

Multilevel Models for Family Data

Jon Rasbash

University of Bristol

Tom O'Connor

University of Rochester Medical Center

Jenny Jenkins

Toronto University

## Abstract

In this paper we highlight some of the more pressing conceptual questions in family research that require additional methodological development. We then outline a basic structure for multilevel models for family data to address each of the issues raised. Models for testing genetic hypotheses, gene-environment interactions, environmental hypotheses of “shared” and “non-shared” effects, and family systems are outlined and demonstrated. Empirical examples are provided for two data sets involving 1-4 children per family. We then discuss a list of useful model extensions and compare our method to existing work that has used multilevel models to estimate genetic effects. The algorithm used (described in full in an appendix) can handle household and extended family data of any size and complexity and can handle a wide array of analytic questions in family research.

## Multilevel Models for Family Data

Multilevel models, also known as variance component models, random effects models and hierarchical linear models (Bryke and Raudenbush, 2001; Goldstein, 2003; Longford, 1993; Snijders & Bosker, 1998), have seen rapid growth and development over the last twenty years and are now becoming a standard part of the quantitative social scientist's toolkit. Multilevel models provide a flexible regression modelling framework for handling data sampled from clustered population structures, such as students within classes within schools, patients within hospitals, repeated measurements within individuals, or children within families. Ignoring the multilevel structure of the data can lead to incorrect inferences because the standard errors of regression coefficients are underestimated. Moreover, if the higher-level units such as neighborhood or family are left out of the model, then we can not explore potentially important questions about their effects, which we refer to as "context" (Earls et al., 1999; Jenkins et al., 2003). It is because most social data have a strong hierarchical structure that multilevel models are becoming so widely used in social science.

One natural application of multilevel models is family studies, where children are nested within families (Jenkins et al., 2002; O'Connor et al., 2001; Raudenbush et al., 1995; Snijders, 1995). This paper extends this work by demonstrating a method for including genetic effects in multilevel models for family data.

Researchers studying family "effects" have struggled to address a number of conceptual issues in an integrated analytic paradigm. We outline what these conceptual issues are before providing empirical examples.

*Testing Genetic hypotheses in family data*

One of the most persuasive alternative explanations for family environmental effects derives from behavioural genetics. Put simply, the rationale is that, because family members

share both genes and family environment, studies showing an association between some aspect of the family environment and child well-being might actually be mediated by shared genetic factors. Thus, the well-replicated finding that, say, parental warmth/support or marital conflict predicts child behavior/emotional problems does not necessarily imply environmental mediation. In fact, many studies have shown that features of the family environment and, more impressively, associations between family environment and child outcomes are at least partly mediated by genetics (Ge et al., 1996; O'Connor et al., 2000; Plomin, 1994). The need to consider hypotheses testing both genetic and environmental mediation is now appreciated in the practice of family research. Nevertheless, conclusions from these studies remain controversial, primarily because of criticisms about the methodology and data analytic approach employed (see, Collins et al., 2000; 2002).

Studies using behavioural genetic methods have been criticized for failing to include measured family environments (family level explanatory variables) and instead inferring environmental effects from a lack of genetic effects, extended family designs (e.g., parents as well as twins), and gene-environment interplay, most notably gene-environment interactions. Although programs that can address these issues are available (e.g., Neale, 2000), empirical examples that counter these criticisms (e.g., Caspi et al., 2000; D'Onofrio et al., 2004) are not typical. Thus, most behavioural genetic reports are still vulnerable to these criticisms. We suggest that a multilevel model framework provides an accessible and broad model to test hypotheses of genetic and environmental influence and is sensitive to detecting the complex pattern of gene-environment interplay, notably gene-environment interactions.

#### *Testing systems and context effects in the family*

A second major challenge to family research concerns the detection of “family level” processes or family context effects. That is, despite the recognized importance of importing

family systems notions into research, few family studies actually assess *family* level differences and sources of variation. Instead, family researchers most commonly assess dyadic relationships within the family – parent-child, marital and sibling, their overlap with each other and child outcomes. Analyses at the dyadic, within-family level do not necessarily indicate what, if anything, makes *families* different from one another. The totality of the family system can not be captured by analysing a set of overlapping dyads separately. In such analyses which include only one dyad per family dyad and family are confounded and no separation can be made between dyad and family components of variation. Furthermore, failure to consider family-level factors, such as whether family stresses such as parental conflict or divorce have shared or common effects on children in the family may be partly reinforced by behavioral genetics findings suggesting that similarity between siblings may be more attributable to shared genes than “shared” environment. However, the general conclusion that “shared environment” has a minimal role may be premature because few studies directly test the hypothesis that there are shared, common or family-wide effects of family stresses.

Few solutions for studying family systems and family-level effects have been proposed. One approach that has been suggested as a possible analytic solution is the social relations model developed by Kenny and colleagues (Cook, 2002; Snijder & Kenny, 1999). When data on relationships within the family are collected on 3 or more family members, the social relations model is able to model variation in relationship quality that is attributable to each member, each dyad, *and* the family. What is significant about the approach for family theory is that it is possible to examine how and why families differ from one another, not just how relationships in different families differ from one another. We elaborate on this

approach below, and show how multilevel model analytic framework can be adapted to incorporate genetics within the social relations model approach.

Family data present a number of methodological challenges that are familiar to researchers in other areas. Accordingly, family data provide a useful laboratory for examining a wide range of data analytic questions, with relevance the extends behind testing hypotheses about “family effects.”

### Method and Results

#### *A basic multilevel model for family data*

If we have data on children within families, we can write the basic multilevel model as

$$\begin{aligned}
 y_{ij} &= \beta_0 + u_j + e_{ij} \\
 u_j &\sim N(0, \sigma_u^2) \quad e_{ij} \sim N(0, \sigma_e^2) \quad (1)
 \end{aligned}$$

$y_{ij}$  is the response for the  $i$ 'th child in the  $j$ 'th family,  $\beta_0$  is the overall mean,  $u_j$  is a departure for the  $j$ 'th family so that  $\beta_0 + u_j$  gives the mean for the  $j$ 'th family,  $e_{ij}$  is the departure of the  $i$ 'th individual from the  $j$ 'th families mean. In the vocabulary of family and behavioural genetics research,  $u_j$  can be described as a “shared environmental” effect, between-family variation, or clustering of measurements at the family level;  $e_{ij}$  can be described as a “non-shared environmental” effect, within-family variation, or child-specific effect.

This model is like a one-way ANOVA except that the  $u_j$  are now random effects as opposed to fixed effects. Usually, we are not interested in making statements about particular families we have sampled. Rather, we wish to make more general statements about the

population of families from which we have drawn our sample. In an ANOVA model we fit one dummy variable to each family. Unfortunately, this means any statements we make about families apply only to the families we have sampled and we lose the generality of inference of the multilevel model. The multilevel model is also more parsimonious because we do not fit a  $u_j$  for every family, but instead we estimate  $\sigma_u^2$  directly. Thus, if we have a 1000 families we have reduced the number of parameters needed to model the between family variation from 1000 to 1.

Once we have estimated the family level variance, a natural avenue of exploration is to add family level variables in order to explain the between family variation. This is not possible in ANOVA models, since having fitted one parameter to each family there are no degrees of freedom remaining to fit further family level explanatory variables. The parsimony of multilevel models allows family level explanatory variables to be fitted as well as directly estimating the residual between family variance. Another key advantage of multilevel models is that they allow complex specifications for the covariance structure of the model. This final point is utilised when we incorporate genetic effects in multilevel models.

#### *Extending the basic model to include genetic effects*

Let us now look at the variance and covariance structure generated by model 1 and see how this changes when we include genetic effects. The variance of the response measurements is

$$\text{var}(y_{ij}) = \sigma_u^2 + \sigma_e^2$$

given the standard independence assumptions of multilevel models:

$$\text{cov}(u_{j_1}, u_{j_2}) = 0$$

$$\text{cov}(u_j, e_{ij}) = 0$$

$$\text{cov}(e_{i_1j}, e_{i_2j}) = 0$$

These three independence assumptions are: random effects for two different families are uncorrelated, family and child random effects are uncorrelated, and random effects for two different children are uncorrelated. The covariance of two children ( $i_1$  and  $i_2$ ) in the same family is

$$\text{cov}(y_{i_1j}, y_{i_2j}) = \text{cov}(u_j + e_{i_1j}, u_j + e_{i_2j}) = \sigma_u^2$$

We can elaborate (1) to incorporate genetic effects as follows:

$$\begin{aligned} y_{ij} &= \beta_0 + u_j + e_{ij} + g_{ij} \\ u_j &\sim N(0, \sigma_u^2) \quad e_{ij} \sim N(0, \sigma_e^2) \quad g_{ij} \sim N(0, \sigma_g^2) \end{aligned} \quad (2)$$

Where  $g_{ij}$  is a genetic effect for the  $i$ 'th child in the  $j$ 'th family. That is, a child's genetic make up contributes to that child's response by an amount  $g_{ij}$ ; the genetic effects are another type of child level residual. For two individuals, ( $i_1, i_2$ )  $\text{cov}(e_{i_1j}, e_{i_2j}) = 0$ ; however, the genetic covariance of two individuals in the same family,  $\text{cov}(g_{i_1j}, g_{i_2j})$ , is clearly not zero since there is a non-zero probability that they share the same genes. Behavioural genetics research makes a number of assumptions in their models. These assumptions are not without criticism, but we merely carry over to the present context. Specifically, these assumptions are: additive genetic variation, random mating, an evolutionary stable population, no gene environment correlations and independent transmission of genes from parents to offspring. The covariance between two individuals in the same family is

$$\begin{aligned} \text{cov}(y_{i_1j}, y_{i_2j}) &= \text{cov}(u_j + e_{i_1j} + g_{i_1j}, u_j + e_{i_2j} + g_{i_2j}) \\ &= \sigma_u^2 + r_{(i_1i_2)j} \sigma_g^2 \end{aligned}$$

Where  $r_{(i_1i_2)j}$  is the relationship coefficient between two individuals and equals (0,0.125,0.25,0.5,1) for unrelated individuals, cousins, half-sibs, full sibs/DZ twins, and MZ twins, respectively. Model (2) therefore gives us a direct estimate of the additive genetic



variation ( $\sigma_g^2$ ), the shared environmental variation ( $\sigma_u^2$ ), and the non-shared environmental variation ( $\sigma_e^2$ ). The model is estimated using an adaptation of the Iterative Generalised Least Squares algorithm described in Goldstein and Rasbash (1993), for details of the adaptation see the appendix.

*Applying the model.* We now illustrate the models (1) and (2) using data collected on Nonshared Environment and Adolescent Development project (NEAD; Reiss et al., 1994, 2000). This data set has 277 full sib pairs in nuclear and stepfamilies, 109 half-sib pairs, 130 unrelated pairs, 93 DZ twins and 99 MZ twins aged between 10 and 18 years. Note that although this data set contains two children per family the method will handle any number of children per family, as we show below. We start by analysing depression scores in the children. Depression scores are formed from a composite of three measures: the Child Depression Inventory (Kovacs, 1983), the depression subscale from the Behaviour Problems Index (BPI; Zill, 1985), and the depression subscale from the Behaviour Events Inventory (BEI; Patterson, 1982). Genetic analyses of child depression in NEAD have been written up elsewhere (O'Connor et al., 1998). In the current paper, we use the child depression data to illustrate the multilevel model analytic approach. The results are shown in Table 1a.

The total variance in the two models is effectively the same 0.284 in model 1 and 0.296 in model 2. In model 2, which includes genetic effects, 80% of the shared environmental variation and 65% of the non-shared environmental variation are re-assigned to the genetic variance. After genetic effects have been added the shared environmental variation (0.018) compared to its standard error (0.017) appears to be non-significant. Caution is needed here because this assumes that this variance is Normally distributed, which is only asymptotically true. With over 700 families it is probably reasonable to assume a Normal distribution for the variances. However, a better test is to remove the parameter from

the model and observe the change in deviance. Under this test, the change in deviance is 1.5 ( $p = 0.22$ ) and which confirms that after genetic effects have been added the shared environmental variance is no longer significant. In other words, there is a clustering of variance in child depression at the family level, and the reason for this is essentially entirely accounted for by genetic resemblance of siblings.

*Including covariates in the model*

We can elaborate the model including covariates for child age, gender, paternal and maternal negativity and whether the child is a member of a stepfamily. The parental negativity measures were formed from composites of the following measures: Parent Discipline Behavior (Hetherington & Clingempeel, 1992), punitive and yielding to coercion subscales; Parent-Child Disagreements (Hetherington & Clingempeel, 1992), conflict subscale and Conflict Tactics Scale (Strauss, 1979), symbolic aggression subscale.

The results of the elaborated model are shown in Table 1b. From the fixed effects we see that depression scores increase with child age, paternal and maternal negativity; girls and children in stepfamilies also have higher depression scores. Adding the covariates reduces the shared environment and genetic variances, but does not reduce the non-shared environment variance. The reduction in the genetic variance (of 29% from 0.209 to 0.148) occurs when maternal and paternal negativity are added to the model.

Pike et al. (1996) analyse the same data using a series of genetically calibrated bivariate structural equations models. Two of the models they consider are bivariate structural equations models for maternal negativity and depression and paternal negativity and depression. In each of these two models they find 15% of the genetic variance in depression is due to a shared genetic component with parental negativity. When we add paternal and maternal negativity to our model as fixed effects we are sweeping out any

common genetic effects shared by parental negativity and adolescent depression. We are also taking account of any environmental correlations whereby sibling pairs of greater relatedness experience more similar parental treatment. Both these factors will reduce the remaining additive genetic variance in the model.

The conceptual issue raised by this approach is whether the interest is in modelling genetic variance/heritability of the “raw” variable, modelling genetic variance/heritability after first accounting for key explanatory variables, or modelling the effects of explanatory variables together with genetic variance. Much of the behavioural genetics research adopts the first approach. More recent behavioural genetic research has focused on the latter approach. We therefore extend the multilevel framework to look at a particular example of how multilevel approaches may be used to test hypotheses about the interplay between environmental and genetic processes, gene-environment interactions.

### *Complex variation and gene environment interactions*

In a multilevel model we simultaneously estimate a model for the mean, specified by the  $\beta$  coefficients (also known as fixed effects) and a model for the variance. Currently our model for the variance partitions the variance into three sources: shared environment, non-shared environment and genetics. The model for the variance can be further elaborated to allow each of the three sources of variation to be modelled as functions of explanatory variables, where the variables may be measured at any level (child or family) and be continuous or categorical. That is

$$\begin{aligned}\sigma_{uij}^2 &= \alpha_0^{(u)} + \alpha_1^{(u)} x_{1ij} + \alpha_2^{(u)} x_{2j} \dots \\ \sigma_{eij}^2 &= \alpha_0^{(e)} + \alpha_1^{(e)} x_{1ij} + \alpha_2^{(e)} x_{2j} \dots \quad (3) \\ \sigma_{gij}^2 &= \alpha_0^{(g)} + \alpha_1^{(g)} x_{1ij} + \alpha_2^{(g)} x_{2j} \dots\end{aligned}$$

We now elaborate the model to allow all three variances to be a function of paternal negativity.

That is:

$$\begin{aligned}
 y_{ij} &= (XB)_{ij} + u_j + e_{ij} + g_{ij} \\
 \text{var}(u_j) &= \sigma_u^2 \quad \text{var}(e_{ij}) = \sigma_e^2 \quad \text{var}(g_{ij}) = \sigma_g^2 \\
 \sigma_{eij}^2 &= \alpha_0^{(u)} + \alpha_1^{(u)} \text{pat\_neg}_{ij} \\
 \sigma_{eij}^2 &= \alpha_0^{(e)} + \alpha_1^{(e)} \text{pat\_neg}_{ij} \\
 \sigma_{gij}^2 &= \alpha_0^{(g)} + \alpha_1^{(g)} \text{pat\_neg}_{ij}
 \end{aligned} \tag{4}$$

The results are presented in model 4 in Table 1c.

Including the three extra parameters reduces the deviance by 40.5. This reduction is almost entirely driven by the gene environment interaction term  $\alpha_1^{(g)}$ ; removing the  $\alpha_1^{(u)}$  and  $\alpha_1^{(e)}$  terms from the model 4 results in a change in only 1.1 in the deviance. The significant  $\alpha_1^{(g)}$  coefficient constitutes a gene-environment interaction because it implies the genetic variance changes as a function of paternal negativity.

Further exploration of the three variances as functions of explanatory variables revealed one other strong interaction, between the genetic variance and gender; the results are given in Table 1c model 5. In model 5 we drop the non-significant  $\alpha_1^{(u)}$  and  $\alpha_1^{(e)}$  and introduce a term modelling the genetic variance as a function of gender. That is:

$$\begin{aligned}
 y_{ij} &= (XB)_{ij} + u_j + e_{ij} + g_{ij} \\
 \text{var}(u_j) &= \sigma_u^2 \quad \text{var}(e_{ij}) = \sigma_e^2 \quad \text{var}(g_{ij}) = \sigma_g^2 \\
 \sigma_{eij}^2 &= \alpha_0^{(u)} \\
 \sigma_{eij}^2 &= \alpha_0^{(e)} + \alpha_1^{(e)} \text{pat\_neg}_{ij} + \alpha_2^{(g)} \text{girl}_j \\
 \sigma_{gij}^2 &= \alpha_0^{(g)}
 \end{aligned} \tag{5}$$

The **girl** explanatory variable only has  $j$  subscript because it is actually a sibling-pair level variable, since in this study only same sex pairs were chosen. This obviously limits the generalisability of the results. However, the main intention of this paper is to demonstrate how multilevel models can be used to fit genetic effects.

Comparing model 3 with model 5, which differ only in two terms modelling the genetic variance as functions of paternal negativity and gender, we see a change in the deviance of 50.5, providing evidence for these gene-environment interactions. Figure 1 shows how the genetic variance changes as a function of paternal negativity and gender.

The response variable exhibited some minor skewness. We ran the models on a range of transformations of the response variable (normal scores, log transform and square root transform) to explore the sensitivity of the findings to the response scale. The same pattern of coefficients (in terms of direction and relative size) was found across all response scales. However, the gene-environment interactions were significant on the raw scale and the square root transform, but not on Normal scores and log transforms. Therefore, the findings of these gene-environment interactions must be suggestive rather than definitive.

#### *Analysing a data set with a range of family sizes*

The NEAD data set contains only sibling pairs of the same gender and over samples of twin, half sib and unrelated sibling pairs. This provides a powerful design for detecting genetic effects but limits the generalisability of the findings. We can also apply these models to large-scale population studies of children within families. One of the advantages of using multilevel models to estimate genetic effects is the models extend naturally to any number of children per family. To illustrate the application of these models to a family population study we use the first wave of the National Longitudinal Study of Canadian Youth (NLSCY, 1995). The NLSCY contains 9333 families and 13953 children. After removing any siblings whose

biological relationship could not be determined from the data we are left with 9243 families and 13,579 children. There are 5555, 3084, 560 and 44 families with 1, 2, 3 and 4 children recorded per family, respectively. Note that this is not always the actual number of children per family as data collection was capped at a maximum of 4 children per family in order to reduce the response burden for families. Genetic relationship between siblings was derived from information about a) which biological parent/s the child lived with and b) a comparison of the date of the commencement of the relationship between the child's biological parents for each child in a sibling dyad. The zygosity of twins was not available. We therefore had to omit one twin from each twin pair. As the genetic relationship of the remaining twin to his/her sibling(s) was known it was possible to include these children in the analysis. Finally, we have 29, 202 and 4797 unrelated, half-sib and full-sib dyads respectively.

We analyse as a response the maternal ratings of children's aggression. This included six items scored on a 1-3 scale from never or not true to often or very true. The same stem was used for all questions: "How often would you say that [name]" with the following items: gets into many fights; when mad at someone says bad things behind the other's back; physically attacks people; threatens people; is cruel, bullies or is mean to others; kicks, bites, hits other children? The internal consistency for the scale was good:  $\alpha = .77$ . This response variable is skewed. However, an analysis using Normal scores, raw scores, square root and log transformations all produced the same patterns in terms of relative size and statistical significance for the coefficients. We present results for the raw data.

Covariates were included in model and were all based on the report of the Person Most Knowledgeable (PMK) about the child (in over 90% of the cases this was the mother). Child age was measured in years. Socioeconomic status was assessed by a composite score based on the education and occupation of the PMK and spouse as well as the household

income. Maternal depression was based on a shortened version (12 items) of the CES-D (Radloff, 1977). The 12 items, rated on a 4-point scale, included the core symptoms of depressive symptomatology, e.g., poor concentration, depressed mood, poor appetite, sleeplessness (internal consistency was  $\alpha = .82$ ). Exposure to violence in the home was based on a 4-point scale with the question rated as follows: “How often does (name) see adults or teenagers in your house physically fighting, hitting or otherwise trying to hurt others?” Parental negativity towards child is based on 7 items rated on a 5-point scale: gets annoyed with child for disobedience, proportion of praise when talks to child (reversed scored), proportion of disapproval when talks to the child, gets angry when punishing child, type of punishment depends on mood, has problems managing the child in general, has to discipline repeatedly for the same thing. Internal consistency of this parental negativity scale was = .71.

Table 2 model 1 shows the results decomposing the variation into shared and non-shared components; model 2 introduces genetic effects and model 3 adds covariates. Models 2 and 3 both estimate a heritability of 53% , however all three variance components are reduced when covariates are included. Increasing child age and family socio-economic status (SES) are associated with decreasing child aggression; boys are more aggressive than girls; maternal depression, being exposed to violence between adults and parental negativity directed towards the child are all associated with increasing child aggression. All these findings accord with what has been shown in other reports.

#### *Other useful model extensions*

Multilevel modelling technology can already estimate a wide range of statistical models for handling a wide range of substantive problems. One advantage of embedding

genetic effects in a multilevel modelling framework is that all these extensions become available. A few of these are mentioned now.

*Extension to population structures involving multiple levels of crossing and nesting.*

Often we have many classifications which induce clustering in our response variable. These classifications may be nested or crossed. For example, in addition to children within families, we may have many raters who are making the measurements or administering questionnaires to the individuals in the study. It can be important to estimate the between rater variance in addition to the shared environment, non-shared environment and genetic variances as failure to take between rater variability into account might result in biased estimates of the other variances. If raters assess entire families then we have a multilevel structure with children nested within families nested within raters. If children in the same family are assessed by different raters and if raters assess many families, then children are contained within a cross classification of family by rater. Whether we have a nested or a crossed structure, we can still readily estimate shared environment, non-shared environment, genetic and between rater components of variation. Indeed any of those variance components can be modelled as a function of any explanatory variables measured at any level.

Unfortunately, in family studies the raters or observers are often not identified in the data sets provided for analysis. The usual procedure is to train raters to an “acceptable” level of reliability; where 70% reliability is typically regarded as acceptable. However this level of reliability means that 30% of the variability of the response being analysed is attributable to the rater and depending on how raters are assigned between and within families different biases in the division between shared, non-shared and genetic components variance components arise. These relative size of these variance components is of crucial substantive interest so the problem is potentially a serious one. By using a multilevel model where raters



are simply included as another random classification these biases are removed from the analysis. Of course the identity of the raters must be provided to the data analyst to make this solution possible and we strongly recommend that this information is provided in the future.

Another example of cross-classification arises in the analysis of directed relationship data on dyads within families; this model and the inclusion of genetics effects is described in section 8. We may have further important levels of clustering in our study for example, geographical areas, schools or hospitals depending on the nature of the study. Again multilevel models allow the variance attributable to these clusters to be modelled.

*Repeated measures data.* Here we have measurement occasions ( $i$ ) nested within individuals ( $j$ ) nested within families ( $k$ ). We can write a multilevel growth curve model with genetic effects as

$$y_{ijk} = \beta_0 + \beta_1 t_{ijk} + \beta_2 t_{ijk}^2 + \beta_3 male_{jk} + \beta_4 famtype_k \dots \\ + v_{0k} + u_{0jk} + u_{1jk} t_{ijk} + g_{jk} + e_{ijk}$$

We can include occasion, individual or family level predictors in the model. Coefficients of growth can vary across individuals and families. The above model allows the linear predictor for growth (rate of growth) to vary across families and individuals. Individuals are now at level 2 in the model so genetic effects are now defined at level 2.

*Multivariate response models.* We can fit multivariate response models within a multilevel framework. We have multiple responses available in the NEAD data set we have been analysing. So we give an example of a simple multilevel, multivariate model with genetic effects here. We use as responses depression and antisocial scores (a more extensive report of these data are written up elsewhere in O'Connor et al., 1998). The basic multivariate model can be written as

$$\begin{aligned}
y_{1i} &= \beta_1 + e_{1i} \\
y_{2i} &= \beta_2 + e_{2i} \\
\begin{bmatrix} e_{1i} \\ e_{2i} \end{bmatrix} &\sim N(0, \Omega_e) \quad \Omega_e = \begin{bmatrix} \sigma_{e1}^2 & \\ \sigma_{e12} & \sigma_{e2}^2 \end{bmatrix}
\end{aligned}$$

where antisocial behaviour is response 1 and depression is response 2. This simply estimates a mean for each response and their covariance matrix:  $\text{var}(\text{antis}) = 0.297(0.01)$ ,  $\text{var}(\text{depression}) = 0.276$ ,  $\text{cov}(\text{antis}, \text{depr}) = 0.113$ . This model (Goldstein, 2003) handles missing responses, whereas conventional multivariate response estimation procedures require incomplete cases to be dropped. Also, we can add covariates and add further structure to the covariance matrix, for example shared environment and genetic random effects :

$$\begin{aligned}
y_{1ij} &= \beta_1 + u_{1j} + e_{1ij} + g_{1ij} \\
y_{2ij} &= \beta_2 + u_{2j} + e_{2ij} + g_{2ij} \\
\begin{bmatrix} u_{1j} \\ u_{2j} \end{bmatrix} &\sim N(0, \Omega_u) \quad \Omega_u = \begin{bmatrix} \sigma_{u1}^2 & \\ \sigma_{u12} & \sigma_{u2}^2 \end{bmatrix} \\
\begin{bmatrix} e_{1ij} \\ e_{2ij} \end{bmatrix} &\sim N(0, \Omega_e) \quad \Omega_e = \begin{bmatrix} \sigma_{e1}^2 & \\ \sigma_{e12} & \sigma_{e2}^2 \end{bmatrix} \quad (6) \\
\begin{bmatrix} g_{1ij} \\ g_{2ij} \end{bmatrix} &\sim N(0, \Omega_g) \quad \Omega_g = \begin{bmatrix} \sigma_{g1}^2 & \\ \sigma_{g12} & \sigma_{g2}^2 \end{bmatrix} \\
\text{cov}(y_{1i_1j}, y_{2i_2j}) &= \sigma_{u12} + r_{(i_1, i_2)j} \sigma_{g12}
\end{aligned}$$

The results for model shown in equation 6 are given in Table 3. The genetic correlation between antisocial behavior and depression is  $\rho_{12}^{(g)} = \frac{0.069}{\sqrt{0.180 * 0.182}} = 0.38$ . Each of these variance/covariance parameters can also be modelled in terms of child or family level variables.

*Inclusion of arbitrarily complex intergenerational pedigree information using a parsimonious model formulation.* The estimation procedure used in this paper, and described in an appendix, subsumes all the biological relationship information in an arbitrary complex

pedigree, potentially running over generations into a single design matrix. This allows the additive genetic variation to be estimated as a single parameter. Of course, if we want to include gene-environment interactions and dominance effects, further design matrices are required and the framework laid out in the appendix readily extends to handle these model elaborations.

*Analysis of dyadic directional relationship data*

Often in psychological studies relationship data take the form of directional measurements on dyads. For example, the amount of aggression *from* individual A *to* individual B. The same structures occur in social network analysis and Snijders and Kenny (1999) develop a cross-classified multilevel model for handling these structures. In this section we apply the Snijders and Kenny model to directional scores from the NEAD data set and extend the model to handle genetic effects.

Each family contained two parents and two children. Families were videotaped interacting in 10-minute problem-solving sessions. Observers analysed the videotapes using a reliable observational rating system (described fully in O'Connor et al., 1995). Observers rated each individual's behavior (e.g., mother behavior to older child and older child behavior to mother). Here we analyse composite negativity measures based on observer codings of anger and coercion.

For each family we have 12 directed measurement of negativity:  $c1 \rightarrow c2$ ,  $c1 \rightarrow m$ ,  $c1 \rightarrow f$ ,  $c2 \rightarrow c1$ ,  $c2 \rightarrow m$ ,  $c2 \rightarrow f$ ,  $m \rightarrow c1$ ,  $m \rightarrow c2$ ,  $m \rightarrow f$ ,  $f \rightarrow c1$ ,  $f \rightarrow c2$ ,  $f \rightarrow m$  (where  $c1$  = child1 [older child],  $c2$  = child2 [younger child],  $m$  = mother and  $f$  = father). Snijders and Kenny describe behaviour from *actors* to *partners*. For example, in the directed score  $m \rightarrow c1$ , mother is the actor and child 1 is the partner. The data can be classified according to actor and partner, it can also be classified according to dyad:

c1→c2, c2→c1:dyad 1; c1→m, m→c1:dyad 2; c1→f, f→c1:dyad 3;

c2→m, m→c2 :dyad 4; c2→f, f→c2 :dyad 5; m→f, f→m:dyad 6;

The relationships between the measurements and the actor, partner, dyad, and family classifications are shown in Figure 2. The negativity measurements lie within a three-way classification of actor by partner by dyad; this cross-classification is nested within family.

Data are clustered by partner, actor, dyad and family. All of these classifications are potentially important from a technical (failure to model them leads to incorrect standard errors for fixed effects) and a substantive point of view. We can write down the multilevel model for this structure as:

$$\begin{aligned}
 y_{i(j,k,l)m} &= \beta_0 + f_m + a_{jm} + p_{km} + d_{lm} + e_{i(j,k,l)m} \\
 f_m &\sim N(0, \sigma_f^2) \quad a_{jm} \sim N(0, \sigma_a^2) \quad p_{km} \sim N(0, \sigma_p^2) \\
 d_{lm} &\sim N(0, \sigma_d^2) \quad e_{i(j,k,l)m} \sim N(0, \sigma_e^2) \\
 \text{cov}(a_{jm}, p_{km}) &= \sigma_{ap} \quad : j = k \\
 \text{cov}(a_{jm}, p_{km}) &= 0 \quad : j \neq k
 \end{aligned} \tag{7}$$

where  $j$  indexes actor,  $k$  indexes partner,  $l$  indexes dyad and  $m$  indexes family and  $i$  is the measurement level.  $(j,k,l)$  denotes cross-classification of classifications  $j,k$  and  $l$ . We also have random effects for family  $m$  ( $f_m$ ), actor  $j$  in family  $m$  ( $a_{jm}$ ), partner  $k$  in family  $m$  ( $p_{km}$ ), dyad  $l$  in family  $m$  and a residual random effect  $e_{i(j,k,l)m}$ . Note that the model estimates a covariance between actor and partner effects for the same individual, however the covariance between actor effect and partner effect is set to zero where the actor and partner are different individuals.

For a family with 3 members this gives rise to the covariance structure shown in figure 3. The NEAD data has families with 4 members and therefore a 12 by 12 covariance matrix for the relational scores, for the sake of compact illustration figure 3 gives the covariance pattern for a family with three members and therefore 6 relational scores. The

results of the model specified in equation (7) are given in model 1 of table 4. The results are given in model 1 of table 4.

The variance of random effects for all classifications are significant. The proportion of the total variance attributable to each classification is family (0.14), actor (0.19), partner (0.09), dyad (0.33) and residual directed score variance (0.24).

The actor effects pick up the extent to which individuals are consistently negative *to all* other family members; partner effects pick up the extent to which individuals consistently receive negativity *from all* other family members. The family effect picks up the extent to which a family contributes to all relationships in the family. It is worth emphasizing the family level effect which assesses “shared” or family-wide effects common to all the relationships within a family. In this case, the findings suggest that, even after accounting for lower level effects at the individual and dyadic level, families differ from one another in their level of negativity.

The dyad effect represents the extent to which individuals within a dyad behave similarly to each other. Snijders and Kenny (1999) point out that the correlation

$\sigma_d^2 / (\sigma_d^2 + \sigma_e^2)$  can be thought of as the **reciprocity** correlation which, after having removed family, actor and partner effects, gives a measure of the correlation between *a*'s behaviour to *b* and *b*'s behaviour to *a*. In this case the reciprocity correlation is 0.58

indicating that within a dyad negative behaviour stimulates a negative response. In other situations where, for example, dominant and placating roles are enacted within dyads, a negative reciprocity correlation would be expected. The actor/partner correlation

$\sigma_{ap} / (\sigma_a \sigma_p) = 0.52$  indicates that individuals who act negatively in their relationships across *all* the other family members also *elicit* a shared high amount of negativity from other family members.

We now consider including genetic effects to the model. Bussell et al (1999) conduct a genetic analysis on a subset of the negativity measurements using a bivariate structural equation model to explore adolescent relationships to siblings and mothers. They use 4 of the twelve directed measurements in their analysis. The first trait they consider is negativity to sibling, the  $c1 \rightarrow c2$  and  $c2 \rightarrow c1$  measures. The second trait they consider is mother's negativity to adolescent, the  $m \rightarrow c1$  and  $m \rightarrow c2$  measurements. This second trait they regard as being a measurement of the child's ability to elicit negativity from the mother, that is, a partner effect. In their conceptualisation both traits are measurements on the children and genetic correlations and cross-correlations can be estimated based on the relationship between the two children being measured. In their analysis they make no separation between actor and partner effects and dyad effects are also not included in the model. They find a large shared environment component of variation and moderate non-shared, additive genetic and dominance genetic components of variation.

We can extend the model given in equation (7) to include additive and dominance genetic effects:

$$y_{i(j,k,l)m} = \beta_0 + f_m + a_{jm} + p_{km} + g_{jm} + g_{km} + d_{lm} + e_{i(j,k,l)m}$$

$$f_m \sim N(0, \sigma_f^2) \quad a_{jm} \sim N(0, \sigma_a^2) \quad p_{km} \sim N(0, \sigma_p^2)$$

$$d_{lm} \sim N(0, \sigma_d^2) \quad e_{i(j,k,l)m} \sim N(0, \sigma_e^2)$$

$$g_{jm} \sim N(0, \sigma_{A_j}^2 + \sigma_{D_j}^2)$$

$$g_{km} \sim N(0, \sigma_{A_k}^2 + \sigma_{D_k}^2)$$

$$\text{cov}(a_{jm}, p_{km}) = \sigma_{ap} \quad : j = k$$

$$\text{cov}(a_{jm}, p_{km}) = 0 \quad : j \neq k$$

Where the actor effect for individual  $j$  in family  $m$  is divided into two parts and environmental effect,  $a_{jm}$  and a genetic effect,  $g_{jm}$ . Likewise the partner effect for individual  $k$  in family  $m$  is divided into two parts, an environmental effect,  $p_{km}$  and a genetic effect  $g_{km}$ . The actor and partner effects represent different behaviours with separate genetic variances. The actor and partner genetic variances are further decomposed into additive and dominance components.

The covariance between two relationship measurements, in family  $m$ , one with actor  $j_1$  and partner  $k_1$  in dyad  $l_1$  and the other with actor  $j_2$  and partner  $k_2$  in dyad  $l_2$  is

$$\begin{aligned} \text{COV}(y_{i_1(j_1, k_1)m}, y_{i_2(j_2, k_2)m}) = \\ \text{COV}(f_m + a_{j_1m} + p_{k_1m} + g_{j_1m} + g_{k_1m} + d_{l_1m} + e_{i_1(j_1, k_1, l_1)m}, \\ f_m + a_{j_2m} + p_{k_2m} + g_{j_2m} + g_{k_2m} + d_{l_2m} + e_{i_2(j_2, k_2, l_2)m}) = \\ \sigma_f^2 + \sigma_a^2 \text{same}(j_1, j_2) + \sigma_p^2 \text{same}(k_1, k_2) + \sigma_{ap} \text{same}(j_1, k_2) + \sigma_{ap} \text{same}(j_2, k_1) \\ + \sigma_d^2 \text{same}(l_1, l_2) + \text{COV}(g_{j_1m} + g_{k_1m}, g_{j_2m} + g_{k_2m}) \\ \text{where } \text{same}(b, c) = 1 : b = c; 0 : b \neq c \end{aligned}$$

Expanding the term for the genetic covariance we have

$$\text{COV}(g_{j_1m} + g_{k_1m}, g_{j_2m} + g_{k_2m}) = r_{j_1j_2}^{(a)} \sigma_{A_j}^2 + r_{j_1j_2}^{(d)} \sigma_{D_j}^2 + r_{k_1k_2}^{(a)} \sigma_{A_k}^2 + r_{k_1k_2}^{(d)} \sigma_{D_k}^2$$

where  $\sigma_{A_j}^2, \sigma_{D_j}^2, \sigma_{A_k}^2, \sigma_{D_k}^2$  are the additive and dominance actor genetic variance components and the additive and dominance partner genetic variance components respectively. The values for the additive and dominance relationship coefficients are those given from standard population genetics theory and are listed in Table 5.

The results are shown in model 2 of table 4. The zero estimates for the actor and partner additive genetic variances come about because they are estimated as negative (but non-significant) and are reset to zero. The total variance in model 1 and model 2 are

effectively identical 0.737 and 0.735. In model 1, the actor variance ( $\sigma_a^2$ ) is 0.130. In model 2, when we introduce genetic effects, the actor variance is splits into two components, environmental ( $\sigma_a^2=0.055$ ) and genetic ( $\sigma_{Dj}^2=0.077$ ). Likewise, the partner variance in model 1 ( $\sigma_p^2$ ) is 0.064, splits almost exactly into environmental ( $\sigma_p^2$  0.014) and genetic ( $\sigma_{Dj}^2=0.049$ ) components in model 2. The analyses of these data indicate that individuals do have common actor and partner effects across all their relationships within a family. The actor effects are stronger than the partner effects. That is, the actor variance components are larger than the partner variance components. Both an individual's propensity to act negatively in relationships and individual's propensity to elicit negativity in relationships have a genetic component. In addition, there are family level factors that affect the quality of all the relationships in a family. For relationship negativity, the dyad is the single most important classification, 31% of the total variability in relationship quality is attributable to dyad level factors.

*Comparison with existing work that has used multilevel techniques to estimate genetic effects*

Guo and Wang (2002) and van den Oord (2001) also use multilevel models to estimate genetic effects. Guo and Wang work in terms of heritabilities and van den Oord in terms of the additive genetic variance. Both of these methods work by estimating different variance components and therefore different correlations for DZ twins, MZ twins, full sibs, half sibs and cousins. From this set of correlations they construct a set of simultaneous equations (in  $h^2$  or  $\sigma_g^2$ ), the solution of which is the desired genetic parameter. Guo and Wang operationalize the model as a two level model (children within families) with dummy variables for MZ and DZ twins, full siblings, half-siblings and cousins defined as random at the family level. These dummy variables are constructed in such a way that the set of



correlations (which are derived from the set of family level variance components) have the following ordering :  $\rho_M > \rho_D = \rho_F > \rho_H > \rho_C$ .

Van den Oord operationalizes the problem by setting up a multilevel model with the following structure:

Level 4: all individuals in a level 4 unit are related at least as cousins

Level 3: all individuals in a level 3 unit are related at least as half-sibs

Level 2: all individuals in a level 2 unit are related as MZ twins, DZ twins or full sibs.

Level 1: individuals

A dummy variable specifying MZ twins is made random at level 2 to distinguish the MZ twin variance component from the DZ twin/full sib variance component. It is not clear how adopted children fit into this framework. Interestingly, in sample used in the van den Oord study, some families included more than 2 children, although in the analysis a maximum of two children per family were included in analyses.

Our method differs in two respects. Firstly, both these methods calculate the genetic parameter from the set of differential correlations derived from variance components for individuals of different degrees of relatedness. Van den Oord estimates four higher level variance components and Guo and Wang estimate five additional variance components (since they estimate different variance components for full-sibs and DZ twins). Both these methods would require still more parameters for more complex pedigree structures. Guo and Wang acknowledge that the technique of calculating the genetic parameter from several model estimated parameters, rather than estimating it directly is “an important issue that needs to be dealt with”, since hypothesis testing of such derived quantities is not straightforward. Our method estimates the additive genetic variance (and the shared and non-shared environment components) *directly*, irrespective of the complexity of the pedigree. This more straight-

forward formulation extends more readily to modelling gene-environment interactions with continuous or categorical environmental variables of the type demonstrated in this paper and to modelling other types of genetic effects, for example, dominance effects.

Guo and Wang and Van Oord do not include genetic effects in their statement of the basic model as we do in equation (2) where we include  $g_{ij}$  the genetic effect for the child  $i$  in family  $j$ . Specifying genetic effects as random effects with given covariance assumptions is in line with the random effects literature, which spans many disciplines and therefore is a useful formulation to communicate these models to a wider audience. Purcell (2002) takes a similar approach within a variance components, structural equation modelling framework. Purcell estimates genetic, shared environment and non-shared environment variance components directly and uses a model that allows any of these variance components to be functions of measured environmental variances. However, Purcell only develops models for twin studies rather than extended families.

### Discussion

In this paper we present a parsimonious method for including genetic effects in multilevel models and, more broadly, for testing multiple conceptual issues in family research. Analysts may prefer these techniques over traditional structural equation genetic models in data sets which have a complex population structure of the kinds detailed in sections 6 and 7.

Key features of this approach include the ability to handle complex population structures – not only children within families, but also families within neighborhoods, observations within children, or a potentially large number of crossed or nested random classifications (Browne et al, 2001). In addition, the model is flexible with regard to the numbers of members per family (e.g., it can handle samples in which families have different

numbers of children) and genetic relatedness to one another, allowing for “extended family” designs in which genetic information is based not only on siblings but also parent-child relationships. A further advantage of the approach is that it estimates genetic and environmental variance directly. This creates improves power for testing gene-environment interactions in which the amount of genetic variance is modelled as a function of an explanatory variable. Also this approach can be used to test hypotheses about family-level processes and effects for individuals or dyadic data. Finally, the model provides an operationalization of “shared,” common, or family-wide effects from risks within the family that is an alternative to that suggested in behavioural genetics modelling. That is significant insofar as several studies that have used multilevel modelling framework have found considerable evidence that some risks, such as parental conflict and parental divorce, have family-wide effects on children’s behavioural/emotional development (e.g., Jenkins et al., in press). Similarly, as we demonstrate above, the model may be similarly more sensitive to gene-environment interactions than alternative methods or typical practice using available methods.

We would not suggest that this analytic framework is the only manner of modelling family data, although its particular profile of strengths may make it especially amenable to apply to clustered family data and to test family theories. At a minimum, further consideration of how multilevel approaches complement and extend – and are extended by – structural equation approaches warrants further examination and development. Van den Oord (2001) also pointed out that a convergence of multilevel and structural equation modelling techniques is occurring, with multilevel modelling packages such as HLM and *MLwiN* (Rasbash et al., 2000) starting to incorporate factor analytic and structural equation models and structural equation modelling packages such as LISREL (Joreskog & Sorbom, 1996),

Mplus (Muthen & Muthen, 1998) and Mx (Neale *et al.*, 1999) incorporating hierarchical structures.

Collating data from multiple family studies with differing structures and complexity will provide a testing ground for replication of conceptual hypotheses. In addition, coordinating data analyses from multiple studies may also identify which sorts of integrated approaches may be most likely to encourage family researchers to tackle questions concerning “family effects” in a more rigorous manner.

## References

- Bussell, D., Neiderhiser J., Pike A., Plomin A., Simmens S., Howe G., et al. (1999). Adolescents' relationships to siblings and mothers: A multivariate genetic analysis. *Developmental Psychology*, 35(5), 1248-1259.
- Browne, W. J., Goldstein, H. and Rasbash, J. (2001). Multiple membership multiple classification (MMMC) models. *Statistical Modelling* 1: 103-124
- Goldstein, H. (2003). *Multilevel statistical models* (3<sup>rd</sup> ed.). London: Edward Arnold.
- Goldstein, H., & Rasbash J. (1993). Efficient computational procedures for estimating parameters in multilevel models based on iterative generalised least squares. *Computational Statistics and Data Analysis*, 13, 63-71.
- Guo G. & Wang W. (2002). The mixed or Multilevel Model for Behaviour Genetic Analysis. *Behaviour Genetics*, 32(1), 37-49.
- Hetherington, E. M. & Clingempeel, W. G. (1992). Coping with marital transitions: A family systems perspective. *Monographs of the Society for Research in Child Development*, 57 (2-3, Serial No. 227).
- Jenkins, J.M., Dunn, J., O'Connor, T.G., Rasbash, J., & Behnke, P. (in press). The mutual influence of marital conflict and children's behavior problems: Shared and non-shared family risks. *Child Development*.
- Jenkins, J.M., Rasbash, J., & O'Connor, T.G. (2003). The role of shared family context in differential parenting. *Developmental Psychology*, 39, 99-113.
- Joreskog, K., & Sorbom D. (1996). *LISREL 8: User's reference guide*. Chicago: Scientific Software International.

- Kovacs, M. (1983). *Children's Depression Inventory: A self-rated depression scale for school-aged youngsters*. Unpublished manuscript, University of Pittsburgh School of Medicine.
- Longford N. T. (1995). *Multilevel statistical models*. New York: Oxford University Press.
- NLSCY. (1995). *User's handbook and microdata guide*. Ottawa, Canada: Statistics Canada & Human Resources Canada.
- O'Connor, T. G., Hetherington, E. M., Reiss, D., & Plomin, R. (1995). A twin-sibling study of observed parent-adolescent interactions. *Child Development*, 66, 812-829.
- O'Connor, T., Dunn, J., Jenkins, J., Pickering, K. & Rasbash, J. (2001). Family settings and children's adjustment: differential adjustment within and across families. *The British Journal of Psychiatry*, 179, 110-115.
- O'Connor, T. G., McGuire, S., Reiss, D., Hetherington, E. M., & Plomin, R. (1998). Co-occurrence of depressive symptoms and antisocial behavior in adolescence: A common genetic liability. *Journal of Abnormal Psychology*, 107, 27-37.
- Paterson, G. R. (1982). *Coercive family process*. Eugene, OR: Castalia.
- Pike A., McGuire S., Hetherington E. M., Reiss D., & Plomin R. (1996). Family environment and adolescent depressive symptoms and antisocial behaviour: A multivariate genetic analysis. *Developmental Psychology*, 32(4), 590-603.
- Purcell S. (2002). Variance components models for gene-environment interactions in twin analysis. *Twin Research*, 5(6), 554-571.
- Radloff, L. S. (1977). The CES-D Scale. A self report depression scale for research in the general population. *Applied Psychological Measurement*, 1, 385-401.
- Rasbash, J., Browne, W., Goldstein., H., Yang, M., et al. (2000). *A user's guide to MLwiN* (2<sup>nd</sup> ed.). London: Institute of Education.

- Raudenbush, S.W. & Bryk, A. S. (2001). *Hierarchical linear modelling* (2<sup>nd</sup> ed.). Newbury Park, CA: Sage.
- Raudenbush, S.W., Brennan, R.T., & Barnett, R.C. (1995). A multivariate hierarchical model for studying psychological change within married couples. *Journal of Family Psychology, 9*(2), 161-174.
- Reiss, D., Plomin, R., Hetherington, E. M., Howe G., Rovine, M., Tyron, A., et al. (1994). The separate social worlds of teenage siblings: An introduction to the study of the nonshared environment and adolescent development. In Hetherington, E.M., Reiss D. & Plomin R. (Eds.), *Separate social worlds of siblings: Impact of the nonshared environment on development* (pp. 63-110). Hillsdale, NJ: Erlbaum.
- Raudenbush, S.W., Bryk, A.S., Cheong, Y.F., & Congdon Jr., R.T. (2001). HLM 5: Hierarchical linear and nonlinear modelling. Chicago: Scientific Software International.
- Snijders, T. & Bosker, R. (1998). *Multilevel analysis: An introduction to basic and advanced multilevel modelling*. London: Sage.
- Snijders, T. A. B. (1995). Multilevel models for family data. In Hox, J. J., van der Meulen, B. E., Janssens, J. M. A. M., ter Laak, J.J.F., & Tavecchio, L.W.C. (Eds.), *Advances in Family Research* (pp. 193-208). Amsterdam: Thesis Publisher.
- Strauss, M. A. (1979). Measuring intrafamily conflict and violence: The Conflict Tactics (CT) Scale. *Journal of Marriage and the family, 41*, 75-85.
- van den Oord. (2001). Estimating effects of latent and measured genotypes in multilevel models. *Statistical Methods in Medical Research, 10*, 393-407.
- Zill, N. (1985). Behaviour problem scales developed from the 1981 Child Health Supplement to National Health Interview Survey. Washington, DC: Child Trends.

## Appendix

*Estimation Algorithm for including arbitrarily complex pedigree information in multilevel models*

The algorithm used to estimate the models in this paper is a minor adaption of the IGLS(iterative generalised least squares) algorithm described in full in Goldstein(1986) and a computationally efficient algorithm is given in Goldstein and Rasbash(1993). For brevity and clarity we describe how the IGLS algorithm works for the two level model

$$y_{ij} = (XB)_{ij} + u_j + e_{ij} \quad (6)$$

$$u_j \sim N(0, \sigma_u^2) \quad e_{ij} \sim N(0, \sigma_e^2)$$

We then describe how genetic effects are included. In (6)

$$V = \text{cov}(Y | XB) = J\sigma_u^2 + I\sigma_e^2$$

Where  $J$  is a block diagonal matrix of ones. For illustration think of two families, each with two kids :

$$V = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix} \sigma_u^2 + \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \sigma_e^2 \quad (7)$$



If  $V$  is known then the usual GLS estimator of  $B$  is

$$B = (X^T V^{-1} X)^{-1} X^T V^{-1} Y \quad (8)$$

If  $V$  is not known estimate (1) assuming  $V = I$ . Now we can get an estimate of  $\theta = \{\sigma_u^2, \sigma_e^2\}$

and therefore  $V$  by forming:

$$\tilde{y}_{ij} = y_{ij} - (XB)_{ij} \quad \tilde{Y} = \{\tilde{y}_{ij}\}$$

The expected value of  $\tilde{Y}\tilde{Y}^T$  is  $V$ . We can form the linear model

$$vec[\tilde{Y}\tilde{Y}^T] = vec[J]\sigma_u^2 + vec[I]\sigma_e^2 + F \quad (9)$$

where  $F$  is a residual term. Showing the elements in (9) explicitly, we have :

$$vec \begin{bmatrix} \tilde{y}_{11}\tilde{y}_{11} & \tilde{y}_{11}\tilde{y}_{21} & \tilde{y}_{11}\tilde{y}_{31} & \tilde{y}_{11}\tilde{y}_{41} \\ \tilde{y}_{21}\tilde{y}_{11} & \tilde{y}_{21}\tilde{y}_{21} & \tilde{y}_{21}\tilde{y}_{31} & \tilde{y}_{21}\tilde{y}_{41} \\ \tilde{y}_{31}\tilde{y}_{11} & \tilde{y}_{31}\tilde{y}_{21} & \tilde{y}_{31}\tilde{y}_{31} & \tilde{y}_{31}\tilde{y}_{41} \\ \tilde{y}_{41}\tilde{y}_{11} & \tilde{y}_{41}\tilde{y}_{21} & \tilde{y}_{41}\tilde{y}_{31} & \tilde{y}_{41}\tilde{y}_{41} \end{bmatrix} = vec \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix} \sigma_u^2 + vec \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \sigma_e^2 + F \quad (10)$$

where the response is modelled by two explanatory variables with coefficients  $\{\sigma_u^2, \sigma_e^2\}$  to be

estimated. This can be estimated by GLS where  $cov(vec[\tilde{Y}\tilde{Y}^T]) = V^{-1} \otimes V^{-1}$ . That is

$$\theta = (Z^{*T} (V^{-1} \otimes V^{-1}) Z^*)^{-1} Z^{*T} (V^{-1} \otimes V^{-1}) Y^*$$

(11)

$$Z^* = \{vec[J], vec[I]\}$$

$$Y^* = vec[\tilde{Y}\tilde{Y}^T]$$

Estimation proceeds by switching between (8) and (11). Variances and covariances between further sets of random effects can be included by adding extra columns into the  $Z^*$  matrix of the appropriate structure. Typically columns of  $Z^*$  are formed from cross-products of explanatory variables that have random coefficients. In (6) only the intercept has a random coefficient (at level 1 and level 2) hence the  $Z^*$  design vectors contain only ones and zeros. The basic multilevel model assumes, within a set of random effects, that effects are independent. That is

$$cov(u_{j_1}, u_{j_2}) = 0$$

$$cov(e_{i_1j}, e_{i_2j}) = 0$$

In some cases, for example spatial or time series models this assumption is relaxed. In time series models the correlation between two measurements made on the same individual at different points in time is modelled as function of the distance in time between the measurements. With genetic effects we have an analogous situation, where the genetic correlation between two individuals is modelled as a function of the *genetic distance* separating the two individuals. Given the basic multilevel model including genetic effects :

$$y_{ij} = (XB)_{ij} + u_j + e_{ij} + g_{ij}$$

$$\text{var}(y_{ij}) = \sigma_u^2 + \sigma_e^2 + \sigma_g^2$$

$$\text{cov}(y_{i_1j}, y_{i_2j}) = \sigma_u^2 + r_{i_1i_2} \sigma_g^2$$

With our example “toy” data set of two families( $j$ ) each of two kids( $i$ ) containing FS and HS, respectively we can construct :

$j$	$i$	$R_j$
1	1	$\begin{bmatrix} 1 \\ \end{bmatrix}$
1	2	$\begin{bmatrix} 0.5 & 1 \\ \end{bmatrix}$
2	1	$\begin{bmatrix} 1 \\ \end{bmatrix}$
2	2	$\begin{bmatrix} 0.25 & 1 \\ \end{bmatrix}$

Equation (7) and (9) now become:

$$V = \begin{bmatrix} 1 & & & \\ 1 & 1 & & \\ 0 & 0 & 1 & \\ 0 & 0 & 1 & 1 \end{bmatrix} \sigma_u^2 + \begin{bmatrix} 1 & & & \\ 0 & 1 & & \\ 0 & 0 & 1 & \\ 0 & 0 & 0 & 1 \end{bmatrix} \sigma_e^2 + \begin{bmatrix} 1 & & & \\ 0.5 & 1 & & \\ 0 & 0 & 1 & \\ 0 & 0 & 0.25 & 1 \end{bmatrix} \sigma_g^2$$

and

$$\text{vec}[\tilde{Y}\tilde{Y}^T] = \text{vec}[J]\sigma_u^2 + \text{vec}[I]\sigma_e^2 + \text{vec}[R]\sigma_g^2 + F$$

Otherwise estimation proceeds as outlined above. Gene environment interactions are estimated by adding a further column(s) to  $Z^*$  which are the product of the environmental variable(s) and  $R$ .

Author Note

We would like to thank the UK Economic and Social Research Council's Research Methods Programme for funding the work leading to the methodological developments described in this paper.

Address for correspondence: Jon Rasbash, Institute of Education, 20 Bedford Way, London WC1H 0AL UK. Email: [j.rasbash@ioe.ac.uk](mailto:j.rasbash@ioe.ac.uk).

Table 1a

## Variance Components for Adolescent Depression

Parameter	Model	
	Model 1	Model 2
Fixed	estm (se)	estm (se)
Intercept( $\beta_0$ )	0.008(0.017)	0.02(0.017)
Random		
Shared env( $\sigma_u^2$ )	0.086(0.011)	0.018(0.017)
Non-shared env( $\sigma_e^2$ )	0.198(0.011)	0.690(0.010)
Genetic( $\sigma_g^2$ )	-	0.209(0.022)
Deviance	2165.88	2129.2

Table 1b

## Variance Components for Adolescent Depression Including Covariates

Parameter	Model 3
Fixed	estm (se)
Intercept	-0.285(0.087)
Age	0.011(0.006)
Maternal negativity	0.157(0.024)
Paternal negativity	0.216(0.026)
Girl	0.158(0.028)
Stepfamily	0.105(0.029)
Random	
Shared env( $\sigma_u^2$ )	0.0035(0.014)
Non-shared env( $\sigma_e^2$ )	0.70(0.096)
Genetic( $\sigma_g^2$ )	0.148(0.020)
Deviance	1780.95

Table 1c

Variance Components for Adolescent Depression Adding Complex Variation and Gene-Environment Interactions

Parameter	Model	
	Model 4	Model 5
Fixed	estm (se)	estm (se)
Intercept	-0.273(0.080)	-0.232(0.079)
Age	0.011(0.005)	0.009(0.005)
Mat_neg	0.170(0.024)	0.166(0.024)
Pat_neg	0.210(0.028)	0.214(0.27)
Girl	0.159(0.027)	0.161(0.028)
Stepfam	0.097(0.028)	0.089(0.028)
Random		
Shared env		
$\alpha_0^{(u)}$	0.0006(0.014)	0.054(0.012)
$\alpha_1^{(u)}$	-0.017(0.019)	-
Non-shared env		
$\alpha_0^{(e)}$	0.073(0.009)	0.071(0.009)
$\alpha_1^{(e)}$	0.0078(0.010)	-
Genetic		
$\alpha_0^{(g)}$	0.155(0.021)	0.131(0.029)
$\alpha_1^{(g)}$	0.093(0.023)	0.077(0.013)



$\alpha_2^{(g)}$ 

0.055(0.017)

Deviance

1740.42

1730.42

Table 2

Decomposing Variation of Child Aggression in NLSCY.

Parameter	Model		
	model 1	model 2	model 3
<b><i>Fixed</i></b>	estm (se)	estm (se)	estm (se)
Intercept	0.835(0.008)	0.836(0.008)	0.524(0.072)
Child age			-0.275(0.0027)
Boy			0.169(0.013)
SES			-0.038(0.010)
Maternal depression			0.011(0.001)
Violent home			0.147(0.017)
Parental negativity			0.091(0.002)
<b><i>Random</i></b>			
Shared env( $\sigma_u^2$ )	0.270(0.011)	0.089(0.083)	0.057(0.065)
Non-shared env( $\sigma_e^2$ )	0.433(0.009)	0.242(0.086)	0.199(0.068)
Genetic( $\sigma_g^2$ )		0.373(0.168)	0.298(0.134)
Deviance	32294.2	32286.8	29109.00.835

Table 3

Results for a Multilevel, Multivariate Response with Genetic Effects

	Variance (antisocial)	Variance (depression)	Covariance (antisocial, depression)
Family	0.068(0.017)	0.025(0.017)	0.027(0.013)
Individual	0.051(0.008)	0.078(0.011)	0.020(0.007)
Genetic	0.180(0.019)	0.181(0.025)	0.069(0.016)
Total	0.299	0.284	0.116

Table 4

Results from Models for Dyadic Relationship Data with Genetic Effects

Parameter	Model	
	Model 1	Model 2
	estm (se)	estm (se)
intercept( $\beta_0$ )	2.85(0.018)	2.85(0.018)
family variance( $\sigma_f^2$ )	0.099(0.015)	0.096(0.015)
actor variance( $\sigma_a^2$ )	0.130(0.011)	0.055(0.015)
partner variance( $\sigma_p^2$ )	0.064(0.010)	0.014(0.013)
actor, partner covariance( $\sigma_{ap}$ )	0.047(0.01)	0.045(0.010)
dyad variance( $\sigma_d^2$ )	0.230(0.013)	0.232(0.013)
directed score variance( $\sigma_e^2$ )	0.167(0.005)	0.167(0.005)
additive actor genetic( $\sigma_{Aj}^2$ )		0.0
dominance actor genetic( $\sigma_{Dj}^2$ )		0.077(0.013)
additive partner genetic( $\sigma_{Ak}^2$ )		0.0
dominance partner genetic( $\sigma_{Dk}^2$ )		0.049(0.011)
-2log like	17669.7	17595.0

Table 5

## Additive and Dominance Relationship Coefficients

Relationship between individuals $b$ and $c$	$r_{(b,c)}^{(a)}$	$r_{(b,c)}^{(d)}$
Parent –offspring	1 / 2	
Half-sibs	1 / 4	
Full sibs, DZ	1 / 2	1 / 4
MZ	1	1

## Figure Captions

*Figure 1.* Gene-Environment Interactions for Depression with Paternal Negativity and Gender.

*Figure 2.* Unit Diagram Showing Relationships between Classifications for NEAD Observational Data.

*Figure 3.* Covariance structure arising from equation(7) for a family with three members(1,2 and 3 where 1->2 is the response with actor family member 1 and partner family member 2 and so on);  $f = \sigma_f^2, d = \sigma_d^2, a = \sigma_a^2, p = \sigma_p^2, e = \sigma_e^2, c = \sigma_{ap}$ .

Figure 1

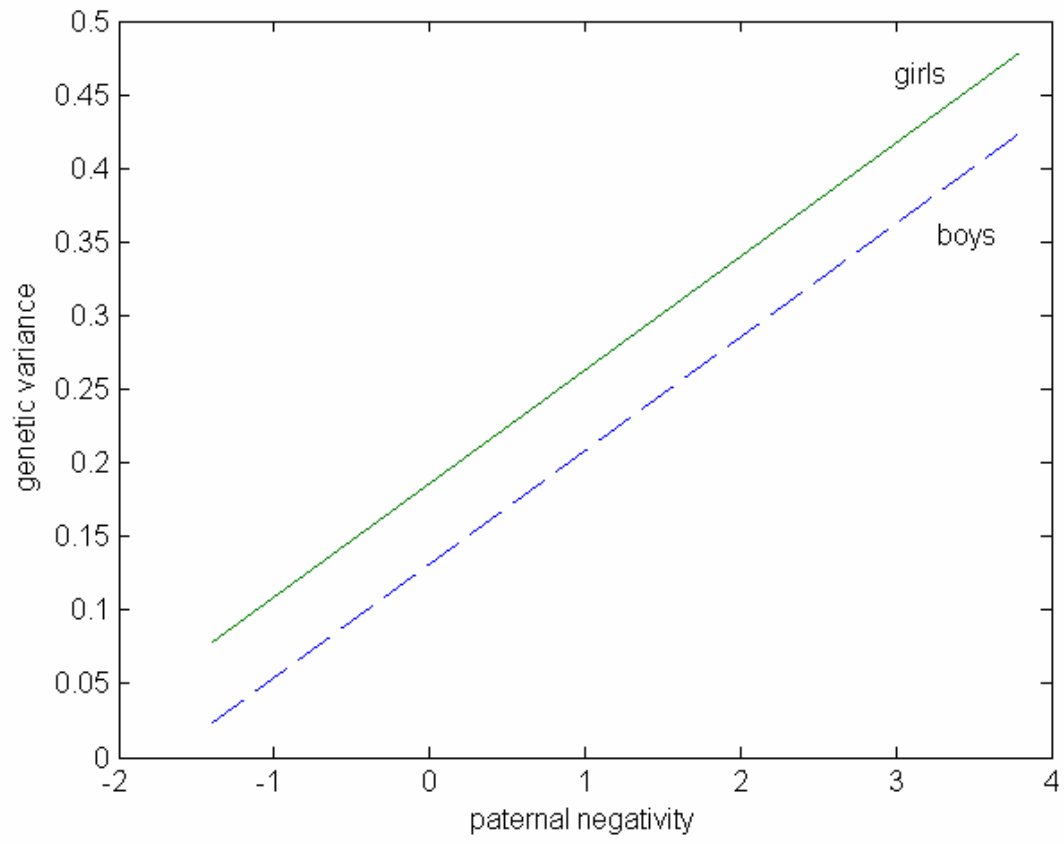


Figure 2

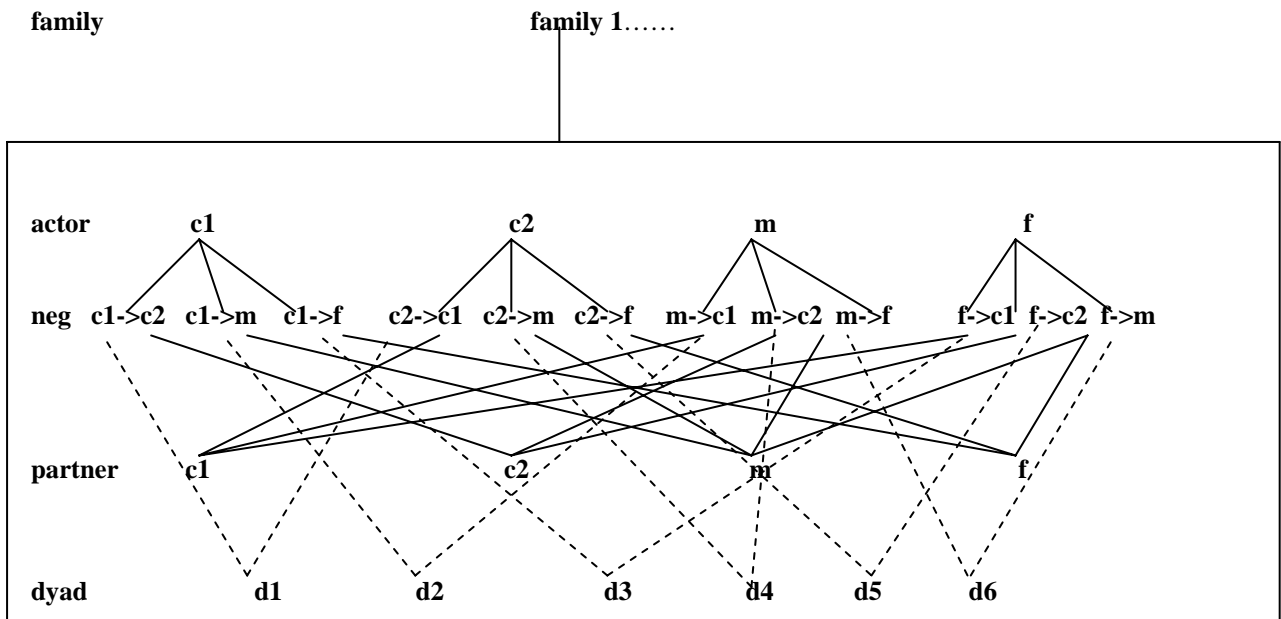




Figure 3

	1->2	2->1	1->3	3->1	2->3	3->2
1->2	$f+d+a+p+e$					
2->1	$f+d+2c$	$f+d+a+p+e$				
1->3	$f+a$	$f+c$	$f+d+a+p+e$			
3->1	$f+c$	$f+p$	$f+d+2c$	$f+d+a+p+e$		
2->3	$f+c$	$f+a$	$f+p$	$f+c$	$f+d+a+p+e$	
3->2	$f+p$	$f+c$	$f+c$	$f+a$	$f+d+2c$	$f+d+a+p+e$

---