Behav Genet
DOI 10.1007/s10519-014-9649-7

ORIGINAL RESEARCH

Assessing Genotype by Environment Interaction in Case of Heterogeneous Measurement Error

Inga Schwabe · Stéphanie M. van den Berg

Received: 8 October 2013 / Accepted: 11 February 2014

Abstract Considerable effort has been devoted to establish genotype by environment interaction (G × E) in case of unmeasured genetic and environmental influences. Although it has been outlined by various authors that the appearance of G × E can be dependent on properties of the given measurement scale, a non-biased method to assess G × E is still lacking. We show that the incorporation of an explicit measurement model can remedy potential bias due to ceiling and floor effects. By means of a simulation study it is shown that the use of sum scores can lead to biased estimates whereas the proposed method is unbiased. The power of the suggested method is illustrated by means of a second simulation study with different sample sizes and G × E effect sizes.

Keywords Genotype by environment interaction · Heterogeneous measurement error · Item response theory · Sum scores · Twin studies

Introduction

Genotype by environment interaction (henceforth referred to as G × E) in its conceptual sense means either that different genotypes respond differently to the same environment or that some genotypes are more sensitive to changes in the environment than others (Cameron 1993; Martin 2000; Sorensen 2010). In the last decade, the assessment of G × E has received increasing attention in twin and family studies (Dick 2011). Various studies have found evidence for the presence of G × E. In the context of educational achievement, Friend et al. (2009) report an interaction between high reading ability and the education of the parents: The heritability of high reading ability was higher for twins when parents were less well educated. Another well-known finding is that heritability of cognitive ability varies with socioeconomic status (Turkheimer et al. 2003; Harden et al. 2006). G × E seems also present for non-cognitive traits . To name a few examples, G × E has been found in the development of depression (Hicks et al. 2009; Lau and Eley 2008; Brendgen et al. 2009), physical and mental health (Johnson and Krueger 2005; Faith et al. 2004; Kim-Cohen et al. 2006) and antisocial behavior (Caspi et al. 2002; Cadoret et al. 1983; Tuvblad et al. 2006). Arguably, G × E is an important phenomenon in complex behavioral traits.

Twin data can be used to investigate the interaction between genotypes and different environmental variables. Often, however, specific environmental variables are not directly measured. Therefore, methods to assess G × E in the case that both genes and environment feature as latent (i.e., unmeasured) variables are needed. A well-known method proposed by Jinks and Fulker (1970) uses data of monozygotic (MZ) twins. Letting T_1 and T_2 denote MZ twin scores, Jinks and Fulker (1970) showed that a correlation between the absolute difference between two twins within a pair ($|T_1 - T_2|$, .i.e., a proxy for variance due to environmental influences) and the sum score of a twin pair ($T_1 + T_2$, i.e., a proxy for variance due to genetic influences) suggests the presence of G × E.

van der Sluis et al. (2006) proposed an alternative method, using MZ twin data and an exponential function to

Edited by Gitta Lubke.

I. Schwabe (✉) · S. M. van den Berg
Department of Research Methodology, Measurement, and Data Analysis, University of Twente, Drienerlolaan 5, 7522 NB Enschede, The Netherlands
e-mail: i.schwabe@utwente.nl

Published online: 20 March 2014

model G × E (cf. SanChristobal-Gaudy et al. 1998). Molenaar et al. (2012) extended this work by including dizygotic (DZ) twin data and modeling G × E for both shared and non-shared environmental variance separately. Furthermore, they extended the univariate approach to a multivariate approach.

G × E in Case of Heterogeneous Measurement Error

There is however a problem in the assessment of G × E that is not tackled by any of the above mentioned methods. In a behavior genetics study, one is typically interested in the origins of observed variance in a phenotypic trait. To this end, often a number of items is presented to respondents. Next, the subject's sum score on the items is computed, assuming that the unweighted summed score can be treated as a proxy for the trait. The variance of the computed sum scores is then decomposed into a number of variance components. In a so-called AE model the variance is decomposed into parts due to additive genetic (A) and unique environmental (E) influences, whereas the so-called ACE decomposition also estimates variance due to common environmental (C) influences (Jinks and Fulker 1970).

However, variance decomposed as due to unique environmental influences does not only capture environmental influences but also measurement error (see e.g. Loehlin and Nichols 1976; Turkheimer and Waldron 2000). Moreover, the amount of information a test (i.e., a set of items) gives, varies for different levels of the phenotypic latent variable, so that measurement error variance is not homogeneous across the scale (see also Lord 1980; Embretson and Reise 2009). For example, while existing IQ tests usually show little measurement error variance for average students, scale scores for high performing students can be very unreliable because of little information provided by only a few very difficult items. Another example comes from clinical scales. If both affected and healthy individuals are assessed with for example a depression scale that contains many extreme items, scale scores may be very reliable for highly depressed participants but very unreliable for healthy controls. In extreme situations such as for high performing students and healthy controls, this leads to ceiling and floor effects, respectively. In case of a ceiling effect a large proportion of subjects receives the highest possible test score, whereas in case of a floor effect a large proportion of subjects receives the lowest possible test score (Lewis-Beck et al. 2004), leading to smaller individual differences at the lower (floor effect) or upper (ceiling effect) end of the measurement scale. This leads to a skewed sum score distributions, which in turn can result in the finding of spurious G × E.

Let us illustrate this with a simple example. Suppose one is interested in the genetic and environmental influences on high general cognitive ability (*g*). To this end, a psychometric cognitive test is administered to MZ and DZ twin pairs selected based on their high school performance. Following the method proposed by Jinks and Fulker (1970), the absolute differences between the test scores within MZ pairs are regressed on the sum of these scores to identify possible G × E. However, in case of a ceiling effect, the test is too easy for the most able twins and most of them will get the highest possible test score, resulting in smaller score differences within highly able twin pairs than within average or less able twin pairs. Twins with a higher sum score seem more alike. In other words, spurious G × E can be expected. In a variance decomposition this results in a lower proportion of variance explained by unique environmental influences for highly able twins than for average or low performing twins. Various authors have tried to draw attention to this potential bias. Eaves et al. (1977) were the first to outline issues and misconceptions surrounding genotype by environment interaction, among other issues stressing the sensitivity of G × E to properties of the measurement scale. This notion has been accentuated by various different authors since then (Martin 2000; van der Sluis et al. 2006; Eaves 2006; Molenaar et al. 2012).

With the increasing attention to G × E and various articles warning for spurious G × E due to scale effects, it is surprising that no method has been proposed yet that assesses G × E that deals with heterogeneous measurement error. Due to spurious G × E, one cannot rely on the validity of research findings concerning G × E. Replication of findings means little, because the same artifacts of a scale may apply to multiple studies. Likewise, a failure to replicate may imply nothing other than the use of a different scale of measurement (Eaves 2006). It is evident that there is the need for a method that can tackle the problem and assess G × E in case of heterogeneous measurement error without bias.

Towards a Solution

Heterogeneous measurement error can be accounted for by explicitly modelling the properties of a scale. This can be done by incorporating an Item Response Theory (IRT) measurement model into the variance decomposition. In IRT models, item scores depend not only on a person's trait level (e.g. intelligence), but also on the properties of the items that were administered (e.g. difficulty). van den Berg et al. (2007) extended the usual AE/ACE variance decomposition with an IRT measurement model. They showed that the simultaneous estimation of an IRT measurement model and a biometric model produced unbiased estimates for heritability coefficients and dominance genetic variance, unlike the sum score approach. Also the proposed method by Molenaar et al. (2012) incorporated a measurement model. They linked

observed item variables first to the underlying construct using a linear factor model and then (in the biometric part of the model) decomposed the phenotypic variances into parts due to additive genetic, common environmental and unique environmental influences. Heteroscedastic residual variances were incorporated in the measurement model to account for possible measurement problems at the level of the observed variables. This led to the absorption of possible floor and ceiling effects and poor scaling effects in the residuals, while the effects of actual genotype by environment interaction were detected in the latent biometrical part of the model. As a factor model was used, the approach is limited to continuous data and cannot be used for dichotomous items (e.g. scored as correct/false). This limitation can be overcome by the combination of an IRT measurement model and a biometric model.

Here, we propose a method that extends the van den Berg et al. (2007) model for dichotomous and polytomous data with a G × E interaction effect. Simulation Study 1 illustrates that the method is superior to the sum score approach, in that the sum score approach leads to spurious G × E, whereas parameter estimates are unbiased with the proposed method. The statistical power of the suggested method to detect actual G × E is illustrated with Simulation Study 2 using different G × E effect sizes and sample sizes.

Biometric Model

The so-called ACE model decomposes observed variance in a phenotypic variable, denoted as σ_P^2, into parts due to additive genetic influences (σ_A^2), common shared environmental influences (σ_C^2) and unique environmental influences (σ_E^2).

In case of G × E, part of the variance due to E varies systematically with additive genetic effect A. Therefore, the E variance component has to be portioned into an intercept (environmental variance when A = 0) and a part that is a function of A, resulting in a variance of σ_E^2 that is different for each individual j:

$$\sigma_{Ej}^2 = \exp(\beta_0 + \beta_1 \mathrm{A_j}) \quad (1)$$

where β_0 denotes the intercept and β_1 is a slope parameter that reflects G × E. G × E is modeled as a (log)linear effect, meaning that the non-shared environmental variance component is larger at either higher or lower levels of the genotype (e.g. larger individual differences). The direction of the effect depends on the sign of the slope parameter. The exponential function is used to avoid negative variances (see also SanChristobal-Gaudy et al. 1998; Bauer and Hussong 2009; Hessen and Dolan 2009; van der Sluis et al. 2006). To take into account the properties of the measurement scale, an IRT measurement model is integrated into the biometric model.

Measurement Model

Whereas in the sum score approach item difficulties are ignored, the IRT approach uses the difficulty of each item as information to be incorporated into the scaling of individual test performance. The probability for a correct answer on item k for individual j is then modelled as a function of the difference between the individual's latent trait score θ_j and the item difficulty parameter b_k. A well-known IRT model is the so called one-parameter logistic model (1PLM), also known as the Rasch model (Rasch 1960). In this model, the odds of passing an item, expressed as the ratio of the number of successes to the number of failures, is modelled using a natural logarithm function (Embretson and Reise 2009):

$$\ln(P_{jk}/(1 - P_{jk})) = \theta_j - b_k \quad (2)$$

The 1PLM is suitable for dichotomous data, as for example data collected from ability tests where item responses are commonly scored correct/false. In the 1PLM, all items are assumed to have the same correlation (factor loading) with the underlying latent trait. That is, all items discriminate equally well between the various levels of the latent trait. It is also possible to estimate factor loadings that differ across items (in the IRT framework referred to as discrimination parameters α_k), which turns the 1PLM into a two-parameter model (2PLM) (see e.g. Embretson and Reise 2009). Furthermore, there are several IRT models that are suitable for ordered categories, as for example Likert scale data (see e.g. Samejima 1970; Masters 1982; Embretson and Reise 2009). In this paper, the 1PLM was used, but extension to other models is straightforward. In case of the 2PLM model for example, the equation changes to:

$$\ln(P_{jk}/(1 - P_{jk})) = \alpha_k\ (\theta_j - b_k) \quad (3)$$

which results in only minor adaptations of the script (described in the next section) used in this article (see Appendix 1). In order to identify the scale, the discrimination parameter for the first item, α_1, can to be fixed to one. Extension to polytomous items is straightforward by applying the method illustrated by van den Berg et al. (2007).

Incorporation of the Measurement Model into the Biometric Model

van den Berg et al. (2007) showed that, in order to take full advantage of the IRT approach, both the IRT measurement model and the variance decomposition model have to be estimated simultaneously, using a so called one-step

approach. However, as this procedure is computationally burdensome, widespread methods of estimating variance components through structural equation modelling reach their computational limit. van den Berg et al. (2007) illustrated that Bayesian statistical modelling with Markov chain Monte Carlo (MCMC) estimation can be a good alternative. In a Bayesian analysis, statistical inference is based on the joint posterior density of the model parameters, which is proportional to the product of a prior probability and the likelihood function of the data (see e.g. Box and Tiao 1972). When analytically deriving the posterior distribution is difficult or impossible, Gibbs sampling (Geman and Geman 1984; Gelfand and Smith 1990; Gelman et al. 2004) can be applied. Here, the MCMC estimation was implemented in the freely obtainable MCMC software package JAGS (Plummer 2003). The JAGS script can be found in the Appendix 1. The script can also be used in the free software package WinBUGS (Lunn et al. 2000).

As in Eaves and Erkanli (2003) and van den Berg et al. (2006, 2007), a Bayesian parameterization of the ACE model was used that only uses univariate distributions. The model is presented for MZ and DZ twins separately.

MZ twins: For each MZ twin pair i a normally distributed common environmental effect was assumed that is the same for both twins:

$$C_i \sim N(\mu, \sigma_C^2) \tag{4}$$

where μ denotes the population mean. Under the assumption that MZ twins have identical genotypic values, the conditional distribution for familial effect F_i for each MZ pair i, given the common environmental effect C_i, is normal:

$$F_i \sim N(C_i, \sigma_A^2) \tag{5}$$

To arrive at the additive genetic effect, the common environmental effect has to be subtracted from F_i:

$$A_i = F_i - C_i \tag{6}$$

The ACE variance decomposition of the latent variable θ_{ij} is complete if we have for individual j of MZ pair i:

$$\theta_{ij} \sim N(F_i, \sigma_E^2) \tag{7}$$

To introduce G × E, the twin pair i specific error variance, σ_{Ei}^2, reflecting unique environmental influences, has to be portioned into an intercept and a scale parameter (see Eq. 1), resulting in a variance of σ_E^2 that is different for each twin pair i:

$$\sigma_{Ei}^2 = \exp(\beta_0 + \beta_1 A_i) \tag{8}$$

Simultaneous with the biometric model above, the latent phenotype θ_{ij} appears in the 1PL IRT model for observed item data Y (see Eq. 2):

$$\begin{aligned} \ln(\mathrm{P_{ijk}}/(1-\mathrm{P_{ijk}})) &= \theta_{\mathrm{ij}} - \mathrm{b_k} \\ Y_{ijk} &\sim Bernoulli(P_{ijk}) \end{aligned} \tag{9}$$

DZ twins: As for MZ twin pairs, a normally distributed common environmental effect is assumed that is the same for both twins (see Eq. 4). While the total genetic variance is the same for DZ and MZ twins, the genetic covariance in MZ twins is twice as large as in DZ twins, assuming random mating. To model a genetic correlation of 0.5 for DZ twins, first a normally distributed familial effect $\mathrm{F_0}$ is assumed with variance $\frac{1}{2}\sigma_A^2$ (cf. Jinks and Fulker 1970):

$$F_{0i} \sim N\left(C_i, \frac{1}{2}\sigma_A^2\right) \tag{10}$$

Then, for each individual twin j from DZ pair i a normally distributed effect F_1 is modeled that includes the Mendelian sampling term:

$$F_{1ij} \sim N\left(F_{0i}, \frac{1}{2}\sigma_A^2\right) \tag{11}$$

so that that F_{1ij} includes the effect of both common environmental and additive genetic influences. To obtain the additive genetic effect, the common environmental effect has to be subtracted from F_1:

$$A_{ij} = F_{1ij} - C_i \tag{12}$$

Similar to Eq. 7 for MZ twins, the ACE decomposition is complete with

$$\theta_{ij} \sim N(F_{1ij}, \sigma_E^2) \tag{13}$$

with the difference that the additive genetic effect is different for each twin. To incorporate G × E into the model, σ_E^2 has to be portioned into different parts, similar to Eq. 8 (MZ pairs). Doing so results in an estimate of σ_E^2 that is different for each individual twin:

$$\sigma_{Eij}^2 = \exp(\beta_0 + \beta_1 A_{ij}) \tag{14}$$

Again, simultaneous to the ACE decomposition the latent phenotype θ_{ij} appears in the 1PLM IRT part of the model (see Eq. 9).

Prior Distributions

With a Bayesian approach, prior distributions have to be made explicit. We use inverse gamma distributions for the additive genetic variance and the common environmental variance ($\sigma_A^2 \sim InvG(1,1)$, $\sigma_C^2 \sim InvG(1,.5)$). These distributions were chosen because they are both flexible and conjugate. In Bayesian probability theory, a prior is called conjugate when the probability distribution of the prior and the posterior distribution have similar forms (in this case the gamma distribution). This results in convenient

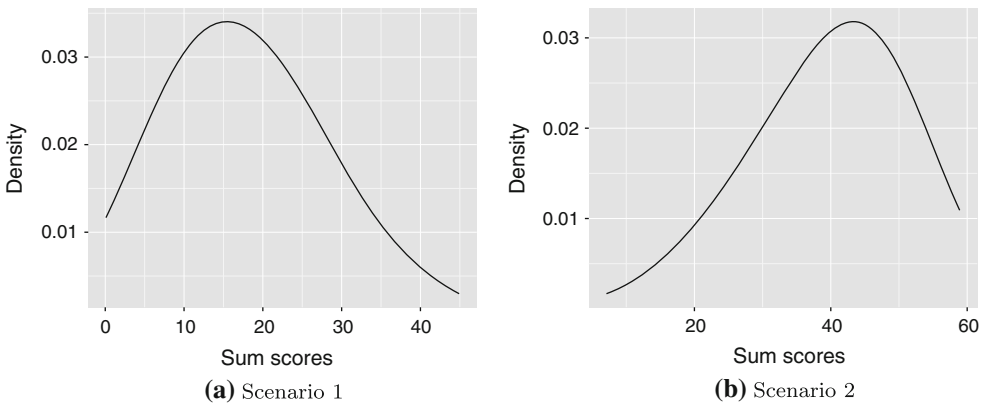
sampling, speeding up the estimation process. The prior for the intercept and the slope parameter can be assumed normal ($\beta_0 \sim N(0,1)$, $\beta_1 \sim N(0,10)$), resulting in relatively and reasonably flat priors in this particular application with IRT models. When item parameters are known, the population mean can be estimated, which can also be given a normal prior distribution ($\mu \sim N(0,10)$).

Simulation Study 1

To illustrate that the sum score approach can lead to the finding of spurious G × E whereas the proposed method is unbiased, a simulation study was conducted. One hundred datasets were generated consisting of 360 DZ twin pairs (72 % of total N) and 140 (28 % of total N) MZ twin pairs. This particular ratio was chosen as it approximately reflects the ratio of MZ and DZ twins in European twin registers. Additive genetic variance was assumed 0.5, common environmental variance was assumed 0.3 and unique environmental variance, $\exp(\beta_0)$, was set to 0.2. The data was simulated without any G × E ($\beta_1 = 0$) and a population mean of 0 for the twins ($\mu = 0$). The 1PLM was used to simulate responses to 60 dichotomous items resulting in a scale with a Cronbach's alpha of 0.90. The data was simulated under two different scenarios. In the first scenario, item parameters were simulated from a normal distribution with a mean of 1 and a standard deviation of 1 to mimic a test with relatively difficult items resulting in a slight floor effect for the distribution of sum scores. In the second scenario, item parameters were simulated from a normal distribution with a mean of -1 and a standard deviation of 1 to mimic a relatively easy test resulting in a slight ceiling effect. The first scenario resulted in a situation that is often encountered in psychopathology studies: a positively skewed sum score distribution. The second scenario resulted in a negatively skewed sum score distribution, a scenario that can be encountered in cognitive ability studies with gifted students. To give an idea of the severity of the skewness, the distributions of the simulated sum scores of all DZ twins are displayed in Fig. 1 for both scenarios. Furthermore, the three different methods for estimating skewness proposed by Joanes and Gill (1998) were used to determine non-normality of the distributions. In the first scenario, the different methods resulted in values in the range $[0.630; 0.632]$ and in the second scenario in the range $[-0.434; -0.435]$.

In both scenarios the item parameters were assumed known in the analysis as this is the case for many existing tests, such as educational tests and in computer-adaptive testing. The simulated data was analysed on the basis of the sum scores approach and on the basis of the suggested method. In the sum score approach, sum scores were calculated from the simulated item data and re-scaled so that they had a mean of 0 and variance 1. This was done to make results of both approaches comparable with respect to the prior distributions. For both approaches, the same prior was used for the population mean ($\mu \sim N(0,10)$). The data was then analyzed with the same JAGS script as in the appendix but without the IRT part.

The simulations were carried out using the software package R (R Development Core Team 2013). As an interface from R to JAGS, the rjags package was used (Plummer 2013). After a burn-in phase of 7,000 iterations, the characterisation of the posterior distribution for the model parameters was based on an additional 12,000 iterations from 1 Markov chain. This choice was based on previous test runs with multiple chains and computing Gelman and Rubin's convergence diagnostic (Gelman and Rubin 1992). All test runs with these numbers of iterations resulted in values < 1.02. The average posterior means of the model parameters for all replicated data sets were calculated, the standard deviation of posterior means, as were the means of all posterior standard deviations. The mean of the posterior standard deviations can be interpreted as the Bayesian analog of the standard error.



(a) Scenario 1 **(b)** Scenario 2

Fig. 1 Distribution of the sum scores of the DZ twins as simulated in Simulation Study 1

Simulation Study 2

A second simulation study was conducted to determine the sample size necessary to find G × E in twin data with the suggested method. As in the first simulation study, the simulated data consisted of DZ (72 % of total N) and MZ (28 % of total N) twin pairs. Additive genetic variance was assumed 0.5, the intercept, $\exp(\beta_0)$, was set to 0.2, the population mean to zero and common shared environmental variance was assumed 0.3. The magnitude of G × E, β_1, was varied. The 1PLM was used to simulate responses to 60 dichotomous items resulting in a scale with a Cronbach's alpha of 0.92. The item parameter values were simulated from a normal distribution with a mean of 0 and a standard deviation of 1 and assumed known in the analysis. To estimate the power to detect G × E, item data were simulated with different sample sizes ($N = 500, N = 1,000, N = 2,000$ twin pairs) and different values for β_1. Effect size of the G × E interaction was defined as the factor with which the environmental variance component increases for an individual with an additive genetic effect of $A_i = \sigma_A$ relative to β_0, and will be henceforth referred to as Δ. To illustrate this, consider an effect size of Δ = 1.1. The environmental variance for a person with an additive genetic effect equal to σ_A can then be computed as

$$\begin{aligned} \sigma^2_{Ei} &= \exp(\beta_0 + \beta_1 A_i) \\ \Delta &= \exp(\beta_1 \sigma_A) \\ \beta_1 \sigma_A &= \ln(\Delta) \\ \beta_1 &= \ln(\Delta)/\sigma_A \end{aligned} \tag{15}$$

resulting in 0.22 ($= 0.2 \times \exp(0.13 \times \sqrt{0.5})$). The slope parameter β_1 then has to be equal to ~ 0.13 ($= \ln(1.1)/\sqrt{0.5}$). With an effect size of Δ = 1.5, β_1 is equal to ~ 0.57 and the environmental variance at $A_i = \sigma_A$ is equal to $0.2 \times 1.5 = 0.3$.

Each condition was repeated 100 times with a different G × E effect size ($\Delta = 1.00, \Delta = 1.30, \Delta = 1.50$ and $\Delta = 1.70$). To estimate the power, the 95 % highest posterior density (HPD, see e.g. Box and Tiao 1972) interval was determined for each parameter. Power was defined as the percentage of simulations in which the 95 % HPD interval did not contain zero.

As in Simulation Study 1, the simulations were carried out using the software package R (R development core team 2013). After a burn-in phase of 7,000 iterations, the characterisation of the posterior distribution for the model parameters was based on 12,000 iterations from 1 Markov chain. The average posterior means of the model parameters for all replicated data sets were calculated, the standard deviation of posterior means, and the means of all posterior standard deviations.

Results

Simulation Study 1

The true parameter values, the average posterior means, and the mean of posterior standard deviations (averaged over 100 replications) are reported in Table 1 for the first scenario.

In the first scenario, a slight floor effect was mimicked, resulting in a positively skewed sum score distribution. It can be seen that the sum score analysis approach resulted in biased parameter estimates. Both genetic variance and common environmental variance were underestimated whereas the intercept (environmental variance when $A = 0$) was overestimated. The sum score approach resulted in an average slope parameter of $\beta_1 = 1.05$, reflecting an effect size of $\Delta = \exp(1.05 \times \sqrt{0.43}) \approx 2.00$.

In the second scenario, a slight ceiling effect was mimicked, resulting in a negatively skewed sum score distribution. Since the second scenario is the mirror image of the first scenario, the parameter estimates were the same but in the opposite direction ($\beta_1 = -1.08$). To save space, results of the second scenario are not tabulated.

Simulation Study 2

The power estimates for the slope parameter β_1 can be found in Table 2. All power estimates for σ^2_A, σ^2_C and $\exp(\beta_0)$ were equal to 1.00 in all conditions, and are therefore not tabulated. The true parameter values, the average posterior means and the average posterior standard deviations can be found in Table 3.

It can be seen that the estimated values are very close to the true values. The power to find G × E in the base-line scenario without any effect ($\Delta = 1.00$) is close to 5 % for $N = 1,000$ and $N = 2,000$. Under the simulated scenario, there is good power to detect an effect size of 1.7, even with only 500 twin pairs.

Table 1 Scenario 1: The average posterior means (SD) averaged over 100 replications

	True value	Sum scores	IRT
σ^2_A	0.50	0.43 (0.05)	0.48 (0.09)
		0.07	0.09
σ^2_C	0.30	0.22 (0.05)	0.32 (0.08)
		0.06	0.08
$\exp(\beta_0)$	0.20	0.26 (0.02)	0.20 (0.03)
		0.03	0.04
β_1	0.00	1.05 (0.15)	0.03 (0.27)
		0.18	0.28

Second line: Mean of posterior standard deviations

Table 2 Estimated power to find G × E for different sample sizes

	$\Delta = 1.00$ β_1	$\Delta = 1.30$ β_1	$\Delta = 1.50$ β_1	$\Delta = 1.70$ β_1
N = 500	0.03	0.48	0.57	0.81
N = 1,000	0.07	0.57	0.92	0.99
N = 2,000	0.07	0.92	1.00	1.00

N refers to the number of twin pairs

Discussion

The aim of this paper was twofold: To illustrate the spurious finding of G × E due to properties of the measurement instrument and to show that the incorporation of an explicit measurement model into the variance decomposition can remedy this potential bias.

In Simulation Study 1, two different scenarios were simulated, mimicking a floor and a ceiling effect. It was shown that the sum score approach in both cases leads to the spurious finding of G × E. This is in line with various publications stressing the sensitivity of G × E to scale properties (Eaves et al. 1977; Martin 2000; Eaves 2006; Molenaar et al. 2012).

Note that in case of a floor effect the sum approach resulted in positive spurious G × E, whereas a ceiling effect evoked negative spurious G × E. This intuitively makes sense. In case of a ceiling effect, a large number of twins get the highest possible test score, resulting in smaller intra-pair differences at the top of the measurement scale. It seems as if the twins at the top of the measurement scale are more similar than the rest of the sample. In the analysis, this is captured as spurious negative G × E: Proportion of variance explained by unique environmental influences decreases with increasing test score. In case of a floor effect, a large number of twins get the lowest possible test score. This results in the exact opposite effect.

In Simulation Study 1, only slight floor and ceiling effects were simulated, such as is often observed in real data. This shows that it is realistic to find spurious effects with the magnitude observed in the simulated data. These results imply that the G × E analysis based on sum scores is very sensitive to scaling issues. Note that the sum score approach does not result in bias when the distribution is not skewed. A simulation study was conducted to show this. One hundred datasets were generated under the same condition as in Simulation Study 1 but with a symmetric sum score distribution (i.e., an expectation of 0 and a standard deviation of 1 for the item parameters). This resulted in an unbiased average posterior mean for β_1 of 0.03 with a standard deviation of 0.24.

We chose to illustrate the finding of spurious G × E due to properties of the measurement scale by mimicking a

Table 3 The average posterior means (SD) averaged over 100 replications

	$\Delta = 1.00$				$\Delta = 1.30$				$\Delta = 1.50$				$\Delta = 1.70$			
	σ^2_A	σ^2_C	$\exp(\beta_0)$	β_1	σ^2_A	σ^2_C	$\exp(\beta_0)$	β_1	σ^2_A	σ^2_C	$\exp(\beta_0)$	β_1	σ^2_A	σ^2_C	$\exp(\beta_0)$	β_1
True value	0.50	0.30	0.20	0.00	0.50	0.30	0.20	0.37	0.50	0.30	0.20	0.57	0.50	0.30	0.20	0.75
N = 500	0.48	0.31	0.21	0.02	0.50	0.31	0.20	0.41	0.48	0.32	0.21	0.55	0.48	0.32	0.21	0.74
	(0.08)	(0.06)	(0.03)	(0.24)	(0.07)	(0.07)	(0.03)	(0.28)	(0.08)	(0.07)	(0.03)	(0.24)	(0.08)	(0.07)	(0.03)	(0.26)
	0.08	0.06	0.03	0.24	0.09	0.06	0.02	0.19	0.09	0.08	0.04	0.27	0.07	0.06	0.03	0.18
N = 1,000	0.49	0.31	0.20	0.02	0.48	0.30	0.21	0.37	0.48	0.32	0.20	0.58	0.50	0.30	0.20	0.74
	(0.07)	(0.06)	(0.02)	(0.19)	(0.06)	(0.05)	(0.02)	(0.18)	(0.07)	(0.06)	(0.03)	(0.15)	(0.07)	(0.05)	(0.03)	(0.16)
	0.07	0.06	0.02	0.18	0.07	0.06	0.02	0.18	0.07	0.06	0.02	0.19	0.07	0.06	0.03	0.18
N = 2,000	0.49	0.31	0.20	−0.01	0.50	0.30	0.20	0.38	0.50	0.30	0.20	0.55	0.50	0.30	0.20	0.75
	(0.05)	(0.05)	(0.02)	(0.13)	(0.05)	(0.04)	(0.02)	(0.11)	(0.07)	(0.05)	(0.03)	(0.16)	(0.05)	(0.05)	(0.02)	(0.14)
	0.05	0.04	0.02	0.12	0.05	0.04	0.02	0.12	0.07	0.06	0.02	0.18	0.05	0.04	0.02	0.13

Second line: Mean of posterior standard deviations. N refers to the number of twin pairs

floor and a ceiling effect. It is important to realize that the problem is however not limited to this situation. A floor or ceiling effect is only an extreme case of a test that does not measure different trait levels equally well. Spurious G × E can also be expected when no floor or ceiling effect has been detected in the data but the distribution is skewed. Although it is of course desirable to make tests more reliable (e.g. adding more difficult items to lower measurement error for highly able students), this does not solve the problem. In practice, tests that discriminate uniformly over the whole range of a trait (e.g. ability) simply do not exist (see also Eaves 1983). Constructing a test with reasonably homogeneous measurement error would involve making a test with a lot of easy items and a lot of difficult items, and no items in between. Such a test might perhaps not result in spurious G × E, but it does not provide a lot of information either, and is therefore not very attractive psychometrically.

Here we proposed to incorporate an explicit measurement model into the variance decomposition in order to remedy potential bias. Molenaar et al. (2012) used a different approach, proposing the incorporation of a linear factor model into variance decomposition. As a linear factor model assumes normally distributed residuals, the linear factor model is inappropriate for categorical variables in general and for binary variables in particular (Bartholomew 2008). Therefore, the method by Molenaar et al. (2012) is limited to continuous data and not suitable for dichotomous or polytomous items. As dichotomous items are often used in ability tests (scored as right/wrong), the incorporation of a measurement model suitable for dichotomous data is relevant for every research field that uses twin data and ability tests to assess G × E (e.g. research in giftedness or educational achievement). In addition, the incorporation of IRT models for polytomous items is straightforward (see e.g. Samejima 1970; Masters 1982; Embretson and Reise 2009). van den Berg et al. (2007) show how k polytomous items with m response categories can be transformed into $k \times (m-1)$ dummy items that can be used in a model for dichotomous items, so our method can also be applied to polytomous items without altering the JAGS script.

Simulation studies 1 and 2 showed that the proposed method does not find any spurious G × E and recovers the true values for the model parameter very well. In addition, Simulation Study 2 showed that the statistical power of the method is sufficient given that large samples are often available from twin registries. Only in case of a very small effect size, one needs 2,000 twin pairs to find G × E. Note that the simulated effect sizes are all smaller than the effect size of the spurious effect that was found when the sum score approach was used. As it is very common in behavior genetic studies to see data with a distribution as simulated in Simulation Study 1 (see Fig. 1), the power of the model seems to be good for G × E effects that can be observed in real data. The results of the power study however apply only to the simulated conditions. The power to detect G × E might be different for traits with a different etiology and studies with a different sample composition.

In all analyses, the item parameters were assumed known. This is the case in, for example, large-scale educational assessment situations (see e.g. Veldkamp and Paap 2013) and in computer adaptive testing (e.g. assessment of quality of life, see e.g. Reeve et al. 2007; Nikolaus et al. 2013). It is straightforward for alternative applications to estimate item parameters as well (see van den Berg et al. 2007). A reasonable approach would in most cases be to use independent standard normal distributions as priors for the difficulty parameters (e.g. $N(0, 10)$). With item parameters unknown, the population mean for the individuals is best fixed to 0, and this makes an expectation of 0 for the item parameters appropriate. A variance of 10 makes the prior relatively and reasonably flat. Of course, additionally estimating difficulty parameters will affect power, but only slightly. If the model is extended with varying factor loadings that need to be estimated (discrimination parameters), power will be affected more severely. Reasonable priors to use would be lognormal with expectation 0 and variance 10. The lognormal distribution constrains the discrimination parameters to be positive. Note that in order to fix the scale, one of the item discriminations should be fixed to 1. For more details, see van den Berg et al. (2007).

In this paper, we focused on variance decomposition in the case that environmental variables are unmeasured. The finding of spurious G × E due to scale properties is however not limited to this situation. Spurious G × E can also arise in case of measured environmental variables. In that situation, measurement error might not only appear at the level of the latent trait but also in the measurement of the environmental variables. Therefore, the method has to be extended to include measured environmental variables as well in future research. Simulation studies have to be conducted to ensure that the extended model is identified and does not result in bias.

This article furthermore focused on G × E only for unique environmental variance. We did not consider any interaction between genetic influences and common environmental influences (G × C) as in Molenaar et al. (2012). We feel doing both would be theoretically tricky as common environmental influences do not necessarily have to be different from unique environmental influences: the distinction is made to allow for the possibility that environmental influences are correlated in twins. How this correlation comes about is for many phenotypes still unknown. The reason that we focused here on the unique environmental influences is because these include all kinds

of measurement error and it is therefore particularly this component that can cause spurious findings related to scale.

Finally, in the present paper, G × E was modelled as a linear effect on the log scale. There is however also the possibility that G × E arises as curvilinear effect (as e.g. modeled by (van der Sluis et al. 2006; Molenaar et al. 2012). Whereas a linear effect on the log scale implies that the effect of the environment is stronger at either higher or lower levels of the genotype (e.g. greater intra-pair differences), a curvilinear effect allows for the possibility that the effect of the environment is stronger at both extreme levels of the genotypic values. In a third simulation study, the proposed model was extended with a curvilinear effect and the power of the model was estimated. Although the power of the model was satisfactory, there was a bias in the estimation of the curvilinear effect. Incorporation of a curvilinear effect seems more complicated and more research is needed to extend the suggested method to include a curvilinear effect as well.

A similar model as introduced in the present article has been proposed in a paper by Molenaar and Dolan (2014). That paper focuses on the same problem (spurious G × E due to scale properties) but was developed independently. A nice feature of the Molenaar and Dolan paper is the addition of additive genetic effects interacting with shared environmental influences and modelling of correlated residuals. In our view, a nice feature of our own implementation in JAGS is that the estimation time of the model is much faster and our parameter recovery is very good: estimates are very close to the true values. So, all in all, the present article and the article by Molenaar and Dolan should be regarded complementary.

Acknowledgments This study was funded by the PROO Grant 411-12-623 from the Netherlands Organisation for Scientific Research (NWO).

Conflict of Interest The authors declare that they have no conflict of interest.

Human and Animal Rights and Informed Consent This article does not contain any studies with human or animal subjects performed by any of the authors.

Appendix 1: JAGS script

JAGS script

```
#Following script incorporates ACE decomposition with
#an 1 PL IRT model

#Ydz = Item responses of DZ twins (matrix)
#Ymz = Item responses of MZ twins (matrix)
#nmz = Number of MZ twin pairs,
#ndz = Number of DZ twin pairs,
#n.items = Number of items administered
#b = Item parameters, assumed known in the analysis

#Required structure of the Ydz/Ymz data matrix:
#Ydz[i,k] = kth datapoint from the ith DZ twin pair
#Ymz[i,k] = kth datapoint from the ith MZ twin pair

#This results in a matrix of nmz (or, in
#case of Ydz, ndz) rows and 2*n.items columns

#e.g. Ymz[1,22] is the response of
#MZ twin 1 from family 1 to item 22
#and Ymz[1,23] is the response of
#MZ twin 2 from family 1 to item 1
#if n.items = 22

#When item parameters are unknown in the analysis,
#following code can be integrated in the script:

#for (i in 1:n.items){
#  b[i] ~ dnorm(0,.1)
#  }

#Mu then has to be set to zero to identify the scale

#JAGS uses precision parameters for the variance
#parameters. Therefore, after running the script,
#these precision parameters should be inverted.
#For example:
#VAR.C <-  1/outputAnalysis$tau.c[,,1]
#with the rjags package

model{

##MZ twins
for (fam in 1:nmz){
  c.mz[fam] ~ dnorm(mu, tau.c)
  f.mz[fam] ~ dnorm(c.mz[fam], tau.a)
  a.mz[fam] <- f.mz[fam] - c.mz[fam]
  tau.e[fam] <- 1/(exp(beta0 + (beta1*a.mz[fam])))

```

```
  for (twin in 1:2){
    pheno.mz[fam,twin]~ dnorm(f.mz[fam],tau.e[fam])
   }

#1pl model twin1
  for (k in 1:n.items){
    logit(p[fam,k]) <- pheno.mz[fam,1] - b[k]
    Ymz[fam,k] ~ dbern(p[fam,k])
   }

#1pl model twin2
  for (k in (n.items+1):(2*n.items)){
          logit(p[fam,k]) <- pheno.mz[fam,2] - b[k-n.items]
          Ymz[fam,k] ~ dbern(p[fam,k])
  }
}

##DZ twins
for (fam in 1:ndz){
   c.dz[fam] ~ dnorm(mu, tau.c)
   f0.dz[fam] ~ dnorm(c.dz[fam], doubletau.a)

  for (twin in 1:2){
          f.dz[fam,twin] ~ dnorm(f0.dz[fam], doubletau.a)
          a.dz[fam,twin] <- f.dz[fam,twin] - c.dz[fam]
          tau.e.dz[fam,twin] <- 1/(exp(beta0 +
                                  (beta1*a.dz[fam,twin])))
          pheno.dz[fam,twin] ~ dnorm(f.dz[fam,twin],
                                  tau.e.dz[fam,twin])
  }


#1pl model twin1 (DZ)
  for (k in 1:n.items){
   logit(p2[fam,k]) <- pheno.dz[fam,1] - b[k]
    Ydz[fam,k] ~ dbern(p2[fam,k])
  }

#1pl model twin2 (DZ)
  for (k in (n.items+1):(2*n.items)){
          logit(p2[fam,k]) <- pheno.dz[fam,2] - b[k-n.items]
          Ydz[fam,k] ~ dbern(p2[fam,k])
         }

}


#Priors
mu ~ dnorm(0,.1)
beta1 ~ dnorm(0,.1)
beta0 ~ dnorm(0,1)

doubletau.a <- 2*tau.a

tau.a ~ dgamma(1,1)
tau.c ~ dgamma(1,.5)
}
```

References

Bartholomew DJ, Steele F, Moustaki I, Galbraith J (2008) Analysis of multivariate social science data. Taylor Francis, New York

Bauer DJ, Hussong A (2009) Psychometric approaches for developing commensurate measures across independent studies: traditional and new models. Psychol Methods 14(2):101–125

Box GEP, Tiao GC (1972) Bayesian inference in statistical analysis. Wiley, New York

Brendgen M, Vitaro F, Boivin M, Girard A, Bukowski WM, Dionne G et al (2009) Gene-environment interplay between peer rejection and depressive behavior in children. J Child Psychol Psychiatr 50(8):1009–1017

Cadoret RJ, Cain CA, Crowe RR (1983) Evidence for gene-environment interaction in the development of adolescent antiscocial behavior. Behav Genet 13(3):301–310

Cameron ND (1993) Methodologies for estimation of genotype with environment interaction. Livest Prod Sci 35(3–4):237–249

Caspi A, McClay J, Moffitt TE, Mill J, Martin J, Craig IW et al (2002) Role of genotype in the cycle of violence in maltreated children. Science 297(5582):851–854

Dick DM (2011) Gene-environment interaction in psychological traits and disorders. Annu Rev Clin Psychol 7:383–409

Eaves LJ (1983) Errors of inference in the detection of major gene effects on psychological test scores. Am J Hum Genet 35(6):1179–1189

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

Eaves LJ, Erkanli A (2003) Markov chain monte carlo approaches to analysis of genetic and environmental change and g x e interaction. Behav Genet 33(3):279–299

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

Embretson SE, Reise SP (2009) Item response theory for psychologists. Psychology Press, Oxford, UK

Faith MS, Berkowitz RI, Stallings VA, Kerns J, Storey M, Stunkard AJ (2004) Parental feeding attitudes and styles and child body mass index: prospective analysis of gene-environment interaction. Pediatrics 114(4):e429–e436

Friend A, DeFries JC, Olson RK, Pennington B, Harlaar N, Byrne B et al (2009) Heritability of high reading ability and its interaction with parental education. Behav Genet 39(4):427–436

Gelfand AE, Smith AFM (1990) Sampling-based approaches to calculating marginal densities. J Am Stat Assoc 85(410):398–409

Gelman A, Carlin JB, Stern HS, Rubin DB (2004) Bayesian data analysis, 2nd edn. Chapman and Hall, London

Gelman A, Rubin DB (1992) Inference from iterative simulation using multiple sequences. Stat Sci 7(4):457–511

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

Harden KP, Turkheimer E, Loehlin JC (2006) Genotype by environment interaction in adolescent's cognitive aptitude. Behav Genet 37(2):273–283

Hessen DJ, Dolan CV (2009) Heteroscedastic one-factor models and marginal maximum likelihood estimation. Br J Math Stat Psychol 62(1):57–77

Hicks BM, DiRago AC, Iacono WG, McGue M (2009) Gene-environment interplay in internalizing discorders: consistent findings across six environmental risk factors. J Child Psychol Psychiatr 50(10):1309–1317

Jinks JL, Fulker DW (1970) Comparison of the biometrical genetical, mava, and classical approaches to the analysis of human behavior. Psychol Bull 73(5):311–349

Joanes DN, Gill CA (1998) Comparing measures of sample skewness and kurtosis. The Statistician 47:183–189

Johnson W, Krueger RF (2005) Higher perceived life control decreases genetic variance in physical health: evidence from a national twin study. Personal Soc Psychol 88(1):165–173

Kim-Cohen J, Caspi A, Taylor A, Williams B, Newcombe R, Craig IW et al (2006) Maoa, maltreatment, and gene-environment interaction predicting children's mental health: new evidence and a meta-analysis. Mol Psychiatr 11(10):903–913

Lau JY, Eley TC (2008) Disentangling gene environment correlations and interactions on adolescent depressive symptoms. J Child Psychol Psychiatr 49(2):142–150

Lewis-Beck MS, Bryman A, Liao TF (2004) The sage encyclopedia of social science research methods. SAGE Publications, Thousand Oaks

Loehlin JC, Nichols PL (1976) Heredity, environment, and personality: a study of 850 sets of twins. University of Texas Press, Austin

Lord FM (1980) Applications of item response theory to practical testing problems. Lawrence Erlbaum Associates, Hilsdale

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

Martin N (2000) Gene-environment interaction and twin studies. In: Spector T, Snieder H, MacGregor A (eds) Advances in twin and sib-pair analysis. Greenwich Medical Media, London, pp 143–150

Masters GN (1982) A rasch model for partial credit scoring. Psychometrika 47(2):149–174

Molenaar D, Dolan CV (2014) Testing systematic genotype by environment interactions using item level data. Behav Genet. doi:10.1007/s10519-014-9647-9

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

Nikolaus S, Bode C, Taal E, Oostveen JC, Glas CA, van de Laar MA (2013) Items and dimensions for the construction of a multidimensional computerized adaptive test to measure fatigue in patients with rheumatoid arthritis. J Clin Epidemiol 66(10):1175–1183

Plummer M (2003) JAGS: a program for analysis of bayesian graphical models using gibbs sampling. In: Proceedings of the 3rd international workshop on distributed statistical computing (DSC 2003), March 20–22, Vienna, Austria. ISSN 1609-395X

Plummer M (2013) rjags: Bayesian graphical models using mcmc. http://cran.r-project.org/package=rjags (R package version 3-10)

R Development Core Team (2013) R: a language and environment for statistical computing. Vienna, Austria. http://www.R-project.org (ISBN 3-900051-07-0)

Rasch G (1960) Probabilistic models for some intelligence and attainment tests. Danish Institute or Educational Research, Copenhagen

Reeve BB, Hays RD, Bjorner JB, Cook KF, Crane PK, Teresi JA et al (2007) Psychometric evaluation and calibration of healthrelated quality of life item banks: plans for the patient reported outcome measurement information system (promis). Med Care 45(5):22–31

Samejima F (1970) Estimation of latent ability using a response pattern of graded scores. Psychometrika 35(1):139

SanChristobal-Gaudy M, Elsen J, Bodin L, Chevalet C (1998) Prediciton of the response to a selection for canalisation of a continous trait in animal breeding. Genet Sel Evol 30:423–451

Sorensen D (2010) The genetics of environmental variation. In: Proceedings of the 9th world congress on genetics applied to livestock. Leipzig, Germany

Turkheimer E, Haley A, Waldron M, D'Onofrio B, Gottesman II (2003) Socioeconomic status modifies heritability of iq in young children. Psychol Sci 14(6):623–628

Turkheimer E, Waldron M (2000) Nonshared environment: a theoretical, methodological, and quantitative review. Psychol Bull 126(1):78–108

Tuvblad C, Grann M, Lichtenstein P (2006) Heritability for adolescent antisocial behavior differs with socioeconomic status: gene-environment interaction. J Child Psychol Psychiatr 47(7): 734–743

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

van den Berg SM, Glas CAW, Boomsma DI (2007) Variance decomposition using an IRT measurement model. Behav Genet 37(4):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 and Fulker test and a new test based on Marginal Maximum Likelihood estimation. Twin Res Hum Genet 9(3):377–392

Veldkamp BP, Paap MCS (2013) Robust automated test assembly for testlet based tests: an illustration with the analytical reasoning section of the lsat (lsac research report, rr 13–02) (Technical Report). Law School Admission Council, Newtown