Journal of Personality 85:1, February 2017
© 2015 Wiley Periodicals, Inc.
DOI: 10.1111/jopy.12227

Modeling Gene-Environment Interactions With Quasi-Natural Experiments

Lauren Schmitz[1] **and Dalton Conley**[2]
[1]University of Michigan
[2]New York University

Abstract

This overview develops new empirical models that can effectively document Gene × Environment (G×E) interactions in observational data. Current G×E studies are often unable to support causal inference because they use endogenous measures of the environment or fail to adequately address the nonrandom distribution of genes across environments, confounding estimates. Comprehensive measures of genetic variation are incorporated into quasi-natural experimental designs to exploit exogenous environmental shocks or isolate variation in environmental exposure to avoid potential confounders. In addition, we offer insights from population genetics that improve upon extant approaches to address problems from population stratification. Together, these tools offer a powerful way forward for G×E research on the origin and development of social inequality across the life course.

The integration of genetic data into large-scale multidisciplinary surveys has transformed the scope of social science research and promises to revolutionize our understanding of the interplay between social and biological forces. Research on Gene × Environment (G×E) interactions—broadly defined as any situation where individual response to environmental risk differs by genotype—has shown gene expression is amplified, or reduced, in the presence of a particular environment; similarly, the effects of the environment are influenced by the presence or absence of specific genetic susceptibilities. In other words, genes operate *through* the environment (Rutter, 2006). The feedback between G and E provides a backdrop for understanding how aspects of the social environment contribute to social inequality and alter the development of human potential across the generational arc.

However, significant methodological hurdles remain in research that uses observational data to explore G×E effects outside of the "lab." Mainly, all but a handful of G×E studies have deployed endogenous measures of the environment, and even for those few exceptions where environment can be said to have been plausibly exogenous, candidate gene approaches have been used that did not control for population stratification or the nonrandom distribution of genes across subpopulations. In this article, we define an environmental measure as endogenous if it is correlated with an outside confounder that is not controlled for in the statistical model, whereas exogenous measures are considered external to the model. For example, prior studies have typically relied on endogenous measures of the environment like educational attainment, which may be correlated with underlying genotype. Or the few studies that used exogenous sources of environmental variation have estimated equations on single candidate genes without adequate controls for population stratification—leading to the possibility that these alleles are acting as proxies for unmeasured ethnic or ancestral influences.

The problem of isolating causal inference in gene-environment research with survey data requires cross-disciplinary work between human geneticists and applied econometricians. Currently, both "camps" are at their limit: Social scientists lack the necessary background in bioinformatics and statistical genetics to incorporate genotype measures into their research, and geneticists are not trained to consider empirical issues like sampling, survey design, measurement of social outcomes, and techniques for isolating causality that form the backbone of microeconometric work (Conley, 2009). New methods that provide adequate identification of exogenous G, E, and thus G×E effects are needed to provide a comprehensive way forward in understanding how the social determinants of health and

Correspondence concerning this article should be addressed to Dalton Conley, 295 Lafayette St., 4th Floor, New York, NY 10012. Email: daltonclarkconley@gmail.com.

behavior interact with the biological determinants (Conley, 2009; Fletcher & Conley, 2013).

To sort this out, we propose a way to properly identify models where both G and E are independent of each other in order to test for G×E interactions. Specifically, we will outline how quasi-natural experimental designs can be used to study whether significant life course events or stressors are magnified or moderated by genotype. Here, quasi-natural experiments refer to the use of instrumental variables (IV), differences-in-differences (DID), or regression discontinuity (RD) designs that either exploit exogenous environmental shocks or can isolate variation in environmental exposure to avoid any potential confounders. These econometric methods are the gold standard for approximating experiments and capturing causal effects with observational data in the social sciences. We then offer insights from population genetics that improve upon extant approaches to address the nonrandom distribution of genotypes across environments while maintaining adequate statistical power. Together, these tools offer a powerful way forward for G×E research.

Sound empirical evidence also has the potential to inform policy recommendations that seek to alter the foundations of social inequality. For example, if findings show phenotypic differences, or observed physical or behavioral differences, in educational attainment are the result of environmental and genetic attributes, then changes to the environment will also affect the distribution of outcomes across generations. On the other hand, if the intergenerational association in education is purely due to genetic characteristics, then even totally equalizing education in a given generation will have little effect on the next generation (Conley et al., 2015). Thus, if we know the extent to which an outcome is related to measurable genotype, we can more precisely target interventions that alleviate the emergence and development of social inequality over the life course.

We begin the article with an overview of how endogenous measures of either G or E can arise and lead to inconsistent estimates of G×E effects in econometric models. Next, we discuss how the latest advances in population genetics can be used to improve measures of genetic variation in G×E studies. These include the use of polygenic scores from genome-wide association studies to measure genotype, as well as approaches to ensure that genotype is not inadvertently proxying environmental differences, including control for principle components, modeling the error structure, holding parental genotype constant, and sibling fixed effects models. Next, we outline how genotype can be incorporated into quasi-natural experimental frameworks, including a discussion of the technical and methodological issues that need to be addressed and how researchers should go about interpreting results from these experiments. Finally, we elaborate on the feasibility, limitations, and best practices for application of this approach to social science research. Throughout, the focus will be on adapting the basic econometric specifications needed to estimate the parameters of interest in applied G×E work.

THE ENDOGENEITY PROBLEM IN APPLIED G×E RESEARCH

Existing efforts to find associations between genetic variation and social behavior in large, multidisciplinary surveys are often unable to support causal inferences because they used endogenous measures of the environment, genotype, or both. Here, we define G or E as endogenous if either term is correlated with the error or disturbance in the econometric model. To illustrate this, consider the following single-equation linear regression:

$$Y_i = \beta_0 + \beta_1 G_i + \beta_2 E_i + \beta_3 G_i \times E_i + u_i, \quad (1)$$

where Y is the biological or behavioral outcome (i.e., phenotype) of interest, G is a measure of genetic variation between individuals, E is the environmental factor, $G \times E$ is their interaction, and u is the unobservable random disturbance or error. For ordinary least squares (OLS) to consistently estimate the betas in this equation, the error term must be uncorrelated with each of the regressors, or $Cov(\boldsymbol{X},\ u) = 0$. When estimating G×E interactions, this is most likely to result from (a) the inability to control for all factors that influence a given phenotype that are also highly correlated with G or E (i.e., omitted variable bias); (b) measurement error, or imperfect measures of either G or E; and (c) simultaneity, or the case where either G or E is determined simultaneously with (or is a function of) the phenotype of interest.

Endogeneity in the environmental factor typically occurs when researchers use perceptual measures of the environment that could be acting as proxies for undetected gene-environment correlations (rGE). Gene-environment correlations operate in a different fashion than G×E interactions and broadly refer to any situation where genotype or genetically influenced behavior affects an individual's selection into environments or experiences. An example would be a verbally precocious child evincing more conversation with her parent than her sibling who is less genotypically endowed to be highly verbal. Later outcomes such as higher reading scores for the more verbal child would not transpire without the environmental input of a willing interlocutor during her development; however, the E in this case is a niche formed as a result of her G. In this example, the E is part of the mediating pathway of G and also moderates it. If rGE is present, it can be of a type that creates spurious effects—that is, omitted variable bias or the case where either or both G or E is acting as a proxy for another G or E. This is the most troubling rGE since it will necessarily lead to false inferences. However, rGE could also introduce a type of simultaneity bias in our G×E model if E is endogenously evoked by G.

Likewise, confounding in G×E models may also occur as the result of undetected G×E, G×G, or E×E phenomena. For example, research has uncovered a G×E interaction between common variants of the glucocorticoid receptor (GR) gene that regulates the hypothalamic-pituitary-adrenal (HPA) axis and self-reported measures of childhood trauma or abuse on the development of depressive symptoms in old age

(Bet et al., 2009). In this example, self-reported measures of childhood trauma could be intertwined with a host of unobserved genetic or environmental influences that are associated with depression in adulthood—on both the environmental and genotypic sides. For instance, unobserved genetic influences on cognition or personality may be highly correlated with—or regulate pathways between—genotype and adult depression. Likewise, a traumatic event in childhood may give rise to self-destructive behaviors in adolescence or other psychiatric disorders that trigger depressive symptoms in old age. Perhaps most troubling is the possibility that childhood trauma is likely to have been caused directly or indirectly by parents, who not only structure the developmental environment for a child but also pass on myriad unmeasured genetic variants. In other words, it may be that childhood trauma is acting as a proxy for parental genotype that itself is passed on to offspring and may interact with measured genotype—creating a latent G×G that is proxied by the measured G×E.

Even if all omitted genetic and environmental variables or pathways between trauma and depression could be accounted for, it would be nearly impossible to find accurate, unbiased measures for all of them. In addition, self-reports of childhood abuse may be determined simultaneously with reports of depression—that is, an individual may misreport the nature or extent of the trauma she experienced if she currently suffers from depression. Due to the endogeneity of the environmental factor, the coefficients in the regression will be biased, and the G×E interaction between childhood abuse and the GR gene cannot confidently be said to have a causal effect on adult depression (though it is still possible that the measured G×E is indeed causal).

On the other hand, endogenous measures of G in G×E studies usually arise in one of two ways. First, studies may suffer from the problem of nonrandom *genetic* assignment. That is, while it is possible that environmental measures are acting as proxies for unobserved genotype, thus leading to biased estimates, it is *also* possible that apparent genetic effects are false positives, the result of the confounding of genotypes and environment through population stratification—a concept popularized by Hamer and Sirota (2000), who used the example of a "chopstick gene" appearing because of data that mix Asians and Caucasians. Population stratification or admixture occurs when a study sample mixes two or more ancestral or ethnic subgroups that have different allele frequencies and, coincidentally, different levels of a particular phenotype. In the chopstick example, the significant association between the "successful-use-of-selected-hand-instruments" (SUSHI) gene and chopstick use is spurious—the result of different allele frequencies in Asians and Caucasians who differ in chopstick use for cultural rather than biological reasons. Therefore, studies must control for the nonrandom distribution of genes across populations to account for differences in genetic structures within populations that could bias estimates. Indeed, even in ethnically homogeneous samples, it turns out that friends and spouses tend to be more genotypically similar than randomly matched individuals (Christakis & Fowler, 2014; Domingue, Fletcher, Conley, & Boardman, 2014) and that even environmental measures such as urbanity are correlated with population structure (Conley et al., 2014).

Exacerbating this problem, the majority of studies have used candidate genes to test for G×E effects. In a candidate gene study, researchers specify ex ante hypotheses about links between a small set of single nucleotide polymorphisms (SNPs), or a single nucleotide location in the DNA that varies between individuals, and a specific phenotype. For example, common polymorphisms of the APOE gene, which codes apolipoprotein, have been found to be a strong predictor of Alzheimer's disease (St George-Hyslop, 2000; Strittmatter et al., 1993). While this approach can be fruitful if there is extensive knowledge about the biological pathway between a given gene and a particular phenotype, it cannot capture the dynamic nature of more complex behavioral traits that are born out of an entire network of interconnected genetic and environmental attributes. Thus, even if proper candidates are found among known pathways, essential genes could still be overlooked if there is incomplete knowledge about other biological systems that are involved in the process (Vink & Boomsma, 2002). More importantly for the present analysis—absent sibling fixed effects or some other form of within-family control such as a transmission disequilibrium test (essentially controlling for parental genotype)—single locus analysis does not offer any way to control for the nonrandom distribution of genotypes across environments.

To overcome these estimation issues, new methods that provide adequate identification of exogenous G, E, and thus G×E effects are needed to provide a comprehensive way forward in understanding how the social factors influencing health and behavior interact with the biological factors that may also influence phenotypes of interest. Randomized control trials or large studies involving human subjects may be costly and limited in their ability to investigate a variety of G×E phenomena over the life course. To utilize the wealth of genotype data that is now available in social surveys, we propose workable models that exploit exogenous sources of environmental variation, comprehensive measures of genetic risk, and controls for population stratification to properly identify G×E effects.

IMPROVING HOW WE MEASURE "G" IN G×E STUDIES

Before addressing how putatively exogenous measures of environmental variables can be used to guard against the possibility that "E" is proxying for unmeasured "G," we discuss how methods in the population genetics literature can be adopted to ensure that G is not correlated with unmeasured G or E. This includes a discussion of why polygenic scores from genome-wide association studies are particularly ripe measures of genotypic variation, as well as approaches to deal with confounding from population stratification, including control for principle components, modeling the error structure, holding parental genotype constant for single locus analysis, and sibling fixed effects models.

Beyond Candidate Genes: The Use of Polygenic Scores

Belsky and Israel (2014) cite two primary reasons why the use of single genetic variants to capture G×E effects is often suboptimal in social science and behavioral research. First, complex health outcomes or behaviors of interest to social scientists are usually highly *polygenic*, or reflect the influence or aggregate effect of many different genes (Visscher, Hill, & Wray, 2008). Individuals fall somewhere on a continuum of genetic risk that reflects small contributions from many genetic loci—even clinically dichotomous outcomes may reflect a shift along a phenotypic continuum known as decanalization (Gibson, 2012). Second, individual genetic loci influencing the etiology of complex phenotypes have *low penetrance*; no single gene produces a symptom or trait at a detectable level, making it difficult to distinguish between environmental and genetic factors (Gibson, 2012). In both cases, the use of single genetic variants in a G×E model would thus result in a form of omitted variable bias, whereby crucial G×G or G×E interactions are obscured and left sitting in the error term, confounding estimates.

Recently, the advent of dense SNP chips has made it possible and relatively inexpensive to measure millions of SNPs in a single study. As a result, researchers are now moving toward using genome-wide association studies (GWASs) to measure genetic risk. A GWAS is a hypothesis-free exercise that looks for associations between a phenotype and millions of singular nucleotide polymorphisms. In the discovery phase, a GWAS will pool large consortia of genetic data using meta-analysis and run regressions testing each SNP at the genome-wide significance level of 5×10^{-8}. In the replication phase, significant associations found in the discovery phase are tested in independent samples.

Using results from a GWAS, researchers can compile a polygenic score for a phenotype that aggregates thousands of SNPs across the genome and weights them by the strength of their association. In essence, a polygenic score is a weighted average or composite score that takes into account information across an individual's entire genome to measure his genetic predisposition to or risk for a particular outcome. Or, a polygenic score (PS) for individual i is a weighted average across the number of SNPs (n) of the number of reference alleles x (0, 1, or 2) at that SNP multiplied by the score for that SNP (β):

$$PS_i = \sum_{j=1}^{n} \left(\beta_j x_{ij}\right). \tag{2}$$

Polygenic scores have several attractive features. First, unlike candidate genes, they are "hypothesis-free" measures—that is, ex ante knowledge about the biological processes involved is not needed to estimate a score for a particular phenotype. Rather, a polygenic score casts a wide net across an individual's entire genome to yield a single quantitative measure of genetic risk, allowing researchers to explore how genes operate within environments where the biological mechanisms are not yet fully understood (Belsky & Israel, 2014). One merely needs to calculate the score and then interact the single variable with an exogenous source of environmental variation to investigate whether G×E effects are at play. Therefore, the strength of the hypothesis-free approach is that it propels knowledge about how genetic mechanisms work by stimulating research outside of the "lab" that can easily test and pinpoint important sources of variation in the social environment.

Second, achieving the statistical power needed to model a candidate gene × environment (cG×E) study for biologically distal, social phenotypes is not possible in existing social surveys that contain the level of detailed information about respondents that motivates G×E inquiry in the social sciences (Belsky, Moffitt, & Caspi, 2013). For example, in order to detect an effect of ex ante reasonable size (i.e., an effect that explains .02% of the variation, which is among the largest of extant effects for single alleles on behavioral outcomes) between a candidate gene and a given phenotype, we would need a study to contain a sample size that provides approximately 93,000 degrees of freedom if we wanted to be sure that it was significant at the conventional genome-wide suggestive significance level of $p < 5 \times 10^{-5}$ (Conley 2015); while we are theoretically only testing "one" hypothesis for the main effect of a candidate gene and one for the interaction effect, experience has shown that alleles found to be predictive in single locus analysis typically fail to replicate if only significant at conventional p-value levels. As a result, since only the most significant findings are usually published, the cG×E literature contains an inflated number of false positives (Duncan & Keller, 2011).

A GWAS, on the other hand, deploys an atheoretical search for alleles that are significantly predictive of an outcome using the raw statistical power from huge consortia such as the Social Science Genetics Association Consortium (SSGAC) to generate the polygenic score. These scores can then be recalculated for participants in a nationally representative panel study with its rich measures to test for G×E effects. For example, polygenic scores from consortia data can be recalculated for a range of phenotypes, including educational attainment (SSGAC; Rietveld et al., 2013), body mass index (GIANT consortium; Yang et al., 2012), cardiovascular disease (CHARGE consortium; Levy et al., 2009), smoking behavior (TAG consortium; Furberg et al., 2010), and psychiatric disorders (PGC consortium; Lee et al., 2013). Finally, unlike most candidate gene studies, GWASs also deploy the wide range of markers to control for confounding from population stratification using principal components—a technique we will discuss in more depth in the next section.

In addition to polygenic risk score analysis, the GWAS "revolution" has spawned a cottage industry of new heritability analysis that deploys a genetic similarity matrix among unrelated individuals in an effort to overcome some of the assumptions (specifically, no rGE) in classical twin-based heritability analysis. This genome-wide-relatedness-matrix estimation maximum likelihood (GREML or GCTA) procedure itself has recently come under scrutiny for perhaps not eliminating rGE (e.g., Conley et al., 2014). This issue aside, the GCTA approach may also prove fruitful for integration with the deployment of exogenous environmental variation, as we propose here. The main problem with stratifying GCTA analysis by potentially genetically correlated "environmental" factors is that if different heritability estimates are obtained for two groups (e.g., from families with varying socioeconomic status) one cannot know whether the observed difference in h^2 is due to differences in the variance of G or E. With exogenous environmental measures that are by definition orthogonal to G, this problem is obviated and one can obtain GCTA estimates that are stratified and reflect a true interaction with E. However, whether or not the G portion of that G×E estimate is itself not biased due to the possible confounding by population structure (and thus environmental variation) is a huge question hanging over such an approach (Conley et al., 2014).

While our proposed strategy of using well-established main effects (i.e., polygenic risk scores) from large multistudy consortia as the grist for our G×E analysis solves many power and replicability issues, it does suffer from one main limitation: Since the genetic main effects arise from meta-analyses of studies that typically span a wide range of (Western) countries and cohorts, it may be the case that the main effects that arise from the extant approaches to pooled analysis are, by design, those that are most robust to local context, thus making them unlikely to show significant interaction effects with exogenous environmental variation. While we recognize this issue, we do not think it will be prohibitive for at least two reasons: First, though main effects are culled from a wide variety of data sets, in such consortia studies the individual parameter estimates for each cohort have typically demonstrated a wide degree of variation, thereby showing the potential importance of environmental moderators. For example, one of the strongest main effects to arise from such consortia studies is the relationship between fat mass and obesity-associated (FTO) genotype and risk for obesity (Cha et al., 2008; Chang et al., 2008; Dina et al., 2007; Frayling et al., 2007; Hunt et al., 2008; Scuteri et al., 2007). However, this very same gene has also been shown to significantly interact with (endogenous) environment (Andreasen et al., 2008; Haworth et al., 2008). Indeed, a recent consortium-based meta-analysis of variability in BMI showed FTO to be genome-wide significant in predicting variation (Yang et al., 2012). Second, the application of consortium data to genome-wide association studies for variability (vGWAS) has become increasingly common. These studies identify loci, genes, and pathways that may be associated with variation in a given phenotype as a way to latently identify potential G×E or G×G interactions without specifying the nature of such an interaction effect. These vGWAS consortia results (already publicly available for height and BMI, with others soon to follow) can be used to enhance or complement our proposed approach by guiding the search for particularly fruitful G×E interplay.

Likewise, another limitation is that although polygenic scores may aggregate and stabilize genetic signals, not all SNPs respond uniformly to the environment, and aggregation may obscure the exact nature of biological pathways. For example, two SNPs may obtain genome-wide significance in a GWAS of psychiatric disorders, but the biological mechanisms of the first SNP may be suppressed in environmental advantage while the second SNP's biological mechanisms may be magnified. Moreover, results from GWAS, as compared to heritability estimates, explain only a small portion of phenotypic variability. For example, the linear polygenic score from all measured SNPs in the Rietveld et al. (2013). GWAS on educational attainment explained approximately 2–3% of the variation in years of schooling. Two to three percent is a relatively small contribution to our understanding of educational outcomes, especially when compared to published meta-analyses that found genetic factors account for up to 40% of the variation (Branigan, McCallum, Kenneth, & Freese, 2013). There are several important explanations for this so-called missing heritability (De los Campos, Vazquez, Fernando, Klimentidis, & Sorensen, 2013), including estimation error in the coefficients from the GWAS, sample size, the role of rare genetic variants, and G×E interactions. As a result, if researchers are faced with a low sample size among treated populations, using power analysis to evaluate whether G×E coefficients are underpowered may be advisable. However, despite these current challenges to molecular genetics research, for the reasons highlighted above, we argue the use of polygenic scores is an important addition to the detection and estimation of genotype × environment relationships.

Addressing Confounding From Population Stratification

With data that contain only a few genetic markers, it is quite difficult to address the problem of population stratification. In studies that have parental genotype for a large proportion of the sample—such as in the Framingham Heart Study (FHS) third generation (Splansky et al., 2007) and Minnesota Twin Family Study (MTFS; Iacono & McGue, 2002)—one solution is adding controls for parental genotype. Essentially, this breaks any population structure through what amounts to a transmission disequilibrium test—that is, variation in offspring genotype is the random result of recombination and segregation of alleles. Meanwhile, if sibling data are available (e.g., FHS third generation, MTFS dizygotic twins and Add Health dizygotic twins), the ideal approach is to conduct an analysis that compares full siblings who are discordant on genotype, where the assignment of genetic differences was also randomized at conception (Harris et al., 2009). Here, sibling fixed effects can be used to estimate the main genetic effects, which also eliminates any possibility of population stratification, even absent parental genotypic information.

However, many large, multidisciplinary studies that have genotyped their participants—such as the Health and Retirement Study (HRS; 2015)—do not have family data (Sonnega et al., 2014). In addition, finding exogenous environmental influences that cut within families (i.e., differ between siblings) with which to interact randomized genotype is an order of magnitude more difficult than finding exogenous environmental variation across a sample of unrelated individuals. If family data are not available, but genome-wide data are available, another approach involves estimating mixed linear models (Liang & Zeger, 1986). Conceptually, such models involve two steps: (a) the genome-wide data are used to estimate the degree of genetic similarity between the individuals in the sample (using GCTA or similar software to estimate the matrix of pairwise genetic similarity), and (b) unlike in a standard regression where the covariance of the error term between any two individuals is assumed to be zero, the covariance is fitted as a linear, increasing function of the individuals' genetic similarity (Kang et al., 2010). In other words, to the extent that two individuals are more recently descended from a common ancestor (as very accurately measured by overall genetic similarity)—and thus are more likely to be similar on unobserved environmental factors—these individuals are not treated as two independent observations on the relationship between the phenotype and the score.

A third, complementary approach to the second one mentioned above that can also easily be applied to studies with hundreds of thousands of genetic markers involves using principal components to control for confounding from population stratification. The principal components measure the uncorrelated variation or dimensions in the data, accounting for ethnic or racial differences in genetic structures within populations that could bias estimates. In essence, if we have data on thousands of SNPs for over 20,000 respondents in a sample, principal component analysis will identify the underlying dimensions in the genotype data where there is a high degree of variance between individuals and will decompose these dimensions into linearly uncorrelated variables. In applied G×E models, this approach provides a simple and efficient solution to the population stratification problem. Using readily available programs like EIGENSTRAT, the first 10 principal components can be calculated and included as controls in a linear regression—a dimensionality that has generally proven adequate in the literature (Price et al., 2006). In particular, when using results from a GWAS to construct polygenic scores for independent samples, controlling for the first 10 principal components accounts for any systematic differences in ancestry that can cause spurious correlations while also maximizing the power that is needed to detect true associations.

INCORPORATING GENOTYPE INTO QUASI-NATURAL EXPERIMENTS

With the above techniques in mind, the following sections modify quasi-natural experimental designs to accommodate heterogeneous effects by genotype. These designs are used in the social sciences to overcome omitted variable and selection problems in estimates of causal relationships. An in-depth review of the theory behind instrumental variables (IV), differences-in-differences (DID), and regression discontinuity (RD) designs can be found in several sources (Angrist & Pischke, 2008; Imbens & Lemieux, 2008; Meyer, 1995). Here, we provide a basic sketch of each econometric framework and how it can be adapted to estimate the parameters of interest in G×E research. Throughout, we emphasize the use of polygenic scores to measure genotypic differences between individuals, and principal components to control for confounding from population stratification. Rather than testing specific polygenic scores and outcomes for each environmental shock, when possible we recommend G×E effects be estimated with more than one quasi-natural experimental design (i.e., IV, DID, or RD).

Modeling G×E Interactions With Instrumental Variables Estimation

IV estimation solves the problem of missing or unknown control variables in the same way a randomized control trial rules out the need for extensive controls in a regression. In a typical IV setup, an instrument is chosen that is (a) highly correlated with the causal variable of interest, or in this case the endogenous environmental factor, but (b) uncorrelated with any other determinants of the outcome of interest. The second condition is known as the "exclusion restriction" since the instrument is excluded from the causal model of interest. Consider the following structural equation:

$$Y_i = \alpha_1 E_i + \alpha_2 G_i \times E_i + X_i' \beta + \epsilon_i, \tag{3}$$

where E is the endogenous environmental factor, G is the polygenic score of interest, $G \times E$ is their interaction, Y is the outcome of interest, X is a vector of exogenous controls, and ϵ is the disturbance term. The vector X includes the main effect of G and the first 10 principal components to account for population stratification in the genotype data. Imagine a suitable instrument Z that meets the above criteria is available for E. Then heterogeneous "treatment" effects by genotype can be tested in an IV framework that instruments E and its interaction with the genetic score G with Z. In a two-stage least squares (2SLS) IV framework, E would be instrumented with Z in the first stage as follows:

$$E_i = \pi_1 Z_i + \pi_2 G_i \times Z_i + X_i' \pi_3 + \eta_i \tag{4}$$

and

$$G_i \times E_i = \gamma_1 Z_i + \gamma_2 G_i \times Z_i + X_i' \gamma_3 + \rho_i, \tag{5}$$

where the model is exactly identified. The first-stage equations can then be substituted into the structural equation to derive the reduced form:

$$\begin{aligned} Y_i = {} & \alpha_1\left[\pi_1 Z_i + \pi_2 G_i \times Z_i + X_i'\pi_3 + \eta_i\right] \\ & + \alpha_2\left[\gamma_1 Z_i + \gamma_2 G_i \times Z_i + X_i'\gamma_3 + \rho_i\right] + \ X_i'\beta + \epsilon_i \\ = {} & Z_i[\alpha_1\pi_1 + \alpha_2\gamma_1] + G_i \times Z_i[\alpha_1\pi_2 + \alpha_2\gamma_2] \\ & + X_i'[\beta + \alpha_1\pi_3 + \alpha_2\gamma_3] + [\alpha_1\eta_i + \alpha_2\rho_i + \epsilon_i]. \end{aligned} \tag{6}$$

$$Y_i = \delta_1 Z_i + \delta_2 G_i \times Z_i + X_i'\delta_3 + \xi_i. \tag{7}$$

Thus, conditional on covariates, 2SLS retains the variation in *E* that is generated by *Z*, or the quasi-experimental variation (Angrist & Pischke, 2008). If the coefficient of interest on the G×E interaction term is significant, then the outcome varies by genotype, or the impact of the environmental exposure on the outcome is influenced by an individual's genotype. Because the instrument is exogenous and only affects the outcome through the first-stage channel, we avoid any potential confounders and can interpret the G×E interaction term as having a causal effect on our outcome of interest.

To illustrate how an IV framework can be used to identify G×E effects, consider the case of military service. Military service is a critical turning point in the lives of young recruits that can have significant consequences on earnings, health, and family dynamics. The range of stressful environmental exposures that could arise as a result of combat coupled with the challenges of post-service life make it a particularly ripe candidate for G×E interplay. However, since selection into the military is far from random, and likely to be correlated with factors like socioeconomic background or prior health status, it would be impossible to sort out the effects of military service from the effects of other gene-environment or gene-gene interactions in a model that uses self-reported veteran status to estimate G×E effects.

To circumvent any bias due to selectivity issues, prior studies have used the Vietnam-era draft lotteries as an instrumental variable for veteran status. Between December 1969 and February 1972, the U.S. Selective Service held four Vietnam draft lotteries. Each of these draft lotteries randomly assigned men in eligible birth cohorts order of induction numbers through a hand drawing of birthdates. The random assignment mechanism of the draft lotteries has been used to identify the effects of wartime military service on a host of outcomes, including economic (Angrist, 1990; Angrist & Chen, 2011; Angrist, Chen, & Song, 2011), family (Conley & Heerwig, 2011; Heerwig & Conley, 2013), and health outcomes (Angrist, Chen, & Frandsen, 2010; Conley & Heerwig, 2012; Dobkin & Shabani, 2009). Since draft eligibility is (a) highly correlated with actual veteran status and (b) only affects outcomes through the first-stage channel, or through its correlation with veteran status, it is considered a valid instrument for military service. In addition, because draft status is orthogonal to standard sociodemographic variables at the time of the lottery, any variation in socioeconomic status after military service is related to the instrument or is a result of the treatment.

For example, to identify whether the effects of military service on depression vary by genotype, instrumented veteran status could be interacted with a polygenic score for psychiatric disorders from the Psychiatric Genomics Consortium (PGC; Lee et al., 2013). Due to the shared genetic etiology for psychiatric disorders, this particular G×E interaction could be used to investigate a number of related pathologies, including schizophrenia, bipolar disorder, autism spectrum disorders, and attention deficit/hyperactivity disorder. If the polygenic score is standardized with a mean of zero and a standard deviation of one, the coefficient on the G×E term (δ_2) can be interpreted as representing the marginal difference in rates of depression between veterans and nonveterans for each one standard deviation increase (or decrease) in the psychiatric score. Therefore, a large and statistically significant coefficient on δ_2 would indicate the existence of a synergistic relationship between genotype and military service on the phenotypic outcome of interest. Similarly, the coefficient on δ_1 represents the local average treatment effect of military status, or the marginal effect of veteran status on depression at the mean polygenic score. In this way, the model allows us to estimate the effects of military service on depression across the entire distribution of genetic risk for psychiatric disorders.

Modeling G×E Interactions With Differences-in-Differences Estimation

DID estimation uses a time or cohort dimension to control for unobserved confounders. In a basic setup, outcomes are observed for two groups in two time periods. One group is exposed to a treatment in the second time period, and the other is never exposed to the treatment. For example, DID can be used to evaluate the effects of an exogenous policy change by comparing outcomes between treatment and control groups before and after a policy is enacted:

$$Y_{igt} = \alpha + \beta_1\delta_g + \beta_2\tau_t + \beta_3\left(\delta_g \times \tau_t\right) + X_{igt}'\beta_4 + \epsilon_{igt}. \tag{8}$$

In this equation, *i* indexes individuals, *g* indexes groups (1 if treatment group, 0 if control group), and *t* indexes years (1 if after the policy change, 0 if before). *X* is a vector of observable characteristics, including the first 10 principal components for population stratification in the genotype data, and *Y* is the outcome of interest. The fixed effects control for the time-invariant characteristics of the treatment group (β_1) and the time-series changes in $Y\ (\beta_2)$. The coefficient of interest on the interaction term (β_3) captures the variation in *Y* specific to the treatment group (relative to the control group) in the years after the law was passed (relative to before the law). Thus, any time- or group-invariant omitted variables that are correlated with being in the treatment group will be "differenced" out, and β_3 represents the causal impact of the policy change. The central assumption is that the average change in the outcome or trend would be the same for both groups in the absence of the treatment.

To accommodate differences by genotype, a differences-in-differences-in-differences (DDD) model can be employed:

$$\begin{aligned}Y_{igt} = {} & \alpha + \beta_1\delta_g + \beta_2\tau_t + \beta_3 G_i + \beta_4(\delta_g \times \tau_t) \\ & + \beta_5(\delta_g \times \tau_t \times G_i) + X'_{igt}\beta_6 + \epsilon_{igt},\end{aligned} \tag{9}$$

where G is the polygenic score of interest. Including the genotype fixed effect both controls for unobserved biological differences across individuals (β_3) and captures any variance in treatment intensity by genotype (β_5).

The quality of the control groups used is crucial to the validity of the estimates; good control groups must evolve similarly to the group experiencing the policy change and react similarly to other changes in the environment that drive policies to change (Besley & Case, 2000). Therefore, care must be taken to ensure group-level fixed effects absorb any potential confounders. For this reason, further interactions between genotype and group fixed effects could be included to account for genotypic differences between treatment and control groups. In addition, differences in exposure to environmental reforms between birth cohorts could be used in place of a time dimension to avoid problems of individual time-varying heterogeneity.

To use an example from the economics literature, suppose we were interested in the impact of earnings increases on employment or health. In their seminal study, Card and Krueger (1994) used an exogenous change in the state minimum wage in New Jersey to estimate a DID model that compared employment outcomes in the fast-food industry before and after the policy was enacted in New Jersey with a nearby state that did not raise its minimum wage (Pennsylvania). They found employment actually *increased* in New Jersey after the minimum wage hike. Here, as long as employment trends would be the same in both states in the absence of the treatment, state and time fixed effects control for any potential differences in geography or industry that could bias estimates. Along these lines, a polygenic score for educational attainment from the SSGAC could be incorporated to assess whether minimum wage increases contribute to better health outcomes for workers who are less likely to obtain a postsecondary education (Rietveld et al., 2013). In this case, a negative and significant result on β_5 would indicate an exogenous increase in wages resulted in better (marginal) health outcomes for individuals with lower scores for educational attainment relative to a control group with similar genetic attributes. This would seem to indicate that minimum wage policies might nurture health and human development by providing a safety net for individuals who are less likely to attend college and therefore more likely to work in lower-wage industries.

Similarly, if sibling data are available, Equation (9) could be transformed into a sibling difference model. Here, if one sibling is exposed to a "treatment" and the other is not, including sibling fixed effects would difference out any observable or unobservable environmental or genetic characteristics that are shared between siblings that might bias estimates. For example, Metzger and McDade (2010) used sibling pairs in which only one sibling was breastfed to evaluate the association between infant breastfeeding history and body mass index (BMI) in late childhood or adolescence. Since siblings share many of the major predictors of childhood obesity (e.g., parental obesity, household income, and family eating habits) a sibling fixed effect model is particularly useful in this context. Their findings indicated breastfeeding in infancy may be an important protective factor against the development of obesity in adulthood—if we can assume that the potential confounders are constant across siblings born to the same mother. If the authors had access to a polygenic score for BMI, they could have added a third difference to the mix and estimated whether the mitigation effects of infancy feeding are greater for individuals with higher than average BMI genetic risk scores.

Modeling G×E Interactions With a Regression Discontinuity Design

A basic "sharp" RD design estimates the causal effect of a treatment by exploiting a distinct cut-off or threshold above or below which a particular intervention is assigned. If treated and untreated individuals are similar near the cut-off point, then it is possible to estimate the local average treatment effect in environments where randomization is unfeasible. A unique feature of a "sharp" RD design is that there is no value of the variable that determines treatment where we can see both treatment and control observations (Imbens & Lemieux, 2008). For example, Hahn, Todd, and Van der Klaauw (1999) studied the effect of an antidiscrimination law on minority hiring that only applies to firms with at least 15 employees. Here, treatment is a deterministic and discontinuous function of the number of employees (i.e., firms with fewer than 15 employees are not subject to the law).

In certain cases, the assignment variable may be directly related to the outcome, and therefore the treatment effect will be related to the outcome as well, even if the treatment had no causal effect on the outcome. For example, the legal age of pension eligibility in a country has been used to identify the causal effect of retirement on health (e.g., Coe & Zamarro, 2011). In this case, the assignment variable (age) is associated with the outcome (health) and the treatment (pension eligibility). Here, the probability of receiving treatment, or retiring, does not change deterministically at the threshold, but instead acts as an exogenous mechanism that increases the probability of being retired.

When the assignment variable is directly related to the outcome, a "fuzzy" RD design that exploits discontinuities in the propensity score, or the probability of treatment conditional on covariates, is needed (e.g., Van der Klaauw, 2002). Basically, in a fuzzy RD design, the discontinuity acts as an instrumental variable for treatment status. Thus, in our example, in order for statutory retirement ages to be valid instruments, they must be predictive of actual retirement behavior. In addition, identification requires that there not be an independent, discontinuous change in the outcome of interest. When looking at how retirement affects health, this means the discontinuity in pension eligibility must be separate from any independent changes in health behaviors or changes in healthcare systems.

For the purpose of investigating G×E interactions, a fuzzy RD design is needed because genotype is likely correlated with both the assignment variable and the outcome of interest. To illustrate how genotype can be incorporated into a fuzzy RD framework, consider the following equation:

$$Y_i=\delta_0+\delta_1 E_i+f(a_i,G_i)+X_i'\varphi+\epsilon_i, \tag{10}$$

where, following our example, E is the endogenous environmental factor (retirement), Y is some health outcome, X is a vector of exogenous controls (including principal components), and ϵ is the disturbance term. The function $f(a_i, G_i)$ is included because policy eligibility is determined by age (a), which is a nonlinear, parametric function of health. The function also includes polygenic score G to allow policy effects to vary by genotype. Because E is endogenous, we exploit the probability of "treatment," or T_i, by using the discontinuity in the legal pension eligibility age, a_0, as an instrument:

$$T_{\mathrm{i}}=1(a_i \geq a_0), \tag{11}$$

where the dummy variable T_i is equal to 1 when an individual is at or above the legal pension age. Subsequently, the propensity score function, or the relationship between the probability of treatment, age, and genotype, can be written as follows:

$$P(E_i=1|a_i,G_i)=f_0(a_i,G_i)+[f_1(a_i,G_i)-f_0(a_i,G_i)]T_i, \tag{12}$$

where age in the trend function is modeled as a second-order polynomial for both the treatment and control groups (higher-order polynomials and semiparametric specifications could also be explored):

$$f_0(a_i,G_i)=\alpha_{00}+\beta_{01}\tilde{a}_i+\beta_{02}\tilde{a}_i^2+\beta_{03}G_i. \tag{13}$$

$$f_1(a_i,G_i)=\alpha_{01}+\rho+\beta_{11}\tilde{a}_i+\beta_{12}\tilde{a}_i^2+\beta_{13}G_i. \tag{14}$$

The age variable is centered, or $\tilde{a}_i \equiv a_i\text{-}a_0$. Centering a_i at a_0 ensures $a_i=a_0$ is the coefficient on T_i in a model with interaction terms. Based on the propensity score function, E can be instrumented with T in the first stage as follows:

$$\begin{aligned}E_i= \; &\alpha_{00}+\beta_{01}\tilde{a}_i+\beta_{02}\tilde{a}_i^2+\beta_{03}G_i+\rho T_i+\beta_1^*\tilde{a}_iT_i\\ &+\beta_2^*\tilde{a}_i^2T_i+\beta_3^*G_iT_i+X_i'\psi+\xi_{1i},\end{aligned} \tag{15}$$

where $\beta_1^*=\beta_{11}-\beta_{01}$, $\beta_2^*=\beta_{12}-\beta_{02}$, and $\beta_3^*=\beta_{13}-\beta_{03}$. Analogous first-stage results must be constructed for each of the polynomial interaction terms in the endogenous set $\{\tilde{a}_iE_i,\ \tilde{a}_i^2E_i,\ G_iE_i\}$ and substituted into the structural equation to derive the reduced form:

$$\begin{aligned}Y_i=&\gamma_0+\gamma_1\tilde{a}_i+\gamma_2\tilde{a}_i^2+\gamma_3G_i+\gamma_4T_i+\gamma_5\tilde{a}_iT_i\\ &+\gamma_6\tilde{a}_i^2T_i+\gamma_7G_iT_i+X_i'\phi+v_i.\end{aligned} \tag{16}$$

In this case, the treatment effect at $a_i-a_o=c>0$ is $\gamma_4+\gamma_5c+\gamma_6c^2+\gamma_7G_i$, whereas the treatment effect at a_0 is $\gamma_4+\gamma_7G_i$. Importantly, the treatment effect includes the G×E interaction, γ_7, which compares treated and untreated groups with the same polygenic score close to the cut-off point, or age of pension eligibility. Because these two groups have essentially the same value for $f(a_i, G_i)$, we can expect individuals just below the cut-off age for pension eligibility to be very similar to individuals just above the cut-off, and thus to have similar average outcomes in the absence of the program as well as similar average outcomes when receiving treatment.

LIMITATIONS OF THE QUASI-NATURAL EXPERIMENT APPROACH TO G×E ANALYSIS

While quasi-natural experimental designs can more effectively isolate exogenous variation in observational data, limitations of this approach should be mentioned. A significant drawback of these frameworks, and a common criticism of the natural experiment approach to econometric analysis in general, is that we cannot fully spell out the underlying theoretical relationships or causal mechanisms at play (e.g., Angrist & Krueger, 2001).

In our example of IV estimation using the Vietnam-era draft lotteries, we cannot pinpoint the impact of specific aspects of the war experience surrounding time in Vietnam (e.g., harshness of military training, combat positions, overseas travel, or number of tours) on mental illness, making it difficult to identify specific cause-effect relationships on the environment side. In addition, good instruments that can properly isolate exogenous statistical variation are challenging to find, and few instruments are generally accepted as solutions to endogeneity in the literature. Natural experiments that are fairly rare or leave few individuals treated may also reduce the potential population of participants, resulting in inadequate statistical power to detect G×E effects.

A related concern is the difficulty in structuring natural experiments that are informative with regard to research on psychological development. For example, in a G×E study on substance abuse, finding an adequate proxy or instrument for the randomization of children to different levels of parental monitoring, which tends to moderate genetic influences on substance use, would be extremely difficult to come by. Yet, even here the discovery of sound natural experiments, though challenging, is possible. A particularly ripe example is the use of exogenous income interventions to measure the mental health of children whose families moved out of poverty compared to those who were never poor or remained poor (Costello, Compton, Keeler, & Angold, 2003; Gennetian & Miller, 2002). In the case of the Costello et al. (2003) study, the influx of income to families of Native American descent in the Great Smoky Mountains Study from the opening of a new casino was used to test the effect of social causation on the trajectory of child and adolescent psychopathology. The authors hypothesized that if poverty had a causal role in inducing mental illness—meaning social causation or a G×E interaction is at play—then relieving poverty would reduce symptoms. Conversely, if social selection or a gene-environment correlation dominates, alleviation would have little effect on symptoms. Results were consistent with a social causation hypothesis, or moving out of poverty was associated with a decrease in the frequency of certain psychiatric symptoms (conduct and oppositional disorder), whereas other symptoms (depression and anxiety) were unaffected.

In this case, researchers used a natural experiment to identify not only the causal effect of income on childhood psychopathology, but also whether the nature of the genetic vulnerability for various psychiatric disorders was a by-product of rGE or G×E. In this article, we discuss the presence of rGE mainly as a methodological confound in G×E interaction models, but rGE is an integral part of the psychological development process (e.g., Bouchard, Lykken, McGue, Segal, & Tellegen, 1990; Moffitt, 1993; Scarr & McCartney, 1983). Since the distinction between rGE and G×E matters when suggesting options for treatment and intervention (Rutter, Pickles, Murray, & Eaves, 2001), quasi-natural experimental methods that can effectively rule out the presence or absence of a G×E interaction will help target proper strategies that can guide individuals toward trajectories of healthy development.

Ultimately, we emphasize that the primary advantage of the natural experiment approach to G×E research is to gain a stronger footing in claims of internal validity. Even though results may not be generalizable to larger populations and the underlying causal relationships may not always be identifiable, because the source of statistical variation is known and isolated, we can begin to use results from these experiments as a stepping stone for future work. For example, if the impact of Vietnam-era service on mental illness displays significant variation by genetic endowment, researchers can use these findings to guide studies that target more specific pathways between military service, genetic inheritances, and psychiatric disorders. In this way, the use of quasi-experimental methods is just one step in the G×E discovery process: Quasi-experiments cull and isolate statistical variation from large observational data sets, whereas theory and other quantitative or qualitative methods in the biological, psychological, or sociological sciences are needed to trace results back to underlying environmental phenomena. Likewise, the natural experiment approach to G×E work should be fundamentally grounded in theory, and a behavioral model should motivate the choice of instruments or experiments, which can in turn be used to support or refute interpretation of the estimates (Angrist & Krueger, 2001, p. 76).

CONCLUSION

We incorporate the latest approaches from population genetics into quasi-natural experimental frameworks to improve the measurement and estimation of G×E interplay in the social and behavioral sciences. We discuss the use of polygenic scores to maximize the amount of genetic information available on an individual into a single, quantitative measure of genetic risk, thus minimizing the possibility that “G” is acting as a proxy for other rGE, G×G, E×E, or G×E interactions. This approach also has the added advantage of using main effect analysis already extant in the literature that benefits from large consortia of adequately powered data to detect individual allelic effects. Testing well-established main effects in independent samples effectively reduces the number of hypotheses tested from millions (of SNPs) times the number of environmental regimes to one index score times the number of environmental factors tested. To avoid any confounding from nonrandom genetic assignment or ancestral differences, we discuss the use of principal components and sibling fixed effects, among others. Given the lack of family data available in nationally representative studies that have genotyped their participants, the use of principal components provides a simple and efficient way to control for population stratification alone or in combination with a mixed linear model that allows for non-independence of error terms based on relatedness between pairs of individuals. Finally, we provide a basic sketch of how these techniques can be incorporated into IV, DID, and RD frameworks to isolate variation in environmental exposure.

While there are several advantages to this approach, the drawback is that one must accept the natural experiments (and polygenic risk scores) one can find. However, we feel it is better to err on the side of good research design rather than on idealized operationalization of environmental variables. Moreover, due to endogeneity issues, current methods being used to uncover G×E interactions are inadequate to support policy inference. Although estimates from quasi-natural experiments may not be externally valid or directly applicable to policy in all cases, their high degree of internal validity may direct practitioners to effective treatments for those health or social outcomes that are the most environmentally responsive or genotypically influenced. Thus, while inducing a military draft lottery, for example, would not be an intervention to promote public health, to the extent that the Vietnam-era draft lottery serves as a proxy for stressful events in young adulthood, or exposure to combat, policymakers may want to design interventions to minimize similar stressful events that may have lasting effects on the development of social inequality over the life course. That is, this approach does not limit the range of policy or intervention options to the particular environmental factor being explored.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

References

Andreasen, C. H., Stender-Petersen, K. L., Mogensen, M. S., Torekov, S. S., Wegner, L., Andersen, G., et al. (2008). Low physical activity accentuates the effect of the FTO rs9939609 polymorphism on body fat accumulation. *Diabetes*, **57**, 95–101.

Angrist, J. D. (1990). Lifetime earnings and the Vietnam-era draft lottery: Evidence from social security administrative records. *American Economic Review*, **80**, 313–336.

Angrist, J. D., & Chen, S. H. (2011). Schooling and the Vietnam-era GI Bill: Evidence from the draft lottery. *American Economic Journal: Applied Economics*, **3**, 96–118.

Angrist, J. D., Chen, S. H., & Frandsen, B. R. (2010). Did Vietnam veterans get sicker in the 1990s? The complicated effects of military service on self-reported health. *Journal of Public Economics*, **94**, 824–837.

Angrist, J. D., Chen, S. H., & Song, J. (2011). Long-term consequences of Vietnam-era conscription: New estimates using social security data. *American Economic Review*, **101**, 334–338.

Angrist, J. D., & Krueger, A. B. (2001). Instrumental variables and the search for identification: From supply and demand to natural experiments. *The Journal of Economic Perspectives*, **15**, 69–85.

Angrist, J. D., & Pischke, J.-S. (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton, NJ: Princeton University Press.

Belsky, D. W., & Israel, S. (2014). Integrating genetics and social science: Genetic risk scores. *Biodemography and Social Biology*, **60**, 137–155.

Belsky, D. W., Moffitt, T. E., & Caspi, A. (2013). Genetics in population health science: Strategies and opportunities. *American Journal of Public Health*, **103**(S1), S73–S83.

Benjamin, D. J., Cesarini, D., Chabris, C. F., Glaeser, E. L., Laibson, D. I., Guðnason, V., et al. (2012). The promises and pitfalls of genoeconomics. *Annual Review of Economics*, **4**, 627–662.

Besley, T., & Case, A. (2000). Unnatural experiments? Estimating the incidence of endogenous policies. *The Economic Journal*, **110**(467), 672–694.

Bet, P. M., Penninx, B. W. J. H., Bochdanovits, Z., Uitterlinden, A. G., Beekman, A. T. F., van Schoor, N. M., et al. (2009). Glucocorticoid receptor gene polymorphisms and childhood adversity are associated with depression: New evidence for a gene–environment interaction. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, **150**, 660–669.

Bouchard, T. J., Lykken, D. T., McGue, M., Segal, N. L., & Tellegen, A. (1990). Sources of human psychological differences: The Minnesota study of twins reared apart. *Science*, **250**(4978), 223–228.

Branigan, A. R., McCallum, K. J., & Freese, J. (2013). Variation in the heritability of educational attainment: An international meta-analysis. *Social Forces*, **92**, 109–140.

Card, D., & Krueger, B. (1994). Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania. *American Economic Review*, **84**, 772–793.

Cha, S. W., Choi, S. M., Kim, K. S., Park, B. L., Kim, J. R., Kim, J. Y., et al. (2008). Replication of genetic effects of FTO polymorphisms on BMI in a Korean population. *Obesity*, **16**, 2187–2189.

Chang, Y.-C., Liu, P.-H., Lee, W.-J., Chang, T.-J., Jiang, Y.-D., Li, H.-Y., et al. (2008). Common variation in the fat mass and obesity-associated (FTO) gene confers risk of obesity and modulates BMI in the Chinese population. *Diabetes*, **57**, 2245–2252.

Christakis, N. A., & Fowler, J. H. (2014). Friendship and natural selection. *Proceedings of the National Academy of Sciences*, **111**(Supplement 3), 10796–10801.

Coe, N. B., & Zamarro, G. (2011). Retirement effects on health in Europe. *Journal of Health Economics*, **30**, 77–86.

Conley, D. (2009). The promise and challenges of incorporating genetic data into longitudinal social science surveys and research. *Biodemography and Social Biology*, **55**, 238–251.

Conley, D. (2015). Genotyping a new, national household panel study: White paper prepared for NSF-sponsored Conference, May 2014. *Journal of Economic and Social Measurement*, **40**, 349–369.

Conley, D., & Heerwig, J. (2011). The war at home: Effects of Vietnam-era military service on postwar household stability. *American Economic Review*, **101**, 350–354.

Conley, D., & Heerwig, J. (2012). The long-term effects of military conscription on mortality: Estimates from the Vietnam-era draft lottery. *Demography*, **49**, 841–855.

Conley, D., Domingue, B. W., Cesarini, D., Dawes, C., Rietveld, C. A., & Boardman, J. D. (2015). Is the effect of parental education on offspring biased or moderated by genotype? *Sociological Science*, **2**, 82–105.

Conley, D., Siegal, M. L., Domingue, B. W., Harris, K. M., McQueen, M. B., & Boardman, J. D. (2014). Testing the key assumption of heritability estimates based on genome-wide genetic relatedness. *Journal of Human Genetics*, **59**, 342–345.

Costello, J. E., Compton, S. N., Keeler, G., & Angold, A. (2003). Relationships between poverty and psychopathology: A natural experiment. *Journal of the American Medical Association*, **290**, 2023–2029.

De los Campos, G., Vazquez, A. I., Fernando, R., Klimentidis, Y. C., & Sorensen, Dl. (2013). Prediction of complex human traits using the genomic best linear unbiased predictor. *PLoS Genetics*, **9**(7), E1003608.

Dina, C., Meyre, D., Gallina, S., Durand, E., Körner, A., Jacobson, P., et al. (2007). Variation in FTO contributes to childhood obesity and severe adult obesity. *Nature Genetics*, **39**, 724–726.

Dobkin, C., & Shabani, R. (2009). The health effects of military service: Evidence from the Vietnam draft. *Economic Inquiry*, **47**, 69–80.

Domingue, B. W., Fletcher, J., Conley, D., & Boardman, J. D. (2014). Genetic and educational assortative mating among U.S. adults. *Proceedings of the National Academy of Sciences*, **111**, 7996–8000.

Duncan, L. E., & Keller, M. C. (2011). A critical review of the first 10 years of candidate gene-by-environment interaction research in psychiatry. *American Journal of Psychiatry*, **168**, 1041–1049.

Fletcher, J. M., & Conley, D. (2013). The challenge of causal inference in gene–environment interaction research: Leveraging research designs from the social sciences. *American Journal of Public Health*, **103**(S1), S42–S45.

Frayling, T. M., Timpson, N. J., Weedon, M. N., Zeggini, E., Freathy, R. M., Lindgren, C. M., et al. (2007). A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science*, **316**(5826), 889–894.

Furberg, H., Kim, Y., Dackor, J., Boerwinkle, E., Franceschini, N., Ardissino, D., et al. (2010). Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nature Genetics*, **42**, 441–447.

Gennetian, L. A., & Miller, C. (2002). Children and welfare reform: A view from an experimental welfare program in Minnesota. *Child Development*, **73**, 601–620.

Gibson, G. (2012). Rare and common variants: Twenty arguments. *Nature Reviews Genetics*, **13**, 135–145.

Hahn, J., Todd, P., & Van der Klaauw, W. (1999). Evaluating the effect of an antidiscrimination law using a regression-discontinuity design (No. w7131). National Bureau of Economic Research.

Hamer, D., & Sirota, L. (2000). Beware the chopsticks gene. *Molecular Psychiatry*, **5**, 11–13.

Harris, K. M., Halpern, C.T., Whitsel, E., Hussey, J., Tabor, J., Entzel, P., et al. (2009). *The National Longitudinal Study of Adolescent to Adult Health: Research design*. Retrieved from http://www.cpc.unc.edu/projects/addhealth/design

Haworth, C., Carnell, S., Meaburn, E. L., Davis, O. S. P., Plomin, R., & Wardle, J. (2008). Increasing heritability of BMI and stronger associations with the FTO gene over childhood. *Obesity*, **16**, 2663–2668.

Health and Retirement Study. (2015). Produced and distributed by the University of Michigan with funding from the National Institute on Aging (grant number NIA U01AG009740). Ann Arbor, MI.

Heerwig, J. A., & Conley, D. (2013). The causal effects of Vietnam-era military service on post-war family dynamics. *Social Science Research*, **42**, 299–310.

Hunt, S. C., Stone, S., Xin, Y., Scherer, C. A., Magness, C. L., Iadonato, S. P., et al. (2008). Association of the FTO gene with BMI. *Obesity*, **16**, 902–904.

Iacono, W. G., & McGue, M. (2002). Minnesota Twin Family Study. *Twin Research*, **5**, 482–487.

Imbens, G. W., & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, **142**, 615–635.

Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S., Freimer, N. B., et al. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics*, **42**, 348–354.

Lee, H. S., Ripke, S., Neale, B. M., Faraone, S. V., Purcell, S. M., Perlis, R. H., et al. (2013). Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nature Genetics*, **45**, 984–994.

Levy, D., Ehret, G. B., Rice, K., Verwoert, G. C., Launer, L. J., Dehghan, A., et al. (2009). Genome-wide association study of blood pressure and hypertension. *Nature Genetics*, **41**, 677–687.

Liang, K.-Y., & Zeger, S.t L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13–22.

Metzger, M. W., & McDade, T. W. (2010). Breastfeeding as obesity prevention in the United States: A sibling difference model. *American Journal of Human Biology*, **22**, 291–296.

Meyer, B. D. (1995). Natural and quasi-experiments in economics. *Journal of Business & Economic Statistics*, **13**, 151–161.

Moffitt, T. E. (1993). Adolescence-limited and life-course-persistent antisocial behavior: A developmental taxonomy. *Psychological Review*, **100**, 674–701.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, **38**, 904–909.

Rietveld, C. A., Medland, S. E., Derringer, J., Yang, J., Esko, T., Martin, N. W., et al. (2013). GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. *Science*, **340**(6139), 1467–1471.

Rutter, M. (2006). *Genes and behavior: Nature-nurture interplay explained*. Malden, MA: Blackwell.

Rutter, M., Pickles, A., Murray, R., & Eaves, L. (2001). Testing hypotheses on specific environmental causal effects on behavior. *Psychological Bulletin*, **127**, 291–324.

Scarr, S., & McCartney, K. (1983). How people make their own environments: A theory of genotype→ environment effects. *Child Development*, **54**, 424–435.

Scuteri, A., Sanna, S., Chen, W.-M., Uda, M., Albai, G., Strait, J., et al. (2007). Genome-wide association scan shows genetic variants in the FTO gene are associated with obesity-related traits. *PLoS Genetics*, **3**(7), e115.

Sonnega, A., Faul, J. D., Ofstedal, M. B., Langa, K. M., Phillips, J. W., & Weir, D. R. (2014). Cohort profile: The Health and Retirement Study (HRS). *International Journal of Epidemiology*, **43**, 576–585.

Splansky, G. L., Corey, D., Yang, Q., Atwood, L. D., Cupples, L. A., Benjamin, E. J., et al. (2007). The third generation cohort of the National Heart, Lung, and Blood Institute's Framingham Heart Study: Design, recruitment, and initial examination. *American Journal of Epidemiology*, **165**, 1328–1335.

St George-Hyslop, P. H. (2000). Molecular genetics of Alzheimer's disease. *Biological Psychiatry*, **47**, 183–199.

Strittmatter, W. J., Saunders, A. M., Schmechel, D., Pericak-Vance, M., Enghild, J., Salvesen, G. S., & Roses, A. D. (1993). Apolipoprotein E: High-avidity binding to beta-amyloid and increased frequency of type 4 allele in late-onset familial Alzheimer disease. *Proceedings of the National Academy of Sciences*, **90**, 1977–1981.

Van der Klaauw, W. (2002). Estimating the effect of financial aid offers on college enrollment: A regression–discontinuity approach. *International Economic Review*, **43**, 1249–1287.

Vink, J. M., & Boomsma, D. I. (2002). Gene finding strategies. *Biological Psychology*, **61**, 53–71.

Visscher, P. M., Hill, W. G., & Wray, N. R. (2008). Heritability in the genomics era—Concepts and misconceptions. *Nature Reviews Genetics*, **9**, 255–266.

Yang, J., Loos, R. J. F., Powell, J. E., Medland, S. E., Speliotes, E. K., Chasman, D. I., et al. (2012). FTO genotype is associated with phenotypic variability of body mass index. *Nature*, **490**(7419), 267–272.