Problems With Using Sum Scores for Estimating Variance Components: Contamination and Measurement Noninvariance

Michael C. Neale,[1] Gitta Lubke,[2] Steven H. Aggen,[1] and Conor V. Dolan[3]

[1] *Virginia Institute for Psychiatric and Behavioral Genetics, Virginia Commonwealth University, Richmond, Virginia, United States of America*
[2] *University of Notre Dame, Notre Dame, Indiana, United States of America*
[3] *University of Amsterdam, Amsterdam, the Netherlands*

Twin studies of complex traits, such as behavior or psychiatric diagnoses, frequently involve univariate analysis of a sum score derived from multiple items. In this article, we show that absence of measurement invariance across zygosity can bias estimates of genetic and environmental components of variance. Specifically, if the item responses are considered as multiple indicators of a latent factor, and the aim is to partition the variance in the latent factor, then the factor loadings relating the items to the factor should be equal for monozygotic (MZ) and dizygotic (DZ) twins. While it seems unlikely, a priori, that these loadings should differ as a function of zygosity, certain special measurement situations are cause for concern. Ratings by parents, or self-ratings of phenotypes which are more easily observed in others than via introspection, may be tainted by the co-twin's phenotype to a greater extent in MZ than DZ pairs. We also show that the analysis of sum scores typically biases both MZ and DZ correlations compared to the true latent trait correlation. These two sources of bias are quantified for a range of values and are shown to be especially acute for sum scores based on binary items. Solutions to these problems include formal tests for measurement invariance across zygosity prior to analysis of the sum or scale scores, and multivariate genetic analysis at the individual item or symptom level.

The individual items in most measurement instruments are designed to measure a single underlying factor or latent trait. However, the items are rarely pure indicators of the underlying factor. For example, a particular symptom used to indicate the presence or absence of depressive disorder may also be sensitive to sleep disorders. If a test is administered to more than one group and if at least one of the items fails to measure the same factors equivalently (i.e., does not have the same factor loading) in those groups, the item is said to lack measurement invariance (MI), or be measurement noninvariant (MNI) across the groups. The present article shows that MNI across monozygotic (MZ) and dizygotic (DZ) twins (hereafter referred to as MNIz; measurement invariance with respect to zygosity will be abbreviated as MIz) can bias estimates of genetic and environmental components of variance. The potential for MNI across gender or across an environmental grouping variable to bias estimates of G × Sex or G × E interaction was demonstrated in a previous report (Lubke et al., 2004).

In factor analysis, a factor is specified to account for the covariance among a set of indicator measures, which may be continuous, ordinal or binary (Mellenbergh, 1994). In the binary case, analyses based on the threshold model are equivalent to the normal ogive item response theory (IRT) model (Takane & de Leeuw, 1987). Such binary item data are common in the study of behavioral traits, including those focusing on complex traits such as psychopathology and substance use. The factor model, in both continuous and ordinal data applications, is the focus of this article. It is the cornerstone of multivariate genetic analyses; its extensions include the popular common and independent pathway models (Kendler et al., 1987; McArdle & Goldsmith, 1990; Neale & Cardon, 1992).

Measurement invariance (Mellenbergh, 1989; Meredith, 1993; Millsap & Everson, 1993) holds with respect to a grouping variable if the probability of an observed test score, or item response, is the same for members of different groups with the same score on the latent factor. In concrete terms, failure of measurement invariance of an item with respect to zygosity would occur if an MZ twin scored higher (or lower) on average on the item than would a DZ twin who has exactly the same score on the latent factor. An item might operate differently between the zygosity groups because of different sensitivity to variables other than those used to define the factor. This

Received 26 July, 2005; accepted 30 September, 2005.

Address for correspondence: Michael C. Neale, Virginia Institute of Psychiatric and Behavioral Genetics, Virginia Commonwealth University USA, 800 East Leigh St. Suite 1–115, Richmond VA 23219–1534, USA. E-mail: mcneale@vcu.edu

https://doi.org/10.1375/twin.8.6.553 Published online by Cambridge University Press

problem could manifest itself as different estimates of the factor loading in the two groups. Other manifestations of MNI include different amounts of item-specific variance and item means that differ for reasons other than a group differences in the factor mean. In general, differences between groups on the latent factor — be they in either the mean or the variance of the factor — are not problematic. Differences in the factor loadings, the item-specific means (see below), or the item-specific variances are indicators of MNI, and are a cause for concern about the comparison of the similarity of MZ and DZ twins.

Tests of measurement invariance can be conducted if and only if the measurement model relating observed item scores to the underlying factor(s) is analyzed simultaneously in all groups (Dolan, 2000; Lubke et al., 2003; Meredith, 1993). Joint, that is, multigroup, analysis of the items allows a test of whether the measurement model is the same across groups. Conversely, if sum scores are precomputed by, for example, summing individual item scores, all information about the relationship between individual items and the underlying factor(s) is lost. Absence of measurement invariance can no longer be detected. This problem also applies to factor scores, which are essentially weighted sum scores.

In genetic epidemiology, it is common practice to derive sum scores by summing individual questionnaire items or symptom scores. The Eysenck Personality Questionnaire scales, Extraversion, Neuroticism and Psychoticism, are computed in this way (Eysenck & Eysenck, 1975), as are the scale scores of many other psychological instruments. Similarly, *Diagnostic and Statistical Manual of Mental Disorders,* (4th ed., text rev.; DSM-IV-TR; American Psychiatric Association, 2000) diagnoses are typically established when a subject has at least r of a given list of s symptoms. Thus, these diagnoses are essentially sum scores that are then subjected to further reduction by a binary filter. In the present article, we focus on variance component models that rely on the comparison of the similarity of MZ and of DZ twin pairs to draw conclusions about the relative impact of genetic and environmental factors. We show that the analysis of sum or factor scores using additive genetic, common environment and specific environment variance components models (ACE models) may result in biased variance components estimates when the individual items from which the sum or factor score is derived are MNIz. More specifically, if MZ twins' responses to items assess the latent factor more accurately, then heritability would be biased upwards.

Sum scores are often regarded as an estimate of the underlying factor score. Obviously, decomposition of a sum score into genetic and environmental variance components only makes sense if the items, which form the sum score, measure the same factor(s) in MZ and DZ twins. Fitting an ACE-type model to a sum score relies on the implicit assumption that the items from which the sum score is derived measure the same underlying factor or latent trait regardless of zygosity, in other words, that the items are MIz. Furthermore, it is assumed that the items index the latent trait equally well. It is, however, an empirical question whether these assumptions hold, and MIz should be established prior to variance component analysis of sum scores. To detect noninvariance with respect to zygosity, the relationship between the individual items and the latent variables can be modeled simultaneously in the MZ and DZ groups, and tested for equality.

A second issue we address is that even when MIz holds, the use of sum scores to estimate variance components of a hypothesized latent trait is likely to be problematic. If the item-specific variance is truly random error, and uncorrelated between twins, then estimates of familial variance components, such as additive genetic or common environment effects, based on the sum score will underestimate the impact of these sources on the latent trait. In the event that some of the item-specific variance is familial (Waller & Reise, 1992), then variance component estimates based on sum scores may either over- or underestimate the latent trait variance components.

In the following sections, we first define the concept of measurement invariance and then show that a variance decomposition of sum scores derived from items that are noninvariant with respect to zygosity may bias estimates of additive genetic and common environment variance components. This bias is quantified algebraically and displayed graphically for several specific sets of factor loadings. In the discussion section, we consider situations in which MNIz is likely to occur and the advantages of multivariate analysis.

Measurement Invariance

The background information on measurement invariance presented in this section is essentially identical to that provided in Lubke et al. (2004); it is reproduced here for convenience because of its central relevance to the present article. Absence of measurement invariance, which is also known as differential item functioning (DIF), has been studied extensively both in the context of confirmatory factor analysis and IRT (Bloxom, 1972; Byrne et al., 1989; Ellis, 1993; Holland & Wainer, 1993; Lubke et al., 2003; Marsh, 1994; McArdle, 1998; Mellenbergh, 1989; Meredith, 1993). Measurement invariance is defined with respect to a grouping variable such as race, gender, or in the present case zygosity, and concerns the measurement model relating observed scores to underlying latent variables (Mellenbergh, 1989; Meredith, 1993). The measurement model must be the same for all groups, in the sense that the probability of observing a given item score is equal for members of different groups who have the same score on the underlying latent variable. More formally, measurement invariance is defined as

$$f(M \mid F,S) = f(M \mid F) \qquad [1]$$

https://doi.org/10.1375/twin.8.6.553 Published online by Cambridge University Press

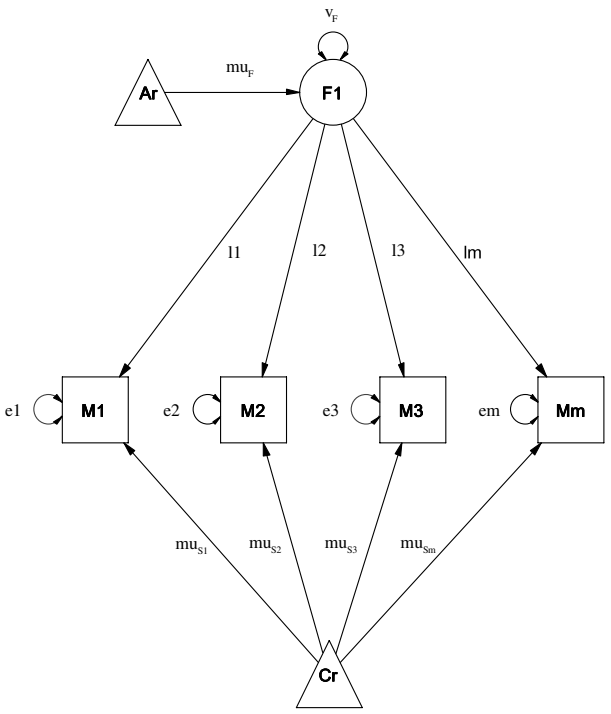

Figure 1

Path diagram of a single latent factor model, illustrating the five parameter classes that may be moderated as a function of group membership: factor mean, mu_F; factor variance, v_F; factor loadings, l_j; item-specific means, mu_j; and item-specific variances, e_j.

where observed variables are denoted as *M*, latent variables as *F*, and the grouping (or *S*election) variable as *S* (Mellenbergh, 1989; see Dolan et al., 2004 for a conceptual explanation). Equation 1 states that, given the scores on the underlying latent variable(s), the probability distribution of the observed scores does not depend on subpopulation membership, for example, being an MZ or a DZ twin, but depends only on the scores on the underlying factor or factors. The distribution of these latent scores may, however, differ between groups.

An extended single factor model for multiple observed measures is shown as a path diagram in Figure 1. In this model, the latent factor F1 is, by convention, shown in a circle as it is not directly observed. It has a variance V_F that is fixed to 1.0 and a mean, shown as the path μ_F from the constant Ar (depicted as a triangle) that is fixed to zero. Variation in the observed scores S_j, j = 1…4 is caused partly by the latent factor and partly by item-specific error variables $E1…E4$. The mean of each observed score is partly a function of the latent factor mean (here zero) and partly by the measure-specific mean mu_{M_j}. In principle, groups can differ with respect to the following five components of this model:

1. The factor mean, mu_F
2. The factor variance, v_F
3. The factor loadings, l_j
4. The item-specific means, mu_{sj}
5. The item-specific variances, e_j

In the case of binary data, the item-specific means are replaced by thresholds, and the item-specific variances cannot be identified and are typically fixed either to 1.0 or to $(1 - l_j^2)$. We note that by using definition variables in Mx (Neale et al., 2003), it is possible to moderate these five components with grouping variables that may be either continuous (such as age) or discrete (such as sex). In the present article, we focus on the effects of moderation of the factor loadings as a function of zygosity.

If the factor loadings differ across groups, the interpretation of the underlying factor or trait may differ as a consequence.[1] Suppose that in one group depression items load strongly on a general factor and anxiety items have weaker loadings on this factor. Suppose also that in a second group these anxiety items have higher loadings on the factor, while the depression items have lower loadings. This different patterning of item loadings is an example of failure of measurement invariance. In the first group, the depression items have a larger weight than the anxiety items, so the general factor would be interpreted as representing liability to depression. By contrast, in the second group, anxiety items have larger loadings so the general factor would be interpreted as an anxiety factor. This difference in interpretation is lost when sum scores are derived from the individual items, because when adding the items, usually the same weights are used for all items in all groups. In other words, sum scores are based on the implicit assumption of measurement invariance and are, in the case of noninvariance, incorrectly interpreted as estimates of the same factor across groups. It would make little sense to estimate heritability of the psychopathology factor by comparing MZ twins' similarity for depression with DZ twins' similarity for anxiety. Thus, there is a strong case for verifying factorial invariance across zygosity groups prior to model fitting.

The incorrect interpretation of a sum score is especially important when sum score variance is decomposed into genetic and environmental variance components. On a conceptual level, it is questionable whether it makes sense to decompose sum scores derived from noninvariant items. However, the problem of analyzing sum scores extends beyond this issue of conceptual interpretation. It is shown below that variance components obtained from modeling sum scores are biased if measurement invariance does not hold with respect to zygosity. Hence, absence of MNI across zygosity is confounded with sources of familial resemblance. Differences in factors loadings or discrimination parameters can be detected only in a multivariate analysis of the individual items carried out simultaneously in all groups. That is, it is best to conduct a multivariate analysis of the items that are measured, rather than of a scale score derived from them.

https://doi.org/10.1375/twin.8.6.553 Published online by Cambridge University Press

Our investigation into the possible effects of failure of measurement invariance on the estimates of variance components from sum scores brings to light another serious potential problem for quantitative genetic studies. When item scores are summed to provide an indicator of the latent trait, the correlation between twins on the sum score can differ substantially from their correlation on the latent trait. As we shall see, the amount of this deviation varies according to the value of the true latent trait correlation, and on the variance components of the residual (item-specific) variances. To a certain extent, this problem is obvious. Bias from this source may be considerable, making the use of sum scores in genetically informative studies (among others) at best questionable.

Theory Behind Variance Component Bias in Sum Scores

Measurement Model

The algebra in the following two subsections is derived to establish two main principles. First, if the items contain error (as is likely to be the case), and this item-specific variance, u^2, is nonfamilial in origin, then the proportion of e^2 estimated in the sum score increases relative to that of the true latent factor. Second, the relative proportions of a^2 and c^2 may also be affected by the analysis of sum scores when there is MNIz.

For reasons of simplicity, we use the common factor model with a single factor F and two continuous observed variables, y_1 and y_2, as a measurement model to demonstrate two main points. Both concern the results obtained when a genetic 'ACE' model (consisting of additive genetic [a^2], common environment [c^2] and unique environment [e^2] variance components [Neale & Cardon, 1992]) is fitted to sum scores.

The arguments made using ordinary algebra in this section are restated in matrix algebra form in the Appendix. We can write the measurement model relating items $j = 1 \ldots m$ to an underlying factor in MZ twin pairs as follows:

$$\begin{aligned} y_{j1\mathrm{mz}} &= \tau_{1j} + \lambda_{j1\mathrm{mz}} F_{\mathrm{mz}} + \varepsilon_{j1\mathrm{mz}} \\ y_{j2\mathrm{mz}} &= \tau_{2j} + \lambda_{j2\mathrm{mz}} F_{\mathrm{mz}} + \varepsilon_{j2\mathrm{mz}} \end{aligned} \qquad [2]$$

where $\lambda_{j1\mathrm{mz}}$ denotes the factor loading of item j measured on twin 1 in an MZ pair, ε_{j1} denotes the residual variance for this item, and τ_{1j} is the item-specific mean or 'regression intercept'. Corresponding equations for the observed scores in DZ twins can be obtained by changing the subscript mz to dz. Thus we have described a simple linear factor model with an intercept term τ.

The sum score of each member of an MZ twin pair is obtained by adding the scores on the two items:

$$S_{\mathrm{mz}} = \lambda_{1\mathrm{mz}} F_{\mathrm{mz}} + \lambda_{2\mathrm{mz}} F_{\mathrm{mz}} + \varepsilon_{1\mathrm{mz}} + \varepsilon_{2\mathrm{mz}} + \tau_{1j} + \tau_{2j} \qquad [3]$$

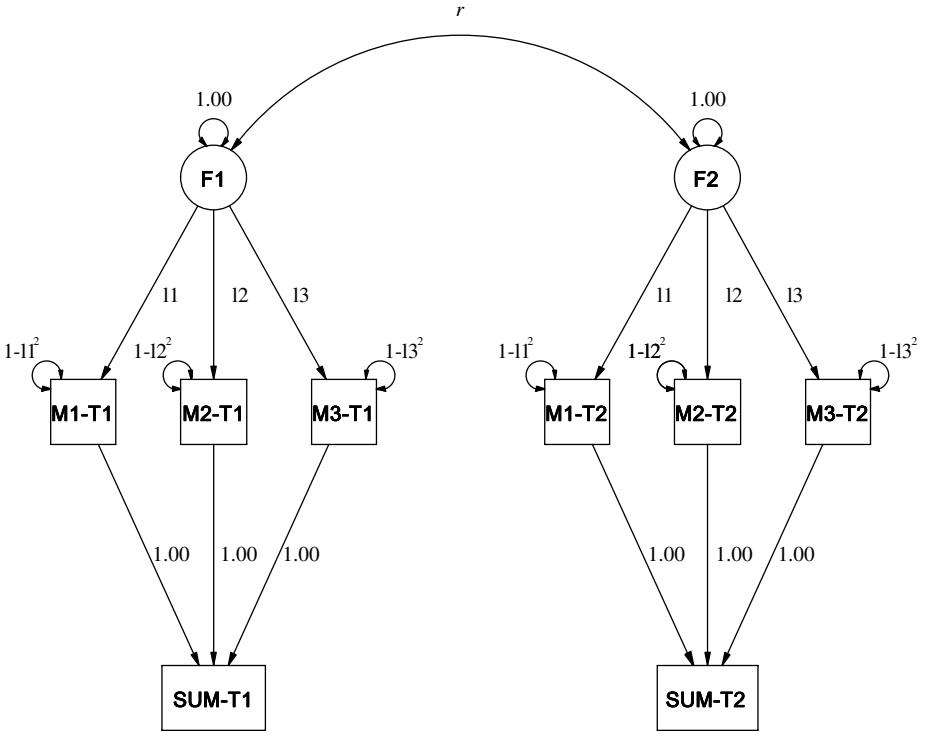

Figure 2

Path diagram illustrating the generation of sum scores Sum-T1 and Sum-T2 from latent traits LT1 and LT1 which correlate *r*.

https://doi.org/10.1375/twin.8.6.553 Published online by Cambridge University Press

Absence of invariance with respect to factor loadings means that at least one of the factor loadings in MZ twins does not equal its counterpart in DZ twins, $\lambda_{imz} \neq \lambda_{idz}$.

The goal here is to demonstrate that the sum score variance component estimates, $\hat{a}_S^2$, $\hat{c}_S^2$ and $\hat{e}_S^2$ are unlikely to equal those of the latent factor, $\hat{a}_F^2$, $\hat{c}_F^2$ and $\hat{e}_F^2$. Furthermore, we show that this is the case *with MIz as well as with MNIz*. Ignoring the measurement model relating individual items and the underlying factor, we write the covariance between the sum scores of a twin pair, t_1 and t_2, according to the ACE-model as:

$$\mathrm{Cov}(S_{t1}, S_{t2}) = \alpha a_S^2 + c_S^2 \quad [4]$$

where, given the usual assumptions of the additive multifactorial model, α is fixed to 1.0 and .5 for MZ and DZ twins, respectively.

If the measurement model for the individual items (Figure 2) is taken into account, the covariance between sum scores of twin 1 and twin 2 can be written as:

$$\mathrm{Cov}(S_{t1}, S_{t2}) = (\lambda_{1t1}\lambda_{1t2} + \lambda_{1t1}\lambda_{2t2} + \lambda_{2t1}\lambda_{1t2} + \lambda_{2t1}\lambda_{2t2}) \times \mathrm{Cov}(F_{t1}, F_{t2}) \quad [5]$$

or when $\lambda_{1t1} = \lambda_{1t2}$ and $\lambda_{2t1} = \lambda_{2t2}$, that is, when measurement invariance with respect to order within a twin pair (twin 1 vs. twin 2) holds, then:

$$\mathrm{Cov}(S_{t1}, S_{t2}) = (\lambda_1 + \lambda_2)^2 \mathrm{cov}(F_{t1}, F_{t2}) \quad [6]$$

where the term involving λ is essentially a weight for the covariance between the factor scores of a twin pair (this argument follows from Lubke et al., 2004). The weight depends on the measurement model, or, more specifically, on the factor loadings. Let k_{mz} and k_{dz} denote the weights of MZ and DZ twins, respectively. Then

$$k_{\mathrm{mz}}^2 = \lambda_{1\mathrm{mz}}^2 + 2\lambda_{1\mathrm{mz}}\lambda_{2\mathrm{mz}} + \lambda_{2\mathrm{mz}}^2 = (\lambda_{1\mathrm{mz}} + \lambda_{2\mathrm{mz}})^2 \quad [7]$$

$$k_{\mathrm{dz}}^2 = \lambda_{1\mathrm{dz}}^2 + 2\lambda_{1\mathrm{dz}}\lambda_{2\mathrm{dz}} + \lambda_{2\mathrm{dz}}^2 = (\lambda_{1\mathrm{dz}} + \lambda_{2\mathrm{dz}})^2 \quad [8]$$

Note that here, for simplicity, we assume that the item-specific variance is not correlated between twins. We return to consider violations of this assumption below. Also worthy of note is that in the general case of m items, we have $k^2 = (\sum_{j=1}^{m} \lambda_j)^2$, that is, the square of the sum of the factor loadings. Let the covariance between twins' factor scores be decomposed according to the ACE model:

$$\mathrm{CovMZ}(F_{t1}, F_{t2}) = a_F^2 + c_F^2 \quad [9]$$

$$\mathrm{CovDZ}(F_{t1}, F_{t2}) = .5a_F^2 + c_F^2 \quad [10]$$

Substituting equations 8 to 10 into equation 6, the covariances of MZ and DZ pairs' sum scores are therefore:

$$\mathrm{Cov}(S_{1\mathrm{mz}}, S_{2\mathrm{mz}}) = k_{\mathrm{mz}}^2(a_F^2 + c_F^2) \quad [11]$$

$$\mathrm{Cov}(S_{1\mathrm{dz}}, S_{2\mathrm{dz}}) = k_{\mathrm{dz}}^2(.5a_F^2 + c_F^2) \quad [12]$$

The key differences between equations 11 and 12 and the usual ones for the resemblance between pairs of twins under the ACE model are the terms k_{mz}^2 and k_{dz}^2. These terms accumulate all the squared factor loadings and cross-products of the factor loadings, and therefore reflect the accuracy with which the observed scores measure the latent trait. Note that the k^2 terms in equations 11 and 12 apply to the whole of the covariance between the twins' factors, regardless of whether this covariance is due to additive genetic or common environment factors. Consequently, if $k_{\mathrm{mz}} = k_{\mathrm{dz}}$, that is, MIz holds, the reduction in correlation of the sum scores is proportionate for MZ and DZ pairs. This induces a proportionate reduction of additive genetic and common environmental (or nonadditive genetic) variance components. We now demonstrate this property by expressing variance components for sum scores, denoted a_S^2, c_S^2 and e_S^2, in terms of the latent factor variance components a_F^2, c_F^2 and e_F^2.

Variance Component Bias in Sum Scores: Measurement Invariance

Suppose that we were able to measure the latent factors directly, and computed the variance of the latent factor, V_P, the covariance of MZ twins, $r_{F\mathrm{mz}}$, and of DZ twins, $r_{F\mathrm{dz}}$. The least squares solution of the system of equations:

$$a_F^2 + c_F^2 + e_F^2 = V_P \quad [13]$$

$$a_F^2 + c_F^2 = r_{F\mathrm{mz}} \quad [14]$$

$$.5a_F^2 + c_F^2 = r_{F\mathrm{dz}} \quad [15]$$

is:

$$a_F^2 = 2(r_{F\mathrm{mz}} - r_{F\mathrm{dz}}) \quad [16]$$

$$c_F^2 = 2r_{F\mathrm{dz}} - r_{F\mathrm{mz}} \quad [17]$$

$$e_F^2 = V_P - r_{F\mathrm{mz}} \quad [18]$$

When there is measurement error, the correlations between twins' sum scores, $r_{S\mathrm{mz}}$ and $r_{S\mathrm{dz}}$, are attenuated relative to the correlations of the factor scores, $r_{F\mathrm{mz}}$ and $r_{F\mathrm{dz}}$. If MI holds ($k = k_{\mathrm{mz}} = k_{\mathrm{dz}}$), then the set of equations for the sum scores S may be written:

$$a_S^2 + c_S^2 + e_S^2 = k^2 V_P + u^2 \quad [19]$$

$$k^2(a_S^2 + c_S^2) = k^2 r_{F\mathrm{mz}} \quad [20]$$

$$k^2(.5a_S^2 + c_S^2) = k^2 r_{F\mathrm{dz}} \quad [21]$$

where u^2 is additional variation in the sum score due to sources of variation that are specific to each of the observed measures.[2] Initially, we assume that this item-specific variance is uncorrelated between twins, though we relax this assumption later. The solution of these new predicted covariances is:

https://doi.org/10.1375/twin.8.6.553 Published online by Cambridge University Press

$$a_S^2 = 2(k^2 r_{F\mathrm{mz}} - k^2 r_{F\mathrm{dz}}) = k^2 a_F^2 \quad [22]$$

$$c_S^2 = 2k^2 r_{F\mathrm{dz}} - k^2 r_{F\mathrm{mz}} = k^2 c_F^2 \quad [23]$$

$$e_S^2 = k^2 V_P + u^2 - k^2 r_{F\mathrm{mz}} = k^2(V_P - r_{F\mathrm{mz}}) + u^2 = k^2 e_F^2 + u^2 \quad [24]$$

Thus, item-specific variance u^2 causes a_S^2 and c_S^2 to be biased downwards relative to the latent trait values but remain in the same ratio to each other, whereas e_S^2 is biased upwards. The actual changes in the sum score variance component estimates will depend on the amount and source (genetic or environmental) of the item-specific variance.

Variance Component Bias in Sum Scores: Measurement Noninvariance

Under MNIz, that is, when $k_{\mathrm{mz}} \neq k_{\mathrm{dz}}$, the covariance between the sum scores for one zygosity may increase or decrease relative to the sum score covariance for the other zygosity. Specifically, if MZ pairs' latent factors are measured more accurately than DZ pairs' ($k_{\mathrm{mz}} > k_{\mathrm{dz}}$), their covariance will increase, which would bias the heritability estimate upwards. Conversely, if MZ pairs are measured less accurately than are DZ pairs, then estimates of the effects of shared environment would be inflated at the expense of additive genetic variance. Algebraically, this bias can be calculated by substituting k_{mz} and k_{dz} into the expectations above, and solving the equations:

$$a_F^2 + c_F^2 + e_F^2 = \begin{cases} k_{\mathrm{mz}}^2 + u^2 & \text{in MZ twins} \\ k_{\mathrm{dz}}^2 + u^2 & \text{in DZ twins} \end{cases} \quad [25]$$

$$k_{\mathrm{mz}}^2 a_F^2 + k_{\mathrm{mz}}^2 c_F^2 = k_{\mathrm{mz}}^2 r_{F\mathrm{mz}} \quad [26]$$

$$.5k_{\mathrm{dz}}^2 a_F^2 + k_{\mathrm{dz}}^2 c_F^2 = k_{\mathrm{dz}}^2 r_{F\mathrm{dz}} \quad [27]$$

In practice, the terms on the right-hand side are the observed variances and covariances of twins' sum scores. Solving these equations yields:

$$a_{MNI}^2 = 2(k_{\mathrm{mz}}^2 r_{F\mathrm{mz}} - k_{\mathrm{dz}}^2 r_{F\mathrm{dz}}) = 2(k_{\mathrm{mz}}^2 r_{F\mathrm{mz}} - k_{\mathrm{mz}}^2 r_{F\mathrm{dz}} + k_{\mathrm{mz}}^2 r_{F\mathrm{dz}} - k^2 r_{F\mathrm{dz}}) = k_{\mathrm{mz}}^2 a_F^2 + 2(k_{\mathrm{mz}}^2 - k_{\mathrm{dz}}^2) r_{F\mathrm{dz}} \quad [28]$$

$$c_{MNI}^2 = 2k_{\mathrm{dz}}^2 r_{F\mathrm{dz}} - k_{\mathrm{mz}}^2 r_{F\mathrm{mz}} = 2k_{\mathrm{dz}}^2 r_{F\mathrm{dz}} - k_{\mathrm{dz}}^2 r_{F\mathrm{mz}} + k_{\mathrm{dz}}^2 r_{F\mathrm{mz}} - k_{\mathrm{mz}}^2 r_{F\mathrm{mz}} = c_d^2 + (k_{\mathrm{dz}}^2 - k_{\mathrm{mz}}^2) r_{F\mathrm{mz}} \quad [29]$$

$$e_{MNI}^2 = 1 - k_{\mathrm{mz}}^2 r_{F\mathrm{mz}} = 1 - r_{F\mathrm{mz}} + r_{F\mathrm{mz}} - k_{\mathrm{mz}}^2 r_{F\mathrm{mz}} = e_F^2 + (1 - k_{\mathrm{mz}}^2) r_{F\mathrm{mz}} \quad [30]$$

where we have assumed that the difference between the MZ variance and the MZ covariance has been used to estimate e^2, because the MZ and DZ variances are expected to differ.

Estimates of e^2 are not directly affected by MNIz, as they are estimated from the difference between the phenotypic variance and the MZ covariance, and only k_{mz} influences the latter. However, in a model-fitting context, where the difference in total variance between MZ and DZ twins has been obscured, the estimate of the phenotypic variance of MZ twins may be incorrect, and the estimate of e^2 may be biased. Again, it must be emphasized that the above algebra is derived only to establish two main principles: (a) when sum scores of items containing error are analyzed, the proportion of e^2 increases relative to that of the true latent factor as long as u^2 is nonfamilial in origin, and (b) the relative proportions of a^2 and c^2 can be affected by the analysis of sum scores when there is MNIz. The amount of bias will depend on the contributions of genetic and environmental sources to the item-specific variance.

Impact of Familial Item-Specific Variance

In the event that u^2 is not entirely due to measurement error or specific environment factors unique to the item, the relationship between the sum score variance component and the trait score variance components loses the simple proportionality identified in equations 23 to 24. Those with a psychometric background might consider familial components to item-specific variance to be an unlikely occurrence. Yet researchers in twin studies will be familiar with the concept of familial test-specific or item-specific variance (Waller & Reise, 1992); and empirical support is found quite regularly, although it is not ubiquitous (contrast agoraphobia with the other anxiety disorders in Hettema et al., 2005). Also evident is that the individual items on questionnaires such as the EPQ (Eysenck & Eysenck, 1975) vary considerably in their variance components (Neale et al., 1986). Therefore, given a sufficiently large item pool, it would be possible to increase or decrease *any* of the variance components of a sum score. Heritability has been proposed as a criterion for the construction of psychological tests (Jones, 1971). However, heritability per se is not what is generally desired; rather, we seek sets of items that accurately define a common construct. It is this construct which is the focus of genetic epidemiological study; variance components of sum scores may therefore prove to be poor indicators of the variance components of the latent trait.

The relative impact of item-specific versus factor variance components on a sum score is influenced by two main factors. First is the size of the factor loadings, as accumulated into the term k^2. Smaller factor loadings decrease the impact of the latent trait variance on the sum score. The item-specific variance components contribute directly to the sum score and are not attenuated by the size of the factor loadings. The second key component is the number of items, because the factor contributes to the sum score variance not only directly (the variance-based terms λ_i^2) but also by virtue of generating covariance between

https://doi.org/10.1375/twin.8.6.553 Published online by Cambridge University Press

the items (terms of the form $\lambda_i\lambda_j$, $j \neq i$). The total number of contributions to the sum score therefore follows the square of the number of items, whereas the number of item-specific components increases linearly with the number of items. However, the item-specific variance components are not attenuated by the size of the factor loadings, which are accumulated into the term k^2. Thus the factor variance to specific variance ratio can be written: k^2: mu^2 where u here includes familial and nonfamilial item-specific variance components.

Multivariate Analysis as an Alternative

The obvious alternative to fitting an ACE-type model to the univariate sum score is to conduct a multivariate analysis in which the measurement model is included in the decomposition into genetic and environmental variance. Instead of deriving the parameters of the ACE model from the variance and the single covariance of the sum score in the twin pairs, the covariance matrix of individual items or test subscales would be analyzed. This covariance matrix is a partitioned matrix containing across twin covariances in the lower left (upper right) block. In practice, this may be done with analysis of the raw data.

In the multivariate model, all factor loadings are estimated, as well as the variance components for the latent factor and for the residual variances (the common pathway model). It is then possible to test whether the factor loadings are equal for MZ and DZ twins — or even for Twin 1s and Twin 2s. If some of the factor loadings for MZs differ from their counterparts in DZ twins, and if, as is typically done in practice, factor loadings are initially fixed to be equal across genders, there is no simple way in which the parameters a_F, c_F and e_F (the variance components of the latent factor) can compensate for the resulting misfit. Given sufficient sample size, fitting a model with factor loadings restricted to be equal across zygosity, when in fact they are not, would result in a poor overall fit of the measurement model. In a multivariate analysis, equality of factor loadings across zygosity is therefore a hypothesis that can be tested by comparing models with and without the equality restrictions using a likelihood ratio test. Furthermore, it is possible to test whether the phenotypic means (or item thresholds) are equal across zygosity and, given ordinal or continuous level of measurement, whether the item-specific variances (θ_i) are equal for the two groups.

The multivariate model discussed here is known as the 'common pathway' or 'psychometric factors' model (Kendler et al., 1987; McArdle & Goldsmith, 1990; Neale & Cardon, 1992). Historically, fitting this model to binary or ordinal data has faced technical difficulties, but these have largely been overcome via approaches such as marginal maximum likelihood (Bock & Aitkin, 1981; Schmitt et al., 2005).

Illustration With Simulated Data

Continuous Data

To illustrate the bias in the variance components estimates of the latent trait that can occur in the analysis of sum scores derived from noninvariant items, we simulated data according to the model shown in Figure 2. This model begins at the top of the diagram with the correlated latent traits, F1 and F2 of a pair of twins. According to the usual factor model, the observed measures of Twin 1 (M1-T1 to M1-T3) are caused partly by the latent factor, via paths l_1 to l_3, and partly by residual components of variance shown as double-headed arrows ($1 - l_1^2$ and so on). These components are set to result in observed measures with unit variance, though this is not a necessary constraint. The observed measures then generate the observed sum score (SUM-T1) of Twin 1 via paths with fixed values of unity. Using the usual rules of path analysis (Neale & Cardon, 1992), it is possible to obtain the predicted variance of SUM-T1 and SUM-T2 and their covariance, whence their correlation can also be computed. A simple Mx (Neale et al., 1999) script for this purpose is provided on the web site http://www.vcu.edu/mx/examples.html under the zygosity measurement invariance link, although the algebra in equations 23 to 24 above could be used instead.

The effect of the size of the factor loadings on the sum score correlation was considered only for the case where all the factor loadings are equal. The loadings were varied from .9 to .1, and the latent trait correlations were varied from .9 to .1. First, nine continuous measures of the latent trait were considered, which is close to a best case scenario, as few studies will have as many continuously distributed indicators of a latent trait. The predicted correlation between twins' sum scores are depicted graphically in Figure 3. Two points are especially noteworthy. First, the attenuation of correlation in the sum scores compared to the true latent trait correlation is not great when the factor loadings are greater than .7. For example, with factor loadings of .8 the sum score correlation is still .78, compared to .9 for the latent trait correlation. Second, the absolute amount of loss of correlation decreases for lower initial correlations. The strict proportionality of the loss of correlation seen in this figure is exactly as predicted by the equations 11 and 12 above. In the context of the classical twin study, therefore, the MZ correlation suffers greater *absolute* attenuation than that of the DZ. However, per equations 23 to 24 above, the familial variance components for the sum scores remain in the same proportion to each other as they are for the latent trait.

Binary Data

Most behavioral data is initially collected in the form of binary items or ordinal items with relatively few categories (Likert scales). Many so-called continuous measures, such as Eysenck Personality Questionnaire Neuroticism (Eysenck & Eysenck, 1975) — to name but one example — are computed as simple sum

https://doi.org/10.1375/twin.8.6.553 Published online by Cambridge University Press

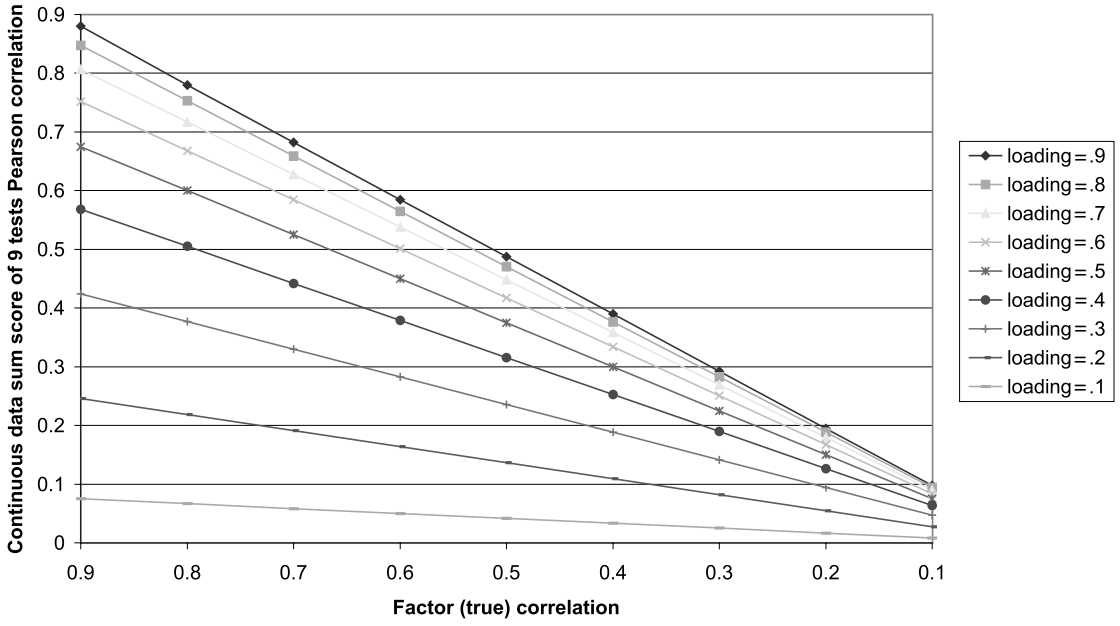

Figure 3

Correlation between sum scores calculated from nine continuous traits as a function of true latent trait correlation and size of factor loading. Sum scores were generated according to the model shown in Figure 2.

scores from a number of binary items. Given that this type of assessment is very common, and that behavior genetic analyses of such measures are popular, it is important to consider the binary case. To address this, we simulated binary data using a simple threshold-based item response probability model. First, pairs of latent trait (factor) scores ($LT1$ and $LT2$) were sampled from a bivariate normal distribution with a unit variances, zero means and correlation r. Second, continuous variable test scores were generated according to the formula

$$MT1 = \lambda_1 LT1 + \sqrt{(1-\lambda_1^2)}ET1$$

where ET1 was sampled from a univariate normal distribution with zero mean and unit variance. Twenty items were sampled, with equal factor loadings (akin to a Rasch model) and a threshold t_j for item j set at $-1.8 + .2j$ to give item difficulties ranging from z scores of –1.8 to + 1.8 at intervals of .2SD. The probability of a positive response for a subject with continuous score $MT1$ was:

$$p\ (\text{response} = 1) = \begin{cases} 0 \text{ if } MT_j < t_j \\ 1 \text{ if } MT_j \geq t_j \end{cases} \qquad [31]$$

The binary item scores were then summed to produce a scale score ranging from zero to 20. This procedure was repeated to generate 100,000 pairs of sum scores for the 81 combinations of latent trait correlations (from .9 to .1) and factor loadings (also from .9 to .1). The polychoric correlation was computed for the twins' sum scores for each dataset, and these results are depicted in Figure 4.

Three features of Figure 4 are noteworthy. First, there is greater attenuation of the correlation with this procedure than in the continuous data case. Even with all 20 items having factor loadings of .9, the latent trait correlation of .9 is estimated to be .85 for the sum score — greater attenuation than for the case with continuous data with only nine measures. Second, this reduction increases quite rapidly as the factor loadings decrease; with factor loadings of .6 the sum score correlation is .6 versus .9 for the latent trait. Third, as in the continuous case, the pattern is linear, so there is less absolute loss of correlation for smaller than for larger correlations.

A scale with 20 items is approaching a best case scenario; in practice, scales are often constructed with a smaller number of items.[3] Therefore, the simulation was repeated for the 10-item and 5-item cases. For the 10-item case, alternate items were selected from the 20-item dataset, starting with the first. The 5-item dataset was constructed by taking every fourth item, starting at the second, to yield items with thresholds at –1.6, –.8, 0, .8 and 1.6 standard deviations from the mean. Results of these simulations are shown in Figures 5a and 5b, respectively. The increase in error due to the smaller number of items is substantial, especially for the

https://doi.org/10.1375/twin.8.6.553 Published online by Cambridge University Press

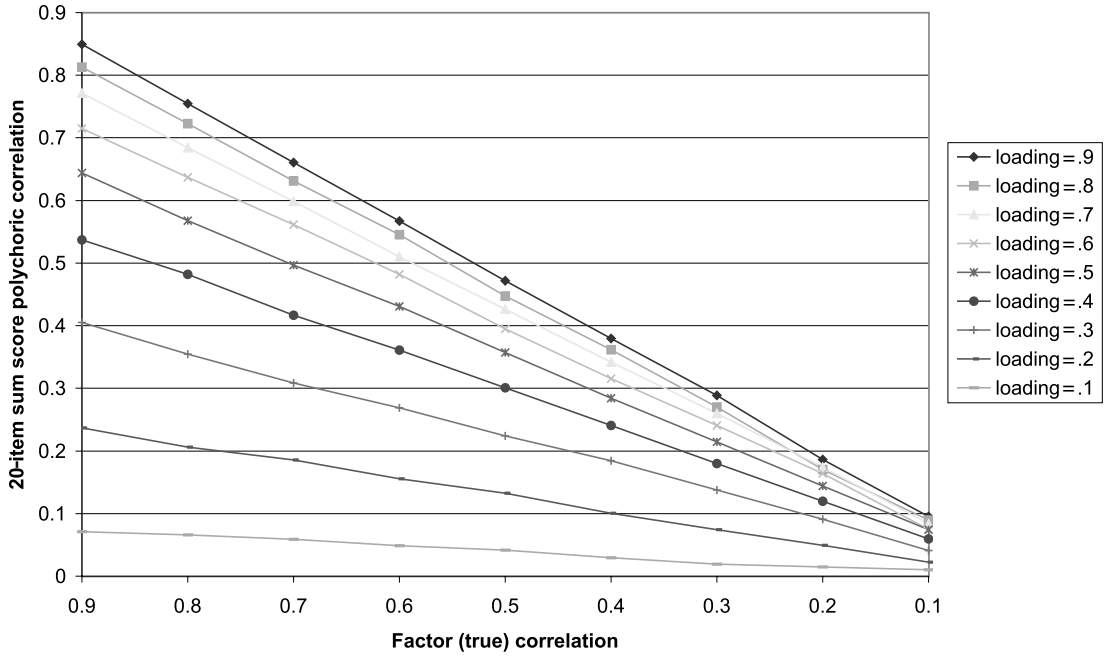

Figure 4

Correlation between sum scores calculated from 20 binary traits as a function of true latent trait correlation and size of factor loading. Sum scores were generated according to the model shown in Figure 2.

5-item case, where the initial factor correlation of .9 is reduced to .53 when factor loadings of .7 are used. Bearing in mind that loadings of .4 or more are often retained during factor analysis for the purposes of scale construction, the difference in familial resemblance for sum scores and for the latent traits which they are supposed to measure could be substantial.

Effects on Variance Components

Measurement Invariance

The implications of these findings for variance components estimates based on sum scores were considered under two conditions. First, even with measurement invariance, where factor loadings are equal across the two zygosities, equations 23 to 24 show that substantial bias may accrue from the use of sum scores. The bias is depicted for the two lines marked a^2 and c^2 in Figure 6. The values of a^2 and c^2 were computed manually using the formulae $a^2 = 2(rMZ - rDZ)$ and $c^2 = rMZ - a^2$ (equivalent to using the least squares fit function and appropriate for illustration). Given factor loadings of .8, the sum score variance components are $\hat{a}_S^2 = .544$ and $\hat{c}_S^2 = .176$ instead of the values of $a_F^2 = .6$ and $c_F^2 = .2$ for the latent trait. This bias increases as the factor loadings decrease.

Measurement Noninvariance

Two forms of MNIz were considered. First, the MZ factor loadings relating the items to the latent trait were set to be .1 greater than those of the DZ twins. Figure 7 overlays two lines, marked $\hat{a2}$ MZ > DZ and $\hat{c2}$ MZ > DZ, which plot the variance components for this form of measurement noninvariance. Adding this form of MNI slightly counteracts the adverse effects of using sum scores. The estimate $\hat{a}_S^2$ is increased compared to the value under measurement invariance condition, and $\hat{c}_S^2$ is decreased. Indeed, $\hat{c}_S^2$ becomes negative when the factor loadings are small; under the usual maximum likelihood estimation procedure, which imposes a zero lower bound on variance components, this parameter would be estimated at zero, and $\hat{a}_S^2$ would be inflated.

Second, we examined the case where DZ twins' factor loadings were .1 greater than those of the MZ twins. Results under this condition are shown in Figure 8. As expected, heritability is biased downwards, and common environment variance is biased upwards in this condition.

Discussion

This article demonstrates that variance component estimates based on sum scores are likely to be biased estimates of latent trait variance components. This is a serious concern for twin studies, given that many psychological scales are sum scores. In addition, many DSM diagnoses — presence or absence of major depression, or of substance use disorder —

https://doi.org/10.1375/twin.8.6.553 Published online by Cambridge University Press

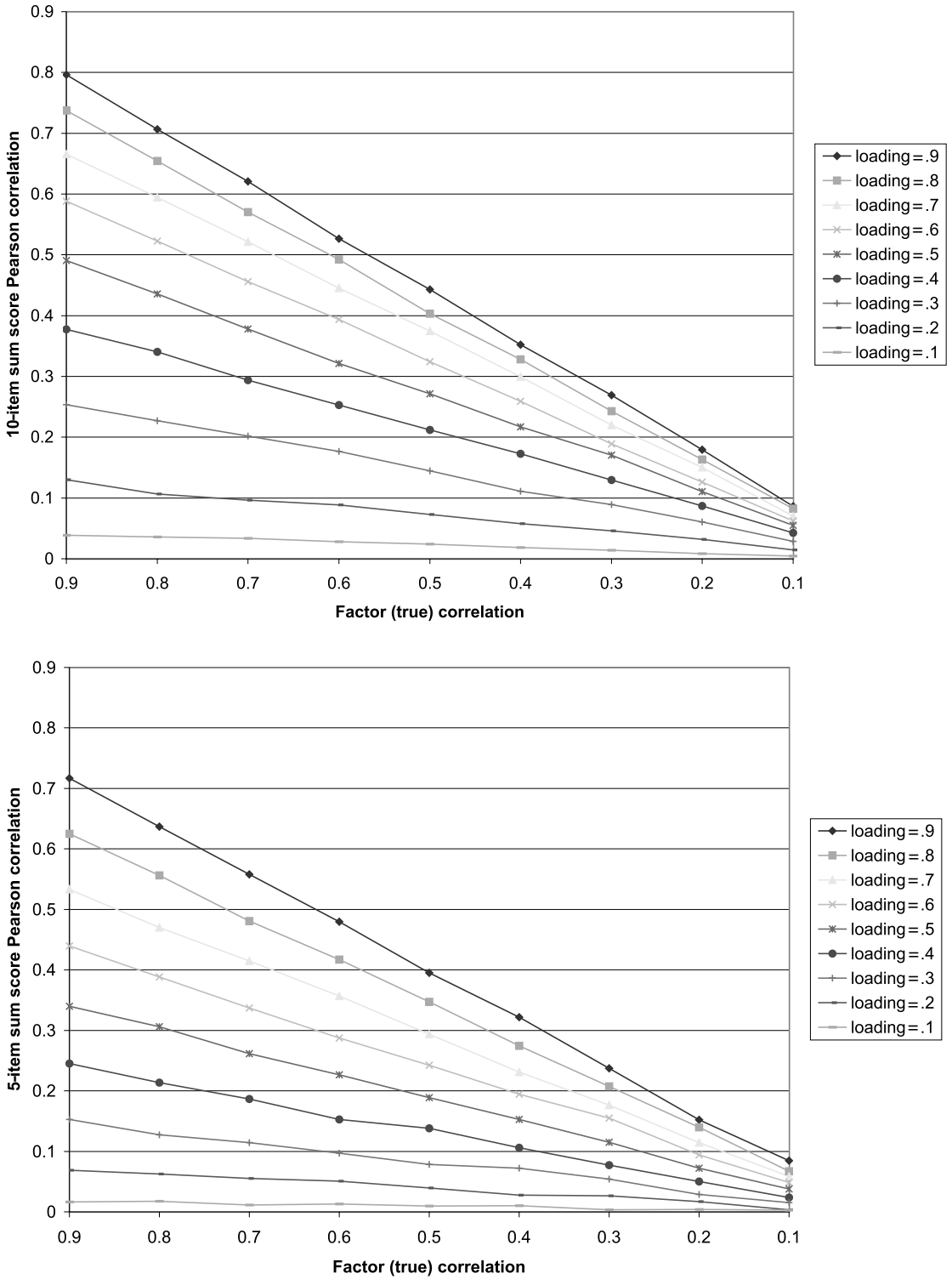

Figure 5

Bias in correlations as a function of factor loading and initial correlation.

Observed variables are sum scores derived from (top) 10 binary items and (bottom) 5 binary items.

https://doi.org/10.1375/twin.8.6.553 Published online by Cambridge University Press

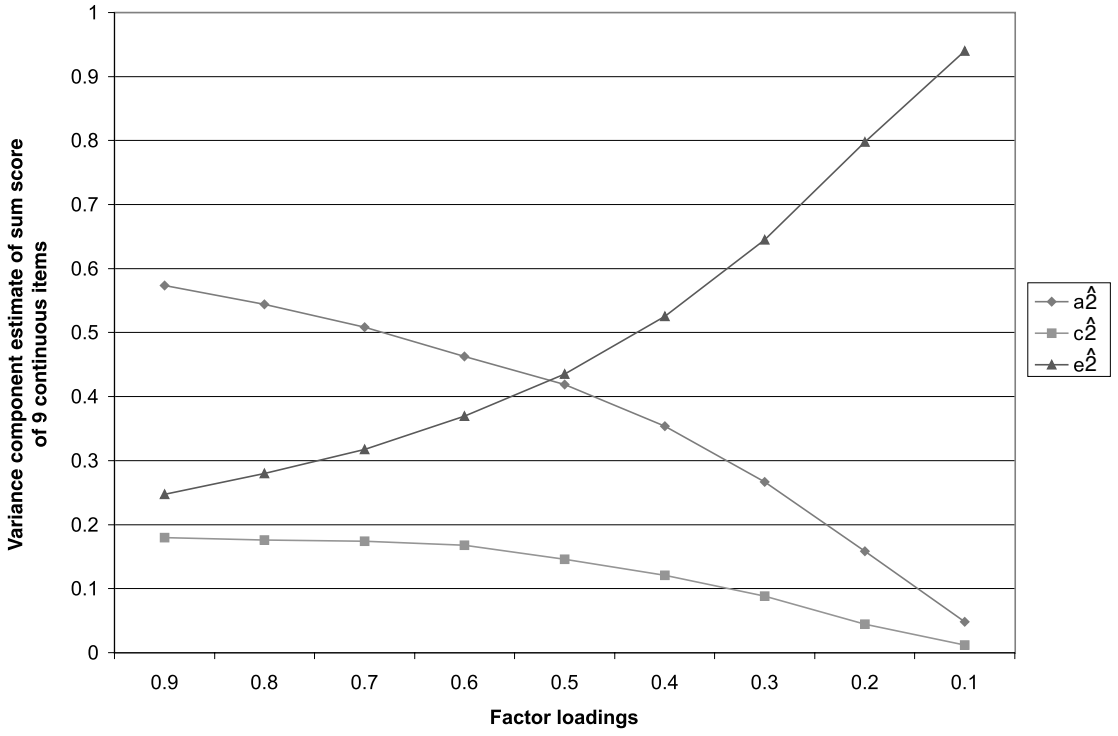

Figure 6

Estimated variance components for a scale based on nine continuous items, as a function of factor loading size.

Latent trait correlations for MZ and DZ twins were set at .8 and .6, corresponding to variance components of $a^2 = .6$ and $c^2 = .2$. Measurement invariance across MZ and DZ twins (MIz) holds.

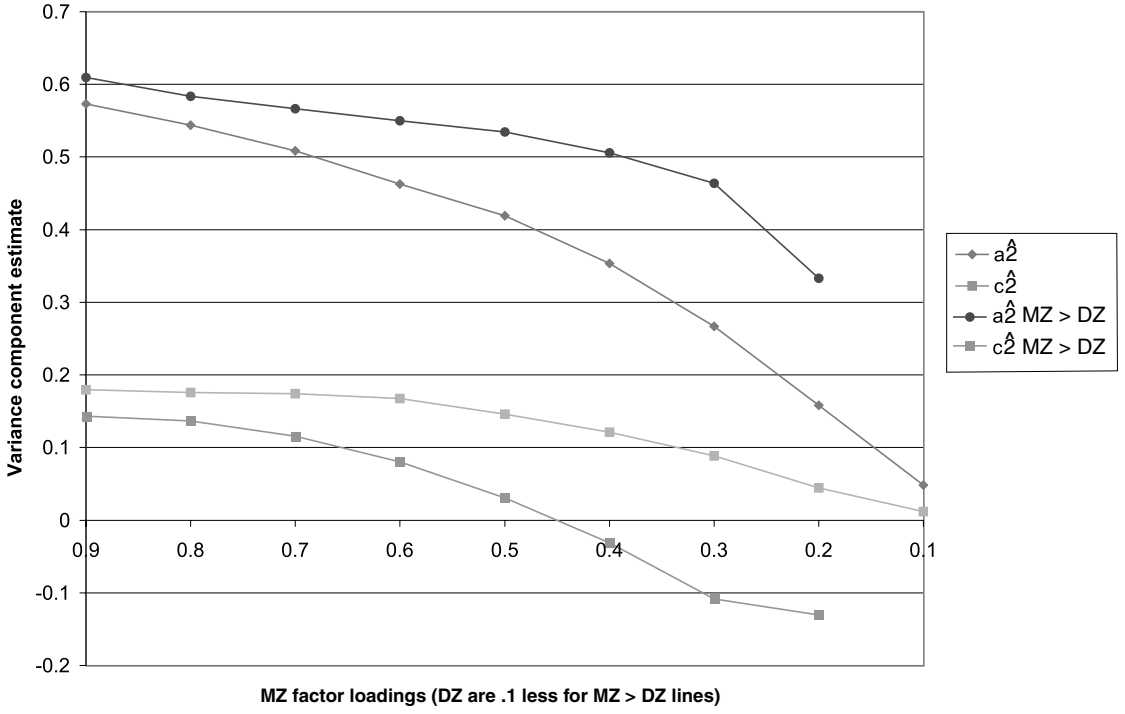

Figure 7

Estimated variance components for a scale based on 20 binary items.

Latent trait correlations for MZ and DZ twins were set at .8 and .6, corresponding to variance components of $a^2 = .6$ and $c^2 = .2$. Lines denoted $\hat{a2}$ and $\hat{c2}$ are the estimates from sum scores; those with MZ > DZ appended are for the case of greater factor loadings for MZ pairs.

https://doi.org/10.1375/twin.8.6.553 Published online by Cambridge University Press

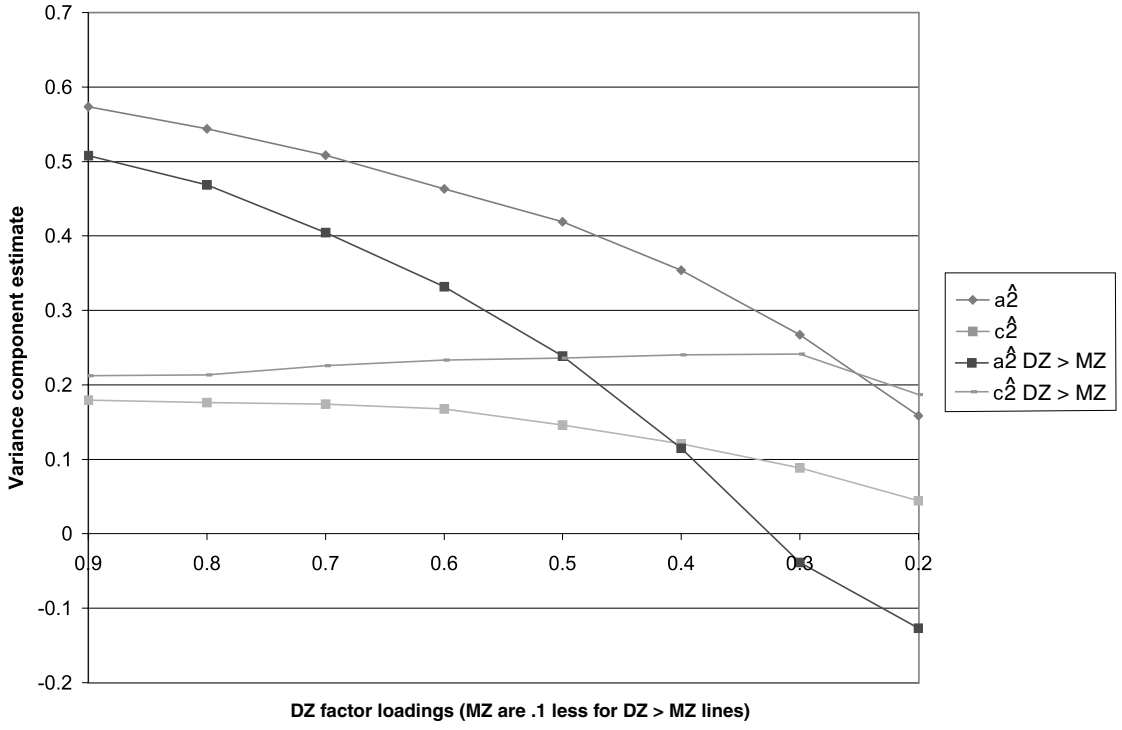

Figure 8

Estimated variance components for a scale based on 20 binary items.

Latent trait correlations for MZ and DZ twins were set at .8 and .6, corresponding to variance components of $a^2 = .6$ and $c^2 = .2$. Lines denoted $\hat{a^2}$ and $\hat{c^2}$ are the estimates from sum scores; those with MZ > DZ appended are for the case of greater factor loadings for MZ pairs.

are themselves based on symptom counts, which are a form of sum score.

We also show that variance component bias occurs when measurement invariance with respect to zygosity fails. A key question for discussion is how likely such failure of MI would be in practice. Four possibilities seem plausible. One is that with self-report data, twins might judge themselves in some relative fashion by comparing themselves to their twin. For example, antisocial behavior is frequently accompanied by little insight when assessed by self-report, whereas external observers — such as their relatives or teachers — appear to observe it with better agreement and precision (Kendler et al., 2002). In the absence of good introspection data, twins may tend to score themselves as similar to their co-twin. If MZ twins are initially more similar (by virtue of genetic factors) the effects of the genetic factors could be amplified by the bias accrued from partly rating the co-twin instead of oneself. A second, perhaps more common, situation is in the analysis of parent or teacher ratings, such as are often used in studies of juvenile twins. If, in cases of doubt about Twin A, the rater uses their knowledge of Twin B to supplement their assessment of Twin A, then again measurement noninvariance may be seen as a function of zygosity. Were such effects operating on the items from which sum-scores or factor scores are derived, it seems likely that a difference in total variance between MZ and DZ pairs may be seen. Unfortunately, such variance differences may be confounded with genuine sibling interaction (Carey, 1986; Eaves, 1976; Neale, 1985; Neale & Cardon, 1992) or other processes, such as parental contrast, which also generate differences in total variance. In fact, MNIz itself would typically generate differences between the total variances of MZ and DZ twins, so it seems unlikely that MNIz contributes to variance component bias for many traits because such variance differences are rarely observed.

For certain measures, two further possibilities exist. For physical traits, such as finger print ridge count or skin-folds from different areas of the body, it seems very unlikely that the measurement would not perform equally well in MZ and DZ twins. Some possible mechanisms remain, however. First, if there are genetic factors involved in MZ or DZ twinning which have a pleiotropic influence on the precision of measurement of certain characteristics, measurement noninvariance might result. This possibility seems, a priori, very unlikely for most traits. Second, a practice effect could occur in some study designs. For example, suppose that the researcher taking the skin-fold (or ridge count, etc.) makes a more accurate assessment on the second twin than on the first, by virtue of

https://doi.org/10.1375/twin.8.6.553 Published online by Cambridge University Press

having recently measured someone morphologically very similar. In this case, we might expect a measurement-order effect, but only in MZ twin pairs. In principle, this effect could be detected quite accurately if the order of assessment within a twin pair is known, by testing for measurement invariance across Twin 1 and Twin 2. Third, suppose that the distribution of age differs between the MZ and DZ twins. If the trait under consideration does not show measurement invariance with respect to age, then MNIz will result. This failure of invariance with respect to a characteristic correlated with zygosity may be the most significant concern.

The analysis of sum scores precludes the detection of group specific factor loadings, which are a form of noninvariance. If undetected, absence of measurement invariance associated with zygosity will cause bias in the estimates of variance components. The direction of this bias is to increase estimates of additive genetic variance and to decrease estimates of common environment variance if MZ twins' measures are better indicators of the latent trait than those of DZ twins. This pattern of bias is reversed when DZ twins' assessments are the better latent trait indicators. Most noteworthy is that even when measurement invariance holds with respect to zygosity, using sum scores of items — even under the relatively ideal circumstances of 20 binary items considered here — there is considerable attenuation of heritability and some inflation of random environment variance components. Multivariate analysis of the items themselves would effectively eliminate this bias, and permit tests of measurement invariance across groups. Recent advances in statistical methods, including generalized marginal maximum likelihood (Aggen et al., 2005; Schmitt et al., 2005), asymptotic weighted least squares including missing values (Muthén & Muthén, 2004), and Bayesian approaches (Eaves & Erkanli, 2003), together with advances in computer hardware performance, make these methods relatively practical today.

It remains common practice to fit genetic variance components models to data which consist of sum scores. Frequently, such scores depart from normality, with floor or ceiling effects generating severe skewness or kurtosis or both. Indeed, for studies of psychopathology it is not uncommon to observe a reverse J-shaped distribution for the sum scores. One approach to analyze such data is to take a logarithm or square root transformation, possibly following age regression, in order to yield a trait distribution which is approximately normal. The analysis of such transformed sum scores may yield variance components estimates which depart even further from those of the latent trait than do sum scores themselves.

Several limitations of the present article should be considered. First, the model used here is akin to a common pathway model, but without residual measure-specific familial components. Clearly, genetic or shared environment factors that influence particular items (as opposed to the general factor) would increase the sum score correlation between relatives. Therefore, when we move from the sum-score approach to estimating the full model, we may not see the large increases of familial correlation that Figures 3 and 4 would imply. However, the variance components estimated for the hypothesized general factor would not be contaminated with item-specific variance, which would be a considerable advantage. Second, it may be argued that the common pathway model typically does not fit very well compared to the 'independent factors' or 'biometric factors' model in which three factors are specified, and the correlation between twins' factors are set to 1.0 for the shared environment, zero for the specific environment, and to 1.0 or .5 for the MZ or DZ additive genetic factor. This well-known model is in fact a submodel of the less widely applied three factor common pathway model. Application of the constraints that: $c = e = 0$ for the first factor, $a = e = 0$ for the second factor, and $a = c = 0$ for the third factor of the three factor common pathway model reduces it to the single factors independent pathway model. Thus, addition of a sufficient number of latent (common pathway) factors would overcome the problem of poor model fit of the single common pathway model.

One valuable feature of using the latent trait model directly is that it should increase statistical power to detect quantitative trait loci (QTL) that influence the latent trait. Genome scanning using a multivariate model with ordinal data still presents statistical challenges, requiring multivariate analysis at repeated intervals on the genome. However, this computational burden may be easily shared by a cluster or grid computing infrastructure, which is growing in popularity. Nevertheless, to run permutation or bootstrap tests to evaluate evidence for linkage, even a large cluster would be heavily burdened. A compromise, in the mean time, would be to only run such tests in and around regions where substantial signals exist. Genetic factor scores (Molenaar et al., 1990) may also provide an efficient starting point, although these suffer from the limitation that, except under unusual ideal measurement conditions, different factor scores have different amounts of intrinsic error.

Although the significance of this article may be considered to apply only to twins, a simple change of label to the latent traits at the top of Figure 2 allows consideration of the relationship between traits — comorbidity in the case of disorder-related latent traits such as psychopathology or substance use — or between them and their risk factors. That is, we could consider Figure 2 as a model for different latent traits in the same individual. Note that in this case, the correlation between the item-specific components may be less across traits within person than they are within trait across twins. This would likely be the case for different traits (say, sum scores of cognitive ability items correlated with sum scores of depression items).

https://doi.org/10.1375/twin.8.6.553 Published online by Cambridge University Press

Here the sum score correlations would be very likely to underestimate the latent trait correlation. On the other hand, repeated measures of the same trait in the same individual could have higher item-specific covariances, in which case the sum score correlations might not underestimate the latent trait correlation. The beauty of item-level analysis is that origin of the correlation between sum scores can be partitioned into that due to the latent trait correlation and that due to item-specific components.

A final point to note is that the multivariate or longitudinal study of psychopathology or substance use — or virtually any aspect of behavior — could benefit from the analysis of the latent trait, as opposed to the sum score. Aggregate scores presuppose that a particular and rather unlikely measurement model holds — that there is no item-specific variance and that all factor loadings are equal. Yet there is no need to make this assumption, and it is clearly practical to test it. At least in principle, *the behavioral scientist should attempt to analyze what has been measured*. Nothing, other than computer time, is lost and much is gained from the analysis of data at their original level of measurement.

Endnotes

1 For the equivalence relation of factor loadings in the common factor model and discrimination parameters in IRT models, see Lord & Novick, 1968.

2 $u^2 = \mathrm{Var}(\varepsilon_1) + \mathrm{Var}(\varepsilon_2)$ in equation [3].

3 Often the factor loadings vary substantially between items, but that is not explored here.

Acknowledgments

This research was supported by NIH grants MH-65322 and DA-018673.

References

Aggen, S. H., Neale, M. C., & Kendler, K. S. (2005). DSM criteria for major depression: Evaluating symptom patterns using latent-trait item response models. *Psychological Medicine, 35,* 475–487.

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., text rev.). Washington, DC: Author.

Bloxom, B. (1972). Alternative approaches to factorial invariance. *Psychometrika, 37,* 425–440.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46,* 443–459.

Byrne, B. M., Shavelson, R. J., & Muthén, B. O. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105,* 456–466.

Carey, G. (1986). Sibling imitation and contrast effects. *Behavior Genetics, 16,* 319–341.

Dolan, C. V. (2000). Investigating Spearman's hypothesis by means of multi-group confirmatory factor analysis. *Multivariate Behavioral Research, 35,* 21–50.

Dolan, C. V., Roorda, W., & Wicherts, J. M. (2004). Two failures of spearman's hypothesis: The GATB in Holland and the JAT in South Africa. *Intelligence, 32,* 155–173.

Eaves, L., & Erkanli, A. (2003). Markov chain Monte Carlo approaches to analysis of genetic and environmental components of human developmental change and g × e interaction. *Behavior Genetics, 33,* 279–299.

Eaves, L. J. (1976). A model for sibling effects in man. *Heredity, 36,* 205–214.

Ellis, J. L. (1993). Subpopulation invariance of patterns in covariance matrices. *British Journal of Mathematical and Statistical Psychology, 46,* 231–254.

Eysenck, H. J., & Eysenck, S. B. G. (1975). *Manual of the Eysenck Personality Questionnaire.* San Diego, CA: Digits.

Hettema, J. M., Prescott, C. A., Myers, J. M., Neale, M. C., & Kendler, K. S. (2005). The structure of genetic and environmental risk factors for anxiety disorders in men and women. *Archives of General Psychiatry, 62,* 182–189.

Holland, P. W., & Wainer, H. (1993). *Differential item functioning.* Hillsdale, NJ: Erlbaum.

Jones, M. B. (1971). Heritability as a criterion in the construction of psychological tests. *Psychological Bulletin, 75,* 92–96.

Kendler, K. S., Heath, A. C., Martin, N. G., & Eaves, L. J. (1987). Symptoms of anxiety and symptoms of depression: Same genes, different environments? *Archives of General Psychiatry, 44,* 451–457.

Kendler, K. S., Prescott, C. A., Jacobson, K., Myers, J., & Neale, M. C. (2002). The joint analysis of personal interview and family history diagnoses: Evidence for validity of diagnosis and increased heritability estimates. *Psychological Medicine, 32,* 829–842.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison Wesley.

Lubke, G. H., Dolan, C. V., Kelderman, H., & Mellenbergh, G. J. (2003). On the relationship between sources of within- and between-group differences and measurement invariance in the context of the common factor model. *Intelligence, 173,* 1–24.

Lubke, G. H., Dolan, C. V., & Neale, M. C. (2004). Implications of absence of measurement invariance for detecting sex limitation and genotype by environment interaction. *Twin Research, 7,* 292–298.

Marsh, H. W. (1994). Confirmatory factor models of factorial invariance: A multi-faceted approach. *Structural Equation Modeling, 1,* 5–34.

McArdle, J. J. (1998). Contemporary statistical models of test bias. In J. J. McArdle & R. W. Woodcock (Eds.),

Human abilities in theory and practice (pp. 157–195). Mahwah, NJ: Erlbaum.

McArdle, J. J., & Goldsmith, H. H. (1990). Alternative common-factor models for multivariate biometric analyses. *Behavior Genetics, 20,* 569–608.

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research, 13,* 127–143.

Mellenbergh, G. J. (1994). Generalized linear item response theory. *Psychological Bulletin, 115,* 300–307.

Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. *Psychometrika, 58,* 525–543.

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement, 17,* 297–334.

Molenaar, P. C. M., Boomsma, D. I., Neeleman, D., & Dolan, C. V. (1990). Using factor scores to detect ge interactive origin of 'pure' genetic or environmental factors obtained in genetic covariance structure analysis. *Genetic Epidemiology, 7,* 83–100.

Muthén, L. K., & Muthén, B. O. (2004). *Mplus 3.11* [Computer program]. Los Angeles: Muthén & Muthén.

Neale, M. C. (1985). *Biometrical genetic analysis of human individual differences*. London: University of London.

Neale, M., Boker, S., Xie, G., & Maes, H. (1999). *Mx: Statistical modeling* (5th ed.). Richmond, VA: Department of Psychiatry, Virginia Commonwealth University.

Neale, M., Boker, S., Xie, G., & Maes, H. (2003). *Mx: Statistical modeling* (6th ed.). Richmond VA: Department of Psychiatry, Virginia Commonwealth University.

Neale, M. C., & Cardon, L. R. (1992). *Methodology for genetic studies of twins and families*. Dordrecht, the Netherlands: Kluwer Academic.

Neale, M. C., Rushton, J. P., & Fulker, D. W. (1986). The heritability of items from the Eysenck Personality Questionnaire. *Personality and Individual Differences, 7,* 771–779.

Schmitt, J. E., Mehta, P. D., Aggen, S. H., Kubarych, T. S., & Neale, M. C. (2005). *Semi-nonparametric methods for detecting latent non-normality: A fusion of latent trait and ordered latent class modeling*. Manuscript submitted for publication.

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52,* 393–408.

Waller, N. G., & Reise, S. P. (1992). Genetic and environmental influences on item response pattern scalability. *Behavior Genetics, 22,* 135–152.

Appendix

The purpose here is to clarify the relationship between genetic and environmental variance components of a latent factor, and the estimates obtained from the analysis of sum scores of items that measure, with error, this latent factor. Beginning with the latent factors of twins, as at the top of Figure 2, we write the covariance matrix of the factors of the twins, F_1 and F_2, as:

$$\Psi_{\text{mz}} = \begin{pmatrix} a_F^2 + c_F^2 + e_F^2 & a_F^2 + c_F^2 \\ a_F^2 + c_F^2 & a_F^2 + c_F^2 + e_F^2 \end{pmatrix} = \begin{pmatrix} \text{V}_{F\text{mz}} & r_{F\text{mz}} \\ r_{F\text{mz}} & \text{V}_{F\text{mz}} \end{pmatrix}$$

The matrix Λ contains the loadings (path coefficients) from these latent factors to the twins' measured items, *M*1 – *T*1 to *M*2 – *T*2 (neglecting *M*3 – *T*1 and *M*3 – *T*1 in the figure as they add nothing to the present discussion) as follows:

$$\Lambda_{\text{mz}} = \begin{pmatrix} l_1 & 0 \\ l_2 & 0 \\ 0 & l_1 \\ 0 & l_2 \end{pmatrix}$$

and matrix Θ is diagonal with elements θ_j for measure j. In principle, Θ may differ between MZ and DZ twins, although we do not explore that possibility here. The covariance of the measured items, Σ, is given by:

$$\Sigma_{\text{mz}} = \Lambda_{\text{mz}} \Psi_{\text{mz}} \Lambda_{\text{mz}}' + \Theta \qquad [32]$$

https://doi.org/10.1375/twin.8.6.553 Published online by Cambridge University Press

$$= \begin{pmatrix} v\lambda_1^2 + \theta_1^2 & & & \\ v\lambda_1\lambda_2 & v\lambda_2^2 + \theta_2^2 & & \\ r_{\mathbf{F}\mathrm{mz}}\lambda_1^2 & r_{\mathbf{F}\mathrm{mz}}\lambda_1\lambda_2 & v\lambda_1^2 + \theta_1^2 & \\ r_{\mathbf{F}\mathrm{mz}}\lambda_1\lambda_2 & r_{\mathbf{F}\mathrm{mz}}\lambda_2^2 & v\lambda_1\lambda_2 & v\lambda_2^2 + \theta_2^2 \end{pmatrix} \quad [33]$$

To compute the covariance between the sum scores, we use an elementary matrix which relates the measured variables to the sums with unit loadings:

$$\mathrm{W} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix}$$

to obtain the covariance of twins's sum scores:

$$\Sigma_{\mathrm{SMZ}} = \mathrm{W}\Sigma\mathrm{W}' \quad [34]$$

$$= \begin{pmatrix} v\lambda_1^2 + \theta_1^2 + v\lambda_2^2 + \theta_2^2 + 2v\lambda_1\lambda_2 & \\ r_{\mathbf{F}\mathrm{mz}}(\lambda_1^2 + \lambda_2^2 + 2\lambda_1\lambda_2) & v\lambda_1^2 + \theta_1^2 + v\lambda_2^2 + \theta_2^2 + 2v\lambda_1\lambda_2 \end{pmatrix} \quad [35]$$

$$= \begin{pmatrix} v(\lambda_1 + \lambda_2)^2 + \theta_1^2 + \theta_2^2 & \\ r_{\mathbf{F}\mathrm{mz}}(\lambda_1 + \lambda_2)^2 & v(\lambda_1 + \lambda_2)^2 + \theta_1^2 + \theta_2^2 \end{pmatrix} \quad [36]$$

Substituting $u^2 = \theta_1^2 + \theta_2^2$ and $k = \lambda_1 + \lambda_2$, we obtain:

$$\begin{pmatrix} vk^2 + u^2 & \\ r_{F\mathrm{mz}}k^2 & vk^2 + u^2 \end{pmatrix}$$

The expected covariances between twins' factor scores under the usual ACE variance components model are:

$$\Psi_{\mathrm{mz}} = \begin{pmatrix} a_F^2 + c_F^2 + e_F^2 & a_F^2 + c_F^2 \\ a_F^2 + c_F^2 & a_F^2 + c_F^2 + e_F^2 \end{pmatrix} = \begin{pmatrix} \mathrm{V}_{F\mathrm{mz}} & r_{F\mathrm{mz}} \\ r_{F\mathrm{mz}} & \mathrm{V}_{F\mathrm{mz}} \end{pmatrix}$$

and

$$\Psi_{\mathrm{dz}} = \begin{pmatrix} a_F^2 + c_F^2 + e_F^2 & .5a_F^2 + c_F^2 \\ .5a_F^2 + c_F^2 & a_F^2 + c_F^2 + e_F^2 \end{pmatrix} = \begin{pmatrix} \mathrm{V}_{Fdz} & r_{Fdz} \\ r_{Fdz} & \mathrm{V}_{Fdz} \end{pmatrix}$$

so the expected covariances between twins' sum scores are therefore:

$$W\Sigma_{mz}W' = \begin{pmatrix} k^2(a_F^2 + c_F^2 + e_F^2) + u^2 & k^2(a_F^2 + c_F^2) \\ k^2(a_F^2 + c_F^2) & k^2(a_F^2 + c_F^2 + e_F^2) + u^2 \end{pmatrix}$$

and

$$W\Sigma_{dz}W' = \begin{pmatrix} k^2(a_F^2 + c_F^2 + e_F^2) + u^2 & k^2(.5a_F^2 + c_F^2) \\ k^2(.5a_F^2 + c_F^2) & k^2(a_F^2 + c_F^2 + e_F^2) + u^2 \end{pmatrix}$$

https://doi.org/10.1375/twin.8.6.553 Published online by Cambridge University Press