-trait correlation in the squared  $Z$  vPGS weights, which are generally better powered. The bottom panel shows the *within-trait* correlation between: (1) the additive weights produced by the HLMM method, which are meant to identify mean effects purged of mean-variance correlations and (2) the non-HLMM vPGS.

<b>Between-trait correlations</b>			
Trait 1	Trait 2	$R_g$	SE
<i>Levels</i>			
BMI	Education	-0.3576	0.0304
BMI	Height	-0.1824	0.0262
BMI	NEB	0.1657	0.0466
Education	Height	0.2708	0.0280
Education	NEB	-0.2850	0.0530
Height	NEB	-0.1048	0.0407
<i>Squared Z score</i>			
BMI	Education	-0.1129	0.1138
BMI	Height	0.0928	0.1887
BMI	NEB	0.3697	0.2395
Education	Height	0.3161	0.3503
Education	NEB	-0.1571	0.3193
Height	NEB	0.0948	0.5428
<b>Within-trait correlations with HLMM additive</b>			
Trait	vPGS	$R_g$	SE
BMI	Levene's	-0.4867	0.09744
BMI	Sibling SD	Not identified	Not identified
BMI	Squared Z	-0.9004	0.04162
Education	Levene's	-1.0709	0.0390
Education	Sibling SD	0.2832	0.2552
Education	Squared Z	-0.8291	0.1019
Height	Levene's	-0.05683	0.283
Height	Sibling SD	Not identified	Not identified
Height	Squared Z	-0.8894	0.3587
NEB	Levene's	0.5181	0.0913
NEB	Sibling SD	0.0989	0.2876
NEB	Squared Z	-0.0951	0.1515

**Table 3: Impact of education reform on health outcomes**

	HLMM vPGS		Sibling vPGS	
	Body Size	Body Size (Threshold)	Body Size	Body Size (Threshold)
	(1)	(2)	(3)	(4)
vPGS	-0.073 (0.049)	-0.044* (0.018)	-0.018 (0.055)	-0.028 (0.021)
Educ16 (Instr.)	-0.225† (0.133)	-0.110* (0.050)	-0.231† (0.134)	-0.111* (0.051)
mPGS	0.198*** (0.049)	0.126*** (0.019)	0.172** (0.055)	0.123*** (0.021)
vPGS x Educ16 (Instr.)	-0.017 (0.055)	0.027 (0.021)	-0.024 (0.063)	0.019 (0.024)
mPGS x Educ16 (Instr.)	0.003 (0.056)	-0.087*** (0.021)	0.017 (0.063)	-0.086*** (0.024)
Constant	-1.764*** (0.263)	-0.348*** (0.100)	-1.748*** (0.264)	-0.347*** (0.100)
Observations	45,961	45,961	45,961	45,961
R <sup>2</sup>	0.061	0.026	0.054	0.023
Adjusted R <sup>2</sup>	0.061	0.025	0.053	0.022

Note:

\*p<0.1; \*\*p<0.05; \*\*\*p<0.01  
 †p < 0.1; \*p < 0.05; \*\* p < 0.01; \*\*\* p < 0.001

**Table 4: Impact of education reform on health outcomes**

	Body Size (1)	Body Size (Threshold) (2)
mPGS	0.159*** (0.045)	0.106*** (0.017)
Educ16 (Instr.)	-0.234† (0.134)	-0.112* (0.051)
mPGS x Educ16 (Instr.)	0.005 (0.051)	-0.074*** (0.019)
Constant	-1.728*** (0.264)	-0.342*** (0.100)
Observations	45,961	45,961
R <sup>2</sup>	0.053	0.022
Adjusted R <sup>2</sup>	0.052	0.021

*Note:* † $p < 0.1$ ; \* $p < 0.05$ ; \*\* $p < 0.01$ ; \*\*\* $p < 0.001$



**Table 5: Genetic moderation of reform’s impact on educational attainment**

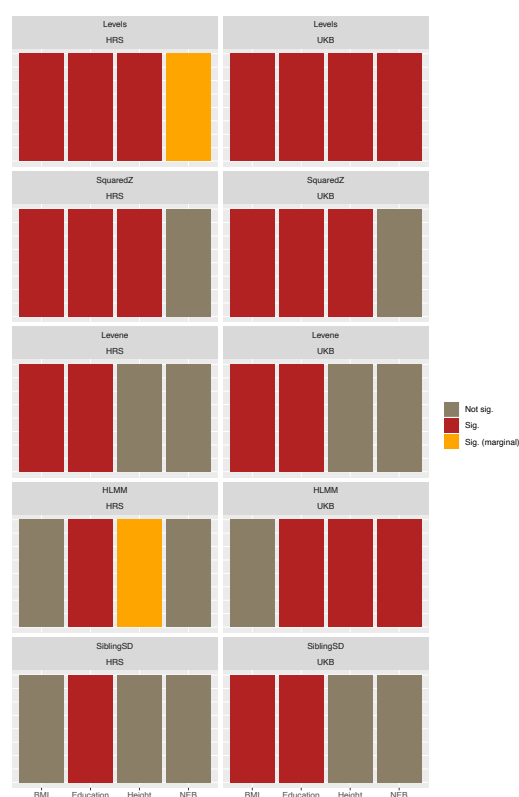
Controls?	<i>HLMM vPGS, Education Outcomes</i>							
	Left School 16 or later		Certification		O-levels or CSE		A-levels	
	N	Y	N	Y	N	Y	N	Y
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
vPGS	-0.059*** (0.003)	-0.040*** (0.005)	-0.063*** (0.003)	-0.044*** (0.005)	-0.008† (0.004)	-0.007 (0.007)	-0.055*** (0.004)	-0.037*** (0.006)
Post Reform	0.153*** (0.008)	0.143*** (0.008)	0.053*** (0.008)	0.053*** (0.008)	0.058*** (0.011)	0.059*** (0.011)	-0.005 (0.009)	-0.005 (0.009)
mPGS		0.022*** (0.005)		0.021*** (0.005)		-0.0002 (0.007)		0.021*** (0.006)
vPGS x Post Reform	0.048*** (0.004)	0.033*** (0.007)	0.036*** (0.005)	0.031*** (0.007)	0.035*** (0.006)	0.034*** (0.010)	0.001 (0.005)	-0.003 (0.009)
mPGS x Post Reform		-0.017* (0.007)		-0.005 (0.007)		-0.001 (0.010)		-0.003 (0.008)
Constant	0.779*** (0.004)	0.478** (0.147)	0.809*** (0.005)	0.004 (0.156)	0.557*** (0.006)	0.247 (0.206)	0.253*** (0.005)	-0.242 (0.178)
Observations	25,690	25,690	26,012	26,012	26,012	26,012	26,012	26,012
R <sup>2</sup>	0.098	0.117	0.055	0.072	0.019	0.027	0.016	0.025
Adjusted R <sup>2</sup>	0.098	0.115	0.055	0.070	0.018	0.025	0.016	0.023

*Note:*

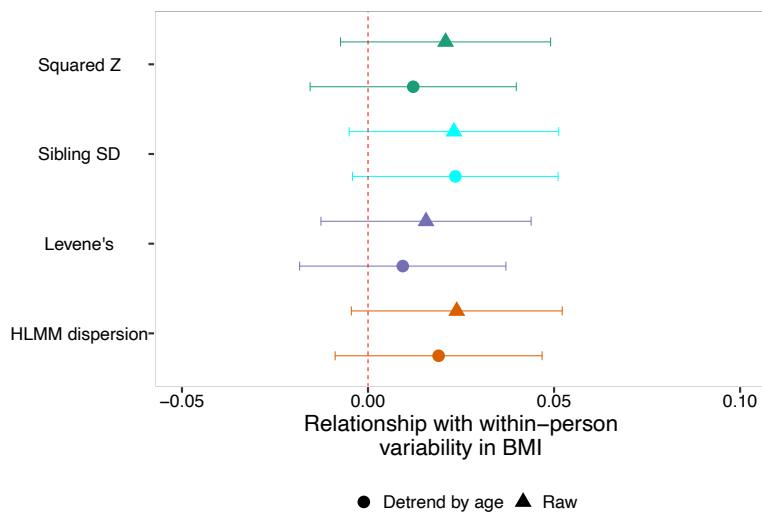
† $p < 0.1$ ; \* $p < 0.05$ ; \*\* $p < 0.01$ ; \*\*\* $p < 0.001$

## 798 Figures

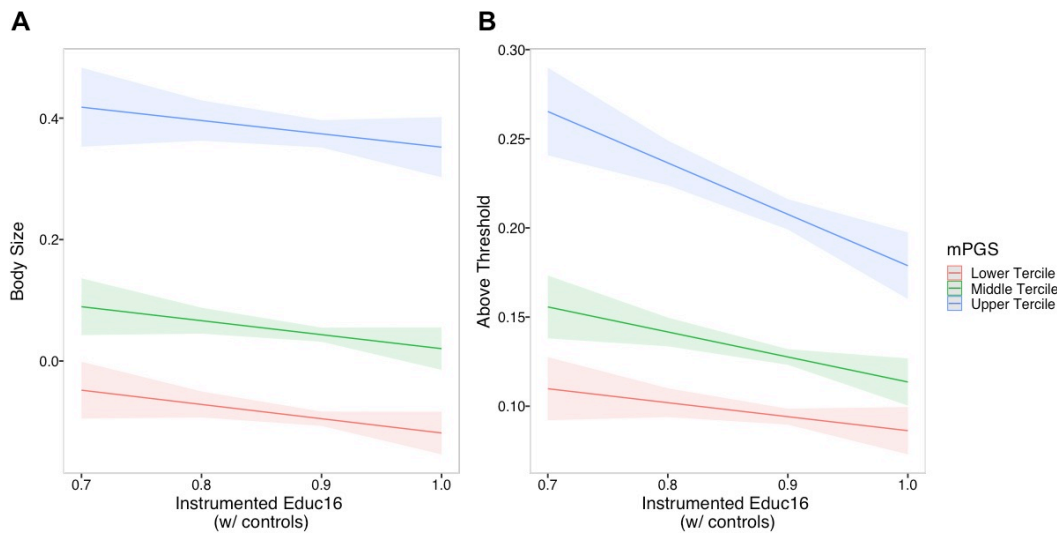
**Fig. 1: Significance of mPGS and plasticity scores in predicting levels of a trait** The figure shows results from the regression specified in Equation 1 in the HRS (left panel) and UKB test set (right panel). The top panel shows that, as expected, the levels PGS predict levels of an outcome (though the relationship with fertility in HRS is weaker than for height, BMI, and education). Moving downwards, across both samples, the Squared Z plasticity score performs the least well in that the plasticity score significantly predicts levels of an outcome for all outcomes except for number of children ever born. The Sibling SD score and HLMM perform best in the HRS test set



**Fig. 2: Relationship between vPGS and within-person variability in BMI: matched N sample** The figure, focusing on the sample where we restrict the estimation sample size for all scores to be equivalent to the estimation sample size for the sibling SD vPGS, shows two versions of the within-person variability analysis: a version with raw BMI over time and a version where BMI is detrended according to age patterns. The figure shows a general correlation between a respondent having a higher vPGS score and them having more variability in their BMI.



**Fig. 3: Interaction between mPGS and Instrumented Educ16 on Body Size outcomes**



**Fig. 4: Interaction between HLMM polygenic score for plasticity and Post Reform on Educational Outcomes**

