# Analytic approaches to twin data using structural equation models

**Frühling V. Rijsdijk and Pak C. Sham**

Date received (in revised form): 13th March 2002

**Fruhling Rijsdijk**
is a lecturer in Quantitative Behavioral Genetics at the Social, Genetic & Development Psychiatry Research Centre of the Institute of Psychiatry, King's College London. She works on the development of statistical methods and computer software for gene-mapping in quantitative traits and the development of structural equation models for (selected) family and twin data.

**Pak Sham**
is Professor of Psychiatric and Statistical Genetics at the Institute of Psychiatry, King's College London. He works on the development of statistical and computational methods for gene mapping in complex disorders, and on the application of these methods to psychiatric conditions such as schizophrenia and affective disorders.

***Keywords:*** *classical twin study, behavioural genetics, Mx, SEM, genetic model fitting*

## Abstract

The classical twin study is the most popular design in behavioural genetics. It has strong roots in biometrical genetic theory, which allows predictions to be made about the correlations between observed traits of identical and fraternal twins in terms of underlying genetic and environmental components. One can infer the relative importance of these 'latent' factors (model parameters) by structural equation modelling (SEM) of observed covariances of both twin types. SEM programs estimate model parameters by minimising a goodness-of-fit function between observed and predicted covariance matrices, usually by the *maximum-likelihood* criterion. Likelihood ratio statistics also allow the comparison of fit of different competing models. The program Mx, specifically developed to model genetically sensitive data, is now widely used in twin analyses. The flexibility of Mx allows the modelling of multivariate data to examine the genetic and environmental relations between two or more phenotypes and the modelling to categorical traits under liability-threshold models.

## INTRODUCTION TO QUANTITATIVE GENETICS

Quantitative genetics is often used to study behaviour, and its development parallels that of behavioural genetics. However, the analyses of non-behavioural traits such as height, body-mass index and brain volume also fall in the scope of the method. The aim of quantitative genetics is to study the relative contribution of genetic and environmental influences to individual differences in traits using family, adoption or twin data or a combination of these different designs. In family studies familial aggregation of a disease or trait is investigated, and, if confirmed, the pattern of disease is used to infer its likely mode of inheritance. A limitation of the family study is its inability to discriminate genetic from shared environmental factors. The reason is that familial resemblance can be due to shared family environment such as diet and social class as well as to shared genes. Adoption studies can overcome this limitation. If adoption occurred early in life, then shared environmental factors may be considered unlikely, and familial resemblance attributed entirely to shared genes. However, adoption data are not easy to obtain, given that information on adoption is often classified. Also, selective placement and prenatal influences can bias adoption data. For these reasons, the classical twin method is the most popular design used in behavioural genetics. The existence of two types of twin pairs, monozygotic (MZ) and dizygotic (DZ), provides a natural experiment for disentangling genetic from environmental factors. In this paper we will focus on the principles and analytical approaches of the twin methodology.

Behavioural genetics is rooted in both psychology and medical sciences (such as psychiatry). Medical sciences traditionally adopt a model where diseases are defined as categorical entities and diagnoses are either present or absent. Psychologists on the other hand prefer quantitative measures of cognitive ability, personality and other traits. The methodology of quantitative genetics and twin research

F. V. Rijsdijk,
SGDP Centre,
Institute of Psychiatry,
111 Denmark Hill,
London SE5 8AF, UK

Tel: +44 (0) 207 848 0890
Fax: +44 (0) 207 848 0866
e-mail: f.rijsdijk@iop.kcl.ac.uk

reflects this duality, although there is now a trend to integrate the two approaches, especially for traits in which both diagnostic criteria and quantitative measures exist (eg depression and anxiety).

## TWIN STUDIES

**Individual differences in a trait can be decomposed into genetic and environmental sources of variance**

The classical twin method uses the information in MZ and DZ twin pairs to disentangle the influences of genetic and environmental factors on a trait.[1–3] There are two main types of twin studies: those based on twin pairs ascertained through affected probands, and those based on population twin registers. The former is appropriate for investigating relatively rare diseases, whereas the latter is better suited for studying common traits and quantitative dimensions. However, they share the basic principles of the twin method, which are described in the next sections.

### Biometrical genetics and the twin method

From biometrical genetic theory it is possible to write structural equations relating observed traits of twins to their underlying genotypes and environments.[4] One can infer the relative importance of these 'latent' factors by comparing the observed correlations (or concordances) between family members with predicted correlations (or concordances) if different sources of genetic and environmental factors were to play a role. The sources of genetic and environmental variation considered in behavioural genetics are as follows:

- Additive genetic influences, **A**, represent the sum of the effects of the individual alleles at all loci that influence the trait.

**Heritability**

- Non-additive genetic influences which represent interactions between alleles at the same locus (dominance, **D**) or on different loci (epistasis).

- Environmental influences shared by family members (common environmental variation, **C**), eg socio-economic status, parenting style, childhood diet or peer influences shared by both adolescent twins.

- Unique environmental influences (**E**) that result in differences among members of one family, eg accidents, differential parental treatment, differential prenatal exposure and measurement error.

The total phenotypic variance, **P**, of a trait is the sum of these variance components ($\mathbf{P} = \mathbf{A} + \mathbf{D} + \mathbf{C} + \mathbf{E}$). To unravel the sources of variance and estimate their contribution, information from genetically informative subjects is essential.

Twin data enable the different variance components to be estimated, because MZ and DZ twins have different degrees of correlation for the genetic components **A** and **D** but the same degrees of correlation for the environmental components **C** and **E**. MZ pairs correlate 1 for both **A** and **D**, whereas DZ pairs correlate $\frac{1}{2}$ and $\frac{1}{4}$ for these components, respectively. Both MZ and DZ pairs correlate 1 for **C** and **E** is uncorrelated for both types of twins. Since the phenotypic differences between MZ twins can only be due to unique environmental influences, this gives us an estimate for **E**. Assuming that MZ and DZ twins experience the same degree of similarity in their environments, any excess of similarity between MZ and DZ twins can be interpreted as due to the greater proportion of genes shared by MZ twins, and thus gives us an estimate for **A**. An estimate for **C** is given by the difference in MZ correlation and the estimated effect of **A**.

#### *Falconer's formula*

Heritability ($h^2$) is an index for the relative contribution of genetic effects to the total phenotypic variance. In the classical twin method, Falconer's formula was used to estimate heritability based on twin correlations: $h^2$ is $2(r_{MZ} - r_{DZ})$,

where $r$ is the intraclass correlation coefficient. The relative contributions of the shared and non-shared environmental effects are: $c^2 = r_{MZ} - h^2$ (or $2r_{DZ} - r_{MZ}$) and $e^2 = 1 - h^2 + c^2$. This approach is not adequate for testing, for example, sex differences and multivariate data and was replaced by a more advanced method using special-purpose software. Covariance structure models are fitted to (multivariate) data from multiple groups (eg MZ and DZ twins) simultaneously by means of maximum likelihood techniques.

**Path diagrams are helpful in depicting twin models**

## Assumptions of the twin method

A number of assumptions are made in the classical twin study. It is important to be aware of the implications of such assumptions and of the extent to which they are realistic in relation to the trait in question. The assumptions include the following:

- MZ and DZ twin pairs share their environments to the same extent.
- Gene–environment correlations and interactions are minimal for the trait.
- Twins are no different from the general population in terms of the trait.
- Matings in the population occur at random (no assortment).

The consequences of violation of these assumptions, and how these might be detected, will be discussed in a later section.

## Path analysis and structural equations

The method of path analysis was first developed by Sewell Wright (1921)[5] with the aim of presenting a method that can be used to interpret observed correlations between a set of variables in terms of an *a priori* model of the causal relations between those variables. In terms of twin studies, a model predicts a series of expectations for correlations between twins based on the hypothesis to be tested (ie whether additive or dominance genetic factors or common environmental factors influence the trait).

The full twin model (for one variable) can be depicted in a path diagram (Figure 1). The observed traits for twin 1 and twin 2 are represented by rectangles, whereas unobserved (latent) genetic and environmental variables are represented by circles. Causal paths are represented by single-headed arrows pointing from the latent variables to the observed traits. The path estimates (or regression coefficients)

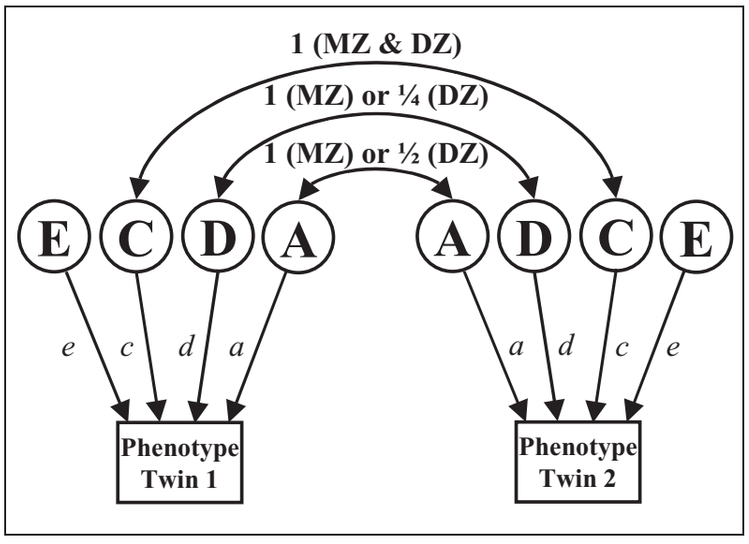


**Figure 1:** Path diagram for the basic univariate twin model. The additive (**A**) and dominance (**D**) factors are correlated 1 between MZ twins and 0.5 and 0.25 between DZ twins, respectively. Shared family environment (**C**) is correlated 1 for both MZ and DZ twins that are reared together in the same home. Unique environment (**E**) is the source of variance that will result in differences among members of one family and is, thus, uncorrelated between members of MZ and DZ pairs. *a*, *d*, *c* and e are the path coefficients for the **A**, **D**, **C** and **E** effects, respectively

indicated by *a*, *c*, *d* and *e* represent the effects of the latent variables on the observed trait. The square of these estimates represents the variance of the trait accounted for by that specific latent factor.

The curved double-headed arrows represent correlations among the latent factors (ie for MZ pairs $r = 1$ for **A**, **D** and **C**; for DZ pairs $r = 0.5$ for **A**, 0.25 for **D** and 1 for **C**). The *genetic* covariance between twin 1 and twin 2 is the product of the paths linking the trait scores via **A** (for MZ: $a * 1 * a = a^2$; for DZ: $a * \frac{1}{2} * a = \frac{1}{2}a^2$). The covariance due to **C** and **D** can be similarly derived by 'path tracing rules'. The total covariance between the two twins is the sum of all the chains (via **A**, **D** and **C**) connecting their observed traits. The expected variances and covariance of the traits within MZ and DZ pairs can then be written in terms of the different variance components:

**Model fitting is a sophisticated method of estimating genetic and environmental effects**

$$\boldsymbol{Cov}\,\mathrm{MZ} = \begin{bmatrix} a^2 + d^2 + c^2 + e^2 & a^2 + d^2 + c^2 \\ a^2 + d^2 + c^2 & a^2 + d^2 + c^2 + e^2 \end{bmatrix}$$

$$\boldsymbol{Cov}\,\mathrm{DZ} = \begin{bmatrix} a^2 + d^2 + c^2 + e^2 & \frac{1}{2}a^2 + \frac{1}{4}d^2 + c^2 \\ \frac{1}{2}a^2 + \frac{1}{4}d^2 + c^2 & a^2 + d^2 + c^2 + e^2 \end{bmatrix}$$

**The effects of C and D are confounded in twin data**

Although both **C** and **D** are included in the diagram and matrices, they are confounded in the classical twin study of MZ and DZ twins reared together and cannot be estimated simultaneously. The twin correlations indicate which of the two components is more likely to be present. When DZ correlations are less than half the MZ correlations, dominance is indicated, because **D** correlates perfectly for MZ but only 25 per cent for DZ twin pairs. Common environmental influences, on the other hand, will make the DZ correlations greater than half the MZ correlations. DZ correlations of about half the MZ correlations suggest additive genetic influences but are also consistent with the presence of both **C** and **D**. In other words, data on twins reared together do not contain enough information to tease out the contrasting effects of the two latent factors. If, for example, data on adoptive siblings are included (which will give us an independent estimate of ***C*** assuming that observed correlations between adoptive siblings are due to shared family environmental effects), we can estimate the effects of both components. The indices of relative contribution of genetic and environmental effects are normally reported as standardised values: that is, if we consider common environmental influences rather than dominance genetic effects, the *heritability* is given by $a^2/(a^2 + c^2 + e^2)$.

## Structural equation model fitting

Whereas path diagrams allow models to be presented in schematic form, they can also be represented as structural equations and covariance matrices and, because all three forms are mathematically complete, it is possible to translate from one to the other.[6,7] Structural equation modelling (SEM) represents a unified platform for path analytic and variance components models and is the current method that is used to analyse twin data. SEM programs involve the use of matrix calculators and numerical optimisation routines. SEM tests hypotheses about relations among observed and latent variables. SEM programs in general perform the following operations:

- make assumptions explicit;
- test the fit of a given model;
- can analyse data from several different familial relationships simultaneously;
- provide estimates of genetic parameters and measurement error;
- allow the comparison of fit of different models;
- are appropriate for human and animal quantitative genetic data.

**Mx: SEM program for free downloading**

### *Mx*

Many SEM programs are available on the market, but the package Mx was specifically developed to model genetically sensitive data in a flexible way.[8] It offers a graphical user interface (MxGUI), which allows path diagrams to be drawn describing the model to be fit to the data. Alternatively, scripts can be written to specify the model. Mx is available for free downloading, supporting various platforms, from the Mx homepage.[9]

Data can be entered as summary statistics for example, covariance matrices and mean vectors, or raw data. Raw data allow greater flexibility: many missing data problems are handled automatically; it is possible to fit finite mixture distributions and it is easy to specify continuous moderator variables. Mx allows for testing both dichotomous moderator effects (eg sex) and continuous moderator variables (eg age). The latter has several advantages: there is no need to partition the sample, and moderating relationships can be directly specified in any part of the model. The Mx homepage contains example scripts for multivariate models, sex differences models, analyses of raw ordinal data and contingency tables. See the web site[10] for a walk through the basic univariate ACE twin model Mx scrip and other behavioural genetic tutorials and modules.

## Fit function

**The maximum-likelihood criterion is commonly used for goodness-of-fit testing**

SEM programs estimate model parameters by minimising a goodness-of-fit statistic between observed and predicted covariance matrices. Different criteria can be used to test goodness-of-fit, but one of the most common and robust is the *maximum-likelihood* criterion. The log-likelihood is maximised by iteratively adjusting the values of the unknown parameters. This process, which is called optimisation or minimisation, is carried out until parameter estimates are obtained that yield the maximum log-likelihood (corresponding in some sense to the smallest possible discrepancies between model and data). Under the assumption of multivariate normality, the contribution of a single twin pair to the log-likelihood function is:

$$-2\ln(2\pi) + \ln|\Sigma| + (x_i - \mu_i)'\Sigma^{-1}(x_i - \mu_i)$$

where $\Sigma$ is the population covariance matrix, $\mu_i$ is the column vector of population means of the variables, and $x_i$ is a column vector of the observed scores of twin pair $i$. As noted above, in the twin model the covariance matrix $\Sigma$ is determined by the parameters $a$, $d$, $c$ and $e$. The overall log-likelihood function is the sum of such contributions from all twin pairs in the sample. In Mx, information about the precision of parameter estimates (and their explained variance) is obtained by likelihood-based confidence intervals (CIs) rather than standard errors. In this method a parameter is progressively moved away from its maximum likelihood estimate in either direction (while the other model parameters are optimised) until the difference in fit, distributed as $\chi^2$ with one degree of freedom, is significant.[11] Unlike standard errors, likelihood-based CIs may be asymmetric around the maximum likelihood estimate.

### *Goodness-of-fit*

The goodness-of-fit of the model relative to a perfectly fitting (saturated) model can be measured by a likelihood ratio chi-square statistic ($\chi^2$). Here, 'perfectly fitting' implies that all the covariances are treated as free parameters, so that their maximum likelihood estimates will be the sample covariances. A non-significant $\chi^2$ value (eg $P > 0.05$) means that the model is consistent with the data, whereas a significant $\chi^2$ value means that the model provides a poor fit to the data and can be rejected. The degrees of freedom (df) for the $\chi^2$ test are the number of observed statistics (which are typically sample variances and covariances) minus the number of parameters being estimated in the model.[6,7]

***Statistical significance***

The statistical significance of the difference between two competing models, provided that the models are nested (ie the set of parameters of one model is a subset of the parameters of the other), can be tested by the difference in $\chi^2$ and the difference in df between the two models. In practice this means that we can test whether the components, **A**, **D**, **C** and **E**, are significantly greater than zero (ie present). For example, it is possible to compare an **AE** model with an **ACE** model, and, in doing so, the significance of the shared environmental component is being tested. If the fit of the simpler, nested model is not significantly worse than that of the full model, the simpler model is preferred, because it provides a more parsimonious explanation of the observed data.[6,7]

**Multivariate twin data allows the estimation of genetic and environmental correlations between traits**

## Multivariate genetic models

If multiple measures have been assessed in twin pairs, the model-fitting approach easily extends to analyse the genetic-environmental architecture of the covariance between the traits. With multivariate models we can investigate the genetic overlap between different disorders, the continuity of genetic factors at different stages of the illness, and the relation between genetic factors and mediating or environmental variables (eg personality, stressful life events) in the development of illness. Within-individual cross-traits covariances imply common aetiological influences. Cross-twin cross-traits covariances imply that these common aetiological influences are familial. Whether these common familial aetiological influences are genetic or environmental is reflected in the MZ/DZ ratio of the cross-twin cross-traits covariances. For depression and anxiety, two very common disorders that often occur together, it was found that the substantial genetic components for both disorders was due to the same genetic factors, whereas the environmental factors were different, and, thus, were responsible for shaping the different outcomes.[12] These results may suggest common strategies for prevention or treatment.

For the simplest, two-variable multivariate **ACE** model the total phenotypic variance is $\Sigma_{\mathbf{P}} = \Sigma_{\mathbf{A}} + \Sigma_{\mathbf{C}} + \Sigma_{\mathbf{E}}$, where $\Sigma_{\mathbf{A}}$, $\Sigma_{\mathbf{C}}$ and $\Sigma_{\mathbf{E}}$ are $2 \times 2$ matrices rather than single values. $\Sigma_{\mathrm{P}}$ can be written as:

$$\begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix} = \begin{bmatrix} \sigma_{\mathrm{A}11} & \sigma_{\mathrm{A}12} \\ \sigma_{\mathrm{A}21} & \sigma_{\mathrm{A}22} \end{bmatrix} + \begin{bmatrix} \sigma_{\mathrm{C}11} & \sigma_{\mathrm{C}12} \\ \sigma_{\mathrm{C}21} & \sigma_{\mathrm{C}22} \end{bmatrix} + \begin{bmatrix} \sigma_{\mathrm{E}11} & \sigma_{\mathrm{E}12} \\ \sigma_{\mathrm{E}21} & \sigma_{\mathrm{E}22} \end{bmatrix}$$

If we re-write the phenotypic correlation in matrix form as follows:

$$\begin{bmatrix} 1 & r_{12} \\ r_{21} & 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{\sigma_{11}}} & 0 \\ 0 & \frac{1}{\sqrt{\sigma_{22}}} \end{bmatrix} * \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix} * \begin{bmatrix} \frac{1}{\sqrt{\sigma_{11}}} & 0 \\ 0 & \frac{1}{\sqrt{\sigma_{22}}} \end{bmatrix}$$

then, for example, the genetic correlation is given by the formula:

$$\begin{bmatrix} 1 & r_{\mathrm{A}} \\ r_{\mathrm{A}} & 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{\sigma_{\mathrm{A}11}}} & 0 \\ 0 & \frac{1}{\sqrt{\sigma_{\mathrm{A}22}}} \end{bmatrix} * \begin{bmatrix} \sigma_{\mathrm{A}11} & \sigma_{\mathrm{A}12} \\ \sigma_{\mathrm{A}21} & \sigma_{\mathrm{A}22} \end{bmatrix} * \begin{bmatrix} \frac{1}{\sqrt{\sigma_{\mathrm{A}11}}} & 0 \\ 0 & \frac{1}{\sqrt{\sigma_{\mathrm{A}22}}} \end{bmatrix}$$

It is possible to partition the phenotypic correlation between two variables in a part determined by common genes ($\sqrt{h_1^2} * r_g * \sqrt{h_2^2}$) and common shared and unique environmental effects.[13]

When we have observed more than three variables per twin, a popular multivariate model is the *common-factor independent-pathway* or *biometric model*. The effect of the genetic and environmental sources of variance is addressed in a structure that is common to the different variables as well as one for the specific portion of variance of these variables

(Figure 2). The *common-factor*, *common-pathway* or *psychometric model* is a more stringent model in which the covariation between variables is caused by a single underlying 'phenotypic' latent variable. This latent variable is caused by genetic and environmental components of variance. As in the independent pathway

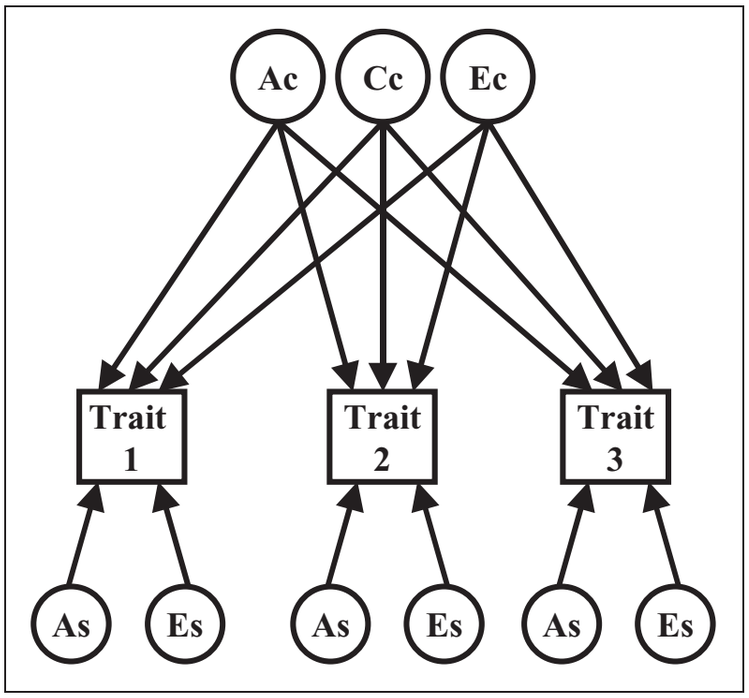


**Figure 2:** Common-factor independent-pathway multivariate model for one twin. **A**, additive genetic, **C**, shared family environment, and **E**, unique environment. Lower-case **c** refers to common effect, and **s** to trait specific effects. Note: **Ac** and all **As** factors are correlated 1 between MZ twins and 0.5 for DZ twins; **Cc** is correlated 1 for both MZ and DZ twins, whereas **Ec** and **Es** are uncorrelated across twins

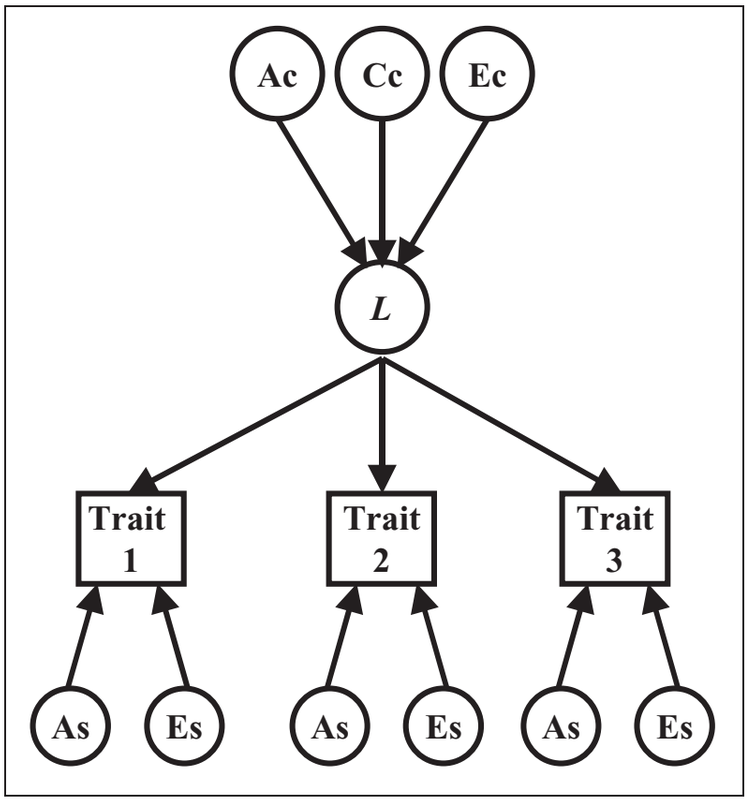


**Figure 3:** Common-factor common-pathway multivariate model for one twin. **A**, additive genetic, **C**, shared family environment, and **E**, unique environment. Lower-case **c** refers to common effect, and **s** to trait specific effects. ***L*** is the latent 'phenotypic' trait

model, there are still variable-specific genetic and environmental sources of variance (Figure 3).

## CATEGORICAL TWIN DATA

Sometimes we are able to discriminate between only a small number of ordered categories with our measuring instrument – for example, the presence or absence of a disease, or the responses to a single item on a questionnaire. In such cases the data take the form of counts – for example, the number of individuals within each response category. Variance component genetic models can be applied to *categorical twin* data by assuming that the ordered categories reflect an imprecise measurement of an underlying *normal distribution of liability*. The liability distribution is further assumed to have one or more *thresholds* (cut-offs) to discriminate between the categories.

**Liability is a hypothetical continuous variable that determines whether an individual will develop a disorder**

When the measured trait is dichotomous – for example, a disorder is either present or not we can partition our observations into pairs concordant for not having the disorder (cell a), pairs concordant for the disorder (cell d) and discordant pairs in which one is affected and one is unaffected (b and c). These frequencies are summarised in a $2 \times 2$ contingency table (CT; Figure 4a). The assumption now is that the joint distribution of liabilities of twin pairs follows a *bivariate normal distribution*, in which both traits have a mean of 0 and standard deviation 1, but the correlation between them is unknown. The shape of such a bivariate normal distribution is determined by the correlation. Figure 5 shows the correlated dimensions for twin pair data with the liabilities of twin 1 and twin 2 on the axes. The correlation (contour) and the two thresholds (cut-offs on the liability distribution) determine the relative

| Twin 1 / Twin 2 | 0 | 1 |
|---|---|---|
| 0 | 00[a] | 01[b] |
| 1 | 10[c] | 11[d] |

(a)

| Twin 1 / Twin 2 | 0 | 1 |
|---|---|---|
| 0 | 00[a] | 01[b] |
| 1 | 10[c] | 11[d] |

(b)

| Proband / Co-twin | 0 | 1 |
|---|---|---|
| 0 | 00[a] | 01[b] |
| 1 | 10[c] | 11[d] |

(c)

**Figure 4:** Contingency tables for categorical twin data under: (a) complete selection; (b) complete ascertainment; (c) single ascertainment. Note: grey areas indicate unobserved cells

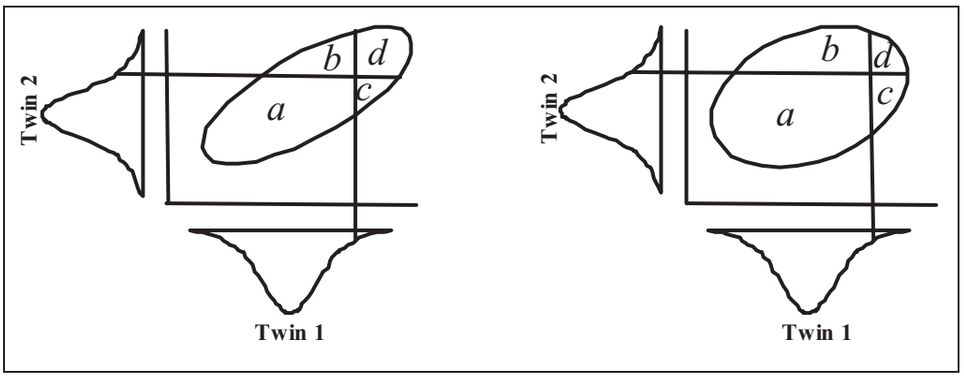


**Figure 5:** Contour plots showing correlated dimensions for twin pair data with the liabilities of twin 1 and twin 2 on the axes. The higher correlation (first plot) will determine different relative proportions (*a*, *b*, *c* and *d*) compared to the lower correlation (second plot)

proportions (*a*, *b*, *c* and *d*) of observations in the four cells of the CT. Conversely, the sample proportions in the four cells can be used to estimate the correlation and the thresholds. The expected proportion under the curve between any ranges of values of the two traits can be calculated by means of numerical integration. Programmed mathematical subroutines can perform these calculations.

**Heritability estimates for categorical traits**

## Genetic model fitting of categorical twin data

When contingency tables of MZ and DZ twin pairs are available, we can estimate the correlation in liability for each type of twins (such correlations are known as *tetrachoric* correlations). Tetrachoric correlations for MZ and DZ twins can be estimated by maximum likelihood, using programs such as Mx or PRELIS. We can also fit a model to the liability that would explain these MZ and DZ tetrachoric correlations. Just as for continuous traits, variance decomposition can be applied to liability, in which correlations in liability are determined by path model. This leads to an estimate of the heritability of the liability to the disorder.

**Contingency Table analyses in Mx**

## Maximum-likelihood analysis of contingency tables

The fit function to CT of twin data is twice the log-likelihood of the observed frequency data calculated by:

$$2\ln L = 2\sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij}\ln(p_{ij})$$

where $n_{ij}$ is the observed frequency in cell *ij*, and $p_{ij}$ is the expected proportion in cell *ij*. Expected cell proportions are calculated by numerical integration of the bivariate normal distribution over two dimensions: the liabilities for twin 1 and twin 2. For example, the probability that both twins are affected (are above a threshold *B*) is equal to the following:

$$\int_B^\infty\int_B^\infty \Phi(L_1,\ L_2;\ 0,\ \Sigma)\mathrm{d}L_1\,\mathrm{d}L_2$$

where $\Phi$ is the multinormal probability density function, $L_1$ is liability of twin 1, $L_2$ is liability of twin 2, 0 signifies that the mean liability for both twins is 0, and $\Sigma$ is the correlation matrix of the two liabilities.

The probability that both twins are unaffected (are below threshold *B*) is given by another integral function with reversed boundaries (from minus infinity to *B*):

$$\int_{-\infty}^B\int_{-\infty}^B \Phi(L_1,\ L_2;\ 0,\ \Sigma)\mathrm{d}L_1\,\mathrm{d}L_2$$

In order to compute a $\chi^2$ statistic the log-likelihood of the data under the model is subtracted from the log-likelihood of the observed data under a perfect fit (the 'saturated model') – that is, the likelihood of expected frequencies that are equal to the observed frequencies, calculated as follows:

$$2 \ln L = 2 \sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij} \ln \left( \frac{n_{ij}}{n..} \right)$$

where $n..$ is the total number of observations in the contingency table. The model's failure to predict the observed data would be reflected in a significant $\chi^2$ for the goodness of fit.

## Comorbidity

Comorbidity is the co-occurrence of two or more disorders in the same patient. To explain this co-occurrence, several models for comorbidity between multifactorial disorders have been developed. These models have implications for the resemblance of the disorders across MZ and DZ twins. Quantitative predictions can be made about the relative proportions of twin-pairs in which each member can fall into one of four categories: neither disorder, disorder A but not B, disorder B but not A, or both A and B.[14] In the correlated liability model for comorbidity individuals have disease A if they are above threshold on the liability to A ($L_A$) and disease B if they are above threshold on the liability to B ($L_B$). Comorbidity arises when the correlation between $L_A$ and $L_B$ is greater than 0. The likelihood function in this case involves integration over four dimensions: the liabilities for A and B in both members of the pair. The data can be summarised in a $4 \times 4$ contingency table. Assuming multivariate normality, each of the predicted proportions in the 16 cells of the contingency table can be expressed as a quadruple integral in which only the limits are altered. For example, the probability that both relatives are comorbid is equal to the following:

$$\int_{t_A}^{\infty} \int_{t_A}^{\infty} \int_{t_B}^{\infty} \int_{t_B}^{\infty} \Phi(L_{A1},\ L_{A2},\ L_{B1},\ L_{B2};\ 0,\ \Sigma)$$

$$\mathrm{d}L_{A1}\, \mathrm{d}L_{A2}\, \mathrm{d}L_{B1}\, \mathrm{d}L_{B2}$$

The extension to more disorders is straightforward: for the comorbidity of three disorders, the predicted proportion of the $9 \times 9$ contingency table can be obtained by six-dimensional integration.

**Twin studies on more categorical traits need efficient sampling strategies**

## Categorical data from proband-ascertained samples

Categorical twin data obtained from random population samples is referred to as *complete selection* because all four cells of the contingency table are represented in the same proportions as the complete population. Sometimes, however, collection of random samples from the population is inefficient. This is the case when we study rare diseases. For example, in schizophrenia, which has a prevalence of less than 1 per cent, the vast majority in a random sample of twins will be pairs in which both are unaffected. A more efficient design in this case would be to ascertain twin pairs through a register of affected individuals. When an affected twin (the *proband*) is identified, the co-twin is followed up to see if he or she is also affected. There are several types of ascertainment:

- **Complete ascertainment**: this refers to the case in which all affected twins in a community sample are registered and selected as probands. In this case twin pairs in which both members are affected will be '*doubly ascertained*' because both members of such twin pairs will be probands (each having the other as an affected co-twin). Twin pairs with one affected member will have one proband and one unaffected co-twin. No concordant unaffected pairs (cell a of the CT) are observed (Figure 4b).

- **Single ascertainment**: if the register is limited, the chance of it containing both members of a twin pair is close to zero. Thus, all twin pairs will have just one proband. The first column of the contingency table (cells *a* and *c*) is not observed, because unaffected individuals cannot be probands (Figure 4c).

- **Multiple incomplete ascertainment**: multiple incomplete

**Proband-assertained data require ascertainment correction for the fit function**

ascertainment lies between *complete* and *single* ascertainment. Depending on the size of the register, the probability of ascertaining more than one proband in the same family increases, and a proportion of twin-pairs will be ascertained more than once. It is possible to express the extent of incomplete ascertainment by means of an ascertainment parameter and to jointly estimate this parameter as well as other parameters in the model.

## Probandwise concordance rate

A simple method of analysing proband-ascertained twin data for categorical traits is the *probandwise concordance rate*. This is the probability that the co-twin of a proband twin will also have the disorder. For complete ascertainment (every affected individual is a proband):

$$\text{Probandwise concordance rate} = \frac{\text{Number of probands whose co-twins are affected}}{\text{Number of probands}}$$

For Figure 4b this is $2d/(2d + b + c)$. Note that the number of probands with an affected co-twin is twice the number of concordant affected pairs.

The inference of a genetic component from proband-ascertained twin pairs is usually based on a difference between MZ and DZ concordance rates. MZ to DZ concordance ratios of 2:1 mean that MZ twins are twice as likely to be affected if their co-twin is affected than DZ twins, indicating the effect of additive genetic influences.

Some earlier twin studies used a 'pairwise' definition of concordance rate.

$$\text{Pairwise concordance rate} = \frac{\text{Number of pairs where both twins are affected}}{\text{Total number of twin pairs}}$$

The pairwise concordance rate cannot be interpreted without knowing the intensity of ascertainment and is now obsolete.

### *Maximum-likelihood analysis of categorical data from proband-ascertained samples*

Proband-ascertained twin samples can be subjected to structural equation modelling, provided that an ascertainment correction is applied. The effect of ascertainment via probands is to distort the frequencies of the cells in the contingency table. In particular, twin pairs without an affected member will not be ascertained, while those with one affected member may be under-represented in comparison to those in which both are affected. Given that the likelihood of observing the different categories are distorted by ascertainment, we need a correction for the fit function. Correction for incomplete ascertainment is achieved by introducing an ascertainment probability for each cell (which may be 0 for concordant unaffected pairs, 1 for concordant affected pairs, and a free parameter between 0 and 1 for discordant pairs). The cell probabilities are adjusted by multiplication with these ascertainment probabilities, divided by a scaling factor so that they sum to 1.[15] This type of analysis can be conducted in Mx using the option for user-defined fit function.

## Checking the assumptions of the twin method

### *Equal environments*

The equal environment assumption across zygosities assumes that environmentally caused similarity is roughly the same for both types of twin pairs reared in the same family. This assumption is the most basic assumption of the twin method and has been the subject of great debate over the years. It is generally agreed that MZ and DZ twins do share their environment to the same extent in many respects: they share the womb at the same time, are exposed to the same environmental factors, are raised in the same family and are the same age. However, there is also some evidence that MZ twins are treated more similarly by their parents and have

**Violation of the assumptions of the twin method will lead to incorrect estimates of $h^2$ and $c^2$**

more frequent contact as adults than DZ twin pairs.[13]

- **Implications:** more similar treatment of MZ twins will increase their correlations relative to DZ correlations, which can result in an overestimation of the genetic effect and an underestimation of the shared environmental effect. (Note: there are also factors that can have the opposite effect and increase variability between MZ twins. One example is when MZ twin pairs are forced to attend different classes at school, while DZ twins are allowed to remain in the same class. This could lead to an underestimation of the genetic effect.)

- **How do we detect this effect?** If parental treatment is more similar for MZ twins, than DZ twins who are mislabelled as MZ twins should be more alike than correctly labelled DZ twins and conversely, MZ twins mislabelled as DZ should be less alike than correctly labelled MZ twins. Little or no effect of mislabelling was found. The effect of degree of contact among twins showed that more frequent contact does not lead to behavioural similarity in same-sex DZ or MZ twins. While in some cases MZ twins in frequent contact were more similar than those with less contact, these correlations tended to be small.[16] Another argument in defence of the equal environments assumption is the fact that studies of MZ twins reared apart have provided correlations for personality variables that are almost the same as those for MZ twins reared together.[17]

***Genotype–environment effects***

*Assortative mating* refers to any non-random pairing of mates on the basis of factors other than biological relatedness. It is included here, since it may be influenced by both genetic and environmental factors and because assortative mating may affect the transmission, magnitude and correlation of both genetic and environmental effects. Apart from assortative mating, social interaction may also cause similarity between mothers and fathers.

- **Implications:** if people choose partners who are phenotypically like themselves, environmental and genetic correlations between relatives are increased. This means that the correlation between DZ twin pairs, relative to that of MZ twin pairs, is increased and, thus, leads to an overestimation of the shared environmental effect.

- **How do we detect this effect?** By looking at the phenotypic correlation between parents for the trait in question. In order to determine whether assortative mating is taking place, it is necessary to trace the change in spouse resemblance over time or analyse the resemblance between the spouses of biologically related individuals.[18]

*Genotype–environment correlation* (or genetic control of exposure to the environment) refers to the fact that exposure to environments is not random, but that genetic factors influence the probability that individuals will select themselves into certain environments. There are different types of G × E correlation, the most common being *active* and *passive correlation*.

*Active* G × E *correlation* arises when an individual creates or invokes environments that are a function of his or her genotype. An example in psychiatry is that the genetic liability to major depression was shown to be associated with a significantly increased risk for experiencing several stressful life events.[19]

- **Implications:** positive correlations will increase, and negative correlations will decrease estimates of genetic components.

- **How do we detect this effect?** There is no way of knowing which genetic effects act directly on the phenotype and which result from the environmental effects that were actually caused by genes unless we have longitudinal trait data and an 'environmental' measure. The first indication is the existence of a genetic overlap between the environmental measures (ie life events) and the trait (see multivariate analyses). It was shown with time survival analysis that 10–15 per cent of the impact of genes on risk for MD is mediated through stressful life events.[19]

*Passive G × C correlation* arises because the environment in which individuals develop is provided by their biological relatives, with whom they are genetically related. An example is that musicians may transmit both genes for musical talent and a family environment rich in music to their offspring.

- **Implications:** positive correlations will tend to increase the estimate for shared family environment effect.

- **How do we detect this effect?** By comparing the correlation between a measure of family environment (parental responsivity, encouragement of developmental advance, provision of toys/books, etc.) and offspring traits in non-adoptive and adoptive families. If the correlation is greater in non-adoptive families, it reflects a genetic origin and thus a passive G × C correlation.

*Gene–environment interaction* (or genetic control of sensitivity to the environment) refers to different genotypes responding differently to the same environment or some genotypes being more sensitive to changes in environment than others. An example in psychiatry is that the depressogenic effect of stressful life events is substantially greater in those at high versus low risk to major depression.[20]

- **G × E interaction implications:** positive interactions will be estimated in **E**.

- **How do we detect this effect?** In practice, it is extremely difficult to detect G × E interactions in humans without explicitly measured environmental indices. However, Jinks and Fulker[1] have shown that G × E interaction can lead to a relationship between the sum and absolute differences of twin pairs' scores (known as heteroscedasticity). This relationship is accompanied by non-normality (skewness), and can often be removed by scale transformation. (Note: sometimes we do not wish to remove this effect, but rather model it as it can give us valuable insight into the aetiology of diseases like the example on major depression given above.) In Neale and Cardon[6] different types of G × E interaction and correlations are discussed in more detail.

- **G × C interaction implications:** positive interactions will be estimated in **A**.

- **How do we detect this effect?** As for G × E interaction it is extremely difficult to detect G × C interactions in humans without explicit environmental measures. G × C interaction may be indicated by a relationship between trait sum and absolute trait difference in DZ but not MZ twins.

### *Generalisability of twins to the general population*

Even if the twin method is a valid way of studying the heritability of a trait, it still needs to be shown that twins are representative of the target population from which the researcher has been sampling. There are some genuine differences between twins and singletons in terms of pregnancy and the birth process. Twins are on average lighter than singletons, are born on average approximately three weeks pre-term, and

have more frequent complications, caesarean sections and malformations. For many diseases, obstetric and paediatric complications do not play an important role and so should not necessarily pose a problem. However, for schizophrenia, there is now substantial evidence that there is an excess of obstetric complications among affected subjects in comparison to controls. Given that obstetric complications are more prevalent among twins and among schizophrenics, this might imply that schizophrenia is also more frequent among twins, and, indeed there is some suggestion that this may be the case.

### Power calculations

An important question in twin study designs is the power one has to detect the effect of a specific variance component given a specific sample size or vice versa. A great advantage of SEM is the possibility to conduct significance tests of nested hypotheses. Power calculations for SEM can be done using Mx, by generating the expected covariance matrices under a specific hypothesis (eg $\mathbf{A} = 0.5$, $\mathbf{C} = 0.2$, $\mathbf{E} = 0.3$), and then run both the true and a restricted (false) model (eg $\mathbf{A} = 0$) on the saved covariance matrices. Since the false model will be a submodel of the true model, the $\chi^2$ statistic from the false model, given the sample size, can be used to determine the power of the test. The power command in Mx uses this $\chi^2$ and the user-supplied significance level, $\alpha$, and degrees of freedom to compute the power of the study to reject the hypothesis. Mx will also print the required sample sizes to reject the alternative models at various power levels.[6]

Results of power studies show that at least 200 pairs are needed for obtaining a reasonable estimate of the degree of genetic influence on a highly heritable trait.[6] For intermediate or low heritable traits, 10–20 times these numbers are required. The same is true for detecting family environmental effects and non-additive genetic effects.

## CONCLUSION

When the assumptions made in the classical twin design are met, it is a powerful tool for partitioning genetic from environmental factors on a trait. The most basic and debated assumption is the equal environment assumption, which assumes that environmentally caused similarity is roughly the same for both types of twin pairs reared in the same family. Violation of this assumption (more similar treatment of MZ twins) will result in increased MZ correlations relative to DZ correlations, which can result in an overestimation of the genetic effect and an underestimation of the shared environmental effect. A strong argument in defence of the equal environments assumption comes from MZ twins reared apart, for whom correlations of personality variables appear to be very similar to those for MZ twins reared together.[17] However, contrary evidence comes from the differential prenatal history of MZ *v*. DZ twins, in particular monochorionic twins. Hypothesised differences in monochoriotic and dichoriotic MZ twin pairs form another challenge to the classical twin study which is a current topic of much debate. The validity of this argument will be solved by twin studies collecting information on placentation.[21]

We have described only the basic biometrical genetic models for twin data, and the common extensions to multivariate and categorical data. However, SEM methodology has the flexibility to be extended in many other ways. Fitting models to multiple groups of MZ and DZ male and female twin pairs allows for testing quantitative sex-differences in, for example, heritability. In addition, the inclusion of opposite-sex pairs allows the test of qualitative sex differences (ie if different genetic factors are operating across sexes by relaxing the constrained of assumed genetic correlation of 0.5). Other examples include the modelling of measurement errors, reporting bias, reciprocal twin interaction and twins–parents data.

Another important application of SEM is the genetic mapping of quantitative trait loci (QTL) using twin and sibling data, by incorporating the effect of measured genotypes on the trait in a variance components model.[22]

Large population-based and volunteer twin registers have been established all over the world (eg Denmark, Norway, Sweden, Finland, Australia, the Netherlands, the USA and the UK) to study the genetic aetiology of health-related and other traits. SEM methodology is an essential analytic tool for extracting maximal information from data collected from these resources.

## *References*


1. Jinks, J. L. and Fulker, D. W. (1970), 'Comparison of the biometrical genetical, MAVA, and classical approaches to the analysis of human behavior', *Psychol. Bull.*, Vol. 73, pp. 311–349.
2. Eaves, L. J. (1977), 'Inferring the causes of human variation', *J. Royal Stat. Soc., Series A*, Vol. 140, pp. 324–355.
3. Martin, N. G. and Eaves, L. J. (1977), 'The genetical analysis of covariance structure', *Heredity*, Vol. 38, pp. 79–95.
4. Mather, K. and Jinks, J.L. (1977), 'Introduction to Biometrical Genetics', Cornell University Press, Ithaca, NY.
5. Wright, S. (1921), 'Correlation and causation', *J. Agric. Res.*, Vol. 20, pp. 557–585.
6. Neale M. C. and Cardon L. R. (1992), 'Methodology for Genetic Studies of Twins and Families', Kluwer Academic Publishers, Dordrecht.
7. Sham, P. C. (1998), 'Statistics in Human Genetics', Oxford University Press, New York.
8. Neale, M. C. (1999), 'Mx: Statistical Modeling', 5th edn, Department of Psychiatry, Medical College of Virginia, Richmond, VA.
9. URL: http://views.vcu.edu/html/mx/mxhomepage.html
10. URL: http://statgen.iop.kcl.ac.uk/bgim/
11. Neale, M. C. and Miller, M. B. (1997), 'The use of likelihood-based confidence intervals in genetic models', *Beh. Genet.,*Vol. 27, pp. 113–120.
12. Kendler, K. S., Neale, M. C., Kessler, R. C. *et al.* (1992), 'Major depression and generalized anxiety disorder: same genes, (partly) different environments?', *Arch. Gen. Psych.*, Vol. 49, pp. 716–722.
13. Plomin R., DeFries, J. C., McClearn, G. E. and McGuffin, P. (2001), 'Behavioral Genetics', 4th edn, Worth Publishers, New York.
14. Neale, M. C. and Kendler, K. S. (1995), 'Models of comorbidity for multifactorial disorders', *Amer. J. Hum. Genet.*, Vol. 57, pp. 935–951.
15. Cardno, A. G., Rijsdijk, F.V., Sham, P.C. *et al.* (2002), 'A twin study of genetic relationships between psychotic symptoms', *Amer. J. Psych.*, Vol. 159, pp. 539–545.
16. Kendler, K. S; Heath, A., Martin, N. G. and Eaves, L.J. (1986), 'Symptoms of anxiety and depression in a volunteer twin population: The etiologic role of genetic and environmental factors', *Arch. Gen. Psych.*, Vol. 43, pp. 213–221.
17. Bouchard, T. J. and McGue, M. (1990), 'Sources of human psychological differences: the Minnesota study of twins reared apart', *Science*, Vol. 268, pp. 223–228.
18. Heath, A. C., Eaves, L. J., Nance, W. E. and Corey, L. A. (1987), 'Social inequality and assortative mating: Cause or consequence?', *Beh. Genet.*, Vol. 17, pp. 9–17.
19. Kendler, K. S. and Karkowski-Shuman, L. (1997), 'Stressful life events and genetic liability to major depression: Genetic control of exposure to the environment?', *Psychol. Med.,* Vol. 27, pp. 539–547.
20. Kendler, K. S. (1998), 'Major depression and the environment: A psychiatric genetic perspective', Anna-Monika Prize Paper, *Pharmacopsychiatry*, Vol. 31, pp. 5–9.
21. *Twin Research* (2001), Special issue: The fetal origins hypothesis, guest editors Lambalk, C. B. and Roseboom, T. J., *Twin Research*, Vol. 4, No. 5.
22. Neale, M. C. (2000), 'QTL mapping with sib-pairs: The flexibility of Mx', in Spector, T. D., Snieder, H. and MacGregor, A. J., Eds, 'Advances in Twin and Sib-pair Analysis', Oxford University Press, London.