Behav Genet (2007) 37:604–616
DOI 10.1007/s10519-007-9156-1

ORIGINAL PAPER

Variance Decomposition Using an IRT Measurement Model

**Stéphanie M. van den Berg · Cees A. W. Glas ·
Dorret I. Boomsma**

Received: 23 August 2005 / Accepted: 12 April 2007 / Published online: 30 May 2007

Abstract Large scale research projects in behaviour genetics and genetic epidemiology are often based on questionnaire or interview data. Typically, a number of items is presented to a number of subjects, the subjects' sum scores on the items are computed, and the variance of sum scores is decomposed into a number of variance components. This paper discusses several disadvantages of the approach of analysing sum scores, such as the attenuation of correlations amongst sum scores due to their unreliability. It is shown that the framework of Item Response Theory (IRT) offers a solution to most of these problems. We argue that an IRT approach in combination with Markov chain Monte Carlo (MCMC) estimation provides a flexible and efficient framework for modelling behavioural phenotypes. Next, we use data simulation to illustrate the potentially huge bias in estimating variance components on the basis of sum scores. We then apply the IRT approach with an analysis of attention problems in young adult twins where the variance decomposition model is extended with an IRT measurement model. We show that when estimating an IRT measurement model and a variance decomposition model simultaneously, the estimate for the heritability of attention problems increases from 40% (based on sum scores) to 73%.

Edited by Stacey Cherny

S. M. van den Berg (✉) · D. I. Boomsma
Department of Biological Psychology, Vrije Universiteit
Amsterdam, Van der Boechorststraat 1, Amsterdam 1081 BT,
The Netherlands
e-mail: SM.van.den.Berg@psy.vu.nl

C. A. W. Glas
Department of Research Methodology, Measurement, and Data
Analysis, University of Twente, Enschede, The Netherlands

Keywords Item response theory · MCMC · Bayesian statistics · Measurement · Attention problems · Sum scores

Introduction

In quantitative genetics, one is interested in the extent to which variation in certain characteristics is heritable. Heritability is expressed in terms of the proportion of the variance of a trait in a population that can be attributed to genetic differences. This genetic variance component can be estimated in, for example, the classical twin design (Boomsma et al. 2002a) in which the covariance structures of monozygotic and dizygotic twins are compared.

However, it is not always straightforward to estimate variance components. A variance component is only meaningful when measures are expressed on a scale of at least interval level. Moreover, many statistical methods require the phenotype to be normally distributed. Many phenotypes are not expressed in clearly defined units and are at best ordinal in character (e.g., conservatism, extraversion). Some traits have even only a nominal character (e.g., psychiatric disorders). There are several ways of dealing with such nominal data. One possibility is to focus on concordance rates and compute recurrence risk ratios (Risch 1990, 2001). Alternatively, one might assume a latent continuous trait with a threshold above which individuals are affected and estimate the heritability on that latent trait (Lynch and Walsh 1998; Falconer 1965; Crittenden 1961). This method can also be used with ordinal data.

For some traits, it is convenient to have multiple indicators (items). For example one might have for a particular disease 10 symptoms that each can be scored as absent (0) or present (1). For each individual one can then compute a sum score that indicates to what extent the individual is affected by the disease. Such sum scores usually show a normal distribution or do so after an appropriate transformation. It is typically assumed that the normally distributed scores or transformations thereof reflect a continuous interval scale and the variance of the sum scores is subsequently decomposed. This approach follows classical test theory (CTT) where it is assumed that the observed score (the sum score) is the aggregate of a true score and a random component, usually referred to as measurement error. When decomposing the variance of sum scores, the measurement error variance (the unreliability) ends up as part of the non-shared environmental variance. As a result, when the reliability of a scale is low (i.e., the measurement error is large) and the analysis is based on sum scores, the heritability of the actual trait is significantly underestimated.

Modelling sum scores is appropriate if the sum scores are highly reliable (for instance because they are based on a large number of correlated items) and well validated. Furthermore, there should be enough variation and the distribution should be more or less normal. Finally, there should be no data missing. If these requirements do not hold, item response theory (IRT) provides a well-established alternative to classical test theory. This paper introduces the basics of the IRT framework, after which its advantages over a sum score approach are discussed. Next, it is argued that IRT models should be estimated simultaneously with the variance decomposition model, which can be done using a Bayesian approach with Markov-chain Monte Carlo estimation. Lastly, a simulation study shows the potential bias when estimating variance components on the basis of sum scores and the Bayesian method is illustrated with an empirical data set on attention problems.

Item response theory models

In IRT models—as opposed to CTT—the influence of the items and the respondents are explicitly modelled by distinct sets of parameters. In these models, an assumed continuous latent variable θ reflects the trait and every item is identified by thresholds β where a response in one category becomes more likely than a response in an adjacent category. It is usually assumed that the latent variables θ_j are drawn from a normal distribution, that is, θ_j are independently and identically distributed $N(\mu, \sigma^2)$, though this assumption is not always necessary to identify the model parameters. The probability of the presence of the symptom i in individual j, $p(Y_{ij} = 1)$, is a function of the difference between the individual's trait score θ_j and the parameter β_i, with β_i indicating the location on the scale where the presence of a symptom becomes more probable than its absence. In the case of multiple symptoms, we have

$$p(Y_{ij} = 1) = \Phi(\theta_j - \beta_i), \tag{1}$$

with $\Phi(.)$ denoting the cumulative standard normal distribution function. That is, the probability of the presence of symptom i in person j is a function of both a person's liability score θ_j and a symptom (or item) parameter β_i. In the IRT framework, this model is referred to as the one-parameter normal ogive model, or 1PNO (Lawley 1943; Lord 1952, 1953). This model is identified with a location restriction, for example, $\mu = 0$. The variance of the latent trait, σ^2, can be estimated and can be interpreted as the covariance of the items: the larger the variance, the higher the reliability of the scale.

An alternative parameterisation replaces the normal ogive by a logistic curve, that is,

$$p(Y_{ij} = 1) = \Psi(\theta_j - \beta_i), \tag{2}$$

where

$$\Psi(x) = \frac{\exp(x)}{1 + \exp(x)}.$$

This version of the model is known as the one-parameter logistic model (1PLM), or Rasch model (Rasch 1960). To illustrate the model, consider an individual with a score θ_j of 1 on the latent trait, and a particular item with parameter $\beta = 1$. Then the probability of a positive response from this individual on this item equals exp(1 – 1)/(1 + exp(1 – 1)) = exp(0)/(1 + exp(0)) = 1/2 = 50%. An individual with a score higher than 1 has a higher probability of showing a positive response, whereas an individual scoring lower than 1 has a lower probability. Individuals with a latent score of –1 have a probability of exp(–2)/(1 + exp(–2)) = 12%. With a simple multiplicative transformation of the scale, the logistic and normal ogive curves are very similar and indistinguishable for all practical work (see, for instance, Lord 1980).

In the Rasch model, as well as in the 1PNO model, all items have the same correlation (''factor loading'') with the underlying latent trait. Analogous to factor models, it is possible to estimate factor loadings that differ across items. In the IRT framework these factor loadings are referred to as discrimination parameters α_i. These parameters indicate the extent to which an item i loads onto the latent trait, and the model becomes

$$p(Y_{ij} = 1|\theta_j, \alpha_i, \beta_i) = \Psi(\alpha_i\theta_j - \beta_i). \tag{3}$$

An alternative form in the literature replaces $\alpha\theta - \beta$ with $\alpha(\theta - \beta)$. This leads to a somewhat different interpretation of the β-parameters (they are scaled differently) but it only involves a reparameterisation.

Essentially, a one-parameter model can be described by a two-parameter model where all α parameters are equal. In order to identify the model and estimate α, however, the variance of the latent trait should be fixed. Thus, a one-parameter model with a large variance of the latent trait is equivalent to a two-parameter model with large discrimination parameter α that is equal for all items together with a fixed variance of the latent trait.

The two-parameter model must be identified by both a location and a scale restriction. The former can be the same restriction as above, that is, $\mu = 0$. The latter can be the additional restriction that the variance of the latent distribution is equal to one, that is, the model is identified by assuming a standard normal distribution, $N(0,1)$, for the latent ability parameters θ_j. Alternatively one fixes one of the discrimination parameters to unity. Generally, however, this identification solution is not advisable, because the standard errors of the parameters blow up if the discrimination parameter chosen for the identification is poorly identified.

IRT models for polytomous data

Often, measurement is based on items or symptoms with more than two categories. For example, answers can be coded as 0 (not at all), 1 (somewhat, sometimes) and 2 (a lot, often). Typically in CTT approaches in behaviour genetics the sum of these item scores is regarded to represent a person's score on the trait of interest and is used for the statistical inference.

There are several IRT models for ordered categories (e.g., Samejima 1969; Masters 1982). These have different rationales and are not reparameterisations of each other, but the practical implications for preferring one over the other are often negligible. Here we describe a continuation-ratio model (Tutz 1990; Verhelst et al. 1997). This model allows the transformation of a polytomous item into a set of dichotomous items, which facilitates model estimation. The response to a polytomous item is viewed as a set of responses to an ordered sequence of virtual dichotomous items: it is assumed that the respondent is administered virtual items until an incorrect or negative response is given. So, in this approach, an item with M categories labelled $m = 0,..., M - 1$, the response is dummy-coded into $M - 1$ dichotomous quasi-items. As an example, for an item with $m = 3$ categories we make two new virtual items. A score of 2 would be coded as correct responses to both virtual items. A score of 1 on the original item would be coded as a correct response to the first virtual item and an incorrect response to the second virtual item. A score of 0 would be coded as an incorrect response to the first virtual item and the second virtual item would be coded as not administered (missing). Now the responses to all virtual items can be modelled by an IRT model for dichotomous items, such as the models given by Eqs. 1, 2 or 3 and can be estimated by any IRT software package that can handle dichotomous items in combination with missing data. There are also IRT packages that estimate models for polytomous items directly (e.g., Multilog; Thissen et al. 2003).

Advantages of using an IRT framework compared to analysing sum scores

We will discuss four advantages of using IRT: (1) it supports construct validity and the scoring rule (e.g., a scoring rule might consist of taking the unweighted sum of symptoms as an estimate of a person's liability), (2) it supports the use of incomplete item administration designs and handling of missing data, (3) it supports accounting for measurement error, and (4) it can handle floor and ceiling effects.

An IRT framework allows one to explicitly model the relationship between item scores and the phenotype of interest. Any combination of items can of course be summed (weighted or unweighted), but this does not guarantee that the sum score reflects a meaningful construct. The meaningfulness of the measurement can be directly assessed in an IRT framework. Fit to an IRT model is empirical evidence that the observed responses can be explained by an underlying structure. The latent variable of the IRT model should, of course, be an appropriate representation of the construct to be measured.

The IRT model that fits the data determines the score rule of the measurement instrument. If, for instance, a one-parameter model does not fit the data, but a two-parameter model does, the sum score where the items scores are weighted with their respective discrimination parameters is a sufficient statistic for θ_j (Lord and Novick 1968). So some items can be more important or sensitive indicators of a trait than others. Modelling the item data in a variance decomposition analysis allows the separate evaluation of model fit regarding the measurement model and the variance decomposition model.

In addition, group differences can be modelled, through differences in means, variances and variance components, and through differences in the way symptoms relate to the latent trait. For instance, one or more symptoms may show a higher incidence rate in one group (indicated by a difference in β-parameters across groups, e.g., females and males), or be a more sensitive indicator for the trait in a particular group (indicated by a difference in α-parameters across groups). Such violations of measurement invariance are usually referred to as differential item functioning (DIF).

A practical advantage of the analysis of data using an IRT framework is the use of incomplete item administration designs and handling of missing data. In some situations, intentionally incomplete item administration designs can greatly improve the efficiency of data collection. With an IRT approach one can also effectively deal with problems specific to longitudinal research where items differ across waves. When using IRT models in a maximum likelihood or a Bayesian framework, it is easy to include individuals that have missing data on one or more items if the data are missing at random (Little and Rubin 1987). When data are not missing at random, the non-randomness can be modelled within an IRT framework by expanding the model with an IRT model that describes the pattern of the missing data (see, for instance, Moustaki and Knott 2000; Moustaki and O'Muircheartaigh 2000; Holman and Glas 2005). The encompassing framework for handling missing data using IRT offers an important advantage over classical test theory. In classical test theory sum scores are only meaningful if the items are the same in all individuals and at all measurement waves.

The third advantage of the analysis of data using an IRT framework is that it accounts for measurement error. Unreliability suppresses the correlation between measurements (attenuation). Particularly when using a scale with only a few items, the correlations amongst sum score variables may be grossly attenuated. Clearly, this has important implications for the estimation of variance components in genetic research. In an IRT framework, the problem can be solved by, instead of focussing on sum scores, considering the correlations between *latent* variables (see, for instance, Béguin and Glas 2001; Fox and Glas 2003). These so-called latent correlations can be seen as estimates of correlations corrected for attenuation. A simulation study and an application of IRT to real data in a later section will show the possible extent of such attenuation effects on the estimation of heritability.

The fourth advantage of IRT has to do with floor and ceiling effects. A problem of analysing sum scores that represent indices of psychopathology is that these scores show a skewed distribution in the general population (Van den Oord et al. 2003; Derks et al. 2004). These skewed distributions result from the fact that many behavioural phenotypes are assessed using questions that relate to symptoms that are relatively rare in the population. These distributional violations may have important implications for the inference regarding relative variance components when analysing sum scores (Derks et al. 2004). In an IRT framework one is essentially free to specify the distribution of the latent trait (in some cases, it can even be estimated). In most cases, with polygenic traits, a normal distribution seems the most reasonable alternative (a mixture approach may be more suitable for traits with only a few large QTL effects). When in turn the variance of the normally distributed latent trait is decomposed into genetic and non-genetic variance, the inference is unbiased if the assumptions of the model are correct.

Variance decomposition: the one-step and the two-step approach

In IRT models, the latent scores θ_j are typically assumed to be random draws from a normal distribution. When we are interested in the extent to which individual differences on the latent trait are heritable, we only need to decompose the variance of the θ_js using, for example, the classical twin design. There are two approaches. The first approach is to first estimate the parameters of the IRT model using standard IRT software (such as, Bilog, Multilog, Parscale, Testfact, ConQuest, OPLM), and then to have the same software estimate each individual score on the latent trait. Next, one uses these estimates of the θ_js as observed values in a standard variance decomposition analysis. This we call the two-step approach.

There are several disadvantages to this two-step approach. First of all, in the IRT model fitting phase, the usual IRT estimation software cannot handle the dependency in the data inherent in twin and family designs. In some cases, with simple designs such as with sibling pairs only, weighting of the data would come a long way in solving this problem, but with more complex family designs, weighting is not a satisfactory solution.

Second, when estimating latent scores for each individual, the estimates of the θ_js, just like sum scores in the CTT tradition, are not simply observations but estimates with error variance. When computing the confidence intervals for the heritability estimates in the second phase, this uncertainty on the latent scores is not taken into account and the heritability confidence intervals are consequently too narrow and the estimates biased downwards. Moreover, in an IRT framework, the confidence intervals for estimates of individual latent scores are dependent on their location on the scale (actually, the number of items with β-parameters that are similar in magnitude to the person score θ and the items' discriminatory power, α), whereas in the variance decomposition, it is assumed that measurement error (as included in the non-shared environmental variance component) is independent of location (cf. CTT). For example, many psychopathology scales have only items that refer to relatively rare symptoms. As a consequence, many individuals in the general population score 0, which does not necessarily imply that all actually have the trait to the exact same degree. In other words, the scale provides very little information on the trait on the low end. In contrast, the upper end of the scale usually shows more variation, which may imply that the measures are more reliable (more items that discriminate between individuals). Thus, a priori it seems likely that psychopathological scales have more discriminatory power at the upper end of the scale than at the lower end. Of course, this is not a bad thing, since these scales were designed to discriminate between the healthy and the sick. Therefore it seems reasonable to forego the assumption of equal reliability across the scale and take differing reliabilities into account.

Actually, using the two-step approach the heritability coefficient estimate will be about the same as when the analysis is carried out on sum scores. This is because IRT estimates and sum scores correlate highly, well over 0.90 in the case of two-parameter models. When applying a one-parameter model, the correlation will be practically one, because a basic assumption of the Rasch model is that a sum score is a sufficient statistic for the score on the latent trait. Therefore, all persons with the same sum score will get the same estimate on the latent trait. Thus, a third problem of the two-step approach is that it neither solves the attenuation problem, nor the non-normality, nor the ceiling effects.

In order to take full advantage of the IRT approach, it is critical to estimate both the measurement model and the variance decomposition model simultaneously, using a one-step approach. However, computationally this is rather challenging. Below, it is shown how this can be done using software for Bayesian estimation procedures. In an application in a later section, we demonstrate the one-step approach for the estimation of heritability with both simulated and empirical data.

Bayesian estimation using a Markov chain Monte Carlo algorithm

In twin studies, a widespread method of estimating variance components is through structural equation modelling (SEM). For continuous traits with normal distributions, this is a flexible approach in that it is able to accommodate all linear models and allows for testing of equality of means, variances, covariances and variance components across subpopulations. However, with more elaborate models with discrete or categorical observed variables, SEM maximum likelihood (ML) estimation or ML procedures for estimating generalised linear mixed models such as GLAMM (Rabe-Hesketh and Skrondal 2005) soon reach computational boundaries. An alternative method is Bayesian statistical modelling with Markov chain Monte Carlo (MCMC) estimation algorithms (see also Eaves et al. 2005).

In the Bayesian approach, inference is based on the posterior density of the model parameters, $P(\eta|Y)$, where η represents the vector of model parameters and Y the observed data. By Bayes' rule, the density $P(\eta|Y)$ is proportional to the product of the likelihood of the data given the model parameters $P(Y|\eta)$ and the marginal density for η, $P(\eta)$, that is,

$$P\ (\eta \mid Y)\ \propto\ P(Y \mid \eta)P(\eta). \tag{4}$$

The marginal distribution of η is termed the *prior* distribution (prior in the sense of before the data have been taken into account), and must be specified by the user. The model provides us with the likelihood function $P(Y|\eta)$, and hence the *posterior* distribution of η is determined (posterior in the sense of after the data have been taken into account). The posterior distribution is a description of the probabilities of possible values for η given the observed data and forms the basis for statistical inference. We may, for example, take the mean or the median of this distribution as our point estimate for η. Further, the interval between the 2.5th and the 97.5th percentile of the posterior distribution provides the so-called central 95% credibility region, which is analogous to a 95% confidence interval in the ML framework. For more on Bayesian statistics, the reader is referred to the introductions by Box and Tiao (1973) and Gelman et al. (2004).

Sometimes it is easy to compute the posterior distribution analytically, but very often this is not possible. One can then use computer simulation to draw a sample of η-values from the posterior distribution. The mean or median of the posterior distribution can then be approximated by the mean or median of the sampled η-values, and approximate credibility regions can be determined in a similar way. In practice, the joint posterior distribution of all model parameters is usually quite complicated. Therefore, the complete set of parameters is split up into a number of subsets in such a way that the conditional posterior distribution of each subset given all other parameters has a tractable form and can be easily sampled from. This approach is known as Gibbs sampling (Geman and Geman 1984; Gelfand et al. 1990; Gelman et al. 2004), which is a special case of an MCMC algorithm. When however the conditional posterior distribution of a subset of the parameters is not easy or even impossible to sample from directly, other MCMC algorithms can be used, where one samples from a similar proposal distribution and uses a decision rule to either accept or reject a sample so that the accepted values can be regarded drawings from the target distribution.

In each iteration of an MCMC algorithm, a sample is taken from each conditional posterior distribution for each subset of the parameter space, given the current values of the other parameters. After a number of so-called ''burn-in'' iterations, necessary for a chain to achieve stationarity (i.e., approaching the target distribution: the joint posterior distribution) sufficiently closely, the subsequent draws can be regarded as sampled from the joint posterior distribution.

The application of the Bayesian approach with MCMC sampling to IRT models is mainly motivated by the fact that IRT models with complex dependency structures require the evaluation of multiple integrals to solve the estimation equations in a likelihood-based framework. This problem is avoided in an MCMC framework. In recent years, the fully Bayesian approach has been adopted to the estimation of IRT models with multiple raters, multiple item types, missing data (Patz and Junker 1999a, b), testlet structures (Bradlow et al. 1999, Wainer et al. 2000), latent classes (Hoijtink and Molenaar 1997), models with a multilevel structure on the ability parameters (Fox and Glas 2001, 2003) and the item parameters (Janssen et al. 2000), and multidimensional IRT models (Béguin and Glas 2001). In behaviour genetics, the approach has been taken up by Eaves and his co-workers (Eaves et al, 2005; Eaves et al. 2004).

In IRT research, the Gibbs sampler is used in two versions: a version with a normal ogive representation such as in Eq. 1, introduced by Albert (1992), and a version with a logistic representation introduced by Patz and Junker (1999a). Below, a logistic version will be used for simulated and real data, implemented in the freely obtainable MCMC software package WinBUGS (http://www.mrc-bsu.cam.ac.uk/bugs/).

Genetic models may be specified in WinBUGS as follows. Under the assumption that an ACE variance decomposition model (additive genetic, shared environmental and non-shared environmental effects) is appropriate for a latent trait θ, the model can be parameterised as a linear random effects model (see also Van den Berg et al. 2006a):

$$\theta_{jk} = a1_k + a2_{jk} + c_k + e_{jk}, \tag{5}$$

where c_k denotes the environmental effect for being a member of family k, and e_{jk} denotes the environmental effect of being individual j in family k. The genetic component is split into $a1$ and $a2$ to model the different genetic correlations amongst monozygotic (MZ) and dizygotic (DZ) twins (cf. Jinks and Fulker 1970). The genetic correlation in MZ twins is usually assumed 1.0 and in DZ twins 0.5, in other words, the genetic covariance in MZ twins is twice as large as in DZ twins. Therefore, if we let the random effect $a1$ be constant within all families and we let $a2$ vary within families only for DZ twins (but be constant for MZ twins), and then fix the variances of $a1$ and $a2$ to be equal, the genetic covariance in MZ twins will be twice as large as in DZ twins. The variance of $a1$ and $a2$ together, VAR($a1$) + VAR($a2$) = 2 * VAR($a1$) can then be interpreted as the variance due to additive genetic effects. We assume that $a1 \sim N(0, ½\ \sigma^2{}_a)$, $a2 \sim N(0, ½\ \sigma^2{}_a)$, $c \sim N(0, \sigma^2{}_c)$, and $e \sim N(0, \sigma^2{}_e)$.

The case for the ADE model can be derived similarly (Van den Berg et al. 2006a). For some estimation problems, it might be computationally more convenient to model sum and differences scores, instead of the latent scores for the twins separately (Van den Berg et al. 2006a; Robert and Casella 2004, p. 396; cf. Boomsma and Molenaar 1986).

Simulation

To illustrate the effect of attenuation on heritability estimates, 101 datasets were generated consisting of 400 MZ twin pairs and 600 DZ twin pairs. A standard normally distributed latent trait was simulated with an additive genetic component of 72% and a non-shared environmental component of 28%. The 1PL IRT model was used to simulate responses to 14 dichotomous items, where the β parameter values ranged from 0.5 to 3.5, with increments of 0.25. This corresponds to questionnaire items that are rarely endorsed by people. The simulated item data were fitted using a model with additive genetic and non-shared environmental effects (AE model) on a latent trait and a 1PL measurement model.

Next, sum scores were computed and these were analysed with an AE model. Since the distribution of the sum scores is positively skewed, the AE analysis was also performed after a logarithmic transformation of the sum scores.

The simulations were carried out using the software package R. For each replicated data set, we computed the twin correlations for the latent scores, the twin correlations of the sum scores and the twin correlations for the log-transformed sum scores. The three types of analyses were carried out in WinBUGS. After a burn-in phase of 1000 iterations, the characterisation of the posterior distribution for the model parameters was based on 1000 iterations from 2 independent Markov chains. From each of the 3 (analyses) * 101 (replicated data sets) marginal posterior distributions for the heritability we took the mean and the median as point estimates.

Further simulations were carried out to illustrate the attenuation effect and the bias in variance components. For simple genetic models, the twin correlations are sufficient statistics for the variance decomposition. Therefore it is enough to show how correlations based on sum scores behave as a function of number of items and beta parameters. Data were simulated using bivariate normally distributed latent values, with correlations 0.9, 0.7, 0.5, 0.3 and 0.1. These latent values were used to simulate corresponding sum scores using a one-parameter logistic IRT measurement model under a variety of conditions. First of all, we used different degrees of discrimination of the items (i.e., the variance of the latent trait: 0.676, 1 and 100). Second, we varied the way in which the items are distributed across the scale, either evenly scattered so that sum score distributions are symmetrical, or only scattered on the upper half part of the scale, that is, using only items that less than 50% of the population endorses, which results in positively skewed sum score distributions (cf. Derks et al. 2004; van den Oord et al. 2003). Third, we varied the number of items (5, 10, 20, 50, 100) to investigate attenuation.

Simulation results

Taking the median parameter values from the 101 data sets, the simulated latent data correlated 0.72 in MZ twins and 0.36 in DZ twins, just as would be expected. The sum scores correlated 0.45 in MZ twins and 0.21 in DZ twins (medians of the 101 data sets) and the log-transformed sum scores correlated 0.41 and 0.20, respectively. Thus, twin correlations are severely attenuated when analysing sum scores, even with 14 items.

Analysing the simulated item data with a 1PL IRT model, using the one-step approach, we recovered the true 72% value for the heritability coefficient closely (see Table 1). When analysing the raw sum scores using a normal AE model, either with or without transformation, the heritability point estimate dropped considerably, to about 42%. Thus, when the true model is an IRT model and the number of items is limited, an analysis of raw or transformed sum scores can lead to extensive underestimation of heritability.

For each condition of latent correlation, number of items, and variance of the latent variable, we simulated 100,000 twin pairs and correlated their sum scores. Figure 1A shows the result for the condition where the variance was 1 and the items were nicely scattered across the distribution of the latent values, between –2½ and 2½ times the standard deviation (1). The attenuation effect is clearly dependent on the number of items: with 100 items, the correlation on the basis of the sum scores is very close to the true correlations. An analysis treating the sum scores as bivariately normal and applying a variance decomposition will approximate the true proportions. Moreover, the degree of the attenuation is proportional to the true correlation: with 5 items, a true correlation of 0.9 will be attenuated to a correlation of 0.55 (61%) and a true correlation of 0.1 will be attenuated to a correlation of 0.06 (60%). Therefore, when the analysis on 5 items is based on the sum score, and the true MZ correlation equals twice the DZ correlation, this ratio is maintained when analysing sum scores. Thus, when applying an AE model, heritability will be underestimated, but no artifactual shared environmental effects or dominance genetic effects will appear as a result of analysing sum scores.

Table 1 Simulation results. Reported heritability values are the medians of the 101 posterior means and medians, standard deviations between parentheses

Method of analysis	Heritability coefficient point estimates	
	Posterior mean	Posterior median
1PL IRT model	0.7232 (0.0585)	0.7245 (0.0589)
Sum scores continuous model	0.4364 (0.0393)	0.4369 (0.0395)
Log-transformed sum scores	0.4046 (0.0403)	0.4047 (0.0406)

Figure 1B shows the result for a scale with slightly worse discrimination: the variance of the trait is now only 0.767. The items are again nicely scattered, between –2½ SD (–2.05) and 2½ SD (2.05). Thus, we retain the spread of the β values in terms of the SD, so that the expected proportion of individuals scoring a particular number of items remains equal across simulation situation; the resultant distribution of the sum scores will be equal. But now, due to the decreased sensitivity of the scale, the number of items has a more pronounced effect on the attenuation. The sum score correlations are now lower than under the model with variance = 1. However, the attenuation effect is still proportional to the true correlations.

Figure 1C shows an extreme situation where the items have high discriminatory power. The variance is now 100, and the items are evenly scattered between –25 and 25. Note that again, we retain the scatter of the beta values in terms of the SD, and again the sum score distribution will not be different from the earlier simulations. However, with such a sensitive scale, practically everybody that scores less than 1 SD below the mean will show a sum score of 16% of the total number of items. Everybody with a latent score higher than 1 SD below the mean will show a sum score of 84% of the number of items. Moreover, the data will show a scalogram pattern, for example with 3 items with increasing difficulty, the only observed patterns will be 111, 110, 100 and 000. Such a pattern will not be observed when the variance is 1, and even less so with a variance of 0.767: more individuals will then show patterns like 101 and 011, etc. Again, attenuation occurs when the number of items is limited, but the effect is much less pronounced, and again the attenuation is proportional across the different correlations. In this situation, an analysis of sum scores will yield reasonable estimates for the variance components given a sufficient number of items.

Actually, when the raw item data follow the scalogram pattern, the true correlations and the corresponding variance components will be recovered when applying a threshold model (Lynch and Walsh 1998). This is also true when the data follow a scalogram pattern but the items are

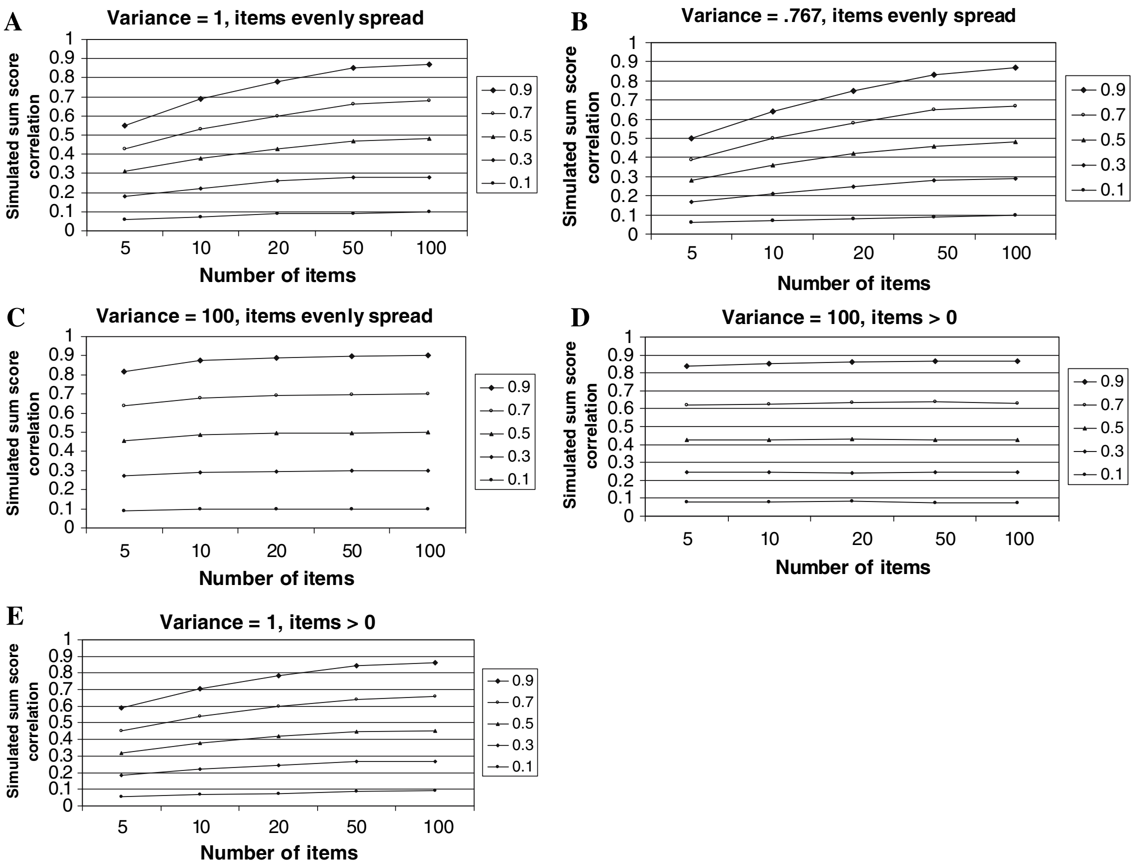

Fig 1 Correlations of simulated sum scores as a function of true correlation at the latent level, variance of the latent trait (quality of the scale), and scatter of the item β parameters (entire scale or only top half, i.e., all > 0)

not evenly scattered across the scale and the sum score distribution is skewed: applying a threshold model will recover the true correlations (cf. Derks et al. 2004). However, when applying an ordinary variance component analysis, ignoring its non-normality will yield biased estimates, underestimating the effects of shared environment and overestimating the effects of dominance (cf. Derks et al. 2004). This because when the items are not evenly scattered and the sum score distribution is skewed, the attenuation effect is no longer proportional to the true correlations (Fig. 1D): small correlations are more severely attenuated than large correlations. In case the true DZ correlation equals half the true MZ correlation, DZ:MZ = 1:2, the correlations of the sum scores will show a smaller ratio, DZ:MZ < 1:2, usually an indication of dominance genetic effects or epistasis. This is hard to see from the Fig. 1D, but with 5 items, the simulated sum score correlation is 0.83667 when the true correlation is 0.9 (92.96%), 0.42429 when the true correlation is 0.5 (84.86%), and 0.076575 when the true correlation is 0.1 (76.58%). Suppose we could analyse the true correlations, 0.9 and 0.5. One would then conclude that additive genetic variance accounts for 80% of the variance, non-shared environmental effects 10% and the shared environmental effects for the remaining 10%. Now if we would base our analysis on the observed sum score correlations 0.84 and 0.42, we would conclude that there are no shared environmental effects. One can imagine that when the true correlations are 0.90 and 0.45 one would conclude dominance effects to be absent, whereas if one would analyse observed sum scores correlations, one would find evidence for dominance genetic variance, the extent of which is dependent on the number of items.

Now, scalogram pattern data that fit a Guttman scale model are extremely rare. More often, item data follow a pattern that can be explained by the more lenient IRT model. Figure 1E shows the attenuation effect when the true model is a one-parameter IRT model with variance 1, where all items are endorsed by fewer than half the participants (i.e., all β parameters larger than the average latent score). Again we see that under the usual IRT model, the attenuation effect depends on the number of items and again we see that due to the skewness of the sum score distribution, the attenuation is not proportional to the true correlation. For example, with five items the simulated sum score correlation equals 0.591256 for true correlation 0.9 (66%), 0.316222 for true correlation 0.5 (63%), and 0.056567 for true correlation 0.1 (57%). When true correlations are again 0.9 and 0.5, the most likely model would be, when based on an analysis of the sum scores with 5 items, 5% dominance genetic variance, 61% additive genetic variance and 34% non-shared environmental variance.

Thus, also under the IRT model, analysing sum scores leads to an underestimation of shared environmental effects and an overestimation of dominance genetic effects when the sum score distributions are skewed.

An application

We illustrate the decomposition of variance using an IRT measurement model with data from the Netherlands Twin Registry (NTR; Boomsma et al. 2002b). Attention problems were measured with the Young Adult Self-Report (YASR; Achenbach 1997). We used data collected in the year 2000 from 460 males and 966 females from MZ twin pairs, 288 males from DZ same-sex twin pairs, 561 females from DZ same-sex twin pairs, and 305 males and 441 females from opposite sex twin pairs. All twins were between 18 and 30 years (inclusive). All available data were used, including data from incomplete pairs and individuals with several items missing. It was assumed that data were missing at random (cf. Van den Berg et al. 2006c).

The attention problems (AP) subscale of the YASR consists of seven items (see Table 2) with three ordered response categories (0 = Not true, 1 = Somewhat or Sometimes True, 2 = Very True or Often True). In children, sum scores typically show a high heritability with a significant non-additive genetic component (Rietveld et al. 2004). In young adults, AP sum scores also showed heritability (40%), but no non-additive genetic component (Van den Berg et al. 2006c).

Here, we estimate A and E variance components using a 1PL measurement model. A main effect of sex, δ, was modelled on the latent trais. The seven original items with three response categories were transformed into 14 dichotomous dummy items for each individual as described above. A separate β-parameter was estimated for each dummy item, so that for each original item there are two β-parameters. For the variance components, locally non-informative (''flat'') inverse gamma priors were used, and for the β and δ parameters we used locally non-informative normal priors. The parameterisation modelled the variances of sum and differences scores for the latent trait (Van den Berg et al. 2006a). The appendix gives the WinBUGS script. Three independent MCMC chains were used with randomised starting values. The chains converged rapidly to the stationary distribution with relatively low autocorrelations. The first 1000 iterations were discarded as burn-in samples, and a further 1000 iterations were used for inference.

Table 2 Items of the attention problems subscale of the young adult self-report (YASR; Achenbach 1997)

Item	Description
1	I act too young for my age
2	I have trouble concentrating or paying attention
3	I daydream a lot
4	My school work or job performance is poor
5	I am too dependent on others
6	I fail to finish things I should do
7	My behaviour is irresponsible

Table 3 Descriptives of marginal posterior distributions for the AE variance decomposition model using the 1PL IRT model for polytomous items with a main effect for sex

Parameter	Mean	SD	2½th percentile	Median	97½th percentile
σ^2_a	0.84	0.07	0.71	0.84	0.99
σ^2_e	0.32	0.06	0.20	0.32	0.44
δ	–0.13	0.05	–0.24	–0.13	–0.02
β_{11}	0.25	0.05	0.15	0.25	0.34
β_{12}	2.76	0.10	2.56	2.76	2.96
β_{21}	–0.76	0.05	–0.86	–0.76	–0.66
β_{22}	2.44	0.08	2.30	2.45	2.60
β_{31}	–0.43	0.05	–0.53	–0.43	–0.33
β_{32}	1.84	0.08	1.71	1.84	1.98
β_{41}	1.90	0.06	1.78	1.90	2.02
β_{42}	3.96	0.20	3.58	3.96	4.36
β_{51}	0.22	0.05	0.13	0.22	0.32
β_{52}	3.03	0.10	2.83	3.02	3.23
β_{61}	0.62	0.05	0.53	0.62	0.73
β_{62}	3.90	0.15	3.63	3.90	4.19
β_{71}	2.57	0.07	2.44	2.57	2.71
β_{72}	4.61	0.31	4.05	4.60	5.27
h^2	0.73	0.05	0.63	0.72	0.82

Note: First index of the betas refers to the item (see Table 1) and the second to the threshold

Results

Table 3 gives the descriptives of the marginal posterior distribution of the parameter values. The estimate for heritability based on the mean of the posterior distribution is 73%. The main effect of sex on the latent trait, with females scoring higher than males, is just significant, as zero is not included in the central 95% credibility region. Values of the β-parameters are all around zero or positive, indicating that the AP scale is most sensitive for individuals with considerable attention problems but has a hard time discriminating individuals with relatively few problems with attention. This results in the severely skewed distributions of sum scores.

The estimate for the heritability (73%) is much larger than the one reported earlier based on sum scores (40%, Van den Berg et al. 2006c). In the current sample, twin correlations for sum scores are very much like those reported earlier (MZ:0.45, DZ:0.17). By applying an IRT measurement model the twin correlation estimates for the latent trait are much higher, 0.76 for MZ twins and 0.30 for DZ twins. For comparison, when using a two-step approach, first estimating IRT model parameters in Multilog and then estimating latent scores for each individual (correlation between sum score and IRT estimate: 0.98), the results showed twin correlations nearly identical to those based on sum scores.

The 1PL IRT measurement model could easily be extended to include discrimination parameters (''factor loadings''). It is most convenient to constrain these to be positive through the specification of lognormal priors where for instance $\alpha = \exp(\gamma)$ and $\gamma \sim N(0, 100)$. In this case, the heritability estimate was not affected by this extension of the model (results not shown).

Discussion

We have compared an IRT model with a sum score approach with indirectly measured phenotypes. Under a range of conditions, the IRT framework is to be preferred over using sum scores. For example, in longitudinal studies with data missing by design or changing measurement instruments, when some items in a questionnaire change across birth cohorts or across different ages or when item data are missing, a sum score approach may no longer be appropriate, but in many cases the analysis can still be meaningfully carried out in an IRT framework using parameter expansion (see, for instance, Glas 1998).

When a simple IRT model does not fit the data, one could consider deleting or changing bad fitting items, and/or deleting bad fitting persons. Alternatively, one could consider using more general IRT models that offer many possibilities of obtaining model fit. General frameworks for multi-level and multi-dimensional IRT models are outlined in Skrondal and Rabe-Hesketh (2004) and De Boeck and Wilson (2004). In the specific context of genetic modelling, it might also occur that a particular subset of items show relatively high genetic correlations compared to the remaining items. In that case a more appropriate model would be an independent pathway model for categorical or ordinal traits (see for instance Van den Berg et al. 2006b).

Good fit to a one-dimensional IRT model is empirical evidence that the observed item responses can be explained by one continuous underlying trait. When it further can be concluded that the scale is meaningful (based on item analysis and association with external measures to assess its validity), and the assumption of measurement invariance across different subpopulations is tenable (Lubke et al. 2004), the approach effectively deals with non-normal distributions of sum scores in for instance psychopathology (Van den Oord et al. 2003).

Moreover, when the measurement model and the variance decomposition model are estimated simultaneously, the variance decomposition deals appropriately with the dependency in the data when estimating IRT model parameters and testing the model's assumptions, and the IRT measurement model deals appropriately with the estimation of the heritability coefficient (correcting for attenuation to obtain an unbiased point estimate) and the reporting of the confidence intervals (correcting for location-dependent uncertainty of person scores on the latent trait).

Our simulations showed the dramatic extent of the attenuation effect and the bias in estimating variance components due to imperfect measurement. Particularly when sum score distributions are skewed, underestimation of shared environmental effects and overestimation of dominance genetic effects may occur. The bias in variance components was also illustrated with an empirical data set: instead of finding a heritability estimate of 40% for attention problems with a sum score (Van den Berg et al. 2006c), a heritability estimate of 73% was obtained when including a measurement model and estimating it simultaneously with the variance decomposition model. This example provides an additional illustration of the bias in variance components due to the analysis of sum scores. However, it should be noted that model fit was not assessed, nor was the assumption of measurement invariance tested. This requires further study.

The crucial element of the one-step approach that leads to unbiased point estimates is the inclusion of the appropriate probabilistic measurement model so that the estimation takes into account the unreliability of the measurement. The probabilistic modelling allows for the fact that twins with identical response patterns may have different scores on the latent trait, and also, that twins with non-identical response patterns may have exactly the same score on the latent trait. Discriminatory power of the items and the number of items are both crucial to the heritability estimated based on sum scores: the fewer the items and the worse the discrimination of the items (i.e., the smaller the variance of the latent trait in the one-parameter model; the smaller the factor loadings in the two-parameter model), the more biased the estimation will be when the analysis is performed on sum scores. High quality scales with a large number of items (say, more than 50) with high discriminatory power that are scattered across the entire scale can indeed be analysed with sum scores, but any other scale should be analysed using the IRT framework if one is interested in an unbiased heritability estimate with trustworthy confidence intervals.

Future work should focus on the assessment of model fit in the context of genetic models. It is only sensible to apply a one-step IRT approach when the data actually conform to an IRT measurement model. If data do not fit an IRT model, for instance when there is differential item functioning across subpopulations, the approach will still lead to biased estimates. A crucial first step therefore is assessing model fit and checking measurement invariance.

Acknowledgments This work was supported by grants from the Netherlands Organisation for Scientific Research, NWO 575-25-006 and NWO 051-02-060, 668-772.

Appendix: WinBUGS script

AE decomposition with 1PL IRT measurement model
and main effect of sex

```
  # The model is parameterised using the sum of latent
  # liabilities in a twin pair and the
  # difference.
  # Nfammz: number of MZ twin pairs, Nfamdz: number
  # of DZ twin pairs, specified in the
  # data matrix
  # Ymz[i, k]: the kth datapoint from the ith MZ twin pair;
  # Ymz[i, 1] is a covariate that is not used in this analysis
  # Ymz[i, 2] is a dummy code for sex of the first twin
  # (1 = male), the next 14 data points relate # to the items
  # for the first twin, Ymz[i, 17] is a dummy code for sex of
  # second twin, the last
  # 14 datapoints relate to the second twin.
  # Ydz[i, k]: the kth datapoint from the ith DZ twin pair;
  # Winbugs uses precision parameters instead of variance
  # parameters for the distributions.
  # tau.summz: the inverse of sigma2.summz, being the
  # variance of the summed latent scores
  # of MZ twin pairs.


Model
   {
   for(i in 1:Nfammz)
      {
      Summz[i] ~ dnorm(0, tau.summz)
      difmz[i] ~ dnorm(0, tau.difmz)
         for( k in 3:16)
         {
         logit(p1[i,k]) <- (summz[i] + difmz[i])/
  2 + Ymz[i,2]*delta – beta[k – 2]
         Ymz[i,k] ~ dbern(p1[i,k])
         }
         for(k in 18:31)
         {
         logit(p1[i,k]) <- (summz[i] – difmz[i])/
  2 + Ymz[i,17]*delta – beta[k – 17]
         Ymz[i,k] ~ dbern(p1[i,k])
         }
      }
   for(i in 1:Nfamdz)
      {
      sumdz[i] ~ dnorm(0, tau.sumdz)
      difdz[i] ~ dnorm(0, tau.difdz)
         for(k in 3:16)
         {
         logit(p2[i,k]) <- (sumdz[i] + difdz[i])/
  2 + Ydz[i,2]*delta – beta[k – 2]
         Ydz[i,k] ~ dbern(p2[i,k])
         }
          for(k in 18:31)
         {
         logit(p2[i,k]) <- (sumdz[i] – difdz[i])/
  2 + Ydz[i,17]*delta – beta[k – 17]
         Ydz[i,k] ~ dbern(p2[i,k])
         }
      }


  # winbugs works with precision parameters, i.e., the
  # inverted variances
  tau.summz <- 1/sigma2.summz
  tau.sumdz <- 1/sigma2.sumdz
  tau.difmz <- 1/sigma2.difmz
  tau.difdz <- 1/sigma2.difdz
  # variance decomposition, see Van den Berg et al., Twin
  # Res Hum Genet, 9(3), 334–342.
  sigma2.summz <- 4*VarA + 2*VarE
sigma2.sumdz <- 3*VarA + 2*VarE
sigma2.difmz <- 2*VarE
sigma2.difdz <- VarA + 2*VarE
# specification of inverse gamma priors for variance
# components
VarA <- 1/invVarA
VarE <- 1/invVarE
invVarA ~ dgamma(0.10, 0.10)
invVarE ~ dgamma(0.10, 0.10)
# sample the heritability parameter
h2 <- VarA/(VarA + VarE)
# remaining priors
delta ~ dnorm(0, 0.01) # the sex effect
for (i in 1:14)
{
beta[i] ~ dnorm(0, 0.01) # item parameters
}
}
```

References

Achenbach TM (1997) Manual for the young adult self-report and young adult behavior checklist. University of Vermont, Department of Psychiatry, Burlington, VT

Albert JH (1992) Bayesian estimation of normal ogive item response functions using Gibbs sampling. J Educ Stat 17:251–269

Béguin AA, Glas CAW (2001) MCMC estimation of multidimensional IRT models. Psychometrika 66:541–562

Boomsma DI, Busjahn A, Peltonen L (2002a) Classical twin studies and beyond. Nat Rev Genet 3:872–882

Boomsma DI, Molenaar PCM (1986). Using LISREL to analyze genetic and environmental covariance structure. Behav Genet 16:237–250

Boomsma DI, Vink JM, Beijsterveldt CEM, De Geus EJC, Beem AL, Mulder EJCM, Riese H et al (2002b) Netherlands Twin Register: a focus on longitudinal research. Twin Res 5:401–406

Box GEP, Tiao GC (1973) Bayesian inference in statistical analysis. Addison-Wesly, Reading, Mass

Bradlow ET, Wainer H, Wang X (1999) A Bayesian random effects model for testlets. Psychometrika 64:153–168

Crittenden LB (1961) An interpretation of familial aggregation based on multiple genetic and environmental factors. Ann New York Acad Sci 91:769–780

De Boeck P, Wilson M (eds) (2004) Explanatory item response models: a generalized linear and nonlinear approach. Springer, New York, NJ

Derks EM, Dolan CV, Boomsma DI (2004) Effects of censoring on parameter estimates and power in genetic modeling. Twin Res 7:659–669

Eaves L, Erkanli A, Silberg J, Angold A, Maes HH, Foley D (2005) Application of Bayesian inference using Gibbs sampling to item-response theory modelling of multi-symptom genetic data. Behav Genet 35:765–780

Eaves L, Silberg J, Foley D, Bulik C, Maes H, Erkanli A, Angold A, Costello EJ, Worthman C (2004) Genetic and environmental influences on the relative timing of pubertal change. Twin Res 7:471–481

Falconer DS (1965) The inheritance of liability to certain diseases, estimated from the incidence among relatives. Ann Hum Genet 29:51–71

Fox JP, Glas CAW (2001) Bayesian estimation of a multilevel IRT model using Gibbs sampling. Psychometrika 66:271–288

Fox JP, Glas CAW. (2003) Bayesian modeling of measurement error in predictor variables. Psychometrika 68:169–191

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

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

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

Glas CAW (1998) Detection of differential item functioning using Lagrange multiplier tests. Stat Sinica 8:647–667

Hoijtink H, Molenaar IW (1997) A multidimensional item response model: Constrained latent class analysis using the Gibbs Sampler and posterior predictive checks. Psychometrika 62:171–189

Holman R, Glas CAW (2005) Modelling non-ignorable missing data mechanisms with item response theory models. Brit J Math Stat Psy, 58:1–17

Janssen R, Tuerlinckx F, Meulders M, de Boeck P (2000) A hierarchical IRT model for criterion-referenced measurement. J Educ Behav Stat 25:285–306

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

Lawley DN (1943) On problems connected with item selection and test construction. P Roy Soc Edinb A 62:74–82

Little RJA, Rubin DB (1987) Statistical analysis with missing data. John Wiley, New York

Lord FM (1952) A theory of test scores. Psychometric Monograph No 7. Psychometric Society, New York

Lord FM (1953) The relation of test score to the trait underlying the test. Educ Psychol Meas 13:517–548

Lord FM (1980) Applications of item response theory to practical testing problems. Erlbaum, Hillsdale, NJ

Lord FM, Novick MR (1968) Statistical theories of mental test scores. Addison-Wesley, Reading

Lubke G, Dolan C, Neale MC (2004) Implications of absence of measurement invariance for detecting sex limitation and genotype by environment interaction. Twin Res 7:292–298

Lynch M, Walsh B (1998) Genetics and analysis of quantitative traits Sinauer, Sunderland, MA

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

Moustaki I, Knott M (2000) Weighting for item non-response in attitude scales using latent variable models with covariates. J R Stat Soc Ser A 163:445–459

Moustaki I, O'Muircheartaigh C (2000) A one dimensional latent trait model to infer attitude from nonresponse for nominal data. Statistica 60:259–276

Patz RJ, Junker BW (1999a) A straightforward approach to Markov chain Monte Carlo methods for item response theory models. J Educ Behav Stat 24:146–178

Patz RJ, Junker BW (1999b) Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. J Educ Behav Stat 24:342–366

Rabe-Hesketh S, Skrondal A (2005) Multilevel and longitudinal modeling using stata. Stata Press, College Station, TX

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

Rietveld MJH, Hudziak JJ, Bartels M, van Beijsterveldt CEM, Boomsma DI (2004) Heritability of attention problems in children: longitudinal results from a study of twins, age 3 to 12. J Child Psychol & Psychiat 45:577–588

Risch N (1990) Linkage strategies for genetically complex traits, I: multilocus models. Am J Hum Genet 46:222–228

Risch N (2001) The genetic epidemiology of cancer: Interpreting family and twin studies and their implications for molecular genetic approaches. Cancer Epidem Biomar 10:733–741

Robert CP, Casella G (2004) Monte Carlo statistical methods, 2nd edn. Springer, New York

Rubin DB, Thomas N (2001) Using parameter expansion to improve the performance of the EM algorithm for multidimensional IRT population-survey models. In: Boomsma A, van Duijn MAJ, Snijders TAB (eds) Essays on item response theory. Springer, New York, NJ, pp 193–204

Samejima F (1969) Estimation of latent ability using a pattern of graded scores. Psychometrika, Monograph Supplement, No. 17

Skrondal A, Rabe-Hesketh S (2004) Generalized latent variable modeling. Chapman and Hall/CRC, London

Thissen D, Chen WH, Bock RD (2003) Multilog (version 7) [Computer software]. Scientific Software International, Lincolnwood, IL

Tutz G (1990) Sequential item response models with an ordered response. Brit J Math Stat Psy 43:39–55

Van den Berg SM, Beem L, Boomsma DI (2006a) Fitting genetic models using MCMC algorithms with BUGS. Twin Res Hum Genet 9:334–349

Van den Berg SM, Setiawan A, Bartels M, Polderman TJC, van der Vaart AW, Boomsma DI (2006b) Individual differences in puberty onset in girls: Bayesian estimation of heritabilities and genetic correlations. Behav Genet 36:261–270

Van den Berg SM, Willemsen G, de Geus EJC, Boomsma DI (2006c) Genetic etiology of stability of attention problems in young adulthood. Am J Med Genet, Part B: Neuropsych Genet 141B:55–60

Van den Oord EJCG, Pickles A, Waldman ID (2003) Normal variation and abnormality: an empirical study of the liability distributions underlying depression and delinquency. J Child Psychol Psych 44:180–192

Verhelst ND, Glas CAW, de Vries HH (1997) A steps model to analyze partial credit. In: van der Linden WJ, Hambleton RK (eds) Handbook of modern item response theory. Springer, New York, pp 123–138

Wainer H, Bradlow ET, Du Z (2000) Testlet response theory: an analog for the 3PL model useful in testlet-based adaptive testing. In: van der Linden WJ, Glas CAW (eds) Computerized adaptive testing: theory and practice. Kluwer Academic Publishers, Boston, MA, pp 245–269