NBER WORKING PAPER SERIES

THE EQUAL ENVIRONMENTS ASSUMPTION IN THE POST-GENOMIC AGE: USING MISCLASSIFIED TWINS TO ESTIMATE BIAS IN HERITABILITY MODELS

Dalton Conley
Emily Rauscher

Working Paper 16711
http://www.nber.org/papers/w16711

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
January 2011

This research uses data from Add Health, a program project directed by Kathleen Mullan Harris and designed by J. Richard Udry, Peter S. Bearman, and Kathleen Mullan Harris at the University of North Carolina at Chapel Hill, and funded by grant P01-HD31921 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development, with cooperative funding from 23 other federal agencies and foundations. Special acknowledgment is due Ronald R. Rindfuss and Barbara Entwisle for assistance in the original design. Information on how to obtain the Add Health data files is available on the Add Health website (http://www.cpc.unc.edu/addhealth). No direct support was received from grant P01-HD31921 for this analysis. This research was funded by the National Science Foundation's Alan T. Waterman Award, SES-0540543. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

The Equal Environments Assumption in the Post-Genomic Age: Using Misclassified Twins to Estimate Bias in Heritability Models
Dalton Conley and Emily Rauscher
NBER Working Paper No. 16711
January 2011
JEL No. I1,I21

ABSTRACT

While it has long been known that genetic-environmental covariance is likely to be non-trivial and confound estimates of narrow-sense (additive) heritability for social and behavioral outcomes, there has not been an effective way to address this concern. Indeed, in a classic paper, Goldberger (1979) shows that by varying assumptions of the GE-covariance, a researcher can drive the estimated heritability of an outcome, such as IQ, down to zero or up close to one. Survey questions that attempt to measure directly the extent to which more genetically similar kin (such as monozygotic twins) also share more similar environmental conditions than, say, dizygotic twins, represent poor attempts to gauge a very complex underlying phenomenon of GE-covariance. Methods that rely on concordance between interviewer classification and self-report offer similar concerns about validity. In the present study, we take advantage of a natural experiment to address this issue from another angle: Misclassification of twin zygosity in a nationally-representative study (Add Health). Since such twins were reared under one "environmental regime of similarity" while genetically belonging to another group, this reverses the typical GE-covariance and allows us bounded estimates of heritability for a range of outcomes of interest to medical and behavioral scientists.

Dalton Conley
New York University
6 Washington Square North #20
New York, NY 10003
and NBER
conley@nyu.edu

Emily Rauscher
New York University
295 Lafayette Street 4th Floor
New York, NY 10012
ekr220@nyu.edu

Introduction

To what extent are social and behavioral outcomes (or phenotypes) due to narrow-sense (additive) genetic heritability (h^2)? Notable researchers such as Richard Plomin or David Rowe, as well as many others, have argued that by comparing social outcomes among genetically identical twins (i.e. monozygotic twins who share 100 percent of their nuclear genes) with those from (same sex) fraternal twins (i.e. dizygotic twins who share, on average, 50 percent of their genes, just like singleton siblings), we can properly estimate the genetic, shared environmental, and non-shared environmental components of traits (see, e.g., Plomin, DeFries, McClearn, McGuffin, 2001). While there are other approaches to estimating heritability among humans, this is by far the most common approach and taken to be the least problematic since, being of a cohort together, both types of twins share uterine environments, experience societal events at the same time and deal with family transitions also at the same point in their development.

In the most naïve approach, narrow-sense (additive) genetic heritability is calculated as two times the difference between the intra-class correlations of identical and fraternal twins. (This is often estimated using an ACE model, where A stands for additive genetic heritability, C for common environment and E for unique environment [essentially an error term].) However, more recently, much more complex structural models have been offered to account for various complications such as the fact that—as a result of assortative mating at the parental level—fraternal twins may share more than 50 percent of their genes. Likewise, non-linear interactions between alleles—such as dominance—have been modeled in attempts to get at broad sense heritability (H^2) (see Purcell 2002 for a review of these models and simulation exercises and Purcell and Pak 2002 for an empirical example). And perhaps most

importantly, the "equal environments" assumption has been relaxed. For the naïve calculation mentioned above, it is necessary to assume that the covariance between environment and genetics is zero—better known as the equal environments assumption (EEA). Put another way, the simple estimation of heritability requires the rather heroic assumption that identical twins experience the same degree of similarity in environment as do (same sex) fraternal twins.

Such newer models include an estimate of the degree to which environmental similarity varies with genetic likeness. However, these are just that: estimates—often based on questions about whether or not respondents were "dressed alike" growing up, whether they were viewed as similarly as "two peas in a pod" and so on (see, e.g., Lichtenstein, Pedersen, and McClearn 1992; Rodgers et al. 1999; Rowe and Teachman 2001; Guo and Stearns 2002). Such questions are likely to capture only some of the ways that environmental similarity differs across identical and fraternal twin pairs, which is troubling since Goldberger (1979) has shown that depending on the GE covariance assumed, estimates of heritability can be driven wildly up or down.

Other more recent work has used adoptees to infer biological estimates of the heritability of social traits. For example, Sacerdote (2004) used a dataset of Korean adoptees in the United States where assignment to families was random (first-come-first-served basis) to examine the intergenerational correlation on important socioeconomic indicators such as educational attainment and income; on behaviors such as drinking and smoking; and on anthropometric measures such as height and weight. The results were then contrasted to intergenerational correlations among biological families from other data sources as well as biological children within those same families (for the subsample that contained biological

children). The results showed that—as might be expected—heritability for physical traits was considerably stronger in biologically intact families. Education (specifically probability of graduating from a four year college) and income were also much more strongly inherited by biological descent. However, health-related behavioral inheritance was similar across the two groups.

Before we accept the putative inference that education and income are predominantly genetically transmitted (while smoking and drinking are culturally transmitted) we must question the external validity of the adoptee sample. While there was adequate variation within the recipient families of adoptees, *on observables*, and while they did not look terribly different on average from non-adopting U.S. families, on observables, we know, *ipso facto*, that families who adopt are a distinct social group on unobservables—as are the adoptees themselves. For example, if socialization is weaker among adoptees who do not feel connected to their adoptive parents, heritability could appear to be weaker by virtue of this fact, not the absence of genetic similarity. There are many other dynamics that could be at work as well, such as increased (or decreased) parental investment, halo effects or stigma and truncated genetic variability among adoptees (or adopters), which may work to bias estimates for this population in unpredictable ways. The only adoption study that would avoid such questions would be one in which adoptees were randomly selected from the newborn population and then randomly assigned to parents, with both groups blind to the treatment (i.e. not knowing whether they were adopted or not)—all while prenatal environment was held constant. In other words, it is an impossibility to reliably estimate genetic heritability using such an approach.

Another intriguing recent study uses sibling identity by descent (IBD) to estimate heritability (for height). The approach of Visscher et al. (2006) was to identify the degree to which siblings shared polymorphisms at about 629 sites (this was the mean number of markers, the range was 201 to 1,717). The correlation of siblings on measured genotype was then compared with the degree of their resemblance on the phenotype—in this case height. The intraclass correlation for siblings ranged from 0.374 to 0.617 (with a mean of 0.498). It is these differences in IBD that they leverage to identify the genetic similarity and thereby estimate heritability. However, they make the assumption that this range of sibling genetic similarity arises from random differences in recombination. However, while this range falls in line with other estimates (c.f., Gagnon, Beise and Vaupel 2005), it does not just result from random variation due to recombination and segregation, it could also result from differential rates of assortative mating at the parental generation. Indeed, in order to arrive at the h2 estimates, Visscher et al. (2006) must assume random mating. This may, in fact, be the case; however, it is an assumption that could easily be tested by comparing the IBD of the parents (or by using sibling sets of three or more individuals and then deploying fixed effects before calculating IBD dyadic correlations on the residual, which would, in fact, be a random result of recombination). We are not suggesting that Visscher et al.'s estimates are necessarily wrong, merely that they would be nice complemented by an alternative approach to deal with GE covariance. We outline such an approach below.

Data and Methods

Given the intractability of adoption studies and the limitations of IBD correlation approaches, in the present analysis we deploy a different approach to improve on the standard

ACE model: We examine the intra-class correlation for monozygotic and (same sex) dizygotic twins who accurately perceive their genetic relatedness and separately for those twin sets who are, in fact, mistaken about their degree of genetic similarity. A non-trivial number of same sex twins are, in fact, incorrect about their zygosity. In Japan, for example, one study that deployed four independent samples found that, in each, between a quarter and 30 percent of MZ twins were misclassified as DZ twins at birth (Ooki, Yokoyama & Asaka 2004). Likewise, in Norway, a study revealed that a questionnaire approach to classifying the zygosity of adult twins was inaccurate 2.4 percent of the time when information from both twins was available and 3.9 percent of the time when information from only one twin was obtained (due to the death of or non-response from the other twin) (Magnus, Berg, & Nance 1983). Finally, a study in Denmark deployed the four traditional questions typically used to assign zygosity and then checked these predictions against genetic test results and found that the overall proportion misclassified was four percent, with the highest error rate among male monozygotic twins (8 percent) (Christiansen et al. 2003). Finally, a study that genotyped 327 Dutch twin pairs found a parental misclassification rate of 19 percent—largely as a result of monozygotic twins perceived as dizygotic (Van den Oord, Boomsma & Verhulst. 2000). So we can imagine the Scandinavian results as lower bounds and the Japanese figure as upper bounds of twin misclassification. In the United States, the National Longitudinal Survey of Adolescent Health is the only nationally representative dataset with self-reported zygosity, researcher-assigned zygosity and "true" genetic zygosity based on genetic testing.

When we examine these data, we find that six twin sets disagree about their collective zygosity (these siblings are excluded from our analysis). Of the remaining 254 same sex twin sets that agree on their zygosity, 45 are incorrect (17.7 percent). The vast majority of these

misperceiving siblings (82.2 percent) are genetically monozygotic twins who thought they were dizygotic. These zygosity assessments are obtained in the first wave of data collection, when the twins range in age from 12 to 18. Thus the 18 percent misclassification rate is understandably lower than the Japanese rate at birth. Likewise, it is understandably higher than the Norwegian or Danish rates, which were asked of adults and were not self-perceived zygosity but rather interviewer assigned zygosity based on a series of questions. Indeed, when Add Health assigns zygosity to twin sets based on a series of questions (such as whether they looked like two peas in a pod as children and were confused by strangers, teachers, or family members), the misclassification rate falls to a mere 5.9 percent. However, a significant additional proportion (6.6 percent) of twin sets remain "undetermined" under this methodology.

Add Health assigned twin zygosity based on a series of questions about similarity. These questions include: growing up, how alike did you and your twin look? Like two peas in a pod or family members; did you and your twin ever confuse strangers?; did you and your twin ever confuse teachers?; did you and your twin ever confuse family members? The similarity score for each pair is the average of these confusability questions for both twins. (These are the same sort of questions typically used to estimate GE covariance.) If a pair was missing answers to these questions, mothers' responses to questions about similarity were used. Comparing similarity score to self-reported zygosity among same-sex twins, Add Health made classification decisions based on a natural cut-point, "a cutoff score where the score distribution seemed to divide naturally" (Rowe and Jacobson 1998: 2).

If a pair claimed they were fraternal, but Add Health would have classified them as identical based on a high similarity score, they were classified as undetermined. Add Health

suggests excluding these pairs or treating them as fraternal. Since we are concerned not with correct classification by the survey researcher, but rather with the lived experience of the twins themselves, we rely primarily on their self-reported zygosity to take advantage of the misclassified twins to interrogate the equal environments assumption.

To question the equal environments assumption, we compare the degree of resemblance among same-sex twins whose genetic and self-reported zygosity match, to those whose identities do not align with their genetic zygosity. Twin self-report is privileged over Add Health classification of zygosity because it better indicates twins' subjective experience. However, intra-class correlations are run multiple times, using both self-reported zygosity and Add Health classification in order to make sure results are not an artifact of our choices. (This sensitivity analysis shows that they are, in fact, similar, though not identical possibly due to differences in sample size; see Table 2.) We are not the first researchers to pursue this "misclassification strategy" to interrogate heritability estimates. Goodman and Stevenson (1989) use this methodology to disentangle genetic and environmental effects among a sample of 13-year-old British twins and find that hyperactivity and attentiveness appear to be about half heritable. They (1989: 694) assign "true" zygosity based on "physical similarity, the number of choria and placentae, and the hospital doctors ascription of zygosity and the parental opinion"; when these sources disagreed, fingerprints were analyzed and blood group was gathered in a few cases. Xian et al. (2000), Scarr and Carter-Saltzman (1979), and Kendler et al. (1993) find evidence to support the equal environments assumption based on a variety of twin data. Kendler and colleagues use female twins from the Virginia Twin Registry, Xian et al. use male twins from the Vietnam Era Twin Registry, and Scarr uses Philadelphia-area twins. Although Scarr and Carter-Saltzman use blood group and Kendler et

al. use DNA data to identify genetic zygosity for pairs of "probable" or "uncertain" status, Xian et al. rely solely on questions about similarity with no molecular evidence. Meanwhile, while innovative for the late 1970s, Scarr's and Carter-Saltzmann's blood group approach is problematic since these loci are not definitive or comprehensive enough. For example, in their data DZ twins differed only on an average of 2.75 blood group loci out of 12. With such high similarity among DZ twins, it implies that many sets who are similar on 12 out of 12 may nonetheless be DZ by chance. Kendler et al.'s approach is the closest to ours. However, they rely on a localized sample and similarity questions and photographs (available for about 80% of twins) to assign zygosity for a majority of their twin pairs. They classified pairs as definite, probable, or uncertain zygosity status based on similarity questions and photographs and then attempted to gather blood samples for the probable and uncertain categories (186 pairs). Blood samples, and therefore genetic zygosity, were available for 119 of these 186 pairs. Genetic information was available for 26 pairs classified as definite zygosity and validated Kendler's assignment in all cases. For the "probable" group, genetic zygosity matched their assignment for 83% of the pairs. To summarize, Kendler's final zygosity assignment relies on DNA data where available (a small portion of their pairs) and definite or probable classification based on similarity questions and photographs. Their DNA data suggests zygosity is assigned with high validity, but some error certainly remains – particularly among pairs in the probable category without genetic data. In contrast to the studies above, genetic zygosity is available for all twins in our data. Like Goodman and Stevenson (1989) and Xian et al., Kendler et al. focus on psychopathologies: major depression, generalized anxiety disorder, phobia, bulimia, and alcoholism. All of these studies find little evidence for significant violations of the equal environments assumption.

Against this backdrop, we are the first to apply this misclassification approach to a recent, nationally representative sample with genetic zygosity information for all twins over a wide range of behavioral and anthropometric outcomes and to address possible bias in the relationship between misclassification and phenotypic similarity due to reverse causation (phenotypic non-resemblance causing misclassification) by comparing perceived zygosity to birth weight discordance.

Table 1: Genetic zygosity by self-reported zygosity among same-sex twins (panel A) and by Add Health zygosity assignment (panel B).

Panel A:

Genetic	Self-Reported			
	MZ	Disagree	DZ	Total
MZ	208	10	74	292
DZ	16	2	210	228
Total	224	12	284	520

Panel B:

Genetic	Add Health Assignment			
	MZ	DZ	Undetermined	Total
MZ	260	18	30	308
DZ	12	220	6	238
Total	272	238	36	546

We focus on the third wave of Add Health panel data for sibling pairs, which surveyed respondents in 2001-2 when they were ages 18-26. Siblings of individuals identified as twins in the stratified (nationally representative) sample were added, yielding 64 percent of sibling pairs from the probability sample and 36 percent from convenience sampling. In other words, to increase the number of pairs, some siblings were added after the random sampling strategy. Sampling weights are therefore not available for genetic data.

Genetic zygosity was determined by 11 "highly polymorphic, unlinked short tandem repeat (STR) markers: D1S1679, D2S1384, D3S1766, D4S1627, D6S1277, D7S1808, D8S1119, D9S301, D13S796, D15S652 and D20S481" and a sex-linked-locus (Harris et al. 2006:992). Twins are classified as genetically monozygotic if they match at all 11 loci. Our sample includes nearly 150 identical twin pairs and over 110 same-sex fraternal twin pairs (although the exact sample size depends on the number of pairs with complete outcome data). Table 1 compares genetic zygosity to perceived zygosity in Panel A and Add Health assigned zygosity in Panel B. Panel A shows that 74 genetically identical twins perceive themselves as fraternal, while 16 genetically fraternal twins believe they are identical. Supplemental tables (S1-S3) provide descriptive measures by zygosity category and compare perceived and assigned zygosity to the similarity index Add Health used to assign zygosity. Mean differences between correctly and incorrectly classified twins are only significant for high school GPA and birth weight.

Our phenotypes include the following: Birth weight; height; weight; BMI; depression score; ADHD; delinquency; and cumulative high school GPA. Birth weight is reported by parents, measured in ounces, and logged. Height and weight, used to calculate body mass index, are self-reported in wave 3. Depression is measured using the Center for Epidemiologic Studies-Depression Scale (CES-D). It consists of 20 questions included in the Add Health survey which ask respondents to rate the frequency of a depressive symptom from 0 (never/rarely) to 3 (most/all of the time). The sum of responses for all 20 items indicates the frequency of depressive symptoms. A scale of attention deficit and hyperactivity disorder (ADHD) behaviors is constructed from 18 questions asked in wave 3 about behavior when the individual was between 5 and 12 years old. The ADHD scale indicates how often

(never/rarely, sometimes, often, or very often) the youth fidgeted, had difficulty sustaining attention in tasks, was forgetful, had difficulty organizing tasks or activities, and left his seat when being seated was expected, among other things. Cumulative high school GPA is gathered from high school transcripts. Heritability for these phenotypes can be simply estimated from the following equation 1:

$$\text{Var}(y) = \text{Var}(g) + \text{Var}(e) + 2\,\text{Cov}(g, e). \quad (1)$$

Where y is an outcome or phenotype, g is the genetic contribution, and e is the environmental contribution. Researchers usually suppose that Cov(g, e) = 0, so the equation reduces to Var(y) = Var(g) + Var(e). Then heritability is the ratio Var(g)/Var(y). To estimate this using MZ and DZ twin correlations, we rely on the following assumptions:

$$r_{mz} = A + C \quad (2)$$

$$r_{dz} = 0.5A + C \quad (3)$$

Where A is shared genetics, C is shared environment, and the 0.5 coefficient for DZ twins echoes the notion that they share, on average, 50 percent of their genes. (If positive assortative mating is at play, then this biases heritability downward). We then difference equations 2 and 3 and solve for A to yield equation 4, below:

$$A = 2\,(r_{mz} - r_{dz}) \quad (4)$$

Finally, we can deduce C from equation 2:

$$C = r_{mz} - A \quad (5)$$

And since we assume that MZ twins reflect maximal environmental and genetic similarity, E (the effect of unique environment) is simply:

$$E = 1 - r_{mz} \quad (6)$$

Again, this model is identified only because we assume away the covariance of A and C (cov GE in our earlier notation). However, in our case, we will estimate two versions of the model, one where we know that the 2*cov(G*E) term is positive—that includes the cases where the genetic and social zygosity match—and one where we assume the 2*cov(G*E) is negative due to the self-misclassification of the twins' zygosity. The covariance should be positive for correctly classified twins (because genetic and environmental similarity are aligned) but negative for misclassified twins (because environmental treatment should not mesh with genetic similarity). Therefore, we hypothesize that heritability estimates among correctly classified twins should overestimate heritability, while estimates among misclassified twins should underestimate heritability. Of course, we do not know a figure for the GE covariance for each group, but its valence is enough to test classically-determined heritability estimates for bias. We will not, then, try to estimate the *true* heritability (or the *true* parameters for components C and E), but merely obtain a sense of whether the bias is substantively and statistically significant.

In a second approach, we use Kendler et al.'s strategy of comparing model fit with and without perceived zygosity in the model. Phenotype is regressed on genetic zygosity and sex for all twins by genetic zygosity (among same-sex, white, same-sex white, and all twins), alternately including perceived zygosity. Genetic zygosity is coded 0.5 for DZ and 1 for MZ twins. Perceived zygosity is 0 for DZ, 0.5 if twins disagree, and 1 for MZ. Akaike (AIC) and Bayesian information criteria (BIC) are used to select the model best balancing data fit and parsimony (Raftery 1986; 1995). In both cases, lower values indicate better fit. BIC alone is insufficient because it may overvalue parsimony or simpler models (Weakliem 1999). Therefore, it is important to consider both statistics in deciding which model fits best.

One concern with our research strategy might be that we are reversing the causal process: Perhaps it is the case that twins who deviate greatly on the phenotypes of interest—say height, weight, GPA, affect—are then socially misclassified? This would then suggest that cov(G,E) is predicted by var(y), confounding our attempts at decomposition. To address this possibility, ideally we would instrument misclassification. As we shall show below, we do have a factor that temporally precedes self-perception of zygosity and strongly predicts it, thus fulfilling the first condition necessary for an instrument. This factor is birth weight differences between the twins. However, birth weight differences are likely to have direct effects on the similarity in phenotypes we consider, net of misclassification status. Birth weight has been shown to affect a range of anthropometric measures (see, e.g., Conley, Strully and Bennett 2003 for a review), and recent work has shown that differences themselves, in fact, have predictive power for the differences between siblings (including twins) (see Conley and Rauscher 2010). Thus, birth weight differences violate the exclusion restriction and would thus fail as an instrument. Indeed, it is likely that any factor that would affect the probability of misclassification would also affect the phenotypes, thus we abandoned the hope for an instrumentation strategy and rely instead on simple comparisons between correctly and incorrectly classified groups. That said, the birth weight analysis gives us some comfort in the notion that misclassification was a result of differences that began at birth and not as a result of the phenotypes under study.

Results

Figures 1 and 2 show intra-class correlations among MZ and DZ twins by perceived zygosity for BMI and high school GPA. In both cases, the correlation among genetic identical

twins is stronger than fraternal twins, whether the identical twins correctly perceive their zygosity or not. BMI shows a stronger distinction between genetically MZ and DZ twins, which supports the argument that BMI is largely heritable (e.g., Allison et al. 1996 find h^2 of BMI is between 0.5 and 0.7 based on twin data from Finland, Japan, and the US). Wide standard error bars illustrate the problem with using genetically fraternal twins who believe they are identical. The small sample sizes for misclassified DZ twins preclude using them.

Table 2 presents intraclass correlations of phenotypes by classification status for identical and fraternal twins. Heritability estimates using all correctly classified twins (column 5) and incorrectly classified MZ twins (column 6) are calculated for each phenotype. Figure 3 graphically compares heritability estimates for these correctly and incorrectly classified twins.

Table 2: Intraclass Correlation and Estimated Heritability by Self-Perceived Zygosity Category

	MZ Correct	MZ Incorrect	DZ Correct	DZ Incorrect	h^2 all Correct	h^2 DZ Correct & Perc DZ-Gen MZ	C Shared Env Correct	E Unique Env Correct	C Shared Env MZ Incorrect	E Unique Env MZ Incorrect
	1	2	3	4	5	6	7	8	9	10
BMI	0.84	0.87	0.35 *†	0.08 *†	0.98	1.00	-0.14	0.16	-0.13	0.13
Height	0.96	0.95	0.72 *†	0.49 *†	0.47	0.46	0.49	0.04	0.49	0.05
ADHD	0.44	0.51	0.24 *†	0.44	0.41	0.54	0.03	0.56	-0.03	0.49
Depression	0.27	0.62 *	0.15 †	.	0.25	0.94	0.40	0.16	0.38	0.15
GPA	0.84	0.85	0.62 *†	0.76	0.44	0.47	0.39	0.72	0.60	0.93

* = significantly different from MZ correct
† = significantly different from MZ incorrect

Figure 1: Twin intraclass correlations for Body Mass Index, by genetic and perceived zygosity; data from genetic subsample of the National Longitudinal Survey of Adolescent Health. Sample sizes are 200 for genetically MZ twins perceived accurately and 69 for MZ twins perceived inaccurately; 194 for same-sex genetically DZ twins perceived accurately and 16 for genetically DZ twins perceived inaccurately.

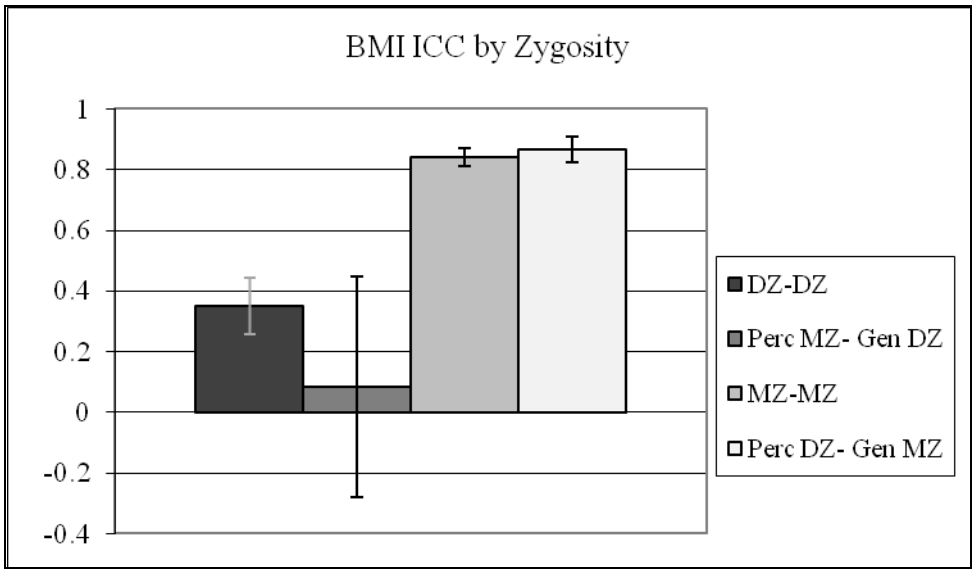


Figure 2: Twin intraclass correlations for cumulative High School GPA, by genetic and perceived zygosity; data from genetic subsample of the National Longitudinal Survey of Adolescent Health. Sample sizes are 185 for genetically MZ twins perceived accurately and 62 for MZ twins perceived inaccurately; 175 for genetically DZ twins perceived accurately and 13 for genetically DZ twins perceived inaccurately.

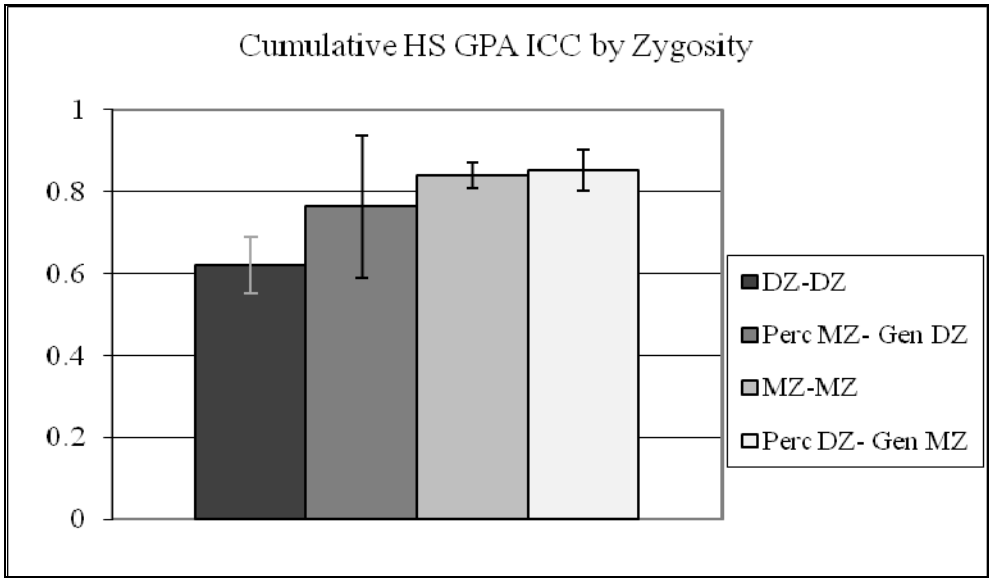
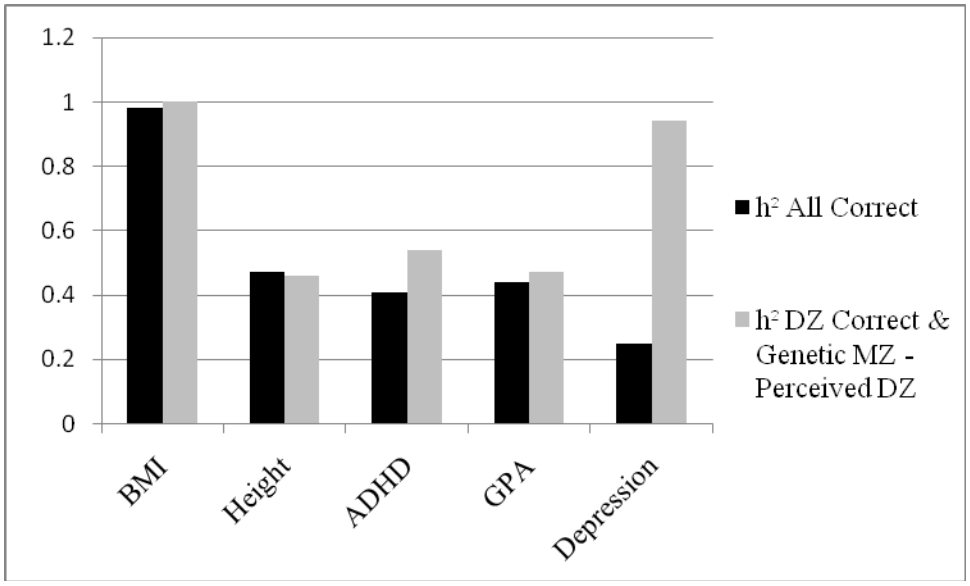


Figure 3: Narrow-sense (additive) heritability estimates (h^2) for correctly and incorrectly classified twins based on figures from Table 2a.



The estimated heritabilities of body mass index and height are about the same among correctly and incorrectly classified twins. Estimated heritability of height is slightly higher among incorrectly identified MZ twins, but in general estimates for BMI and height do not provide evidence that correctly classified twins underestimate heritability.

In contrast to these largely inherited outcomes, behavioral outcomes such as depression, ADHD, and GPA show higher heritability among incorrectly classified twins. Estimated heritability is only slightly higher for GPA, but substantially for ADHD and depression among misclassified twins. Oddly, identical twins who believe they are fraternal are more similar in GPA, depression, and ADHD symptoms than other MZ twins. (The difference is only significant for depression, however.) There could, of course, be a complicated behavioral response to similarity and difference across measures. For example, MZ twins who perceive themselves as DZ may be more similar in their psychological reactions to what they may sense as some discrepancy (perhaps that they are more "similar" on physical measures than they might expect to be given their belief that they are dizygotic—

however, mean levels of depression are not different for this misclassified group, complicating this story). Alternatively, it could be that MZ twins who correctly perceive themselves to be MZ psychologically seek to individuate more than those who perceive themselves as DZ and thus do not feel compelled to form psychological niches. Overall, comparing estimates for correctly and incorrectly classified twins suggest traditional heritability estimates are not overestimated, and may in fact be underestimated for behavioral phenotypes - particularly depression.

Columns 7-10 in Table 2 list estimated shared and unshared environmental contribution to phenotypes. Similar to the heritability estimates, shared environmental estimates are quite similar using correctly and incorrectly classified MZ correlations, except for depression and to a small extent ADHD. Depression and ADHD estimates suggest shared environment is less important among identical twins who believe they are fraternal. This suggests the equal environments assumption may be problematic, because shared environment is more important for twins who believe they are identical. Correctly classified identical twins may be treated more similarly than genetically MZ twins who believe they are fraternal. Shared environment estimates of ADHD and depression are negative, however, for incorrectly classified MZ twins, which makes this evidence weak. Estimated individual environmental contributions (E) are generally larger than shared environment (C). Only height and GPA have smaller individual environmental contributions – for both correctly and incorrectly classified identical twins.

Table 3 compares model fit for regressions predicting individual depression and pair depression difference using AIC and BIC statistics. Regressions include genetic zygosity (.5 for DZ and 1 for MZ twins) and an indicator for male (indicators for both male and opposite

sex in twin pair models including non-same-sex twins). Following Kendler et al., model fit is compared to a model including perceived zygosity (0 for DZ, .5 if twins disagree, and 1 for MZ). An AIC or BIC difference of 5 indicates a significant difference in model fit and lower values are better. In every case, using both AIC and BIC, perceived zygosity significantly improves model fit among white, same sex, same sex white, and all twins. Perceived zygosity improves prediction of both individual depression and twin pair depression difference, suggesting environmental differences due to perception are nontrivial.

Table 3: Model Fit with and without Perceived Zygosity

		White			Same-sex			Same-sex White			All		
	Individuals	Depression											
		AIC	BIC		AIC	BIC		AIC	BIC		AIC	BIC	
MZ	w/o	1234.6	1241.3		1824.5	1831.9		1234.6	1241.3		1824.5	1831.9	
	w/ perceived	1184.3	1194.2	*	1748.9	1759.9	*	1184.3	1194.2	*	1748.9	1759.9	*
DZ	w/o	1528.7	1535.8		1427.5	1434.4		849.7	855.7		2490.2	2498.5	
	w/ perceived	1434.8	1445.2	*	1353.9	1364.1	*	804.9	813.7	*	2347.4	2359.3	*
Twins	w/o	2764.2	2776.6		3252.1	3264.9		2084.2	2095.8		4315.2	4328.9	
	w/ perceived	2618.7	2635.1	*	3104.3	3121.3	*	1988.2	2003.5	*	4098.8	4116.8	*
	Pairs	Depression Difference											
		AIC	BIC		AIC	BIC		AIC	BIC		AIC	BIC	
MZ	w/o	564.4	569.7		842.0	848.0		564.4	569.7		842.0	848.0	
	w/ perceived	541.8	549.6	*	805.6	814.6	*	541.8	549.6	*	805.6	814.6	*
DZ	w/o	760.5	769.0		705.1	710.6		425.2	429.8		1231.1	1241.0	
	w/ perceived	716.6	727.7	*	675.7	683.9	*	402.8	409.5	*	1168.1	1181.2	*
Twins	w/o	1332.4	1346.2		1552.7	1563.5		991.3	1000.9		2081.5	2097.1	
	w/ perceived	1263.3	1280.3	*	1485.4	1499.7	*	943.8	956.3	*	1981.1	2000.2	*

Table 4 offers evidence that twin misclassification is driven at least partially by very early differences. Twins who are genetically identical, but misperceive themselves as fraternal, have significantly higher differences in birth weight. The sample size for incorrectly classified DZ twins is only 7 pairs, so results for this group are not conclusive. Among MZ twins, however, perceived zygosity is related to birth weight differences.

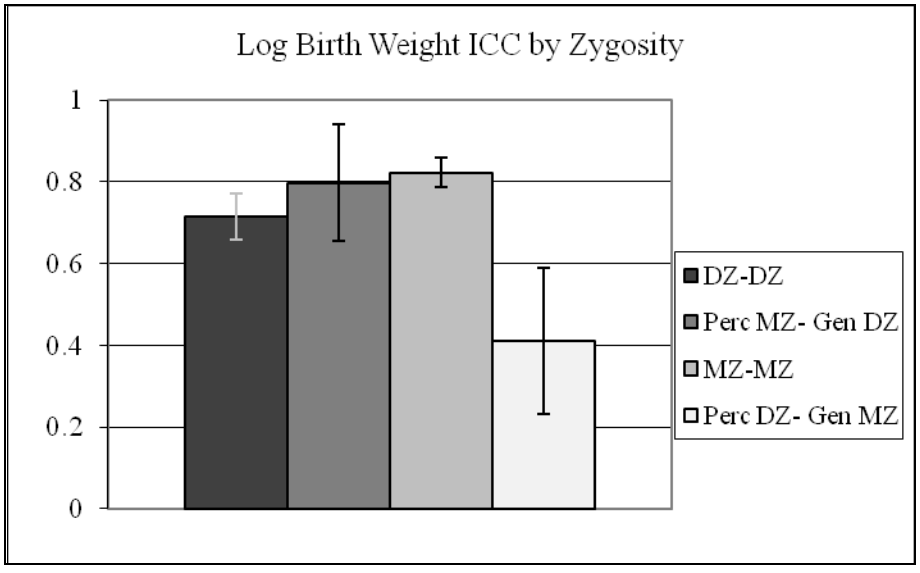
Figure 5 illustrates the relationship between birth weight and perceived zygosity. Misclassified MZ twins have substantially lower similarity in birth weight than all other twin types and likely encouraged their identification as DZ twins. Misclassified DZ twins had slightly higher birth weight similarity than their correctly classified counterparts, but the sample size is too small to reach significance.

Table 4: Birth weight differences by zygosity among same sex twins

	Birth Weight Difference	N (pairs)	Std Dev
MZ Correct*	0.08	74	0.07
DZ Correct	0.10	73	0.10
MZ Incorrect*	0.13	22	0.12
DZ Incorrect	0.08	7	0.09

* indicates significant difference between groups; birth weight differences are only significant between twin pairs who correctly and incorrectly identified as identical twins

Figure 5: Twin intraclass correlations for birth weight, by genetic and perceived zygosity; data from genetic subsample of the National Longitudinal Survey of Adolescent Health. Sample sizes are 159 for genetically MZ twins perceived accurately and 48 for MZ twins perceived inaccurately; 157 for genetically DZ twins perceived accurately and 14 for genetically DZ twins perceived inaccurately.



Discussion

Although it may be partially endogenous to the phenotypes studied here, misperception is significantly related to a much earlier difference – birth weight – which suggests misperception is not primarily codetermined with phenotypic similarity. Overall, the evidence suggests that typical twin heritability estimates of behavioral outcomes are not upwardly biased by failing to address the covariance between genes and environment. In other words, our evidence supports the equal environments assumption. Our results therefore build on previous research to suggest that phenotypic similarity and perceived zygosity are not co-determined. Perceived zygosity is influenced by differences as early as birth. Therefore, evidence supports methods used here and in previous studies which compare similarity based on actual and perceived zygosity to assess the equal environments assumption.

Results suggest that heritability estimates may be higher if we deploy comparisons among twins who misperceive their zygosity – but mainly for behavioral phenotypes. While we may not make much of these differences, they at least give us comfort that by setting the GE covariance term to zero in standard heritability models, we are not significantly biasing results. A number of approaches—ranging from the misclassification strategy pursued here to using IBD sibling resemblance models—seem to be converging on the results that the old narrow-sense heritability estimates are not far off. This assumes, of course, that the other assumption of random mating holds. However, if parents tend to be more alike genetically than they would be if mating were random (a likely case, especially if we believe genes are related to phenotypes and the same phenotypes that researchers tend to study are those on which mates also sort), then heritability estimates would be downwardly biased. There are instances where we might expect genetic opposites to attract, such as the major

histocompatibility complex where genetic diversity increases the chances of species or population survival through an epidemic. The phenotypes of interest to most social scientists and those studied here, however, are likely to see assortative mating (educational assortative mating – related to GPA, ADHD, delinquency, and depression – offers the most obvious example). So all in all, it seems reasonable to take results from an ACE model more or less at face value. In fact, we come to this conclusion grudgingly, having set out on this empirical exercise with the assumption that we were going to show h^2 to be overstated for our range of phenotypes due to omitted, positive GE covariance.

Work Cited

Allison, D.B., J. Kaprio, M. Korkeila, M. Koshenvuo, M.C. Neale, and K. Hayakawa. 1996. "The heritability of body mass index among an international sample of monozygotic twins reared apart." *International Journal of Obesity* 20: 501-506.

Christiansen L, Frederiksen H, Schousboe K, Skytthe A, von Wurmb-Schwark N, Christensen K, Kyvik K. 2003. "Age- and sex-differences in the validity of questionnaire-based zygosity in twins." *Twin Research*. 6:275-8.

Conley, Dalton, Kate W. Strully, and Neil G. Bennett. 2003. *The Starting Gate: Birth Weight and Life Chances*. Berkeley: University of California Press.

Conley, Dalton and Emily Rauscher. "Genetic Interactions with Prenatal Social Environment: Effects on Academic and Behavioral Outcomes." NBER Working Paper 16026 www.nber.org/papers/w16026

Gagnon, Alan, Jan Beise and J.W. Vaupel. 2005. "Genome-wide identity-by-descent sharing among CEPH siblings." *Genetic Epidemiology*. 29:215-224.

Goldberger, Arthur S. 1979. "Heritability." *Economica*, New Series 46(184):327-347.

Goodman, R. and J. Stevenson. 1989. "A twin study of hyperactivity—II. The aetiological role of genes, family relationships and perinatal adversity." *Journal of Child psychology and Psychiatry*. 30: 691 – 709.

Guo, Guang and Elizabeth Stearns. 2002. "The Social Influences on the Realization of Genetic Potential for Intellectual Development." *Social Forces* 80(3):881-910.

Harris, Kathleen Mullan, Carolyn Tucker Halpern, Andrew Smolen, and Brett C. Haberstick. 2006. "The National Longitudinal Study of Adolescent Health (Add Health) Twin Data." *Twin Research and Human Genetics* 9, 6: 988-997.

Kendler, Kenneth S., Michael C. Neale, Ronald C. Kessler, Andrew C. Heath, and Lindon J. Eaves. 1993. "A Test of the Equal-Environment Assumption in Twin Studies of Psychiatric Illness." *Behavior Genetics* 23, 1: 21-27.

Lichtenstein, P., N.L. Pedersen, and G.E. McClearn. 1992. "The origins of individual differences in occupational status and educational level: A study of twins reared apart and together." *Acta Sociologica* 35: 13-31.

Magnus P, Berg K, & Nance WE. 1983. "Predicting zygosity in Norwegian twin pairs born 1915-1960." *Clinical Genetics*. 24:103-12.

Ooki, S., Y. Yokoyama & A. Asaka. 2004. "Zygosity misclassification of twins at birth in Japan." *Twin Research*. 7:228-232.

Plomin, Robert, John C. DeFries, Gerald E. McClearn, and Peter McGuffin. 2001. *Behavioral Genetics* (4th ed.). New York: Worth Publishers.

Purcell, Shaun. 2002. "Variance Components Models for Gene-Environment Interaction in Twin Analysis." *Twin Research*. 5:554-571.

Purcell, Shaun and Pak. Sham. 2002. "Variance Components Models for Gene-Environment Interaction in Quantitative Trait Locus Linkage Analysis." *Twin Research*. 5:572-576.

Raftery, Adrian E. 1995. "Bayesian Model Section in Social Research." *Sociological Methodology* 25: 111-163.

Raftery, Adrian E. 1986. "Choosing Models for Cross-Classifications." *American Sociological Review* 51, 1: 145-146.

Rodgers, J.L., D.C. Rowe, and M. Buster. 1999. "Nature, nurture, and first sexual intercourse in the USA: Fitting behavioral genetic models to NLSY kinship data." *Journal of Biosocial Science* 31: 29-41.

Rowe, David C. and Kristen C. Jacobson. 1998. "National Longitudinal Study of Adolescent Health: Pairs Code Book." Carolina Population Center: Chapel Hill, NC.

Rowe, D. and J. Teachman. 2001. "Behavioral genetic research designs and social policy studies." In A. Thornton (ed.) *America's Families and Children: Research Needed in the Coming Millennium*. Ann Arbor, MI: University of Michigan Press: 157-187.

Sacerdote, Bruce. 2004. "What Happens When We Randomly Assign Children to Families?" NBER Working Paper No. W10894.

Scarr, Sandra and Louise Carter-Saltzman. 1979. "Twin Method: Defense of a Critical Assumption." *Behavior Genetics* 9, 6: 527-542.

Van den Oord, E., D.I. Boomsma & F.C. Verhulst. 2000. "A study of genetic and environmental effects on the co-occurrence of problem behaviors in three-year-old twins." *Journal of Abnormal Psychology*. 109:360-372.

Visscher, Peter M.,, Sarah E. Medland, Manuel A. R. Ferreira, Katherine I. Morley, Gu Zhu, Belinda K. Cornes, Grant W. Montgomery, Nicholas G. Martin. 2006. "Assumption-Free Estimation of Heritability from Genome-Wide Identity-by-Descent Sharing between Full Siblings." *PLoS Genetics.* 2: e41. doi:10.1371/journal.pgen.0020041

Weakliem, David L. 1999. "A Critique of the Bayesian Information Criterion for Model Selection." *Sociological Methods and Research* 27, 3: 359-397.

Xian, Hong, Jeffrey F. Scherrer, Seth A. Eisen, William R. True, Andrew C. Heath, Jack Goldberg, Michael J. Lyons, and Ming T. Tsuang. 2000. "Self-Reported Zygosity and the Equal-Environments Assumption for Psychiatric Disorders in Vietnam-Era Twin Registry." *Behavior Genetics* 30, 4: 303-310.

Supplementary Material

Supplemental tables provide descriptive measures by zygosity category (S1) and compare perceived and assigned zygosity to the similarity index Add Health used to assign zygosity (S2 and S3). Mean differences between correctly and incorrectly classified twins are only significant for high school GPA and birth weight. Identical twins who believe they are fraternal have significantly higher high school GPAs than correctly identified identical twins. The same pattern does not hold among fraternal twins who believe they are identical. Overall, all misclassified twins have significantly higher GPAs than all correctly classified twins. Birth weight is significantly higher among fraternal twins who believe they are identical than correctly perceived fraternal twins. This difference is not significant among identical or all twins.

Table S1: Means by Classification Category – Same Sex Twins

	MZ-MZ		DZ-Actual MZ			DZ-DZ		MZ-Actual DZ			Any Misclass		Correct Class			All		
	Mean	N	Mean	N		Mean	N	Mean	N		Mean	N	Mean	N		Mean	Std Dev	N
Male	0.47	208	0.41	74		0.53	210	0.38	16		0.40	90	0.50	418	+	0.48	0.5	508
HS GPA	2.67	185	3.01	62	**	2.62	175	2.59	13		2.93	75	2.64	360	**	2.69	0.79	435
Depression	5.90	207	5.62	74		5.37	208	4.00	16		5.33	90	5.63	415		5.58	4.98	505
ADHD	12.80	203	11.56	72		13.26	205	13.47	15		11.89	87	13.03	408		12.83	8.87	495
BMI	25.02	200	25.58	69		25.74	194	27.83	16		26.00	85	25.37	394		25.48	6.09	479
Obese	0.14	200	0.14	69		0.14	194	0.38	16	+	0.19	85	0.14	394		0.15	0.36	479
Height	66.86	202	66.41	71		67.50	198	66.74	16		66.47	87	67.18	400		67.05	4.19	487
Birth Weight (log oz)	4.48	159	4.50	48		4.51	157	4.61	14	*	4.53	62	4.49	316		4.5	0.18	378
Birth Weight (oz)	89.60	159	91.29	48		92.02	157	101.57	14	*	93.61	62	90.80	316		91.26	16.28	378

Differences between correctly and incorrectly classified twins are significant at: + p<.10; * p<.05; ** p<.01.

Tables S2 and S3 show Add Health zygosity assignment by similarity score (based on responses to questions about how similar the twins are). Table S2 illustrates the main cut-off in similarity for Add Health-assigned zygosity. Table S3 shows that similarity score and self-perceived zygosity is not as strongly related.

Table S2. Add Health zygosity assignment of same-sex twins by similarity score

Similarity Score	Add Health Assignment			
	MZ	DZ	Undetermined	Total
0	4	232	0	236
33.3	2	36	0	38
50	4	38	0	42
60	4	28	0	32
66.7	16	4	6	26
71.4	16	0	4	20
75	68	2	22	92
80	12	4	6	22
83.3	14	2	4	20
85.7	10	2	4	16
87.5	84	0	14	98
100	186	0	0	186
Total	420	348	60	828

Table S3: Self-reported zygosity of same-sex twins by similarity score

Similarity Score	Self-Reported Zygosity			
	MZ	Disagree	DZ	Total
0	8	4	222	234
33.3	4	0	34	38
50	10	2	30	42
60	10	0	22	32
66.7	12	2	12	26
71.4	16	0	4	20
75	58	2	32	92
80	10	2	10	22
83.3	12	0	8	20
85.7	10	0	6	16
87.5	78	2	18	98
100	140	6	40	186
Total	368	20	438	826

Table S4 offers the same measures as Table 2 in the main text, but based on zygosity assigned by Add Health rather than perceived zygosity. Samples sizes are smaller for mis-assigned than misperceived twins (18 and 12 vs. 74 and 16 as shown in Table 1), but results are generally similar. Exceptions (differences of more than 0.10) are highlighted, but probably reflect the small number of mis-assigned twins.

Table S4: Intraclass Correlation and Estimated Heritability by Add Health-Assigned Zygosity Category

	MZ Correct	MZ Incorrect	DZ Correct	DZ Incorrect	h^2 All Correct	h^2 DZ Correct & Perc DZ-Gen MZ	C Shared Env Correct	E Unique Env Correct	C Shared Env MZ Incorrect	E Unique Env MZ Incorrect
BMI	0.84	0.57	0.36	.	0.96	0.42	-0.12	0.16	0.15	0.43
Height	0.96	0.93	0.71	0.47	0.5	0.44	0.46	0.04	0.49	0.07
ADHD	0.39	0.49	0.23	0.09	0.32	0.52	0.07	0.61	-0.03	0.51
Depression	0.31	0.48	0.19	.	0.24	0.58	0.07	0.69	-0.1	0.52
GPA	0.84	0.18	0.63	0.44	0.42	-0.9	0.42	0.16	1.08	0.82

Highlighted values differ from those in Table 2 (using perceived zygosity) by 0.10 or more.