Can Psychometric Measurement Models Inform Behavior Genetic Models? A Bayesian Model Comparison Approach

Ting Wang, Phillip K. Wood, and Andrew C. Heath

As methodologists have increasingly noted, the role of psychometrics in operationalizing a construct is often overlooked when evaluating research claims (Borsboom 2006). In a related vein, others have noted that psychological research appears to move away from assessment and interpretation of a single a priori statistical model to a more nuanced comparison of models which assess the trade-off between a model's parsimony and complexity in explaining behavior (e.g., Rodgers 2010). The genetic factor model is one such statistical model often used to estimate the relative contributions of genetic and environmental components of observed behavior in genetically informative designs (Heath, Neale, Hewitt, Eaves, & Fulker 1989; Martin & Eaves 1977; Neale & Cardon 1992). Mathematically, the genetic factor model decomposes observed phenotypic variability into additive genetic (A), common (C), and unique (E) environmental components and is, for that reason, often referred to as the ACE model.

Recently, Franić et al. (2013) discussed how the genetic factor model can be used in the service of psychometrics by informing researchers about the different patterns of dimensionality and factor structure associated with genetic and environmental components of the ACE model. They note that adjudication of dimensionality is obviously not possible based on phenotypic factor analysis which does not take into account the genetically informative nature of the data. In their paper, Franić et al. propose conducting a Cholesky decomposition for the genetic and

Electronic supplementary material: The online version of this chapter (doi: 10.1007/978-3-319-20585-4_10) contains supplementary material, which is available to authorized users.

T. Wang • P.K. Wood (✉) • A.C. Heath
Department of Psychological Sciences, University of Missouri, Columbia, MO, USA
e-mail: phillipkwood@gmail.com

M. Stemmler et al. (eds.), *Dependent Data in Social Sciences Research*, Springer
Proceedings in Mathematics & Statistics 145, DOI 10.1007/978-3-319-20585-4_10

environmental components of the ACE model and decide on dimensionality of each of the environmental and genetic components of the model. They then rotate this solution to a more substantively meaningful form using Promax rotation.

Absent a strong a priori rationale for the factor structure of the environmental and genetic components of the ACE model, this approach appears reasonable and reflects the general practice of behavior genetic models which involve several items (e.g., Heath, Eaves, & Martin 1989; Heath, Jardin, Eaves, & Martin 1989), repeated measurements across successive occasions (e.g., Chang, Lichtenstein, Asherson, & Larsson 2013; Roberson-Nay et al. 2013), and multivariate studies of simultaneously measured variables in which some rationally defined order or priority exists across the manifest variables (e.g., Ludeke, Johnson, & Bouchard 2013). It is well appreciated that such Cholesky decompositions are not unique and that models consisting of other triangular orderings, models with common factors and residual subfactors, or autoregressive factors may fit such data equally well (Loehlin 1996). Rotation of initial Cholesky factorization to more conceptually meaningful form such as simple structure is also a reasonable procedure (e.g., Carey & DiLalla 1994).

Although the strategy outlined by Franić et al. is quite promising, the present paper proposes four reasons why a more fine-grained Bayesian psychometric approach may prove useful. First, for reasons discussed below, multifactor ACE models sometimes encounter empirical under-identification problems. Second, in some research contexts (such as, for example, the genetic analysis of body mass index data considered below where a variety of ages are considered but for which any one individual is assessed at multiple, but not all, measurement occasions), Cholesky factorization across all measurement occasions is not mathematically possible. Third, rotation of the identified solution to simple structure and the original Cholesky decomposition may obscure the psychometric measurement model underlying the construct of interest. Finally, there is reason to believe that Bayesian estimation may be preferable to ML or eigenvalue decomposition. This is particularly the case when sample sizes are small (Boomsma 1982; Chou, Bentler, & Satorra 1991; Hoogland & Boomsma 1998; Hu, Bentler, & Kano 1992; Lee & Song 2004). Additionally, Carey, Goldsmith, Tellegen, and Gottesman (1978) speculate that discrepant estimates of genetic and environmental effects in personality and psychiatric traits may be due to over-extraction of factors or to factors which describe weak effects which limit the generalizability of exploratory factor loadings in the ACE model. Again, these concerns are not meant to criticize the general approach outlined by Franić et al., but instead to highlight that refining the set of candidate psychometric measurement models provide researchers with models which may not be immediately obvious in some situations or estimable in other contexts.

Empirical Under-Identification

The issue of empirical under-identification is not unique to the estimation of genetic models and can occur when researchers attempt to fit a factor model which is more complex than the true model which generated the data, when small sample sizes are examined, and when the factor loadings of the model describe weak or non-existent effects (Kenny, Kashy, & Bolger 1998; Kenny & Milan 2013). Within genetic factor models, the problem of empirical under-identification manifests itself in convergence failures or improper solutions (such as negative variance estimates or estimation correlations which exceed one; see e.g., Phillips & Matheny 1997). Rietveld, Posthuma, Dolan, and Boomsma (2003) discuss the identification issue as it bears on the statistical power of a given genetic model, noting that a given behavior genetic model is mathematically identified if and only if the null space of the Jacobian is zero (i.e., has full column rank). This is, however, only a necessary but not sufficient condition for a specific model within the context of a particular data set.

As Kenny and Milan (2013) note, researchers who encounter empirical under-identification problems usually make post-hoc changes to the model such as redacting individual parameters thought superfluous or adding indicator variables to improve the resolution of the factor structure or instrumental variables which help resolve erroneously specified directions of causality in the model. Researchers using genetic models often constrain parameters of the model to equality or set other parameters to zero (Henderson 1982). Other strategies have included reducing the number of factors considered due to the presumed lack of statistical power associated with the sample (e.g., Martin, Scourfield, & McGuffin 2002). Rietveld et al. (2003) have noted that this state of affairs can be somewhat confusing given that at times researchers have claimed particular genetic models are over-parameterized and not identified while others have investigated the model and found this not to be the case.

Measurement Models

That notions of strictly parallel, tau-equivalent, and congeneric measurement models can be expressed as structural equation models has been noted since Lord, Novick, and Birnbaum's (1968) classical test theory text. In the case of measurement equivalence across a set of manifest variables, strictly parallel measurement requires that both error variances and factor loadings are identical for all variables. Tau equivalence, by contrast, assumes only that the loadings are identical and congeneric measurement permits the factor loadings and error variances across items to be different. Mathematically identified exploratory factor models correspond to a congeneric measurement model, while the tau equivalent model constitutes a more parsimonious model because the loadings across manifest variables are constrained to equality.

In other cases, however, a behavioral or genetic component may be poorly represented by a single congeneric factor, requiring more complex measurement alternatives. Although in many situations multiple oblique or orthogonal factors may be appropriate, measurement models which are intermediary between the one- and two-factor models may be appropriate in other situations. The random intercept model (Maydeu-Olivares & Coffman 2006) is one such model, consisting of both a freely estimated factor and an orthogonal general random intercept factor. The interpretational status of the random intercept factor depends on the particular constructs under investigation: Maydeu-Olivares and Coffman, for example, interpreted the random intercept factor they found in questionnaire data as a general response bias method factor and interpreted the remaining congeneric factor as the construct of interest. When the manifest variables under consideration consist of repeated measurements of the same variable, the factor pattern of the random intercept factor model corresponds to those which would be observed under the free basis growth curve model of Meredith and Tisak (1990). The random intercept factor model differs from the free basis model only in that the random intercept model estimates separate intercepts for each manifest variable and assumes that the latent variables of interest have a zero mean, while the growth curve model assumes that such intercepts are constrained to zero and mean levels of the manifest variables are explained by estimated latent variable means. Taken together, the tau equivalent, congeneric, and random intercept factor models constitute a more fine-grained set of measurement models which are simpler (in the case of the tau-equivalent model) or intermediate models between the dimensions considered under traditional factor analytic models. It is hoped that such a process will result in a “right-sizing” of the statistical model which will result in models which are easier to fit and may well be more generalizable across replications.

Specifically, we speculate that the standard single-factor model may be an over-complex measurement model when effects are relatively week. Specifically, estimation of the distinct individual loadings of the common factor model assumes a congeneric measurement model for a particular genetic or environmental component while the tau-equivalent measurement model which constrains loadings to be equal across variables may be more appropriate. Mathematical derivations (Davis-Stober 2011) also support the idea that predictor weights in the general linear model fail to replicate across samples because of just such over-complexity. This effect is found to be especially true when the sample size is small ($N < 150$) and the effect size of interest is moderate or small (R^2 is smaller than 0.6). Since the measurement model of factor analysis is a type of regression as well (although admittedly one in which the predictor variable for all observed variables is missing), it seems reasonable that similar difficulties in generalizability would be found. Although the sample sizes for behavior genetic studies are frequently quite large, in some contexts (such as the assessment of multiple cohorts of twins measured prospectively), the sample sizes associated with the data in some contexts may be rather small and comparable to the values considered by Davis-Stober. Because phenotypical behavior is frequently thought to entail expression of multiple genes, with each gene exhibiting only a small unique effect (Joseph & Ratner 2013; Turkheimer 2000), the effect sizes of interest may well fall into the “moderate to small” criterion considered by Davis-Stober. In any event, exploration of genetic and environmental components under a tau equivalent model may provide a useful parsimonious comparison to the estimates from the congeneric factor model.

Robustness to Small Sample Size and/or Small Experimental Effects

Finally, as reviewed above, there is some reason to believe that exploration of more parsimonious measurement models using Bayesian estimation may be preferable to congeneric ML estimates when the effect of interest is small or when measurement is based on relatively few observations. If, for example, the additive genetic components of a model consist of a random intercept factor model, but the remaining environmental components are congeneric factor models, a researcher who fits a random intercept or congeneric factor model to all components will likely find that the resulting model is not empirically identified under maximum likelihood (ML) estimation. Even assuming congeneric measurement across all components, this predicament would also occur under triangular factorization if some components consist of multiple factors while others are well-represented by single factors. As another example, assuming a tau equivalent factor model may be appropriately parsimonious when summarizing effects which appear to be small across all manifest variables. As described below, we propose that Bayesian models which compare measurement models for the individual components of the genetic factor model may inform researchers of the relative explanatory power of different measurement models across genetic and environmental components (Lee 2007; Lindley 1977).

We will now present the formal definitions of the three measurement models we wish to consider in the genetic factor model, the tau equivalent, congeneric (i.e., standard factor), and random intercept factor models. We will then describe how such measurement models can be estimated and compared using a Bayesian conjugate approach. This approach will then be illustrated using simulated and real-world data.

Psychometric Models: Tau-Equivalent and Congeneric Factor Models

The standard factor model for N individuals measured across k variables in which j latent variables are assessed can be represented in matrix notation as follows (using Sörbom’s 1974 notation but with the small adaptation that models are presented so that rows of observed scores correspond to individuals and columns correspond to variables):

$$Y = \alpha + \eta\Lambda + \varepsilon$$

The matrix Y contains N rows of individuals, and k variables (which can consist of repeated measurements of a single variable or different manifest variables at a single occasion). α represents an N by k column scalar matrix of intercepts. η is an N by j matrix of values on the latent variable(s) of the model, and Λ is a j by k matrix of factor loadings. ε is an N by k matrix of errors under the assumption that each column of ε is *i.i.d.* across the N rows. When only one factor is present, the variance/covariance matrix is constrained to unity to mathematically identify the model. Identification of multiple orthogonal factors via triangular decomposition was discussed above. The variance/covariance matrix associated with the matrix of errors of predictions, ε, is usually referred to as Ψ and is most often specified as a diagonal, freely estimated matrix. When all possible factor loadings are freely estimated the resulting measurement model is referred to as a congeneric factor model and is the standard measurement model used in the ACE model.

As noted above, the tau-equivalent factor model (Lord et al. 1968, pp. 47–50) assumes that factor loadings in λ are equal. Mathematically, this model is equivalent to the random intercept component employed in some hierarchical linear models, except that in these models, the variance of the factor is assumed to be freely estimated and the factor loadings in Λ are fixed to 1.

Complex Alternative Models: Random Intercept Model

In the random intercept model (RI) (Maydeu-Olivares & Coffman 2006), two orthogonal factors are estimated, with one factor consisting of freely estimated parameters as in the single-factor congeneric model, and the remaining factor's loadings constrained to equality (or equivalently, to unity with a freely estimated factor variance). As noted earlier, in terms of the number of estimated parameters, the RI model is more complex than the single-factor congeneric model (by estimating a single loading across all manifest variables on the second factor), but more parsimonious than the orthogonal two-factor model (which has k-2 more degrees of freedom than the RI model due to the k-1 freely estimated loadings on the second factor). The RI model also differs from the usual multifactor orthogonal models (such as Cholesky or other triangular decomposition) in that each manifest variable is assumed to load on both factors.

Specifically, using the factor model notation defined above each row vector of Λ can now be written as Λ_0, Λ_1, $\Lambda_2, \ldots,$ Λ_k. Λ_0 represents the random intercept factor and all Λ_0 are constrained to 1 with the variance associated with the intercept factor freely estimated or, equivalently, with all Λ_0 constrained to equality and the intercept variance constrained to unity. Λ_1 through Λ_k are defined as before for the multifactor congeneric measurement model. For those more accustomed to path diagram representations, Fig. 1 shows the random intercept model for the case of

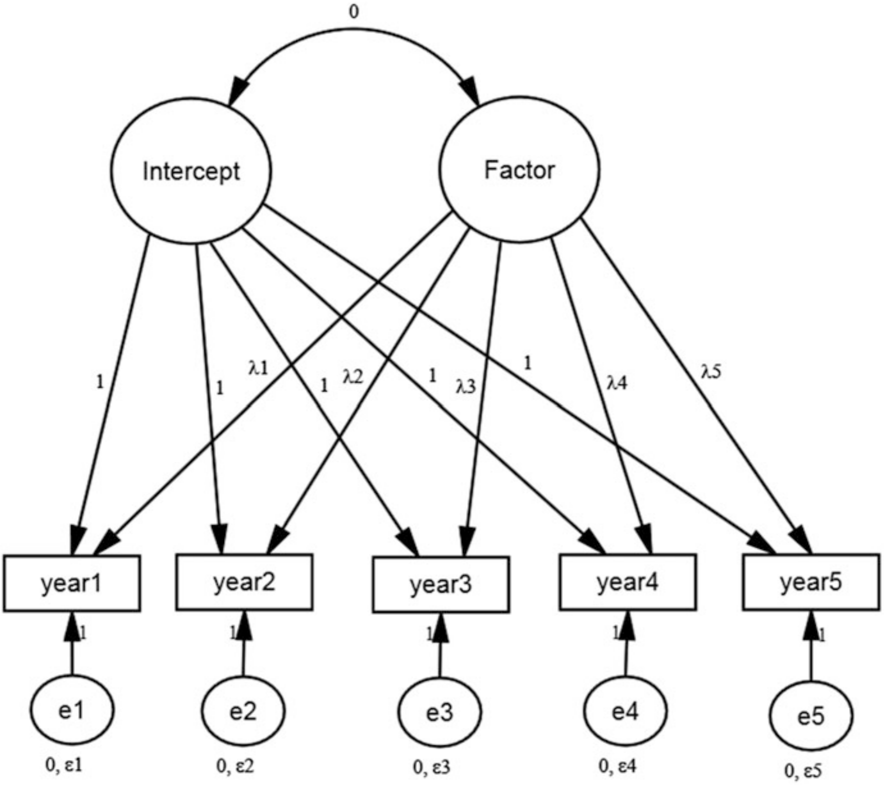

Fig. 1 RI model path diagram

five time measurements. In this Figure λ_1–λ_5 indicate loadings in Λ_1. The random intercept's loadings are represented by the λ_{RI} which are fixed to equality over the measurement occasions. Given that the ACE model frequently assumes unit variances, we chose this strategy to identify the random intercept factor.

The interpretational status of the random intercept factor depends on the particular research situation. As Maydeu-Olivares and Coffman (2006) note, one source of such variability in cross-sectional data may be due to response format, such as systematic negative (or positive) wording of items or a general method factor associated with response. In their analysis of optimistic orientation, Maydeu-Olivares and Coffman found that the random intercept factor resulted in better fit to the data than the traditional one-factor model and was also a parsimonious alternative to a two-factor simple structure model. They interpreted the random intercept factor as a general endorsement or acquiescence factor or, more generally, as a method factor associated with the Likert assessment format. Within the context of longitudinal data, however, the RI model is identical in structure to the free basis growth curve model (Meredith & Tisak 1990) except that, in the growth curve model, mean level information in the manifest variables is used to estimate factor means for both factors while in the random intercept model, factor means are assumed to be zero and individual manifest variable intercepts are estimated. As such, the RI model could represent such a growth process, but the statistical model relies only on the variance/covariance matrix for the identification of such change patterns. As such, when a single group of monozygous and dizygous twins is analyzed (as for the female twin data considered below), the random intercept factor model loadings are identical to those associated with the reference group considered in Dolan, Molenaar, and Boomsma's (1989 1992) multigroup structured means genetic factor model. An explication of an approach to the structuring of mean effects models for genetic data involves a survey of several articles by Dolan and colleagues as well as consideration of additional psychometric models and is the object of a companion article.

Genetic Factor Model in Factor Analysis Notation

As described in Heath et al. (Heath, Eaves & Martin 1989; Heath, Jardin et al. 1989; Heath, Neale et al. 1989), the genetic factor model for twin data is an extension of the factor model described above, except that η is an $n*6$ matrix, with distinct η_A, η_C, and η_E representing the additive genetic, common environmental and unique environmental components for each member of the twin pairs under consideration. Variances across all latent variables are fixed to unity and three additional constraints are placed across the three factors associated with on the ACE structural model: For monozygotic twins, the correlation between genetic components across twins is fixed to 1; for dizygotic twins, this correlation is fixed to 0.5. Finally, the correlation between common factors across both twins is constrained to 1.

Random Intercept Factor Model Applied to ACE Model

One general model for assessment of the psychometric properties of the ACE model occurs when all the three components of the ACE model are modeled as random intercept factors. We therefore differentiate six factors for the resulting genetic model in which we subscript intercept factors to indicate their status as random intercept factors. Accordingly, the terms A, $A_{\text{Intercept}}$, C, $C_{\text{Intercept}}$, E, and $E_{\text{Intercept}}$ denote the congeneric and tau-equivalent components of the genetic factor model for the additive genetic, common environmental, and unique environmental effects respectively. Matrices of the resulting genetic factor model consist of the observed scores of Y as an *n by 2k* matrix for k measurement occasions. The column scalar matrix of α has dimensions *n by 2k* matrix, η is an *n by 12* matrix of factor values, and λ is a patterned *12 by 2k* matrix of factor loadings. This random intercept genetic factor model is the same as the traditional genetic factor model except that each

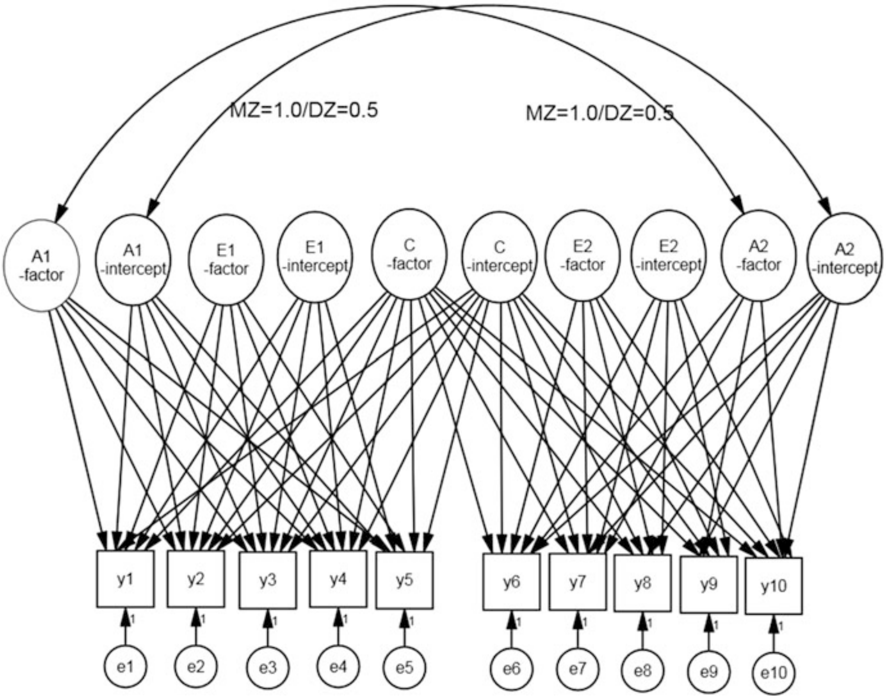

Fig. 2 Proposed model path diagram

genetic and environmental component is represented by two, rather than one latent variable due to the addition of a random intercept model. Model constraints for this model are identical to those for the genetic factor model described above, with the A, $A_{\text{Intercept}}$, C, $C_{\text{Intercept}}$, E, and $E_{\text{Intercept}}$ components assumed uncorrelated. The proposed full model path diagram is shown in Fig. 2.

Bayesian Estimation

Basic Principles and Concepts of Bayesian SEM

As noted above, different measurement models may be appropriate across the genetic and environmental components of the model. Some components may be modeled best as tau equivalent, for example, while the congeneric or random intercept factor models may be most appropriate for other components. If this is the case, researchers attempting to estimate the full RI measurement model for all components are "over-factoring" the data (Rindskopf 1984; Sato 1987) and are likely to find that the model is empirically under-identified due to the nonuniqueness of the solution space (Savalei & Kolenikov 2008). Such estimation difficulties are not present, however, in Bayesian approaches and parameters which are zero or very close to zero are simply estimated as any other parameters in the model (e.g., Lee 2007). The methodological benefit of this approach, however, is that prior distributions may exclude improper values by definition. (For example, variances estimated in the Bayesian approach using the inverse Gamma distribution can never take negative values, thereby preventing one type of improper solution.) In addition, because the matrix of parameter estimates does not need to be inverted, locally degenerate solutions are not encountered during the process of estimation (Shi & Lee 1998). This permits researchers to compare the relative fit of models with different measurement models across components.

Several excellent treatments of Bayesian inference and use of the Gibbs Sampler are available in both systematic (Gelman, Carlin, Stern, Dunson, & Vehtari 2013) and didactic presentations. For Winbugs applications of the Gibbs Sampler, Eaves et al. (2005) present a Bayesian genetic IRT analysis of questionnaire items and Zhang, Hamagami, Wang, Nesselroade, and Grimm (2007) present Winbugs specifications of growth models. Muthén (2010) presents a similar discussion of example analyses and technical aspects using Mplus. In the interests of space, we will not repeat these presentations, but will limit our discussion to those topics which deal with the basic logic of Bayesian SEM and those technical aspects of estimation which proved most important to the estimation of the random intercept genetic factor model.

Bayes' Theorem

Let M be an arbitrary structural equation model consisting of both parameter specifications of the model with a vector of unknown parameters θ. For brevity of presentation, we will take M to represent both the structural equations representing the model as well as any (possibly informative) prior beliefs of the researcher about these parameters expressed via an appropriate probability distribution. Let Y again be the observed data defined as in Equation 1 above. Based on a well-known identity in probability (Gelman et al. 2013), the posterior probability density function associated with θ given the observed data and structural model may be defined as:

$$p\left\langle \theta | Y, M \right\rangle = p\left(Y\left|\theta, M\right.\right) p\left(\theta\right)$$

$p(\theta|Y,M)$ represents the posterior density function of the researcher's beliefs about the parameters of the model. $p(Y|\theta,M)$ can be regarded as the likelihood function. The posterior density function incorporates the sample information and the prior density function $p(\theta)$ (Lee & Song 2004).

Gibbs Sampler

The joint analytic form of the posterior distribution poses difficulties to a formal evaluation of the density (Lee 1980). As a result, data augmentation procedures involving Markov Chain Monte Carlo (MCMC) methods such as the Gibbs Sampler are used to obtain the posterior distribution of $p(\theta|Y)$. Such techniques involve a successive iterative approach to generating estimates of the posterior distributions of the parameters and also provide some indication of the reasonableness of the distributional assumptions of the model. Let η be the set of latent variables in the model. The rationale is that adding latent variables η could turn the conditional distribution $p(\theta|Y, \eta)$ and $p(\eta|Y, \theta)$ into simpler form. Given a sample $\{\theta^{(t)}, \eta^{(t)}\}$ draws from $p(\theta, \eta|Y)$, an iteration

$$\theta^{(t+1)} p\left(\theta \left|Y, \eta^{(t)}\right.\right)$$

$$\eta^{(t+1)} p\left(\eta \left|Y, \theta^{(t+1)}\right.\right)$$

samples a new state $\left\{\theta^{(t+1)}, \eta^{(t+1)}\right\}$. In the end, we could get enough samples in the chain and observe the posterior distribution of θ (Geman & Geman 1984). At convergence, different chains generated with different starting values are merged together (after discarding a number of iterations during the beginning phases of each, which are treated as burn-in iterations). If successive observations are highly positively correlated (as was frequently found in several of the genetic factor models we considered) values are taken only from successive intervals (such as every 20th iteration), a process known as “thinning” (Gelman et al. 2013).

Data from the MCMC iterations used in estimation can also be plotted as a diagnostic of whether the parameter of interest appears to take the form assumed by the distributions chosen by the researcher to represent beliefs about the parameter, a method known as Posterior Predictive Checking (PPC, Gelman et al. 2013). As described below, in the data sets considered in this paper, PPC of the estimated posterior distributions alerted us to the fact that the Gibbs Sampler was prone to produce multi-modal posterior distributions symmetric about zero for random intercept factor loadings, particularly if the size of the effect was modest. We discuss this issue and solutions below.

Model Fit

In addition to providing posterior distributions about the parameters of interest, the Bayesian approach also permits the researcher to evaluate the fit of the structural model based on its likelihood. Although several approaches to assessing model fit can be taken (Gelman et al. 2013; Lee 2007) we will discuss three here. The BIC (Schwarz 1978) is popular within structural modeling because it penalizes models for their complexity (expressed as the number of parameters in the model).

The Deviance Information Criterion (DIC, Spiegelhalter, Best, Carlin, & van der Linde 2002) is a Bayesian generalization of information criteria such as the AIC and BIC which penalizes models based on the effective number of parameters in the model. For both the DIC and BIC, smaller values are considered better-fitting. Posterior predictive checking of the likelihood of the model as a whole also permits the researcher to estimate the Posterior Predictive p-value, an estimate based on the PPC of the likelihood ratio chi-square statistic for the model (Meng 1994). This represents a rough estimate of the probability that the data could have been generated under the candidate model. The proposed model may be considered as plausible if the PP p-value estimate is not far from 0.5 (acceptable range 0.3–0.7). Meng (1994) notes that the PP p-value is not suitable for comparing different models but is a reliable index of stand-alone model fit.

Model Comparison: Bayes Factors

In addition to providing stand-alone measures of model fit, it is also possible to assess the relative fit of candidate structural models for the data. In addition to simply comparing the incremental fit, the Bayesian approach also permits the researcher to assess the relative informative power associated with increases in model complexity. Most generally, this comparison is made using the Bayes factor, which we now introduce in some greater detail given the need to understand its basic logic and the fact that its estimation is the object of ongoing study. From the Bayes theorem comparing the odds ratio associated with the comparison of a base model, M_0 with a more complex model, M_1, we can obtain:

$$\frac{p\left(M_1\left|Y\right.\right)}{p\left(M_0\left|Y\right.\right)}=\frac{p\left(Y\left|M_1\right.\right)p\left(M_1\right)}{p\left(Y\left|M_0\right.\right)p\left(M_0\right)}$$

which permits us to define the Bayes factor as

$$B_{10}=\frac{p\left(Y\left|M_1\right.\right)}{p\left(Y\left|M_0\right.\right)}$$

Thus we see that posterior odds = Bayes factor*prior odds (Lee 2007). Larger Bayes factors mean stronger evidence for M_1 relative to M_0. $p\langle Y|\, M_1\rangle, p\langle Y|\, M_0\rangle$ is obtained by integrating $p\langle Y|\,\theta, M_1\rangle, p\langle Y|\,\theta, M_0\rangle$ over the parameter space, respectively. It is, however, often difficult to obtain Bayes factor analytically using a path sampling approach (Gelman & Meng 1998) and, for that reason, another easy and quick way to calculate Bayes factor is by using BIC (Muthén & Asparouhov 2011):

$$BF=\frac{p\left(M_1\right)}{p\left(M_0\right)}=\frac{\exp\left(-0.5BIC_{M_1}\right)}{\exp\left(-0.5BIC_{M_0}\right)}$$

Although there is some dispute about the validity of calculating Bayes factor by using BIC (Gelman et al. 2013), in the simulation study presented below, we found that this criterion worked well in practice. Generally speaking, Bayes factors less than 3 represent minimal support for the alternative model, values between 3 and 20 positive support for the alternative model, values between 20 and 150 strong support, and values larger than 150 decisive support.

Simulated Data Example

We will now illustrate our general approach of fitting a general random intercept genetic factor model and assessing the relative fit of more parsimonious measurement models for the genetic and environmental components using simulated data (generated from SAS). We generated simulated twin data for 1000 hypothetical twin pairs using the following factor loadings: Across all variables, $A_{\mathrm{Intercept}}$, $C_{\mathrm{Intercept}}$, and $E_{\mathrm{Intercept}} = 0.4$, 0.4, and 0.3, respectively. A factor loadings were zero. C factor loadings were chosen as 0.4, –0.4, 0.3, –0.3 and 0.2 across the five variables. E factor loadings were chosen as –0.3, –0.3, 0.3, 0.3, and 0.3.

To demonstrate the ability of the procedure to correctly arrive at a more parsimonious model and to highlight the empirical under-identification issues associated with more parsimonious models under ML estimation, we chose to simulate data in which an intercept model was appropriate for the additive genetic component, but for which RI models were appropriate for the shared and unique environmental components. 1000 replication data sets were generated to investigate the sampling behavior of the approach using SAS. Models were estimated using Mplus (Muthén and Muthén 1998–2010). The Gibbs Sampler iteration number was set at 5000 to allow a generous amount of iterations for the MCMC chains in the Bayesian analyses. By default, the first half of these iterations was used as a burned-in phase. Initial inspection of the MCMC chains revealed marked auto-correlation across iterations of the Gibbs Sampler, and so a thinning value of 50 was chosen for the analyses which appeared to remedy the auto-correlation problem (Albert & Chib 1993). All 1000 replications met the convergence criteria by Bayesian estimation (PSR close to 1 for each parameter) (Muthén & Asparouhov 2011). The PP-p value associated with the general RI model had a mean of 0.53, standard deviation 0.25 across replications, indicating good fit. In addition, the genetic slope factor loadings are all non-significant. Under ML estimation, however, all 1000 samples failed to converge which we take as evidence that they were not empirically identified. When the correct model is fit to the data, however, ML models did converge. There was little difference between the Bayesian and ML estimates under the correct model, with bias estimates not exceeding 2 % across the estimated loadings (See Table S1 in supplemental materials accompanying the manuscript.)

Table 1 Percentages of replications with Bayes factors > 20 favoring column model over row model across 1000 simulated samples

Model	Two-factor model	RI model	ACE model	True model
Two-factor model	NA	91.6 %	47.4 %	99.8 %
RI model	7.70 %	NA	10.4 %	99.4 %
ACE model	52.0 %	89.3 %	NA	96.3 %
True model	0.20 %	0.60 %	3.50 %	NA

Table 2 Summary of parameter bias in traditional ACE genetic factor model

	Bayesian estimation	ML estimation
Parameter	Bias Mean (S.D.)	Bias Mean (S.D.)
A loadings	45.8 % (0.106)	45.5 % (0.110)
C loadings	−39.7 % (0.071)	−39.1 % (0.073)
E loadings	−27.7 % (0.379)	−27.7 % (0.377)

Model Comparison

Table 1 presents proportions of model comparisons across replications in the simulated data which exceed criteria for strong support in comparisons of the true model, the traditional ACE model, and a freely estimated two-factor solution across all genetic and environmental components. As can be seen from the fifth column of the table, the true model is preferred over the competing two-factor, random intercept, and traditional ACE models in 96.3–99.8 % of the cases. The two-factor model is preferred over the traditional ACE model in only 52 % of the cases, and the random intercept model is preferred over the traditional ACE model in 89.3 % of the cases. This high latter percentage is unsurprising, given that the random intercept model differs from the true model only in that the A factor is redacted from the RI model to produce the true model. Taken together, model comparisons based on the simulated data reveal that Bayesian estimation appears able to correctly identify the correct model and, even when the model under consideration is slightly over-complex, the factorial complexity of the genetic and environmental components in these data is detected.

Genetic Factor Model

When these data were analyzed with the (mis-specified) traditional genetic factor model in which all three components are assumed to have a congeneric measurement model (i.e., have only A, C, and E factors), all 1000 replication yielded a zero PP-p value, indicating poor model fit. The bias summary associated with the ACE model is presented in Table 2. Results suggest that, for the simulated data considered here, failure to correctly include intercept components for the common unique environmental effects introduces substantial bias in the estimated additive effects of the model.

Two-Factor Model

As noted above, the RI measurement model is a two-factor model, but one with considerably more parsimony than the traditional two-factor congeneric model. The question therefore remains as to whether the Bayes factor can also correctly reward the greater parsimony of the RI model relative to the more complex traditional two-factor model. When the traditional two-factor model is estimated from the data, the PP-p value has mean of 0.50, with standard deviation 0.23, indicating the high degree of model fit found for the (true) RI model. To secure a mathematically identified solution for the two-factor model, the first loading associated with each of the A2, C2, and E2 factors was set to zero. Bias estimates for the two-factor model are of necessity quite pronounced, given that the Cholesky form of the two-factor model represents an affine rotation of the true structure of the data. If calculated as a percent bias relative to the true model, bias estimates of the two-factor Cholesky model averaged 37.2 %, with bias across the particular types of loading ranging from 12.5 to 97.5 % (See Table S2 in the supplemental materials accompanying the manuscript.)

Alternatively, if the approach outlined by Franić et al. (2013) is followed, the correct dimensionality of the genetic and environmental components is identified as a two-factor solution. However, the structure of the random intercept model is not correctly specified due to the fact that the resulting decomposition is triangular in nature. Even if the two-dimensional factor structure is rotated via an affine transformation to a form most closely resembling the true factor structure, two of the recovered loadings still deviate by approximately .05 due to sampling variability. Since it is difficult to judge empirically in real-world applications whether such variation represents sampling variability or a true multifactor structure in which factor loadings of one factor are unequal to each other, we believe it reasonable to directly compare the two-factor and random intercept models as outlined here.

Other Alternative Models: Bayesian Estimation

In addition to these selected model comparisons, we also compared the true model with all other combinations of the three possible measurement models (tau, congeneric, and random intercept) for each component of the genetic factor model. Model fit indices for the models are shown in Table 3 as well as the Bayes factor comparing the true model to each candidate. Although it would be possible to compare all of these candidate models using Bayes factors, the evaluation of such a matrix of pairwise comparisons would be both tedious and liable to substantial experiment-wise error given the number of contrasts. If, however, researchers compare the relative fit of the random intercept model to models which redact intercept or factor models from the genetic model, a relatively proscribed set of model comparisons results. Well-fitting parsimonious models can then be compared to the random intercept model in an attempt to identify a more parsimonious model. As can be seen in Table 3, when Bayes factors are calculated relative to the random intercept

Table 3 Candidate models' PP-p value and percent of Bayes factors strongly preferring the RI and true models

Model	PP-p value (S.D.)	BIC (S.D.)	BF Perc. Over 20 (RI) (%)	BF Perc. Over 20 (True) (%)
RI model	0.53 (0.25)	49,327.21 (234.73)	NA	99.4
True model	0.52 (0.26)	49,249.37 (215.04)	0.06	NA
RI without $A_{\text{Intercept}}$ model	0.52 (0.25)	49,279.5 (215.83)	4.00	99.3
RI without $C_{\text{Intercept}}$ model	0.07 (0.11)	51,798.18 (2371.75)	95.3	99.8
RI without $E_{\text{Intercept}}$ model	0.00 (0.00)	49,892.57 (712.31)	96.0	99.3
$ACEE_{\text{Intercept}}$ model	0.07 (0.11)	50,167.56 (1492.76)	79.6	100
$ACC_{\text{Intercept}}E$ model	0.00 (0.00)	49,515.99 (220.87)	92.8	100
$AA_{\text{Intercept}}CE$ model	0.00 (0.00)	50,304.43 (1602.06)	96.3	100

model, only the true model and the random intercept model without the $A_{\text{Intercept}}$ factor were not significantly worse fitting than the RI model, as shown in the fourth column. When the true model is considered as a base model, the evidence strongly supporting the true model is found between 99.3 and 100 % of the replications.

Summary Remarks for Simulation Study

Under ML estimation, estimating an over-complex RI measurement model for all three components results in empirical under-identification. When the random intercept is present for the genetic component but the data are analyzed using the traditional ACE model, estimates of heritability of the genetic component are over-estimated under both ML and Bayesian estimation. When the true model is known, however, ML and Bayesian parameter estimates appeared similar. Because of the empirical under-identification problems in ML estimation, comparison of candidate measurement models was only possible under the Bayesian approach. For these data, the correct model was identified using the Bayes factor. Significantly, the random intercept measurement model was also found to be a parsimonious alternative to the traditional two-factor model.

Care must be taken in conducting Bayesian analyses, however. Even with the simulated data under consideration, large thinning values were necessary to reduce autocorrelation across iterations of the Gibbs Sampler and bimodality was observed in some of the PPC plots which indicated possibly misleading estimates and confidence intervals for the Bayesian approach. Once identified, however, these bimodality issues were successfully addressed. In the next section, the general RI genetic factor model and its more parsimonious alternatives are considered in an empirical data example. In addition to the didactic value of a real-world example, use of a real-world example also permits exploration of the effects of the non-normality and unmodeled causal effects on model fit, comparison, and parameter estimation.

Empirical Study

The genetic and environmental effects on body mass index (BMI) have been investigated across several studies. Allison et al. (1996), in a study of Japanese, Finnish, and American twins, reported that additive genetic effects appeared more pronounced at early ages, that the genetic effects did not appear due to shared environmental effects during this time, and that heritability coefficients ranged between .5 and .7 for the data sets considered. Elks et al. (2012), in a review of 88 estimates of the heritability of BMI across twin studies, found heritability estimates ranging from .47 to .90. It is worth noting that most of these estimates (61) were based on AE models (i.e., a model with no common environmental effects), while 15 were based on the traditional ACE model. (The remainder were based on direct comparisons of within and between twin correlations or the non-additive genetic model.) Estimates of the genetic heritability of BMI using the ACE model were generally .12 higher than estimates from the AE model. Readers are referred to Elkes et al. for a discussion of the genome-wide association studies investigating the loci associated with BMI.

The BMI data we wish to analyze are taken from the Missouri Adolescent Female Twin Study (MOAFTS), a genetic-epidemiological, prospective twin-family study of alcohol use in young females. (For full details, including response rates, see Waldron, Bucholz, Lynskey, Madden, & Heath 2013.) Using a cohort sequential design, twins were aged 13, 15, 17, and 19 when first enrolled in the study. In analyses presented here, we exclude African-American twins, because of small numbers but significant mean differences in BMI distribution. A total of 3416 Missouri female adolescent twins (85 % participation rate, approximately 55 % MZ and 45 % DZ) were interviewed from 1995 to 2012 with a telephone version of the Child Semi-Structured Assessment for the Genetics of Alcoholism. In this study, we only concentrated on the body mass index (BMI) variable. Observations from twin pairs with at least five measurement occasions were selected for this longitudinal analysis. Descriptive statistics by age groups are listed in Table 4. Since all observed variables are positively skewed, even after fitting the model, we transformed the data by taking the log of the original data. The following analyses were based on the transformed data.

Bayesian Model Comparison

As in the simulation study, two-factor, RI, and simpler alternatives were considered for the BMI data. Table 5 presents the Bayes factor (relative to the final model), PP-p value, and DIC for each reduced model as well as the two-factor model. A model consisting of a RI model for the additive genetic effect, a tau equivalent model for the unique environmental effect, and no common environmental effect was chosen as the final model based on its Bayes factor relative to the RI model

Table 4 Descriptive statistics of body mass index by age group

	N	Twin 1			Twin 2		
Age	Twin pairs	Mean	S.D.	Skewness	Mean	S.D.	Skewness
13	58	19.9	2.684	0.82	20.26	3.244	1.323
14	71	20.75	3.109	1.061	20.1	3.019	1.465
15	150	21.1	3.222	1.542	21	3.188	1.819
16	110	21.05	3.04	1.437	21.16	3.419	1.842
17	188	21.74	3.113	1.222	21.58	3.731	2.199
18	160	21.97	3.359	1.156	22	3.646	1.434
19	89	23.09	4.443	1.81	22.7	3.523	1.351
20	117	23.07	4.334	1.458	22.53	3.653	1.315
21	31	23.28	4.925	1.763	23.22	5.379	2.387
22	80	23.6	4.833	2.298	23.38	4.736	1.793
23	86	24.84	5.068	1.059	24.23	4.587	1.372
24	68	23.89	4.782	1.221	23.83	4.705	0.997
25	65	24.95	5.538	1.777	24.84	5.35	1.437
>25	113	26.54	5.804	1.125	25.88	6.08	1.152

($1.84*10^{19}$). Although such a choice of models may seem somewhat unusual, it is a choice consonant with other research on BMI during young adulthood; Elks et al. (2012) report 26 studies of BMI spanning both young and older samples compared to nine studies reporting the traditional ACE model. Although such a contrast does not ensure correctness via democratic vote, it does speak to the fact that a decision to redact the common environmental component is not without precedent.

ML Model Comparison

The model comparison results using ML estimation were similar to their Bayesian counterparts and are shown in Table S3 in the supplemental materials for the manuscript. Model fit index such as RMSEA and CFI were very similar across the different models. Moreover, chi-square test cannot be used to compare all models given that they are not nested models. However, based on examination of the BIC values, the $AA_{\text{Intercept}}C_{\text{Intercept}}EE_{\text{Intercept}}$ model demonstrated the best fit ($\text{BIC} = -5335.9$), with the $AA_{\text{Intercept}}C_{\text{Intercept}}E_{\text{Intercept}}$ model showing a value only slightly larger than this ($\text{BIC} = -5304.5$). The model chosen under Bayesian estimation, $AA_{\text{Intercept}}E_{\text{Intercept}}$ ($\text{BIC} = 5165.7$) was larger than these other two models but still lower than the other models considered. On examination of the

It should be noted that when all Bayesian models which included a common environmental effect failed to find environmental effects greater than zero, regardless of whether a tau equivalent, congeneric or random intercept model was used to model the component.

Table 5 PP-p and BIC values for candidate models and Bayes factor of body mass index data

Model	PP-p	BIC	Bayes factor	Bayes factor vs. AA$_{\text{Intercept}}$,E$_{\text{Intercept}}$
AA$_{\text{Intercept}}$,E$_{\text{Intercept}}$	0.291	−5157.05	1.84E + 19	1
Two-factor	0.401	−4729.84	3.14E−74	1.71E−93
RI full model	0.301	−5068.33	1.00E + 00	5.43E−20
A,C,C$_{\text{Intercept}}$,E,E$_{\text{Intercept}}$	0.309	−5075.19	3.09E + 01	1.67E−18
A,A$_{\text{Intercept}}$,C,E,E$_{\text{Intercept}}$	0.317	−4864.42	5.27E−45	2.86E−64
A,A$_{\text{Intercept}}$,C,C$_{\text{Intercept}}$,E	0.333	−4840.35	3.12E−50	1.70E−69
A,A$_{\text{Intercept}}$,C$_{\text{Intercept}}$,E$_{\text{Intercept}}$	0.299	−4918.28	2.61E−33	1.42E−52
A,C,C$_{\text{Intercept}}$,E	0.289	−5055.06	1.31E−03	7.13E−23
A,A$_{\text{Intercept}}$,C,E	0.293	−4910.09	4.34E−35	2.36E−54
A, C, E (Genetic Factor Model)	0.285	−5042.8	2.85E−06	1.55E−25
A$_{\text{Intercept}}$,C,C$_{\text{Intercept}}$,E,E$_{\text{Intercept}}$	0.289	−5130.3	2.85E + 13	1.55E−06
A,A$_{\text{Intercept}}$, C$_{\text{Intercept}}$,E,E$_{\text{Intercept}}$	0.323	−5131.53	5.28E + 13	2.87E−06
A,A$_{\text{Intercept}}$,C,C$_{\text{Intercept}}$, E$_{\text{Intercept}}$	0.285	−5067.28	5.90E−01	3.21E−20
A$_{\text{Intercept}}$, C$_{\text{Intercept}}$,E,E$_{\text{Intercept}}$	0.133	−5130.5	3.16E + 13	1.72E−06
A$_{\text{Intercept}}$,C,C$_{\text{Intercept}}$, E$_{\text{Intercept}}$	0.275	−5136.93	7.87E + 14	4.28E−05
A$_{\text{Intercept}}$C$_{\text{Intercept}}$E$_{\text{Intercept}}$	0.122	−5000.67	2.03E−15	1.10E−34
AE	0.124	−4963.992	2.03E−15	1.19E−42

ML estimates, the C$_{\text{Intercept}}$ and E factor loadings were, although significant, modest in magnitude (all λ’s $<$ .05). Because of the advantages of the Bayesian estimation approach to model comparison and because the additional factors, if present, appeared to represent modest effects, we chose to report ML and Bayesian estimates for this model.

Parameter Estimation

Bayesian parameter estimates based on the final model are shown in Table 6. (Corresponding ML parameter estimates for this model were almost identical in value.) Consistent with Allison et al.’s (1996) finding, the genetic intercept appears to explain more variability than the genetic factor in early years, especially from ages 13 through 18. During later years (from ages 21 through 26 and later), the genetic factor appears to explain roughly the same proportion of variability as the intercept. The pattern of loadings for the genetic factor appears to be roughly nonlinear and suggests systematic differences in the genetic component associated with BMI during the adolescent, young adult, and adult years.

Also consistent with the majority of the twin studies reviewed by Elks et al. (2012), common environmental effect was either not statistically significant (based on Bayesian estimates). Given that dropping CI gave a similar model fit index

Table 6 Bayesian parameter estimates for body mass index data

Age	λ^a	Std. Dev.[a]	$p\|H_0 = 0$	Intercept	Std. Dev.	$p\|H_0 = 0$
	A					
13	−3.40	1.70	0.023	3.00	0.01	0.00
14	−0.30	1.10	0.408	3.02	0.01	0.00
15	1.60	1.00	0.063	3.04	0.01	0.00
16	0.50	1.00	0.285	3.05	0.01	0.00
17	1.80	1.00	0.034	3.07	0.01	0.00
18	3.40	1.00	0	3.08	0.01	0.00
19	5.50	1.30	0	3.11	0.01	0.00
20	5.10	1.10	0	3.12	0.01	0.00
21	8.70	1.80	0	3.14	0.01	0.00
22	8.10	1.20	0	3.14	0.01	0.00
23	10.20	1.30	0	3.17	0.01	0.00
24	9.40	1.50	0	3.18	0.01	0.00
25	9.90	1.40	0	3.18	0.01	0.00
>26	13.00	1.40	0	3.22	0.01	0.00
	$A_{intercept}$					
All Ages	12.90	0.50	0			
	$E_{intercept}$					
All Ages	4.10	0.30	0			

[a] Values in columns multiplied by 100 for ease of presentation

(PP-p value is 0.293) and the Bayes factor is 2.81 favoring the model without CI, we conclude that dropping the CI factor from the model seems reasonable and we note that inclusion of the effect does not seem to affect other parameters and explains at most a minimal amount of variability.

The proportion of variability explained by genetic and environmental effects by age is shown in Table 7. For these data, heritability estimates for the final model (shown in the column labeled "Additive" under the Heading "Final Model") ranged from 0.72 to 0.82 with an average of 0.77 across years, which compares favorably with the 0.75 median estimate from Elks et al.'s (2012) meta-analysis. In contrast to the heritability estimates based on the traditional ACE model and models used in Elks et al.'s study, heritability does not appear to be more pronounced in younger ages than in older ages. Heritability estimates from the traditional ACE model (shown in the same column under the heading "ACE") for these data are somewhat lower (mean = .68, range 0.49–0.82 across years) and appear to be slightly lower for twins older than 21. The difference in average heritability between the final and traditional ACE model of .09 is similar to the 0.12 increase noted by Elks et al. when models are fit which do not include an environmental effect. It is also worth noting that statistically significant common environmental effects using the single-factor ACE model were only found for ages 21 through 25 and, even for these, the proportion of variability in BMI explained was on average 7 %. The discrepancy between the final model and the traditional one-factor ACE model does not appear to be due entirely, however, to the estimation of a common environmental effect, because when a two-factor ACE model is estimated from these data, the average heritability across ages is 0.59, and none of the factor loadings associated with the common environmental effects is statistically significant.

A comparison of unique environmental effects of the final model with the one- and two-factor traditional ACE models reveal that the average unique environmental effect was slightly smaller for the final model (0.06) than for either the one- or two-factor ACE models (.09 and .12, respectively).

Summary Remarks of Empirical Study

Taken together, estimates from the final model under Bayesian estimation produce estimates of heritability consonant with the Elks et al. (2012) review and replicate the conclusion made by many researchers that common environmental effects in BMI appear to be negligible. The pattern of differential common environmental effects found under a one-factor ACE model is not replicated by either the random intercept model selected as most reasonable or by a freely estimated two-factor model. Although, as Visscher, Gordon, and Neale (2008) note, small sample studies may be underpowered to detect a statistically significant common environmental effect, the existence of such differential effects were not found using the Bayesian model comparison procedure outlined here and, even if thought to exist, their magnitude appears to be confined to older ages and to be minimal in comparison to the magnitude of heritability coefficients during these ages. For these data, the proposed model comparison approach appears to yield a model which is both parsimonious and reasonably similar to the larger literature on the magnitude of environmental and genetic effects.

Discussion

The measurement model which researchers choose to operationalize environmental and genetic components of behavior genetic models has important implications for the estimation and interpretations of such models. When psychometric alternatives to the traditional factor model such as the tau equivalent and random intercept models are considered, substantially different estimates of the relative salience of genetic and environmental contributions are obtained. Comparison of candidate measurement models seems warranted given that the psychometric complexity of the true model is largely unknown to the researcher prior to analysis and, even if it were, such exploration can inform the researcher about possible alternate estimates for genetic and environmental components that a reasonable skeptic might raise. Consideration of overly complex genetic models, however, is often prevented in maximum likelihood estimation because such models are not empirically identified,

Table 7 Estimated proportion of variability explained in BMI by final, traditional ACE, and two-factor models

Source	Additive genetic			Common Env.			Unique Env.		
Age	Final model	ACE	Two-factor	Final model[a]	ACE	Two-factor	Final model	ACE	Two-factor
13	0.72	0.79	0.69	0	0.05	0.00	0.07	0.01	0.11
14	0.78	0.81	0.69	0	0.01	0.00	0.08	0.01	0.10
15	0.77	0.82	0.68	0	0.00	0.00	0.08	0.01	0.06
16	0.82	0.81	0.70	0	0.01	0.00	0.08	0.05	0.06
17	0.77	0.77	0.66	0	0.00	0.00	0.08	0.07	0.07
18	0.76	0.66	0.58	0	0.01	0.00	0.07	0.15	0.20
19	0.73	0.57	0.54	0	0.03	0.00	0.06	0.25	0.17
20	0.80	0.77	0.73	0	0.01	0.00	0.07	0.11	0.09
21	0.81	0.69	0.50	0	0.10	0.00	0.06	0.05	0.13
22	0.79	0.59	0.55	0	0.12	0.00	0.06	0.16	0.11
23	0.79	0.57	0.53	0	0.12	0.00	0.05	0.18	0.18
24	0.77	0.68	0.54	0	0.11	0.00	0.05	0.05	0.08
25	0.74	0.49	0.49	0	0.17	0.00	0.05	0.14	0.16
>25	0.76	0.51	0.45	0	0.24	0.01	0.04	0.03	0.11

[a] Fixed to zero

which leaves open the question of whether complex alternatives were simply not numerically obtained given the software or whether they are empirically unidentified due to being over-complex. Such model estimation was, however, possible under Bayesian estimation and estimated parameters for the final model appeared largely similar to corresponding estimates from ML estimation for both the simulated and real-world data.

Given that traditional behavior genetic models often involve the assessment of only different numbers of freely estimated latent variables, this paper seeks to highlight the fact that a greater number of models are possible, given parsimonious patterning of the factor loadings involved. The extension of such models to convey mean effects makes it possible for the researcher to specify the patterning of such variance components as growth curve models. As noted above, a great variety of models are possible under the model comparison procedure described above, which may prompt some researchers to wonder how best to limit the specification and search of models to a more tractable number in practice. In the fortunate cases where the researcher is in the position of having some knowledge concerning the functional form of growth over time, it would be possible to specify nonlinear constraints on the estimated factor loadings so that the underlying estimated curve corresponds to a parametric growth model such as the logistic or Gompertz curve (Grimm & Ram 2009). Increasingly, however, it appears that the patterns of growth observed over time in empirical data do not follow such tidy mathematical specifications, leading some to adopt the nonparametric growth curve as a reference curve for characterizing the form of growth over time. For example, Ram and Grimm (2009), in a study of longitudinal finite mixture models, advocate for initial specification of a free curve growth model as a model of functional change over time which can then serve as a reference form for the identification of finite mixtures. In general, however, adoption of such a “nonparametric” growth curve model raises questions concerning the interpretability of the identified curves.

It is quite possible that one reason for the failure of identified patterns of growth over time to follow a parametric form is due to the fact that Alessandri, Caprara, and Tisak (2012) point out that the presence of a single, but nonparametric pattern of loadings for growth data may indicate the presence of several, rather than one source of stability of time which may include environmental effects, age-related effects or turning points. Given that the genetic and environmental effects identified through behavior genetic models are genetically multi-determined, it is probably more reasonable to expect that identified effects should probably exhibit such composite patterns over the lifespan.

As such, it is important to recall that the proposed model comparison approach to fitting genetic and environmental effects is no panacea and that behavior genetic growth curve models, as with any latent variable model, are subject to the “naming problem” in that the latent variables identified may not represent the constructs initially intended. Although it is possible to attempt a remedy of this by modeling one of the factors of the model according to some agreed upon parametric form and to identify “residual factors” which would model additional covariation due to the extraneous effects, the fact that there is at least in the context of much behavior genetic models, little agreement as to the functional form of growth over time makes such an approach untenable. In the end analysis, probably the best remedy for this ambiguity lies in the identification of such possible confounding effects and data collection strategies designed to provide a less ambiguous portrait of change over time.

Model Support

Rather than basing model comparisons in terms of probability statements common under the frequentist approach, the Bayesian approach permitted adjudication between candidate models based on a quantification of the relative support for a particular model relative to other candidate measurement models. Operationally, measurement models which include random intercept components permit the researcher to consider the model of tau equivalence as a parsimonious alternative to the single-factor model usually considered in genetic factor models. The possibility that the manifest variables of the study constitute equivalently scaled measures would seem an attractive one to researchers, especially if the manifest variables in the model constitute longitudinal assessments. More generally, inclusion of a random intercept model makes more fine-grained comparisons of models intermediary between those usually considered by researchers which are based on factor dimensionality. For example, in both simulated and real-world data, factor models with random intercept components were estimated and selected which were more complex than the single-factor model but yet more parsimonious than the freely estimated two-factor model. The question of whether the random intercept model or multi-factor measurement model better describe the data also has important implications for the investigation of multigroup invariance. Equality constraints have sometimes been used across Cholesky factors to test for invariance across groups (e.g., Loehlin & Martin 2013). Different hypotheses about equality constraints of factor loadings and component variability are implied under the RI model, however, suggesting that different conclusions about partial invariance may be made under the RI and Cholesky factor models.

Model Support Varies as a Function of Study Design

When such comparisons are considered across studies in an area, such model comparisons provide statements of what measurement models seem reasonable based on characteristics of the study. When the statistical power of the data is low (i.e., when effect size is small or small sample sizes are analyzed), researchers are more likely to find the tau equivalent measurement and its associated standard error are a parsimonious summary. In the body mass data, such intercept factors seemed sufficient to explain variability due to common and unique environmental effects. Congeneric measurement models, by contrast, provide information that markedly different effect sizes across manifest variables have been found. Similarly, multivariate studies of relatively few manifest variables are similarly less likely than studies with several manifest variables to recover random intercept models or multiple factors due to the lower power associated with smaller samples of the multivariate space. Accordingly, the ability to identify method variability, response set, or a complex measurement model such as that underlying a growth process varies as a function of study design.

Limitations

Although use of Bayesian estimation for genetic modeling has promise, it is not without its difficulties. The bimodality of estimated factor loadings across MCMC replications was one difficulty most often encountered when estimated factor loadings were modest and successive MCMC iterations varied between small positive and equally well-fitting small negative values. This problem is equivalent to the reflection problem in factor analysis in general (i.e., that a factor model with loadings multiplied by -1 fits the data as well as the original factor loadings). In Bayesian analysis, researchers can detect the resulting bimodal posterior distribution using different starting values across chains. If the estimated loading is not far away from zero, the convergence criteria or K–S test are unlikely to detect such bimodality. Such bimodality can, however, be remedied by constraining one or more such marginal loadings to be positive across those parameters which appear to exhibit bimodality (Congdon & Congdon 2003). Although such a remedy is appropriate in many situations (Erosheva & Curtis 2013), it should be noted that it is not a universal solution and requires further research (Chan & Jeliazkov 2009).

Future Directions

Use of the model comparison approach outlined here can be readily extended to a greater variety of genetic models for twin and family data. Although the models considered here assumed that the manifest variables were continuously measured variables, extensions of the models presented here to genetic factor models using categorical data (Cho, Wood, & Heath 2009) would appear straightforward, subject to additional identification requirements of the latent response variable approach required for categorical data. Developments in both behavior genetic modeling and Bayesian statistics have also extended structural models using generalized linear mixed models, enabling researchers to specify random effects for variables with other known distributions such as Poisson or other exponential link functions (e.g., Bolker et al. 2009). Additionally, given that multilevel behavior genetic models have also been proposed for genetic data (e.g., Guo & Wang 2002), modeling a random intercept term within a factor model provides the researcher with the ability to assess the relative fit of such models within a factor analytic framework, with the added benefit that the multilevel models can be modeled as special cases of the general genetic models considered here, as they result when factor loadings are constrained to fixed values. Finally, as mentioned above, the random intercept model can also be extended to the case of estimation of growth curve models, although the psychometric measurement alternatives are slightly more complex in those situations.

Conclusion

The exploration of more fine-grained model comparisons motivated by psychometric models for the environmental and genetic components of behavior genetic models appears promising when Bayesian estimation is considered. Bayesian models appear less susceptible to problems of empirical under-identification frequently encountered under ML estimation. The tau equivalent and random intercept models in particular appear to be two parsimonious alternatives to the factor components usually considered under a Cholesky decomposition or other exploratory factor approaches. Although care must be taken to assure that estimation difficulties related to multi-modality and serial correlation in the MCMC estimation procedure are identified and remedied, use of the Bayes factor appears to be a promising means for assessing the relative support of candidate psychometric behavior genetic models.

Electronic supplementary material

Below is the link to the electronic supplementary material.Mplus Program for Fitting Bayesian One-Factor ACE model (DOCX 21 kb)Mplus Program for Fitting Final Bayesian Random Intercept Model for Simulated Data (DOCX 25 kb)

References

Albert, J. H., & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association, 422*, 669–679.

Alessandri, G., Caprara, G. V., & Tisak, J. (2012). A unified latent curve, latent state-trait analysis of the developmental trajectories and correlates of positive orientation. *Multivariate Behavioral Research, 47*, 341–368. doi:10.1080/00273171.2012.673954.

Allison, D. B., Kaprio, J., Korkeila, M., Koskenvuo, M., Neale, M. C., & Hayakawa, K. (1996). The heritability of body mass index among an international sample of monozygotic twins reared apart. *International Journal of Obesity, 20*, 501–506.

Bolker, B., Brooks, M., Clark, C., Geange, S., Poulsen, J., Stevens, M., et al. (2009). Generalized linear mixed models: A practical guide for ecology and evolution. *Trends in Ecology & Evolution, 24*(3), 127–135. doi:10.1016/j.tree.2008.10.008.

Boomsma, A. (1982). The robustness of LISREL against small sample sizes in factor analysis models. In K. G. Jöreskog & H. Wold (Eds.), *Systems under indirect observation: Causality, structure, prediction* (pp. 149–173). Amsterdam: North-Holland.

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika, 71*, 425–440.

Carey, G., & DiLalla, D. L. (1994). Genetics, personality, and psychopathology. *Journal of Abnormal Psychology, 103*, 32–43.

Carey, G., Goldsmith, H. H., Tellegen, A., & Gottesman, I. I. (1978). Genetics and personality inventories: The limits of replication with twin data. *Behavior Genetics, 8*(4), 299–313.

Chan, J. C. C., & Jeliazkov, I. (2009). Efficient simulation and integrated likelihood estimation in state space models. *International Journal of Mathematical Modelling and Numerical Optimisation, 1*(1), 101–120.

Chang, Z., Lichtenstein, P., Asherson, P. J., & Larsson, H. (2013). Developmental twin study of attention problems: High heritabilities throughout development. *JAMA Psychiatry, 70*(3), 311–318.

Cho, S. B., Wood, P. K., & Heath, A. (2009). Decomposing group differences of latent means of ordered categorical variables with the genetic factor model. *Behavior Genetics, 39*, 101–122.

Chou, C. P., Bentler, P. M., & Satorra, A. (1991). Scaled test statistics and robust standard errors for non-normal data in covariance structure analysis: A Monte Carlo study. *British Journal of Mathematical and Statistical Psychology, 44*, 347–357.

Congdon, P., & Congdon, P. (2003). *Applied Bayesian modelling* (Vol. 394). New York: Wiley.

Davis-Stober, C. P. (2011). A geometric analysis of when fixed weighting schemes will outperform ordinary least squares. *Psychometrika, 76*(4), 650–669.

Dolan, C., Molenaar, P., & Boomsma, D. (1989). LISREL analysis of twin data with structured means. *Behavior Genetics, 19*(1), 51–62.

Dolan, C. V., Molenaar, P. C., & Boomsma, D. I. (1992). Decomposition of multivariate phenotypic means in multigroup genetic covariance structure analysis. *Behavior Genetics, 22*(3), 319–335.

Dolan, C. V., Molenaar, P. C. M., & Boomsma, D. I. (1994). Simultaneous genetic analysis of means and covariance structure: Pearson-Lawley selection rules. *Behavior Genetics, 24*, 17–24.

Eaves, L., Erkanli, A., Silberg, J., Angold, A., Maes, H. H., & Foley, D. (2005). Application of Bayesian inference using Gibbs sampling to item-response theory modelling of multi-symptom genetic data. *Behavior Genetics, 35*(6), 765–780. doi:10.1007/s10519-005-7284-z.

Elks, C. E., Den Hoed, M., Zhao, J. H., Sharp, S. J., Wareham, N. J., Loos, R. J., et al. (2012). Variability in the heritability of body mass index: A systematic review and meta-regression. *Frontiers in Endocrinology, 3*, 29. doi:10.3389/fendo.2012.00029.

Erosheva, E. A., & Curtis, S. M. (2013). Dealing with rotational invariance in Bayesian confirmatory factor analysis. Technical report #589, Seattle, WA: Department of Statistics, University of Washington. http://www.stat.washington.edu/research/reports/2011/tr589.pdf

Franić, S., Dolan, C. V., Borsboom, D., Hudziak, J. J., van Beijsterveldt, C. E. M., & Boomsma, D. I. (2013). Can genetics help psychometrics? Improving dimensionality assessment through genetic factor modeling. *Psychological Methods, 18*, 406. doi:10.1037/a0032755.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., & Vehtari, A. (2013). *Bayesian data analysis* (3rd ed.). New York: CRC press.

Gelman, A., & Meng, X. L. (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical Science, 13*(2), 163–185.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 5*(6), 721–741. doi:10.1109/TPAMI.1984.4767596.

Grimm, K. J., & Ram, N. (2009). Nonlinear growth models in Mplus and SAS. *Structural Equation Modeling, 16*(4), 676–701. doi:10.1080/10705510903206055.

Guo, G., & Wang, J. (2002). The mixed or multilevel model for behavior genetic analysis. *Behavior Genetics, 32*(1), 37–49.

Heath, A. C., Eaves, L. J., & Martin, N. G. (1989). The genetic structure of personality III. Multivariate genetic item analysis of the EPQ scales. *Personality and Individual Differences, 10*(8), 877–888.

Heath, A. C., Jardin, R., Eaves, L. J., & Martin, N. G. (1989). The genetic structure of personality II: Genetic item analysis of the EPQ. *Personality & Individual Differences, 10*, 615–624.

Heath, A., Neale, M., Hewitt, J., Eaves, L., & Fulker, D. (1989). Testing structural equation models for twin data using LISREL. *Behavior Genetics, 19*(1), 9–35.

Henderson, N. D. (1982). Human behavior genetics. *Annual Review of Psychology, 33*, 403–440.

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling: An overview and a meta-analysis. *Sociological Methods and Research, 26*, 329–367.

Hu, L., Bentler, P. M., & Kano, Y. (1992). Can test statistics in covariance structure analysis be trusted? *Psychological Bulletin, 112*, 351–362.

Joseph, J., & Ratner, C. (2013). The fruitless search for genes in psychiatry and psychology: Time to re-examine a paradigm. In S. Krimsky & J. Gruber (Eds.), *Genetic explanations: Sense and nonsense* (pp. 94–106). Cambridge, MA: Harvard University Press.

Kenny, D. A., Kashy, D. A., & Bolger, N. (1998). Data analysis in social psychology. In D. Gilbert, S. Fiske, & G. Lindsey (Eds.), *Handbook of social psychology* (4th ed., Vol. 1, pp. 233–265). Boston: McGraw-Hill.

Kenny, D. A., & Milan, S. (2013). Identification: A non-technical discussion of a technical issue. In R. Hoyle, D. Kaplan, G. Marcoulides, & S. West (Eds.), *Handbook of structural equation modeling* (pp. 145–163). New York: Guilford.

Lee, S. Y. (1980). Estimation of covariance structure models with parameters subject to functional restraints. *Psychometrika, 45*(3), 309–324.

Lee, S. Y. (2007). *Structural equation modelling: A Bayesian approach*. New York: John Wiley.

Lee, S. Y., & Song, X. Y. (2004). Evaluation of the Bayesian and maximum likelihood approaches in analyzing structural equation models with small sample sizes. *Multivariate Behavioral Research, 39*(4), 653–686.

Lindley, D. V. (1977). A problem in forensic science. *Biometrika, 64*(2), 207–213.

Loehlin, J. C. (1996). The Cholesky approach: A cautionary note. *Behavior Genetics, 26*, 65–69.

Loehlin, J. C., & Martin, N. G. (2013). General and supplementary factors of personality in genetic and environmental correlation matrices. *Personality and Individual Differences, 54*, 761–766.

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

Ludeke, S., Johnson, W., & Bouchard, T. J. (2013). "Obedience to traditional authority:" A heritable factor underlying authoritarianism, conservatism and religiousness. *Personality and Individual Differences, 55*, 375–380.

Martin, N. G., & Eaves, L. J. (1977). The genetical analysis of covariance structure. *Heredity, 38*(1), 79–95.

Martin, N., Scourfield, J., & McGuffin, P. (2002). Observer effects and heritability of childhood attention-deficit hyperactivity disorder symptoms. *British Journal of Psychiatry, 180*, 260–265.

Maydeu-Olivares, A., & Coffman, D. L. (2006). Random intercept item factor analysis. *Psychological Methods, 11*(4), 344.

Meng, X. L. (1994). Posterior predictive p-values. *The Annals of Statistics, 22*(3), 1142–1160.

Meredith, W., & Tisak, J. (1990). Latent curve analysis. *Psychometrika, 55*(1), 107–122.

Muthén, B. (2010). *Bayesian analysis in Mplus: A brief introduction*. Retrieved from https://www.statmodel.com/download/IntroBayesVersion%203.pdf

Muthén, B., & Asparouhov, T. (2011). Bayesian SEM: A more flexible representation of substantive theory. *Psychological Methods, 17*(3), 313–335. doi:10.1037/a0026802.

Muthén, L. K., & Muthén, B. O. (1998–2010). *Mplus user's guide* (6th ed.). Los Angeles, CA: Muthén & Muthén.

Neale, M. C., & Cardon, L. R. (1992). *Methodology for genetic studies of twins and families*. New York: Springer.

Phillips, K., & Matheny, A. P. (1997). Evidence for genetic influence on both cross-situation and situation-specific components of behavior. *Journal of Personality and Social Psychology, 21*(1), 129–138.

Ram, N., & Grimm, K. J. (2009). Growth mixture modeling: A method for identifying differences in longitudinal change among unobserved groups. *International Journal of Behavioral Development, 33*(6), 565–576. doi:10.1177/0165025409343765.

Rietveld, M. J., Posthuma, I. D., Dolan, C. V., & Boomsma, D. I. (2003). ADHD: Sibling interaction or dominance: An evaluation of statistical power. *Behavior Genetics, 33*(3), 247–255.

Rindskopf, D. (1984). Structural equation models. *Sociological Methods & Research, 13*(1), 109–119.

Roberson-Nay, R., Moruzzi, S., Ogliari, A., Pezzica, E., Tambs, K., Kendler, K. S., et al. (2013). Evidence for distinct genetic effects associated with response to 35% CO_2. *Depression and Anxiety, 30*(3), 259–266. doi:10.1002/da.22038.

Rodgers, J. L. (2010). The epistemology of mathematical and statistical modeling: A quiet methodological revolution. *American Psychologist, 65*(1), 1–12. doi:10.1037/a0018326.

Sato, M. (1987). Pragmatic treatment of improper solutions in factor analysis. *Annals of the Institute of Statistical Mathematics, 39*(1), 443–455.

Savalei, V., & Kolenikov, S. (2008). Constrained versus unconstrained estimation in structural equation modeling. *Psychological Methods, 13*(2), 150–170.

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*(2), 461–464.

Shi, J. Q., & Lee, S. Y. (1998). Bayesian sampling-based approach for factor analysis models with continuous and polytomous data. *British Journal of Mathematical and Statistical Psychology, 51*(2), 233–252.

Sörbom, D. (1974). A general method for studying differences in factor means and factor structure between groups. *British Journal of Mathematical and Statistical Psychology, 27*, 229–239.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B, 64*, 583–639.

Turkheimer, E. (2000). Three laws of behavior genetics and what they mean. *Current Directions in Psychological Science, 9*(5), 160–164.

Visscher, P. M., Gordon, S., & Neale, M. C. (2008). Power of the classical twin design revisited: II Detection of common environmental variance. *Twin Research and Human Genetics, 11*, 48–54.

Waldron, M., Bucholz, K. K., Lynskey, M. T., Madden, P. A. F., & Heath, A. C. (2013). Alcoholism and timing of separation in parents: Findings in a Midwestern birth cohort. *Journal of Studies on Alcohol and Drugs, 74*, 337–348.

Zhang, Z., Hamagami, F., Wang, L., Nesselroade, J. R., & Grimm, K. J. (2007). Bayesian analysis of longitudinal data using growth curve models. *International Journal of Behavioral Development, 31*(4), 374–383. doi:10.1177/016502540707776.