Implications of Absence of Measurement Invariance for Detecting Sex Limitation and Genotype by Environment Interaction

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

[1] *Virginia Commonwealth University, USA*

[2] *University of Amsterdam, the Netherlands*

Using univariate sum scores in genetic studies of twin data is common practice. This practice precludes an investigation of the measurement model relating the individual items to an underlying factor. Absence of measurement invariance across a grouping variable such as gender or environmental exposure refers to group differences with respect to the measurement model. It is shown that a decomposition of a sum score into genetic and environmental variance components leads to path coefficients of the additive genetic factor that are biased differentially across groups if individual items are non-invariant. The arising group differences in path coefficients are identical to what is known as "scalar sex limitation" when gender is the grouping variable, or as "gene by environment interaction" when environmental exposure is the grouping variable. In both cases the interpretation would be in terms of a group-specific effect size of the genetic factor. This interpretation may be incorrect if individual items are non-invariant.

The individual items in most measurement instruments are designed to measure an underlying factor or latent trait. However, the items are rarely pure indicators of the underlying factor. For example, an item designed to measure attention deficit/hyperactivity disorder (ADHD) may also be sensitive to a learning disorder. If a test is administered to more than one group and an individual item does not measure the same factors in those groups, the item is not measurement invariant over the groups. Measurement invariance refers to the situation in which a measurement instrument measures the same underlying factor(s) or latent trait(s) in different groups. The present article shows that absence of measurement invariance can be confounded with sex limitation or gene by environment interaction.

Measurement invariance holds with respect to a grouping variable if the probability of an observed item response is the same for members of different groups with the same score on the factor that a given measurement instrument is designed to measure. An observed item is not measurement invariant with respect to, say, gender, if one gender scores higher on average on the item than the other gender without actually scoring higher on the underlying factor. The higher observed scores are because the non-invariant items are sensitive to other variable(s) in addition to the factor, and that these additional variables increase the observed scores. Suppose that a questionnaire is designed to measure ADHD and that one of the items is also sensitive to a learning disorder in males but not in females. The resulting observed average score on this item is higher in males than in females because the item measures an additional factor in males. If relations between items and underlying factors are represented in a measurement model, such as the common factor model or an item response model for categorical outcomes, then one or more of the parameter(s) concerning the non-invariant items are not the same across groups.

Measurement invariance can only be investigated if the measurement model relating observed item scores to the underlying factor(s) is analyzed simultaneously in all groups. Simultaneous analysis of the item scores allows a test of whether the measurement model is the same across groups. If sum scores are computed by adding 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. If factor scores are computed in more than one group using group-invariant factor loadings, then the same problem applies because factor scores are essentially weighted sum scores.

In genetic analyses it is common practice to derive sum scores by summing individual questionnaire

Received 13 December, 2003; accepted 25 February, 2004.

Address for correspondence: Gitta H. Lubke, Virginia Institute of Psychiatric and Behavioral Genetics, Virginia Commonwealth University, 800 East Leigh St. Suite 1–115, Richmond VA 23219-1534, USA. Email: glubke@vcu.edu

https://doi.org/10.1375/twin.7.3.292 Published online by Cambridge University Press

item or symptom scores. The Eysenck Personality Questionnaire scores of Extraversion, Neuroticism, and Psychoticism are computed in this way (Eysenk & Eysenk, 1975). Similarly, Diagnostic and Statistical Manual (DSM) diagnoses, which are typically established when a given subject displays at least n of a given list of m symptoms, where $n < m$, are essentially sum scores (American Psychological Association, 2000). Sum scores are often used to investigate the genetic and environmental contributions to the total sum score variance by applying ACE-type models. The present paper focuses on genetic analyses that are meant to compare heritabilities across gender or environmental exposure groups. It is shown that a decomposition of sum scores using ACE-type models results in biased path coefficients when the individual items from which the sum score is derived are not measurement invariant. More specifically, if gender differences are investigated, the arising pattern of path coefficients of the genetic factor across same and opposite sex zygosity groups is identical to what is known as scalar sex limitation. Similarly, if individual items are not measurement invariant with respect to a variable representing environmental exposure, the decomposition of sum scores can result in a pattern of path coefficients that would lead to the conclusion that the genetic factors have different effect sizes across exposure groups.

Sum scores are usually regarded as an estimate for the underlying factor score. Obviously, a 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 in all groups that are investigated. Fitting an ACE-type model to a sum score is based on the implicit assumption that the items from which the sum score is derived measure the same underlying factor or latent trait — in other words, that the items are measurement invariant. However, measurement invariance has to be established empirically. To detect non-invariance with respect to a grouping variable, the relationship between individual items and underlying latent variables has to be modeled simultaneously in all groups in a multivariate analysis.

After briefly introducing the concept of measurement invariance, it is shown that a variance decomposition of sum scores derived from non-invariant items results in a pattern of path coefficients for opposite sex twin pairs that is usually interpreted in terms of different effect sizes of the genetic factor across genders or in case of gene by environment interactions, as different effect sizes of the genetic factor across environmental exposure groups. Hence, when analyzing sum scores, absence of measurement invariance is easily confounded with this type of sex limitation or gene by environment interaction. It is shown that detection of absence of measurement invariance can be carried out in a multivariate analysis where the relation between individual items and underlying factor is modeled explicitly.

Measurement Invariance

Absence of measurement invariance, which is also known as differential item functioning, has been studied extensively both in the context of confirmatory factor analysis and item response theory (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 gender, and concerns the measurement model relating observed scores to underlying latent variables (Mellenbergh, 1989; Meredith, 1993). The measurement model has to 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 has been defined by Mellenbergh (1989) as

$$f(Y \mid \eta, s) = f(Y \mid \eta), \tag{1}$$

where observed variables are denoted as Y, latent variables as η, and the grouping variable as S. Equation 1 shows that, given the scores on the underlying latent variable(s) (i.e., the latent traits or factor scores), the probability distribution of observed scores does not depend on sub-population membership such as gender, but depends only on the scores on the underlying factor(s).

Consider a situation where measurement invariance is absent: that is, $f(Y \mid \eta, s) \neq f(Y \mid \eta)$. An observed variable Y is non-invariant with respect to a grouping variable S if the observed score depends not only on the latent variables η but also on S, or variable(s) related to S. There are three different types of effects of S or variable(s) related to S, that may or may not occur simultaneously. First, the effect can be constant for all possible scores on η. This results in a group difference in the intercept of the regression of Y on η. The regression lines in the common factor model or regression curves in item response models are parallel. Second, the effect can vary independently of η, which results in a group difference with respect to the precision with which the latent variables η are measured. The regression lines (in case of linear regression of Y on η) or the regression curves (in case of non-linear regression on η) are equal across groups, but the residuals of the regression differs. In the common factor model, this is modeled as group differences in measurement residuals. Thirdly, the effect can increase or decrease as a function of η, resulting in a group difference with respect to the steepness of the regression. In the common factor model this is manifest in group-specific factor loadings, whereas in item response models the discrimination parameters would be group-specific.

https://doi.org/10.1375/twin.7.3.292 Published online by Cambridge University Press

The third type of effect is the primary focus of the present article.

If factor loadings or discrimination parameters differ across groups, the interpretation of the underlying factor or trait differs as a consequence.[1] Suppose that in one group math items load strongly on a general Intelligence Quotient (IQ)-test factor and verbal items have weaker loadings. If the other group shows the reverse pattern, then the test is not measurement invariant. The math items have a larger weight than the verbal items in the first group and the general factor has to be interpreted accordingly, namely as a predominantly math-related factor. The interpretation is different in the second group, showing that the test does not measure the same general factor in the two groups. 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 case of non-invariance, incorrectly interpreted as an estimate of the same factor across groups.

The incorrect interpretation of a sum score is especially important if 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 non-invariant items. However, the problem of analyzing sum scores extends beyond the issue of a conceptual interpretation. It is shown below that in the case of non-invariance the path coefficients of the genetic factor resulting from a decomposition of sum scores are biased differentially across the groups (e.g., gender groups in the case of sex limitation models or environmental exposure groups in the case of G × E interaction models). Hence, absence of measurement invariance is confounded with sex limitation or gene by environment interaction. Differences in factor loadings or discrimination parameters can be detected only in a multivariate analysis of the individual items that is carried out simultaneously in all groups.

Biased Path Coefficients of the Genetic Factor

Univariate Case

For reasons of simplicity, the common factor model with a single factor η and two continuous observed variables Y_1 and Y_2 is chosen as a measurement model to demonstrate how absence of measurement invariance can be confounded with scalar sex limitation or gene by environment interaction if sum scores are analyzed instead of the individual items. It should be noted that the argument is not restricted to this type of measurement model. As shown above, the definition of measurement invariance only refers to the distribution of observed variables given latent variables and hence is not restricted to a specific type of measurement model. The argument presented here can be extended easily to models with more factors, or to models for ordered categorical data such as the proportional odds model (Agresti, 1990), or a two-parameter item response model. In the context of the present paper, the differences between these models are not essential.

The rationale is cast in the context of sex limitation, but similarly applies to other grouping variables. For G × E interaction the subscripts f and m indicating females and males could be changed to $e1$ and $e2$ indicating two environmental exposure groups.

The measurement model relating items to an underlying factor in females is

$$\begin{aligned} y_{1f} &= \lambda_{1f}\eta_f + \varepsilon_{1f} \\ y_{2f} &= \lambda_{2f}\eta_f + \varepsilon_{2f}, \end{aligned} \tag{2}$$

where λ and ε indicate factor loadings and regression residuals, respectively. Corresponding equations for the observed scores in males can be obtained by changing the subscript f to m.

Sum sores for females are obtained by adding the scores on the two items,

$$Sf = \lambda_{1f}\eta_f + \lambda_{2f}\eta_f + \varepsilon_{1f} + \varepsilon_{2f}. \tag{3}$$

Figure 1 shows the ACE model for opposite sex twins. Part A shows the common pathway model, where the ACE decomposition is applied to a factor underlying observed scores Y, and part B shows the ACE model applied to a sum score S.

Absence of measurement invariance with respect to factor loadings means that the matrix containing the factor loadings in females does not equal the matrix of factor loadings in males, $\Lambda_f \neq \Lambda_m$. The goal here is to demonstrate that the estimate of the additive genetic variance, a^2, is biased differentially across gender if the ACE model is applied to the sum score and the factor loadings $\Lambda_f \neq \Lambda_m$. For ease of presentation, and without loss of generality, it is assumed that the parameter c equals 0 in both genders such that the common environment variance component is 0. Hence, we consider an AE-model rather than an ACE-model.

Ignoring the measurement model relating individual items and the underlying factor, the covariance between the sum scores of a twin pair, t_1 and t_2, according to the AE-model is given by

$$\mathrm{Cov}(S_{t1}, S_{t2}) = a_{t1}a_{t2}\alpha, \tag{4}$$

where, given the usual assumptions, α is fixed to unity and 0.5 for monozygotic (MZ) and same sex dizygotic (DZ) twins, respectively, and may be estimated in opposite sex twin pairs.

If the measurement model for the individual items is taken into account, the covariance between sum scores of twin 1 and twin 2 can be derived using equation 3,

$$\begin{aligned} \mathrm{Cov}(\mathrm{S}_{t1}, \mathrm{S}_{t2}) = (&\lambda_{1t1}\lambda_{1t2} + \lambda_{1t1}\lambda_{2t2} + \\ &\lambda_{2t1}\lambda_{1t2} + \lambda_{2t1}\lambda_{2t2}) \times \mathrm{Cov}(\eta_{t1}, \eta_{t2}), \end{aligned} \tag{5}$$

https://doi.org/10.1375/twin.7.3.292 Published online by Cambridge University Press

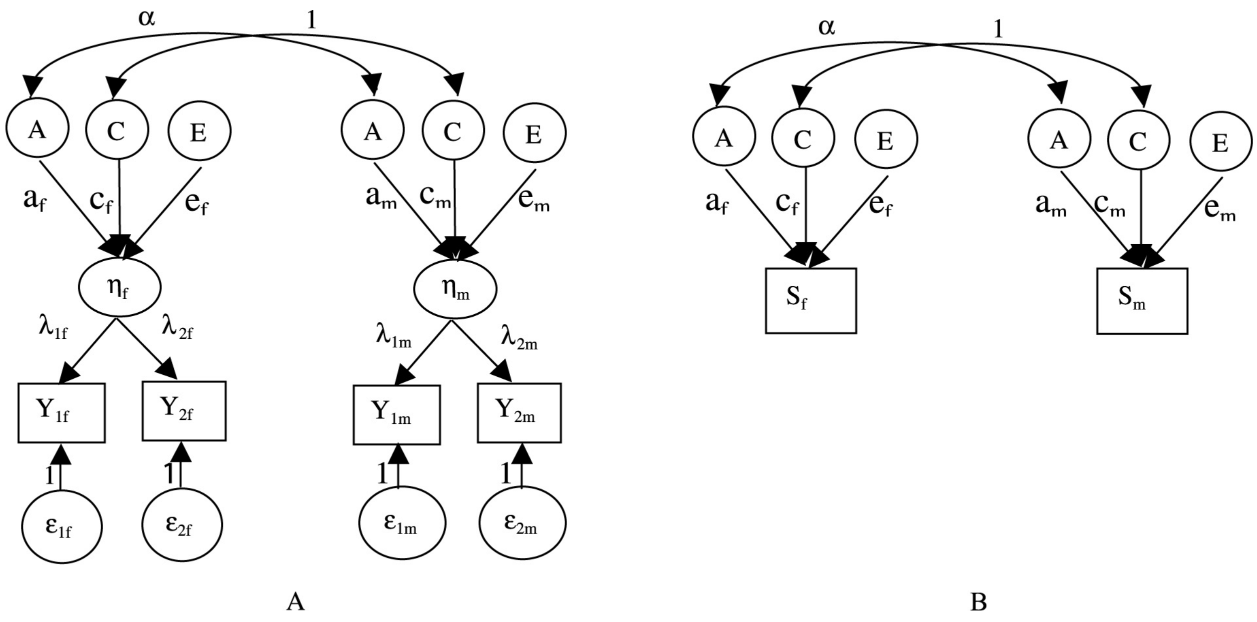

Figure 1

Opposite sex ACE model for a latent factor (Part A) and for a sum score (Part B). The variances of the latent variables A, C, and E are fixed to unity, whereas those of ε_{1m}, ε_{2m}, ε_{1f}, and ε_{2f} are estimated as free parameters.

where the term between brackets is basically a weight for the covariance between the factor scores of a twin pair. The weight depends on the measurement model or, more specifically, on the factor loadings. Let k_f, k_m, and k_o denote the weights of same sex female twins, same sex male twins, and opposite sex twins, respectively, then

$$\begin{aligned} k_f &= \lambda_1{}^2{}_f + 2\lambda_{1f}\lambda_{2f} + \lambda_2{}^2{}_f = (\lambda_{1f} + \lambda_{2f})^2 \\ k_m &= \lambda_1{}^2{}_m + 2\lambda_{1m}\lambda_{2m} + \lambda_2{}^2{}_m = (\lambda_{1m} + \lambda_{2m})^2 \\ k_o &= \lambda_{1f}\lambda_{1m} + \lambda_{1f}\lambda_{2m} + \lambda_{2f}\lambda_{1m} + \lambda_{2f}\lambda_{2m}, \end{aligned} \qquad (6)$$

which shows that the weight k_o is the geometric mean of the weights k_f and k_m.

The covariance between factor scores is decomposed according to the AE-model as

$$\mathrm{Cov}(\eta_{t1}, \eta_{t2}) = a_{t1}a_{t2}\alpha. \qquad (7)$$

Combining equations 5 through 7, the covariances of MZ and DZ same sex and DZ opposite sex twins are

$$\begin{aligned} \mathrm{Cov}(S_{1fMZ}, S_{2fMZ}) &= k_f \times a^2{}_f \times \alpha_{fMZ} \\ \mathrm{Cov}(S_{1fDZ}, S_{2fDZ}) &= k_f \times a^2{}_f \times \alpha_{fDZ} \\ \mathrm{Cov}(S_{1mMZ}, S_{2mMZ}) &= k_m \times a^2{}_m \times \alpha_{mMZ} \\ \mathrm{Cov}(S_{1mDZ}, S_{2mDZ}) &= k_m \times a^2{}_m \times \alpha_{mDZ} \\ \mathrm{Cov}(S_{1fDZ}, S_{2mDZ}) &= k_o \times a_o \times \alpha_{oDZ}, \end{aligned}$$

Where $a_o = a_f \times a_m$. In case of no sex limitation, the parameter a is equal for males and females, and $\alpha_{fMZ} = \alpha_{mMZ} = 1$ and $\alpha_{fDZ} = \alpha_{mDZ} = \alpha_{oDZ} = 0.5$ (i.e, no sex limitation). If the two items Y_1 and Y_2 are measurement invariant with respect to gender, then $\Lambda_f = \Lambda_m$ and $k_f = k_m = k_o$. If factor loadings are not equal for males and females, for instance if item 1 is non-invariant with respect to gender, $\lambda_{1f} > \lambda_{1m}$, then $k_f > k_o > k_m$.

In a simple decomposition of the sum score the measurement model relating items to the factor is ignored, which means that the ks are implicitly fixed to unity (see Equation 4). Differences in these weights due to differences in the factor loadings are absorbed by the path coefficients a, resulting in $a_f^2 > a_o > a_m^2$. Since k_o is the geometric mean of k_f and k_m, fixing the ks to unity results in a_o being the geometric mean of $a^2{}_f$ and $a^2{}_m$. This pattern of differences in additive genetic variance components is identical to scalar sex limitation (Neale & Cardon, 1992). Since the sum score is regarded as an estimate of the factor score η underlying the individual items, the results would usually be interpreted as gender-specific effect sizes of the genetic factor that is derived from the decomposition of the factor η. In the case of G × E interaction, the results would be interpreted in terms of different heritabilities across environmental exposure groups. However, such an interpretation would be incorrect if the group differences are due to absence of measurement invariance.

The correct interpretation would be not that the heritabilities differ across groups, but that factor η is not the same factor in the different groups. The factor differs because the importance of individual items differs across groups. The correct interpretation can only be obtained if the differences in factor loadings are detected. Analyzing sum scores precludes this possibility.

https://doi.org/10.1375/twin.7.3.292 Published online by Cambridge University Press

Multivariate Case

An alternative to fitting an ACE-type model to the univariate sum score is a multivariate analysis where 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 is considered. This covariance matrix is a partitioned matrix with across twin covariances in the lower left (upper right) block. Denote the block with the covariances of twin 1 and twin 2 as $\Sigma_{t1,t2}$, then for opposite sex twins we have

$$\Sigma_{t1,t2} = \Lambda_f(a_f a_m \alpha_a)\Lambda^t_m + \Theta_\varepsilon, \qquad (8)$$

where Θ_ε denotes the residual covariance matrix. In the multivariate model, all factor loadings are estimated. If some elements of Λ_f differ from the corresponding elements in Λ_m and if, as is often done in practice, factor loadings are initially fixed to be equal across genders, there is no simple way in which the parameter a can compensate for the resulting misfit. Given sufficient sample size, fitting a model with factor loadings restricted to be equal across gender will result in rejection of the model due to lack of fit if the population factor loadings differ. In a multivariate analysis equality of factor loadings across gender is clearly a hypothesis that can be tested by comparing models with and without the equality restrictions using a likelihood ratio test.

Illustration with Simulated Data

To illustrate the bias in path coefficients of the genetic factor that occurs in the analysis of sum scores derived from non-invariant items, 500 data sets according to the common pathway model were created for five groups (i.e., MZ female and male, DZ female and male, and opposite sex) with the following characteristics. The a and e parameters are $\sqrt{.5}$ for both males and females, whereas c equals 0. Furthermore, $\alpha_{fMZ} = \alpha_{mMZ} = 1$ and $\alpha_{fDZ} = \alpha_{mDZ} = \alpha_{oDZ} = 0.5$, which is the parameterization for absence of sex limitation. The variance of the factors A and E is unity. The measurement model is a single common factor model with four observed continuous items, factor loadings for females and males are .3, .4, .3, .4 and .6, .7, .8, .9, respectively, and residual variances for the two genders are .91, .84, .91, .84 and .64, .51, .36, .19, such that the variance of all observed items is approximately unity. The 500 data sets differ with respect to the factor scores A and E and with respect to the residual scores. Sample sizes are 200 for MZ and DZ male and female twin pairs and for opposite sex twin pairs. Sum scores are derived by summing the observed item scores using unity weights for all items in all groups.

Table 1 shows the sum score correlations between twins for the five zygosity groups averaged over the 500 data sets. The empirical standard errors of the averaged sum score correlations are small, indicating that these results do not vary greatly over the 500 data sets. The pattern of the correlations clearly indicates scalar sex limitation. The table also shows the estimated correlation between factors of twin pairs after fitting a single factor model to the multivariate data, without restricting the factor loadings to be equal across groups. The model with factor loadings equated across groups is always rejected due to lack of fit. Not surprisingly, the correlation between factors when the factor loadings are not equated corresponds to the true values.

Table 1

Twin Sum Score Correlations Versus Factor Correlations Averaged Over 500 Data Sets

	MZ_f	DZ_f	MZ_m	DZ_m	DZ_o
Sum score	.18 (.07)	.09 (.07)	.42 (.06)	.21 (.07)	.14 (.05)
Unequal loadings	.50 (.20)	.25 (.19)	.50 (.06)	.25 (.08)	.25 (.09)

Note: Standard errors of the mean correlations capturing the variation across the 500 replications are given between brackets. The standard errors of the factor correlations of female twins are higher due to the lower reliability of the item scores.

Conclusion

This article shows the confounding between absence of measurement invariance and scalar sex limitation that occurs when sum scores are analyzed instead of individual items. By replacing gender by grouping variables such as environmental exposure or age, the argument is easily extended to $G \times E$ or $G \times age$ interaction. Analyzing sum scores precludes the detection of group-specific factor loadings which are a form of non-invariance. If undetected, absence of measurement invariance will be confounded with group differences in heritabilities. In the present article, sum scores are assumed to be derived by assigning the same weight to each of the individual items across groups. We note that any weighted sum score with weights differing across items (such as factor scores) will suffer from the same problem if the same scoring procedure is used for all groups.

Measurement invariance is a broad concept, which can be applied even if the assumption of an underlying latent trait or factor is not based on theoretical considerations. Generally, measurement invariance can be assessed in any scores with common variation by modeling the scores using an item response model such as the common factor model. An example is counts of hair in different sub-areas of the male and female body, which can be modeled to be due to one or more factors. Simply summing over all counts and decomposing the resulting sum score may lead to a confounding of sex limitation and measurement invariance.

Caution is therefore required in the interpretation of any study that tests for the effects of sex-limitation

https://doi.org/10.1375/twin.7.3.292 Published online by Cambridge University Press

in pre-computed scale scores. A finding of scalar sex-limitation could arise from genuine differences in the scale of genetic effects, or from failure of measurement invariance, or both. Conversely, the absence of evidence for scalar sex-limitation in a sum score could reflect the absence of sex-limitation as well as measurement invariance for that scale. However, it is also possible that both sex-limitation and absence of measurement invariance exist and that their effects counterbalance each other — for example, when both higher heritability and lower factor loadings exist in one group.

These arguments extend beyond simple sex-limitation and genotype by environment interaction where the environmental variable is simply binary. Models for continuous $G \times E$ interaction (Neale, 1998; Neale & Cardon, 1992; Purcell, 2002) that involve an ordinal or continuous environmental measure would be similarly affected when the observed sum scores to be analyzed fail to meet the assumption of measurement invariance. This problem will also be manifest in the linkage analysis of quantitative trait loci (QTLs) whose action varies as a function of environmental indices (Purcell & Sham, 2002).

The multivariate analysis of item-level data may pose significant computational burdens, especially in the case of linkage analysis. Therefore, it is important to consider alternative approaches. One obvious approach is to conduct the test for measurement invariance prior to the computation of sum scores or factor scores. If the data exhibit the property of measurement invariance, then it is safe to proceed with the variance components or linkage analysis of sum scores or factor scores. However, if the data exhibit non-invariance, then the investigator should consider conducting the multivariate analysis and hope that sufficiently rapid hardware and software are available for the task. Another possibility to consider would be to revise the scale to establish measurement invariance. Obviously, this latter approach is not always feasible; many measures (such as the DSM) have established scales on which much prior research has been done. At the very least, reporting of absence of measurement invariance and consideration of its likely consequences as a limitation, is warranted.

The type of absence of measurement invariance discussed here (e.g., differences in factor loadings) concerns an interaction between the non-invariant item and the grouping variable (e.g., gender or environmental exposure), where the regression of the item on the factor is steeper for one group than for the other. The interaction is due to the sensitivity of the item to other variables in addition to the factor. It is an interaction on the item level between the underlying factor and the additional variable and not, as one would conclude in an analysis of sum scores, between the genetic factor and gender or environmental exposure. Analysis of sum scores is therefore to be avoided if individual item scores are available, and multivariate item-level analysis should be used instead. If a non-invariant item is detected in a multivariate analysis, further research may trace the nature of the additional variable(s) that influence that item.

Acknowledgments

Gitta H. Lubke, Department of Psychiatry; Michael C. Neale, Departments of Psychiatry and Genetics; Conor V. Dolan, Department of Psychology. The research of the first two authors was supported through grant MH65322 by NIMH. The research of Conor V. Dolan was supported by Netherlands Organization for Scientific Research (NWO).

Endnote

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

References

Agresti, A. (1990). *Categorical data analysis*. New York: Wiley.

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

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

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.

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: Digits.

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

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.

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.

Mellenbergh, G. J. (1989). Item Bias and Item Response Theory. *International Journal of Educational Research, 13,* 127–143.

https://doi.org/10.1375/twin.7.3.292 Published online by Cambridge University Press

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

Neale, M. C. (1998). Modeling interaction and non-linear effects with Mx: A general approach. In G. Marcoulides & R. Schumacker (Eds.), *Interaction and non-linear effects in structural equation modeling* (pp. 43–62). New York, NY: Lawrence Erlbaum Associates.

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

Purcell, S. (2002). Variance components models for gene-environment interaction in twin analysis. *Twin Research, 5*, 554–571.

Purcell, S., & Sham, P. (2002). Variance components models for gene-environment interaction in quantitative trait locus linkage analysis. *Twin Research, 5*, 572–576.

https://doi.org/10.1375/twin.7.3.292 Published online by Cambridge University Press