Published in final edited form as:

Behav Genet. 2013 January ; 43(1): 71–84. doi:10.1007/s10519-012-9568-4.

Operating characteristics of alternative statistical methods for detecting gene-by-measured environment interaction in the presence of gene-environment correlation in twin and sibling studies

Carol A. Van Hulle^{1,4}, Benjamin B. Lahey², and Paul J. Rathouz³

¹Waisman Center, University of Wisconsin-Madison

²University of Chicago

³University of Wisconsin School of Medicine and Public Health

Abstract

It is likely that all complex behaviors and diseases result from interactions between genetic vulnerabilities and environmental factors. Accurately identifying such gene-environment interactions is of critical importance for genetic research on health and behavior. In a previous article we proposed a set of models for testing alternative relationships between a phenotype (P) and a putative moderator (M) in twin studies. These include the traditional bivariate Cholesky model, an extension of that model that allows for interactions between M and the underlying influences on P, and a model in which M has a non-linear main effect on P. Here we use simulations to evaluate the type I error rates, power, and performance of the Bayesian Information Criterion under a variety of data generating mechanisms and samples sizes ($n=2000$ and $n=500$ twin pairs). In testing the extension of the Cholesky model, false positive rates consistently fell short of the nominal Type I error rates ($\alpha=.10, .05, .01$). With adequate sample size ($n=2000$ pairs), the correct model had the lowest BIC value in nearly all simulated datasets. With lower sample sizes, models specifying non-linear main effects were more difficult to distinguish from models containing interaction effects. In addition, we provide an illustration of our approach by examining possible interactions between birthweight and the genetic and environmental influences on child and adolescent anxiety using previously collected data. We found a significant interaction between birthweight and the genetic and environmental influences on anxiety. However, the interaction was accounted for by non-linear main effects of birthweight on anxiety, verifying that interaction effects need to be tested against alternative models.

Keywords

gene-environment correlation; gene-environment interaction; gene-environment moderation; simulation study; twin study

Statistical methodologies for testing and estimating the degree of gene-by-measured environment interaction in quantitative behavior genetic designs have been a major focus for the past decade, beginning with an article by Dick, Rose, Viken, Kaprio & Koskenvuo (2001) and especially popularized by Purcell (2002). There, he proposed an important extension of the classic bivariate biometric model to allow testing interactions between a measured environment and each of the variance components (A, C, or E), while accounting

⁴Corresponding author: Carol Van Hulle, University of Wisconsin-Madison, Madison, WI 53705, vanhulle@waisman.wisc.edu.

for A-, C-, or E-by-measured environment correlations arising from the influence of genes (A) and environmental factors (C and E) common to both the phenotype and the measured environment. Since the publication of Purcell's article, researchers have relied on his model to test gene-by-environment interactions for a wide range of phenotypes, including perceived control and physical health (Johnson & Krueger, 2005), family income and intelligence scores (Turkheimer, Haley, Waldron, D'Onofrio, & Gottesman, 2003), prenatal complications and asthma (van Beijsterveldt & Boomsma, 2008), protein intake and body composition (Silventoinen et al., 2009), marital quality and anxiety (South & Krueger, 2008), and others (Johnson, McCue, & Iacono, 2009; Lau & Eley, 2008). Whereas interactions between candidate moderators and additive genetic influences (A) were usually the focus of these studies, researchers typically tested the potentially important interactions between the candidate moderator and shared (C) and unshared (E) environmental influences as well. Therefore, in this article, we refer to "GxM" in the *generic* sense of interactions between the moderator M and A, C or E, and refer in the same generic sense to correlations between A, C, or E and the measured environment as " r_{GM} ".

Recently, we examined statistical aspects of Purcell's approach and demonstrated that, under some plausible conditions, his model incorrectly identifies GxM when it does not exist (Rathouz, Van Hulle, Rodgers, Waldman, & Lahey, 2008). Because of the central importance of accurately identifying GxM interactions, there is a need to have robust statistical procedures available for testing and quantifying GxM. In particular, such procedures should not only identify GxM in data in which GxM is operating, but should also provide comparisons to plausible and equally parsimonious alternative models that do not include GxM. That is, any procedure should consider a sufficiently wide class of statistical models to allow the data to distinguish between GxM and equally parsimonious non-GxM mechanisms that could lead to the observed joint distribution of phenotypes. We (Rathouz et al., 2008) proposed a broad class of such models and showed how various members of that class, both involving and not involving GxM, could be directly compared via standard statistical procedures. Tests of the proposed new models for identifying GxM have not however been thoroughly evaluated to confirm that Type I error rates are correct with realistic sample sizes and to establish sample sizes needed to adequately power such comparisons. Working within this class of proposed models (Rathouz et al., 2008), the *four aims* of the current study are to assess, under a variety of data generating mechanisms and sample sizes, (a) whether likelihood ratio tests have correct Type I error rates for comparing nested alternative models, (b) the power of likelihood ratio tests for comparing nested models, (c) the ability of the Bayesian information criterion (BIC) to choose the better of two models (whether nested or not), and (d) the ability of BIC to choose the correct among several alternative models.

In the next section, we review both the central model proposed by Purcell (2002) and a subset of the alternative models that we proposed in Rathouz et al. (2008). Following that, we describe the design and results of a large simulation study that addresses the four aims outlined above. Then, we illustrate the utility of including our alternative models in an analysis of gene-by-environment moderation in the presence of gene-environment correlation using data from the Tennessee Twin Study (Lahey, Waldman, Loft, Hankin, & Rick, 2004).

Models for Testing Gene-by-Moderator Interaction

In this section, we present models proposed both by Purcell (2002) and by Rathouz et al. (2008) that are extensions of the classical bivariate twin model. We use "M" to refer to the putative moderator variable, and "P" to refer to the phenotype of interest. We acknowledge that M may not always be strictly environmental in nature, but nevertheless may play a

moderating role in influencing the phenotype P. Because each alternative model reflects different underlying biology, distinguishing between them is essential to understanding the biological processes relating M to P, and the way in which genetic and environmental factors jointly influence M and P.

The classical ACE twin model, shown here for the putative moderator, is given by

$$M = \mu_M + a_M A_M + c_M C_M + e_M E_M, \quad (1)$$

where A_M , C_M , and E_M are standard normal latent variables, uncorrelated with one another, and μ_M is the mean of M (Neale & Cardon, 1992). The more familiar variance components specification of the model is derived from (1). In general, A refers to additive genetic influences, C to shared environmental influences that reflect similarity among relatives, and E to non-shared environmental influences that are unique to each individual. Designs including both mono- and dizygotic twins render the model identifiable.

One common extension of (1) is the bivariate Cholesky model (Neale & Cardon, 1992). That model augments (1) with a model for P, viz.,

$$P = \mu_P + a_C A_M + c_C C_M + e_C E_M + a_U A_U + c_U C_U + e_U E_U. \quad (2)$$

Model (2), together with (1), allows for genetic (a_C), shared (c_C), and non-shared (e_C) environmental influences *common* to M and P, denoted here by the subscript “c”. Corresponding influences *unique* to P, denoted by the subscript “U”, are given by a_U , c_U , and e_U in (2). The model described by Purcell (2002), shown in Figure 1a, allows for both common genetic and common environmental influences as in (2), and possible interactions between the moderator and genetic (AxM) and/or environmental (CxM or ExM) influences on P, viz.,

$$P = \mu_P + (a_C + \alpha_C M) A_M + (c_C + \kappa_C M) C_M + (e_C + \varepsilon_C M) E_M + (a_U + \alpha_U M) A_U + (c_U + \kappa_U M) C_U + (e_U + \varepsilon_U M) E_U. \quad (3)$$

In (3), interactions between the moderator and various genetic and environmental influences on P are captured by coefficients α_C , κ_C , ε_C , α_U , κ_U , ε_U . For example, the presence of AxM (additive genetic-by-measured environment interaction), jointly captured by α_C and α_U , can be tested by the hypothesis that $\alpha_C = \alpha_U = 0$. Assuming (without loss of generality) that $\alpha_U > 0$, then when $\alpha_C > 0$, the unique genetic influences on P are stronger for larger values of M. If $\alpha_C < 0$, then when $\alpha_C > 0$, the genes that influence both M and P have a stronger influence on P at larger values of M. The degree to which there are genetic influences on P that also impact M, i.e. additive genetic-by-measured environment correlation (r_{AM}), is captured jointly by the parameters α_C and α_U , and can be tested via the null hypothesis that $\alpha_C = \alpha_U = 0$.

Note that (3) captures the effect of M on P indirectly through α_C , κ_C , and ε_C . However, as demonstrated by Rathouz et al. (2008), by imposing some constraints on (3) we can re-express the model equation in terms of direct effects of M on P. If it is true that $\alpha_C / a_M = e_C / e_M = c_C / c_M = \beta_1$ and $\alpha_U / a_M = \varepsilon_C / e_M = \kappa_C / c_M = \beta_2$, then we can rewrite (3) as

$$P = \mu_P + \beta_1 M + \beta_2 M^2 + (a_U + \alpha_U M) A_U + (c_U + \kappa_U M) C_U + (e_U + \varepsilon_U M) E_U. \quad (4)$$

Here the factors common to M and P operate directly through M to influence P (shown in Figure 1b). When $\alpha_U = \kappa_U = \varepsilon_U = 0$, which we refer to as model(4*), then there is no true

interaction of A_U , C_U or E_U with M . Whereas models (4) and (4*) are subsets of model (3), they have a qualitatively different interpretation about the biological processes giving rise to P . However, if (3) is tested without considering (4) or (4*) then nonlinear main effects β_I and β_2 may be detected as non-zero interaction effects α_C , κ_C or ϵ_C .

Finally, (3) can be expressed as an extension of the correlated factors model (McArdle & Goldsmith, 1990) rather than as an extension of the Cholesky model. The Cholesky model depends on the ordering of the variables and is therefore most appropriate when clear temporal or causal reasons for ordering M and P exist. For cases where the ordering of M and P is arbitrary, a correlated factors model is more appropriate. We (Rathouz et al., 2008) extended the correlated factors model for $G \times M$ by setting

$$P = \mu_p + (a_p + \alpha_p M)A_p + (c_p + \kappa_p M)C_p + (e_p + \epsilon_p M)E_p; \\ r_a = \text{corr}(A_M, A_p), r_c = \text{corr}(C_M, C_p) \text{ and } r_e = \text{corr}(E_M, E_p). \quad (5)$$

We showed that (5) is a special case of (3) when $\alpha_C/\alpha_C = \alpha_U/\alpha_U \equiv \gamma_A$, $\kappa_C/\kappa_C = \kappa_U/\kappa_U \equiv \gamma_C$, and $\epsilon_C/\epsilon_C = \epsilon_U/\epsilon_U \equiv \gamma_E$. Model (3) can be recovered from (5) by setting

$$a_p = \sqrt{a_C^2 + a_U^2}, c_p = \sqrt{c_C^2 + c_U^2}, e_p = \sqrt{e_C^2 + e_U^2}, r_a = a_C / \sqrt{a_C^2 + a_U^2}, r_c = c_C / \sqrt{c_C^2 + c_U^2}, r_e = e_C / \sqrt{e_C^2 + e_U^2}, \alpha_p = \gamma_A / \sqrt{a_C^2 + a_U^2}, \kappa_p = \gamma_C / \sqrt{c_C^2 + c_U^2} \\ \text{and } \epsilon_p = \gamma_E / \sqrt{e_C^2 + e_U^2}. \text{ Model (5) allows for interaction effects with } M \text{ with three fewer parameters than model (3), but does not permit decomposition of } G \times M \text{ into common and unique parts as in (3).}$$

Rathouz et al. (2008) proposed other models as well. However, only the subset of models reviewed here can be estimated and tested using existing software, in particular the popular Mplus modeling environment (Muthén & Muthén, 1998). This is because the other proposed models involve quadratic or multiplicative functions of latent variables. We will address these other models in future studies.

Simulation Design

Model specification and data generation

To address aims (a)–(d) of this study, simulated data were generated (i) for M under model (1) and (ii) for P under models (2), (3), (4), (4*), (4†) described below, and (5). We simulated data using Stata 12.1 (StataCorp, 2011) under multiple specifications for each model, described in Table 1, for a total of 15 different data generating mechanisms (DGM). DGM's varied by the strength of correlation between the latent quantities and the moderator (high r_{AM} and low r_{EM} or low r_{AM} and high r_{EM}), and by the strength of interaction with the moderator (high $A \times M$ and low $E \times M$ or low $A \times M$ and high $E \times M$). For the non-linear effects models, we simulated data with a large or small quadratic term. We also simulated data under a model that included the linear effect of M on P but dropped the non-linear effect of M on P (4†). All values are shown in Table 1. For each of the 15 scenarios listed in Table 1, we simulated sample sizes of $n=2000$ pairs (1000 each of MZ and DZ pairs) and of $n=500$ pairs. All simulations were performed with 2000 replicates.

For all DGM's the moderator (M) had mean of zero and variance of one; the variance components of M were $a_M^2 = .45$, $c_M^2 = .1$ and $e_M^2 = .45$, which reflect values common in behavior genetics. Likewise, for all specifications, the phenotype (P) had a mean of zero and variance of two *in the absence of interactions with M* (the moments will change when

interactions are present). For DGM's (2) through (4†), $a_U = \sqrt{0.9 - a_C^2}$, $e_U = \sqrt{0.9 - e_C^2}$, and $c_U = \sqrt{0.2}$; for DGM (5), $a_p = e_p = \sqrt{0.9}$, and $c_p = \sqrt{0.1}$. For DGM (2) and DGM (3), high r_{AM}

(r_{EM}) was set at 0.5 and low r_{AM} (r_{EM}) was set at 0.1. For DGM 3, we simulated interactions between M and the genetic effects common to M and P (a_C) and interactions between M and genetic effects unique to P (a_U). We included analogous ExM (e_C , e_U) and CxM (κ_C , κ_U) interactions in all simulations. High AxM was defined as a_C or a_U being one half of the main effect of common (a_C) or unique (a_U) genetic influences on P. Doing so ensured that the effect of A_M or A_U on P would be absent when M is two standard deviations below its mean. Low AxM was defined as one quarter of the main effect of common or unique genetic influences. In this condition, the effect of A_M or A_U on P would be reduced by half when M is two standard deviations below its mean. An analogous definition was used to specify high and low ExM. Similarly, for DGMs (4) and (4*), the quadratic main effect of M on P (β_2), was set such that the effect of M on P was absent (large) or reduced by half (small) when M is two standard deviations below its mean. All values are collected in Table 1. We chose to focus on genetic and non-shared environmental correlations and interactions between M and P for simplification. Therefore, where applicable, the shared environment influences, c_C , κ_C , κ_U , r_C , and, κ_P , were set to small positive values (see Table 1 Note) and were not considered further.

Data analysis

For each DGM, the six models listed in the first column of Table 1 were fitted to the data using the popular structural equations modeling software Mplus 5.21 (Muthén & Muthén, 1998)¹. For each model we used three random sets of starting values. For reasons we were not able to determine, fitting model (5) in Mplus was exceptionally computationally intensive. Therefore, we fitted Model (5) using only one set of starting values. We computed a likelihood ratio test (LRT) statistic for all pairs of nested models (see Table 1). For each hypothesis test, when data were generated under the null model, we present the empirical (i.e., simulated) Type I error rates for nominal rates of 0.1, 0.05, and 0.01. When data were generated under the alternative model, the simulations allow an examination of empirical power. These experiments evaluate the ability of the maximum likelihood statistical procedure to detect when non-linear model terms, including GxM, are needed. Hence several models were compared via the LRT as alternatives to the null hypothesis of the bivariate Cholesky model (2). Interactions (AxM and ExM) were additionally tested by comparing the Cholesky with AxM and ExM model (3) to the non-linear main effects with AxM and ExM model (4), and to the non-linear main effects only model (4*). Finally, the non-linear main effects model (4*) was compared to a linear effects only model (i.e. $\beta_2 = 0$), which we refer to as (4†). For power, we used the results from the simulations with $n=2000$ to estimate the chi-square non-centrality parameter of the LRT statistic, and from there, obtained empirical estimates of sample sizes needed for power of 70% and 90% (Saunders, Bishop, & Barrett, 2003); we considered this form to be more useful to the reader than presenting simulated power. Mathematical details are given in an Appendix A.

We also empirically assessed the degree to which BIC was able to differentiate nested or non-nested models when neither model reflected the true DGM. Raftery (1995) showed that a BIC difference of 10 corresponds to a Bayesian odds of 150:1 that the model with the more negative value is the better fitting model and hence that a difference of 10 should be considered “very strong” evidence in favor of the model with the more negative value. We computed the difference in BIC for each pair of models. Models were determined to be equivocal, that is, describe the data equally well, if the BIC difference was between -10 and 10. For each replicate, we determined the best model among all the alternatives according to

¹Stata and Mplus scripts for data generation and model fitting are available from the first author at <http://www.waisman.wisc.edu/twinresearch/researchers/vanhullecv.shtml>

lowest BIC, allowing us to see how often the correct model was chosen and, when it was not, which other models were chosen.

Simulation Study Results

Type I error rate

The number of replicates that exceeded χ^2_{crit} for $\alpha=.10$, $.05$, and $.01$ for each pair of nested models are given in Table 2 for samples sizes of $N=2000$ and $N=500$. The first column lists the alternative model. The second column lists the true model (i.e. model used to generate the simulated data). Models are listed in order of complexity. As an illustration of how to read the table, when data are generated under the Cholesky model (2) with high r_{AM} and low r_{EM} , the number of false positives was slightly lower (3.5%) than the expected 5% when comparing model (3) to model (2). In general, the LRT procedure for Cholesky GxM model (3) is poorly approximated by the chi-square distribution when testing null models (2), (4) and (5).

We now consider the results in more detail. First, when testing various sub-models versus the Cholesky GxM model (3), we found that the chi-square distribution was generally not well-calibrated to the empirical LRT distribution. False positive rates were generally lower than expected, leading to conservative, and hence underpowered, tests. There were some improvements when moving from $n=500$ to $n=2000$, but they were not uniform. For the basic Cholesky DGM (2A-B), another 2000 replicates were simulated under each condition, with sample size increased to 4000 MZ and 4000 DZ pairs. The rate of false positives increased under DGM 2A with low r_{AM} and high r_{EM} , but remained unchanged under DGM 2B with high r_{AM} and low r_{EM} . Similar results were obtained when testing other null models with model (3) as the alternative; these results are described in Appendix B. We did, however, obtain the expected rate of false positives when we compared model (3) specifying AxM effects only (dropping CxM and ExM parameters) to the traditional Cholesky (2); error rates equaled 10.1%, 4.6%, and 1.0% for $\alpha=.10$, $.05$, and $.01$ respectively when $n=2000$ (results not shown in table).

The rate of false positives was closer to expected at $n=2000$ when comparing other sets of nested models. However, the test for GxM in the non-linear main effects model, (4*) versus (4), and that for the Cholesky (2) versus the linear main effects (4 $\dagger$), were still underpowered at the $\alpha=.05$ level at $n=2000$.

Second, tests of the correlated factors GxM model (5) as the null hypothesis performed poorly, without consistent improvement from sample size $n=500$ to $n=2000$. Testing model (5) indirectly by imposing the constraints, noted earlier, on model (3) required to reproduce model (5) reduced the number of replicates that failed to converge considerably, though the failure rate was still quite high. For these reasons (see Computational issues for more details) we deemed the results from fitting model (5) to be untrustworthy and did not consider that model further.

Sample Size

In general, for $n=2000$, across nearly all nested model comparisons, whether they involved Purcell's model or not, when the sub-model was not the true model, it was rejected for most of the replicates (results shown in Appendix C). To cast these results in a form that is more illustrative to the reader, they were used to estimate sample sizes needed for rejecting the null hypothesis with 70% and 90% power (Table 3). To do so, we are assuming that the LRT follows a non-central chi-square distribution. The results from examining Type I error rates suggest that using the chi-square as a reference distribution will likely lead to conservative results. That is, suppose that the investigator used these results for study planning, but when

analyzing data, used simulation to obtain accurate p-values. Then, we expect that the ultimate test will have higher power than that estimated here. Very small samples are necessary when conducting an omnibus test either of all possible moderator-by-variance components interactions (AxM, CxM, and ExM) or of non-linear main effects. Larger, but not infeasible, sample sizes are necessary for rejecting non-linear main effects in favor of interaction effects with the moderator when the latter is the true mechanism of action.

Because researchers tend to be most interested in the ability to detect AxM (versus CxM or ExM), we simulated 2000 replicates with samples sizes of 1000 MZ and 1000 DZ twin pairs as before but now with only AxM effects present, and no ExM or CxM effects. Model (3) with only AxM interactions (dropping CxM and ExM parameters) was then compared to the traditional Cholesky (2). Under DGM's with high AxM, approximately 150–180 participants were needed to ensure adequate power. However, under DGM's with low AxM, samples sizes jumped to 480–540 twin pairs (results not shown in table).

BIC

The results on power were largely supported by results using BIC as the criterion to select better-and best-fitting models. In Table 4, bold-faced font indicates the model corresponding to the DGM; underline font indicates the settings in which GxM is frequently detected when it is not actually present in the simulated data. Overall, for pairwise comparisons, BIC generally was highly likely to select the correct model when one of the two models reflected the true DGM. The noted exceptions to this were some comparisons between the Cholesky GxM model and the non-linear main effects with GxM model. A similar pattern was observed for selecting the best out of the four models; in the few settings in which BIC failed for $n=500$, there was substantial improvement for $n=2000$. Among non-nested models, when the DGM contained interactions (DGMs 3A-D and DGMs 4A-B), models specifying interaction effects had significantly lower (better) BIC values than those without. Under DGMs 3A-D, the non-linear main effects with GxM model (4) was preferred over model (3) with much greater frequency when modeling fewer twin pairs ($n=500$), and model (4) had the lowest BIC across models (2)-(4†) in the majority (>75%) of replicates. Why? Model (4) captures interactions between M and unique influences on M and P as does model (3), but in model (4), interactions between M and influences *common* to M and P are captured by a single parameter, β_2 . These results show that, with smaller samples sizes, it is difficult to detect significant differences among α_C , κ_C , and ε_C . Finally, when GxM does not exist in the simulated data, but there are substantial non-linear effects of M on P (DGMs 4*A-B), GxM is often nonetheless detected if one compares GxM models to linear models without GxM such as the Cholesky (2) or the linear main effects model (4†). This shows the importance of accounting for possibly nonlinear relationships that give the misimpression of GxM when doing a thorough analysis to test for GxM.

Computational issues

As noted, tests of the correlated factors GxM model (5) as the null hypothesis performed poorly. Several issues arose when fitting the data to model (5). When we fitted model (5) directly, the rate of non-convergence was surprisingly high (often >20% of replicates). Therefore, instead we fitted model (5) indirectly by specifying model (3) with the appropriate constraints. This reduced the number of replicates that failed to converge, though the failure rate was still quite high (3–258 of 2000 replicates) in comparison to the fit of other models (0–12 of 2000 replicates). In addition, both directly and indirectly fitting model (5) was computationally intensive in Mplus, necessitating use of only a single set of starting values rather than multiple sets of starting values as was done with models (2)–(4†).

We also encountered difficulty obtaining the best log-likelihood value for model (3) under certain DGM's. Specifically, under DGM's 3B, 3D and 5D (low AxM, high ExM) model (4) had higher, log-likelihoods than model (3) for 59, 25, and 14 replicates respectively. Changing the number of sets of starting values from 3 to 10 when fitting model (3) resulted in higher model (3) log-likelihoods for all of the replicates in question, but did not have any impact on the log-likelihood from model (4). For all replicates, the log-likelihood after refitting model (3) was greater than the log-likelihood from model (4). Therefore, we recommend using around 10 starting values when fitting model (3) to ensure an optimal fit.

Simulation Discussion

We conducted these simulations in order to characterize type I error rates and power for a subset of the model comparisons laid out in Rathouz et al. (2008). We draw three main conclusions from this study. First, Type I error rates are consistently low when comparing nested models; consequently, the alternative model would be erroneously accepted at greater rates than is generally accepted. Second, the correlated factors with GxM model proposed in Rathouz et al. (2008) is very difficult to fit in Mplus. Third, data generated under a non-linear main effects model can lead to incorrect detection of GxM if GxM is not tested against the non-linear main effects model, but adding such tests is straightforward to implement. We discuss these three points in more detail in the Discussion section.

Illustrative Application: Birthweight and Anxiety

Here we illustrate a prototypical analysis, with sample sizes and variables typical of larger studies, examining gene-by-environment interactions. We specifically highlight the ways in which nonlinear effects can be modeled and how doing so informs the overall conclusions one may draw. We examine the relationship between birthweight (M) and child/adolescent anxiety (P). A number of studies report that low birthweight (LBW) infants show elevated rates of anxiety. For example, Asbury, Dunn, and Plomin (2006) found that among 7 year old MZ twins who were discordant for birthweight, the lighter twin was rated by teachers as more anxious than the heavier twin. This finding of a relationship between birthweight and anxiety controls for confounding due to common genetic influences and to shared environment, but it does not tell us whether that relationship is due to (i) confounding by unshared environmental factors impacting both birthweight and anxiety, (ii) to GxM interaction, (iii) to the direct influence of birthweight on anxiety, or (iv) some combination of all three. The following analysis aimed to tease this out.

Sample and Data

The Tennessee Twin Study (TTS) is a representative sample of 6–17-year-old twins born in Tennessee and living in one of the state's five metropolitan statistical areas (MSAs) in 2000–2001 (Lahey et al., 2004). A random sample of identified families was selected stratified on the age of the twins and geographic area. Interviews were completed with 2,063 adult caretakers, with a response rate for caretakers of 70%. When the adult caretaker was interviewed, 98% of the twin pairs were also interviewed. The caretaker classified 71% of the twins as Non-Hispanic white, 24% as African American, 2% as Hispanic, and 3% in other groups. Parents and guardians who agreed to participate gave written informed consent and twins who were old enough to be interviewed (≥ 9 years of age) gave oral assent.

Adult caretakers and youth were interviewed separately using the Child and Adolescent Psychopathology Scale (Lahey et al., 2004) to assess the youth's DSM-IV symptoms of attention-deficit/hyperactivity disorder, oppositional defiant disorder, conduct disorder, major depression, generalized anxiety disorder, separation anxiety disorder, agoraphobia, social phobia, specific phobia, and obsessive-compulsive disorder. We analyzed self-

reported total anxiety symptoms, completed by twins aged 9–17 years (N=1582). Mothers reported child's weight at birth in ounces. To improve accuracy, only reports from biological mothers (90.9%) were used, resulting in N=1429 pairs (541 monozygotic twins, 888 dizygotic twins).

Application Results

We focused on models (2)–(4†), specifying birthweight as the moderator (M) and anxiety as the phenotype of interest (P). After residualizing birthweight and anxiety on gender, ethnicity, and age (anxiety only), and standardizing to unit variance, birthweight and anxiety were modestly, negatively correlated ($r = -.09$, $p < .001$). We then fit model (2) to determine if birthweight and anxiety share any common underlying genetic or environmental influences. Results from this model (BIC=14406) suggested that shared and non-shared environmental factors in particular influence both anxiety and birthweight. An omnibus test of genetic and environmental influences common to anxiety and birthweight that set a_C , c_C , and e_C parameters from model (2) to 0 produced a significant loss in fit ($\chi^2_{\text{diff}} = 17.9$, 3df, $p < .001$) although BIC only dropped modestly (BIC=14402). It is not entirely uncommon for LRT and BIC to “disagree” as BIC imposes a greater penalty on complexity. Because our interest is on a flexible model for the *joint* distribution of birthweight and anxiety in order to provide a basis for exploring GxM, we conclude that the bivariate Cholesky with influences common to birthweight and anxiety is the better fitting model. To examine common effects further, each of common parameters a_C , c_C , and e_C were dropped from the model in turn. Covariation between birthweight and anxiety seemed to be entirely due to common shared environmental influences, as dropping c_C resulted in a significant decrement in fit ($\chi^2_{\text{diff}} = 7.9$, 1df, $p = .005$, BIC=14407), whereas either the common genetic (a_C) or the common non-shared environmental parameters (e_C) could be dropped from model (2) without a loss in fit ($\chi^2_{\text{diff}} = .58$, 1df, $p = .44$, BIC=14399; $\chi^2_{\text{diff}} = 3.5$, 1df, $p = .06$, BIC=14402, respectively). We next fitted model (3) to test for possible interactions between birthweight and the common genetic and environmental influences on anxiety and as well as interactions between birthweight and the unique genetic and environmental influences on anxiety. Model (3) (BIC=14422) fitted significantly better than model (2) ($\Delta\chi^2 = 28.1$, 6df, $p = .001$), but with a substantial increase in BIC. We also found that unshared genetic and environmental influences on anxiety did not vary with birthweight (jointly testing parameters a_U , κ_U , and e_U ; $\Delta\chi^2 = 2.9$, 3df, $p = .40$, BIC=14402). In contrast, dropping the interaction parameters a_C , κ_C , and e_C from model 3 significantly reduced the fit of the model ($\Delta\chi^2 = 16.2$, 3df, $p = .001$, BIC=14415). That is genetic influences on anxiety appear to be stronger, and environmental influences weaker, at lower birthweights, suggesting the presence of GxM (see Table 5).

As noted in the simulation study, it is possible that the apparent interaction between birthweight and common genetic and environmental influences on anxiety noted above is an artifact of a non-linear association of birthweight and anxiety in this sample of twins. We found significant linear and non-linear (quadratic) effects of birthweight on later anxiety ($p = .04$ and $p = .015$ respectively) in OLS regression. Children with either very high or very low birthweights tended to have somewhat higher anxiety than those born at an average weight (see Figure 2). Therefore we proceeded with model (4). We found that (4) (BIC=14399), which includes interactions between birthweight and unique influences on anxiety (a_U , κ_U , and e_U), fit the data as well as the full version of Model (3) ($\Delta\chi^2 = 6.6$, 4df, $p = .16$), indicating that the significant interactions between birthweight and influences common to birthweight and anxiety may be better explained by the non-linear association of birthweight and anxiety. Parameter estimates are shown in Table 5. As before, dropping the interactions between birthweight and *unique* influences on anxiety from model, (4*), resulted in a non-significant loss in fit ($\Delta\chi^2 = 7.2$, 3df, $p = .07$, BIC=14385) compared to model (4). In contrast, dropping the non-linear effect of birthweight on anxiety, model (4†),

resulted in a highly significant loss in fit ($\Delta\chi^2 = 15.8$, 1df, $p < .001$, BIC=14393) compared to model (4*). We conclude, based both on the series of likelihood ratio χ^2 tests between nested models and on BIC values, that the best fitting model is the non-linear main effects association without GxM (4*).

Application Conclusion

Our goal was to illustrate the utility of the proposed set of models by quantifying GxM for a set of variables that are exemplary of the types of phenomena typically studied using behavior genetic approaches. Here we found that, had we stopped at fitting models (2) and (3), we would have concluded that birthweight moderates the common (to both birthweight and anxiety) genetic and environmental influences on anxiety. However, adding model (4) to our set of analyses suggests that in fact the apparent moderation can be explained equally well by a non-linear association of birthweight and anxiety. Although models (3) and (4) fit equally well, model (4) offers a simpler explanation of the data. Given that birthweight and unique genetic and environmental influences on birthweight could be also be dropped from the models without a loss in fit, evidence in favor of any interaction effect between M (birthweight) and the underlying genetic and environmental influences on P (anxiety) is weak at best.

Discussion

Powerful study designs have been developed that move beyond the classical BG twin method for bivariate data to better capture the interplay between genetic and environmental influences on behavior (Medland, Neale, Eaves, & Neale, 2008; Price & Jaffee, 2008; Purcell, 2002; Rathouz et al., 2008). Whereas other designs may be more powerful for distinguishing genetic and environmental interactions (such as twins reared apart or children of twins), twin pairs reared together remain more readily available, easier to ascertain, and therefore the dominant BG approach to testing GxM. It is imperative to continue to refine statistical methods that allow the analyst to compare alternative models that include different mechanisms by which a putative moderator may influence the genetic and environmental influences on the behavior of interest.

In a large simulation study, we started with a generous sample size that was likely to provide adequate power. In all settings wherein the null model was correct, however, the null model was not rejected in favor of model (3) as often as expected. This will lead to underpowered tests for GxM at any practical sample size. We do not believe the problem is due to the specific algorithm or software (Mplus) for fitting the models because we have independently developed our own fitting algorithm in R (R Development Core Team, 2011), and in the vast majority of cases obtained likelihood values that were nearly identical to those obtained in Mplus. The problematic tests are not on the boundary of the parameter space, which is where one would often encounter problems. Finally, an examination of the asymptotic behavior did not resolve the issue; we increased the sample size to 4000 each of MZ/DZ pairs and obtained similar results. The statistical literature has little in the way of methodological investigations of the asymptotics of non-linear structural equations models; thus, at this time we cannot give an explanation for why tests for GxM are underpowered. It appears unlikely that tests using the chi-square reference distribution are liberal.

Computational issues arose with the correlated factors GxM model (5). The correlated factors GxM model allows one to look at relationships between two (or more) variables when there is no specific ordering. It has three fewer terms than the Cholesky GxM model (3), but proved to be more difficult to fit in Mplus. We specified the model two different ways: first we directly specified a correlated factors GxM model, and second we imposed restrictions on model (3) in order to recover model (5). Both methods were computationally

intensive and may not have resulted in an optimal fit. We encountered similar problems with the correlated factors GxM model [i.e., long convergence times and unstable results] when using our own fitting algorithm in R. Whereas model (5) seems like an interesting alternative GxM model to model (3), at this point, we cannot recommend its use.

We also found that across all models, several sets of starting values were needed in order to obtain the optimal log-likelihood. Model (3) was particularly sensitive to mis-specified starting values. Another statistical program for fitting non-linear structural equations models may do a better job of fitting these complex models. When using these models in applications with Mplus, we recommend at least 10 sets of starting values to ensure a global maximum is achieved.

We found that GxM can be erroneously detected when it does not hold under non-linear main effects of M on P, unless the alternative non-linear main effects GxM model (4) is specifically tested as a comparison to model (3). The problem arises specifically when detected GxM is with respect to the common influences on M and P rather than the influences unique to P. As researchers apply these methods to more highly correlated candidate moderators (e.g., endophenotypes) and phenotypes, we anticipate that r_{GM} will be greater and that candidate GxM will be more often concentrated among the common (versus the unique) influences on M and P. We therefore recommend that when looking at possible GxM interactions, researchers fit model (4) and subsets thereof, as well as models (3) and (2), particularly when there is clear evidence of a correlation between the moderator and the phenotype of interest.

Testing alternative model (4) has not been the practice in published studies of GxM to date. This does not mean that results derived from those studies are always, or even ever, incorrect, but the present simulation study shows ways in which incorrect conclusions could have been reached. The restriction that the genetic and environmental influences on the moderator contribute to the phenotype to the same degree is indeed a strong assumption. But this assumption can be directly tested with the data. Moreover, if the assumption holds and (3) is tested without considering (4) or (4*) then non-linear main effects β_1 and β_2 may be detected as non-zero interaction effects α_C , κ_C , or e_C . Alternatively, if (4) and (4*) are rejected in favor of (3), then the conclusion of GxM (in 3) is even more convincing than if (4) and (4*) had not been examined.

It is difficult to evaluate the likelihood of non-linear main effects in published studies using Purcell's (2002) procedures because most do not report the parameter estimates from fitted models. We highly recommend that the practice of reporting parameter estimates replace the practice of plotting variance components, as the former yields much greater interpretability of the analyses on the part of the reader (Rathouz et al., 2008).

It is important to note that all latent variables were normally distributed in the present simulation study. Performance of model fitting, testing and comparison procedures rely to an unknown degree on the distributional properties of the responses M and P. Because many measures of behavior, especially in psychiatry, have a substantial number of zeros and are skewed to the right, it will be important to determine how the procedures perform when data do not derive from normally distributed latent variables. This is the subject of ongoing work.

Finally, in Rathouz et al. (2008), we proposed several other alternatives to Model (3) that include multiplicative gene-gene or environment-environment effects that could be mistaken for GxM. Because these alternative models involve quadratic or multiplicative functions of latent variables, they cannot be fitted in available structural equation modeling software such as Mplus, and were therefore excluded from the current study. We are in the process of developing and testing computational algorithms in R to be able to fit these models. It will

then be necessary to determine how they compare to models (3) and (4). The R package under development will also allow researchers to fit the models described in the current study. The bivariate Cholesky and correlated factors models (without GxM) can currently be implemented using the R package OpenMx (Boker et al., 2011).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This study was funded by the NIH grant R21 MH086099 from the National Institute for Mental Health. Infrastructure support was provided by the Waisman Center via a core grant from the National Institute of Child Health and Human Development (P30 HD03352).

References

- Asbury K, Dunn JF, Plomin R. Birthweight-discordance and differences in early parenting relate to monozygotic twin differences in behaviour problems and academic achievement at age 7. *Developmental Science*. 2006; 9(2):F22–F31.10.1111/j.1467-7687.2006.00469.x [PubMed: 16472310]
- Boker S, Neale M, Maes H, Wilde M, Spiegel M, Brick T, Spies J, et al. OpenMx: An Open Source Extended Structural Equation Modeling Framework. *Psychometrika*. 2011; 76(2):306–317.10.1007/s11336-010-9200-6 [PubMed: 23258944]
- Dick DM, Rose RJ, Viken RJ, Kaprio J, Koskenvuo M. Exploring gene environment interactions: Socioregional moderation of alcohol use. *Journal of Abnormal Psychology*. 2001; 110(4):625–632.10.1037/0021-843X.110.4.625 [PubMed: 11727951]
- Johnson W, Krueger RF. Higher Perceived Life Control Decreases Genetic Variance in Physical Health: Evidence From a National Twin Study. *Journal of Personality and Social Psychology*. 2005; 88(1):165–173.10.1037/0022-3514.88.1.165 [PubMed: 15631582]
- Lahey BB, Waldman ID, Loft JD, Hankin B, Rick J. The structure of child and adolescent psychopathology: Generating new hypotheses. *Journal of Abnormal Psychology*. 2004; 113:358–385. [PubMed: 15311983]
- McArdle JJ, Goldsmith HH. Alternative common factor models for multivariate biometric analyses. *Behavior Genetics*. 1990; 20(5):569–608.10.1007/BF01065873 [PubMed: 2288547]
- Medland SE, Neale MC, Eaves LJ, Neale BM. A Note on the Parameterization of Purcell's G×E Model for Ordinal and Binary Data. *Behavior Genetics*. 2008; 39(2):220–229.10.1007/s10519-008-9247-7 [PubMed: 19083089]
- Muthén, L.; Muthén, B. *Mplus User's Guide* (Sixth). Los Angeles CA: Muthén & Muthén; 1998.
- Neale, M.; Cardon, L. *Methodology for genetic studies of twin and families*. Boston, MA: Kluwer Academic Publishers; 1992. NATO ASI Series D: Behavioral and social sciences (Vol. 67)
- Price T, Jaffee S. Effects of the family environment: Gene-environment interaction and passive gene-environment correlation. *Developmental Psychology*. 2008; 44(2):305–315.10.1037/0012-1649.44.2.305 [PubMed: 18331124]
- Purcell S. Variance Components Models for Gene–Environment Interaction in Twin Analysis. *Twin Research*. 2002; 5(6):554–571.10.1375/136905202762342026 [PubMed: 12573187]
- R Development Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2011. Retrieved from <http://www.r-project.org>
- Raftery AE. Bayesian model selection in social research. *Sociological Methodology*. 1995; 25:111–163.
- Rathouz PJ, Van Hulle CA, Rodgers JL, Waldman ID, Lahey BB. Specification, Testing, and Interpretation of Gene-by-Measured-Environment Interaction Models in the Presence of Gene–Environment Correlation. *Behavior Genetics*. 2008; 38(3):301–315.10.1007/s10519-008-9193-4 [PubMed: 18293078]

- Saunders C, Bishop D, Barrett J. Sample size calculations for main effects and interactions in case-control studies using Stat's `nchi2` and `npnchi2` functions. *The Stata Journal*. 2003; 3(1):47–56.
- Silventoinen K, Hasselbalch AL, Lallukka T, Bogl L, Pietilainen KH, Heitmann BL, Schousboe K, et al. Modification effects of physical activity and protein intake on heritability of body size and composition. *American Journal of Clinical Nutrition*. 2009; 90(4):1096–1103.10.3945/ajcn.2009.27689 [PubMed: 19710197]
- South SC, Krueger RF. Marital quality moderates genetic and environmental influences on the internalizing spectrum. *Journal of Abnormal Psychology*. 2008; 117(4):826–837.10.1037/a0013499 [PubMed: 19025229]
- StataCorp. *Stata Statistical Software: Release 12*. College Station, TX: StataCorp LP; 2011.
- Turkheimer E, Haley A, Waldron M, D'Onofrio B, Gottesman II. Socioeconomic status modifies heritability of IQ in young children. *Psychological Science*. 2003; 14(6):623–628.10.1046/j.0956-7976.2003.psci_1475.x [PubMed: 14629696]
- van Beijsterveldt T, Boomsma DI. An exploration of gene-environment interaction and asthma in a large sample of 5-year-old Dutch twins. *Twin Research and Human Genetics*. 2008; 11(2):143–149. [PubMed: 18361714]

Figure 1. Path diagrams for models with (a) and without (b) moderation of the influences common to M and P . (a) Purcell's (2002) model for testing and quantifying GxM in the presence of rGM. (b) alternative model containing main effects of M and M^2 instead of α_C , κ_C , and ϵ_C terms.

Figure 2. Scatterplot of childhood/adolescent total anxiety by birthweight (standardized after residualizing on age, gender, and race).

\$watermark-text

\$watermark-text

\$watermark-text

Table 1

Models and data generation mechanism (DGM) parameter values used in simulation.

Model	DGM	Correlation	Simulation Interaction	Simulation Condition	Non-Linear Main Effect	Simulation Parameter Values	Nested in
Cholesky (Chol)	(2A)	High r_{AM}	---	---	---	$a_C=.5; a_U=0.806$	(3), (5)
		Low r_{EM}				$e_C=0.1; e_U=0.94$	
	(2B)	Low r_{AM}	---	---	---	$a_C=0.1; a_U=0.94$	(3), (5)
		High r_{EM}				$e_C=0.5; e_U=0.806$	
Cholesky with GxM (CholGxM)	(3A)	High r_{AM}	High AxM	---	---	$a_C=0.5; a_C=0.25; a_U=0.806; a_U=0.403$	
		Low r_{EM}	Low ExM			$e_C=0.1; e_C=0.025; e_U=0.94; e_U=0.235$	
	(3B)	High r_{AM}	Low AxM;	---	---	$a_C=0.5; a_C=0.125; a_U=0.806; a_U=0.25$	
		Low r_{EM}	High ExM			$e_C=0.1; e_C=0.05; e_U=0.94; e_U=0.47$	
	(3C)	Low r_{AM}	High AxM	---	---	$a_C=0.1; a_C=0.05; a_U=0.94; a_U=0.47$	
		High r_{EM}	Low ExM			$e_C=0.5; e_C=0.125; e_U=0.806; e_U=0.202$	
	(3D)	Low r_{AM}	Low AxM	---	---	$a_C=0.1; a_C=0.025; a_U=0.94; a_U=0.235$	
		High r_{EM}	High ExM			$e_C=0.5; e_C=0.25; e_U=0.806; e_U=0.403$	
Nonlinear Main Effects with GxM (NLMainGxM)	(4A)	---	Low AxM	Large	Large	$\beta_1=0.51; \beta_2=0.127$	(3)
			Low ExM			$a_U=0.806; e_U=0.94; a_U=0.201; e_U=0.235$	
	(4B)	---	Low AxM	Small	Small	$\beta_1=0.51; \beta_2=0.0637$	(3)
			Low ExM			$a_U=0.806; e_U=0.94; a_U=0.201; e_U=0.235$	
Nonlinear Main Effects only (NLMain)	(4*A)	---	---	Large	Large	$\beta_1=0.51; \beta_2=0.127$	(3), (4)
						$a_U=0.806; e_U=0.94$	
	(4*B)	---	---	Small	Small	$\beta_1=0.51; \beta_2=0.0637$	(3), (4)
						$a_U=0.806; e_U=0.94$	
Linear Main Effects (LinMain)	(4†)	---	---	---	---	$\beta_1=0.51; a_U=0.806; e_U=0.94$	(2), (4*)
	(5A)	High r_{AM}	High AxM	---	---	$r_{AM}=0.527; a_P=0.474; a_P=e_P=0.94$	(3)
		Low r_{EM}	Low ExM			$r_{EM}=0.105; e_P=0.237$	
	(5B)	High r_{AM}	Low AxM	---	---	$r_{AM}=0.527; a_P=0.237; a_P=e_P=0.94$	(3)
Correlated factors with GxM (CorrGxM)		Low r_{EM}	High ExM			$r_{EM}=0.105; e_P=0.474$	

Model	DGM	Simulation Condition		Simulation Parameter Values	Nested in
		Correlation	Interaction		
	(5C)	Low r_{AM}	High AxM	$r_{AM} = 0.105$; $\alpha_P = 0.474$; $a_P = e_P = 0.94$	(3)
		High r_{EM}	Low ExM	$r_{EM} = 0.527$; $e_P = 0.237$;	
	(5D)	Low r_{AM}	Low AxM	$r_{AM} = 0.105$; $\alpha_P = 0.237$; $a_P = e_P = 0.94$	(3)
		High r_{EM}	High ExM	$r_{EM} = 0.527$; $e_P = 0.474$	

Note: DGM = data generating mechanism. DGM numbers correspond to model numbers in text, with A–D enumerating specific instances. For all simulation conditions, $a_M = e_M = \sqrt{0.9}$, $c_M = \sqrt{0.1}$, and $c_U = \sqrt{0.2}$; for DGM's 3A–D, interaction between shared environment and the moderator $\kappa_C = \kappa_U = 0.01$ and shared environment common to M and P, $c_C = 0.01$; for DGM's 4A–B values for the interaction between the variance components unique to P and the moderator were $\alpha_U = 0.25^*a_U$, $\kappa_U = 0.01$, $e_U = 0.25^*e_U$; for DGM's 5A–D, interaction between shared environment and moderator $\kappa_P = 0.01$ and correlations between the shared environment and the moderator $r_{CM} = 0.01$.

Table 2

Percent of LRT statistics under the null hypothesis exceeding critical value for pairs of nested models based on 2000 replicates of either $n=2000$ twin pairs or $n=500$ twin pairs.

% Type I error rates												
Model for H_A	Model for H_0	DGM	Condition of H_0	LRT df	N = 2000			N = 500				
					10	5	1	10	5	1		
Cholesky with GxM	(3)	Correlated factors with GxM	high r_{AM}	high AxM	3	10.9	5.9	1.2	5.1	2.8	0.9	
			high r_{AM}	low AxM	3	5.7	2.6	0.5	3.4	1.4	0.8	
			low r_{AM}	high AxM	3	26.9	20.2	10.0	8.9	4.6	1.9	
			low r_{AM}	low AxM	3	6.5	2.7	0.8	3.5	2.3	1.6	
NL main effects with GxM	(4)	NL main effects with GxM	large β_2		4	5.1	2.6	0.7	4.3	1.8	0.1	
			small β_2		4	5.5	2.9	1.0	4.3	1.9	0.2	
			large β_2	(4*)	7	6.6	3.5	0.7	4.4	1.7	0.5	
			small β_2		7	6.6	3.5	.07	4.4	1.7	0.5	
Cholesky	(2)	Cholesky	high r_{AM}		6	5.8	3.5	0.7	4.1	1.7	0.2	
			low r_{AM}		6	6.9	2.9	0.6	4.6	1.8	0.4	
			high r_{AM}	(2)	3	11.7	5.7	1.4	9.8	4.7	1.1	
			low r_{AM}		3	11.9	6.4	1.6	11.6	5.9	1.0	
Correlated Factors with GxM	(5)	Cholesky	high r_{AM}		3	11.7	5.7	1.4	9.8	4.7	1.1	
			low r_{AM}		3	11.9	6.4	1.6	11.6	5.9	1.0	
NL main effects with GxM	(4)	NL main effects	large β_2	(4*)	3	9.4	2.7	1.1	6.5	1.7	0.4	
			small β_2		3	9.4	4.8	1.1	6.5	2.8	0.4	
NL main effects	(4*)	Lin main effects		(4†)	1	9.6	4.7	1.1	11.1	5.3	1.1	
Cholesky	(2)	Lin Main effects		(4†)	2	8.4	3.8	0.9	8.4	4.6	0.9	

Note: For all DGM's high r_{AM} denotes corresponding low r_{EM} and high AxM denotes low ExM. The same rule applies to low r_{AM} and low ExM.

Table 3

Estimated sample size under the alternative hypothesis needed to reject the null hypothesis with 70% or 90% power ($\alpha=.05$).

DGM		Power	
Model for H_A	Condition of H_A	Model for H_0	
Cholesky with GxM	high r_{AM}	high AxM	Chol
	high r_{AM}	low AxM	
	low r_{AM}	high AxM	
	low r_{AM}	low AxM	
	high r_{AM}	high AxM	NLMMain
	high r_{AM}	low AxM	
	low r_{AM}	high AxM	
	low r_{AM}	low AxM	
	high r_{AM}	high AxM	NLMMainGxM
	high r_{AM}	low AxM	
	low r_{AM}	high AxM	
	low r_{AM}	low AxM	
NL Main Effects with GxM			
		large β_2	NLMMain
		small β_2	
NL Main Effects			
		large β_2	LinMain
		small β_2	

Table 4
Percent each model is favored via comparison of BIC values^a for all pairwise comparisons; n=2000 and n=500

	DGM											
	CHOL				CHOLGxM				NLMainGxM			
	2A	500	2000	500	2000	500	2000	500	2000	500	2000	500
CholGxM	2000	500	2000	500	2000	500	2000	500	2000	500	2000	500
Equivocal												
Chol	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	95.8	2.1
											4.2	33.2
												14.7
												1.0
											64.7	84.6
												100.0
CholGxM	33.4		99.0	8.4	25.7	0.3	52.8	0.6	99.0	8.7	95.4	5.0
Equivocal	60.0	25.6	1.0	67.7	63.8	21.6	41.3	31.3	1.0	68.8	2.8	61.1
NLMainGxM	6.6	74.3		23.9	10.5	78.1	5.9	68.1		22.4	1.8	33.9
											100.0	100.0
											100.0	100.0
											100.0	100.0
CholGxM	4.1		85.6		100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Equivocal	30.0	1.0	13.4	16.4								
NLMain	65.9	99.0	1.1	83.3							100.0	100.0
											100.0	100.0
											100.0	100.0
NLMainGxM												
Equivocal	0.8	1.9	1.5	4.0							0.7	1.8
NLMain	99.2	98.1	98.5	96.0							99.3	98.2
											99.3	98.2
NLMainGxM												
Equivocal											42.9	36.8
Chol	100.0	94.1	100.0	100.0							56.2	61.5
											0.8	1.7
												22.6
NLMain											97.3	98.9
Equivocal	5.8	86.2		38.7							100.0	45.7
Chol	94.2	13.6	100.0	61.3								1.1
											2.7	54.3
Lowest Overall BIC ^b												
CholGxM												
NLMainGxM												
NLMain	0.2	4.0		36.2							100.0	99.4
											100.0	100.0
											100.0	100.0

\$watermark-text

\$watermark-text

\$watermark-text

\$watermark-text

\$watermark-text

\$watermark-text

\$watermark-text

Table 5

Parameter estimates from fitting Cholesky with GxM and Non-linear Main Effects with GxM to birthweight and child/adolescent anxiety.

	Birthweight (M)	Anxiety (P)	Common Effects	Interaction Common to M and P	Interaction Unique to P
CholGxM	$a_M = 0.42 (0.36 - 0.47)$	$a_U = 0.57 (0.45 - 0.67)$	$a_C = 0.02 (-0.11 - 0.16)$	$\alpha_C = -0.16 (-0.23 - -0.08)$	$\alpha_U = 0.004 (-0.13 - 0.13)$
	$c_M = 0.83 (0.75 - 0.86)$	$c_U = 0.47 (0.37 - 0.58)$	$c_C = -0.08 (-0.15 - -0.02)$	$\kappa_C = 0.09 (0.04 - 0.14)$	$\kappa_U = -0.01 (-0.13 - 0.12)$
	$e_M = 0.38 (0.36 - 0.40)$	$e_U = 0.64 (0.60 - 0.67)$	$e_C = -0.04 (-0.08 - 0.01)$	$e_C = 0.08 (0.04 - -0.13)$	$e_U = -0.03 (-0.06 - 0.01)$
NLMMainGxM	$a_M = 0.42 (0.36 - 0.47)$	$a_U = 0.59 (0.47 - 0.71)$	$\beta_1 = -0.07 (-0.11 - 0.04)$	$\beta_2 = 0.04 (0.02 - 0.07)$	$\alpha_U = -0.10 (-0.19 - -0.01)$
	$c_M = 0.83 (0.75 - 0.86)$	$c_U = 0.46 (0.33 - 0.58)$			$\kappa_U = -0.08 (-0.03 - 0.18)$
	$e_M = 0.38 (0.36 - 0.40)$	$e_U = 0.64 (0.61 - 0.67)$			$e_U = -0.10 (-0.03 - 0.02)$