Behav Genet (2014) 44:212–231
DOI 10.1007/s10519-014-9647-9

ORIGINAL RESEARCH

# Testing Systematic Genotype by Environment Interactions Using Item Level Data

**Dylan Molenaar · Conor V. Dolan**

Received: 27 June 2013 / Accepted: 5 February 2014 / Published online: 22 February 2014


**Abstract** Investigating genotype by environment interactions (GxE) is generally considered challenging due to the scale dependency of the interaction effect. The present paper illustrates the problems associated with testing for GxEs on summed item scores within the well-known ACE model. That is, it is shown how genuine GxEs may be masked and how spurious interactions can arise from scaling issues in the data. A solution is proposed which explicitly distinguishes between a measurement model for the ordinal item responses and a biometric model in which the GxE effects are investigated. The new approach is studied in a simulation study using both a scenario in which the measurement instrument suffers from mild scaling problems and a scenario in which the measurement instrument suffers from severe scaling problems. Results indicate that the severity of the scale problems affects the power to detect GxE, but it rarely results in false positives. We illustrate the new approach on a real dataset concerning affect.

**Keywords** Genotype by environment interaction · Poor scaling · Item response theory · ACE-model

D. Molenaar (✉)
Psychological Methods, Department of Psychology, University of Amsterdam, Weesperplein 4, 1018 XA Amsterdam, The Netherlands
e-mail: D.Molenaar@uva.nl

C. V. Dolan
Department of Biological Psychology, VU University, Amsterdam, The Netherlands

## Introduction

> The existence of (…) genotype-environmental interactions of a systematic variety may alert the psychometrician to areas in which further test development may take place. Marked directional non-additivity, for example, may indicate a threshold in the scale beyond which measurement is difficult, or simply has not been attempted (Eaves et al. 1977).

Systematic genotype by environment interaction (GxE) concerns the situation in which the strength of genetic influences on a given phenotypic construct differs across environments. A well-known instance of GxE is the finding of Turkheimer et al. (2003), who showed that the heritability of cognitive ability increases for increasing levels of socioeconomic status. Other instances of GxE include the genotype by parental closeness interaction on alcoholism (Miles et al. 2005), the genotype by income interaction on physical health (Johnson and Krueger 2005), and the genotype by negative life events interaction on depressive symptoms (Lau and Eley 2008).

Given the popularity of GxE as a research topic, it is interesting that investigating GxE remains associated with major challenges. The main problem is that GxE effects can be conflated with artificial effects due to (1) genotype-environment correlation (see e.g., Turkheimer et al. 2009; Purcell 2002); and (2) due to poor measurement scaling (Eaves et al. 1977, 2002; Eaves 2006; Kamin 1974; Mather and Jinks 1971; Purcell 2002; van der Sluis et al. 2006). With respect to the former, solutions to test for GxE while correcting for possible genotype by environment correlation have been proposed by Rathouz et al. (2008), Purcell (2002), and van der Sluis et al. (2012), although this issue still remains a topic of investigation. With respect to the

problem of measurement scale, few solutions have been proposed. The present paper focuses on this problem of measurement in the detection of GxE.

As pointed out with respect to interactions in general (e.g., Loftus 1978; Wagenmakers et al. 2012; Zand Scholten 2011, chapter 5), and with respect to GxEs in particular (e.g., Eaves et al. 1977; Eaves 2006; Jinks and Fulker 1970; van der Sluis et al. 2006, 2012; Molenaar et al. 2012), statistical interaction effects are generally scale dependent, which means that (1) interaction effects can be removed or indeed created by monotonic non-linear transformation of the data; and (2) interactions may spuriously arise due to non-normality in the observed data as a result of arbitrary properties of the measurement scale. Within the field of behavior genetics, the problem of measurement scale dependency is clearly illustrated in a recent study by Molenaar et al. (2013), in which IQ measures from 14 different twin studies were tested for GxE. They found that interaction effects fluctuated greatly across studies while controlling for age and gender. As the test batteries used to obtain the IQ measures differed considerably across the different studies, they concluded that the GxE effects observed were likely attributable to measurement problems.

The most important measurement problems in the detection of GxE are floor and ceiling effects (van der Sluis et al. 2006), and poor scaling of the measurement (Eaves 2006; Eaves et al. 1977). These problems all boil down to the same principle, i.e., the amount of information about the phenotypic construct varies across the scale used to measure the construct. In case of floor and ceiling effects this is due to the fact that a disproportionate number of subjects receive the lowest or highest possible score. This censoring causes individual differences to be smaller at the lower end of the scale (floor effect) or at the upper end of the scale (ceiling effect). In the case of poor scaling of measurement, the amount of information concerning the phenotypic construct varies across the scale because the resolution of the scale varies across its range. For instance, depression questionnaires commonly discriminate relatively well at the upper range, but less well over the lower end of the scale. Note that poor scaling of the measurement does not necessarily imply a floor or ceiling effect, but it does imply non-normality. Poor scaling commonly arises by summing individual items and treating the resulting variable as a continuous variable in the analysis. As pointed out and illustrated by Tucker-Drob (2009), these composite scores are highly likely to be poorly scaled if items are disproportionally hard or disproportionally easy or when the items do not follow a one-parameter model (Molenaar and Borsboom 2013). That is, sum scores will be non-normally distributed despite the underlying normally distributed construct.

A possible approach to account for the measurement issues above is to explicitly take the measurement properties of the scale into account. This can be done by using an appropriate measurement model for the items that are used to measure the phenotypic construct. Popular measurement models include the 1 and 2 parameter models for dichotomous items (Rasch 1960; Birnbaum 1968) and the graded response model for Likert scale items (Samejima 1969). With respect to heritability analyses, van den Berg et al. (2007) demonstrated that neglecting measurement properties of dichotomous items generally results in underestimated heritability coefficients. They showed that the use of a measurement model produced unbiased estimates.

With respect to GxE research, Tucker-Drob et al. (2009) and Molenaar et al. (2012) also advocated the use of measurement models. Specifically, a model was proposed in which the phenotypic variables were first linked, as indicators, to the underlying phenotypic construct using a factor model as measurement model. Next, in the biometric part of the model, the phenotypic variance was decomposed into parts due to additive genetic (A), common environment (C), and unique environment (E) influences. In Tucker-Drob et al. (2009), measurement problems were accounted for by introducing non-linear effects in the measurement model. Next, GxE was introduced in the biometric model by using the moderation approach of Purcell (2002).

Molenaar et al. (2012) introduced GxE using the classical conceptualization of systematic GxE of Jinks and Fulker (1970; Eaves et al. 1977), in which GxE is operationalized as heteroscedastic latent environmental variance, which varies systematically with the environment (see also van der Sluis et al. 2006; see Molenaar and Boomsma 1987, for an alternative approach in case of GxE with unmeasured environment). Using this method, within the ACE model, both AxE and AxC interactions can be detected as heteroscedastic E or C variance, respectively. To account for possible measurement problems at the level of the observed variables, heteroscedastic residual variances were incorporated in the measurement model. In this way, possible floor/ceiling or poor scaling effects (i.e., effects uncorrelated across variables) are absorbed in the residuals, while the effects of GxE—if any—are detected in the latent biometrical part of the model (as a genuine GxE effect should be common to all indicators of the common latent phenotype). Note that this approach is thus suitable to test for GxE in sum score variables that are poorly scaled (as discussed above), as the scale problems will be captured as heteroscedastic residuals. However, if the majority of the observed variables are subject to floor/ceiling or poor scaling effects, these effects may still arise as artificial

AxE or AxC (see Molenaar et al. 2012). In addition, as the measurement model is a linear factor model, this approach is limited to continuous data. In the cases of polytomous items (e.g., 3 or 5 point self-report scales, common in the assessment of personality) and dichotomous items (e.g., correct/false scores, common in ability tests items), this approach is inappropriate (see Dolan 1994; van den Berg et al. 2007).

Therefore, a generalization of the methodology of Molenaar et al. (2012) is needed for categorical item scores. This is challenging as due to the categorical nature of the data, the marginal maximum likelihood framework is numerically demanding (although we note that it is possible, see below). As demonstrated by Eaves and Erkanli (2003), the Bayesian framework is computationally more tractable in case of complex likelihood functions. Inspired by Eaves and Erkanli (2003), van den Berg et al. (2007) employed the Bayesian approach to the genetic analyses of dichotomous items using a 1 parameter measurement model, applying a Bayesian hierarchical parameterization of the ACE decomposition in the latent part of the model (Eaves and Erkanli 2003; van den Berg et al. 2006). This model was recently extended by Schwabe and van den Berg (2014) to include an AxE effect in the biometric model.[1]

The aims of this paper are twofold: (1) we demonstrate that minor scale problems readily give rise to spurious GxE effects in sum scores; and (2) we present a general approach to test for GxE in categorical twin data that does not suffer from spurious GxE, even in the presence of severe scale problems. We focus on dichotomous and Likert scale items as these are common in personality and cognitive ability research. The main advantages of the approach outlined in this paper over that of Schwabe and van den Berg (2014) are that: (1) we do not restrict ourselves to dichotomous data, but also consider Likert scale data; (2) we model violations of local independence (residual correlations between the scores of the twins in the same twin pair) which are common in twin data; (3) we also consider AxC interactions as Molenaar et al. (2012) show that this could benefit the power to detect AxE; and (4) we present a full estimation approach, that is, we estimate all model parameters from the data. Schwabe and van den Berg (2014) assume that the parameters from the measurement model are known constants, which could be the case in—for instance—large scale assessments, but which could be problematic in smaller datasets. An advantage of the approach by Schwabe and van den Berg (2014) is that it is mathematically more elegant and less computationally demanding.

The outline of this paper is as follows: First we present suitable measurement models for Likert data and dichotomous data, and we discuss how we account for violations of local independence. Next, in the biometric model, we introduce an alternative to the Bayesian parameterization of the ACE decomposition used by Eaves and Erkanli (2003) and van den Berg et al. (2006) in which we introduce GxE in terms of heteroscedastic environment influences (as in Jinks and Fulker 1970; van der Sluis et al. 2006; Molenaar et al. 2012). Then we present a simulation study to (1) demonstrate the problems associated with neglecting the measurement properties of the individual items in GxE analyses; and (2) to show that our approach provides a feasible solution to this problem. Finally, we apply the model to a real dataset on affect, and end with a discussion.

## Model derivation

In the traditional univariate twin ACE model (Jinks and Fulker 1970; Eaves et al. 1977; Eaves 1977; Neale and Cardon 1992), the scores on a phenotypic variable as observed in the twin pairs, $Y_1$ and $Y_2$, are assumed to be an additive function of the A, C, and E components, where by definition, the environmental twin correlations are $\text{cor}(C_1,C_2) = 1$ and $\text{cor}(E_1,E_2) = 0$. Assuming assortative mating, the additive genetic twin correlations are $\text{cor}(A_1, A_2) = 1$ in MZ twins, and $\text{cor}(A_1,A_2) = .5$ in DZ twins. Under the assumption that the A, C, and E factors are mutually uncorrelated, the variance of $Y_j$ is decomposed as

$$\text{var}(Y_j) = \sigma_A^2 + \sigma_C^2 + \sigma_E^2, \quad j = 1,\ 2. \tag{1}$$

In operationalizing GxE, one could make the effect of A to depend on E and C, or one could make the effect of C and E to depend on A. As Jinks and Fulker (1970) defined systematic GxE as heteroscedastic environmental variance across genotypes, van der Sluis et al. (2006) and Molenaar et al. (2012) used the latter operationalization (the effect of E and C depends on A) and formalized AxC and AxE interactions as follows

$$\begin{aligned}\text{var}(Y_j|A_j) &= \sigma_C^2|A_j + \sigma_E^2|A_j \\ &= \exp(\gamma_0 + \gamma_1 A_j) + \exp(\beta_0 + \beta_1 A_j)\end{aligned} \tag{2}$$

That is, the variance of C and E conditional on A is modelled as a function of A. We denote this *systematic* GxE as the conditional environmental variance varies systematically with A. The choice of the exponential function is purely pragmatic: it ensures that the variance is strictly positive (see also Bauer and Hussong 2009; Hessen and Dolan 2009). In Eq. 2, parameters $\gamma_0$ and $\beta_0$ are

[1] The work by Schwabe and van den Berg (2014) was conducted parallel to—but independent from—the present undertaking. Only at a late stage did we learn about each others work on this topic.

baseline parameters for the variance of C and E, respectively, and parameters $\gamma_1$ and $\beta_1$ account for the AxC and AxE interaction in terms of heteroscedasticity. Note that the operationalization of AxC in Eq. 2 implies a scalar effect of $A_j$ on the variance of C which places the effect of C on different scales for two members of a DZ twin pair. As shown by Molenaar et al. (2012), the parameters ($\beta_0$, $\beta_1$, $\gamma_0$, $\gamma_1$, and $\sigma^2_A$) can be estimated using marginal maximum likelihood (Bock and Aitkin 1981) in the Mx software package (Neale et al. 2006). Relevant Mx input files are available at http://www.dylanmolenaar.nl.

The presence of AxE or AxC complicates the calculation of heritability. To standardize $\sigma^2_A$, one needs:

$$h^2 = \frac{\sigma_A^2}{\sigma_A^2 + \exp\left(\beta_0 + \frac{1}{2}\beta_1^2\right) + \exp\left(\gamma_0 + \frac{1}{2}\gamma_1^2\right)} \quad (3)$$

where the two terms in the denominator denote the marginal variance of E and C, respectively (see Hessen and Dolan 2009). The standardized marginal contributions of C and E to the phenotype variance can be calculated in the same way.

The statistical properties of the univariate model in Eq. 2 were studied by van der Sluis et al. (2006) and Molenaar et al. (2012). Generally, parameter recovery is satisfactory, AxE and AxC interactions are well separable, and power to detect AxE is good in terms of required sample size. On the other hand, large sample sizes are needed to detect AxC. It was also shown that, if A is the predominant source of variation, presence of an unmodelled CxE interaction will not result in spurious AxE.

Below we develop a methodology based on the idea above, but for the case in which we do not have a single phenotype variable, but multiple ordinal item responses that measure an underlying phenotypic construct, θ. First, we introduce the measurement model in which the responses to the items are linked to θ. Next, in the biometric model, we present a Bayesian formulation of the multi-item ACE model that is suitable to apply the model in Eq. 2 to θ.

## Measurement model

In the measurement model, the responses of twin $j$ in twin pair $p$ on item $i$, $X_{pij}$, are linked to the underlying phenotypic construct or latent variable $\theta_{pj}$. As $j = 1, 2$ we have two latent variables, $\theta_{p1}$ and $\theta_{p2}$ that are expected to be dependent due to common genetic and environmental influences underlying the phenotypic construct. This dependency can be accommodated by the application of standard twin modeling (e.g., specification of an ACE model), as mentioned above. Standard measurement models, however, assume local independence. In the present case, this means that that the item responses of the twins are independent conditional on the latent variables $\theta_{p1}$ and $\theta_{p2}$. This assumption will generally not hold in twin data, as responses of twins of the same twin pair, conditional on $\theta_{p1}$ and $\theta_{p2}$, are still likely to be correlated due to shared item-specific genetic and environmental influences. These residual correlations should be taken into account to avoid bias in the other parameter estimates. In a factor analytic framework, this is straightforward, as the residual covariances are explicitly model parameters, which can be freely estimated. In an item response theory framework, however, this is less straightforward as the item responses are treated as realizations of a Bernoulli distribution (dichotomous responses) or multinomial distribution (ordinal responses), which precludes incorporation of a residual covariance matrix. It could be argued that we should resort to the item factor model (Wirth and Edwards 2007), in which residual correlations can be incorporated straightforwardly. However, this cannot easily be combined with the decomposition that we wish to carry out for the latent phenotypes. We therefore propose to model possible violations of conditional independence by introducing additional latent variables that account for the covariance unique to a given item within a twin pair. Specifically, in case of ordinal responses (e.g., personality tests where items are Likert scales) the model is given by

$$X_{pij} \sim \text{cat}\left(\pi_{pij0}, \ldots, \pi_{pij(c-1)}\right), \quad (4)$$

with, for $c = 0, \ldots, C_i - 1$,

$$\begin{aligned}\pi_{pijc} &= P\left(X_{pij} = c|\theta_{pj}\right) \\ &= \Phi\left(\frac{a_i\theta_{pj} + r_i\delta_{pi} - b_{ic}}{\sqrt{1 - r_i^2}}\right) \\ &\quad - \Phi\left(\frac{a_i\theta_{pj} + r_i\delta_{pi} - b_{i(c+1)}}{\sqrt{1 - r_i^2}}\right)\end{aligned} \quad (5)$$

In Eq. 5, φ(.) denotes the cumulative normal distribution function, $a_i$ is the discrimination parameter of item $i$, $b_{ic}$ is the category difficulty parameter of category $c$ in item $i$ with $b_{i0} = -\infty$ and $b_{iC} = \infty$, $\theta_{pj}$ is the position of twin $j$ of twin pair $p$ on the latent variable, $C_i$ denotes the number of response categories of item $i$, and $\delta_{pi}$ is the additional latent variable to model the conditional associations among the scores within twin pair $p$ on item $i$ with item specific slope parameter $r_i$. Figure 1 shows a schematic representation of the measurement model. As can be seen from the figure, $r_i$ can be interpreted as the factor loading of item $i$ on the additional latent $\delta_{pi}$ variables. This factor loading is equal across twin 1 and twin 2 item scores. By assuming that $\delta_{pi}$ has a standard normal distribution (see also below), the residual polychoric correlation between the item scores of twin 1 and twin 2 is given by $r_i^2$.

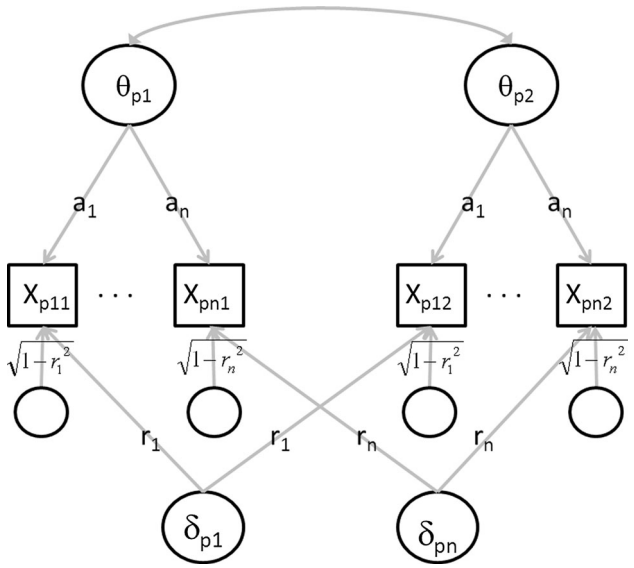


**Fig. 1** Schematic representation of the measurement model in Eq. 5. Note the *straight arrows* represent probit regressions. The category difficulty parameters, $b_{ic}$ are not depicted. The *blank circles* represent the polychoric variance of the item scores conditional on the latent variables

The term $\sqrt{1-r_i^2}$ (Eq. 5) is not standard in item response models, but it is necessary here to retain the original scaling of the item parameters and to ensure that $r_i^2$ can be interpreted as a correlation.[2]

In the case of dichotomous data (e.g., in ability tests where item responses are commonly scored correct/false), $C_i$ can be set to 2 and the model above simplifies to

$$X_{pij} \sim \text{bern}(\pi_{pij}) \tag{6}$$

$$\pi_{pij} = P(X_{pij} = 1|\theta_{pj}) = \Phi\left(\frac{a_i\theta_{pj} + r_i\delta_{pi} - b_i}{\sqrt{1-r_i^2}}\right) \tag{7}$$

All item parameter in Eqs. 5 and 7 are equal across MZ and DZ twin members except $r_i$, which varies over zygosity. That is, we estimate separate $r_{DZ,i}$ and $r_{MZ,i}$ as the shared genetic and environment item specific influences are expected to be different across MZ and DZ twins.

## Biometric model

In the biometric model, the correlations between $\theta_{p1}$ and $\theta_{p2}$ are modeled as a function of the A, C, and E factors in the MZ and DZ twins. In this decomposition, we introduce AxE and AxC interactions according to Eq. 2. Eaves and Erkanli (2003) implemented a Bayesian parameterization of the univariate ACE model in Eq. 1 (see also van den Berg et al. 2006). This parameterization consists of a hierarchical structure, in which at each level of the hierarchy, all variables are combined from the previous level. An alternative to the hierarchical parameterization might be to specify a multivariate normal distribution for $\theta_{p1}|A$ and $\theta_{p2}|A$, and make the diagonal elements of the conditional covariance matrix a function of A. However, in our experience this approach worked poorly as for the traditional ACE model, sampling from the posterior distribution was extremely slow, and not all chains displayed random intermixing.

Below, we propose an alternative parameterization of the ACE model that only uses univariate distributions as in Eaves and Erkanli (2003) and van den Berg et al. (2006). Within this parameterization, we include the AxC and AxE interactions, following Eq. 2. We present the model details for the MZ and DZ twins separately.

### MZ twins

As in MZ twins, $\text{cor}(A_1,A_2) = 1$, we use $A_1 = A_2 = A$ with

$$A \sim N(0; \sigma_A^2) \tag{8}$$

where $\sigma_A^2$ can be used to determine heritability given appropriate standardization (see Eq. 3). We now impose an ACE decomposition on the latent variable $\theta_{pj}$. First, conditioning on A yields the following conditional distribution for $\theta_{pj}$ in the MZ twin sample:

$$\theta_{pj}|A \sim N(\mu_j|A; \sigma_j^2|A) \tag{9}$$

where

$$\mu_1|A = A \tag{10}$$

and

$$\mu_2|A = A + \rho|A \times (\theta_{p1}|A - A) \tag{11}$$

The expression for $\mu_2|A$ in Eq. 11 follows from standard results of the bivariate normal distribution, i.e., $E(X|Y) = \mu_X + \sigma_X\rho\ (Y - \mu_Y)/\sigma_Y$ where in this case, $X = \theta_{p2}|A$ and $Y = \theta_{p1}|A$. In addition, $\rho|A$ is the correlation between $\theta_{p1}|A$ and $\theta_{p2}|A$ which is given by

$$\rho|A = \frac{\sigma_C^2|A}{\sigma_C^2|A + \sigma_E^2|A} \tag{12}$$

Now $\sigma_j^2|A$ in the conditional distribution of $\theta_{pj}|A$ above is given by

$$\sigma_1^2|A = \sigma_C^2|A + \sigma_E^2|A \tag{13}$$

for $j = 1$ and

[2] The cumulative standard normal distribution function, $\Phi(.)$, in Eq. 5 assumes unit polychoric variances of the item scores conditional on $\theta_{p1}$ and $\theta_{p2}$. Due to the presence of $\delta_{pi}$, the conditional polychoric variance will depart from 1 if $r_i \neq 0$. Therefore, to prevent parameter bias, the scaling term is added to ensure that the conditional polychoric variance will be equal to 1.

$$\sigma_2^2|A = \sigma_1^2|A \times (1 - \rho^2|A) \quad (14)$$

for j = 2. At this point, the model above is a standard ACE model. We introduce GxE by making $\sigma_C^2|A$ and $\sigma_E^2|A$ a function of A, as in Eq. 2, i.e.,

$$\sigma_C^2|A = \exp\left(\gamma_0 + \gamma_1 \frac{A}{\sigma_A}\right) \quad (15)$$

$$\sigma_E^2|A = \exp\left(\beta_0 + \beta_1 \frac{A}{\sigma_A}\right) \quad (16)$$

Note that here we divide A by $\sigma_A$ to standardize the $\gamma_1$ and $\beta_1$ parameters.

DZ twins

For DZ twins, the model is slightly more complicated as it does not hold that $A_1 = A_2$. However, $A_2$ can be modelled as a function of $A_1$. If we introduce a new random variable $A'_2$ with

$$A'_2 \sim N(0; \sigma_A^2) \quad (17)$$

which is uncorrelated to $A_1$, we can transform $A'_2$ in such a way that its transformation is correlated 0.5 with $A_1$. If we denote the transformed variable by $A_2$, we obtain

$$A_2 = \frac{1}{2}A_1 + \sqrt{1 - \left(\frac{1}{2}\right)^2} A'_2 \quad (18)$$

Note that the above follows from the Cholesky decomposition of a bivariate correlation matrix with a 0.5 correlation. Now if we assume that

$$A_1 \sim N\left(0; \sigma_A^2\right) \quad (19)$$

$A_2$ is also normally distributed with zero mean and variance $\sigma_A^2$. To obtain the ACE decomposition for $\theta_{pj}$ we condition on $A_1$ in the twin 1 sample, and on both $A_1$ and $A'_2$ in the twin 2 sample (due to Eq. 18). This results in

$$\theta_{p1}|A_1 \sim N\left(\mu_1|A_1; \sigma_1^2|A_1\right) \quad \text{for } j = 1 \quad (20)$$

and

$$\theta_{p2}|A_1, A'_2 \sim N\left(\mu_2|A_1, A'_2; \sigma_2^2|A_1, A'_2\right) \quad \text{for } j = 2 \quad (21)$$

where

$$\mu_1|A_1 = A_1 \quad (22)$$

and

$$\mu_2|A_1, A'_2 = A_2 + \rho|A_1, A'_2 \times \left(\theta_{p1}|A_1 - A_2\right). \quad (23)$$

where $A_2$ is given by Eq. 18. Note that Eq. 23 is based on the same idea as Eq. 11 in case of the MZ twins. In Eq. 23, the conditional correlation between the latent variable in the twin 1 sample, $\theta_{p1}|A_1$, and in the twin 2 sample, $\theta_{p2}|A_1, A'_2$, is given by

$$\rho|A_1, A'_2 = \frac{\sigma_{C1}|A_1 \times \sigma_{C2}|A_1, A'_2}{\sqrt{\sigma_{C1}^2|A_1 + \sigma_{E1}^2|A_1} \times \sqrt{\sigma_{C2}^2|A_1, A'_2 + \sigma_{E2}^2|A_1, A'_2}} \quad (24)$$

again this follows the idea put forward in Eq. 12 for the MZ twins. Now $\sigma_1^2|A_1$ in the conditional distribution of $\theta_{p1}|A_1$ in Eq. 20 is given by

$$\sigma_1^2|A_1 = \sigma_C^2|A_1 + \sigma_E^2|A_1 \quad (25)$$

and $\sigma_2^2|A_1, A'_2$ in Eq. 21 is given by

$$\sigma_2^2|A_1, A'_2 = \left(\sigma_C^2|A_1, A'_2 + \sigma_E^2|A_1, A'_2\right) \times \left(1 - \rho^2|A_1, A'_2\right) \quad (26)$$

Following the MZ case in Eqs. 15 and 16, GxE can be introduced similarly by making $\sigma_C^2|A_1$ and $\sigma_E^2|A_1$ from Eq. 25 and $\sigma_C^2|A_1, A'_2$ and $\sigma_E^2|A_1, A'_2$ from Eq. 26 exponential functions of $A_1$ and $A_2$, i.e.,

$$\sigma_C^2|A_1 = \exp(\gamma_0 + \gamma_1 A_1) \quad (27)$$

$$\sigma_C^2|A_1, A'_2 = \sigma_C^2|A_2 = \exp(\gamma_0 + \gamma_1 A_2) \quad (28)$$

and

$$\sigma_E^2|A_1 = \exp(\beta_0 + \beta_1 A_1) \quad (29)$$

$$\sigma_E^2|A_1, A'_2 = \sigma_E^2|A_2 = \exp(\beta_0 + \beta_1 A_2), \quad (30)$$

where $A_2$ is a function of $A_1$ and $A'_2$ given by Eq. 18.

Prior distributions

As we use a Bayesian approach to model fitting, prior distributions need to be specified on the parameters in the model. Distributions for the person parameters, $A_1$, $A'_2$, and $\delta_{pi}$ have already been specified above. The remaining parameters in the measurement model are $a_i$, $b_i$, and $r_i$. For these parameters we use

$$b_{ic} \sim \text{unif}(b_{i(c-1)}; b_{i(c+1)})$$

with $b_{i0} = -\infty$ and $b_{iC} = \infty$ as noted above. This prior ensures that the $b_{iC}$ parameters are strictly increasing for increasing $c$. An alternative way to evoke this order constraint is to use the 'ranked' function in BUGS (see Curtis 2010). In addition we use

$$a_i \sim \text{unif}(-5,5)$$

and

$$\delta_{pi} \sim \text{normal}(0,1),$$

$$r_{MZ,i} \sim \text{unif}(0,1),$$

$$r_{DZ,i} \sim \text{unif}(0,1).$$

Note that parameter range of $r_{MZ,I}$ and $r_{DZ,i}$ is in the interval $-1$, 1 as $r_i^2$ is a correlation (as discussed above). We restrict $r_{MZ,I}$ and $r_{DZ,i}$ to be positive to avoid sign switching (i.e., to avoid that the scale of the $\delta_{pi}$ reverts during the sampling). This restriction implies that the residual correlation could not be negative. In present case, this is what we expect for twin data, i.e., the correlation $r_i^2$ between the responses to a given item of twin 1 and twin 2 conditional on the latent variables, $\theta_{p1}$ and $\theta_{p2}$ is due to shared environmental and genetic influences specific to that item. However, this restriction of positive residual correlation can be relaxed in principle (but this will require a minor reparameterization of the $r_i$ parameter in Eq. 5). In specifying the priors for $r_{MZ,i}$ and $r_{DZ,I}$ we did not impose the restriction that $r_{MZ,I} > r_{DZ,I}$ as we prefer uninformative priors to ensure that the results obtained using the present approach will be similar to those results that would have been obtained using a frequentist framework (in a frequentist framework this restriction, although plausible, is commonly not specified). However, we note that incorporating this restriction on the prior of $r_{MZ,I}$ and $r_{DZ,I}$ is straightforward.

The remaining parameters in the biometric model are $\beta_0$, $\beta_1$, $\gamma_0$, $\gamma_1$, and $\sigma_A^2$. We estimate $\omega = \ln \sigma_A^2$ instead of $\sigma_A^2$, so that all these parameters can be submitted to the same prior distribution, i.e.,

$$\beta_0, \beta_1, \gamma_0, \gamma_1, \omega \sim \text{unif}(-5,5).$$

Note that these priors are quite uninformative.

### Bayesian estimation and convergence

Bayesian model estimation revolves around the posterior distribution of the parameters given the data. If $\boldsymbol{\tau}$ denotes the vector of free model parameters and $\mathbf{X}$ denotes the matrix of item scores, then the posterior distribution of the model parameters is proportional to

$$p(\tau|\mathbf{X}) \propto l(X|\tau)g(\tau)$$

where $l(.)$ is the likelihood of the data given the model parameters and $g(.)$ is the prior distribution of the parameters. For relatively simple models, $p(\boldsymbol{\tau}|\mathbf{X})$ is analytically tractable, and parameters could be calculated directly. For relatively complex models, parameter estimates are obtained by drawing samples from $p(\boldsymbol{\tau}|\mathbf{X})$ and determining the mean for each parameter in $\boldsymbol{\tau}$ across these samples.

There are many sampling schemes that differ in the exact way in which samples from the posterior are obtained. In Gibbs sampling (Geman and Geman 1984), samples are not drawn from $p(\boldsymbol{\tau}|\mathbf{X})$ directly but from the conditional distributions of a (set of) parameter(s) given the data and the remaining parameters. Samples from this procedure have been shown to converge to $p(\boldsymbol{\tau}|\mathbf{X})$. However, the procedure requires enough samples to ensure convergence. Therefore, an important aspect of this sampling based approach is that enough samples from the posterior distribution are considered to enable reliable inferences about the parameters in the model. As samples are only from the posterior after the procedure converged, the first samples are commonly omitted from the analysis (which is the so called burn-in). For the samples that follow, convergence can be checked. In this paper we use the Gelman and Rubin (1992) diagnostic to investigate convergence. For this diagnostic multiple sampling sequences need to be run using different starting values. Then, the within and between sequence variance is compared similarly to an ANOVA. If the sequences have converged, the ratio of the variances is close to one. We apply this diagnostic in the illustration section.

Another issue associated with the reliability of the sampling results concerns the correlations of a given parameter across subsequent draws from the posterior. These so-called autocorrelations should ideally be small as this indicates that the samples cover the whole range of the posterior density and not only a sub part. In the illustration section we show how autocorrelation can be used to asses reliability of the sampling results.

We implemented the model above in the freely available OpenBUGS software (Lunn et al. 2009). See Appendix A and B for the syntax given dichotomous and Likert items, respectively. In the syntax, we include references to the formula above. This implementation is based on the classic BUGS (Bayesian inference Using Gibbs Sampling) language, thus, with minor adaptation, it can also be used in the free software packages WinBUGS (Lunn et al. 2000) and JAGS (Plummer 2003). Although here we focused on a Bayesian implementation of the model, we note that present undertaking is equally amenable in a frequentist framework. That is, the model above could be fitted by maximum likelihood using software packages like SAS (SAS Institute 2011), OpenMX (Boker et al. 2010), and Mx (Neale et al. 2006). An Mx script to fit the model is available from www.dylanmolenaar.nl.

## Simulation study

### Design

We consider 4 settings: settings 1–3 involve 15, 25, and 35 dichotomously scored items, respectively, and setting 4 involves 15 Likert scale items with 5 response categories. In the measurement model, all discrimination parameters $a_i$ are set to 1, $r_{MZ,i}^2$ is set to 0.6 and $r_{DZ,i}^2$ is set to 0.3.

In each setting, we create poor scaling according to a mild and a severe scenario. In case of dichotomous items,

**Table 1** Number of item with a difficulty parameter $b_i$ located within different intervals of the θ scale for the mild and severe scenario's in the simulation study

| Scenario | Setting | Interval of θ | | |
|---|---|---|---|---|
| | | [−2.5,−1) | [−1,1) | [1,3] |
| Mild | 15 dichotmous items | 5 | 5 | 5 |
| | 25 dichotmous items | 9 | 8 | 8 |
| | 35 dichotmous items | 13 | 11 | 11 |
| | | [−2.5,−1.5] | (−1.5,1) | [1,3] |
| Severe | 15 dichotmous items | 15 | 0 | 0 |
| | 25 dichotmous items | 25 | 0 | 0 |
| | 35 dichotmous items | 35 | 0 | 0 |

the mild scenario involved the introduction of a slightly disproportional number of easy items. That is, the item difficulties are slightly disproportionally distributed across the θ range such that the sum score of these items will be poorly scaled to a minor extent. Specifically, we divided the θ range into three intervals, i.e., [−2.5,−1), [−1,1), and [1,3]. Within each interval we chose equally spaced $b_i$ values for a given number of items. See Table 1 for the exact number of items with within each interval in case of 15, 25, and 35 dichotomous items. As a result of this setup, scaling of θ within each interval is good, however, over the whole θ range, scaling is slightly uneven due to the somewhat smaller range of the first interval and due to the difference in the number of items in case of 25 and 35 items (see Table 1). In case of the Likert items, item category parameters, $b_{i1}$ to $b_{i4}$ were fixed to −1, 0, 1, and 2, which results in a disproportionate number of responses in the first category.

In the severe scenario, we specified all dichotomous item difficulties to be equally spaced within the interval [−2.5,−1.5], see Table 1. In case of Likert items, we fixed the item category parameters, $b_{i1}$ to $b_{i4}$, to 0, 1, 3, and 4, which results in a severe floor effect. See Fig. 2 for examples of the resulting sum (dichotomous items) and item (Likert items) score distribution in the two scenarios. Note that in both scenarios, for dichotomous items, the sum score is poorly scaled as it contains more information about θ at the lower θ range, because of the disproportional number of items with a low item difficulty in the mild scenario, and no intermediate or hard items at all in case of the severe scenario. In addition, all Likert items display a floor effect due to the majority of responses being in the lower answer categories. In the mild scenario the effects are just visible (but detectable, as we show below). The severe scenario is quite extreme but by no means uncommon. For instance, intelligence tests are commonly characterized by relatively more easy items while depression questionnaires commonly suffer from floor effects in normal populations.

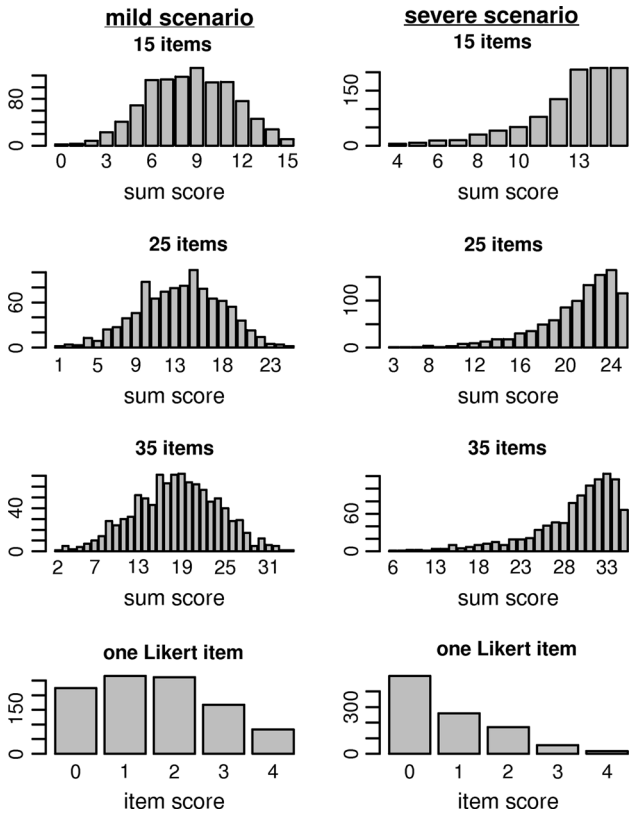


**Fig. 2** Example of resulting (sum) score distributions in case of 15, 25, and 35 dichotomous items, and the resulting item score distribution in case of a single Likert item. *Plots* are generated using the parameters as chosen for the simulation study. *Left hand column* concerns the mild scenario and the *right hand column* concerns the severe scenario

In the biometric model, parameters for ω, $\beta_0$, and $\gamma_0$ were chosen to equal log(0.5), log(0.25), and log(0.25), respectively, so that heritability equaled 0.5 in de absence of GxE. For each setting we simulated 50 datasets without GxE, i.e., $\beta_1 = 0$ and $\gamma_1 = 0$, and 50 datasets with GxE, using $\beta_1 = 0.25$ and $\gamma_1 = 0.25$. Note that this effect size corresponds to a 'moderate' effect in the Molenaar et al. (2012) study. For each replication, we simulated data for 1,000 MZ and 1,000 DZ twins. To each dataset we fitted the model as described above to investigate (1) parameters recovery, (2) the rate with which true AxE and AxC are detected ('hit rate'), and (3) to investigate the rate with which spurious GxE arises ('false positives rate'). In case of dichotomous items (setting 1–3), we used a 1 parameter measurement model (i.e., $a_i$ in Eq. 7 is fixed to equal 1 for all *i*). In case of the Likert items, we used the 2-parameter model in Eq. 7, in which we fixed the discrimination parameter of the first item to equal 1 for identification purposes ($a_1 = 1$) and estimated the remaining discrimination parameters.

In addition to applying the methodology above, we also fitted the univariate model to the sum scores of the items using the marginal maximum likelihood routine from Molenaar et al. (2012). By doing so, we aimed to illustrate (1) that spurious GxE can arise at sum score level (i.e., increased false positive rate), and (2) true GxE effects can be masked at sum score level (i.e., decreased hit rate). In case of the full Bayesian GxE model, hit rates and false positives rates were determined by assessing the percentage of the replications in the 95 % highest posterior density (HPD) regions of the GxE parameters $\beta_1$ and $\gamma_1$ included the value 0. For instance, when AxE is present in the data, and the 95 % HPD of $\beta_1$ does not include 0, the hit rate increases as the AxE is correctly detected. Similarly, if AxE is not in the data, but the 95 % HPD of $\beta_1$ does not include 0, this is a false positive. In case of the application of the univariate GxE model to the sum scores, hit rates and false positive rates were determined using the power of the likelihood ratio test to detect AxE and AxC effects (see Satorra and Saris 1985; Saris and Satorra 1993; Dolan and van den Berg 2008). Specifically, power to detect AxE was assessed by determining the power of the likelihood ratio test to reject a model with only AxC in favor of a model with both AxE and AxC. Similarly, power to detect AxC was calculated by determining the power to reject a model with only AxE in favor of a model with both AxE and AxC. If GxE is truly present, the power coefficient is an estimate of the hit rate; if GxE is truely absent, the power coefficient is an estimate of the false positive rate. For the power analyses in the case of the sum scores, we used a 0.05 level of significance.

For each full model application, we drew 2,000 samples from the posterior distribution as burn-in. Next, we drew an additional 2,000 samples from which we determined the mean of all parameters in the model. From experiences with fitting the model to simulated data, we knew that this number of draws is sufficient to ensure that the chains are converged to their stationary distributions. However, we note that this does not imply that in practice this scheme (4,000 samples; 2,000 burn-in) will ensure reliable sampling results. Therefore, in practical applications, we recommend that convergence criteria are carefully considered as we will discuss more fully in the illustration section.

For the full model we used the OpenBUGS code in the Appendices. For the univariate model applications on the sum scores, we used the Mx software program (Neale et al. 2006) using the scripts from Molenaar et al. (2012).

## Results

### Scenario 1: mild scale problems

#### *Parameter recovery*

The true parameter values and the mean and standard deviation of the posterior parameter distributions, averaged

**Table 2** Mild scenario: posterior means (standard deviation) of the parameters in the measurement model part of the full model

| Items | GxE | $a_1$ | $a_2$ | $a_3$ | $a_4$ | $a_5$ | MZ | | | | | DZ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | $r_1$ | $r_2$ | $r_3$ | $r_4$ | $r_5$ | $r_1$ | $r_2$ | $r_3$ | $r_4$ | $r_5$ |
| *True values* | | *1.00* | *1.00* | *1.00* | *1.00* | *1.00* | *0.60* | *0.60* | *0.60* | *0.60* | *0.60* | *0.30* | *0.30* | *0.30* | *0.30* | *0.30* |
| 15 dicho | No | – | – | – | – | – | 0.54 | 0.55 | 0.55 | 0.58 | 0.58 | 0.22 | 0.23 | 0.25 | 0.27 | 0.24 |
| | | – | – | – | – | – | (0.13) | (0.11) | (0.08) | (0.07) | (0.06) | (0.13) | (0.13) | (0.14) | (0.11) | (0.11) |
| | Yes | – | – | – | – | – | 0.52 | 0.55 | 0.57 | 0.58 | 0.58 | 0.25 | 0.22 | 0.26 | 0.30 | 0.29 |
| | | – | – | – | – | – | (0.13) | (0.11) | (0.09) | (0.07) | (0.06) | (0.16) | (0.11) | (0.12) | (0.12) | (0.09) |
| 25 dicho | No | – | – | – | – | – | 0.57 | 0.50 | 0.56 | 0.57 | 0.56 | 0.23 | 0.23 | 0.27 | 0.24 | 0.28 |
| | | – | – | – | – | – | (0.12) | (0.13) | (0.10) | (0.11) | (0.10) | (0.14) | (0.15) | (0.13) | (0.13) | (0.12) |
| | Yes | – | – | – | – | – | 0.54 | 0.55 | 0.55 | 0.56 | 0.56 | 0.21 | 0.23 | 0.25 | 0.25 | 0.24 |
| | | – | – | – | – | – | (0.14) | (0.12) | (0.11) | (0.11) | (0.08) | (0.13) | (0.12) | (0.15) | (0.12) | (0.12) |
| 35 dicho | No | – | – | – | – | – | 0.53 | 0.54 | 0.55 | 0.57 | 0.56 | 0.20 | 0.23 | 0.24 | 0.26 | 0.26 |
| | | – | – | – | – | – | (0.16) | (0.15) | (0.11) | (0.11) | (0.09) | (0.14) | (0.15) | (0.12) | (0.14) | (0.14) |
| | Yes | – | – | – | – | – | 0.50 | 0.54 | 0.56 | 0.56 | 0.57 | 0.24 | 0.24 | 0.25 | 0.26 | 0.22 |
| | | – | – | – | – | – | (0.14) | (0.13) | (0.10) | (0.12) | (0.10) | (0.16) | (0.13) | (0.15) | (0.15) | (0.11) |
| Likert | No | 1[a] | 1.00 | 1.01 | 1.01 | 1.00 | 0.60 | 0.60 | 0.60 | 0.60 | 0.60 | 0.29 | 0.29 | 0.30 | 0.30 | 0.29 |
| | | – | (0.03) | (0.03) | (0.03) | (0.04) | (0.03) | (0.03) | (0.02) | (0.02) | (0.03) | (0.03) | (0.04) | (0.04) | (0.04) | (0.04) |
| | Yes | 1[a] | 1.02 | 1.01 | 1.01 | 1.01 | 0.59 | 0.60 | 0.60 | 0.59 | 0.59 | 0.29 | 0.29 | 0.29 | 0.29 | 0.30 |
| | | – | (0.04) | (0.03) | (0.04) | (0.04) | (0.03) | (0.03) | (0.03) | (0.02) | (0.03) | (0.04) | (0.04) | (0.04) | (0.04) | (0.03) |

[a] The discrimination parameter of the first item, $a_1$, is fixed to equal 1

over the 50 replications within each cell, are depicted in Table 2 (measurement model) and in Table 3 (biometric model). We did not tabulate the $b_{ic}$ parameters to save space. However, results indicated that these parameters are well recovered in all settings.

As can be seen from Table 2, the residual correlations, $r^2_{MZ,i}$ and $r^2_{DZ,i}$ are well recovered in case of the Likert items, but slightly underestimated in case of dichotomous items. In addition, in the Likert case, the discrimination parameters, $a_i$, are recovered well. In Table 3, it can be seen that parameter recovery in the biometric model is generally acceptable. Most importantly, GxE parameters $\beta_1$ and $\gamma_1$ appear to be recovered well. Specifically, in the absence of AxE and AxC, the posterior means of $\beta_1$ and $\gamma_1$ are nearly 0, and in the presence of AxE and AxC, the posterior means of $\beta_1$ and $\gamma_1$ are close to their true value 0.25. From the table, it also appears that $\ln \sigma^2_A$ tends to be slightly overestimated while $\gamma_0$ tends to be slightly underestimated. We think this is not a great problem as it is well established in ACE twin modeling that it is relatively more difficulty to resolve C and A than A and E (Martin et al. 1978).

Generally, we conclude that parameter recovery of the full model is good in the case of Likert items, and acceptable for the dichotomous items. In the latter case, residual correlations are somewhat underestimated, but given the size of the model and the relative little information available concerning individual differences (i.e., only the 0's and 1's of the binary items), we think that these results are tolerable.

### *False positive rate and hit rate*

Table 4 contains results concerning the hit rates and the false positive rates. As can be seen, in the case of dichotomous items, hit rates in the sum score analyses are close to the level of significance, 0.05, meaning that the true GxE effects in the data are rarely detected. This can also be seen in Table 5, where parameter estimates for the sum score

**Table 3** Mild scenario: posterior means (standard deviation) averaged over replications for the parameters in the biometrical part of the full model

| GxE | Items | $\ln \sigma^2_A$ | $\beta_0$ | $\beta_1$ | $\gamma_0$ | $\gamma_1$ |
|---|---|---|---|---|---|---|
| Yes | *True values* | *−0.73* | *−1.39* | *0.25* | *−1.39* | *0.25* |
| | 15 dichotomous | −0.69 (0.20) | −1.41 (0.08) | 0.25 (0.10) | −1.65 (0.59) | 0.29 (0.17) |
| | 25 dichotomous | −0.67 (0.14) | −1.40 (0.06) | 0.27 (0.08) | −1.52 (0.34) | 0.27 (0.14) |
| | 35 dichotomous | −0.68 (0.14) | −1.40 (0.05) | 0.27 (0.07) | −1.56 (0.42) | 0.27 (0.16) |
| | 15 Likert | −0.71 (0.13) | −1.41 (0.08) | 0.25 (0.07) | −1.53 (0.29) | 0.29 (0.13) |
| No | *True values* | *−0.73* | *−1.39* | *0.00* | *−1.39* | *0.00* |
| | 15 dichotomous | −0.67 (0.19) | −1.40 (0.07) | 0.02 (0.11) | −1.72 (0.72) | −0.01 (0.16) |
| | 25 dichotomous | −0.66 (0.13) | −1.41 (0.06) | −0.01 (0.09) | −1.60 (0.42) | 0.02 (0.12) |
| | 35 dichotomous | −0.67 (0.13) | −1.40 (0.06) | −0.01 (0.06) | −1.60 (0.45) | 0.00 (0.13) |
| | 15 Likert | −0.66 (0.16) | −1.40 (0.06) | −0.01 (0.06) | −1.98 (1.22) | 0.00 (0.11) |

**Table 4** Mild scenario: hit rate (rate with which AxE or AxC are correctly detected) and false positive rate (rate with which AxE and AxC are falsely detected) with 95 % confidence intervals

| Items | Hit rate | | | | False positive rate | | | |
|---|---|---|---|---|---|---|---|---|
| | Sum score | | Full model | | Sum score | | Full model | |
| | AxE | AxC | AxE | AxC | AxE | AxC | AxE | AxC |
| 15 dicho | 0.13 | 0.05 | 0.70 | 0.44 | 0.77 | 0.13 | 0.12 | 0.04 |
| | (0.04; 0.22) | (0.00[a]; 0.11) | (0.57; 0.83) | (0.30; 0.58) | (0.65; 0.89) | (0.04; 0.22) | (0.03; 0.21) | (0.00; 0.09) |
| 25 dicho | 0.08 | 0.05 | 0.96 | 0.56 | 0.89 | 0.19 | 0.08 | 0.02 |
| | (0.00; 0.16) | (0.00[a]; 0.11) | (0.91; 1.00[a]) | (0.42; 0.70) | (0.80; 0.98) | (0.08; 0.30) | (0.00; 0.16) | (0.00; 0.06) |
| 35 dicho | 0.05 | 0.05 | 1.00 | 0.60 | 0.90 | 0.23 | 0.02 | 0.04 |
| | (9.01; 0.11) | (0.00[a]; 0.11) | – | (0.46; 0.74) | (0.82; 0.98) | (0.11; 0.35) | (0.00; 0.06) | (0.00; 0.09) |
| 15 likert | 1.00 | 0.99 | 0.94 | 0.70 | 1.00 | 0.88 | 0.04 | 0.02 |
| | – | (0.96; 1.02) | (0.87; 1.00[a]) | (0.57; 0.83) | – | (0.79; 0.97) | (0.00; 0.09) | (0.00; 0.06) |

[a] These bounds are fixed to 0 or 1, because the actual bound exceeded the theoretical [0,1] interval. All other occurrences of lower bounds equal to 0 are due to rounding. Note that in the case of a 1.00 false positive or hit rate, the confidence interval is not calculated as the standard error equals 0. In the case of the sum score analysis, hit rates and false positive rates are based on the power of the likelihood ratio test to detect the corresponding effect

**Table 5** Mild scenario: mean marginal maximum likelihood parameter estimates (standard deviation) of the parameters in the univariate sum score analysis

| GxE | Items | ln $\sigma_A^2$ | $\beta_0$ | $\beta_1$ | $\gamma_0$ | $\gamma_1$ |
|---|---|---|---|---|---|---|
| Yes | *True values* | −0.73 | −1.39 | 0.25 | −1.39 | 0.25 |
| | 15 dichotomous | −0.71 (0.09) | −1.11 (0.06) | −0.08 (0.06) | −1.82 (0.59) | −0.02 (0.09) |
| | 25 dichotomous | −0.73 (0.09) | −1.20 (0.05) | −0.07 (0.06) | −1.57 (0.28) | −0.03 (0.07) |
| | 35 dichotomous | −0.71 (0.07) | −1.24 (0.05) | −0.05 (0.05) | −1.53 (0.25) | −0.05 (0.08) |
| | 15 Likert | −0.69 (0.04) | −1.40 (0.06) | 0.48 (0.04) | −1.58 (0.13) | 0.39 (0.04) |
| No | *True values* | *−0.73* | *−1.39* | *0.00* | *−1.39* | *0.00* |
| | 15 dichotomous | −0.73 (0.07) | −1.13 (0.04) | −0.19 (0.05) | −1.71 (0.25) | −0.14 (0.09) |
| | 25 dichotomous | −0.71 (0.09) | −1.22 (0.06) | −0.22 (0.05) | −1.63 (0.30) | −0.16 (0.07) |
| | 35 dichotomous | −0.71 (0.06) | −1.27 (0.05) | −0.24 (0.04) | −1.55 (0.17) | −0.16 (0.08) |
| | 15 Likert | −0.69 (0.06) | −1.35 (0.05) | 0.35 (0.04) | −1.58 (0.17) | 0.29 (0.04) |

analyses are given. It can be seen that in case of dichotomous items, when the GxE effect is present, $\beta_1$ and $\gamma_1$ estimates are close to 0. Thus, the skewness and change in twin correlations due to the GxE effect in the data are masked by the poor scaling of the sum score. For the full model, hit rates of the AxE effect are acceptable in the case of 15 dichotomous items (i.e., .70) and good in the other cases (at least .94). Hit rates of the AxC effect are poor for all the cases with dichotomous items (0.60 at most). This is in line with the results by Molenaar et al. (2012) who showed that power to detect AxC is generally low. However, in present approach, for the Likert items, hit rates are acceptable (0.70), which is much better than in the univariate results of the Molenaar et al. paper. This is because here we use at least 30 Likert variables (15 for each twin), while in the univariate model of Molenaar et al. only 2 continuous variables are used.

The false positive rates in Table 4 should ideally all be close to the 0.05 level (reflecting either the level of significance or the probability of the HPD region). The results in Table 4 show that this is not the case for tests of AxE based on the sum score. False positive rates ranged between .77 and 1.00 (see also Eaves 2006). The mean parameter estimates of $\beta_1$ and $\gamma_1$ (see Table 5)—in the absence of GxE—clearly depart from 0. That is, both are smaller than 0 in the case of dichotomous items, and larger than 0 for the Likert items. With respect to AxC, the false positive rate of the sum score is not too bad in the case of dichotomous items (.23 at most). In addition, in the case of the Likert items the test of AxC is also associated with a large false positive rate (.88). In case of the full model, all false positive rates are reasonably close to the 0.05 rate.

## Scenario 2: severe scale problems

### *Parameter recovery*

The true parameter values and the mean and standard deviation of the posterior parameter distributions, averaged over the 50 replications within each cell, are shown in Table 6 (measurement model) and in Table 7 (biometric model) for the severe scenario. Again, we did not tabulate the $b_{ic}$ parameters, but results indicated that these parameters are well recovered in all settings.

The pattern of results concerning the measurement model parameter recovery in Table 6 is generally the same as in the mild scenario discussed above. That is, the residual correlations, $r^2_{MZ,i}$ and $r^2_{DZ,i}$ are well recovered in case of the Likert items, but slightly underestimated in case of dichotomous items. In the Likert case, where we introduced discrimination parameters, $a_i$, these parameters are recovered well.

Results concerning the biometric model parameters in Table 7 indicate that in the severe scenario, variance of A is overestimated and the variance of C is underestimated for both dichotomous and Likert items. This was also evident in the mild scenario. However in the present scenario, the bias is greater. Parameter recovery of the AxE parameter, $\beta_1$, was good in both the GxE condition (true value 0.25) and the no-GxE condition (true value 0). On the contrary, AxC parameter $\gamma_1$ is somewhat overestimated both when GxE is present and when GxE is not present in case of dichotomous items. As judged by the standard deviations of the mean parameter estimates of $\gamma_1$, parameter variability is larger as compared to the mild scenario in Table 3. Judged by the large standard deviation, this overestimation is still within a reasonable range. This is also evident in the false positive rate, as discussed below. In addition, the overestimation decreases when using more dichotomous items. In case of Likert items, parameter $\gamma_1$ is recovered well.

Taken together, parameter recovery is not as accurate as in the mild scenario. Most notably, the A and C components are somewhat less well resolved. In addition, $\gamma_1$ is somewhat overestimated in the case of dichotomous items. However, the AxE parameter $\beta_1$ is still well recovered.

**Table 6** Severe scenario: posterior means (standard deviation) of the parameters in the measurement model part of the full model

| Items | GxE | $a_1$ | $a_2$ | $a_3$ | $a_4$ | $a_5$ | MZ | | | | | DZ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | $r_1$ | $r_2$ | $r_3$ | $r_4$ | $r_5$ | $r_1$ | $r_2$ | $r_3$ | $r_4$ | $r_5$ |
| *True values* | | *1.00* | *1.00* | *1.00* | *1.00* | *1.00* | *0.60* | *0.60* | *0.60* | *0.60* | *0.60* | *0.30* | *0.30* | *0.30* | *0.30* | *0.30* |
| 15 dicho | No | – | – | – | – | – | 0.53 | 0.55 | 0.52 | 0.55 | 0.55 | 0.23 | 0.23 | 0.24 | 0.25 | 0.25 |
| | | – | – | – | – | – | (0.13) | (0.12) | (0.14) | (0.13) | (0.12) | (0.13) | (0.14) | (0.14) | (0.18) | (0.14) |
| | Yes | – | – | – | – | – | 0.56 | 0.54 | 0.52 | 0.53 | 0.55 | 0.21 | 0.23 | 0.23 | 0.22 | 0.28 |
| | | – | – | – | – | – | (0.13) | (0.12) | (0.13) | (0.12) | (0.08) | (0.12) | (0.13) | (0.15) | (0.13) | (0.13) |
| 25 dicho | No | – | – | – | – | – | 0.52 | 0.55 | 0.52 | 0.55 | 0.57 | 0.23 | 0.21 | 0.22 | 0.23 | 0.20 |
| | | – | – | – | – | – | (0.14) | (0.12) | (0.14) | (0.11) | (0.11) | (0.14) | (0.12) | (0.15) | (0.14) | (0.12) |
| | Yes | – | – | – | – | – | 0.54 | 0.52 | 0.53 | 0.54 | 0.53 | 0.28 | 0.20 | 0.23 | 0.26 | 0.24 |
| | | – | – | – | – | – | (0.14) | (0.15) | (0.15) | (0.12) | (0.15) | (0.18) | (0.14) | (0.13) | (0.12) | (0.13) |
| 35 dicho | No | – | – | – | – | – | 0.55 | 0.55 | 0.53 | 0.53 | 0.58 | 0.21 | 0.22 | 0.27 | 0.24 | 0.25 |
| | | – | – | – | – | – | (0.11) | (0.14) | (0.12) | (0.13) | (0.12) | (0.12) | (0.13) | (0.15) | (0.12) | (0.14) |
| | Yes | – | – | – | – | – | 0.58 | 0.55 | 0.57 | 0.53 | 0.56 | 0.24 | 0.20 | 0.22 | 0.24 | 0.24 |
| | | – | – | – | – | – | (0.12) | (0.13) | (0.12) | (0.15) | (0.13) | (0.14) | (0.14) | (0.11) | (0.14) | (0.14) |
| Likert | No | 1[a] | 1.01 | 1.01 | 1.01 | 1.01 | 0.60 | 0.60 | 0.59 | 0.59 | 0.59 | 0.30 | 0.29 | 0.30 | 0.29 | 0.28 |
| | | – | (0.04) | (0.04) | (0.05) | (0.05) | (0.03) | (0.03) | (0.03) | (0.03) | (0.03) | (0.04) | (0.04) | (0.05) | (0.05) | (0.05) |
| | Yes | 1[a] | 1.01 | 1.01 | 1.01 | 1.00 | 0.60 | 0.60 | 0.60 | 0.59 | 0.59 | 0.29 | 0.29 | 0.28 | 0.28 | 0.30 |
| | | – | (0.04) | (0.04) | (0.05) | (0.04) | (0.03) | (0.03) | (0.03) | (0.03) | (0.03) | (0.04) | (0.05) | (0.05) | (0.04) | (0.05) |

**Table 7** Severe scenario: posterior means (standard deviation) averaged over replications for the parameters in the biometrical part of the full model

| GxE | Items | $\ln \sigma^2_A$ | $\beta_0$ | $\beta_1$ | $\gamma_0$ | $\gamma_1$ |
|---|---|---|---|---|---|---|
| Yes | *True values* | *−0.73* | *−1.39* | *0.25* | *−1.39* | *0.25* |
| | 15 dichotomous | −0.63 (0.24) | −1.39 (0.14) | 0.31 (0.15) | −1.73 (0.66) | 0.48 (0.32) |
| | 25 dichotomous | −0.62 (0.21) | −1.39 (0.12) | 0.27 (0.16) | −1.73 (0.65) | 0.43 (0.27) |
| | 35 dichotomous | −0.68 (0.18) | −1.41 (0.11) | 0.24 (0.14) | −1.55 (0.45) | 0.42 (0.27) |
| | 15 Likert | −0.67 (0.12) | −1.41 (0.10) | 0.24 (0.08) | −1.54 (0.41) | 0.30 (0.13) |
| No | *True values* | *−0.73* | *−1.39* | *0.00* | *−1.39* | *0.00* |
| | 15 dichotomous | −0.56 (0.30) | −1.38 (0.18) | 0.05 (0.21) | −1.79 (0.79) | 0.47 (0.53) |
| | 25 dichotomous | −0.61 (0.30) | −1.41 (0.14) | 0.00 (0.15) | −1.77 (0.63) | 0.27 (0.49) |
| | 35 dichotomous | −0.63 (0.22) | −1.41 (0.12) | 0.01 (0.13) | −1.76 (0.66) | 0.10 (0.31) |
| | 15 Likert | −0.65 (0.17) | −1.42 (0.09) | 0.01 (0.09) | −1.77 (0.96) | 0.00 (0.19) |

*False positive rate and hit rate*

Table 8 contains results concerning the hit rates and the false positive rate. In the table, only results for the full model are presented as it was computationally not possible to conduct univariate analyses on the sum scores due to the severe floor and ceiling effects in the data (in the majority of the cases, the estimation did not converge).

As can be seen in Table 8, for dichotomous items, hit rates are moderate with rates roughly around 0.40. False positive rates are acceptable for AxE with rates close to the 0.05 rate. For AxC, the false positive rates tend to be inflated in case of 15 and 25 dichotmous items, with rates of 0.20. However, this is only a minor deviance of the 0.05 level as judged by the confidence interval. In case Likert scale items, the hit rate is good for AxE (0.84) and moderate for AxC (0.64). In addition, false positives are close to 0.05.

## Illustration

We analyzed responses of 308 MZ twin pairs and 447 DZ twin pairs (mean age 45.50; min. 25; max. 74) to an affect questionnaire (Mroczek and Kolarz 1998), which was administered in the National Survey of Midlife Development in the United States (MIDUS) in 1995–1996 under the auspices of the Inter-university Consortium for Political and Social Research (ICPSR; Brim et al. 2010).[3] The

[3] The opinions expressed in this article are those of the authors and do not necessarily reflect the views of the ICPSR.

questionnaire consists of 12 items. Each item describes a particular affect. Respondents had to indicate on a 5 point Likert scale to what degree they had experienced that affect in the last 30 days. Items 1–6 concerned positive affect and items 7–12 concerned negative affect.

## Convergence diagnostics

Similar to the simulation study, we used 4,000 draws from the posterior distribution omitting the first 2,000 as burn-in.

**Table 8** Severe scenario: hit rate (rate with which AxE or AxC are correctly detected) and false positive rate (rate with which AxE and AxC are spuriously detected) with 95 % confidence bounds for the full model

| Items | Hit rate | | False positive | |
|---|---|---|---|---|
| | AxE | AxC | AxE | AxC |
| 15 dichotomous | 0.42 | 0.34 | 0.06 | 0.20 |
| | (0.28; 0.56) | (0.21; 0.47) | (0.00[a]; 0.13) | (0.09; 0.31) |
| 25 dichotomous | 0.40 | 0.28 | 0.10 | 0.20 |
| | (0.26; 0.54) | (0.16; 0.40) | (0.02; 0.18) | (0.09; 0.31) |
| 35 dichotomous | 0.38 | 0.46 | 0.02 | 0.12 |
| | (0.25; 0.51) | (0.32; 0.60) | (0.00[a]; 0.06) | (0.03; 0.21) |
| 15 likert | 0.84 | 0.64 | 0.06 | 0.06 |
| | (0.74; 0.94) | (0.51; 0.77) | (0.00[a]; 0.13) | (0.00[a]; 0.13) |

[a] For these cases the lower bound of the confidence interval is fixed to 0 as the actual lower bound is smaller than 0

To ensure that this scheme is sufficient for convergence, we considered (1) trace plots of the draws from the posterior distribution; (2) autocorrelations between the parameters across subsequent subsets ('lags') of draws; and (3) the Gelman and Rubin statistic (Gelman and Rubin (1992) discussed above. All of these diagnostics are available in the R package 'coda' (Plummer et al. 2005). As we have many parameters, we focus on the two GxE parameters, $\beta_1$ and $\gamma_1$, that are of main interest to present application. As the Gelman and Rubin statistic needs multiple sequences of draws from the posterior, we used three such chains using different starting values.

First, see Fig. 3 for the trace plots of the samples for one of the chains. As can be seen, these seem to vary randomly around a stable average for both parameters. In case of non-convergence, the samples would have been drifting away from the running average. Next, the autocorrelations are plotted for the GxE parameters in Fig. 4 for 100 lags. As can be seen, for increasing lags, these correlations approach zero for both parameters. Finally, the Gelman and Rubin diagnostic equaled 1.02 and 1.07 for $\beta_1$ and $\gamma_1$ respectively which is judged to be sufficiently close to 1 (commonly, thresholds of 1.1 or 1.2 are used). From the above, we conclude that our scheme (4,000 samples, 2,000 burn-in) is sufficient for present purposes. Below, we present results from a new sequence of samples from the posterior using this scheme.

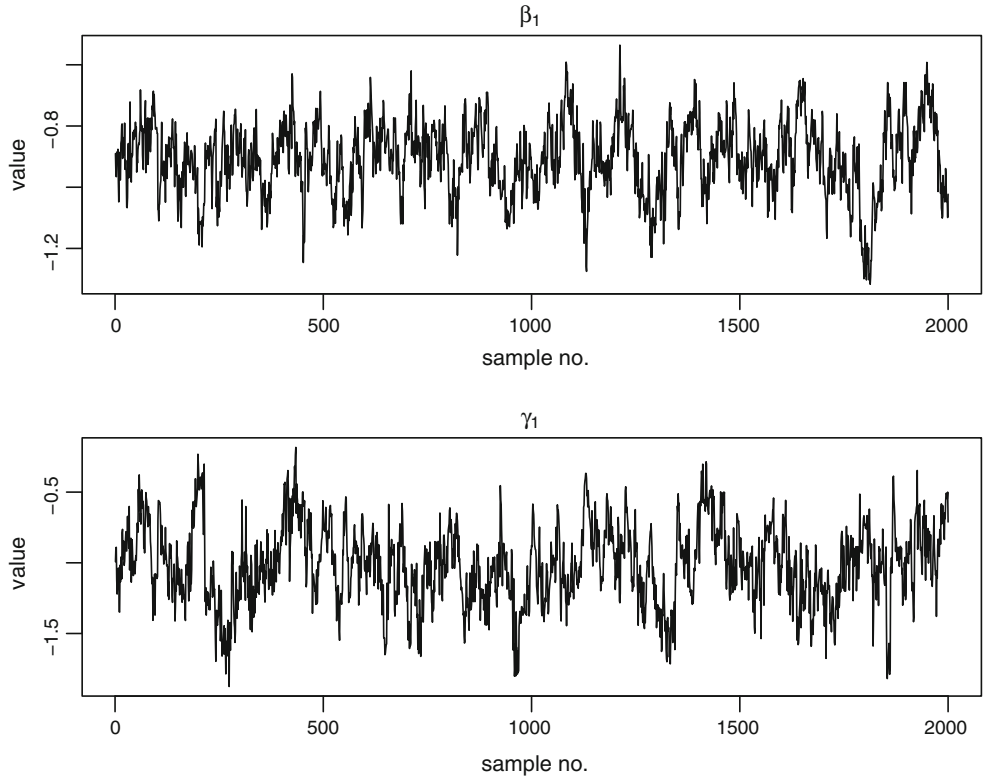


**Fig. 3** *Trace plots* of the draws from the posterior distribution for $\beta_1$ and $\gamma_1$

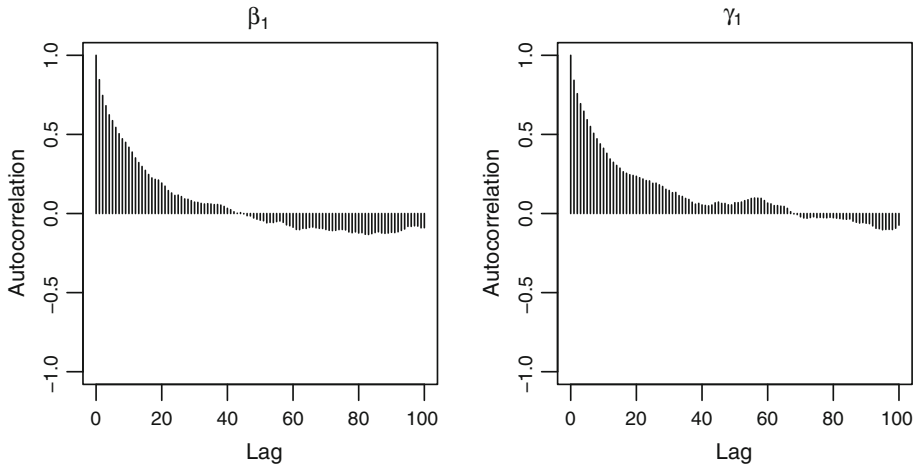


**Fig. 4** Autocorrelations for $\beta_1$ and $\gamma_1$ across 100 lags

**Table 9** Means, standard deviations, and 95 % HPD regions of the posterior parameters in the measurement model

| *i* | Discrimination, $a_i$ | | | | DZ residual cor, $r_i^2$ | | | | MZ residual cor, $r_i^2$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Posterior | | 95 % HPD | | Posterior | | 95 % HPD | | Posterior | | 95 % HPD | |
| | Mean | SD | Lower | Upper | Mean | SD | Lower | Upper | Mean | SD | Lower | Upper |
| 1 | 1 | – | – | – | 0.09 | 0.07 | 0.00 | 0.24 | 0.11 | 0.09 | 0.00 | 0.31 |
| 2 | 0.82 | 0.06 | 0.70 | 0.94 | 0.10 | 0.06 | 0.00 | 0.23 | 0.39 | 0.07 | 0.23 | 0.51 |
| 3 | 0.78 | 0.06 | 0.67 | 0.91 | 0.11 | 0.06 | 0.00 | 0.22 | 0.19 | 0.09 | 0.01 | 0.37 |
| 4 | 1.47 | 0.10 | 1.27 | 1.67 | 0.21 | 0.13 | 0.00 | 0.49 | 0.25 | 0.15 | 0.01 | 0.55 |
| 5 | 0.95 | 0.08 | 0.82 | 1.12 | 0.05 | 0.05 | 0.00 | 0.16 | 0.10 | 0.08 | 0.00 | 0.27 |
| 6 | 1.17 | 0.09 | 1.00 | 1.36 | 0.22 | 0.11 | 0.01 | 0.43 | 0.36 | 0.15 | 0.03 | 0.62 |
| 7 | −1.64 | 0.13 | −1.94 | −1.42 | 0.07 | 0.06 | 0.00 | 0.22 | 0.32 | 0.10 | 0.13 | 0.50 |
| 8 | −2.03 | 0.18 | −2.43 | −1.73 | 0.06 | 0.06 | 0.00 | 0.23 | 0.08 | 0.08 | 0.00 | 0.28 |
| 9 | −1.37 | 0.11 | −1.59 | −1.19 | 0.15 | 0.08 | 0.00 | 0.29 | 0.32 | 0.08 | 0.16 | 0.46 |
| 10 | −1.44 | 0.11 | −1.67 | −1.26 | 0.02 | 0.02 | 0.00 | 0.08 | 0.28 | 0.09 | 0.11 | 0.45 |
| 11 | −1.49 | 0.12 | −1.74 | −1.30 | 0.02 | 0.02 | 0.00 | 0.08 | 0.09 | 0.07 | 0.00 | 0.26 |
| 12 | −1.61 | 0.12 | −1.86 | −1.42 | 0.08 | 0.07 | 0.00 | 0.23 | 0.10 | 0.08 | 0.00 | 0.27 |

$a_1$ is fixed to equal 1 for identification reasons

## Results

The means, standard deviations, and 95 % HPD regions for the posterior distributions of the parameters $a_i$ and $r_i^2$ of the measurement model are depicted in Table 9. To save space, we did not tabulate the $b_{ic}$ parameter (as these are 48 parameters), but to illustrate: for item 1 the posterior means (sd) were −3.83 (0.21), −2.96 (0.13), −1.88 (0.08), and −0.75 (0.06) for $\beta_{11}$ to $\beta_{14}$, respectively. As can be seen in Table 9, in the measurement model, DZ residual correlations, $r_i^2$ are small. For all but item 6, the HPD contains 0 (note that 0 is also the lower boundary of the parameter, the lower bound of the HPD can thus not be smaller than 0). In the MZ twins, the HPD of $r_i^2$ does not include 0 for 7 of the 12 items. As can also be seen from the table, the posterior mean of $a_i$ is negative for items 7–12, which is to be expected as these items concern negative affect.

In Table 10, the mean, standard deviation, and 95 % HPD region for the posterior parameter values of the biometric model are depicted. From these estimates, marginal heritability was calculated to be 0.08 [using Eq. 3, but replacing $\sigma_A^2$ with $\exp(\omega)$], standardized marginal variances of C and E were respectively 0.28 and 0.63. As can be seen, both AxE and AxC interactions are detected in the data, as indicated by the HPD of $\beta_1$ and $\gamma_1$. Posterior means of both $\beta_1$ and $\gamma_1$ are negative, indicating that the variance of E and C is decreasing for increasing levels of A. See Fig. 5 for a graphical representation of how $\sigma_E^2|A$ and $\sigma_E^2|A$ vary across the additive genetic factor A.

**Table 10** Means, standard deviations, and 95 % HPD regions of the posterior parameters in the biometric model

| Par | Posterior | | 95 % HPD | |
|---|---|---|---|---|
| | Mean | SD | Lower | Upper |
| $\omega$ | −2.53 | 0.29 | −3.04 | −1.92 |
| $\beta_0$ | −0.89 | 0.16 | −1.19 | −0.60 |
| $\beta_1$ | −0.88 | 0.12 | −1.11 | −0.66 |
| $\gamma_0$ | −1.77 | 0.35 | −2.62 | −1.26 |
| $\gamma_1$ | −0.97 | 0.35 | −1.75 | −0.41 |

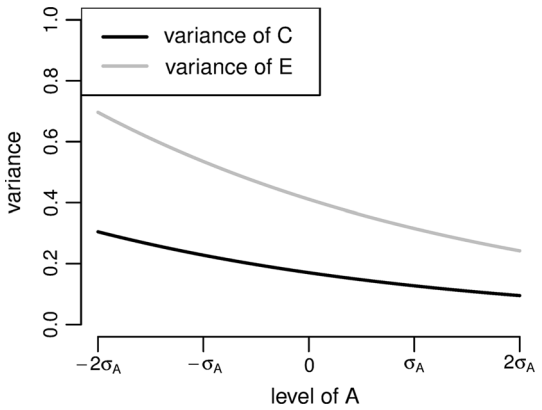


**Fig. 5** GxE in the affect data: environmental variances E and C as a function of additive genetic effects, A

## Discussion

We illustrated the well-know finding (Eaves 2006; Eaves et al. 1977; Purcell 2002) that spurious GxE can arise in sum score analyses due to poor scaling. As a solution, we proposed a model which takes the measurement properties of the individual items into account. Based on the results of simulation study, we consider the method to be viable. Using this method we showed that poor scaling lowers the power to detect GxE, but does not affect the false positive rate, except for a small effect concerning AxC. That is, present approach does not result in spurious GxE. In that sense, we illustrated that—by specifying an appropriate measurement model for the data—the problem of poor scaling in GxE research can be overcome. However, the price that one has to pay is that, as scaling problems increase, larger sample sizes and more items are necessary to ensure an acceptable power to detect GxE, especially in the case of dichotomous items.

From the simulation study, we draw four main conclusions: First, AxC is most sensitive to scaling problems. That is, with increasing scaling problems, the hit rates of AxC declines substantially, and false positives rate increase slightly. AxE is only affected in terms of a decreased hit rate, i.e., the effect is adequately detected. Second, for greater scaling problems, the A and C factor are more difficult to resolve, with the A factor being somewhat overestimated and the C factor being somewhat underestimated. Third, Likert scale items are more robust to scaling problems than dichotomous items, with moderate to good hit rates, and no excessive false positives. Fourth, in case of severe scale problems in the dichotomous case, increasing the number of items facilitates the accurate detection of GxE.

The severity of the problem of poor scaling of the sum score depends upon the item characteristics of the individual items. We showed that even a slight increase in the number of disproportionally easy items can result in spurious GxE, or mask genuine GxE. Thus, when focusing solely on sum scores without consideration of the item characteristics, it is not clear whether skewness is due to GxE or due to poor scaling. Thus, one should be cautious in applying the univariate GxE methodology (Molenaar et al. 2012; van der Sluis et al. 2006; Purcell 2002) to sum scores. If one can ensure that a given sum score is well scaled, results of the univariate method could in principle be interpreted in terms of genuine GxE effects. To investigate scaling of the sum score, an item-level analysis can serve to check the item properties. If the scaling is considered poor, the model proposed in this paper can be used to take the item characteristics explicitly into account. However, we stress that proper scaling of the measurement is still an important goal to pursue, as the results of present paper clearly illustrate that—even with an appropriate measurement model—poor scaling of the measurement can severely affect power to detect GxE.

We think that another interesting result from present paper is that it offers a solution to the problem of the univariate approach where power to detect AxC was found to be low (Molenaar et al. 2012). As here we used the same parameters as in the Molenaar et al. study, we are able to compare results. Due to the use of multiple items in present approach, power to detect AxC was appreciably larger, and even acceptable in case of 15 Likert items and mild scaling problems.[4] This suggests that researchers interested in detection of AxC should analyze item level data to ensure acceptable power.

In this paper we did not consider the possibility of curvilinear interactions. As discussed in van der Sluis et al. (2012), additional interaction parameters can be included (in addition to $\beta_1$ and $\gamma_1$) to make the variance of C and E a

[4] Because we used a Bayesian model fit approach, we omitted the term 'power' in discussion of the results and spoke of 'hit rate' instead. However, to be able to compare results of the frequentist results we take the 'hit rate' as found in present study as an indication for what the 'power' would have been when we would have implemented present model in a frequentist framework.

polynomial function of A. Here we did not consider this an option as the additional multinomial parameters are likely to require larger sample sizes than the ones considered here. However, when large datasets are available, the script in the Appendix can be easily extended to incorporate quadratic terms.

We think problems identified with the sum score pertain to both tests of genotype by unmeasured environment as discussed in this paper (see Jinks and Fulker 1970), and to the moderation approach (Purcell 2002). See Tucker-Drob et al. (2009) for an illustration of how the measured moderation approach could be confounded by measurement problems. Fortunately, the moderation approach has been extended to accommodate ordinal and binary data (see Medland et al. 2009). However, this extension concerns a univariate model, i.e., one item at the time. In case of multiple items, moderation of the parameters in a biometric model with an appropriate measurement model for the items can be considered following the procedure as outlined in this paper.

Some problems associated with the detection of GxE remain in present approach. That is, as GxE implies non-normality of the phenotypic scores, other sources of non-normality could produce artificial GxE (see Eaves 2006; van der Sluis et al. 2012). For instance, non-normality on the level of the latent variable can arise due to ability differentiation, the phenomenon that highly intelligent subjects display smaller individual differences than do less intelligent ones (Spearman 1927), which may cause spurious GxE as is illustrated in Tucker-Drob et al. (2009). In addition, non-normality in the phenotypic scores could arise due to unrepresentative sampling. For instance, if high intelligent subjects are overrepresented in a given sample, it is likely that spurious GxE results arise because of skewness in the latent variable due to sampling bias. However, as this problem is not explicitly investigated in present paper, it remains a topic for further investigation.

Finally, as models get more complex, the sample sizes needed to obtain reliable results also increase. Whereas the univariate GxE model may require as few as 50 MZ and 50 DZ twin pairs (see Molenaar et al. 2013), the present approach requires far larger samples.

Another drawback of present approach is that the method is computationally intensive. Particularly in case of Likert items, application of the model can take several hours to complete. For samples larger than studied here, this can be problematic. In a Molenaar et al. (2013), we have therefore used a two-stage procedure. In this procedure, an appropriate factor model is fitted to the data of the twin 1 and twin 2 members separately and factor scores are calculated. These factor scores are then submitted to the univariate approach by Molenaar et al. (2012). This method appeared to work well despite the data being highly skewed. Specifically, we found the same pattern of results using the model as outlined in present paper, and the two-stage procedure described above. However, as in the two-stage procedure the standard errors of the factor score estimates are neglected; it is unclear how this affects the results in a new application. We therefore strongly recommend to validate the results of the two-stage procedure in (a subset of) the data using the full model as described in present paper.

**Acknowledgment** Conor V. Dolan is supported by the European Research Council (Genetics of Mental Illness; Grant number: ERC-230374).

**Conflict of Interest** Dylan Molenaar and Conor V. Dolan declare that they have no conflict of interest.

**Human and Animal Rights and Informed Consent** All procedures followed were in accordance with the ethical standards of the responsible committee on human experimentation (institutional and national) and with the Helsinki Declaration of 1975, as revised in 2000 (5). Informed consent was obtained from all patients for being included in the study. As we analyse data that was administered in the National Survey of Midlife Development in the United States (MIDUS) in 1995–1996 under the auspices of the Inter-university Consortium for Political and Social Research (ICPSR), we kindly refer to this source for details concerning ethical procedures and informed consent.

## Appendix A: OpenBUGS code to fit the model in case of dichotomous item scores

```
model{

for(p in 1:Nmz){
        AMZ[p]~dnorm(0,pa2)                                                     # Eq 8
        c2MZ[p] <- exp(y0 + y1 * AMZ[p]/exp(.5*a2))                             # Eq 15
        e2MZ[p] <- exp(b0 + b1 * AMZ[p]/exp(.5*a2))                             # Eq 16
        precRESmz[p] <-1/(c2MZ[p]+e2MZ[p])                                      # inverse of Eq 13 as
                                                                                # precision is used by BUGS
        corMZ[p] <- c2MZ[p]  / (c2MZ[p] + e2MZ[p] )                             # Eq 12
        thetaMZ[p,1]~dnorm(AMZ[p], precRESmz[p])                                # Eq 9 for twin 1
        muMZ[p]<-AMZ[p]+corMZ[p]*(thetaMZ[p,1]-AMZ[p])                          # Eq 11
        prMZ[p]<-min(precRESmz[p]*1/(1-corMZ[p]*corMZ[p]),50)                   # inverse of Eq 14
        thetaMZ[p,2]~dnorm(muMZ[p],prMZ[p])                                     # Eq 9 for twin 2

        for(i in 1:nit){
                resMZ[p,i]~dnorm(0,1)

                for(k in 1:2 ){
                        probMZ[ p , i+(k-1)*nit] <-                             # Eq 7
                                phi((a[i]*thetaMZ[p,k]+rhoMZ[i]*resMZ[p,i]-b[i])/sqrt(1-rhoMZ[i]*rhoMZ[i]))
                        datMZ[p,i+(k-1)*nit] ~ dbern(probMZ[p,i+(k-1)*nit])     # Eq 6

                }
                }
}

for(p in 1:Ndz){

        ADZ[p,1]~dnorm(0,pa2)                                                   # Eq 8
        dummy1[p]~dnorm(0,pa2)                                                  # Eq 17
        ADZ[p,2]<-.5*ADZ[p,1]+sqrt(1-.5*.5)*dummy1[p]                           # Eq 18
        c2DZ[p,1] <- exp(y0 + y1 * ADZ[p,1]/exp(.5*a2))                         # Eq 27
        c2DZ[p,2] <- exp(y0 + y1 * ADZ[p,2]/exp(.5*a2))                         # Eq 28
        e2DZ[p,1] <- exp(b0 + b1 * ADZ[p,1]/exp(.5*a2))                         # Eq 29
        e2DZ[p,2] <- exp(b0 + b1 * ADZ[p,2]/exp(.5*a2))                         # Eq 30
        precRESdz[p,1] <-1/(c2DZ[p,1]+e2DZ[p,1])                                #inverse of Eq 25
        precRESdz[p,2] <-1/(c2DZ[p,2]+e2DZ[p,2])                                #inverse of Eq 25 for DZ
        corDZ[p] <-exp(y0+.5*y1/exp(.5*a2)*(ADZ[p,1]+ADZ[p,2])) /               # Eq 24
                (sqrt(c2DZ[p,1]+e2DZ[p,1])*sqrt(c2DZ[p,2]+e2DZ[p,2]))
        thetaDZ[p,1]~dnorm(ADZ[p,1], precRESdz[p,1])                            # Eq20
        muDZ[p]<-ADZ[p,2]+corDZ[p]*sqrt((c2DZ[p,2]+e2DZ[p,2])/(c2DZ[p,1]+e2DZ[p,1]))*(thetaDZ[p,1]-ADZ[p,1])
                                                                                # Eq 23
        prDZ[p]<-min(precRESdz[p,2]*1/(1-corDZ[p]*corDZ[p]),50)                 # Eq 26
        thetaDZ[p,2]~dnorm(muDZ[p],prDZ[p])                                     # Eq 21

        for(i in 1:nit){
                resDZ[p,i]~dnorm(0,1)

                for(k in 1:2 ){
                        probDZ[ p , i+(k-1)*nit] <- phi((a[i]*thetaDZ[p,k]+rhoDZ[i]*resDZ[p,i]-b[i])/sqrt(1-
                                rhoDZ[i]*rhoDZ[i]))                             # Eq 7
                        datDZ[p,i+(k-1)*nit] ~ dbern(probDZ[p,i+(k-1)*nit])     # Eq 6
                }
        }
}


for(i in 1:nit){
        rhoMZ[i]~dunif(0,1)
        rhoDZ[i]~dunif(0,1)
        b[i]~dunif(-5,5)
}
a[1] <-1

for(2 in 1:nit){
        a[i]~dunif(-5,5)
}
pa2<-1/exp(a2)
a2~dunif(-5,5)
y0~dunif(-5,5)
y1~dunif(-5,5)
b0~dunif(-5,5)
b1~dunif(-5,5)
}
```

## Appendix B: OpenBUGS code to fit the model in case of Likert item scores

```
model{

for(p in 1:Nmz){
        AMZ[p]~dnorm(0,pa2)
        c2MZ[p] <- exp(y0 + y1 * AMZ[p])
        e2MZ[p] <- exp(b0 + b1 * AMZ[p])
        precRESmz[p] <-1/(c2MZ[p]+e2MZ[p])
        corMZ[p] <- c2MZ[p]  / (c2MZ[p] + e2MZ[p] )
        thetaMZ[p,1]~dnorm(AMZ[p], precRESmz[p])
        muMZ[p]<-AMZ[p]+corMZ[p]*(thetaMZ[p,1]-AMZ[p])
        prMZ[p]<-min(precRESmz[p]*1/(1-corMZ[p]*corMZ[p]),50)
        thetaMZ[p,2]~dnorm(muMZ[p],prMZ[p])

        for(i in 1:nit){
                resMZ[p,i]~dnorm(0,1)

                for(k in 1:2 ){
                        for(j in 1:(ncat) ){
                                probMZ[ p , (ncat)*(i-1)+(j) +(k-1)*nit*ncat] <-
                                        phi((a[i]*thetaMZ[p,k]+rhoMZ[i]*resMZ[p,i]-b[j,i])/sqrt(1-rhoMZ[i]*rhoMZ[i]))
                                        –phi((a[i]*thetaMZ[p,k]+rhoMZ[i]*resMZ[p,i]-b[j+1,i])/sqrt(1-
                                        rhoMZ[i]*rhoMZ[i]))
                                }
                        datMZ[p,i+(k-1)*nit] ~
                        dcat(probMZ[p,((ncat)*(i-1)+1+(k-1)*nit*ncat): ((ncat)*(i-1)+ncat+(k-1)*nit*ncat)])
                }
        }
}

for(p in 1:Ndz){
        ADZ[p,1]~dnorm(0,pa2)
        dummy1[p]~dnorm(0,pa2)
        ADZ[p,2]<-.5*ADZ[p,1]+sqrt(1-.5*.5)*dummy1[p]
        c2DZ[p,1] <- exp(y0 + y1 * ADZ[p,1])
        c2DZ[p,2] <- exp(y0 + y1 * ADZ[p,2])
        e2DZ[p,1] <- exp(b0 + b1 * ADZ[p,1])
        e2DZ[p,2] <- exp(b0 + b1 * ADZ[p,2])
        precRESdz[p,1] <-1/(c2DZ[p,1]+e2DZ[p,1])
        precRESdz[p,2] <-1/(c2DZ[p,2]+e2DZ[p,2])
        corDZ[p] <- exp(y0+.5*y1*(ADZ[p,1]+ADZ[p,2])) / (sqrt(c2DZ[p,1]+e2DZ[p,1])*sqrt(c2DZ[p,2]+e2DZ[p,2]))
        thetaDZ[p,1]~dnorm(ADZ[p,1], precRESdz[p,1])
        muDZ[p]<-ADZ[p,2]+corDZ[p]*sqrt((c2DZ[p,2]+e2DZ[p,2])/(c2DZ[p,1]+e2DZ[p,1]))*(thetaDZ[p,1]-ADZ[p,1])
        prDZ[p]<-min(precRESdz[p,2]*1/(1-corDZ[p]*corDZ[p]),50)
        thetaDZ[p,2]~dnorm(muDZ[p],prDZ[p])

        for(i in 1:nit){
                resDZ[p,i]~dnorm(0,1)

                for(k in 1:2 ){
                        for(j in 1:(ncat) ){
                                probDZ[ p , (ncat)*(i-1)+(j) +(k-1)*nit*ncat] <-
                                        phi((a[i]*thetaDZ[p,k]+rhoDZ[i]*resDZ[p,i]-b[j,i])/sqrt(1-rhoDZ[i]*rhoDZ[i])) -
                                        phi((a[i]*thetaDZ[p,k]+rhoDZ[i]*resDZ[p,i]-b[j+1,i])/sqrt(1-rhoDZ[i]*rhoDZ[i]))

                                }
                        datDZ[p,i+(k-1)*nit] ~
                                dcat(probDZ[p,((ncat)*(i-1)+1+(k-1)*nit*ncat): ((ncat)*(i-1)+ncat+(k-1)*nit*ncat)])
                }
        }
}

for(i in 1:nit){
        rhoMZ[i]~dunif(0,1)
        rhoDZ[i]~dunif(0,1)

        for( j in 2:(ncat)){
                b[j,i]~dunif(b[j-1, i],b[j+1, i])
        }
        b[1,i]<--1/0
        b[(ncat+1),i]<-1/0
}

for(i in 2:nit){
        a[i]~dunif(-5,5)
}
a[1]<-1
pa2<-1/exp(a2)
a2~dunif(-5,5)
y0~dunif(-5,5)
y1~dunif(-5,5)
b0~dunif(-5,5)
b1~dunif(-5,5)
}
```

## References


Bauer DJ, Hussong AM (2009) Psychometric approaches for developing commensurate measures across independent studies: traditional and new models. Psychol Methods 14:101–125

Birnbaum A (1968) Some latent trait models and their use in inferring an examinee's ability. In: Lord EM, Novick MR (eds) Statistical theories of mental test scores, vol 17–20. Addison Wesley, Reading

Bock RD, Aitkin M (1981) Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. Psychometrika 46:443–459

Boker S, Neale MC, Maes HH, Wilde M, Spiegel M, Brick T et al (2010) OpenMx: an open source extended structural equation modeling framework. Psychometrika 76:306–317

Brim OG, Baltes PB, Bumpass LL, Cleary PD, Featherman DL, Hazzard WR et al (2010) National Survey of Midlife Development in the United States (MIDUS), 1995–1996. Inter-university Consortium for Political and Social Research, Ann Arbor, MI

Curtis SM (2010) BUGS code for item response theory. J Stat Softw 36(1):1–34

Dolan CV (1994) Factor analysis of variables with 2, 3, 5 and 7 response categories: a comparison of categorical variable estimators using simulated data. Br J Math Stat Psychol 47:309–326

Dolan CV, van den Berg SM (2008) Power analysis. In: Neale, et al. (eds) Statistical genetics: gene mapping through linkage and association. Taylor & Francis, London

Eaves LJ (1977) Inferring the causes of human variation. J R Stat Soc A 140:324–355

Eaves LJ (2006) Genotype x environment interaction in psychopathology: fact or artifact? Twin Res Human Genet 9:1–8

Eaves LJ, Erkanli A (2003) Markov chain Monte Carlo approaches to analysis of genetic and environmental components of human developmental change and GxE interaction. Behav Genet 33:279–299

Eaves LJ, Last K, Martin NG, Jinks JL (1977) A progressive approach to non-additivity and genotype-environmental covariance in the analysis of human differences. Br J Math Stat Psychol 30:1–42

Evans DM, Gillespie NA, Martin NG (2002) Biometrical genetics. Biol Psychol 1–2:33–51

Gelman A, Rubin DB (1992) Inference from iterative simulation using multiple sequences. Stat Sci 7:457–511

Geman S, Geman D (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Trans Pattern Anal Mach Intell 6:721–741

Hessen DJ, Dolan CV (2009) Heteroscedastic one-factor models and marginal maximum likelihood estimation. Br J Math Stat Psychol 62:57–77

Jinks JL, Fulker DW (1970) Comparison of the biometrical genetical, mava, and classical approaches to the analysis of human behavior. Psychol Bull 73:311–349

Johnson W, Krueger RF (2005) Genetic effects on physical health: lower at higher income levels. Behav Genet 35:579–590

Kamin L (1974) The sciencs and politics of IQ. Erlbaum, Potomac

Lau JYF, Eley TC (2008) Disentangling gene-environment correlations and interactions on adolescent depressive symptoms. J Child Psychol Psychiatry 49:142–150

Loftus GR (1978) On interpretation of interactions. Mem Cognit 6:312–319

Lunn DJ, Thomas A, Best N, Spiegelhalter D (2000) WinBUGS—a Bayesian modeling framework: concepts, structure, and extensibility. Stat Comput 10:325–337

Lunn D, Spiegelhalter D, Thomas A, Best N (2009) The BUGS project: evolution, critique, and future directions. Stat Med 28:3049–3067

Martin NG, Eaves LJ, Kearsey MJ, Mavies P (1978) The power of the classical twin design. Heredity 40:97–116

Mather K, Jinks JL (1971) Biometrical genetica: the study of continous variation. Chapman & Hall, London

Medland SE, Neale MC, Eaves LJ, Neale BM (2009) A note on the parameterization of Purcell's GxE model for ordinal and binary data. Behav Genet 39:220–229

Miles DR, Silberg JL, Pickens RW, Eaves LJ (2005) Familial influences on alcohol use in adolescent female twins: testing for genetic and environmental interactions. J Stud Alcohol 66:445–451

Molenaar PCM, Boomsma DI (1987) Application of nonlinear factor-analysis to genotype environment interaction. Behav Genet 17:71–80

Molenaar D, Borsboom D (2013) The formalization of fairness: issues in testing for measurement invariance using subtest scores. Educ Res Eval 19:223–244

Molenaar D, van der Sluis S, Boomsma DI, Dolan CV (2012) Detecting specific genotype by environment interactions using marginal maximum likelihood estimation in the classical twin design. Behav Genet 42:483–499

Molenaar D, van der Sluis S, Boomsma DI, Haworth CM, Hewitt JK, Martin NG et al (2013) Genotype by environment interactions in cognitive ability: a survey of 14 studies from four countries covering four age groups. Behav Genet 43:208–219

Mroczek DK, Kolarz CM (1998) The effect of age on positive and negative affect: a developmental perspective on happiness. J Pers Soc Psychol 75(5):1333–1349

Neale MC, Cardon LR (1992) Methodology for genetic studies of twins and families. Kluwer Academic, Dordrecht

Neale MC, Boker SM, Xie G, Maes HH (2006) Mx: statistical modeling, 7th edn. Department of Psychiatry, VCU, Richmond

Plummer M (2003) JAGS: a program for analysis of Bayesian graphical models using Gibbs sampling. Proceedings of the 3rd international workshop on distributed statistical computing (DSC 2003), pp 20–22

Plummer M, Best N, Cowles K, Vines K (2005) CODA: output analysis and diagnostics for MCMC. R package version 0.9–2

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

Rasch G (1960) Probabilistic models for some intelligence and attainment tests. Danmarks Paedogogiske Institut, Copenhagen

Rathouz PJ, van Hulle CA, Rodgers JL, Waldman ID, Lahey BB (2008) Specification, testing, and interpretation of gene-by-measured-environment models in the presence of gene environment correlation. Behav Genet 38:301–315

Samejima F (1969) Estimation of ability using a response pattern of graded scores. Psychometrika Monograph 17:253

Saris WE, Satorra A (1993) Power evaluations in structural equation models. In: Bollen KA, Long JS (eds) Testing structural equation models. Sage, Newbury Park, pp 181–204

SAS Institute (2011) SAS/STAT 9.3 user's guide. SAS Institute, Cary

Satorra A, Saris WE (1985) The power of the likelihood ratio test in covariance structure analysis. Psychometrika 50:83–90

Schwabe I, van den Berg SM (2014) Assessing genotype by environment interaction in case of heterogeneous measurement error. Behav Genet. doi:10.1007/s10519-014-9649-7

Spearman CE (1927) The abilities of man: their nature and measurement. Macmillan, New York

Tucker-Drob EM (2009) Differentiation of cognitive abilities across the life span. Dev Psychol 45:1097–1118

Tucker-Drob EM, Harden KP, Turkheimer E (2009) Combining nonlinear biometric and psychometric models of cognitive abilities. Behav Genet 39:461–471

Turkheimer E, Haley A, Waldorn M, D'Onofrio B, Gottesman II (2003) Socioeconomic status modifies heritability of IQ in young children. Psychol Sci 14:623–628

Turkheimer E, Harden KP, D'Onofrio B, Gottesman II (2009) The Scarr Rowe interaction between measured socioeconomic status and the heritability of cognitive ability. In: McCartney K, Weinberg RA (eds) Experience and development: a festschrift in honor of Sandra Wood Scarr. Psychology Press, New York, pp 81–97

van den Berg SM, Beem L, Boomsma DI (2006) Fitting genetic models using Markov chain Monte Carlo algorithms with BUGS. Twin Res Human Genet 9:334–342

van den Berg SM, Glas CA, Boomsma DI (2007) Variance decomposition using an IRT measurement model. Behav Genet 37:604–616

van der Sluis S, Dolan CV, Neale MC, Boomsma DI, Posthuma D (2006) Detecting genotype-environment interaction in monozygotic twin data: comparing the Jinks & Fulker test and a new test based on marginal maximum likelihood estimation. Twin Res Hum Genet 9(3):377–392

van der Sluis S, Posthuma D, Dolan CV (2012) A note on false positives and power in GxE modelling of twin data. Behav Genet 42:170–186

Wagenmakers E-J, Krypotos A-M, Criss AH, Iverson G (2012) On the interpretation of removable interactions: a survey of the field 33 years after Loftus. Mem Cognit 40:145–160

Wirth RJ, Edwards MC (2007) Item factor analysis: current approaches and future directions. Psychol Methods 12:58–79

Zand Scholten A (2011) Admissible statistics from a latent variable perspective. Unpublished doctoral dissertation. University of Amsterdam, Amsterdam