Behav Genet (2013) 43:415–426
DOI 10.1007/s10519-013-9602-1

ORIGINAL RESEARCH

Heritability and the Equal Environments Assumption: Evidence from Multiple Samples of Misclassified Twins

Dalton Conley · Emily Rauscher · Christopher Dawes · Patrik K. E. Magnusson · Mark L. Siegal

Received: 1 August 2012 / Accepted: 16 July 2013 / Published online: 1 August 2013

Abstract Classically derived estimates of heritability from twin models have been plagued by the possibility of genetic-environmental covariance. 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 represent poor attempts to gauge a complex underlying phenomenon of GE-covariance. The present study exploits a natural experiment to address this issue: Self-misperception of twin zygosity in the National Longitudinal Survey of Adolescent Health (Add Health). Such twins were reared under one "environmental regime of similarity" while genetically belonging to another group, reversing the typical GE-covariance and allowing bounded estimates of heritability for a range of outcomes. In addition, we examine twins who were initially misclassified by survey assignment—a stricter standard—in three datasets: Add Health, the Minnesota Twin Family Study and the Child and Adolescent Twin Study in Sweden. Results are similar across approaches and datasets and largely support the validity of the equal environments assumption.

Keywords Equal environments · Twin misclassification · Heritability · ACE model

Edited by Chandra Reynolds.

Electronic supplementary material The online version of this article (doi:10.1007/s10519-013-9602-1) contains supplementary material, which is available to authorized users.

D. Conley (✉)
Department of Sociology, New York University & NBER, 6 Washington Square North Room 20, New York, NY 10003, USA
e-mail: conley@nyu.edu

E. Rauscher
Department of Sociology, University of Kansas, Fraser Hall Room 716, 1415 Jayhawk Blvd, Lawrence, KS 66045, USA

C. Dawes
Wilf Family Department of Politics, New York University, 269 Mercer Street 3rd Floor, New York, NY 10003, USA

P. K. E. Magnusson
Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden

M. L. Siegal
Department of Biology, Center for Genomics and Systems Biology, New York University, 12 Waverly Place, New York, NY 10003, USA

Introduction

Research has claimed to measure the heritabilities of a wide variety of traits and behaviors, from height (Visscher et al. 2006) to autism (Liu et al. 2010) and even food preferences (Breen et al. 2006). Many estimates, however, are based on twin pair analysis and therefore reliant on strong assumptions about the relative environmental similarity of identical (monozygotic, MZ) and fraternal (dizygotic, DZ) twins (the equal environments assumption that identical and fraternal twins experience the same degree of environmental difference and/or influence each other's outcomes to the same extent) (see, e.g., Plomin et al. 2001). If society treats identical twins more similarly than fraternal twins, for example, the resulting unequal twin environments could cause traditional twin analyses to overestimate heritability.

Although there are other approaches to estimating heritability in humans, twin comparisons are by far the most

common approach and taken to be the least problematic because, 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 (h^2) is calculated as two times the difference between the intra-class correlations of identical and fraternal twins. Narrow-sense heritability 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[1] at the parental level—fraternal twins may share more than 50 % 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 Sham 2002 for an empirical example). Perhaps most importantly, the "equal environments" assumption (EEA) has been relaxed. The naïve calculation mentioned above is based on the EEA. That is, it assumes that the covariance between environment and genetics is zero. Put another way, the simple estimation of heritability requires the rather heroic assumption that identical twins experience the same degree of similarity in environment (including reciprocal effects on each other) as do (same sex) fraternal twins.[2]

The 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 et al. 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.

While alternative heritability approaches are emerging, such as those that use sibling identity by descent (IBD) to estimate phenotypic similarity (Visscher et al. 2006) or those that use genetic covariance among non-related individuals (<2.5 % genetic relatedness) (Davies et al. 2011; Yang et al. 2010), questions about the EEA—and the viability of traditional twin-based heritability estimates more broadly—remain. In the present study, we offer an approach to deal with GE covariance that relies on traditional twin methods. Exploiting variability in whether or not the twins accurately perceived their own zygosity, we putatively reverse the direction of social environmental similarity (and confounding) that is typically present in twin studies relying on the EEA. In other words, we take advantage of twins who believe they are fraternal when they are in fact identical (or vice versa) and assess whether the degree of twin similarity differs from twins who accurately perceive their zygosity. Thus, we are able to replicate the standard ACE model—the workhorse of behavioral genetics—and interrogate a key assumption of the paradigm.

If heritability estimates and twin similarity are similar regardless of perceived zygosity, results would support the EEA and lend credence to traditional twin ACE heritability strategies. If, on the other hand, results indicate a strong relationship between perceived zygosity and heritability—with lower heritability based on genetic zygosity—it would cast doubt on traditional twin heritability strategies and suggest that many traits might be more socially malleable than previous research based on such strategies would suggest.

Misclassified twin research

We are not the first researchers to pursue this "misclassification strategy" to interrogate heritability estimates. Goodman and Stevenson (1989) used this methodology to disentangle genetic and environmental effects in a sample of 13-year-old British twins and estimated that hyperactivity and attentiveness are about half heritable. They assigned "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 (Goodman and Stevenson 1989). Xian et al. (2000), Scarr and Carter-Saltzman (1979), and Kendler et al. (1993) found evidence to support the EEA for other behavioral traits based on a variety of twin data. Kendler et al. (1993) examined major depression, generalized anxiety disorder, phobia, bulimia

[1] Assortative mating is the non-random selection of mates in a population. For example, brunettes may be more likely to pair with other brunettes (positive assortative mating) or non-brunettes (negative assortative mating).

[2] Technically, if their genetic similarity in appearance, for instance, is causing the twins to be confused and/or treated more similarly, then that is an effect of genes and thus should unproblematically be part of the overall "genetic" effect (Jencks 1980). However, this logic flies in the face of common sense understandings of what we mean by genetic effects and makes the estimates less externally valid to the rest of the non-twin population. Moreover, bias is introduced by any increased cross-sibling interaction that leads to increased similarity in phenotypes.

and alcoholism using female twins from the Virginia Twin Registry. Xian et al. (2000) examined alcohol and drug dependence, nicotine dependence, major depression, and posttraumatic stress disorder using male twins from the Vietnam Era Twin Registry. Scarr and Carter-Saltzman (1979) examined personality, cognitive and physical development using Philadelphia-area twin adolescents. Although Scarr and Carter-Saltzman (1979) used blood group and Kendler et al. (1993) used DNA data to identify genetic zygosity for pairs of "probable" or "uncertain" status, Xian et al. (2000) relied solely on questions about similarity with no molecular evidence.

Although innovative for the late 1970s, the blood group approach of Scarr and Carter-Saltzman (1979) is problematic because these loci are not definitive or comprehensive enough. For example, in their data, DZ twins differed only at an average of 2.75 blood group loci out of 12. Such high similarity among DZ twins implies that many sets who match at 12 out of 12 may nonetheless be DZ by chance. The approach of Kendler et al. (1993) is the closest to ours. However, they relied 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 pair zygosity as definite, probable or uncertain 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 the original assignment in all cases. For the "probable" group, genetic zygosity matched the original assignment for 83 % of the pairs. To summarize, for final zygosity assignments, Kendler et al. (1993) relied 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 suggest zygosity was assigned with high validity, but some error certainly remained—particularly among pairs in the probable category without genetic data.

Against this backdrop, we are the first to apply the misclassification approach to a recent sample with accurate genetic zygosity information for all twins as well as a wide range of measured behavioral and anthropometric outcomes. We are also the first 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.

We use data from the National Longitudinal Survey of Adolescent Health (Add Health) and analyze both physical and behavioral phenotypes, including: Height; body mass index (BMI); attention deficit hyperactivity disorder (ADHD); depression; cumulative high school grade point average (GPA); and birth weight. Each of these phenotypes has a justification for its inclusion. Height is highly heritable and has been the focus of several new strategies for estimating heritability (Visscher et al. 2006; Yang et al. 2010). The heritability of BMI has garnered more attention in the wake of research about the relationship between social networks and obesity (Christakis and Fowler 2007). Violations of the EEA could explain why high heritability estimates do not match arguments about the social contagion of obesity. Previous research using misclassified twins (Goodman and Stevenson 1989; Xian et al. 2000; Kendler et al. 1993) studied attention deficit and depression—two behavioral phenotypes that are widely available in surveys to allow replication—and yielded evidence to support the EEA. However, our strategy might improve on earlier research and support arguments that ADHD is largely dependent on social environment (e.g., Timimi and Taylor 2004). We also include high school GPA. The high putative influence of social factors on GPA makes it an especially good phenotype to test the EEA. For instance, classic research on teacher perception has shown that grades are strongly dependent on perception and social labeling (e.g., Rist 1977). Teachers may be more likely to either confuse twins (making it difficult to assign different grades) or assess their achievement more similarly if twins perceive themselves as identical rather than fraternal. If any behavior provides evidence against the EEA, high school GPA should.

We compare perceived zygosity to birth weight discordance as a potential instance of phenotypic non-resemblance causing misclassification.[3] Previous research has taken steps to try to address violation of the EEA, but has given less attention to the mechanisms through which this could occur. For example, phenotypic similarity and perceived zygosity could be co-determined over the life course. Perhaps it is the case that twins who deviate greatly on the phenotypes of interest—say height, weight, GPA—are then socially misclassified? This would represent a

[3] Ideally we would instrument misclassification. Birth weight differences temporally precede self-perception of zygosity and strongly predict it, thus fulfilling the first condition necessary for an instrument. 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 et al. 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 2013). 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.

problem for our approach by reversing the causal arrow from phenotype to perceived zygosity. Alternatively, perceived zygosity could be influenced by differences as early as birth. This would be better for our models because such a dynamic would suggest that once a label is applied, it renders (or mitigates) phenotypic similarity or difference. Because newborns have not been subjected to a conscious, social environmental regime of treatment yet, differences in phenotypic distance by misclassification suggest that this is a moment when the causality does indeed go from phenotype to perceived zygosity. So if we do find that the EEA upwardly biases our estimates of heritability for the other phenotypes, the birth weight analysis would serve as an important check (though by no means prove) that causality is going in the direction we posit: from birth weight differences, to perceived zygosity at birth, to phenotypic similarity later in life. In fact, we find a significant relationship between misclassification and twin birth weight differences, which occur before social classification. This does not completely rule out reverse causality, but 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.

Data and methods

To build on previous research, we examine the intra-class correlation[4] for MZ and (same sex) DZ twins who accurately perceive their genetic relatedness and separately for those twin sets who are, in fact, mistaken about their degree of genetic similarity. We calculate heritability estimates (using a standard additive ACE model) as twice the difference between the intra-class correlations of MZ and DZ twins. Again, the ACE model is identified only because we assume away the covariance of A and C. However, in our case, we 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 based on correctly classified twins should overestimate heritability, whereas estimates based on 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 substantive and statistically significant. We achieve this by comparing naïve heritability estimates based on self-reported and survey-assigned zygosity to estimates based on genetically determined zygosity—separately for correctly and incorrectly perceiving twins. We conduct sensitivity analyses using the revised DeFries-Fulker regression technique (Lazzeroni and Ray 2013).

A non-trivial number of same sex twins are, in fact, incorrect about their zygosity. In Japan, for example, one study that assayed four independent samples found that, in each, between a quarter and 30 % of MZ twins were misclassified as DZ twins at birth (Ooki et al. 2004). Likewise, in Norway, a study revealed that a questionnaire approach to classifying the zygosity of adult twins was inaccurate 2.4 % of the time when information from both twins was available and 3.9 % of the time when information from only one twin was obtained (due to the death of or non-response from the other twin) (Magnus et al. 1983). Similarly, a study in Denmark used four questions to assign zygosity and then checked these predictions against genetic test results and found that the overall proportion misclassified was 4 %, with the highest error rate among male MZ twins (8 %) (Christiansen et al. 2003). Finally, a study that genotyped 327 Dutch twin pairs found a parental misclassification rate of 19 %—largely as a result of MZ twins perceived as DZ (Van den Oord et al. 2000). So we can consider the Scandinavian results as lower bounds and the Japanese figure as upper bounds of twin misclassification. In the United States, Add Health is the only national 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 pairs are incorrect (17.7 %). The vast majority of these misperceiving siblings (82.2 %) are MZ twins who thought they were DZ. These zygosity assessments were obtained in the first wave of data collection, when the twins ranged in age from 12 to 18. Thus the 18 % 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 one uses Add Health zygosity assignments, the

[4] Intra-class correlation is the proportion of the variance between pairs, measured as the variance between twin pairs divided by the sum of the variance within pairs and the variance between pairs. ICC = $\sigma_B/(\sigma_B+\sigma_W)$.

misclassification rate falls to a mere 5.9 %. However, a significant additional proportion (6.6 %) 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. 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 cutoff score where the score distribution seemed to divide naturally" (Rowe and Jacobson 1998). If a pair claimed they were DZ, but Add Health would have classified them as MZ based on a high similarity score, they were classified as undetermined. Add Health suggests excluding these pairs or treating them as DZ (Harris et al. 2006).

This discussion illustrates the complexity of attempting to assign zygosity without genetic information. As supplementary Tables S2 and S3 show, there is a great deal of variation in similarity score and any cut point is arbitrary. Furthermore, similarity scores do not always match self-reported zygosity. 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 question the EEA, 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.

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 sample were added, yielding 64 % of sibling pairs from the probability sample and 36 % 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 all twins in the genetic data and are not used. Winship and Radbill (1994) argue against using analytic weights in multivariate analysis.

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). A STR is a stretch of adjacent copies of a DNA sequence; because copy numbers mutate at a high rate and vary considerably within the population, STRs can be used to identify an individual genetically. Twins are classified as genetically MZ if they match at all 11 loci. Our sample includes nearly 150 MZ twin pairs and over 110 same-sex DZ 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 to Add Health assigned zygosity in Panel B. Panel A shows that 74 genetically MZ twins perceive themselves as DZ, whereas 16 genetically DZ twins believe they are MZ. (Supplemental Table S4 further breaks down this split by Add Health classification.) This leaves a small sample of misclassified twins, which is a limitation of this analysis. In an effort to address this limitation, we calculate heritability estimates using a variety of twin samples, including naïve estimates based on twin self-report and Add Health classification, in addition to estimates based on genetic zygosity and misperceived zygosity. We take all of these estimates into account and interpret results from the smaller, misclassified sample in conjunction with others. These steps slightly reduce concern about the smaller number of misclassified twins, but do not solve the problem. A rough power calculation of the difference between correlation coefficients for two groups of 37 and 104 subjects (the number of correctly and incorrectly classified MZ pairs) based on Fisher's Z-test suggests that, if one correlation is 0.7, the other must be at least 0.84 to be statistically different with a one-tailed test (0.85 two-tailed). If one correlation is 0.5, the other must be 0.71 to be statistically different with a one-tailed test. Thus, given a relatively strong ICC, the difference between incorrectly and correctly classified MZ twin correlations must be about 0.15 or 0.2 to yield a significant difference.

To further address concerns about sample size, we also include replication studies from two other surveys: the Sweden Twin Registry and the Minnesota Twin Family Study. These surveys allow replication of analyses based on survey-assigned zygosity, but unlike Add Health, do not include self-perceived zygosity. The Swedish Twin Registry data (Magnusson et al. 2012) we use is based on the CATSS study (Child and Adolescent Twin Study in Sweden), which includes individuals born after 1992. Zygosity among same sex twins was assigned based on questions about physical similarity during childhood: (1) Are your twins like two peas in a pod? (2) How often did people have difficulty distinguishing between your twins? (3) How alike were you and your twin partner during childhood considering eye color? (4) How alike were you and your

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

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

twin partner during childhood considering hair color? The Minnesota Twin Family Study (MTFS) includes same sex twins born since 1971, who were ages 11 and 17 in 1990 when the study began (Iacono et al. 2006). MTFS assigned zygosity based on parental responses to a zygosity questionnaire about twin similarity, staff rating of physical similarity, and an algorithm based on height/weight ratio, head width/length ratio, and fingerprint ridge count (Iacono and McGue 2002).

Phenotypes used in the analysis of Add Health data include the following: height; weight; BMI; depression score; ADHD; delinquency; cumulative high school GPA; and birth weight. Height and weight, used to calculate body mass index, are self-reported in wave 3 of the Add Health survey. Measured height and weight have higher rates of missing values so we use self-reports to maintain as many respondents as possible. Depression is measured using nine items of the Center for Epidemiologic Studies-Depression Scale (CES-D). CES-D normally includes more items that were omitted from wave 3. Therefore we also include the other six questions about the frequency of depressive symptoms in wave 3. The sum of responses for all items (listed in the supplemental section) 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 transcript information in the Add Health data.

Birth weight is reported by parents (in the wave 3 survey), measured in ounces. Birth weight is usually approximately normally distributed with a long tail at the low end, but we rely on twins, who fall at the low end of the distribution. We therefore take the natural log of birth weight. Although this measure is retrospective, when children are teens, parents typically remember birth weight well (e.g., Walton et al. 2000 report an 85 % accurate recall rate when children are teenagers). A limitation of this retrospective measure is that parents could mis-report birth weight based in part on twin zygosity classification. We cannot definitively identify the causal direction, but evidence of an association between the two could inform future research. Of course, other factors could influence the likelihood of misperceived twin zygosity. Potential examples include sex, family history of twinning, or even family socioeconomic status. For example, Christiansen et al. (2003) found a higher zygosity error rate among males whereas misperceived zygosity is somewhat higher among females in our sample. We focus on birth weight because it offers variation *within* the twin pair.

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.

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 MZ twins is stronger than DZ twins, whether the MZ twins correctly perceive their zygosity or not. The similar correlations regardless of perceived zygosity support the EEA. 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 sample-size problem with using genetically DZ twins who believe they are MZ.

Table 2 presents intra-class correlations of phenotypes by classification status for MZ and DZ 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.

The estimated heritabilities of BMI and height are about the same for correctly and incorrectly classified twins.

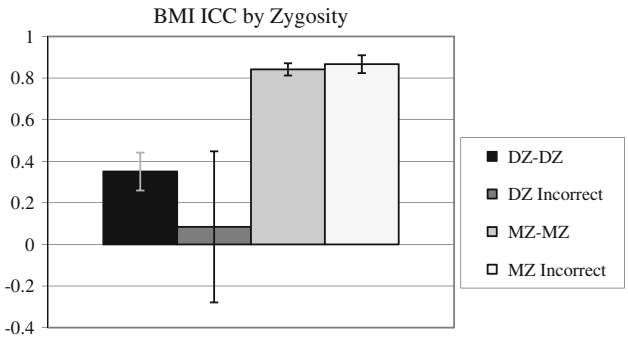

Fig. 1 Twin intraclass correlations for body mass index, by genetic and perceived zygosity; data from genetic subsample of the National Longitudinal Survey of Adolescent Health. Sample sizes are 196 for genetically MZ twins perceived accurately and 66 for MZ twins perceived inaccurately; 186 for same-sex genetically DZ twins perceived accurately and 16 for genetically DZ twins perceived inaccurately. DZ Incorrect indicates genetic DZ twins who perceived themselves as MZ and MZ Incorrect indicates genetic MZ twins who perceived themselves as DZ

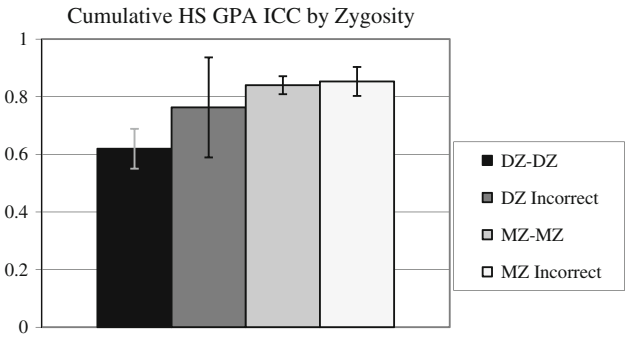

Fig. 2 Twin intraclass correlations for cumulative High School GPA, by genetic and perceived zygosity; data from genetic subsample of the National Longitudinal Survey of Adolescent Health. Sample sizes are 172 for genetically MZ twins perceived accurately and 56 for MZ twins perceived inaccurately; 152 for genetically DZ twins perceived accurately and 12 for genetically DZ twins perceived inaccurately. DZ Incorrect indicates genetic DZ twins who perceived themselves as MZ and MZ Incorrect indicates genetic MZ twins who perceived themselves as DZ

Estimated heritability of BMI 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 symptoms, ADHD symptoms, and GPA show higher heritability among incorrectly classified twins. Estimated heritability is only slightly higher for GPA, but substantially higher for ADHD and depression symptoms among misclassified twins. Oddly, MZ twins who believe they are DZ 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 DZ—however, mean levels of depression symptoms 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.

In every case, as Fig. 3 illustrates, naïve heritability estimates based on perceived zygosity among all twins are lower than those based on genetic zygosity. Twin classification error seems to underestimate heritability for all of these traits. Heritability based on Add Health classification (Table 3) is generally similar to estimates based on twins who accurately perceived their genetic zygosity, but lower than estimates for those who incorrectly perceived zygosity. Heritability estimates are robust to choice of estimation procedure, as the generalized DeFries-Fulker regression method (Lazzeroni and Ray 2013) yields similar results. As Fig. 4 shows, DeFries-Fulker heritability estimates based on perceived zygosity are consistently lower than those based on genetic zygosity. Overall, Add Health results 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 contributions to phenotypes. Similar to the heritability estimates, shared environmental estimates are quite similar using correctly and incorrectly classified MZ correlations, except for symptoms of depression and to a small extent ADHD. Depression and ADHD estimates suggest shared environment is less important among MZ twins who believe they are DZ. This suggests the EEA may be problematic, because shared environment is more important for twins who believe they are MZ. Correctly classified MZ twins may be treated more similarly than genetically MZ twins who believe they are DZ. Shared environment estimates of ADHD and depression symptoms 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. In some cases C appears to be negative. If this were the case, it would suggest that a common environmental regime actually leads to greater phenotypic

Table 2 Intraclass correlation and estimated heritability by self-perceived zygosity category

Phenotype	MZ correct	MZ incorrect	DZ correct	DZ incorrect	h^2 All correct	h^2 DZ correct and MZ incorrect	C shared Env correct	E unique Env correct	C shared Env MZ incorrect	E unique Env MZ incorrect	Naïve h^2 based on perceived zygosity
	1	2	3	4	5	6	7	8	9	10	11
BMI	0.84 (196)	0.87 (66)	0.35 (186) *†	0.08 (16) *†	0.98	1.00	−0.14	0.16	−0.13	0.13	0.67
Height	0.96 (198)	0.95 (68)	0.72 (190) *†	0.49 (16) *†	0.47	0.46	0.49	0.04	0.49	0.05	0.33
ADHD	0.44 (198)	0.51 (70)	0.24 (198) *†	0.44 (14)	0.41	0.54	0.03	0.56	−0.03	0.49	0.30
Depression	0.27 (206)	0.62 (74) *	0.15 (204) †	n/a (16)	0.25	0.94	0.40	0.16	0.38	0.15	0.06
GPA	0.84 (172)	0.85 (56)	0.62 (152) *†	0.76 (12)	0.44	0.47	0.39	0.72	0.60	0.93	0.35
Birth Weight	0.82 (148)	0.41 (44) *	0.72 (144) *†	0.80 (14)	0.21	−0.61	0.61	0.18	1.02	0.59	0.32

In all cases, heritability estimates based on perceived zygosity are lower than estimates based on genetic zygosity. MZ Incorrect indicates genetic MZ twins who perceived themselves as DZ and DZ Incorrect indicates genetic DZ twins who perceived themselves as MZ. Sample sizes are in parentheses

*significantly different from MZ correct, † significantly different from MZ incorrect, n/a indicates a value could not be calculated with the sample and data available

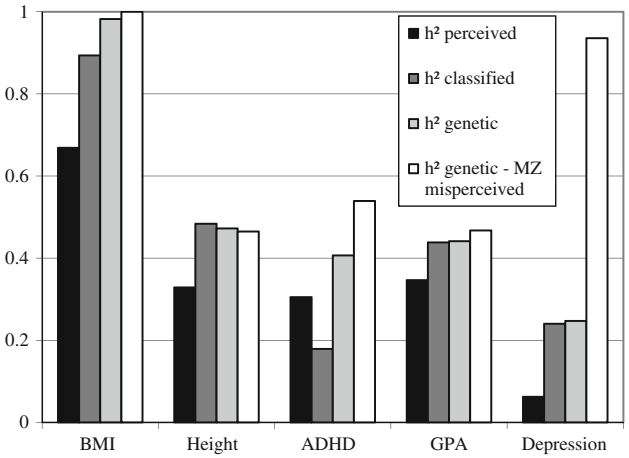

Fig. 3 Narrow-sense (additive) heritability estimates (h^2) by twin zygosity derived from ICC differences: naïve self-perceived, Add Health classified, correctly perceived genetic, and incorrectly perceived genetic zygosity based on figures from Table 2 (columns 5, 6, 11) and Table 3 (column 11)

distance, which is entirely possible in a niche-formation model where common environmental regimes foster divergent developmental responses. That said, the estimates for C in these cases are not statistically significantly different from zero, so it would be premature to suggest any particular dynamic. Readers should note that the ICCs for MZ twins are not more than double those for DZ twins, suggesting there are only additive effects. Thus, the intraclass correlation results suggest that dominance is not a concern in this study. We therefore deploy only additive models.

Results based on the CATSS and the MYFS studies are provided in Table 4. Similar to Add Health results, in nearly every case—with the only exception being birth weight in the CATSS data—heritability estimates based on assigned zygosity are lower than those based on genetic zygosity. In most cases, ICC estimates are lowest for correctly assigned DZ twins, followed by incorrect DZ, incorrect MZ, and correct MZ. This pattern could be consistent with violations of the EEA, although we might expect incorrect DZ to be closer to correct MZ and incorrect MZ to be more similar to correct DZ. Notably, this pattern is not found for birth weight, which suggests a different relationship for this phenotype.

To summarize results so far, heritability estimates based on genetically confirmed twin zygosity are generally higher than estimates based on perceived or survey-assigned zygosity in all three samples. Thus, with the exception of birth weight, heritability based on perceived or assigned zygosity is likely to be substantially underestimated. This result supports the EEA, which would expect heritability based on genetic zygosity to be lower because it accounts for environmental differences.

Table 3 Intraclass correlation and estimated heritability by Add Health-assigned zygosity category

Phenotype	MZ correct	MZ incorrect	DZ correct	DZ incorrect	h^2 all correct	h^2 DZ correct and MZ incorrect	C shared Env correct	E unique Env correct	C shared Env MZ incorrect	E unique Env MZ incorrect	Naïve h^2 based on Add Health-classified zygosity
	1	2	3	4	5	6	7	8	9	10	11
BMI	0.84 (246)	**0.57** (18)	0.36 (196)	n/a (12)	0.96	**0.42**	−0.12	0.16	**0.15**	**0.43**	**0.89**
Height	0.96 (248)	0.93 (18)	0.71 (200)	0.47 (12)	0.5	0.44	0.46	0.04	0.49	0.07	**0.48**
ADHD	0.39 (248)	0.49 (14)	0.23 (208)	**0.09** (12)	0.32	0.52	0.07	0.61	−0.03	0.51	**0.18**
Depression	0.31 (256)	**0.48** (18)	0.19 (216)	n/a (12)	0.24	**0.58**	**0.07**	**0.69**	**−0.1**	**0.52**	**0.24**
GPA	0.84 (212)	**0.18** (16)	0.63 (164)	**0.44** (8)	0.42	**−0.9**	0.42	**0.16**	**1.08**	**0.82**	0.44
Birth Weight	0.75 (184)	n/a (14)	0.74 (152)	0.72 (8)	**0.02**	n/a	**0.73**	0.25	n/a	n/a	**0.22**

The measures are the same as in Table 2, but based on zygosity assigned by Add Health rather than perceived zygosity. Samples sizes are smaller for mis-assigned than misperceived twins (see Table S4), but results are generally similar. Exceptions (differences of more than 0.10) are in boldface, but probably reflect the small number of twins whose Add Health assignment does not match their genetically confirmed zygosity. Sample sizes are in parentheses

Values in bold differ from those in Table 2 (using self-perceived zygosity) by 0.10 or more.

n/a indicates a value could not be calculated with the sample and data available

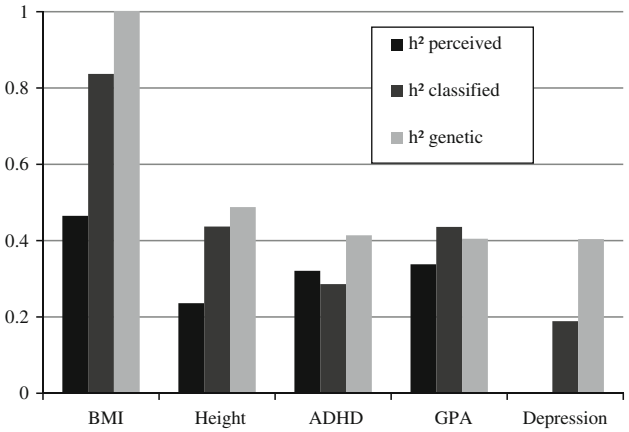

Fig. 4 Narrow-sense (additive) heritability estimates (h^2) by twin zygosity derived from DeFries-Fulker regressions: naïve self-perceived, Add Health classified, and genetic zygosity

Table 5 presents evidence that twin misclassification may be driven at least partially by very early differences. Twins who are genetically MZ, but misperceive themselves as DZ, 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. We infer that their lower similarity likely encouraged their identification as DZ twins. Misclassified DZ twins had slightly higher birth weight similarity than their correctly classified counterparts, but this difference is not significant.

Discussion

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 EEA and lends credence to methods used here and in previous studies that compare similarity based on actual and perceived zygosity to assess the EEA. Further, our results build on previous research to suggest that phenotypic similarity and perceived zygosity are not co-determined. Perceived zygosity appears to be influenced by differences as early as birth. Other factors—such as sex, family history of twinning, or even family socioeconomic status—could of course influence the likelihood of misperceived twin zygosity. However, our evidence suggests that phenotypic distance later in life is not driving misclassification and that our putative causal model is oriented

Table 4 Intraclass correlation and estimated heritability by zygosity category for replication studies

Swedish Twin Registry Data								
Phenotype	MZ correct	DZ correct	MZ incorrect	DZ incorrect	h^2 All correct	h^2 DZ correct and MZ incorrect	h^2 MZ correct and Perc MZ-Gen DZ	Naïve h^2 based on survey-assigned zygosity
BMI	0.870 (1796)	0.537 (1664)	0.693 (82)	0.571 (26)	0.666	0.598	0.312	0.244
Height	0.970 (1828)	0.824 (1704)	0.940 (82)	0.921 (26)	0.292	0.098	0.194	0.038
ADHD	0.663 (1892)	0.183 (1762)	0.530 (88)	0.345 (26)	0.960	0.636	−0.266	0.370
GPA	0.900 (336)	0.659 (270)	0.630 (14)	0.782 (4)	0.482	0.236	−0.058	−0.304
Birth weight	0.790 (1874)	0.748 (1750)	0.750 (88)	0.688 (26)	0.064	0.204	−0.080	0.124

Minnesota Twin Family Study						
Phenotype	MZ correct	DZ correct	MZ incorrect	DZ incorrect	h^2 All correct	h^2 DZ correct and MZ incorrect
BMI	0.801 (1074)	0.429 (564)	0.023 (10)	n/a	0.744	−0.812
Height	0.950 (1076)	0.750 (554)	0.874 (10)	n/a	0.400	0.258
Years of educ.	0.567 (742)	0.460 (376)	0.359 (6)	n/a	0.214	−0.202
Birth weight	0.786 (1034)	0.728 (536)	0.653 (8)	n/a	0.116	−0.075

The correlation and heritability measures are the same as in Table 3, but based on data from the Swedish Twin Registry and the Minnesota Twin Family Study. These additional analyses help address concern about the small number of misclassified twins in the Add Health data. Sample sizes are in parentheses

Table 5 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

Birth weight is measured in log ounces, so differences represent the natural log of the ratio of birth weights within each twin pair. Standard deviations measure dispersion of the birth weight differences of all pairs within each zygosity category. Differences are only significant between twin pairs who correctly and incorrectly identified as MZ twins

* indicates significant difference between groups

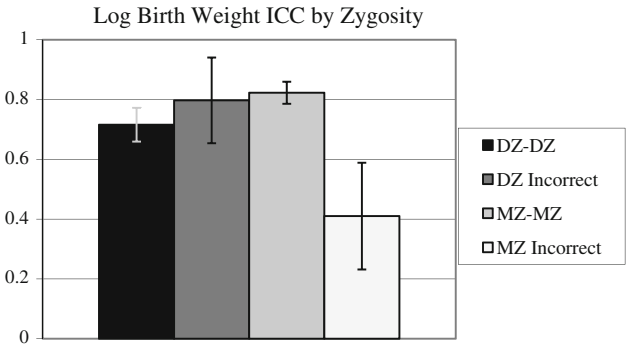

Fig. 5 Twin intraclass correlations for birth weight, by genetic and perceived zygosity; data from genetic subsample of the National Longitudinal Survey of Adolescent Health. Sample sizes are 148 for genetically MZ twins perceived accurately and 44 for MZ twins perceived inaccurately; 144 for genetically DZ twins perceived accurately and 14 for genetically DZ twins perceived inaccurately. DZ Incorrect indicates genetic DZ twins who perceived themselves as MZ and MZ Incorrect indicates genetic MZ twins who perceived themselves as DZ

correctly: Perceived zygosity is influenced by birth weight and this labeling process lingers at least through adolescence—assuming that recall bias of birth weight is random and not further influenced by downstream phenotypic distance. This suggests that, had we found significant upward bias in heritability due to GE covariance, our approach to eliminate that bias would not have suffered from endogeneity (dependence of perceived zygosity on phenotypic similarity). However, because the regime of social treatment (based on classification of zygosity) does not seem to lower heritability estimates, this issue is moot.

In fact, results suggest that heritability estimates may be *higher* if we compare twins who misperceive their zygosity—but mainly for behavioral phenotypes. Specifically, MZ twin perceived zygosity appears to be more important than actual zygosity for depression symptoms, GPA, and ADHD symptoms—which could indicate that perceptions have greater impact than genes on these outcomes. This suggests that even while heritability did not ultimately appear to be upwardly biased, there is still an important role of socialization in determining psychological and developmental outcomes.

Further, the fact that h^2 calculated from genetically confirmed zygosity is higher than that calculated from perceived zygosity may result from early developmental

divergences among misclassified MZ twins. We might also expect misclassified DZ twins to be more similar than correctly classified DZ pairs, perhaps due to greater genetic similarity. (This may be due to the fact that although DZ twins have ~50 % of IBD genes on average, the variance in the distribution is large, and indeed the proportion of DZ twins sharing as high as ~65 % IBD is not negligible.) Thus, these misclassified twins would narrow the gap between DZ twins and MZ twins and thereby result in lower h^2 estimates. Nonetheless, based on our results, we expect this "bias" to be small in magnitude. This finding deserves replication tests and further analysis, but this will require self-perceived zygosity to be recorded in more studies.

A number of approaches—ranging from the misclassification strategy pursued here to using IBD sibling resemblance models—seem to be converging on the conclusion that longstanding narrow-sense heritability estimates are fairly accurate (Visscher et al. 2008). In addition to the EEA, this conclusion also rests on an assumption of random mating. If parents tend to be more alike genetically than they would be if mating were random (a likely case, especially if 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, as has been proposed for example for the major histocompatibility complex where genetic diversity might increase the chances of surviving infectious disease at the individual or population level (Hedrick 1999). However, the phenotypes of interest to most social scientists, including those studied here, are likely to see positive assortative mating [educational assortative mating (see, e.g., Mare 1991)—related to GPA, ADHD, delinquency, and depression—offers the most obvious example]. Overall, therefore, it seems reasonable to take results from an ACE model more or less at face value. In fact, we were surprised by this conclusion, having expected to find h^2 was overstated for our range of phenotypes due to omitted, positive GE covariance. Still, as our results show, the misclassified-twin approach has value in revealing cases where an indicator of socialization—perceived zygosity—is important relative to genetic differences in determining behavioral-trait outcomes.

Acknowledgments 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.

References

Allison DB, Kaprio J, Korkeila M, Koshenvuo M, Neale MC, Hayakawa K (1996) The heritability of body mass index among an international sample of monozygotic twins reared apart. Int J Obes 20:501–506

Breen FM, Plomin R, Wardle J (2006) Heritability of food preferences in young children. Physiol Behav 88:443–447

Christakis NA, Fowler JH (2007) The spread of obesity in a large social network over 32 years. N Engl J Med 357(4):370–379

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 Res 6:275–278

Conley D, Rauscher E (2013) Genetic interactions with prenatal social environment: effects on academic and behavioral outcomes. J Health Soc Behav 54(1):109–127

Conley D, Strully KW, Bennett NG (2003) The starting gate: birth weight and life chances. University of California Press, Berkeley

Davies G, Tenesa A, Payton T, Yang J, Harris SE (2011) Genome-wide association studies establish that human intelligence is highly heritable and polygenic. Mol Psychiatry 16:996–1005

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

Guo G, Stearns E (2002) The social influences on the realization of genetic potential for intellectual development. Soc Forces 80(3):881–910

Harris KM, Halpern CT, Smolen A, Haberstick BC (2006) The national longitudinal study of adolescent health (Add Health) twin data. Twin Res Human Genet 9(6):988–997

Hedrick PW (1999) Balancing selection and MHC. Genetica 104:207–214

Iacono WG, McGue M (2002) Minnesota twin family study. Twin Res 5(05):482–487

Iacono WG, McGue M, Krueger RF (2006) Minnesota center for twin and family research. Twin Res Human Genet 9(6):978–984

Jencks C (1980) Heredity, environment, and public policy reconsidered. Am Sociol Rev 45:723–736

Kendler KS, Neale MC, Kessler RC, Heath AC, Eaves LJ (1993) A test of the equal-environment assumption in twin studies of psychiatric illness. Behav Genet 23(1):21–27

Lazzeroni LC, Ray A (2013) A generalized Defries-Fulker regression framework for the analysis of twin data. Behav Genet 43:85–96

Lichtenstein P, Pedersen NL, McClearn GE (1992) The origins of individual differences in occupational status and educational level: a study of twins reared apart and together. Acta Sociol 35:13–31

Liu K, Zerubavel N, Bearman P (2010) Social demographic change and autism. Demography 47(2):327–343

Magnus P, Berg K, Nance WE (1983) Predicting zygosity in Norwegian twin pairs born 1915–1960. Clin Genet 24:103–112

Magnusson PKE, Almqvist C, Rahman I, Ganna A, Viktorin A, Walum H, Halldner L, Lundström S, Ullén F, Långström N, Larsson H, Nyman A, Gumpert CH, Råstam M, Anckarsäter H, Cnattingius S, Johannesson M, Ingelsson E, Klareskog L, de Faire U, Pedersen NL, Lichtenstein P (2012) The Swedish twin registry: establishment of a biobank and other recent developments. Twin Res Human Genet. doi:10.1017/thg.2012.104

Mare RD (1991) Five decades of educational assortative mating. Am Sociol Rev 56(1):15–32

Ooki S, Yokoyama Y, Asaka A (2004) Zygosity misclassification of twins at birth in Japan. Twin Res 7:228–232

Plomin R, DeFries JC, McClearn GE, McGuffin P (2001) Behavioral genetics, 4th edn. Worth Publishers, New York

Purcell S (2002) Variance components models for gene-environment interaction in twin analysis. Twin Res 5:554–571

Purcell S, Sham P (2002) Variance components models for gene–environment interaction in quantitative trait locus linkage analysis. Twin Res 5:572–576

Rist RC (1977) On understanding the process of schooling: contributions of labeling theory. In: Karabel J, Halsey AH (eds) Power and ideology in education. Oxford University Press, New York, pp 292–305

Rodgers JL, Rowe DC, Buster M (1999) Nature, nurture, and first sexual intercourse in the USA: fitting behavioral genetic models to NLSY kinship data. J Biosoc Sci 31:29–41

Rowe DC, Jacobson KC (1998) National longitudinal study of adolescent health: pairs code book. Chapel Hill, NC, Carolina Population Center

Rowe D, Teachman J (2001) Behavioral genetic research designs and social policy studies. In: Thornton A (ed) America's families and children: research needed in the coming millennium. University of Michigan Press, Ann Arbor, pp 157–187

Scarr S, Carter-Saltzman L (1979) Twin method: defense of a critical assumption. Behav Genet 9(6):527–542

Timimi S, Taylor E (2004) ADHD is best understood as a cultural construct. Br J Psychiatry 184:8–9

Van den Oord E, Boomsma DI, Verhulst FC (2000) A study of genetic and environmental effects on the co-occurrence of problem behaviors in three-year-old twins. J Abnorm Psychol 109:360–372

Visscher PM, Medland SE, Ferreira MAR, Morley KI, Zhu G, Cornes BK, Montgomery GW, Martin. NG (2006) Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. PLoS Genet 2:e41. doi:10.1371/journal.pgen.0020041

Visscher PM, Hill WG, Wray NR (2008) Heritability in the genomics era—concepts and misconceptions. Nat Rev Genet 9:255–266

Walton KA, Murray LJ, Gallagher AM, Cran GW, Savage MJ, Boreham C (2000) Parental recall of birthweight: a good proxy for recorded birthweight? Eur J Epidemiol 16(9):793–796

Winship C, Radbill L (1994) Sampling weights and regression analysis. Sociol Methods Res 23(2):230–257

Xian H, Scherrer JF, Eisen SA, True WR, Heath AC, Goldberg J, Lyons MJ, Tsuang MT (2000) Self-reported zygosity and the equal-environments assumption for psychiatric disorders in Vietnam-era twin registry. Behav Genet 30(4):303–310

Yang J et al (2010) Common SNPs explain a large proportion of the heritability for human height. Nat Genet 42(7):565–569