

Running Head: THE ENVIRONMENT

The Environment in Behavioral Genetics Research  
Gregory Carey  
University of Colorado

## Abstract (120 words)

Contemporary behavioral genetics models the environment as two uncorrelated variables— $C$  (common environment) and  $E$  (unique environment). This paper demonstrates two bizarre assumptions of this model. First, if one randomly selects a specific, concrete environmental variable then that variable must correlate either 0 or 1 between siblings. It is not possible to have a correlation of, say, .38. Second, if one randomly selected concrete environmental variable is part of  $C$  and another is part of  $E$ , then the two environmental variables must correlate exactly 0 within an individual. A simple solution is proposed—substitute a single latent variable for the total environment in place of  $C$  and  $E$ .

### The Environment in Behavioral Genetics Research

Traditional behavioral genetic analysis has three types of latent variables: additive genetic values ( $A$ ), common environmental genetic values ( $C$ ), and unique environmental values ( $E$ ).<sup>1</sup> Neale and Cardon (1992, p. 14) define the unique environment as “all those environmental influences that are so random in their origin, and idiosyncratic in their effects, as to contribute to differences between members of the same family.” They discuss variable  $C$  in the following way (p. 15): “Any environmental factors that are shared by family members will create differences between families and make family members relatively more similar. The environment between families is sometimes called the *shared environment*, the *common environment*, or just the *family environment*.” Similar definitions may be found in introductory texts to behavioral genetics (Carey, 2003; Plomin, DeFries, McClearn, & McGuffin, 2008) as well as more advanced quantitative texts (Neale, Ferreira, Medland & Posthuma, 2008; Sham, 1998).

In path analysis/structural equation modeling (henceforth, simply referred to as “path analysis”), a latent variable is an unmeasured variable hypothesized to be responsible, along with other variables, for the covariances among yet other variables. In principle, however, a latent variable is potentially measurable were one to have the appropriate technology and measurement instruments.

Contemporary technology prevents us from observing additive genetic values for a polygenic trait. One can, however, follow the logic of traditional quantitative genetic models (e.g., Crow & Kimura, 1970; Falconer & Mackay, 1996; Lynch & Walsh, 1998; Mather & Jinks, 1982) and perform a thought experiment that allows us to see the meaning of this genetic value. Let us start with the simple additive genetic model.

Imagine a future technology that can identify all the polymorphic loci that contribute to a phenotype and can genotype a very large sample on those loci. Every individual has two alleles at a locus, so we can use a form of quasi dummy coding to score the alleles. If the first individual has alleles 3 and 5 at the first locus, we assign the values of 1 to variable  $A_{13}$  and to variable  $A_{15}$  where  $A_{ij}$  denotes the  $j$ th allele at the  $i$ th locus. All other  $A_{1j}$ s are coded 0 at the first locus for that individual. If that person is a homozygote for allele 4 at the second locus, the person receives a score of 2 on  $A_{24}$  with 0s for all other  $A_{2j}$ s at that locus.

After scoring the whole sample, regress the phenotypic scores on all of the  $A$ s for all loci and all alleles.<sup>2</sup> The predicted values from this regression would constitute additive genetic values or numerical estimates for latent variable  $A$  in the traditional behavioral genetic model.

The environment lacks the equivalent of genes and alleles as well as a rigorous science of how fundamental environmental units are transmitted. Here, it is assumed that behavioral phenotypes are the result of a large number of individual, specific, and concrete environmental events and variables that may have complicated causal and statistical relationships among themselves but act within the context of linear models. Let  $X_1, X_2, \dots, X_k$  denote these environmental variables for a phenotype.

A thought experiment can then be used to give us environmental values. Imagine a future technology that can catalogue and measure all the  $X$ s on a large population of individuals. We could then regress observed phenotypic values on the vector of all  $X$ s. The predicted value from such a regression is a person’s total environmental score and

the regression coefficients for the  $X$ s are weights indicating the extent to which each  $X$  uniquely contributes to the total environmental value.

According to these “genetic” and “environmental” thought experiments, a model for sib pairs would have four latent variables—the additive genetic values for sibs 1 and 2 (say,  $A_1$  and  $A_2$ ) and the total environmental values for sibs 1 and 2 (say,  $T_1$  and  $T_2$ ). In the traditional behavioral genetic analysis, however, the two latent environmental variables are depicted as three latent variables—the “common” environment ( $C$ ) and two “unique” environments ( $E_1$  and  $E_2$ ), one for each sibling.

If  $C$  and the two  $E$ s act like variables in path analysis, then they too must be linear functions of the  $X$ s. Here, I demonstrate that  $C$  and  $E$  can indeed be expressed as linear functions of the  $X$ s but the assumptions necessary to do so are so highly restrictive that they border on the absurd.

#### *Necessary conditions of the ACE Model*

Figure 1 presents the ACE model used for twin and sibling data.<sup>3</sup> Subscripts 1 and 2 in Figure 1 refer respectively to sib 1 and sib 2, so that  $A_1$  denotes the additive genetic value for sib 1 and  $P_2$ , sib 2’s phenotypic value. The ACE model assumes that the common environmental values for sib 1 are the same as those for sib 2. Hence, instead of two latent variables,  $C_1$  and  $C_2$  with a correlation of 1.0 between them, only a single latent variable,  $C$ , is depicted. The quantity  $\alpha$  on the double-headed arrow connecting  $A_1$  with  $A_2$  is the genetic correlation for the pair. Under simple conditions, this quantity equals 1.0 for identical twins, .50 for fraternal twins and full sibs, .25 for half-siblings, and 0 for adoptive siblings. A key assumption of this model is that all correlations among  $C$ ,  $E_1$ , and  $E_2$  are 0.

[Insert Figure 1 about here]

We now want to develop a model of  $C$  and  $E$  in terms of the  $X$ s. Figure 2 provides a path diagram for the first sib where that sib’s common environment ( $C_1$ ) and unique environment ( $E_1$ ) are expressed as a function of two  $X$ s. The notation  $X_{1i}$  denotes the  $i$ th variable for sib 1. For simplicity, variable  $A_1$  is not depicted.

[Insert Figure 2 about here]

Note that if any  $X$  predicts both the common and unique environment for an individual, then  $C$  and  $E$  will be correlated. That is, in Figure 2, if  $a \neq 0$ , then  $b$  must equal 0. Similarly, if  $c \neq 0$ , then  $d$  must equal 0. We now arrive at the first necessary condition for the ACE model: every concrete environmental variable must be placed into one of two mutually exclusive categories, a  $C$  variable or an  $E$  variable. To permit an  $X$  to influence both latent variables implies one of two consequences. First,  $C$  and  $E$  could be correlated in nature—which, as it turns out, may not be a bad idea. Second, the environment possesses a magical property so that positive and negative paths cancel each other out, leaving  $C$  and  $E$  to be uncorrelated.

Let us apply this by mentally erasing arrows  $b$  and  $c$  in Figure 2. To preserve the orthogonal nature of  $E_1$  and  $C_1$ , it is necessary that  $w = 0$ . This leads to the second necessary condition of the ACE model: every  $X$  that impinges on  $E$  is uncorrelated with every  $X$  that impinges on  $C$ . This condition, however, leads to a quandary. The purpose of an ACE model is to explain the similarities and differences among relatives. Why should relationships among the causal variables of sib pairs *force* concrete environmental factors to be uncorrelated *within an individual*? Nevertheless, let us accept these two conditions and move on.

Figure 3 adds sib 2 to the model. There are two  $X$  variables for each individual, the first for the unique environment (denoted by subscript  $e$ ) and the second for the common environment (subscript  $c$ ). Several potential paths are not shown because they are inconsistent with the ACE model. There can be no path originating in an  $X_e$  for one sib and entering either the  $C$  or the  $E$  for the other sib. Otherwise, the  $E$  for the first sib would be correlated with the  $C$  or  $E$  for the other sib. Similarly, there cannot be a path originating in an  $X_c$  for one sib and entering the  $E$  for the other sib.

Concentrate on the double-headed arrows connecting the  $X$ s of sib 1 with those of sib 2. For  $E_1$  to remain uncorrelated with  $E_2$ , the double-headed arrow between  $X_{1e}$  and  $X_{2e}$  must be 0. This gives the third necessary condition of the ACE model: every  $X$  that defines the unique environment for one sib is uncorrelated with every  $X$  that defines the unique environment of the other sib.

[Insert Figure 3 about here]

We can now mentally erase the double-headed arrow in Figure 3 connecting  $X_{1e}$  with  $X_{2e}$ . Because the  $E$  of one sib cannot be correlated with the  $C$  of the other sib, the double-headed arrows between  $X_{1e}$  and  $X_{2c}$  and between  $X_{1c}$  and  $X_{2e}$  be 0 (necessary condition number four). Again, mentally erase these oaths.

The final necessary condition derives from the requirement that  $C_1$  and  $C_2$  are perfectly correlated and requires some algebra. Note that the  $X_c$ s have been defined in terms of *individuals*, not in terms of *pairs*. Hence, it is not possible to have an  $X_c$  that has arrows going into *both*  $C_1$  and  $C_2$ . (To allow for the possibility of a single variable influencing both sib 1 and sib 2, write the variable twice—once for sib 1 and the second time for sib 2—and let the correlation between the two  $X_c$ s be 1.0.)

Hence, we can write  $C_1 = \mathbf{c}'\mathbf{x}_{1c}$  and  $C_2 = \mathbf{c}'\mathbf{x}_{2c}$ , where  $\mathbf{c}$  is a vector of weights and  $\mathbf{x}_{ic}$  denotes the vector of  $X_c$ s for sib  $i$ . Because of the intraclass relationship, the variances of the two  $C$ s are the same and will equal

$$\text{var}(C_1) = \text{var}(C_2) = \mathbf{c}'\mathbf{W}\mathbf{c}$$

where  $\mathbf{W}$  is the within-individual correlation matrix for the  $X_c$ s. The correlation between  $C_1$  and  $C_2$  may be written as

$$\text{corr}(C_1, C_2) = \mathbf{c}'\mathbf{S}\mathbf{c}$$

where  $\mathbf{S}$  is the correlation matrix between the  $X_c$ s of one sib and the  $X_c$ s of the other sib. In order for this correlation to equal 1.0,

$$\mathbf{c}'\mathbf{S}\mathbf{c} = \mathbf{c}'\mathbf{W}\mathbf{c},$$

or  $\mathbf{S}$  must equal  $\mathbf{W}$ . This reveals the fifth and last condition: the sibling correlation matrix must equal the within-individual correlation matrix. Again, this poses the sticky situation of trying to rationalize why sibling correlations for the  $X_c$ s force the values of within-individual correlations (or *visa versa*). But let us accept this in order to discuss an implication of the necessary condition.

Note that the diagonals of  $\mathbf{W}$ , and hence, the diagonals of  $\mathbf{S}$ , must be 1.0. Thus, each and every concrete environmental variable that defines the common environment *must* have the same value for both sibs. In other words, every  $X_c$  for one sib must correlate 1.0 with its homologue for the other sib.

We have successfully completed the task of writing  $C$ ,  $E_1$ , and  $E_2$  as a linear function of the  $X$ s. Figure 4 summarizes these results. But also note that this effort has resulted in unrealistic assumptions.

[Insert Figure 4 about here]

Select a single random  $X$  from Figure 4. That  $X$  can be either an  $X_c$  or an  $X_e$ . If it is an  $X_c$ , then the correlation between that  $X_c$  for sib 1 and the  $X_c$  for sib 2 must be 1. If the  $X$  is an  $X_e$ , then the correlation between that  $X_e$  for sib 1 and the  $X_e$  for sib 2 must be 0. In short, the ACE model assumes that any concrete, environmental influence on behavior must correlate either 0 or 1 among siblings. It is not possible for a specific environmental variable to have a sibling correlation of .21, .34 or .46.

These assumptions stretch the imagination. It is clear that some environmental variables *may* have the attributes implied by the ACE model. Random measurement error may be legitimately modeled as a “unique environmental variable.” It is also obvious that other environment variables can be treated as “common environmental variables.” Maternal age at birth for twin siblings is an example. But should *all* environmental variables for a phenotype correlate either 0 or 1 across siblings? That is a testable hypothesis but should never be treated as a necessary assumption for behavioral genetic analysis.

*An absurd situation: peer influence*

One particularly absurd situation for the ACE model is peer influence. Let  $X_{c1}$  and  $X_{c2}$  denote the best friend of, respectively, sib 1 and sib 2. If we allow for the possibility that sib 1’s best friend may also influence sib 2, we have the model depicted in Figure 5. For simplicity, all other  $X_c$ s are not shown in the figure.

[Insert Figure 5 about here]

Assume that all variables are standardized. In order that  $\sigma_{C_1}^2 = \sigma_{C_2}^2 = \text{corr}(C_1, C_2) = 1$ , the contribution of  $X_{c1}$  and  $X_{c2}$  to the variances of  $C_1$  and  $C_2$  must equal their contribution to the covariance between  $C_1$  and  $C_2$ . The contribution of  $X_{c1}$  and  $X_{c2}$  to the variance of a  $C$  is

$$(b + \delta)^2 + b^2 + 2\rho b(b + \delta),$$

and the contribution to the covariance is

$$2b(b + \delta) + \rho[b^2 + (b + \delta)^2].$$

In order for the two to be equal

$$(b + \delta)^2 + b^2 + 2\rho b(b + \delta) - 2b(b + \delta) - \rho[b^2 + (b + \delta)^2] = 0$$

which reduces to

$$\delta(1 - \rho) = 0.$$

Hence, the two necessary and sufficient conditions for the ACE model are that  $\delta = 0$  and/or  $\rho = 1$ .

But consider the implications of either of these two conditions. The requirement that  $\delta = 0$  implies that my brother’s best friend influences me as much as he influences my brother. That is not a reasonable assumption. Could it be so? Yes. *Must* it be so? Absolutely not. The mathematical assumption of the ACE model is that it *must* be so.

Similarly, the condition that  $\rho = 1$  implies that your best friend and your sister’s best friend must behave identically. Again, this is unreasonable. One could extend the model from a single best friend to a network of friends, but this only compounds the problem. In each case (second best friend, etc.) one would arrive at a similar conclusion.

*How did this happen?*

How did this curious perspective on the environment arise? I offer the hypothesis that legitimate, statistical orthogonal *variance components* from ANOVA were translated into orthogonal *variables* in path analysis because they gave the correct answer in simple cases. I further hypothesize that this is legitimate when there is interest in abstract statistical quantification. On the other hand, when path analysis is aimed at causal understanding, those “variance component variables” can be misleading.

To understand this, one must recall that current methodology and terminology in behavioral genetics was heavily influenced by the ANOVA-based approach used in the Department of Genetics at the University of Birmingham during the 1960s and 1970s. In psychology, the seminal article was Jinks and Fulker (1970) that compared three contemporaneous methods and advocated the biometrical approach used at Birmingham. In the Jinks and Fulker (1970) notation, there are four variance components for a twin design: within-family genetic variance component ( $G_1$ ), between-family genetic variance component ( $G_2$ ), within-family environmental variance component ( $E_1$ ) and between-family environmental variance component ( $E_2$ ). Within-family components estimate the extent to which siblings *differ* because of environmental ( $E_1$ ) or genetic ( $G_1$ ) reasons. Between-family components estimate the extent to which siblings *are similar* for environmental ( $E_2$ ) or genetic ( $G_2$ ) reasons. These variance components can be estimated from the between-pair and within-pair means squares of the ANOVAs for various types of siblings (identical twins, fraternal twins, ordinary siblings, etc.).

In the transition to path analysis/structural equation modeling from ANOVA, the biometrical variance component  $E_2$  became variable  $C$  and variance component  $E_1$  became variable  $E$ . Curiously, variance component  $G_2$  did not become variable  $G_C$  (the “common” or “shared” genotype) and variance component  $G_1$  never become variable  $G_U$  (the “unique” or “nonshared” genotype).

Consider how  $G_C$  and  $G_U$  would have been viewed had they been interpreted analogously to  $C$  and  $E$  in causal modeling. The definition of  $G_C$  would have been “the effect of all those genetic factors that siblings share and make them similar.”  $G_U$  would have been defined as “the effect of all those genetic factors that siblings do not share and make them different.” This is nonsensical. The same snippets of polymorphic DNA that make siblings similar also make them different. It is just that, on average, one quarter of sib pairs will share both snippets, one half will share one snippet, and the remaining quarter will share no snippet at a locus.

It is helpful to repeat the exercise given above about the environment but this time substituting  $G_C$  and  $G_U$  for, respectively,  $C$  and  $E$ . Instead of  $X_s$ , we would have the quasi dummy-coded  $A_{ijs}$  as the exogenous genotypic variables. To preserve the orthogonality of  $G_C$ ,  $G_{U1}$ , and  $G_{U2}$ , we would come to the analogous conclusions that we arrived at earlier about the environment. Any single “allele” at a locus can influence either  $G_C$  or  $G_U$  but not both. Any allele that influences  $G_C$  cannot be correlated with any other allele (both within a locus and across loci) that influences  $G_U$ . This holds both within an individual and across sibs. Finally, if we select any allele at any locus for sib 1 and another random allele at any locus for sib 2, then the two must correlate either 0 or 1.

Finally, I offer the following challenge to anyone who insists that  $G_C$  and  $G_U$  can be treated as variables. Assume a single locus with two alleles,  $A$  and  $a$ , and with the additive genetic values of  $-1$ ,  $0$ , and  $1$  for, respectively, genotypes  $aa$ ,  $Aa$ , and  $AA$ . Let  $p$

denote the frequency of allele  $A$  and assume Hardy-Weinberg-Castle equilibrium. It is a trivial exercise in quantitative genetics to write down all of the genotypic combinations for sib pairs and their frequencies. Here is the challenge: perform this and then come up with real numbers for  $G_C$ ,  $G_{U1}$ , and  $G_{U2}$  such that the covariance matrix among these variables is diagonal with elements  $p(1-p)$ —half of the genetic variance—on the diagonal.

The problem with  $C$  and  $E$  may also be viewed from the perspective of correlation versus causality. The situation is analogous to students (sibs) nested within schools (families). One can estimate orthogonal variance components for school and for student within school, compute the intraclass correlation, and test for significance. This is a valid and legitimate use of ANOVA, but it is *descriptive* and akin to computing a correlation between two variables and testing for significance. Just like the correlation, the variance component does not necessarily imply causality.

In a causal model, variables akin to the environmental  $X$ s could: (1) influence students within schools (which then influence the school mean); (2) influence the school mean without influencing individual differences within schools; (2) jointly influence school means and individual differences within school. In a *statistical* sense, the variance component for school is independent of the variance component for student within school. In a *causal* sense, however, family background variables could independently influence both school means and individual differences within schools. Hence, from a causal perspective, variables  $C$  and  $E$  may actually be correlated in the real world.

*Does the ACE model lead to errors of inference?*

What is the effect of fitting an ACE model to data? Would we make serious errors of inference? There will be no difficulty in trivial cases—e.g., there is no environmental influence on the phenotypes or the environment does not contribute to the phenotypic correlations among relatives.

For non-trivial cases, we can explore potential errors of inference by comparing the equations of the ACE to those of the model using the  $X$ s. Refer to the latter as the AT model and let it have the following structural equation:

$$P_{ij} = aA_{ij} + tT_{ij}$$

where subscript  $i$  denotes the  $i$ th sib ( $i = 1, 2$ ), subscript  $j$  denotes the  $j$ th phenotype, and  $T$  is the total environment. The structural equation for the ACE model is

$$P_{ij} = aA_{ij} + cC_j + eE_{ij}.$$

Assume that all latent variables are standardized. The equations for the four generic quantities required for a multivariate behavioral genetic analysis are provided in Table 1.

[Insert Table 1 about here]

If there is only one phenotype, then only Equations (1) and (3) in Table 1 apply. Using Equation (3) and omitting subscript  $i$  gives  $c^2 = st^2$ . Substituting this into Equation 1, again omitting subscript  $i$ , gives  $e^2 = (1-s)t^2$ . Hence,  $c^2$  may legitimately be interpreted as the extent to which the environment contributes to the sibling correlation for a phenotype. The quantity  $e^2$  is the total environmental variance less the the environmental contribution to the sibling phenotypic correlation. Thus, there will be no errors of inference when only a single phenotype is analyzed.

When more than one phenotype is measured, Equation (4) requires that  $r_{C_j} c_i c_j = s_{ij} t_i t_j$ . This implies that the quantity  $r_{C_j} c_i c_j$  equals the extent to which the environment

contributes to the correlation between, say, IQ in sib 1 and educational level in sib 2. There is no problem here other than a slight difference in wording.

From Equation (2),  $r_{e_{ij}} e_i e_j = (w_{ij} - s_{ij}) t_i t_j$  so this quantity equals the difference between the extent to which the environment contributes to the cross-trait correlation within an individual and the cross-trait correlation across siblings. Again, this quantity is meaningful even though it requires tortuous phrasing.

On the other hand, the “common environmental correlation” in the ACE model becomes

$$r_{c_{ij}} = \frac{s_{ij}}{\sqrt{s_{ii} s_{jj}}}$$

This expression makes sense in only limited situations—e.g.,  $s_{ii}$  and  $s_{jj}$  are variances or both are reliability estimates.

The unique environmental correlation equals

$$r_{e_{ij}} = \frac{(w_{ij} - s_{ij}) t_i t_j}{e_i e_j} = \frac{(w_{ij} - s_{ij}) t_i t_j}{\sqrt{(a_i^2 + s_i t_i^2)(a_j^2 + s_j t_j^2)}}.$$

It is difficult to make sense of this expression.

The fact that quantities  $r_{c_{ij}} c_i c_j$  and  $r_{e_{ij}} e_i e_j$  make sense fits with the hypothesis given above. Both are covariance components. The first quantity is the between-pair covariance component between phenotypes  $i$  and  $j$ , and the second is the within-pair covariance component. That the quantities  $r_{c_{ij}}$  and  $r_{e_{ij}}$  do *not* have sensible meanings also agrees with the hypothesis. These quantities treat  $C_i$  and  $C_h$ , as well as  $E_i$  and  $E_j$ , as *variables* in path analysis and join the members within each pair with double-headed arrows.

In addition to multivariate models, there are two other areas in which the ACE model has the possibility of leading to incorrect inferences: gene-environment interaction and gene-environment correlation. They are discussed below.

#### *Multivariate factor models*

In multivariate ACE models, it is also customary to speak of “common environmental factors” and “unique environmental factors.” Unless the strict conditions of the ACE models hold, these “factors” can impede research.

To illustrate the problem, I took the data set from Ullman’s (2007) chapter in Tabachnick and Fidell (2007) on the subscales of the Weschler Intelligence Scale for Children (WISC).<sup>4</sup> In an exploratory principal components analysis, both the eigenvalue criterion and the scree test suggested that a three-factor solution was satisfactory. Hence, the first three principal components along with the estimates of the variance unexplained by these three components were taken as the basis for a simulation of 250 adoptive sib-pairs. This number was based on the size of the adoptive families in the Colorado Adoption Project or CAP (Plomin & DeFries, 1984; Plomin, DeFries & Fulker, 1988, 1994). The sibling correlations for the three principal components and the 11 scale-specific residuals were generated from random, uniform distributions on intervals that would give “observed” adoptive sibling correlations in the range of .15 to .25 for the WISC subscales.

One thousand data sets were simulated. For each data set, four AT models were fitted, modeling one, two, three, and four environmental factors. (Naturally, because only adoptive sib-pairs were simulated, no genetic effects were modeled). Sixteen ACE models were fitted, allowing from one to four common environmental factors and one to four unique environmental factors. Table 2 provides a summary of the best fitting models using the Akaike information criterion (AIC).

[Insert Table 2 about here]

In the AT model, there was never a case in which a one-factor or two-factor solution was favored. The correct three-factor solution was chosen in slightly less than three-fourths of the cases.

Like the AT model, the ACE models never gave a situation in which a one or two-factor unique environmental was selected. Almost half the time, a three-factor unique environment and a one-factor common environment solution was preferred. In thirty percent of the cases, a three-factor unique environment and two-factor common environment finding was accepted.

To begin the process of uncovering the “true” state of affairs, the ACE model would have first had to select a model giving the *same* number of common and unique factors and then followed up with an analysis demonstrating that the factor pattern matrices were similar for the two types of environment. But the situation necessary to perform this latter step occurred in only 6.2% of the simulations, a percent almost identical to the conventional alpha level in psychological research. Hence, most researchers would incorrectly conclude that there are a different number of common and unique environmental.

These results clearly demonstrate that the ACE model *may* give incorrect inferences in multivariate behavioral genetic analysis. Hence, this problem may not be an academic curiosity.

*Phenotypic moderation (gene-environment interaction)*

There is another area in which the ACE model may impede research—gene-environment interaction or, in a more appropriate terminology, phenotypic moderation (see Purcell, 2002; Rathouz, Van Hulle, Rodgers, Waldman & Lahey, 2008). Typically, these models test whether genetic and environmental influences in a phenotype (say, IQ) systematically vary as a function of a moderating phenotype (parental education). Three interaction terms are modeled—(1) the moderator and *A*, (2) the moderator and *C*, and (3) the moderator and *E*. When the assumptions of the ACE model are not met, then there can be only two potential moderating effects—(1) between the moderator and *A* and (2) between the moderator and *T*, the total environment. What is the effect of fitting an ACE moderation model when the assumptions are not met?

This is an area that deserves systematic exploration, but here I demonstrate that ACE models may reduce the power for detecting effects involving *A*. I simulated data from the following model

$$P = A + T + \beta_A XA + \beta_T XT$$

where *P*, *A*, and *T* are defined as before, *X* is an “observed” moderator for an individual, and the two  $\beta$ s are regression coefficients. Heritability for *X* and for *P* (in the absence of moderation) was set at .50, the genetic correlation between *X* and *P* (again, in the absence of moderation) was set to .50, and the following environmental correlations were used: (1)  $\text{corr}(X_i, T_i) = .35$ ; (2)  $\text{corr}(X_1, X_2) = .25$ ; (3)  $\text{corr}(T_1, T_2) = .25$ ; (4) and  $\text{corr}(X_1, T_2) =$

.20. To simplify the case, I let  $\beta_A = \beta_T = \beta$  and varied  $\beta$  from .02 to .20 by .02. For each value of  $\beta$ , 1,000 replicate simulations were generated for 500 identical and 500 fraternal twin pairs.

The power for detecting a moderating effect for the environment was similar for the AT and the ACE model (see panel A of Figure 6), although the AT model was consistently and significantly more powerful. The power for detecting moderation with  $A$ , however, was very different between the two (Panel B) and always favored the AT model. These results reinforce the case that at least in some circumstances, the ACE model might lead to incorrect conclusions about phenotypic moderation and gene-environment interaction.

[Insert Figure 6 about here]

#### *Gene-environment correlation*

A final area that may prove troublesome comprises those models positing a gene-environment correlation with variables  $A$  and  $C$  but not one with  $A$  and  $E$ . Recall the two thought experiments at the beginning of this paper that “defined values” for a latent additive genetic variable and a latent total environmental variable. One could compute the correlation between these two variables and arrive at an estimate of gene-(total) environment correlation. But how does this imply that  $A$  will be correlated with  $C$  but not with  $E$ ? Here, I merely point out gene-environmental correlation as an area ripe for future study.

#### *A proposed solution*

Is there a solution to the ACE model’s problems? Yes. Furthermore, the proposed solution is very simple—do not use  $C$  and  $E$  as variables in a path model and substitute a symbol like  $T$  to denote the total environment. The alternative is to catalogue all those cases in which the ACE model does not give misleading answers and all those that lead to substantive errors of inference. This alternative, however, could lead to a large and potentially confusing set of rules to distinguish misleading from other cases. Please note carefully—I do *not* propose the abandonment of variance components, estimation and hypothesis-testing of variance component parameters, nor theoretical explorations in behavioral genetics using variance components. I only suggest that “variables”  $C$  and  $E$  be eliminated in situations where they do not belong—causal linear models.

The proposed solution, albeit radical, has the decided advantage of simplicity. The model would have a variable for total environment and then allow the total environments of various relatives to be correlated. There is, however, a minor notational problem. The best letter to denote the total environment is  $E$ , but its use has been usurped by the ACE model to denote the unique environment. In ACE terminology, an AE model is one in which the environment does not influence the correlation between sibling phenotypes. Retaining  $E$  but giving it a different meaning will sow some perplexity in the short run, but probably avoid long-term confusion. Hence, that is the proposed notation.

This model for sib pairs is illustrated in Figure 7. Here, the quantity  $s$  is used to denote the correlation between the total environments of two siblings. The actual value of  $s$  may be allowed to vary as a function of the type of sibling. For example, it may differ for full sibs, fraternal and identical twins. Note that the model in Figure 7 is not novel. It was used by Loehlin & Nichols (1976) for twins and by Carey & Rice (1983),

Eaves and colleagues (see Eaves, Eysenck & Martin, 1989) and Vogler & Fulker (1983) for more complex pedigrees and other relationships.

[Insert Figure 7 about here]

The model has also been extended to the analysis of multiple phenotypes (e.g., Rice, Carey, Fulker & DeFries, 1989).

### *Discussion*

The purpose of this article was to explicate the hidden assumptions of the traditional ACE model and propose a simple solution that does not require those stringent and unrealistic casual conditions. There is a very large empirical literature in behavioral genetics using the ACE model and its multivariate generalizations. Must it all be reassessed? Some of it may indeed need re-evaluation but certainly not all of it. Some phenotypes (e.g., many adult personality traits) do not exhibit environmental correlations for twins and siblings (Bouchard & Loehlin, 2001; Eaves et al., 1989; Loehlin & Nichols, 1976), so the ACE results will be valid. Hence, reassessment will depend on the phenotype. Also, as demonstrated above, analyses of a single phenotype require only the reinterpretation of parameters. The parameter values and statistical significance remain unchanged.

The most likely area for reassessment is that literature using multivariate models reporting significant differences between “common environment” and “unique environment” factor structure. Included in this body of work would be developmental studies of such phenotypes as cognitive abilities or antisocial behavior in which the correlations between sibling environments change over time. Only an empirical comparison between the results of ACE models and the proposed models will tell if different substantive conclusions are warranted.

Perhaps the major reason for abandoning *C* and *E* is that they have operated as taken-for-granted, set-piece structures that have prevented us behavioral geneticists from thinking deeply about the environment. We behavioral geneticists have followed the lead of other behavioral scientists in focusing on large, global aspects of the environment with the eye of placing them into *C* or *E*. Consider parental education. In terms of causal processes in the real world of physics and chemistry, we have treated this variable as if its causal mechanism were akin to simple “exposure” such as the acquisition of sunburn or radiation poisoning.

Instead, we might think of parental education as a composite label describing individual differences in parent-child interactions over a very large number of small, specific, and concrete events (i.e., the *Xs*). The real causal factors are the specific interactions such as a parent’s response to third grade Suzie’s question, “What is a molecule?” Well-educated parents may give a more appropriate and understandable response than less educated parents. Furthermore, if Suzie—for whatever reasons—is more interested in school than her sister Sylvia, she might ask more such questions of her parents resulting in increased academic performance relative to Sylvia. Here, the cumulative effect of such *Xs* may create differences between Suzie and Sylvia. Parental education may be just as much part of “the unique environment” as it was thought to have been something associated with “the common environment.”

Viewing parental behavior in terms of *Xs* may lead to a different view than that proposed by Rowe (1994) and Harris (1995, 1998). Because of small empirical estimates of the variance of *C*, these researchers argued that parents have little causal effect on their

children's behavior. Their conjecture may be true, but it is not a necessary conclusion from the data. Small estimates of shared environmental variance imply that whatever parents are doing, the net result does not produce great *similarities* in their children's behavior. Perhaps some parental Xs induce similarities whereas other Xs produce differences, what has been termed idiosyncratic parental effects (Carey, 2003, pp. 404-405).

Finally, serious reappraisal of the environment might explain one of the most enigmatic results in behavioral genetics and redirect efforts at measuring the environment. For many adult phenotypes, the overall effect of all the environmental Xs does not produce similarities among twins or siblings. If we were to measure even a tiny subset of these Xs, they should result in identical twin correlations indistinguishable from 0. The enigma is the following—where are these variables? Attempts to measure aspects of “the idiosyncratic environment” have proved unsuccessful (Pike, Manke, Reiss & Plomin, 2000; Reiss, Neiderhiser, Hetherington & Plomin, 2000). Can you think of *any* variable—putatively environmental or not—on which identical twins are uncorrelated?

Perhaps we require “micromeasures” of environment to detect these variables. For example, we may have to spend considerable time within the Smith household cataloguing all the times and circumstances under which Suzie and Sylvia ask for help with their homework as well as the intensity, depth, and clarity of all the parental responses. We may also have to monitor Suzie's and Sylvia's grades more often than intervals of a year or two. Whatever the case, the best control for such a study would come about if Suzie and Sylvia were identical twins.

References

- Bartels, M., Rietveld, M. J. H., Van Baal, G. C. M., & Boomsma, D. I. (2002). Genetic and environmental influences on the development of intelligence. *Behavior Genetics, 32*(4), 237-249.
- Bouchard, T. J., Jr., & Loehlin, J. C. (2001). Genes, evolution, and personality. *Behavior Genetics, 31*(3), 243-273.
- Carey, G., & J., R. (1983). Genetics and personality temperament: Simplicity or complexity? *Behavior Genetics, 13*, 43-63.
- Carey, G. (. (2003). *Human genetics for the social sciences*. Thousand Oaks, CA: Sage Publications Ltd.
- Crow, J. F., & Kimura, M. (1970). *An introduction to population genetics theory*. New York: Harper & Row.
- Eaves, L. J., Eysenck, H. J., & Martin, N. G. (1989). *Genes, culture and personality: An empirical approach*. San Diego CA: Academic Press.
- Falconer, D. S., & Mackay, T. F. C. (1996). *Introduction to population genetics* (4th ed.). Longman: Essex, England.
- Harris, J. R. (1998). *The nurture assumption: Why children turn out the way they do*. New York: Touchstone.

- Harris, J. R. (1995). Where is the child's environment? A group socialization theory of development. *Psychological Review*, *102*(3), 458-489.
- Jinks, J. L., & Fulker, D. W. (1970). Comparison of the biometrical genetical, MAVA, and classical approaches to the analysis of the human behavior. *Psychological Bulletin*, *73*(5), 311-349.
- Loehlin, J. C., & Nichols, R. C. (1976). *Heredity, environment, and personality*. Austin TX: University of Texas Press.
- Lynch, M., & Walsh, B. (1998). *Genetics and analysis of quantitative traits*. Sunderland, MA: Sinauer.
- Neale, B. M., Ferreira, M. A. R., Medland, S. E., & Posthuma, D. (Eds.). (2008). *Statistical genetics: Gene4 mapping through linkage and association*. London: Taylor & Francis.
- Neale, M. C., & Cardon, L. R. (1992). *Methodology for genetic studies of twins and families (NATO ASI series D: Behavioural and social sciences-vol. 67)*. Dordrecht, The Netherlands, 1992: Kluwer Academic Publishers B.V.
- Pike, A., Manke, B., Reiss, D., & Plomin, R. (2000). A genetic analysis of differential experiences of adolescent siblings across three years. *Social Development*, *9*(1), 96-114.
- Plomin, R., DeFries, J. C., McClearn, G. E., & McGuffin, P. (2008). *Behavioral genetics* (5th ed.). New York: Worth Publishers.

- Plomin, R., DeFries, J. C., & Fulker, D. W. (Eds.). (1994). *Nature and nurture during middle childhood*. Oxford, UK: Blackwell.
- Plomin, R., & DeFries, J. C. (1985). *Origins of individual differences in infancy: The colorado adoption project*. Orlando, FL: Academic Press.
- Plomin, R., DeFries, J. C., & Fulker, D. W. (1988). *Nature and nurture during infancy and early childhood*. New York: Cambridge University Press.
- Purcell, S. (2002). Variance components models for gene-environment interaction in twin analysis. *Twin Research*, 5(6), 554-571.
- Rathouz, P. J., Van Hulle, C. A., Rodgers, J. L., Waldman, I. D., & Lahey, B. B. (2008). Specification, testing, and interpretation of gene-by-measured-environment interaction models in the presence of gene-environment correlation. *Behavior Genetics*, 38(3), 301-315.
- Reiss, D., Neiderhiser, J. M., Hetherington, E. M., & Plomin, R. (2000). *The relationship code: Deciphering genetic and social influences on adolescent development*. Cambridge, MA, US: Harvard University Press.
- Rowe, D. C. (1994). *The limits of family influence: Genes, experience, and behavior*. New York: Guilford Press.
- Sham, P. (1998). *Statistics in human genetics*. London: A Hodder Arnold.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics*. New York: Pearson.

Ullman, J. B. (2007). Structural equation modeling. In B. G. Tabachnick, & L. S. Fidell (Eds.), *Using multivariate statistics* (5th ed., pp. 676-780). New York: Pearson.

Vogler, G. P., & Fulker, D. W. (1983). Familial resemblance for educational attainment. *Behavior Genetics*, *13*(4), 341-354.

Table 1. Comparison of the equations for the ACE and the AT models.

| Equation | Quantity                     | ACE Model  | AT Model                                     |
|----------|------------------------------|--|--|
| 1        | $\text{var}(P_i)$            | $a_i^2 + c_i^2 + e_i^2$  | $a_i^2 + t_i^2$                              |
| 2        | $\text{cov}(P_{1i}, P_{1j})$ | $r_{A_{ij}} a_i a_j + r_{C_{ij}} c_i c_j + r_{E_{ij}} e_i e_j$ | $r_{A_{ij}} a_i a_j + w_{ij} t_i t_j$        |
| 3        | $\text{cov}(P_{1i}, P_{2i})$ | $\alpha a_i^2 + c_i^2$   | $\alpha a_i^2 + s_{ii} t_i^2$                |
| 4        | $\text{cov}(P_{1i}, P_{2j})$ | $\alpha r_{A_{ij}} a_i a_j + r_{C_{ij}} c_i c_j$               | $\alpha r_{A_{ij}} a_i a_j + s_{ij} t_i t_j$ |

Table 2. Percent of best fitting models for the number of environmental factors for the AT and the ACE models.

| AT Model:                       |      | ACE Model:                             |   |      |     |     |  |
|---------------------------------|------|--|---|------|-----|-----|--|
| Number of Environmental Factors |      | Number of Unique Environmental Factors | Number of Common Environmental Factors: |      |     |     |  |
|                                 |      |  | 1                                       | 2    | 3   | 4   |  |
| 3                               | 72.8 | 3                                      | 46.3                                    | 30.6 | 6.0 | 0.4 |  |
| 4                               | 27.2 | 4                                      | 9.1                                     | 5.8  | 1.6 | 0.2 |  |

## Figure Captions:

Figure 1. The traditional behavioral genetic model for sibling resemblance.  $A$  = additive genetic value,  $C$  = common environmental value,  $E$  = unique environmental value,  $P$  = phenotypic value, subscripts 1 and 2 denote respectively sibling 1 and sibling 2.

Figure 2. Model for individual sibling 1 expressing the common and unique environment in terms of two concrete environmental variables,  $X_{11}$  and  $X_{12}$ .

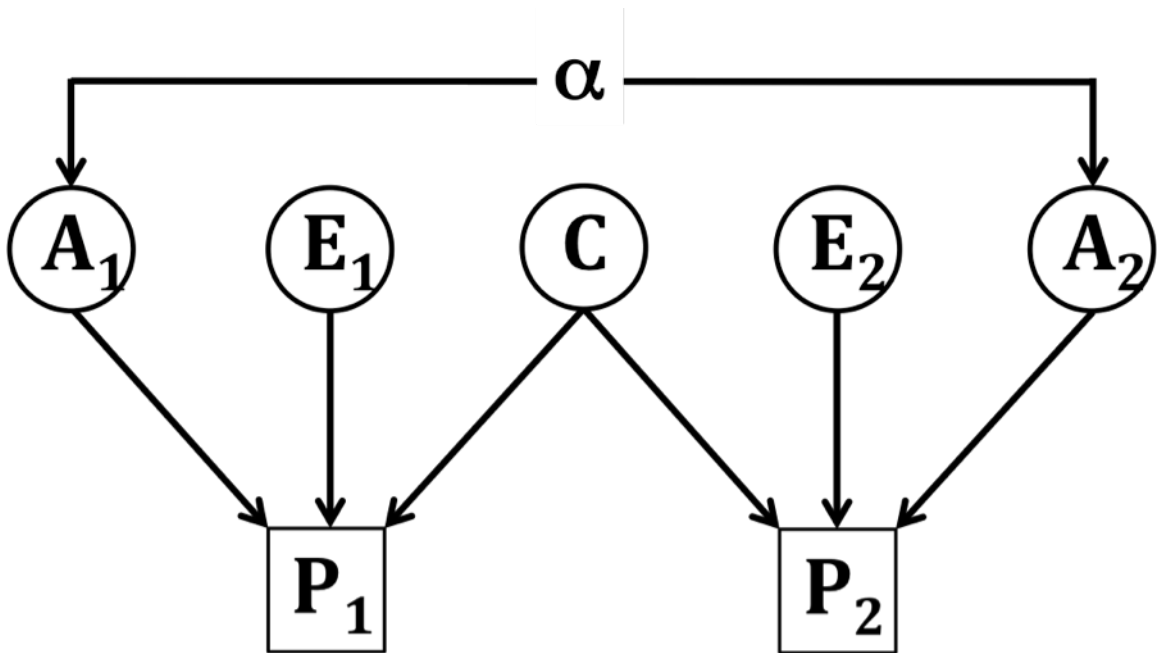
Figure 3. A model expressing the common environment and the two unique environments for siblings using the necessary conditions already derived (see text).

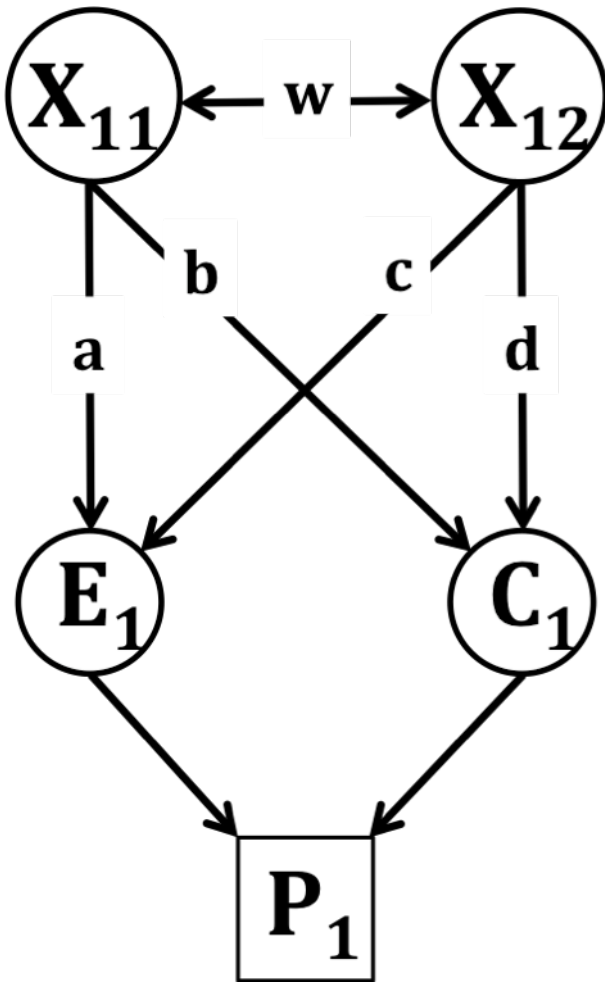
Figure 4. Schematic model illustrating the necessary conditions of the ACE model.

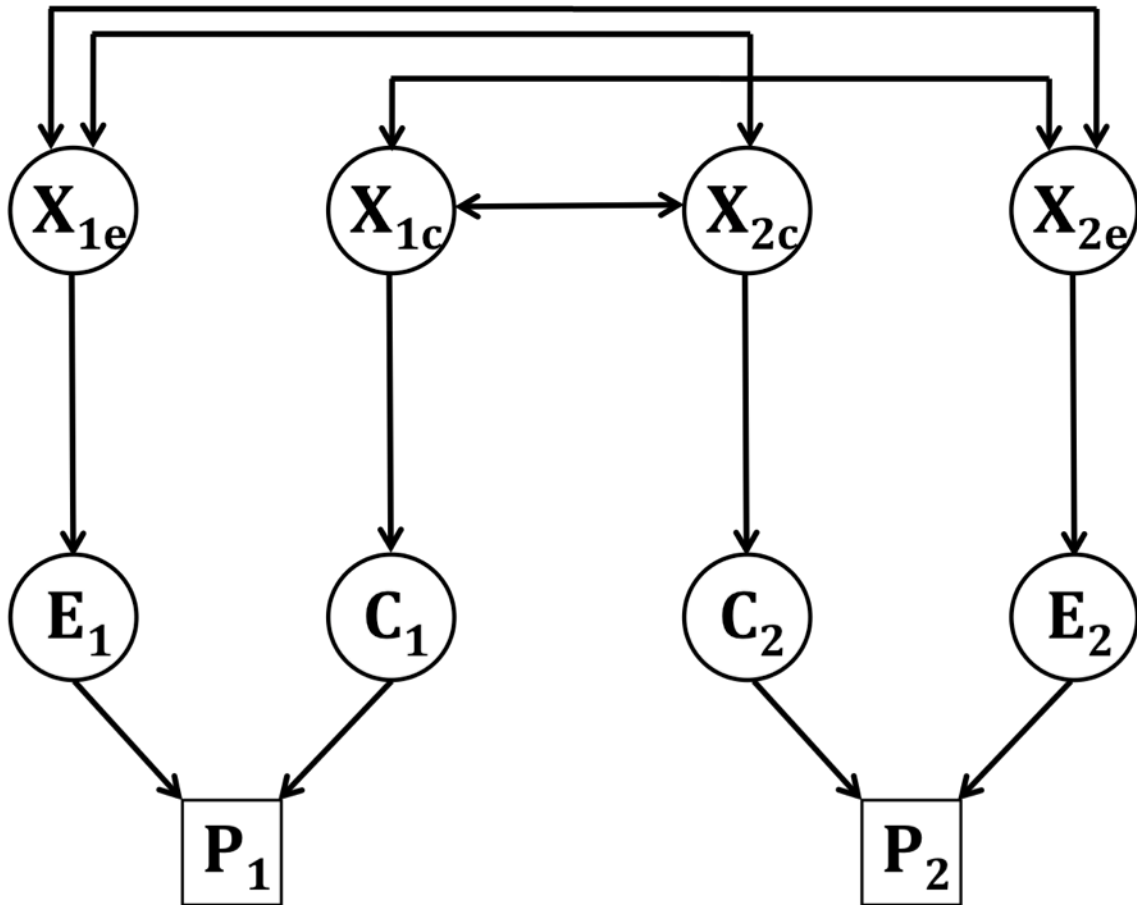
Figure 5. A model for peer influence.

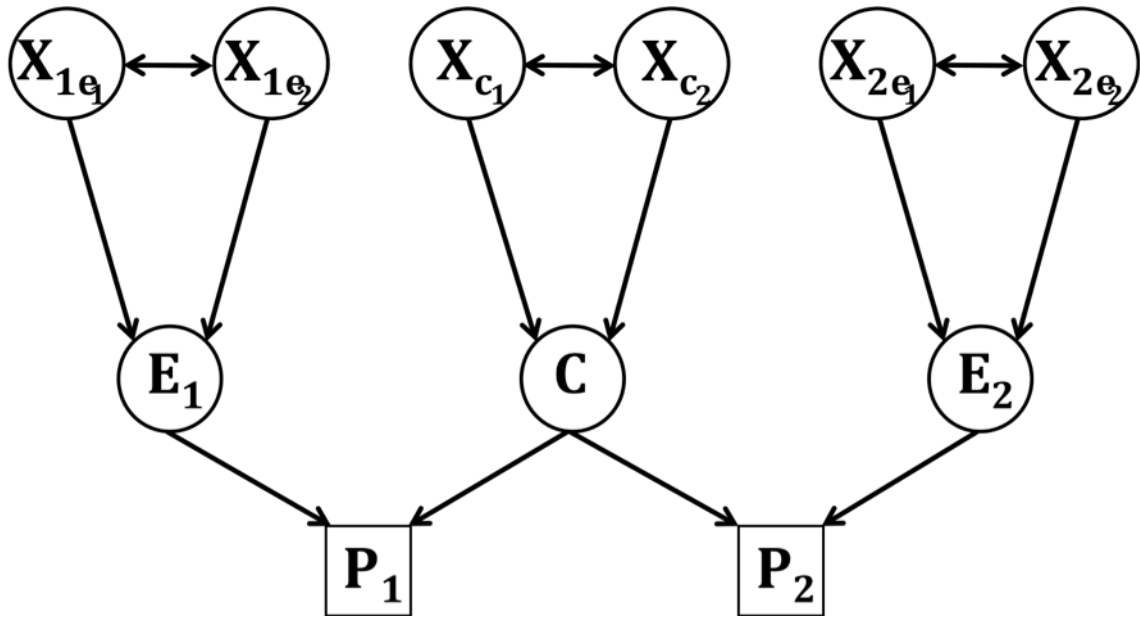
Figure 6. Power for detecting phenotypic moderation involving the environment (panel A) and the genotype (panel B) for the ACE and the AT models.

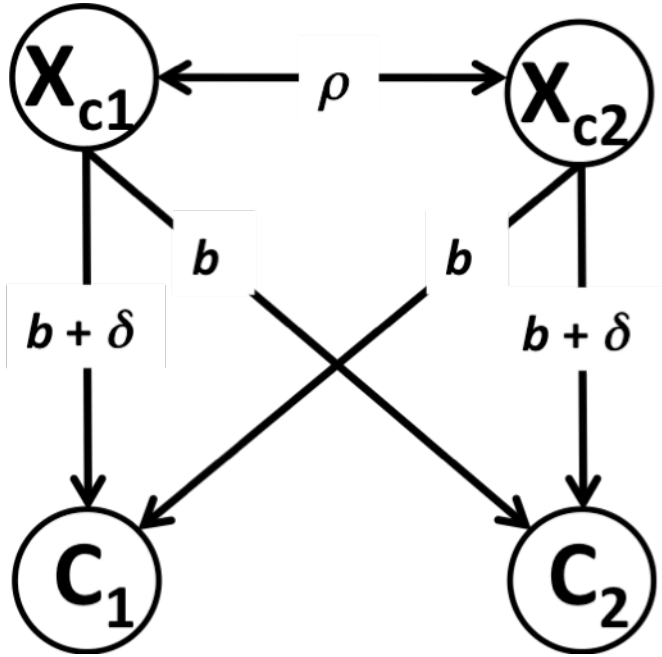
Figure 7. Model for sibling resemblance that avoids the restrictive and possibly misleading assumptions of the ACE model.

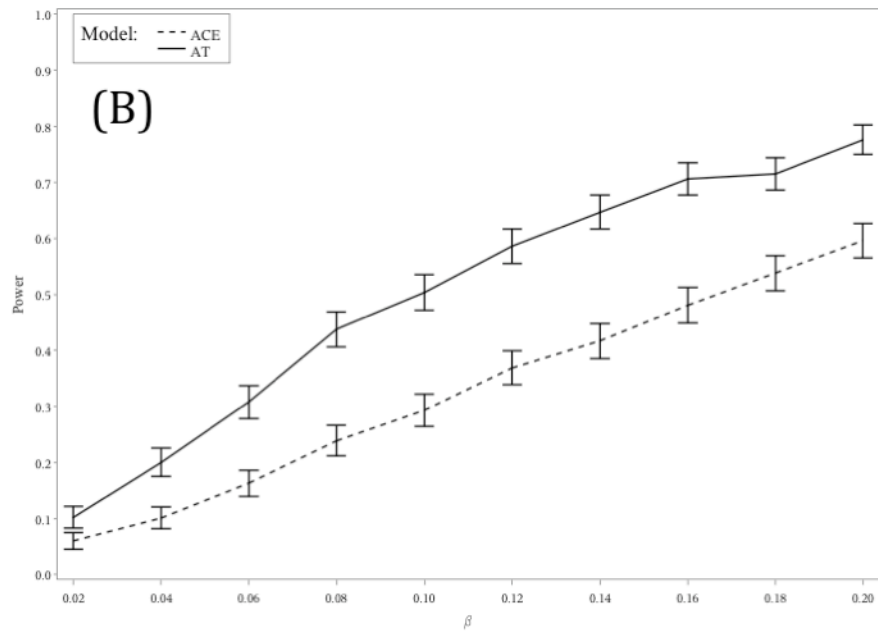
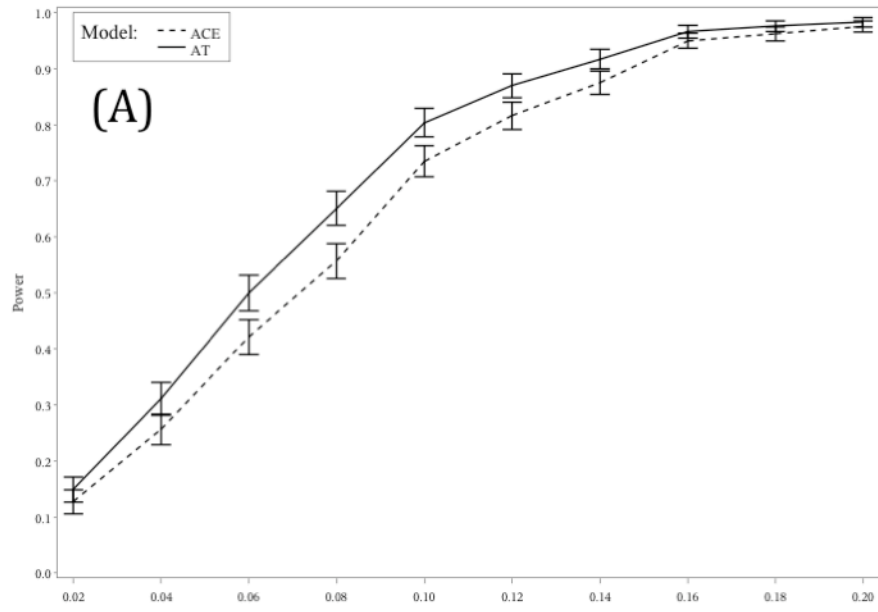


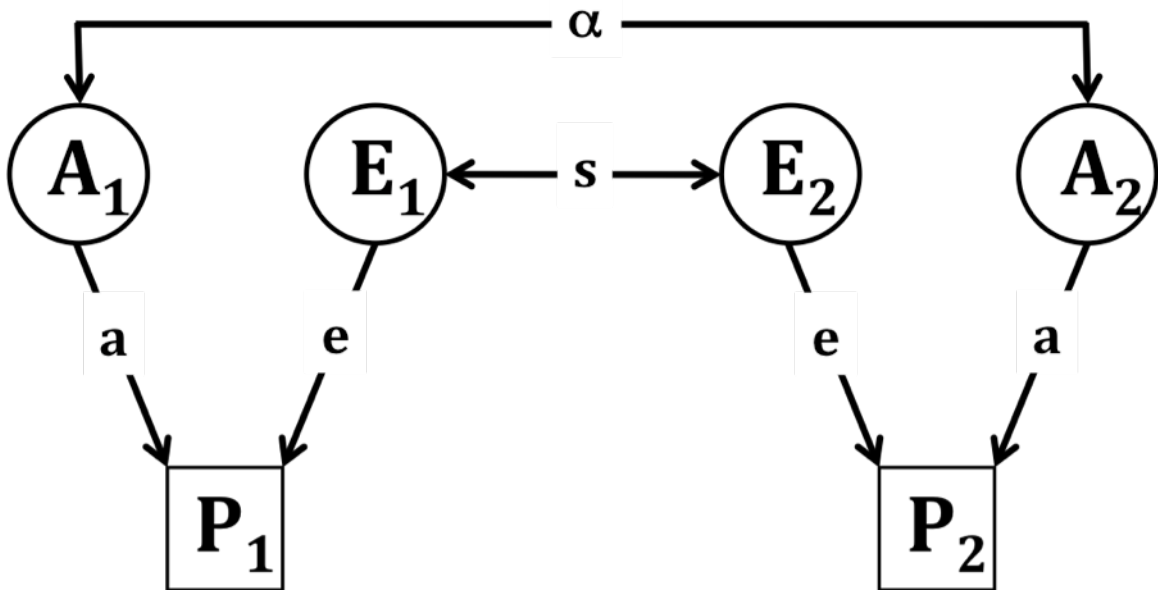












ACKNOWLEDGEMENTS:

Supported in part by grant RR 025780 from NCRR. I thank John Loehlin, Nick Martin, Mike Neale, and Lindon Eaves for helpful comments on an earlier draft.

FOOTNOTES:

---

<sup>1</sup> In some cases a fourth latent variable ( $D$ ) is used for dominance genetic variance.

<sup>2</sup> It is assumed that one  $A$  from each locus is dropped to avoid a singular matrix.

<sup>3</sup> From here on, I speak of sibling pairs and assume that the sibs are in an intraclass relationship, i.e., the assignment of the sibs to “sib 1” and “sib 2” is arbitrary. The arguments can be developed for parent-offspring and other relationships, but that adds an unnecessary layer of complexity to the exposition.

<sup>4</sup> The data set is the WISCSEM.DAT data set and can be downloaded from <http://www.ablongman.com/tabachnick/stats/data.html>.