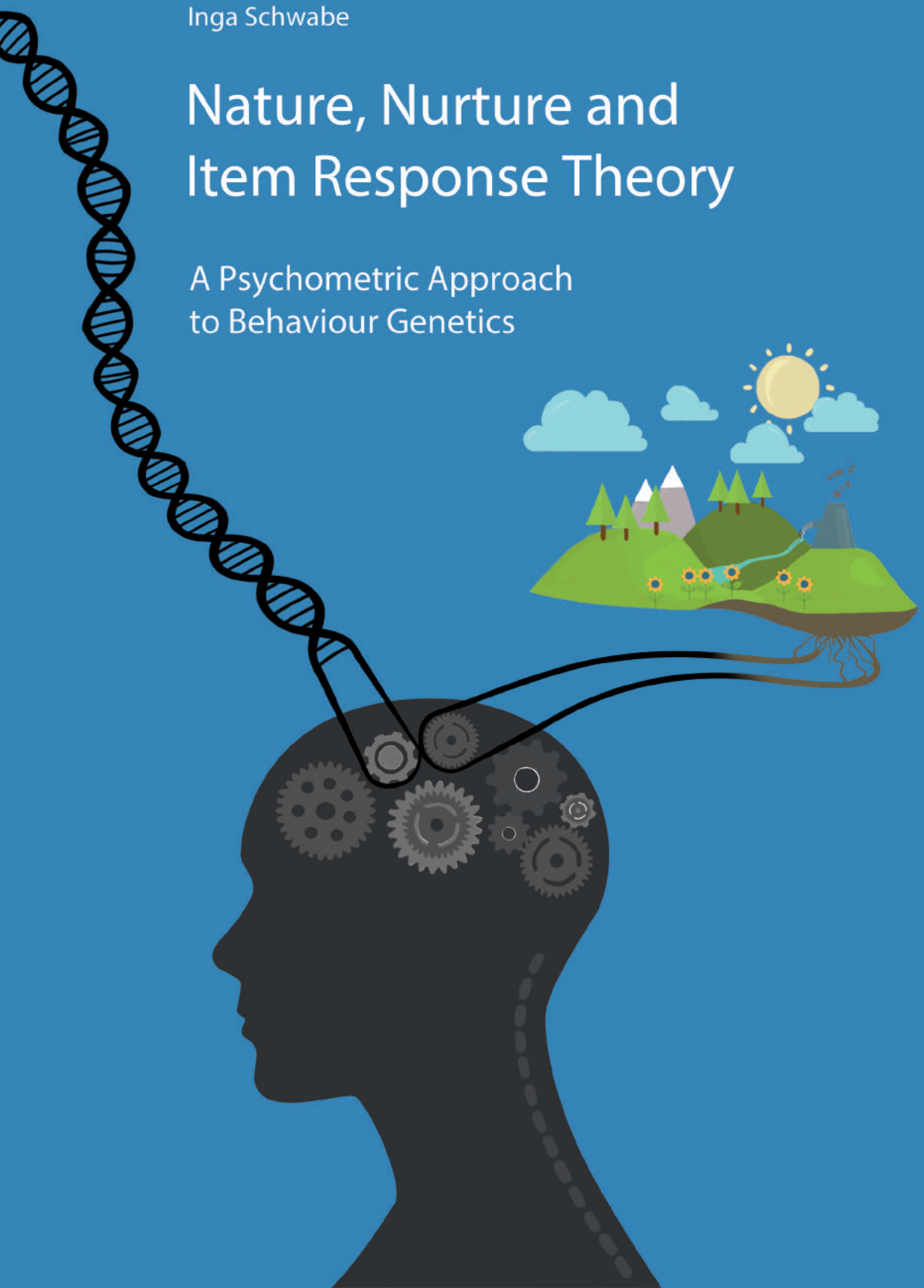
Inga Schwabe

Nature, Nurture and Item Response Theory

A Psychometric Approach to Behaviour Genetics



Nature, Nurture and Item Response Theory

A Psychometric Approach to Behaviour Genetics

Inga Schwabe

Graduation Committee:

Chair	Prof. dr. T.A.J. Toonen
Promotor	Prof. dr. C.A.W. Glas
Assistant promotor	Dr. S.M. van den Berg
Members	Prof. dr. M. Bartels
	Prof. dr. I. Klugkist
	Dr. G.H. Lubke
	Dr. S. van der Sluis
	Prof. dr. J.H. Walma van der Molen

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

Schwabe, Inga
Nature, Nurture and Item Response Theory - A Psychometric Approach to Behaviour Genetics
PhD Thesis University of Twente, Enschede. - Met samenvatting in het Nederlands.
ISBN: 978-90-365-4073-5
doi: 10.3990/1.9789036540735
Printed by Ipskamp Printing, Enschede
Cover design and illustration: Inga Schwabe

Nature, Nurture and Item Response Theory
A Psychometric Approach to Behaviour Genetics

Dissertation

to obtain
the degree of doctor at the University of Twente,
on the authority of the rector magnificus,
Prof. dr. H. Brinksma,
on account of the decision of the graduation committee
to be publicly defended
on Thursday, March 24th, 2016 at 14:45

by

Inga Schwabe

born on June 29th, 1988
in Oldenburg, Germany

This dissertation is approved by the following promotores:

Promotor: Prof. dr. C.A.W. Glas
Assistant promotor: Dr. S.M. van den Berg

Contents

CHAPTER 1

Introduction

One of psychology's defining questions involves the origin of individual differences in behaviour: Why are some people happy and other depressed? Why do some children seem to be born to solve mathematical equations while other struggle to pass exams? The *nature-nurture debate* is concerned with the extent to which these differences are inherited (i.e., genetic) or acquired (i.e., learned, environmental). *Behaviour genetics*, a field within psychology, aims to provide insights into this debate by studying the relative importance of genetic and environmental influences in explaining variability in a trait. Their variance can be inferred from resemblance among family members. One of the methods that adopts this approach is the twin design, which compares resemblance in identical (monozygotic, MZ) and non-identical (dizygotic, DZ) twin pairs. MZ twin pairs share the exact same genomic sequence and the same rearing environment, including prenatal environmental conditions. DZ twins also share the same prenatal and rearing environment but on average only share half of the segregating genes. Based on these known differences in genetic similarity, the relative impact of nature and nurture can be estimated by comparing covariance in MZ and DZ twin pairs. When MZ twins are more similar in a trait (i.e., phenotype) than DZ twins, this implies that genetic influences are important.

1.1 Genetic models

In a typical twin study, the total variance in a trait (e.g. mathematical ability) is assessed in a large and representative sample of twins. The variance, referred to as phenotypic variance (σ^2_P) is then decomposed into a number of variance components. In the AE decomposition, phenotypic variance is decomposed into parts due to additive genetic (A) and unique-environmental (E) influences whereas the ACE model also estimates variance

due to common-environmental (C) influences (Jinks & Fulker, 1970). A graphical representation of the ACE model in structural equation model (SEM) notation can be found in Figure 1.1. Additive genetic influences refer to the total effect on a trait stemming from all gene loci. Common-environmental influences are shared influences (e.g., familial influences) and are parametrized as being perfectly correlated in a twin pair whereas unique-environmental influences are unique to a twin and parametrized as being uncorrelated within a twin pair. It is also possible to fit an ADE model in which the C component is replaced by a D component to estimate dominance effects (non-additive genetic influences). Dominance effects arise when (part of) the inheritance of a trait is governed by dominant genetic mechanisms − a dominant gene inherited from one parent trumps a recessive gene inherited from the other parent. For example, when a twin inherits a recessive gene for blue eyes from the mother and a dominant gene for brown eyes from the father, then the dominant gene determines the trait and the twin's eyes are brown.

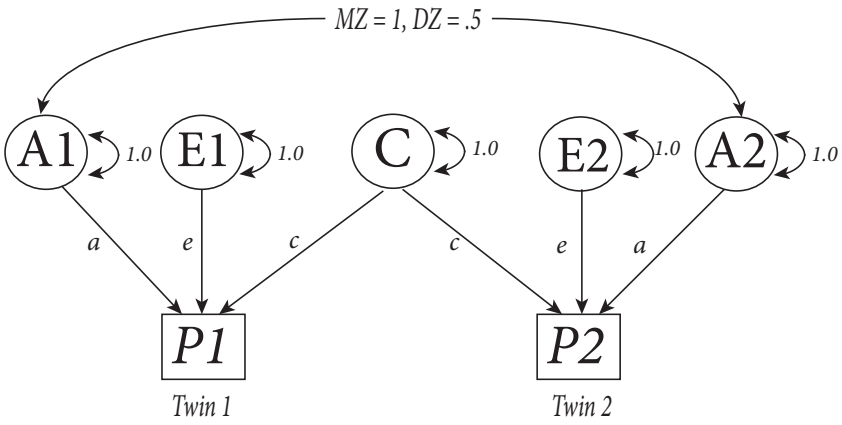


Figure 1.1: The ACE model in structural equation model (SEM) notation. P denotes the phenotypic values of the first ($P1$) and second ($P2$) twin and A refers to additive genetic influences for the first ($A1$) and second ($A2$) twin, which are correlated 0.5 in DZ twin pairs and 1 in MZ twin pairs. $E1$ and $E2$ denote unique-environmental influences of the first and second twin respectively and are assumed to be uncorrelated. C, common-environmental influences, are the same for the first and second twin of a twin pair. Double-headed arrows denote (co-)variances. The path coefficients a, c and e represent regression coefficients that express the estimated effect of the respective influences.

1.1.1 Genotype-environment interaction

Research in the field of behaviour genetics has shown that genetic influences make a substantial contribution to individual trait differences while the part of the variance that is explained by common-environmental influences is much smaller. A non-trivial proportion of the variance can be attributed to unique-environmental influences. These findings are supported by an extensive literature and so universal that Turkheimer (2000) coined them as the *"three laws of Behaviour Genetics"*. These laws apply to a broad range of observable traits such as mathematical ability, one's well-being or political attitudes to name only a few examples.

However, it is also generally acknowledged that a simple distinction into "nature" on the one hand and "nurture" on the other hand is often too simplistic to explain individual differences in a trait. Research has shown that they often go hand in hand - a phenomenon that is referred to as *genotype-environment interaction* formerly. For example, research suggests that additive genetic influences on depression interact with marital status in women, where genetic influences are more important for unmarried women (Heath, Eaves & Martin, 1998). Another well-known finding is that genetic influences on IQ are more important in families with a high socioeconomic status (e.g. Turkheimer, Haley, Waldron, D'Onofrio & Gottesman, 2003). Genotype-environment interaction has also been found in the development of depression (e.g. Hicks, DiRago, Iacono & McGue, 2009; Lau & Eley, 2008), physical and mental health (e.g. Johnson & Krueger, 2005; Faith et al., 2004; Kim-Cohen et al., 2006) and antisocial behavior (e.g. Caspi et al., 2002; Cadoret, Cain & Crowe, 1983; Tuvblad, Grann & Lichtenstein, 2006).

1.2 Measurement of behavioural traits

In order to apply the twin design to investigate genotype-environment interaction, we have to measure the trait first. The measurement of a physical trait such as length is easy: Using the measurement tape, we can *directly* measure the length of a person and anyone will agree with us that the result (e.g., 175 centimetres) resembles the physical length. The measurement of a behavioural trait such as mathematical ability, however, is more complicated. That is, behavioural traits can only be measured *indirectly*. To measure mathematical ability, we can use a test that consists of twenty mathematical problems of differing type and difficulty. An individual's solutions to these problems can then be used to obtain a score on the test. For example, for every mathematical problem that was solved, students score one point, assuming that a mathematically talented child should be able to solve all problems whereas one without any mathematical talent can solve only a few problems. The score on such a test, often referred to as the *sum score* is then assumed to resemble mathematical ability, measured *indirectly* by the test questions (items).

Indirectly measured attributes such as mathematical ability are referred to as *latent traits* in the field of psychology. *Psychometrics* is a branch of psychology that is concerned with the measurement of these latent traits. This dissertation approaches the nature-nurture debate from a psychometric angle - that is, it is investigated whether the field of psychometrics can improve research practices in the field of behaviour genetics.

1.3 A psychometric approach to behaviour genetics

There are a number of psychometric issues that require special attention in the analysis of genetically-informative data. These include heterogeneous measurement error, scaling and scale transformations, the handling of missing data and harmonization of phenotypes. In this dissertation, it is shown how ignoring these psychometric issues can lead to biased results and it is demonstrated how item response theory (explained in more detail below), a method from the field of psychometrics, can be used to prevent potential bias. In the following, a short summary is given of the psychometric issues that are addressed in this dissertation.

1.3.1 Heterogeneous measurement error

As most tests consist of a lot of items of average difficulty, it is usually easy to differentiate between average scoring individuals. Often, however, there are only a few very simple or very difficult items. Therefore, tests are not evenly reliable across the entire range of sum scores and it is more difficult to investigate individual differences in very low- or very high-scoring individuals. That is, the measurement error is *heterogeneous* - higher for the left and right tail of the trait continuum. This can result in the finding of spurious genotype-environment interaction effects. In Chapter 2, it is explained why and when this happens and how this problem can be solved. While Chapter 2 is concerned with an *omnibus* test of genotype-environment interaction to assess whether there is any statistically significant interaction, this method is extended to include one or more measured moderator variable(s) in Chapter 5.

1.3.2 Scaling

While most will agree on the scale to measure a person's length, this is not necessarily true for the measurement of psychological traits. For example, what should be the metric to measure a construct like mathematical ability? Should this be a scale from zero to ten or from ten to thirty? Usually, there is no consensus on what scale should be used for a given trait. Likewise, many psychological tests change over time, for example a mathematical ability test with thirty items might be shortened to a test with only ten items, or additional items might be added after a re-evaluation of the scale.

What items are included in a test version, however, has direct impact on the distribution of the sum scores of the different versions of a test. For example, Figure 1.2 shows the distribution of the sum scores of different versions of a scale that was used to measure the school motivation of 4220 individual twins from the Netherlands Twin Register, including the full scale with all items and two different subscales consisting each of only five items.

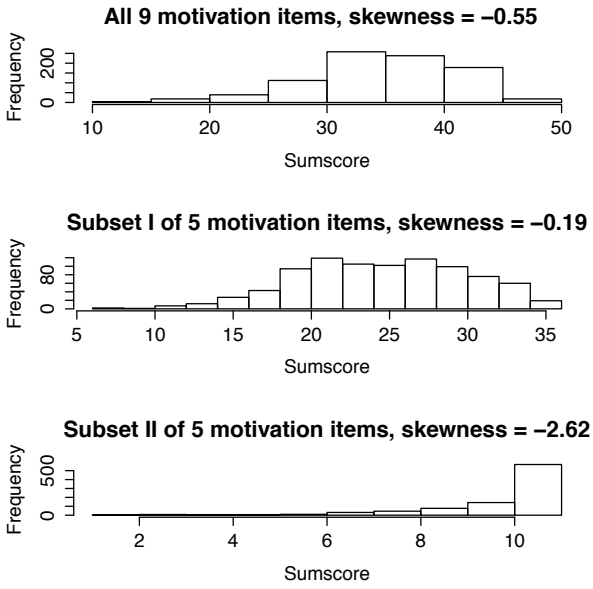


Figure 1.2: Distribution of sum scores on three different versions of the motivation scale: All nine motivation items (skewness=-0.55), subset I of five items (skewness=-0.19) and subset II of five items (skewness=-2.62).

We can see that a different choice of items leads to a different distribution of sum scores and therefore to a different skewness. Given that statistical findings are dependent on the measurement scale, this might mean that, using the same data, researcher A finds a genotype-environment interaction effect while researcher B cannot replicate this effect when she or he uses another scale (e.g., consisting of another subset of items). The methods introduced in Chapter 2 and Chapter 5 model the twin data such that the findings are independent of scale properties, meaning that, as long as a set of items measures a particular trait, biometric results (i.e., conclusions regarding heritability or genotype-environment interaction) are the same regardless of the particular (sub)set of items that is used.

1.3.3 Missing data

Handling missing data is an important topic in the measurement of traits. Due to time limits, a test taker might not reach the end of the test or a respondent might not answer all questions on a questionnaire but for example skip items on sensitive topics (e.g. drug abuse). In case of missing data on a subset of items, a decision needs to be made about the handling of these missing item scores. When sum scores are used, often, one of the following approaches are applied: a) Imputing the respondent's mean response on all available items or b) Imputing the item's mean of the missing item. More complex methods to handle missing data exist, but they are seldom used in the field of behaviour genetics.

A problem of most traditional approaches is that the uncertainty of the imputed values is not taken into account: Standard errors and confidence intervals are calculated as if there were no missing item scores. The item response theory approach (explained in more detail below) provides a flexible method to handle missing item data in which also the uncertainty of estimates is taken into account.

The twin model can be extended to include covariates, which can index (but are not restricted to) common-environmental or unique-environmental influences. The collection of these data can however also lead to missing data. For example, a twin researcher might link twin data from a twin registry to data from the same twin from another (external) source to retrieve covariate data and entities cannot be uniquely linked to a common identifier such as the name or address of a family. Likewise, a questionnaire that is used to gather covariate data might not be fully completed. In the usual approach to handle missing covariate data, only phenotypic and covariate data of individual twins with complete data can be used, leading to reduced power to detect statistical effects. In Chapter 6 of this dissertation, it is shown how *all* observed data can be used by including covariates in the expected covariance matrix of a twin analysis.

1.3.4 Harmonization of phenotypes

In behaviour genetics research, the data of different cohorts or different twin registers are often combined to increase statistical power. However, often not the same test or questionnaire was used in all cohorts or registers. The different test versions may differ with respect to their overall difficulty and as a result, sum scores are not comparable across the different samples. For example, a mathematical ability test used in one twin register might be composed of very difficult items, while the test used in another twin register might be relatively easy. The item response theory approach (explained in more detail below) can be used to harmonize measures such that data from individual twins is comparable. For example, in Chapter 3 and 5, item data on the mathematics subscale of a national educational achievement

test of twins from the Netherlands Twin Register (NTR) was used. As the test was administered using different test versions, sum scores were not comparable across versions. Measures needed to be harmonized such that data from individual twins assessed by a different test version could be compared meaningfully.

1.4 Item response theory

To tackle above described psychometric issues, we depart from earlier work in twin research by modelling raw item data instead of sum scores. Item data is modelled using the item response theory (IRT) approach that will be explained in more detail in the following.

An indirect assumption of the sum score approach is that every item measures the trait equally well and is equally difficult to answer. Whereas this traditional approach thus ignores properties of the items, these are explicitly modelled in the IRT approach. The IRT approach is model-based measurement in which a person's latent trait level on a certain scale (e.g. mathematical ability), is estimated using not only trait levels (e.g., an individuals performance on a test), but also test item properties such as the difficulty of each item. So, both, performance as well as item properties, are used as information to be incorporated into the scaling of individual test performance. The simplest IRT model is the Rasch model, also known as the one-parameter logistic model (1PLM). This IRT model is suitable for dichotomous data (e.g., scored as correct = 1 and false = 0), as for example collected from ability tests. In the Rasch model, the probability of a correct answer to item k (e.g. on a mathematics test) by twin j from family i, $P(Y_{ijk} = 1)$, is modelled as a logistic function of the difference between the twin's latent trait score (e.g., representing mathematical ability) and the difficulty of the item:

$$P(Y_{ijk} = 1) = \frac{\exp(\theta_{ij} - b_k)}{1 + \exp(\theta_{ij} - b_k)} \tag{1.1}$$

where θ_{ij} represents the latent trait (e.g., mathematical ability) score of individual twin j from family i such that, in case of mathematical ability, a twin with a high latent trait score has a high mathematical ability. A higher latent trait score results in a higher probability to answer the item correctly. Parameter b_k represents the difficulty of item k which is parametrized as the trait level associated with a 50% chance of answering the item correctly. When the difficulty b of an item k increases, the probability of answering the item correctly decreases.

The IRT approach can be illustrated by means of *item characteristic curves* which display the probability of a correct response as a function of the latent trait scores (abilities). The item characteristic curves of two items with different difficulty can be seen in Figure 1.3. The left-hand

curve represents an easier item because the probability of a correct response is higher for low-ability twins than it is in case of the second item. It furthermore approaches a probability of 1 for a correct response faster than the curve of the right-handed item does.

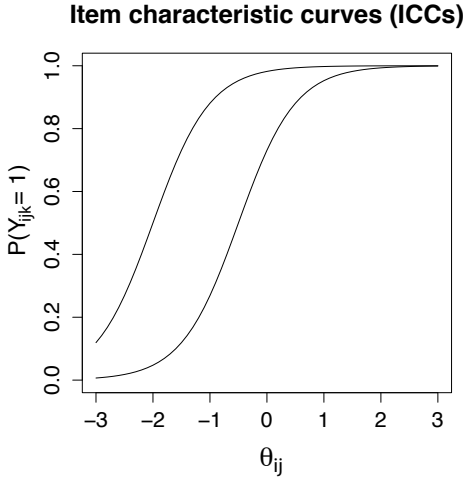


Figure 1.3: Two item characteristic curves (ICCs) for items with the same discrimination but different levels of difficulty (based on simulated test data).

An underlying assumption of the Rasch model is that all items discriminate equally well between varying abilities. An extension of the Rasch model, the two-parameter model (2PLM), estimates discrimination parameters (comparable to factor-loadings) that differ across items (see e.g. Embretson and Reise, 2009). There are several IRT models that can be used for non-dichotomous data such as ordered categories (e.g. Likert scale data). In this dissertation, both kinds of IRT models (suitable for dichotomous and non-dichotomous data respectively) were applied.

A large part of this dissertation was devoted to the development of new methodology in which both IRT and genetic model are estimated *simultaneously*. In Chapter 2, the IRT approach is used to model genotype-environment interaction at the latent level of the phenotype. In Chapter 5, this method is extended to include one or more measured variable(s) to model moderation of variance decomposition at item level. Another part of this dissertation was concerned with applications of the new methodology, summarized shortly in the following.

1.5 Applications

In a collaboration with the Netherlands Twin Register at the VU University and the psychometric group of Cito, the twin data from a subset of the NTR were linked to their item scores on the mathematics subscale of a Dutch national educational achievement test (*Eindtoets Basisonderwijs*) that is administered by Cito yearly in the last year of primary school. In Chapter 3, the method that is introduced in Chapter 2, is applied to these item scores to investigate genotype-environment interaction in mathematical ability while correcting for heterogeneity in the measurement of mathematics performance through the application of an IRT model.

In Chapter 4, item data of twins and their parents from the Health and Life-Style Survey for Twins assessed in the Virginia 30K sample (Eaves et al., 1999; Hatemi et al., 2009) on the Wilson-Patterson conservatism scale are used to psychometrically evaluate this scale. Based on the results, a new scale is devised and used to investigate genotype-environment interaction, extending the method introduced in Chapter 2 to ordinal data.

The method introduced in Chapter 5 that incorporates an IRT model into the modelling of variance decomposition moderation, is applied to the the data of 2110 12-year old Dutch twin pairs to test moderating effects of a family's socio-economic status on individual differences in mathematical ability. As in Chapter 3, the twins' item scores on the mathematics subscale of the *Eindtoets Basisonderwijs* were used. The method that is introduced in Chapter 6 to model missing covariate data is applied to the test scores on the *Eindtoets Basisonderwijs* test of 990 twin pairs to investigate the effects of school-aggregated measures and the sex of a twin on these scores.

CHAPTER 2

Assessing Genotype by Environment Interaction in Case of Heterogeneous Measurement Error

Based on:
Inga Schwabe and Stéphanie M. van den Berg
Behavior Genetics, 44(4), 394-406

Abstract

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

2.1 Introduction

Genotype by environment interaction (henceforth referred to as G×E) in its conceptual sense means either that different genotypes respond differently to the same environment or that some genotypes are more sensitive to changes in the environment than others (Cameron, 1993; Martin, 2000; Sorensen, 2010). In the last decade, the assessment of G×E has received increasing attention in twin and family studies (Dick, 2011). Various studies have found evidence for the presence of G×E. In the context of educational achievement, Friend et al. (2009) report an interaction between high reading ability and the education of the parents: The heritability of high reading ability was higher for twins when parents were less well educated. Another well-known finding is that heritability of cognitive ability varies with socioeconomic status (Turkheimer et al., 2003; Harden, Turkheimer & Loehlin, 2007). G×E seems also present for non-cognitive traits. To name a few examples, G×E has been found in the development of depression (Hicks et al., 2009; Lau & Eley, 2008; Bukowski et al., 2009), physical and mental health (Johnson & Krueger, 2005; Faith et al., 2004; Kim-Cohen et al., 2006) and antisocial behaviour (Caspi et al., 2002; Cadoret et al., 1983; Tuvblad et al., 2006). Arguably, G×E is an important phenomenon in complex behavioural traits.

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

van der Sluis et al. (2006) proposed an alternative method, using MZ twin data and an exponential function to model G×E (cf. SanChristobal-Gaudy, Elsen, Bodin & Chevalet, 1998). Molenaar et al. (2012) extended this work by including dizygotic (DZ) twin data and modelling G×E for both shared and non-shared environmental variance separately. Furthermore, they extended the univariate approach to a multivariate approach.

2.1.1 G×E in case of heterogeneous measurement error

There is however one problem in the assessment of G×E that is not tackled by any of the above mentioned methods. In a behaviour genetics study, one is typically interested in the origins of observed variance in a phenotypic trait. To this end, often a number of items is presented to respondents.

Next, the subject's sum score on the items is computed, assuming that the unweighted summed score can be treated as a proxy for the trait. The variance of the computed sum scores is then decomposed into a number of variance components. In a so-called AE decomposition, the variance is decomposed into parts due to additive genetic (A) and unique-environmental (E) influences, whereas the so-called ACE model also estimates variance due to common-environmental (C) influences (Jinks & Fulker, 1970).

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

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

among other issues stressing the sensitivity of G×E to properties of the measurement scale. This notion has been accentuated by various different authors since then (Martin, 2000; van der Sluis et al., 2006; Eaves, 2006; Molenaar et al., 2012).

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

2.1.2 Towards a solution

Heterogeneous measurement error can be accounted for by explicitly modelling the properties of a scale. This can be done by incorporating an Item Response Theory (IRT) measurement model into the variance decomposition. In IRT models, item scores depend not only on a person's trait level (e.g. intelligence), but also on the properties of the items that were administered (e.g. difficulty). van den Berg, Glas and Boomsma (2007) extended the usual AE/ACE variance decomposition with an IRT measurement model. They showed that the simultaneous estimation of an IRT measurement model and a biometric model produced unbiased estimates for heritability coefficients and dominance genetic variance, unlike the sum score approach. Also the proposed method by Molenaar et al. (2012) incorporated a measurement model. They linked observed item variables first to the underlying construct using a linear factor model and then (in the biometric part of the model) decomposed the phenotypic variances into parts due to additive genetic, common-environmental and unique-environmental influences. Heteroscedastic residual variances were incorporated in the measurement model to account for possible measurement problems at the level of the observed variables. This led to the absorption of possible floor and ceiling effects and poor scaling effects in the residuals, while the effects of actual genotype by environment interaction were detected in the latent biometrical part of the model. As a factor model was used, the approach is limited to continuous data and cannot be used for dichotomous items (e.g. scored as correct/false). This limitation can be overcome by the combination of an IRT measurement model and a biometric model.

Here, we propose a method that extends the van den Berg et al. (2007) model for dichotomous and polytomous data with a G×E interaction effect. Simulation study 1 illustrates that the method is superior to the sum score approach, in that the sum score approach leads to spurious G×E, whereas

parameter estimates are unbiased with the proposed method. The statistical power of the suggested method to detect actual G×E is illustrated with simulation study 2 using different G×E effect sizes and sample sizes.

2.2 Biometric model

The ACE model decomposes observed variance in a phenotypic variable, denoted as σ^2_P, into parts due to additive genetic influences (σ^2_A), common-environmental influences (σ^2_C) and unique-environmental influences (σ^2_E).

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

$$\sigma^2_{Ej} = \exp(\beta_0 + \beta_1 A_j) \tag{2.1}$$

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

2.3 Measurement model

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

$$\ln(P_{jk}/(1 - P_{jk})) = \theta_j - b_k \tag{2.2}$$

The 1PLM is suitable for dichotomous data, as for example data collected from ability tests where item responses are commonly scored correct/false. In the 1PLM, all items are assumed to have the same correlation (factor loading)

with the underlying latent trait. That is, all items discriminate equally well between the various levels of the latent trait. It is also possible to estimate factor loadings that differ across items (in the IRT framework referred to as discrimination parameters α_k), which turns the 1PLM into a two-parameter model (2PLM) (see e.g. Embretson & Reise, 2009). Furthermore, there are several IRT models that are suitable for ordered categories, as for example Likert scale data (see e.g. Samejima, 1969; Masters, 1982; Embretson & Reise, 2009). In this paper, the 1PLM was used, but extension to other models is straightforward. In case of the 2PLM model for example, the equation changes to:

$$\ln(P_{jk}/(1 - P_{jk})) = \alpha_k \ (\theta_j - b_k) \tag{2.3}$$

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

2.4 Incorporation of the measurement model into the biometric model

van den Berg et al. (2007) showed that, in order to take full advantage of the IRT approach, both the IRT measurement model and the variance decomposition model have to be estimated simultaneously, using a one-step approach. However, as this procedure is computationally burdensome, widespread methods of estimating variance components through structural equation modelling (SEM) reach their computational limit. van den Berg et al. (2007) illustrated that Bayesian statistical modelling with Markov chain Monte Carlo (MCMC) estimation is a good alternative. In a Bayesian analysis, statistical inference is based on the joint posterior density of the model parameters, which is proportional to the product of a prior probability and the likelihood function of the data (see e.g. Box & Tiao, 1973). When analytically deriving the posterior distribution is difficult or impossible, Gibbs sampling (Geman & Geman, 1984; Gelfand & Smith 1990; Gelman et al. 2004) can be applied. Here, the MCMC estimation was implemented in the freely obtainable MCMC software package JAGS (Plummer, 2003). The JAGS script can be found in Appendix A. The script can also be used in the free software package WinBUGS (Lunn, Thomas, Best & Spiegelhalter, 2000).

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

MZ twins

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

$$C_i \sim N(\mu, \sigma_C^2) \tag{2.4}$$

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

$$F_i \sim N(C_i, \sigma_A^2) \tag{2.5}$$

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

$$A_i = F_i - C_i \tag{2.6}$$

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

$$\theta_{ij} \sim N(F_i, \sigma_{Ei}^2) \tag{2.7}$$

To introduce G×E, the twin pair i specific error variance, σ_{Ei}^2, reflecting unique-environmental influences, has to be portioned into an intercept and a scale parameter (see Equation 2.1), resulting in a variance of σ_E^2 that is different for each twin pair i:

$$\sigma_{Ei}^2 = \exp(\beta_0 + \beta_1 A_i) \tag{2.8}$$

Simultaneous with the biometric model above, the latent phenotype θ_{ij} appears in the 1PL IRT model for observed item data Y (see Equation 2.2):

$$\ln(P_{ijk}/(1 - P_{ijk})) = \theta_{ij} - b_k \tag{2.9}$$

$$Y_{ijk} \sim Bernoulli(P_{ijk}) \tag{2.10}$$

DZ twins

As for MZ twin pairs, a normally distributed common-environmental effect was assumed that is the same for both twins (see Equation 2.4). While the total genetic variance is the same for DZ and MZ twins, the genetic covariance in MZ twins is twice as large as in DZ twins, assuming random mating. To model a genetic correlation of 0.5 for DZ twins, first a normally

distributed familial effect F_0 is assumed with variance $\frac{1}{2}\sigma_A^2$ (cf. Jinks & Fulker, 1970):

$$F_{0i} \sim N(C_i, \frac{1}{2}\ \sigma_A^2) \tag{2.11}$$

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

$$F_{1ij} \sim N(F_{0i}, \frac{1}{2}\ \sigma_A^2) \tag{2.12}$$

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

$$A_{ij} = F_{1ij} - C_i \tag{2.13}$$

Similar to Equation 2.7 for MZ twins, the ACE decomposition is complete with

$$\theta_{ij} \sim N(F_{1ij}, \sigma_{Eij}^2) \tag{2.14}$$

with the difference that the additive genetic effect is different for each twin. To incorporate G×E into the model, σ_E^2 has to be portioned into different parts, similar to Equation 2.8 (MZ pairs). Doing so results in an estimate of σ_E^2 that is different for each individual twin:

$$\sigma_{Eij}^2 = \exp(\beta_0 + \beta_1 A_{ij}) \tag{2.15}$$

Again, simultaneous to the ACE decomposition the latent phenotype θ_{ij} appears in the 1PLM IRT part of the model (see Equations 2.9 and 2.10).

2.4.1 Prior distributions

With a Bayesian approach, prior distributions have to be made explicit. We use inverse gamma distributions for the additive genetic variance and the common-environmental variance ($\sigma_A^2 \sim InvG(1,1)$, $\sigma_C^2 \sim InvG(1,1)$). These distributions were chosen because they are both flexible and conjugate. In Bayesian probability theory, a prior is called conjugate when the probability distribution of the prior and the posterior distribution have similar forms (in this case the gamma distribution). This results in convenient sampling, speeding up the estimation process. The prior for the intercept and the slope parameter can be assumed normal ($\beta_0 \sim N(0,1)$, $\beta_1 \sim N(0,10)$), resulting in relatively and reasonably flat priors in this particular application. When item parameters are known, the phenotypic population mean has to be estimated, which can also be given a normal prior distribution ($\mu \sim N(0,10)$). When item parameters are not known but estimated, the phenotypic population mean should be fixed (e.g., $\mu = 0$) to identify the scale.

2.5 Simulation study 1

To illustrate that the sum score approach can lead to the finding of spurious G×E whereas the proposed method is unbiased, a simulation study was conducted. One hundred datasets were generated consisting of 360 DZ twin pairs (72% of total N) and 140 (28% of total N) MZ twin pairs. This particular ratio was chosen as it approximately reflects the ratio of MZ and DZ twins in European twin registers. Additive genetic variance was assumed 0.5, common-environmental variance was assumed 0.3 and unique-environmental variance, $\exp(\beta_0)$, was set to 0.2. The data was simulated without any G×E ($\beta_1 = 0$) and a phenotypic population mean of 0 ($\mu = 0$). The 1PLM was used to simulate responses to 60 dichotomous items resulting in a scale with a Cronbach's alpha of 0.90. The data was simulated under two different scenarios. In the first scenario, item parameters were simulated from a normal distribution with a mean of 1 and a standard deviation of 1 to mimic a test with relatively difficult items resulting in a slight floor effect for the distribution of sum scores. In the second scenario, item parameters were simulated from a normal distribution with a mean of -1 and a standard deviation of 1 to mimic a relatively easy test resulting in a slight ceiling effect. The first scenario resulted in a situation that is often encountered in psychopathology studies: a positively skewed sum score distribution. The second scenario resulted in a negatively skewed sum score distribution, a scenario that can be encountered in cognitive ability studies with gifted students. To give an idea of the severity of the skewness, the distributions of the simulated sum scores of all DZ twins are displayed in Figure 2.1 for both scenarios. Furthermore, the three different methods for estimating skewness proposed by Joanes and Gill (1998) were used to determine non-normality of the distributions. In the first scenario, the different methods resulted in values in the range $[0.630; 0.632]$ and in the second scenario in the range $[-0.434; -0.435]$.

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

The simulations were carried out using the software package R (R development core team, 2013). As an interface from R to JAGS, the rjags package was used (Plummer, 2013). After a burn-in phase of 7,000 iterations, the characterisation of the posterior distribution for the model parameters was based on an additional 12,000 iterations from 1 Markov

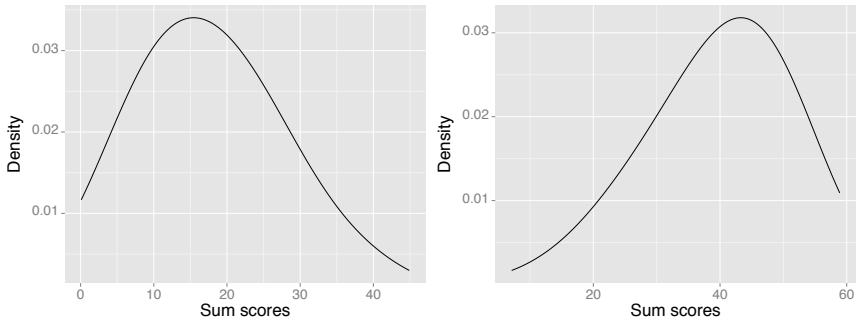


Figure 2.1: Distribution of the sum scores of the DZ twins as simulated in simulation study 1.

chain. This choice was based on previous test runs with multiple chains and computing Gelman and Rubin's convergence diagnostic (Gelman & Rubin, 1992). All test runs with these numbers of iterations resulted in values < 1.02. The average posterior means of the model parameters for all replicated data sets were calculated, the standard deviation of posterior means, as were the means of all posterior standard deviations. The mean of the posterior standard deviations can be interpreted as the Bayesian analog of the standard error.

2.6 Simulation study 2

A second simulation study was conducted to determine the sample size necessary to find G×E in twin data with the suggested method. As in the first simulation study, the simulated data consisted of DZ (72% of total N) and MZ (28% of total N) twin pairs. Additive genetic variance was assumed 0.5, the intercept, $\exp(\beta_0)$, was set to 0.2, the phenotypic population mean to 0 and common-environmental variance was assumed 0.3. The magnitude of G×E, β_1, was varied. The 1PLM was used to simulate responses to 60 dichotomous items resulting in a scale with a Cronbach's alpha of 0.92. The item parameter values were simulated from a normal distribution with a mean of 0 and a standard deviation of 1 and assumed known in the analysis. To estimate the power to detect G×E, item data were simulated with different sample sizes ($N = 500, N = 1000$ and $N = 2000$ twin pairs) and different values for β_1. The effect size of the G×E interaction was defined as the factor with which the environmental variance component increases for an individual with an additive genetic effect of $A_i = \sigma_A$ relative to β_0, and will be henceforth referred to as Δ. To illustrate this, consider

an effect size of $\Delta = 1.1$. The environmental variance for a person with an additive genetic effect equal to σ_A can then be computed as

$$\begin{aligned} \sigma^2_{Ei} &= \exp(\beta_0 + \beta_1 A_i) \\ \Delta &= \exp(\beta_1 \sigma_A) \\ \beta_1 \sigma_A &= \ln(\Delta) \\ \beta_1 &= \ln(\Delta)/\sigma_A \end{aligned} \tag{2.16}$$

resulting in 0.22 ($= 0.2 \times \exp(0.13 \times \sqrt{0.5})$). The slope parameter β_1 then has to be equal to ~ 0.13 ($= \ln(1.1)/\sqrt{0.5}$). With an effect size of $\Delta = 1.5$, β_1 is equal to ~ 0.57 and the environmental variance at $A_i = \sigma_A$ is equal to $0.5 \times 1.5 = 0.75$.

Each condition was repeated 100 times with a different G×E effect sizes ($\Delta = 1.00, \Delta = 1.30, \Delta = 1.50$ and $\Delta = 1.70$). To estimate the power, the 95% highest posterior density (HPD, see e.g. Box and Tiao 1973) interval was determined for each parameter. Power was defined as the percentage of simulations in which the 95% HPD interval did not contain zero.

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

2.7 Results

Simulation study 1

The true parameter values, the average posterior means, and the mean of posterior standard deviations (averaged over 100 replications) are reported in Table 2.1 for the first scenario.

In the first scenario, a slight floor effect was mimicked, resulting in a positively skewed sum score distribution. It can be seen that the sum score analysis approach resulted in biased parameter estimates. Both genetic variance and common-environmental variance were underestimated whereas the intercept (environmental variance when $A = 0$) was overestimated. The sum score approach resulted in an average slope parameter of $\beta_1 = 1.05$, reflecting an effect size of $\Delta = \exp(1.05 \times \sqrt{0.5}(0.43)) \approx 2.00$.

In the second scenario, a slight ceiling effect was mimicked, resulting in a negatively skewed sum score distribution. Since the second scenario is the mirror image of the first scenario, the parameter estimates were the

same but in the opposite direction (β_1 = -1.08). To save space, results of the second scenario are not tabulated.

Table 2.1: Scenario 1: The average posterior means (SD) averaged over 100 replications. Second line: Mean of posterior standard deviations.

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

Simulation study 2

The power estimates for the slope parameter β_1 can be found in Table 2.2. All power estimates for σ_A^2, σ_C^2 and $\exp(\beta_0)$ were equal to 1.00 in all conditions, and are therefore not tabulated. The true parameter values, the average posterior means and the average posterior standard deviations can be found in Table 2.3.

It can be seen that the estimated values are very close to the true values. The power to find G×E in the base-line scenario without any effect ($\Delta = 1.00$) is close to 5% for $N = 1000$ and $N = 2000$. Under the simulated scenario, there is good power to detect an effect size of 1.7, even with only 500 twin pairs.

Table 2.2: Estimated power to find G×E for different sample sizes. N refers to the number of twin pairs

	$\Delta = 1.00$ β_1	$\Delta = 1.30$ β_1	$\Delta = 1.50$ β_1	$\Delta = 1.70$ β_1
N = 500	0.03	0.48	0.57	0.81
N = 1000	0.07	0.57	0.92	0.99
N = 2000	0.07	0.92	1.00	1.00

Table 2.3: The average posterior means (SD) averaged over 100 replications. Second line: Mean of posterior standard deviations. N refers to the number of twin pairs

	$\Delta = 1.00$				$\Delta = 1.30$				$\Delta = 1.50$				$\Delta = 1.70$			
	σ_A^2	σ_C^2	$\exp(\beta_0)$	β_1	σ_A^2	σ_C^2	$\exp(\beta_0)$	β_1	σ_A^2	σ_C^2	$\exp(\beta_0)$	β_1	σ_A^2	σ_C^2	$\exp(\beta_0)$	β_1
True value	0.50	0.30	0.20	0.00	0.50	0.30	0.20	0.37	0.50	0.30	0.20	0.57	0.50	0.30	0.20	0.75
N = 500	0.48 (0.08)	0.31 (0.06)	0.21 (0.03)	0.02 (0.24)	0.50 (0.07)	0.31 (0.07)	0.20 (0.03)	0.41 (0.28)	0.48 (0.08)	0.32 (0.07)	0.21 (0.03)	0.55 (0.24)	0.48 (0.08)	0.32 (0.07)	0.21 (0.03)	0.74 (0.26)
	0.08	0.06	0.03	0.24	0.09	0.06	0.02	0.19	0.09	0.08	0.04	0.27	0.07	0.06	0.03	0.18
N = 1000	0.49 (0.07)	0.31 (0.06)	0.20 (0.02)	0.02 (0.19)	0.48 (0.06)	0.30 (0.05)	0.21 (0.02)	0.37 (0.18)	0.48 (0.07)	0.32 (0.06)	0.20 (0.03)	0.58 (0.15)	0.50 (0.07)	0.30 (0.05)	0.20 (0.03)	0.74 (0.16)
	0.07	0.06	0.02	0.18	0.07	0.06	0.02	0.18	0.07	0.06	0.02	0.19	0.07	0.06	0.03	0.18
N = 2000	0.49 (0.05)	0.31 (0.05)	0.20 (0.02)	-0.01 (0.13)	0.50 (0.05)	0.30 (0.04)	0.20 (0.02)	0.38 (0.11)	0.50 (0.07)	0.30 (0.05)	0.20 (0.03)	0.55 (0.16)	0.50 (0.05)	0.30 (0.05)	0.20 (0.02)	0.75 (0.14)
	0.05	0.04	0.02	0.12	0.05	0.04	0.02	0.12	0.07	0.06	0.02	0.18	0.05	0.04	0.02	0.13

2.8 Discussion

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

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

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

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

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

test might perhaps not result in the finding of spurious G×E, but it does not provide a lot of information either, and is therefore not very attractive psychometrically.

Here we proposed to incorporate an explicit measurement model into the variance decomposition in order to remedy potential bias. Molenaar et al. (2012) used a different approach, proposing the incorporation of a linear factor model into variance decomposition. As a linear factor model assumes normally distributed residuals, the linear factor model is inappropriate for categorical variables in general and for binary variables in particular (Bartholomew et al., 2008). Therefore, the method by Molenaar et al. (2012) is limited to continuous data and not suitable for dichotomous or polytomous items. As dichotomous items are often used in ability tests (scored as right/wrong), the incorporation of a measurement model suitable for dichotomous data is relevant for every research field that uses twin data and ability tests to assess G×E (e.g. research in giftedness or educational achievement). In addition, the incorporation of IRT models for polytomous items is straightforward (see e.g. Samejima, 1969; Masters, 1982; Embretson & Reise, 2009). van den Berg et al. (2007) show how k polytomous items with m response categories can be transformed into $k \times (m-1)$ dummy items that can be used in a model for dichotomous items, so our method can also be applied to polytomous items without altering the JAGS script.

Simulation studies 1 and 2 showed that the proposed method does not find any spurious G×E and recovers the true values of the model parameter very well. In addition, simulation study 2 showed that the statistical power of the method is sufficient given that large samples are often available from twin registries. Only in case of a very small effect size, one needs 2000 twin pairs to find G×E. Note that the simulated effect sizes are all smaller than the effect size of the spurious effect that was found when the sum score approach was used. As it is very common in behaviour genetic studies to see data with a distribution as simulated in simulation study 1 (see Figure 1), the power of the model seems to be good for G×E effects that can be observed in real data. The results of the power study however apply only to the simulated conditions. The power to detect G×E might be different for traits with a different etiology and studies with a different sample composition.

In all analyses, the item parameters were assumed known. This is the case in, for example, large-scale educational assessment situations (see e.g. Veldkamp & Paap, 2013) and in computer adaptive testing (CAT) (e.g. assessment of quality of life, see e.g. Reeve et al., 2007; Nikolaus et al., 2013). It is straightforward for alternative applications to estimate item parameters as well (see van den Berg et al., 2007). A reasonable approach would in most cases be to use independent standard normal distributions as priors for the difficulty parameters (e.g. $N(0, 10)$). With item parameters unknown, the phenotypic population mean for the individuals is best fixed to 0, and this makes an expectation of 0 for the item parameters appropriate.

A variance of 10 makes the prior relatively and reasonably flat. Of course, additionally estimating difficulty parameters will affect power, but only slightly. If the model is extended with varying factor loadings that need to be estimated (discrimination parameters), power will be affected more severely. Reasonable priors to use would be lognormal with expectation 0 and variance 10. The lognormal distribution constrains the discrimination parameters to be positive. Note that in order to fix the scale, one of the item discriminations should be fixed to 1. For more details, see van den Berg et al. (2007).

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

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

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

A similar model as introduced in the present article has been proposed in a paper by Molenaar & Dolan (2014). That paper focuses on the same problem (spurious G×E due to scale properties) but was developed

independently. A nice feature of the Molenaar and Dolan paper is the addition of additive genetic effects interacting with shared environmental influences and modelling of correlated residuals. In our view, a nice feature of our own implementation in JAGS is that the estimation time of the model is much faster and our parameter recovery is very good: estimates are very close to the true values. So, all in all, the present article and the article by Molenaar and Dolan should be regarded complementary.

CHAPTER 3

Increased Environmental Sensitivity in High Mathematics Performance

Based on:
Inga Schwabe, Dorret I. Boomsma and Stéphanie M. van den Berg,
Under revision

Abstract

The results of international comparisons of students such as PISA (Program for International Student Assessment) and TIMSS (Trends in International Mathematics and Science Study) are often taken to indicate that mathematical education in Dutch schools is not appropriate for mathematically talented students. However, there has been no empirical study yet that investigated this hypothesis. If indeed, Dutch students with a (genetic) predisposition for high mathematical ability are not nurtured to their full potential, their mathematics performance should be more affected by environmental factors than that of children with a (genetic) predisposition for low mathematical ability. In behaviour genetics such a situation is termed *genotype-environment interaction*: the relative importance of environmental influences differs depending on students' genotypic values. To investigate genotype-environment interaction, we analyzed mathematics performance of 2110 Dutch twin pairs on a national achievement test. The analysis was corrected for heterogeneity in the measurement of mathematics performance through the application of an item response theory (IRT) measurement model. As hypothesized, results suggest that environmental influences were relatively more important in explaining individual differences in students with a genetic predisposition for high mathematical ability than

in students with a genetic predisposition for low mathematical ability (effect size = 1.63). Thus, performance in low-ability students is better predicted by their genotypic value than performance in the high-ability students.

3.1 Introduction

While some children seem to be born with the ability to solve complex mathematical equations, others are terrified of equations and mathematical symbols. Dutch teachers usually focus on the latter group of students: the weakest (Dekker, 2014). Often criticized as a "culture of C-grades", education in the Netherlands has the reputation of being traditionally less focused on students with high mathematics performance levels. In an ideal school system, however, also the talented child should be nurtured to its full potential. After all, the brightest students may be the ones who make important contributions to science, find cures for diseases or invent new technologies.

International comparisons such as the Program for International Students Achievement (PISA) and the Trends in International Mathematics and Science Study (TIMSS) show that, in the Netherlands, the average mathematical performance level in primary education is relatively high. This observation can, however, be attributed mainly to the high performance in the left tail of the achievement continuum: the weakest students are performing better than the weakest students from all other countries participating in PISA and TIMMS. However, the variance of test scores is, compared to other high-scoring countries, very small: the performance levels of Netherlands' lowest- and highest-scoring students are relatively close. In other words, whereas Holland's weakest students perform exceptionally well, Netherlands' top students are outperformed by the brightest students from Asian and other western countries (see e.g. Meelissen et al., 2012; van der Steeg, Vermeer & Lanser, 2011). This appears to be a persistent phenomenon: similar patterns have been found over the years for different age groups (see e.g. Minne, Rensman, Vroomen & Webbink, 2007). These findings are often presented as underperformance in the high-ability students (see e.g. van der Steeg et al., 2011) and interpreted as an indication that mathematical education in Dutch schools is better tailored to the weaker students than to the mathematically talented students. However, one cannot draw conclusions on underlying processes based on the test score distribution alone. There are alternative explanations for the relatively poor performance of the top students in the Netherlands. For example, they might be genetically different from students from other countries or not motivated enough to push themselves to reach their full potential.

In this article, the underperformance of Dutch mathematically talented students was investigated from a behaviour genetics perspective. A child's mathematical talent was defined as its *genotypic value*, representing the

summated effect of all genes that affect mathematical ability (Falconer & MacKay, 1995). The absence of inequalities in educational opportunities would predict that individual differences in scores are mainly explained by genetic differences (nature) rather than environmental influences (nurture) (see also Shakeshaft et al., 2013). This means that, if indeed, in primary education, mathematically talented children are not nurtured to their full potential, their performance should be more affected by situational factors than the performance of average or weak students of the same age. For example, they might be at the mercy of random events like having a teacher that is interested in their abilities. In the behaviour genetics literature, such a situation is formally described as *genotype-environment interaction*: conditional on a child's *genotypic value* for mathematical ability, environmental influences can be more or less important (e.g. Cameron, 1993).

3.1.1 Genetic analysis

One of the methods used in behaviour genetics to estimate the relative influence of genetic and environmental factors is the twin design. Twin pairs are either identical (monozygotic, MZ) or non-identical (dizygotic, DZ). MZ twins (largely) share the same genomic sequence and the same rearing environment, including prenatal environmental conditions. DZ twins also share the same prenatal and rearing environment but on average only share half of the segregating genes. By using the twin design, the relative contributions of genetic variability and environmental variability can be estimated, where the heritability is defined as the ratio of genetic variance divided by total variance in a measured trait (phenotypic variance).

3.1.2 Prior research

Although a considerable number of twin studies have been conducted on the heritability of mathematical ability (see e.g. Alarcon, Knopik & DeFries, 2000; Markowitz, Willemsen, Trumbetta, van Beijsterveldt & Boomsma, 2005; Oliver et al., 2004; Kovas, Haworth, Petrill & Plomin, 2007; Hart, Petrill, Thompson & Plomin, 2009; Shakeshaft et al., 2013; Davis et al., 2014), to our knowledge, there is only one twin study that compared the relative contributions of genetic and environmental influences in mathematically high-scoring children and children in the normal range. In a population-based sample of 10-year-old British twins, Petrill, Kovas, Hart, Thompson and Plomin (2009) defined mathematically high-scoring twins as those who scored at or above the 85th percentile and analyzed twin concordance rates (i.e., whether an individual twin meets the high mathematics cutoff or not) and estimated genetic and environmental variance components. In the top 15% of students, results were similar to those obtained across the normal range of ability. Similar results were reported

for high cognitive performance and high reading performance (Petrill et al., 1998; Ronald, Spinath & Plomin, 2002; Saudino, Plomin, Pedersen & McClearn, 1994; Boada et al., 2002), traits that are highly correlated with mathematical ability (Davis, Haworth & Plomin, 2009; Plomin & Deary, 2015; Davis et al., 2014). These findings seem to argue against the presence of a genotype-environment interaction, at least in the populations studied. If there were indeed genotype-environment interaction, studies focusing on the high extreme of mathematical ability (or a related trait), should reveal that environmental influences are more (or less) important than for the normal range of ability.

Although the comparison of concordance rates in MZ and DZ twin pairs provides a simple test of the etiology of extreme performance scores, the continuous measure of mathematical performance needs to be transformed into a dichotomous variable. As a result, statistical power is low and information is lost regarding variability along the entire performance continuum (see also Boada et al., 2002). Furthermore, an arbitrary cutoff has to be chosen (e.g., as in Petrill et al. (2009), the 85th percentile). Instead, here, we model genotype-environment interaction continuously, letting the size of environmental variance components vary as a function of the genotypic value (see below for more detail). Thus, rather than using a dichotomized variable, this method takes advantage of the continuous dimension of mathematical performance. Using this approach, however, we have to correct for the increased measurement error in the upper tail of the test score distribution. While most achievement tests show little measurement error for average scoring students, scores can be very unreliable for high performing students due to the small amount of information provided by only a few very difficult items. In other words: measurement error is not the same across the ability continuum (heterogeneity). The relative lack of reliability in the upper and lower tails leads to smaller correlations among sum scores (attenuation), which leads to bias when estimating genetic and environmental variance components (see also van den Berg et al., 2007) and furthermore can lead to the finding of spurious genotype-environment interaction effects (see Schwabe & van den Berg, 2014; Molenaar & Dolan, 2014). The problem of heterogeneous measurement error can be solved by, instead of focusing on observed test scores, modelling latent variables (see for instance Béguin & Glas, 2001; Fox & Glas, 2003), that is, estimating variance components by correcting for attenuation (see also van den Berg et al., 2007).

In earlier research on genotype-environment interaction in mathematics performance, a non-Dutch population was studied, using a method that is low in statistical power. Here, we model genotype-environment interaction continuously, by applying a recently developed method (Schwabe & van den Berg, 2014; Molenaar & Dolan, 2014) that corrects for measurement error through the application of an item response theory (IRT) measurement model. By incorporating an IRT model into the analysis, the results regarding genotype-environment interaction presented here are free

of artefacts due to heterogeneous measurement error across the performance continuum. The method was applied to data from 2110 12-year-old Dutch twin pairs on the 60 items of the mathematical subscale of the *Eindtoets Basisonderwijs* test, a Dutch national achievement test administered in the final year of primary education. If the primary educational system in the Netherlands really is better suited for students without a special talent (i.e., low genotypic value) for mathematics than for talented students (i.e., high genotypic value), results should show more random environmental variation in children genetically predisposed towards high mathematical ability than for children genetically predisposed towards low mathematical ability.

3.2 Method

3.2.1 Data

The sample of twins for this study comes from the Netherlands Twin Register (NTR, Boomsma et al., 2002), established in 1986 by the Department of Biological Psychology at the VU University in Amsterdam. This register includes approximately 40% of all multiple births in the Netherlands. Data on 12-year-old twins from birth cohorts 1998-2000 were analysed to study genotype-environment interaction in mathematical achievement on the *Eindtoets Basisonderwijs* test. Conducted and analysed by the testing company Cito, the *Eindtoets Basisonderwijs* is a Dutch national educational achievement that assesses what a child has learned during the past eight years of primary education. The test consists of 290 multiple choice items in four different subjects (language, arithmetic/mathematics, study skills and world orientation [optional]). For this paper, the 60 dichotomous item scores (coded as 0=incorrect, 1=correct) of the mathematics subscale of this test were used. The method used in this study required item data, whereas at the NTR only total test scores are available. The NTR data on twins therefore had to be linked to item data that is available at Cito. This linking was only done for twins for which signed informed consent forms for database linking were available. To further protect the privacy of the twins, the linking was not done by the authors of this manuscript, nor the researchers associated with the NTR, but by an ICT co-worker at Cito who was not involved in the study. For the linking, the co-worker received data on 7031 individual twins from the NTR on name, sex, birth year, name of the school, and total Cito score, if available. The first step of the linking was then to link the NTR data to a BRIN code, a 6-digit number that is given to educational institutes by the Dutch ministry. The first four numbers distinguish between Dutch educational institutes and the last two numbers are optional and only used when an educational institutes has more than one location (e.g. "00" to indicate that it is the main location). Then 12 different queries with a different combination of the BRIN code, birth year, sex, surname and initials of a twin were used to identify the item

data associated with an individual twin. 1017 individual twins had more than one unique match and 2427 individual twins could not be matched at all, reducing the dataset to 3587 individual twins consisting of 2149 families. To link individual twins that could be linked to their item scores to the NTR data of their co-twin, a unique family ID in the NTR dataset was used to find the NTR record of the co-twin. Excluding triplets (N=63 individual twins, 21 families), this led to a dataset of 4238 individual twins (2119 families). Twin pairs with unknown zygosity (N pairs = 9) were excluded from the analysis, leading to a total of 4220 individual twins (2110 twin pairs), forming 581 MZ pairs and 1529 DZ pairs. Of the monozygotic twins, 282 pairs were male and 299 were female; of the dizygotic twins, 360 pairs were male, 309 were female, and 860 were of opposite sex. For 711 twins, item scores were unknown. The reasons that the scores were missing were either that the child had not reached final grade yet (N twins = 52), the child was attending special education (N twins = 34), a different test was used at the school the twin was attending (N twins = 13), the child (N twins =2) or the whole school (N twins = 1) did not attend the test or the reason was unknown (N twins = 609).

The *Eindtoets Basisonderwijs* was administered using different test versions. In each year a regular test (paper-based) and an anchor test (paper-based) was used. The anchor test is an adapted version of the regular test in which 20 items are replaced by anchor items that are common between the years. This creates an internal anchor with which the tests from different years can be linked. Thus, 40 of the 60 items were the same in the regular test version and the anchor test version of the particular year (2010, 2011 & 2012) whereas 20 items were unique in the regular and anchor test version of a particular year. The 20 unique items were the same every year in the anchor test version but these unique items differed from year to year in the regular test version. Furthermore, there were three different digital test versions (computer-based tests) of which two shared 45 items while the rest of the items were unique. In this particular twin sample, the different combination of items from the regular test, the anchor test and the digital test led to nine different combinations (test versions).

These different test forms may differ with respect to their overall difficulty and, as a result, the sum scores are not comparable across versions. To make them comparable, measures needed to be harmonized such that data from individual twins assessed by a different test version could be compared meaningfully. However, instead of equating sum scores from different versions to make them comparable, we equated item parameters across versions by an IRT measurement model. To link the different regular and anchor test versions, the psychometric group at Cito made a concurrent estimate of all the item parameters in the twin data. Note that the data allowed for a test linking design that is a combination of common item equating and common person equating. That is, the regular test and the anchor test item parameters were linked via common item blocks and the

anchor item blocks and the digital tests were linked via common persons. Test linking using item equating allows correcting for differences in difficulty at the item level, rather than at the total test score level and makes it easy to correct for heterogeneous measurement error. The resulting item difficulty and discrimination parameters for all items were imputed in the measurement model which is described in further detail below.

3.2.2 Genetic models

Using twin data, we can estimate different genetic models. The most commonly used model is the ACE model, which decomposes observed phenotypic variance, σ_P^2, into variance due to additive genetic influences (denoted as σ_A^2), variance due to common-environmental influences (denoted as σ_C^2) and variance due to unique-environmental influences (residual variance, denoted as σ_E^2). Common-environmental influences are influences that are shared in a twin pair (e.g. family environmental influences) and are parametrized to be perfectly correlated within a twin pair. Unique-environmental influences are not shared within one family and are parametrized to be uncorrelated for members of a twin pair (i.e., non-shared environmental influences). It is also possible to use an AE decomposition in which common-environmental variance is fixed to zero or an ADE model in which the C component (common-environmental influences) is replaced by a D component to estimate dominance effects (non-additive genetic variance). In this paper, all three biometric models (with and without genotype-environment interaction) were fitted in order to find the biometric model that fitted the data well, while, at the same time, being parsimonious.

All biometric models were fitted simultaneously with a measurement model (IRT model). In the following, the modelling of genotype-environment interaction and the IRT model will be discussed separately and it will then be shown how they can be fitted simultaneously. The illustration of the modelling of genotype-environment interaction will be based on the most commonly used model, the ACE model.

Genotype-environment interaction

In case of genotype-environment interaction, the amount of variance due to environmental influences varies systematically with genotypic value A. In the twin design, the genotypic value is parametrized as latent (e.g., unobserved) factor that is hypothesized to be responsible for the covariance in MZ and DZ twins. We can distinguish between two different types of interaction effects: There can be an interaction with unique-environmental influences (henceforth referred to as A$\times$E) and there can be an interaction with common-environmental influences (henceforth referred to as A$\times$C).

In case of A$\times$E, we portion variance due to unique-environmental influences into an intercept (representing unique-environmental variance when

$A = 0$) and a part that is a function of A. This makes unique-environmental variance different for each individual j:

$$\sigma^2_{Ej} = \exp(\beta_0 + \beta_1 A_j) \tag{3.1}$$

where β_0 denotes the intercept (i.e., unique-environmental variance when $A = 0$) and β_1 is a slope parameter that represents A×E. Likewise, to model A×C, we portion variance due to common-environmental influences into an intercept (i.e., common-environmental variance when $A = 0$) and a part that is a function of A (cf. Molenaar & Dolan, 2014):

$$\sigma^2_{Cj} = \exp(\gamma_0 + \gamma_1 A_j) \tag{3.2}$$

where γ_0 denotes the intercept and γ_1 represents A×C.

Both interaction effects are modelled here as (log)linear effects, meaning that environmental variance is larger at either higher or lower levels of the genotypic value (i.e., larger differences among individuals with similar genotypic value). The sign of the slope determines the direction of the interaction effect. The exponential function is used in order to avoid negative variances (see also e.g. SanChristobal-Gaudy et al., 1998; van der Sluis et al., 2006; Bauer & Hussong, 2009; Hessen & Dolan, 2009).

Measurement model

Whereas in the sum score approach, scale properties are ignored, the IRT approach uses properties of each item as information to be incorporated into the scaling of individual test performance. The probability of a correct answer of individual j on item k is modelled as a function of the difference between the individual's latent trait score θ_j and the item difficulty, b_k, representing the trait level associated with a 50 % chance of answering an item correctly. Furthermore, a discrimination parameter, a_k, can be incorporated into the measurement model. The discrimination parameter indicates how rapidly the probabilities of giving a correct answer change with varying levels of the latent trait and is comparable to a factor loading in factor analysis. Latent trait value θ_j can be interpreted as the true theoretical performance level for individual j that is corrected for the difficulty levels of the items that were in a student's test version.

Here the one parameter logistic model (OPLM, Verhelst, Glas & Verstralen, 1995) version of an IRT model was used at Cito that is suitable for dichotomous data where item responses are scored as correct/false. In the OPLM, item difficulty parameters b_k are estimated and item discrimination parameters are imputed as known constants. In this model, the odds that person j answers item k correctly is modelled using a natural logarithm function:

$$\ln(P_{jk}/(1 - P_{jk})) = a_k(\theta_j - b_k) \tag{3.3}$$

The item-response curves (IRC) resulting from Equation 3.3 can be transformed into item information curves. For a dichotomous item, the item information curve for item k can be calculated as follows:

$$I_k(\theta) = P_k(\theta)(1 - P_k(\theta)) \tag{3.4}$$

where $P_k(\theta)$ refers to the conditional probability of answering item k correctly (i.e, the IRC) given θ and item parameters a_k and b_k (Embretson & Reise, 2009). Under the assumption that conditional on θ, the item responses are uncorrelated (i.e. local independence), item information curves are additive across items and can be used to determine the test information function (Embretson & Reise, 2009):

$$TI(\theta) = \sum_{k=1}^{K} I_k(\theta) \tag{3.5}$$

The test information curve represents the amount of psychometric information a test contains for all points along the continuum of latent traits (abilities). The test information has an exact relationship with an individuals' standard error of measurement, which can be calculated as:

$$SE(\theta) = \frac{1}{\sqrt{TI(\theta)}} \tag{3.6}$$

3.2.3 Incorporating biometric and measurement model

Van den Berg et al. (2007) showed that, to take full advantage of the IRT approach, the genetic model and the IRT model have to be fitted concurrently (using a one-step approach). Here we use Bayesian statistical modelling to estimate all parameters simultaneously. In a Bayesian analysis, statistical inference is based on the joint posterior density of the model parameters, which is proportional to the product of a prior probability and the likelihood function of the data (for further reading see e.g. Box & Tiao, 1992). We use a Markov chain Monte Carlo (MCMC) algorithm called Gibbs sampling (Geman & Geman, 1984; Gelfand & Smith, 1990; Gelman et al., 2004) to obtain this joint posterior density. The algorithm works by iteratively drawing samples from the full conditional distributions of all unobserved parameters of a model. The full conditional distribution relates to the distribution of a parameter given the current values of all other relevant parameters of the model (Gilks, Richardson & Spiegelhalter, 1996). In each iteration of the Gibbs sampling algorithm, a sample is taken from the conditional distribution of every parameter in the model, given the current values of the other relevant parameters of the model. It can be shown that after a number of "burn-in" iterations, subsequent draws can be seen as draws from the joint posterior distribution of all parameters.

ACE model

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

Under the assumption that MZ twins have identical genotypic values as they largely share the same genomic sequence, we assumed for each MZ twin pair i a normally distributed additive genetic effect A_i:

$$A_i \sim N(0, \sigma_A^2) \tag{3.7}$$

Then, for each MZ twin pair i, a normally distributed common-environmental effect C_i that correlates perfectly within one family was modelled:

$$C_i \sim N(0, \sigma_{Ci}^2) \tag{3.8}$$

To incorporate A×C, common-environmental variance, σ_{Ci}^2, was defined as a function of the additive genetic effect of family i ($\sigma_{Ci}^2 = \exp(\gamma_0 + \gamma_1 A_i)$, see also Equation 3.2). An ACE decomposition on the latent variable θ_{ij} for individual twin j from pair i then yields:

$$\theta_{ij} \sim N(\mu + A_i + C_i, \sigma_{Ei}^2) \tag{3.9}$$

where μ denotes the phenotypic population mean and σ_{Ei}^2 is portioned into an intercept β_0 and a slope parameter β_1 ($\sigma_{Ei}^2 = \exp(\beta_0 + \beta_1 A_i)$, see also Equation 3.1) to estimate A×E. Simultaneously with the biometric model described above, the latent phenotype θ_{ij} appears in the OPLM for observed item data Y_{ijk} (see also Equation 3.3):

$$\ln(P_{ijk}/(1 - P_{ijk})) = a_k(\theta_{ij} - b_k) \tag{3.10}$$

$$Y_{ijk} \sim Bernoulli(P_{ijk}) \tag{3.11}$$

where a_k refers to the discrimination parameter of item k and b_k to the difficulty of item k. These parameters were estimated beforehand by the psychometric group at Cito and imputed in our model.

While the total genetic variance is the same for DZ and MZ twins, the genetic covariance in MZ twin pairs is twice as large as in DZ twins (assuming random mating). To model these different genetic correlations, first a normally distributed familial effect $A1_i$ was modelled (cf. Jinks & Fulker, 1970):

$$A1_i \sim N(0, \tfrac{1}{2}\, \sigma_A^2) \tag{3.12}$$

Then, for each individual twin j from DZ family i a normally distributed additive genetic effect $A2_{ij}$ with expectation $A1_i$ was used:

$$A2_{ij} \sim N(A1_i, \tfrac{1}{2}\, \sigma_A^2) \tag{3.13}$$

To model common-environmental influences C_i that are perfectly correlated within one DZ pair i, we used a standard normal distribution:

$$C_i \sim N(0, 1) \tag{3.14}$$

In case of A×C, the variance of C_i can differ depending on genotype $A2_{ij}$. Therefore, the common-environmental effect C_i was scaled by multiplying it with the standard deviation σ_{Cij}, where $\sigma^2_{Cij} = \exp(\gamma_0 + \gamma_1 A2_{ij})$ (see also Equation 3.2). This yields a common-environmental effect $C2_{ij}$ that is unique for every individual twin j from family i:

$$C2_{ij} = C_i \sqrt{\exp(\gamma_0 + \gamma_1 A2_{ij})} \tag{3.15}$$

An ACE decomposition then yields:

$$\theta_{ij} \sim N(\mu + A2_{ij} + C2_{ij}, \sigma^2_{Eij}) \tag{3.16}$$

As for MZ twins, σ^2_{Eij} was portioned into an intercept β_0 and a slope parameter β_1 ($\sigma^2_{Eij} = \exp(\beta_0 + \beta_1 A2_{ij})$, see also Equation 3.1) to model A×E. Again, simultaneously with the ACE decomposition the latent phenotype θ_{ij} was estimated based on the OPLM (see Equations 3.3 and 3.10-3.11).

ADE model

In addition to the ACE model, an ADE model was estimated. The ADE model decomposes phenotypic variance σ^2_P into variance due to additive genetic influences (denoted as σ^2_A), variance due to non-additive genetic (dominance) influences (denoted as σ^2_D) and variance due to unique-environmental influences (denoted as σ^2_E). Under this model, for MZ twins, we assumed for each family i total (additive and non-additive) genetic effects, G_i, with additive genetic value A_i (see Equation 3.7) as expectation and a variance of σ^2_D. The A×E interaction effect was then conditioned on the complete genotype instead of only additive genetic effects (i.e., $\sigma^2_{Ei} = \exp(\beta_0 + \beta_1 G_i)$; henceforth referred to as G×E) and the expected mean of θ_{ij} was $\mu + G_i$. In order to model a correlation of $\frac{1}{4}$ for dominance genetic effects in DZ twins, we split up σ^2_D into two parts, $\frac{1}{4}\sigma^2_D$ and $\frac{3}{4}\sigma^2_D$:

$$G_i \sim N(A1_i, \tfrac{1}{4}\sigma^2_D) \tag{3.17}$$

$$G1_{ij} \sim N(G_i, \tfrac{1}{2}\sigma^2_A) \tag{3.18}$$

$$G2_{ij} \sim N(G1_{ij}, \tfrac{3}{4}\sigma^2_D) \tag{3.19}$$

where $A1_i \sim N(0, \frac{1}{2}\sigma^2_A)$ (see Equation 3.12). As for MZ twins, the genotype-environment interaction (G×E) was modelled on the complete genotype of individual j from family i ($\sigma^2_{Eij} = \exp(\beta_0 + \beta_1 G2_{ij})$) and the expected value of θ_{ij} was $\mu + G2_{ij}$.

3.2.4 Prior distributions

As a Bayesian approach was used, prior distributions had to be specified. A reasonable choice for the prior distribution of variance components would be an inverse gamma distribution (e.g., $\sigma^2_A, \sigma^2_C \sim InvG(1,1)$), because this distribution is both flexible and conjugate. A prior is called conjugate when the probability distribution of the prior and the posterior distribution have similar forms, resulting in convenient sampling and a faster estimation process. The discrimination parameters for the OPLM were however quite high (in the range of [1;9]) which resulted in a very small total variance of θ. Gelman (2006) showed that in case of small variances, an inverse gamma distribution cannot be regarded as non-informative (i.e., expressing only vague information about the variance components), which might make the posterior distribution very dependent on the choice of the prior distribution. Therefore, we used a uniform prior distribution for all variance components ($\sigma^2_A,\ \sigma^2_D \sim U(0,100)$ and $\sigma^2_E,\ \sigma^2_C \sim U(0,100)$) in the models without interaction effects. In the biometric models with interaction effects, the prior for the intercepts and slope parameters was normal and relatively non-informative ($\beta_0, \gamma_0 \sim N(-1,2)$, $\beta_1, \gamma_1 \sim N(0,10)$). A normal prior distribution was placed on the phenotypic population mean ($\mu \sim N(0,10)$). As noted above, item parameters were assumed known.

3.2.5 Analysis

To find the genetic model that fits the data well and, at the same time, is parsimonious, we fitted all genetic models without interaction effects (simple ACE and ADE model), an ACE model with one (either A×E or A×C) interaction effect and with both interaction effects and an ADE model with a G×E interaction effect. Furthermore, we fitted an AE model in which common-environmental influences were set to zero. A simple AE without A×E interaction was fitted as well as an AE model with A×E interaction. In order to asses model fit, the deviance information criterion (DIC, Spiegelhalter, Best, Carlin & van der Linde, 2002) was calculated for each model. The DIC is a measure that estimates the amount of information that is lost when a given model is used to represent the process that generates the data. It takes account of both the goodness of fit and the complexity of a model and can be seen as a Bayesian analogue of Akaike's Information Criterion (AIC).

For the MCMC estimation, we used the freely obtainable software package JAGS (Plummer, 2003). For further data handling, the statistical programming language R (R development core team, 2013) was used. As an interface from R to JAGS, we used the rjags package (Plummer, 2013).

After a burn-in phase of 12,000 iterations for each separate chain, the characterization of the posterior distribution for the model parameters was based on a total of 75,000 iterations from five different Markov chains. The

burn-in period was chosen on basis of previous test runs with multiple chains, calculating Gelman and Rubin's convergence diagnostic (Gelman & Rubin, 1992). The mean and standard deviation of the posterior point estimates was calculated for each parameter as was the 95 % highest posterior density (HPD, see e.g. Box & Tiao, 1973) interval, which can be interpreted as the Bayesian analogue of a confidence interval (CI). The influence of model parameters can be regarded as significant when the respective HPD interval does not contain zero. This does not hold for the variance components, as these are bounded at zero. Furthermore, for all test versions, the test information and standard error of measurement was calculated for a range of latent trait θ values. Furthermore, the effect size, defined as the factor with which the environmental variance component increases for an individual with an additive genetic effect of $A = \sigma_A^2$ (Schwabe & van den Berg, 2014), was determined for both interaction parameters.

3.3 Results

Table 3.1 presents the DIC for all fitted biometric models. The ACE model with A×E and A×C showed the lowest DIC. Therefore, the ACE model with an A×E as well as A×C interaction effect was chosen as the preferred model for our data.

Table 3.1: Model fit (DIC) for all fitted biometric models. DIC = deviance information criterion.

Biometric model	DIC
I. AE	204356
a) with A×E	204337
II. ACE	204357
a) with A×E	204338
b) with A×C	204344
c) with A×E + A×C	204334
III. ADE	204356
a) with A×E	204337

Based on the ACE model with A×E and A ×C, the posterior means and standard deviations of all parameters as well as narrow-sense heritability can be found in Table 3.2.

Narrow-sense heritability is the proportion of the additive genetic variance of the phenotypic variance and was defined here as $\frac{\sigma_A^2}{(\sigma_A^2+\exp(\gamma_0)+\exp(\beta_0))}$. The results suggest that most of the phenotypic variance can be explained

Table 3.2: Estimates of variance components and narrow-sense heritability, based on the ACE model with A×C and A×E interactions. Total phenotypic variance, defined as $\sigma_A^2 + \exp(\gamma_0) + \exp(\beta_0)$, was 0.0757. HPD refers to the 95% highest posterior density interval.

	Posterior point estimate (SD)	HPD
σ_A^2	0.0552 (0.0035)	[0.0480;0.0617]
$\exp(\gamma_0)$	0.0054 (0.0024)	[0.0012;0.0099]
$\exp(\beta_0)$	0.0151 (0.0016)	[0.0120;0.0183]
β_1	2.0837 (0.4939)	[1.0886;3.0344]
γ_1	2.6159 (1.3503)	[-0.1358;5.2189]
h^2	0.7286 (0.0396)	[0.6448;0.7988]

by genetic influences, resulting in a narrow-sense heritability h^2 of 0.7286. A substantial part of the phenotypic variance could be explained by unique-environmental influences while variance due to common-environmental influences was negligibly small. The results showed a positive A×E interaction effect such that individuals individuals having high genotypic values for mathematical ability show more variance due to unique-environmental influences than individuals with lower genotypic values. The 95% HPD interval shows that this effect was significant. Furthermore, the results suggest that there is a positive A×C effect such that individuals having high genotypic values for mathematical ability show more variance due to common-environmental influences than individuals with lower genotypic values. The 95% HPD interval however shows that this interaction effect was not significant. Effect sizes of the interaction effects were 1.63 (A×E) and 1.85 (A×C).

To illustrate the magnitude of the A×E and A×C interaction effects, the latent trait values for 10,000 MZ twin pairs were simulated based on the parameter values found under the ACE with A×E and A×C model (i.e. the posterior point estimates, see Table 3.2). We then plotted the latent trait value of the first twin of all MZ twin pairs against the latent trait value of the second twin of all MZ twin pairs. Furthermore, the 95% credibility region of the A×E interaction effect is displayed for the entire range of estimated genotypic values (see Figure 3.1).

Figure 3.2 shows the test information curve and standard errors of measurement for each different test version for a range of latent trait values ([-0.20;0.90]). This range was chosen based on the 95% HPD interval of the θ values that occurred in the posterior based on the ACE model with A×E and A×C. Investigating the test information curves, it can be seen that

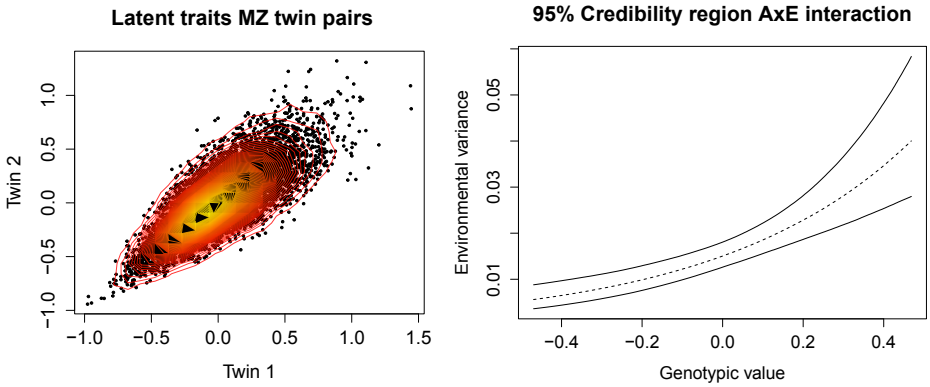


Figure 3.1: Left: Simulated latent trait values for 10,000 MZ twin pairs, conditional on the posterior means of the ACE model with A×E and A×C interaction (see Table 2). Right: 95% credibility region of the A×E interaction effect for the entire range of estimated genotypic values.

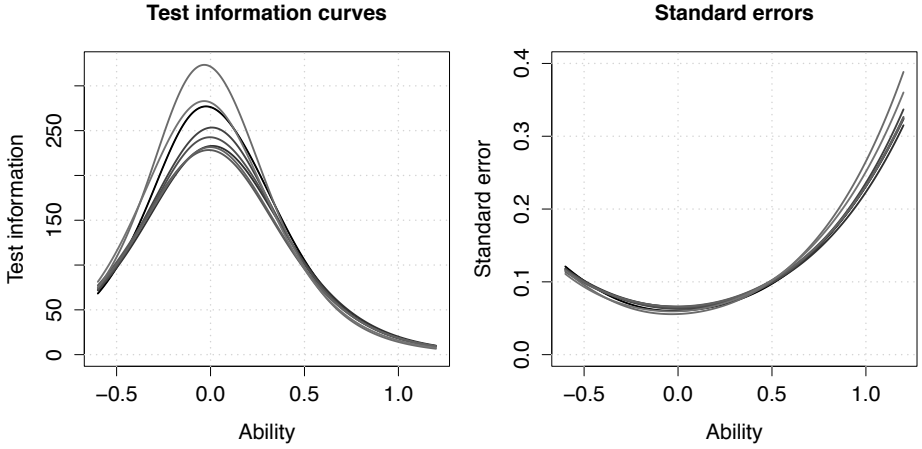


Figure 3.2: Test information curves (left) and standard errors (right) for all different test versions for a representative part of the latent trait continuum.

all test versions provide substantial information for the average and even low-performing range of trait values, but that there is less information for the very-high performing students. This is also reflected in the standard errors of measurement. Although standard errors are generally low, they are higher for high-performing students. This is also reflected in the distribution

of the sum scores (see Figure 3.3). Although only 35 (<1%) students got a perfect score of 60 correct items, only 828 students (20%) scored above the mode of 52 correct item answers.

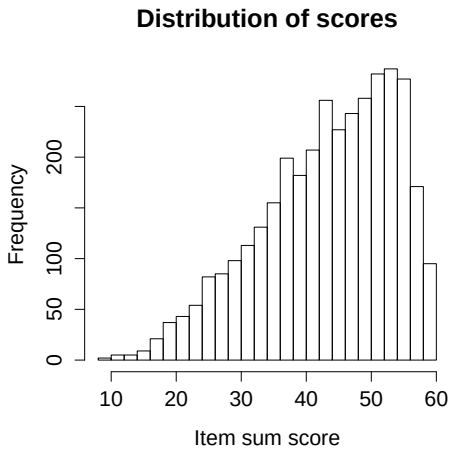


Figure 3.3: Histogram of the raw sum scores (not corrected for different test versions).

3.4 Discussion

The results of international comparisons are often interpreted as an indication that education in Dutch schools is more appropriate for the weakest students than for the mathematically talented students. However, until now there has been no empirical study that tested this hypothesis. From a behaviour genetics perspective, we can translate this hypothesis into the presence of *genotype-environment interaction*: if indeed children with a special inherited talent for mathematics are not nurtured to their full (genetic) potential, we expect their academic performance to show more environmental variability.

Genotype-environment interaction was investigated in Dutch students on the continuous dimension of mathematical performance. As hypothesized, we found a significant genotype-environment interaction (A×E): Unique-environmental influences were relatively more important in explaining individual differences in children with a genetic predisposition towards high mathematical ability than in explaining individual differences in children with a genetic predisposition towards low mathematical ability. In other words, the heritability was higher in low-ability students than in high-ability students.

We have, however, to be cautious in drawing conclusions. The distribution of the twin's mathematics sum scores was negatively skewed. Although only 1% obtained a perfect score of 60 correct items, a relatively low percentage scored higher than the mode. While correction for attenuation through the IRT model does address the problem of heterogeneous measurement error, it does not solve the more extreme problem of lack of reliability. If the test does not discriminate well in the right tail of the distribution, then the estimates of genotype-environment interaction effects are based foremost on the information on the rest of the population. To draw reliable conclusions on children with extreme high ability, a larger sample of mathematically talented twins who took a test with very difficult items would be required.

In recent years, there has been a remarkable increase in attention for the education of talented students in the Netherlands. The goal of the government is to encourage excellent performance by offering talented students education tailored to their individual needs (see Dekker, 2014). Appropriate education for gifted students has become an important issue since 2009 (Boer, Minnaert & Kamphof, 2013). The present results add to our understanding on this issue, but drawing policy conclusions based on current findings is too premature. Underperformance of Dutch students in international comparisons with other countries is a complex issue that needs much further research. Twin studies can be a valuable complement to the findings of educational research, because the twin design makes it possible to correct for genetic influences when the effect of specific environmental influences is investigated. Furthermore, variance that cannot be explained by measured environmental influences can be attributed to one of two sources − influences that are shared in a twin pair (common-environmental influences) and influences unique for a twin. A twin study alone that considers only latent genetic and environmental influences, however, likely underestimates influences involved in the process. Factors that impact this complex issue likely concern many variables at the student level, but also on the school and family level and the larger policy level of each country.

As genotype-environment interaction was investigated in case of unmeasured genetic and environmental influences, no conclusions can be drawn on the nature and importance of specific environmental influences that are more important for students genetically predisposed towards high mathematical ability than for students with a genetic predisposition towards low or average mathematical ability. Still, one important implication can be drawn: The finding of a significant A×E effect compared to a non-significant A×C interaction effect suggests that factors that are not shared in children from the same family are the most important source of individual differences in high-ability students. Therefore, future research should aim to explain within-family variance rather than between-family variance.

One factor that could be important is of course the school environment. As many twin pairs shared the same classroom, it is however not very likely that school influences contributed a lot to unique-environmental variance.

Still, however, the school environment might be important. Although children in the same school appear to receive the "same" school environment, this does not necessarily mean that they perceive this environment similarly (Plomin & Daniels, 1987). A process referred to as *genotype-environment correlation* might be important. A scenario that could potentially explain our results would be the following: mathematically talented twins ask more questions which leads to an amplification of effects (e.g. talented twins become even more talented because of a better understanding), which would result in the finding of a positive A×E interaction effect.

There is a broad range of influences that can contribute to differences in twin pairs, ranging from prenatal differences to different perceptions of the environment to subtle differences in brain structure. Future research should first focus on variables that have proven to be important for talented students, such as peer influences (Austin & Draper, 1981), personality characteristics (Ackerman, 1997) and motivation (Vallerand, 1994).

CHAPTER 4

Genes, Culture and Conservatism - A Psychometric-Genetic Approach

Based on:
Inga Schwabe, Wilfried Jonker and Stéphanie M. van den Berg,
Behavior Genetics, in press

Abstract

The Wilson-Patterson conservatism scale was psychometrically evaluated using homogeneity analysis and item response theory (IRT) models. Results showed that this scale actually measures two different aspects in people: on the one hand people vary in their agreement with either conservative or liberal catch-phrases and on the other hand people vary in their use of the "?" response category of the scale. A 9-item subscale was constructed, consisting of items that seemed to measure liberalism, and this subscale was subsequently used in a biometric analysis including genotype-environment interaction, correcting for non-homogeneous measurement error. Biometric results showed significant genetic and shared environmental influences, and significant genotype-environment interaction effects, suggesting that individuals with a genetic predisposition for conservatism show more non-shared variance but less shared variance than individuals with a genetic predisposition for liberalism.

4.1 Introduction

The term conservatism is used in many ways (Pedhazur & Schmelkin, 1991), but most often refers to politic-economic conservatism. More generally, conservatism can be seen as a generalized resistance to change and ambiguity which is expressed as a preference for safe, traditional and conventional forms of institutions and behaviour. Wilson and Patterson (1968) developed a conservatism scale to measure social attitudes related to the conservative personality. They regarded the then existing scales to be of poor psychometric quality because of susceptibility to agreement-response bias and complex, double-barrelled and/or confusing questions. Instead, their conservatism scale consists of very short catch-phrases. Examples of catch-phrases include "Liberals" and "Living together". The test taker is then asked "Please indicate whether or not you agree with each topic by circling "Yes" or "No" as appropriate. If uncertain please circle "?". These catch-phrases were expected to activate the respondent's affective system, the system Wilson and Patterson (1968) hypothesised to be the most influential component for conservative attitudes and behaviours. The affective system seems indeed to be important, since conservatives tend to have stronger disgust reactions than liberals (Inbar, Pizarro & Bloom, 2009) and brain data suggest that conservatives and liberals process risk and fear differently (Schreiber et al., 2013). Furthermore, Hibbing, Smith and Alford (2014) found that conservatives tend to have stronger physiological responses to features of the environment that are negative and also devote more psychological resources to these stimuli.

The development of the scale was based on seven characteristics that Wilson and Patterson (1968) expected to be present in highly conservative individuals: 1) religious fundamentalism, 2) right-wing political orientation, 3) insistence on strict rules and punishments, 4) intolerance of minority groups, 5) preference for conventional art, clothing and institutions, 6) anti-hedonistic outlook, and 7) superstition and resistance to science. A large pool of more than 130 catch-phrases were created that Willson and Patterson (1968) regarded to be effective discriminators for these seven characteristics. Based on three successive item analyses (Wilson & Patterson, 1968), 50 items were selected from this pool. To control for response bias, half of the items were phrased in the affirmative direction of conservatism and half of the items were liberally phrased. Although initially conceived of as a unidimensional scale (Wilson, 1973), subsequent research on the structure of the scale showed that four factors were required to explain most of the observed variance. These factors were named 1) Militarism-punitiveness (12 items), 2) Anti-hedonism (12 items), 3) Ethnocentrism and out-group hostility (12 items) and 4) Religion-puritanism (12 items; Wilson, 1973). Eaves et al. (1999) devised a shortened and somewhat altered conservatism scale consisting of 28 items. Most of the items were taken from the original conservatism scale, with a few items added that were

regarded relevant at the time of data collection. The eigenvalues of the inter-item correlations for the 28 items suggested, according to Eaves et al. (1999), that a general "conservative - liberalism" factor was substantial but not exhaustive to explain the observed variance. An exploratory factor analysis with oblique rotation suggested that five factors explained most of the observed variance in 24 of the 28 items. These factors were named: 1) sexual permissiveness (8 items), 2) economic liberalism (5 items), 3) militarism (5 items), 4) political preference for democrats or republicans (2 items) and 5) religious fundamentalism (5 items). Note that in all these psychometric analyses, linear relationships were assumed between the items, treating the "?" response as exactly midway between a "yes" and "no" answer.

4.1.1 Prior genetic research

Using various versions of the Wilson and Patterson conservatism scale, research has shown that both, genetic and cultural, influences are responsible for the observed variance in conservatism. Based on a conservatism measure derived from the original scale (Wilson & Patterson, 1968), Martin et al. (1986) reported monozygotic (MZ) twin correlations of 0.60 for males and 0.64 for females, assessed in a large sample from the Australian Twin Registry. Eaves et al. (1997) reported on MZ and dizygotic (DZ) twin correlations across age (9.5 to 75 years). They found that prior to age 20 all variance due to individual differences was age-related, implicating environmental influences. However, after age 20, age effects vanished and there were significant differences between MZ and DZ twin correlations, suggesting genetic influences. In a later study, Eaves et al. (1999) reported heritability estimates of 0.65 for males and 0.45 for females based on the Virginia 30,000 study of twins and their relatives. Bouchard et al. (2003) assessed the 28-item conservatism scale in the Minnesota Study of Twins Reared Apart (MISTRA) and reported a heritability of 0.56. Hatemi et al. (2014) published results of a genome-wide association study (GWAS) meta-analysis, where several cohorts and, among other measures, various versions of the Wilson-Patterson conservatism scale were used. They also reported on variance components. They found a combined weighted mean of relative influences across measures and cohorts of 0.40 for genetic influences, 0.18 for common-environmental influences and 0.42 for unique-environmental influences. The GWAS meta-analysis showed no genome-wide significant hits, which may be partly related to the heterogeneity of the measures used across cohorts, but could also be due to multidimensionality of the conservatism measure (see also van der Sluis, Verhage, Posthuma & Dolan, 2010).

4.1.2 Need for psychometric evaluation

Establishing a measure with good psychometric properties is important for finding genomic signals for personality traits such as conservatism (van der Sluis et al., 2010; van den Berg & Service, 2012). Prior research on the psychometric dimensionality of the conservatism scale was based on linear factor analysis with "yes", "?" and "no" coded as 3,2 and 1 respectively, thus assuming that a "?" response is exactly midway between a "yes" and "no" response. This assumption, however, is not necessarily true - a "?" response might mean something else than being psychologically (exactly) between a "yes" and "no" answer. Reactions like "I don't know what I think" or "I don't know what is meant by busing", are psychologically different from for example an "I don't care" reaction, or a reluctance to convey the true affective response. Converse (1964) and other political scientists (e.g. Campbell, Converse, Miller & Stokes, 1960) have demonstrated that the general American public is largely uninformed about current political affairs and has gaps in knowledge of political systems. Arguably, it is likely that respondents do not understand or care about (some of) the catch-phrases of the Wilson-Patterson scale.

In addition, it is important to know to what extent scores on this scale reflect true trait variability and to what extent they reflect measurement error. Measured as split-half internal consistency, the conservatism scale has been reported to have a high reliability of 0.94 (Wilson & Patterson, 1968). This finding is supported by several studies. For example, Henningham (1996) reports alpha reliability of 0.81 on a 27-item version based on the original scale and an alpha reliability of 0.74 on a simplified and modernized 12-item version. In this paper, the psychometric properties of the conservatism scale were assessed more rigorously by using homogeneity analysis (de Leeuw & Mair, 2009) and item response theory (IRT), thereby greatly relaxing the assumption of prior research of linear relationships among items. With the establishment of a good scale, a biometric analysis including genotype-environment interaction was done.

4.1.3 Genotype-environment interaction

Genotype-environment interaction refers to the situation that some genotypes are more sensitive to changes in the environment than other or, conversely, that genotypes respond differently to the same environment (see e.g. Cameron, 1993; Martin, 2000; Sorensen, 2010). Although various studies suggest that genotype-environment interaction is an important phenomenon in complex behavioural traits (e.g., anti-social behaviour, Caspi et al. (2002); cognitive ability, Turkheimer et al. (2003) or depression, Hicks et al. (2009)), research on genotype-environment interaction has not been a focus of genetic studies on conservatism. Present study was concerned with an *omnibus* test to assess whether there is any statistically signifi-

cant genotype-environment interaction. Therefore, the method that we use here to model genotype-environment interaction is parametrized such that both, genetic as well as environmental, influences are modelled as latent (i.e., unmeasured) variables. If indeed, genotype-environment interaction is found, future research on the etiology of conservatism can focus on the exact nature of this effect by collecting specific, environmental measures at the family or individual level, depending on the results of this research.

Schwabe and van den Berg (2014; see also Molenaar & Dolan, 2014) recently developed a method that models genotype-environment interaction in such a way that statistical findings are independent of scale properties. This means that as long as a set of items measures a particular trait, such as conservatism, biometric results (i.e., conclusions regarding heritability and genotype-environment interactions) are the same regardless what particular (sub)set of items is used. This is important since it is generally recognized that statistical findings regarding non-linear effects such as genotype-environment interaction are dependent on the scale at which the analysis takes place; a simple transformation such as taking the logarithm or computing the root of a particular measure (e.g., a sum score) either obscures or reveals interaction effects (see e.g. Eaves et al., 1977; Martin, 2000; van der Sluis et al., 2006; Eaves, 2006; Molenaar et al., 2012; Schwabe & van den Berg, 2014; Molenaar & Dolan, 2014). Schwabe and van den Berg (2014; see also Molenaar & Dolan, 2014) showed that the skewness in the phenotype distribution in large part determines finding a genotype-environment effect, even when that skewness in sum scores is only due to response frequencies in the items' response categories. For instance, a relatively large proportion of "yes" responses on dichotomous yes-no questions leads to a skewed distribution of the *number* of "yes" answers (total test score). Slightly rephrasing the questions might cause no real change in item content (e.g., changing "Do you like peanut butter?" into "Do you like peanut butter very much?"), but can cause a change in proportion of yes-answers and thus change the skewness of the test score. Therefore, a rewording can lead to obscuring or revealing genotype-environment interaction effects even when the measured construct is the same. By applying the method by Schwabe and van den Berg (2014), that involves item response theory (IRT) modelling while modelling genotype-environment interaction at the level of the latent construct, our results regarding genotype-environment are free of any statistical artefacts due to response category frequencies. Still, the scale at which we model the interaction effect is arbitrary, but at least it is identified by using an IRT model. This makes our results comparable to other studies with perhaps slightly different items or a subset of the items, but where the scale was identified in the same way (i.e. the same IRT model).

4.1.4 This research

The first part of this study consists of a psychometric evaluation of the 28-item conservatism scale as used in Eaves et al. (1997, 1999), Bouchard et al. (2003) and the adult cohort in Hatemi et al. (2009). Item response models were used that take into account the categorical nature of the responses by modelling non-linear relationships between item responses and the trait being measured. In an exploratory analysis, multidimensional homogeneity models (Gifi, 1990) that assume nominal response categories were fitted in order to re-evaluate the psychometric dimensionality of the Wilson-Patterson scale. The Gifi method relaxes the assumption that a "?" answer falls exactly halfway between a "yes" and a "no" answer. Based on the results, a new scale was devised. Item response theory (IRT) models were then used to confirm the results of the homogeneity analysis and to evaluate the psychometric quality of the new scale. In the second part of this study, the new scale was used to investigate genotype-environment interaction. For the genotype-environment analysis, a Bayesian approach was used in which the biometric model and an IRT model were fitted simultaneously.

4.2 Method

4.2.1 Data

The data come from the Health and Life-Style Survey for Twins assessed in the Virginia 30K sample (Eaves et al., 1999; Hatemi et al., 2009), selecting data on twins and their parents. Part of this survey was the 28-item scale described above. Zygosity status was based on self-reported resemblance with a reported percentage correct of 95% (Eaves et al., 1999). Total sample size was 14454. Mean age was 52.13 ($SD = 17.8$, range 16-94). For the psychometric analyses in the first part of this study, we used all available data from twins and their parents that had complete data for the 28 items ($N = 12315$, of which 10405 twins). For the biometric modelling in the second part of this research we only used twin data (2795 monozygotic twin pairs, 3280 dizygotic twin pairs). Item data that was missing was assumed missing at random.

4.2.2 Part I: psychometric analyses

For the psychometric analyses, only data from twins and their parents were used with complete data on all 28 items with "no" coded as 1, "?" coded as 2 and "yes" coded as 3. Items associated with conservatism (as reported by Eaves et al. (1999), i.e. items 1 (Death penalty), 9 (Military Drill, 10 (Draft), 16 (Capitalism), 17 (Segregation), 18 (Moral Majority), 20 (Censorship), 21 (Nuclear Power), 23 (Republicans), 25 (School Prayer) and 28 (Busing)) were reverse-coded, so that a high sum score is associated

with low conservatism (high liberalism). The analyses were done using SPSS (IBM, 2013) and R (R development core team, 2013). R is an open source language and environment for statistical computing, which is freely available at `http://cran.r-project.org`.

First, using SPSS (IBM, 2013), a classical assessment of psychometric quality was performed on the scale as proposed by Eaves et al. (1999): computing item-total correlations and estimating reliability. Next, the responses were assumed nominal and a homogeneity analysis was done. Homogeneity analysis can be seen as a principal components analysis for nominal data. The analysis positions both individuals and item answer categories into one geometric space. It then uses alternating least squares to minimize the distances between the position of an item's particular answer category ("no", "?", "yes") and individuals that chose that particular category (see e.g., Heiser & Meulman, 1994; van der Kloot, 1997). Using SPSS (IBM, 2013), the dimensionality of the geometric space was determined. A two-dimensional homogeneity model was then further analysed with the R package homals (de Leeuw & Mair, 2009). Based on the homogeneity analysis results, a unidimensional conservatism scale was constructed.

The reliability of the new scale was calculated in SPSS and IRT models were used to confirm the results of the homogeneity analysis and to further evaluate the new scale. For the IRT modelling, the R package mirt (Chalmers, 2012) was used. The IRT analysis was done using a generalized partial credit IRT model (GPCM, Muraki, 1992), which is an IRT model that is suitable for polytomous, ordinal data. The GPCM model has parameters both for difficulty (i.e., thresholds) as well as discrimination parameters that are the IRT analogue of factor loadings. For our current data with three ordered response categories ("no", "?", "yes") the GPCM specifies two threshold parameters for each item, one for the location on the scale where the probability of a "?" equals the probability of a negative response, and one parameter for the location on the scale where the probability of a positive response equals the probability of a "?" response. Also a Partial Credit Model (PCM) was applied, which is a restricted version of the GPCM where the discrimination parameters (factor loadings) are all assumed equal to one. For model comparison purposes, the Akaike information criterion (AIC) and the Bayesian information criterion (BIC) were computed. At item level, goodness of fit was evaluated using chi-square statistics, comparing observed and expected response frequencies for different bins of test scores.

4.2.3 Part II: biometric analysis

In the second part of this study, the newly constructed scale was used in a biometric analysis including genotype-environment interaction. Here we follow the new method that integrates an IRT model into biometric modelling of genotype-environment interaction at the latent construct (see Schwabe & van den Berg, 2014; Molenaar & Dolan, 2014). As biomet-

ric model, the ACE model was used which decomposes total phenotypic variance, σ_P^2, into variance due to additive genetic influences (σ_A^2), variance explained by common-environmental influences (σ_C^2) and variance due to unique-environmental influences (σ_E^2). Common-environmental influences were parametrized to be perfectly correlated in a twin pair and unique-environmental influences were parametrized to be uncorrelated in one family.

Bayesian approach

van den Berg et al. (2007) showed that, in order to take full advantage of the IRT approach, both the IRT measurement model and the biometric model have to be estimated simultaneously, using a one-step approach. However, as this procedure is computationally burdensome, widespread methods of estimating variance components through structural equation modelling (SEM) reach their computational limit. van den Berg et al. (2007) showed that Bayesian statistical modelling can be an alternative to enable the simultaneous modelling of an IRT measurement and variance decomposition model. In the Bayesian approach, statistical inference is based on the posterior density of the model parameters which is proportional to the product of a prior probability and the likelihood function of the data (for further reading see e.g. Box & Tiao, 1973). Here we use Gibbs sampling (Geman & Geman, 1984; Gelfand & Smith, 1990; Gelman et al., 2004), a Markov chain Monte Carlo (MCMC) algorithm, to study the posterior densities of model parameters. This method was applied using the freely obtainable MCMC software package JAGS (Plummer, 2003). The JAGS script can be found in Appendix B. As similar syntax is used, the script can be used also in the free software package WinBUGS (Lunn et al., 2000) with minor adaptations. As an interface from R to JAGS, the R package rjags was used (Plummer, 2013).

As in Eaves and Erkanli (2003) and van den Berg et al. (2006, 2007), a Bayesian version of the ACE model was used that only specifies univariate distributions. This model is an extension of the Schwabe and van den Berg (2014) model to a (Generalized) Partial Credit model ((G)PCM, Muraki 1992) version at the measurement level. Furthermore, the model was extended to include, besides an interaction with unique-environmental influences, also an interaction with common-environmental influences, following Molenaar and Dolan (2014).

Biometric and IRT model

In the following, the full model, consisting of both variance decomposition (ACE model) and measurement model (IRT model), will be described for MZ and DZ twins separately.

We assumed that the additive genetic effect A and the common-environmental effect C are both normally distributed. Thus, we have for individual twin j from MZ twin pair i:

$$A_i \sim N(0, \sigma_A^2) \tag{4.1}$$
$$C_i \sim N(0, \sigma_{Ci}^2) \tag{4.2}$$
$$\theta_{ij} \sim N(A_i + C_i, \sigma_{Ei}^2) \tag{4.3}$$

where θ_{ij} is a person-specific latent variable that can be interpreted as the conservatism trait that is being assessed by the k items (i.e., the phenotype). To model genotype-environment interaction, we let the amount of variance due to environmental variance vary systematically with genotypic value A. Using this parametrization, we can distinguish between two different types of interactions effects. There can be an interaction with unique-environmental influences (henceforth referred to as A×E), but there can also be an interaction with common-environmental influences (henceforth referred to as A×C). To introduce A×C and A×E, σ_E^2 and σ_C^2 are portioned into an intercept and a slope parameter. This results in an estimate of σ_E^2 and σ_C^2 that is different for each twin pair i:

$$\sigma_{Ei}^2 = \exp(\beta_0 + \beta_1 A_i) \tag{4.4}$$
$$\sigma_{Ci}^2 = \exp(\gamma_0 + \gamma_1 A_i) \tag{4.5}$$

where β_0 and γ_0 denote the intercepts (i.e., unique-environmental variance when $A = 0$ and common-environmental variance when $A = 0$) and β_1 and γ_1 denote linear interaction effects (representing A×E and A×C respectively), allowing that the environmental variance components are larger at either higher or lower levels of the genotypic value. The direction of both interaction effects depends on the sign of the interaction parameters, β_1 and γ_1. The exponential function was used to avoid negative variances (c.f. SanChristobal-Gaudy et al., 1998; Bauer & Hussong, 2009; Hessen & Dolan, 2009; van der Sluis et al., 2006; Molenaar et al., 2012).

To take into account properties of the measurement scale, simultaneously with the variance decomposition, the latent phenotype θ_{ij} appeared in the GPCM (for three response categories) for observed item data on item k of twin j from family i, Y_{ijk}. This was based on the results of the first part of the study, which suggested that this model was the best fitting IRT model for our data. Let p_{ijkl} be the conditional probabilities of a particular response $l \in \{\text{“}no\text{”}, \text{“?”}, \text{“}yes\text{”}\}$ to an item k by twin member j from twin pair i, given the latent variables θ_{ij} and item parameters:

$$p(y_{ijk} = \text{“no”}|\theta_{ij}, \alpha_k, \beta_{kl}) = \left[1 + e^{\alpha_k(\theta_{ij}-\beta_{k1})}e^{\alpha_k(\theta_{ij}-\beta_{k2})}\right]^{-1} \quad (4.6)$$

$$p(y_{ijk} = \text{“?”}|\theta_{ij}, \alpha_k, \beta_{kl}) = p(y_{ijk} = \text{“no”}|.) \times e^{\alpha_k(\theta_{ij}-\beta_{k1})} \quad (4.7)$$

$$p(y_{ijk} = \text{“yes”}|\theta_{ij}, \alpha_k, \beta_{kl}) = p(y_{ijk} = \text{“no”}|.) \times p(y_{ijk} = \text{“?”}|.) \quad (4.8)$$

$$\times e^{\alpha_k(\theta_{ij}-\beta_{k2})} \quad (4.9)$$

where α_k is the discrimination parameter for item k (factor loading) and β_{kl} is the lth threshold parameter for item k, representing the item “difficulties” that an individual has to “step through” in order to reach the next response category. The conditional probabilities can be intuitively interpreted as though an individual twin “passes through” each of the preceding answer categories before finally stopping at one response category (Li & Baser, 2012). In order to identify the model, we assumed the first threshold to be zero for all items k, $\beta_{k1} = 0$, the phenotypic mean μ to be zero and set α_3 to one. We estimated the thresholds for response categories “?” and “yes”, β_{k2} and β_{k3}. Observed item data, Y_{ijk} was assumed to have a multinomial distribution:

$$Y_{ijk} \sim Multinomial(p(y_{ijk} = y|\theta_{ij}, \alpha_k, \beta_{kl})) \quad (4.10)$$

The model is similar for DZ twins, but the genetic covariance in MZ twins is twice as large as in DZ twins. To model these different genetic correlations among MZ and DZ twins, first a normally distributed additive genetic effect $A1$ was modelled and then, for each individual twin j from DZ pair i, a normally distributed additive genetic effect $A2$ was assumed. Furthermore, in order to model common-environmental influences C, we used a standard normal distribution. We then have for DZ twins:

$$A1_i \sim N(0, \tfrac{1}{2}\,\sigma_A^2) \quad (4.11)$$

$$A2_{ij} \sim N(A1_i, \tfrac{1}{2}\,\sigma_A^2) \quad (4.12)$$

$$C_i \sim N(0, 1) \quad (4.13)$$

In order to model A×C, the common-environmental effect C was scaled by multiplying it with the standard deviation σ_{Cij}, where $\sigma_{Cij}^2 = \exp(\gamma_0 + \gamma_1 A2_{ij})$, yielding a common-environmental effect $C2$ that was unique for every individual twin j from DZ pair i:

$$C2_{ij} = C_i\sqrt{\exp(\gamma_0 + \gamma_1 A2_{ij})} \quad (4.14)$$

To model A×E, the residual term was different for every individual twin:

$$\sigma_{Eij}^2 = \exp(\beta_0 + \beta_1 A2_{ij}) \quad (4.15)$$

A variance decomposition on the latent conservatism variable θ_{ij} for individual twin j from DZ pair i then yielded:

$$\theta_{ij} \sim N(A2_{ij} + C2_{ij}, \sigma^2_{Eij}) \tag{4.16}$$

As for MZ twins, simultaneous to the variance decomposition, the latent phenotype, θ_{ij}, appeared in the GPCM IRT model for three response categories (see Equations 4.6-4.9) and observed item data was assumed to have a multinomial distribution (see Equation 4.10).

Prior distributions

As prior distribution for the additive genetic variance, we chose an inverse gamma distribution ($\sigma^2_A \sim InvG(1,1)$). We chose independent normal distributions for both intercepts (β_0, $\gamma_0 \sim N(-1,2)$) as well as for both slope parameters (γ_1, $\beta_1 \sim N(0,10)$). For the item thresholds, we used a normal distribution ($\beta_k \sim N(0,10)$) and a lognormal distribution for the item discrimination parameters ($\log(\alpha_k) = N(0,10)$).

In order to find the biometric model that fits the data well and, at the same time, is parsimonious, we estimated different biometric models. These included a biometric model without any interaction effects (simple ACE model), an ACE model with one (either A×E or A×C) interaction effect and a model with both interaction effects. The deviance information criterion (DIC, Spiegelhalter et al., 2002), a measure that estimates the amount of information that is lost when a given model is used to represent the data-generating process, was calculated to assess the model fit of each model. The DIC takes account of both the complexity of a model and the goodness of fit. It can be seen as a Bayesian analogue of Akaike's information criterion (AIC). In the models without interaction effect(s), the same, independent, prior distributions were chosen for common-environmental and unique-environmental variance (σ^2_C, $\sigma^2_E \sim InvG(1,1)$).

After a burn-in phase of 20,000 iterations for each separate chain, the characterisation of the posterior distribution for the model parameters was based on a total of 120,000 iterations from six different Markov chains. This was chosen on the basis of previous test runs with multiple chains and computing Gelman and Rubin's convergence diagnostic (Gelman & Rubin, 1992). The mean and standard deviation of the posterior point estimates was calculated for each parameter as was the 95 % highest posterior density (HPD, see e.g. Box & Tiao, 1973) interval. The HPD can be interpreted as the Bayesian analog of a confidence interval (CI). When the HPD does not contain zero, the influence of a parameter can be regarded as significant.

Sum score analysis

In order to compare biometric results gained by this methodology with results gained by the sum score approach, the biometric model that was

chosen as the best model for our data was also estimated using sum scores instead of item scores. In this analysis, sum scores were calculated from the twin data with answer categories coded as 0,1 and 2 respectively and re-scaled so that they had a mean of zero and variance of one in order to make results of both approaches comparable with respect to the prior distributions. Sum scores were then analyzed with the same JAGS script (see Appendix B) but without the IRT part. After a burn-in period of 10,000 iterations, the characterisation of the posterior distribution for the sum score analysis was based on 15,000 iterations from 1 Markov chain, based on previous test runs with multiple chains and computing Gelman and Rubin's convergence diagnostic (Gelman & Rubin, 1992).

4.3 Results

Based on the original 28 item scale with reverse coding (following (Eaves et al., 1999)), the reliability estimate was 0.73 (Guttman's lambda 2; Cronbach's alpha = 0.71). Item 28 (Busing) had a negative correlation (r = -.18) with the total score, as did item 16 (Capitalism, r=-0.02).

4.3.1 Homogeneity analysis results

A homogeneity analysis was performed on the original item responses (no reverse coding). To evaluate the dimensionality of the scale, the eigenvalues associated with the first five dimensions were calculated, displayed in Figure 4.1.

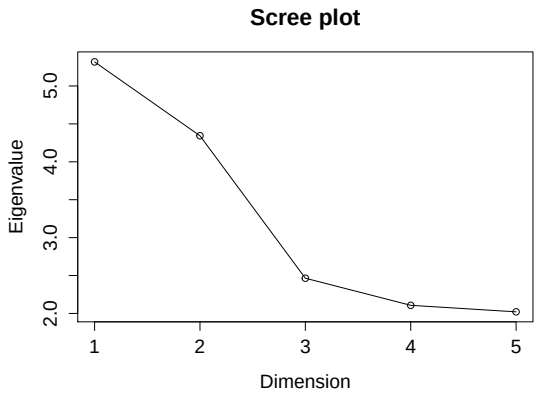


Figure 4.1: Homogeneity analysis: Scree plot that displays the eigenvalues associated with the first five dimensions.

It can be seen that, while the first two dimensions have a relatively large eigenvalue, the eigenvalues rapidly decrease when more dimensions are added to the scale. Based on these results, a model with two dimensions was chosen for further analysis. Item loadings on both dimensions can be seen in Figure 4.2.

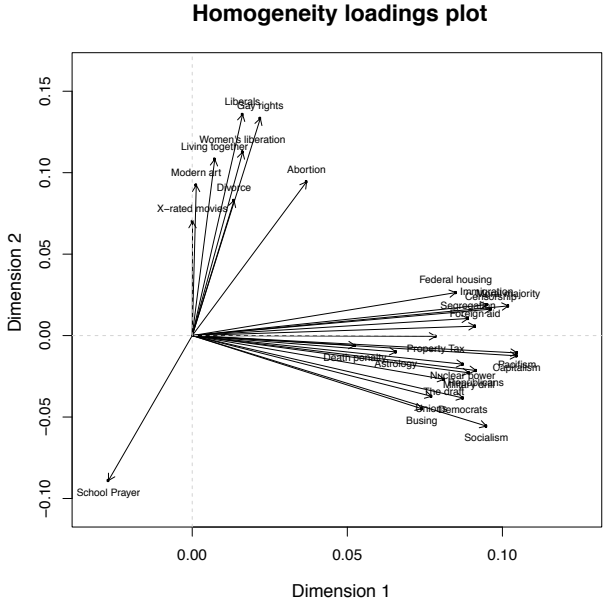


Figure 4.2: Plot of itemloadings based on a two-dimensional homals model.

It can be seen that the item "School prayer" has a negative loading on dimension 2. Furthermore, a large number of items have strong and positive loadings on dimension 1 while a smaller subset of items have positive and strong loadings on dimension 2. Thus, some items mostly discriminate among individuals along the first dimension and some items discriminate among individuals along the second dimension.

In order to interpret the two dimensions, we plotted category points for the "yes", "no" and "?" answer categories, which can be found in Figure 4.3. To save space, the category points plots of only two items are displayed, but category points plots for all items can be found in the online supplementary material of this article. A category point can be seen as the centre of gravity of all individuals who gave that particular response to a catch-phrase. When the category points are far apart on the x-axis (y-axis), this means that this item discriminates well between individuals on the first dimension (second dimension).

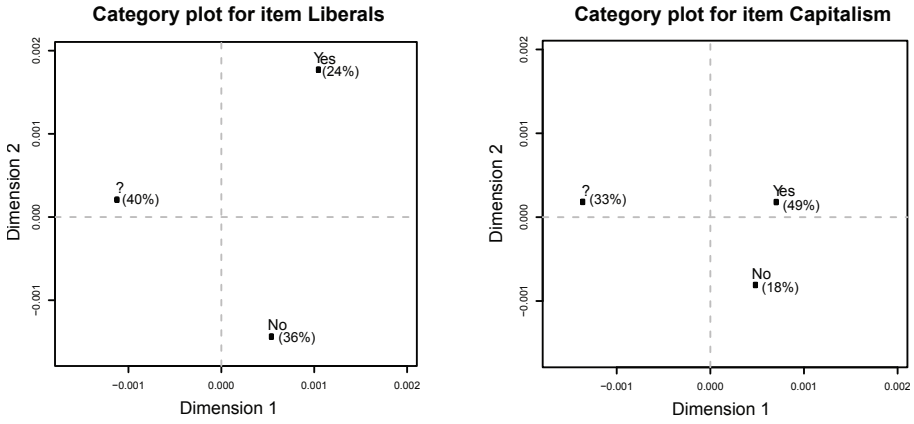


Figure 4.3: Category points plots for the catch-phrases "Liberals" (high loading on dimension 2) and "Capitalism" (high loading on dimension 1). Proportion of "yes", "?" and "no" answers were added to the plots.

Investigating the category points on the second dimension of the catch-phrase "Liberals", an item with a higher loading on the second dimension, we can see that the "?" answer category falls roughly between the "yes" and "no" answer categories. This implies that the second dimension distinguishes between liberal and conservative persons, assuming that someone with a high latent trait value for liberalism would be more inclined to approve of liberals (i.e., by answering "yes") while someone with a very low latent trait for liberalism would be more inclined to disapprove of liberals (i.e. by answering "no"). When we look at the first dimension, we can observe that the distance between the "yes" and "no" answer categories is very small, while their distance to the "?" category point is relatively large. This suggests that the first dimension mainly distinguishes between people who answered with "?" or with either "yes" or "no". When we investigate the category points plot for the catch-phrase "Capitalism", an item with a higher loading on the first dimension, we can see the same pattern, but also that, although the "?" falls roughly in between the "yes" and "no" answer categories on the second dimension, the distance between the "yes" and "no" answer categories is much smaller on this dimension (compared to the "Liberals" item) while the distance between the "?" category point and the "yes" and "no" category points is still large on the first dimension. This suggests that this item better distinguishes between people who answered with "?" and people who answered with "yes" or "no" than between conservative and liberal persons.

Interpretation of the dimensions

The category points plots suggest that the second dimension can be interpreted as liberalism (positive direction) - conservatism (negative direction) dimension while the first dimension seems to distinguish mainly between respondents with a "?" answer and respondents with either a "yes" or "no" answer. To get more insight into the nature of the two dimensions, we looked at the relationship between the itemloadings and the proportion of "?" answers on each item, which can be seen in Figure 4.4 for dimension 1 (left) and dimension 2 (right) separately.

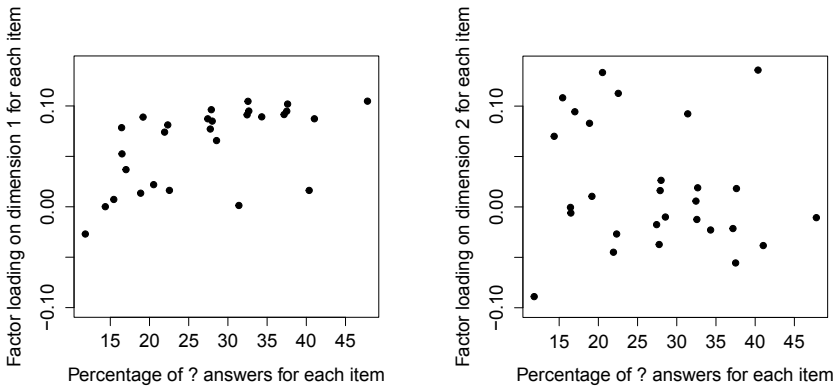


Figure 4.4: Item loadings on dimension 1 (left) and dimension 2 (right) for each item and percentage of "?" answers for each item, excluding missing data.

We can see that this relationship seems to be more or less random for the second dimension, but the proportion of "?" answers increases with higher itemloadings for dimension 1. A simple regression shows that, indeed, this relationship is positive and significant for the first dimension (T(26) = 3.75, $p < 0.01$), but non-significant for the second dimension (T(26) = -0.936, $p = 0.36$).

The relationship between itemloadings and proportion of "?" answers on each item implies that the tendency of items to distinguish between individuals with a "?" answer and respondents who either answered "yes" or "no" is based in a high level of "?" responses, which might indicate that these items measure concepts on which many people simply do not have well-formed attitudes. The interpretation of this dimension however remains unclear, as we do not know participants' true reason to give a "?" answer. Not having a well-formed attitude might indeed be the reason to give a "?" answer, but there are other possible reasons - for example, participants might not have understood what a particular catch-phrase means (e.g., item "Busing").

How can we handle this multidimensionality?

A way to handle the multidimensionality of the Wilson-Patterson conservatism scale could be to use a weighted sum score, weighting items differently based on their itemloading on the second (liberalism-conservatism) dimension. However, although it is not possible to retrieve participants' true motivation to give a "?" answer, category points plots as well as the relationship between factor loadings and the proportion of "?" answers suggest that items with a high loading on this dimension do not give much relevant information to distinguish between conservative and liberal persons. Therefore, we decided to use only items with a higher itemloading on dimension 2 (conservatism-liberalism) than on dimension 1 (ambiguous interpretation). For further psychometric analysis, we consequently selected the remaining 9 liberalism-conservatism items Gay rights, Women's liberation, Living together, Modern art, Divorce, X-rated movies, School prayer (reverse-coded), Liberals and Abortion. In all items, the "?" category point was in between the "yes" and "no" category points on the second dimension, a requirement for the application of the ordinal G(PCM) IRT model.

4.3.2 Evaluation of the new scale

Item-total correlations showed positive signs and were in the range 0.50-0.68. The reliability was estimated at 0.78 (lambda 2).

A GPCM and a PCM were estimated using the 9 items. AIC and BIC favoured the GPCM over the PCM. Table 4.1 gives the estimated GPCM model parameters.

Table 4.1: GPCM parameter estimates and item fit statistics.

	α	β_2	β_3	X^2	df	p
X-rated movies	0.58	-1.52	-1.47	253.53	27	<0.01
Modern art	0.60	0.11	0.09	232.92	27	<0.01
Women's liberation	1.00	0.22	0.80	204.93	27	<0.01
Abortion	1.01	-0.65	-0.16	112.60	27	<0.01
Gay rights	1.65	-0.90	-1.95	235.90	27	<0.01
Liberals	1.44	0.57	-0.67	357.63	27	<0.01
Living together	1.08	-0.91	-0.53	205.03	27	<0.01
Divorce	0.74	-0.44	0.23	147.16	27	<0.01
School Prayer (rev. coded)	0.65	-1.86	-2.04	109.54	27	<0.01

Item fit statistics showed the largest chi-square value for the Liberals item. Observed number of responses for each response category as well as expected number of responses for each response category under the GPCM were plotted for ordered bins of total scores (see online supplementary material). Supplementary Figure 1 shows that there is no systematic misfit

for the Liberals item: the red lines (observed number of responses) largely overlap with the corresponding black lines (number of responses predicted by the fitted GPCM), as they do for all items. Supplementary Figure 2 shows model fit based on twin data only.

4.3.3 Biometric modelling

For the biometric modelling, only twin data were used. The DIC for all fitted biometric models can be found in Table 4.2. Based on these results, the model with both interaction terms (A×E and A×C) was chosen as the preferred model for our data.

Table 4.2: Model fit (DIC) for all fitted biometric models. DIC = deviance information criterion.

Biometric model	DIC
No interaction effects (simple ACE model)	181853
ACE model with A×E	181782
ACE model with A×C	181849
ACE model with A×E and A×C	181776

The results based on the model with an A×E and an A×C interaction effect are displayed in Table 4.3.

Table 4.3: Posterior mean (standard deviation) and HPD of all parameters of the ACE model with A×E and A×C interaction effects.

	σ_A^2	$\exp(\gamma_0)$	$\exp(\beta_0)$	β_1	γ_1
Mean (SD)	0.43 (0.04)	0.29 (0.03)	0.07 (0.01)	-2.81 (0.21)	0.54 (0.14)
HPD	[0.33;0.51]	[0.22;0.35]	[0.05;0.10]	[-3.22;-2.39]	[0.31;0.84]

The results suggest substantial A and C components, as well as a negative A×E interaction effect such that individuals having low genotypic values for liberalism show *more* residual variance than individuals with high genotypic values for liberalism. The HPD interval shows that this effect is significant. Furthermore, a significant and positive A×C interaction effect was found such that individuals having low genotypic values for liberalism show *less* common-environmental variance than individuals with a high genotypic value for liberalism. The 95% credibility region for both interaction effects is displayed in Figure 4.5 for the entire range of estimated genotypic values.

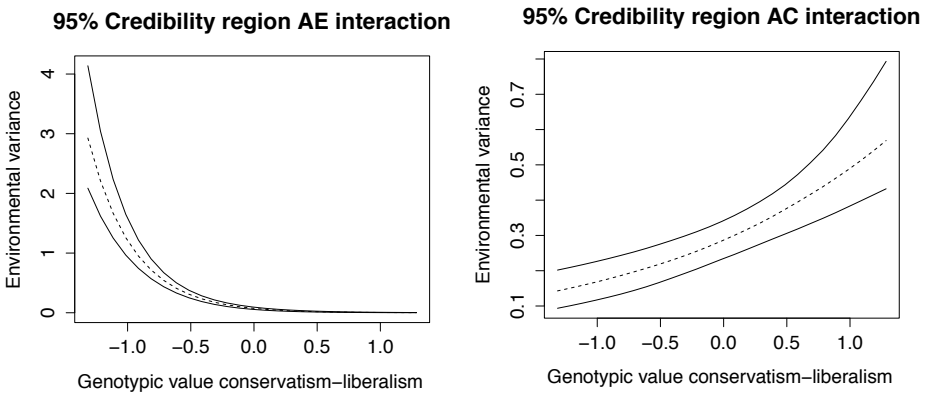


Figure 4.5: 95 % credibility region for A×E interaction (left) and A×C interaction (right), separately for each genetic value. Based on the results of the ACE model with A×E and A×C.

Defined as $\frac{\sigma_A^2}{\sigma_P^2}$ where $\sigma_P^2 = \sigma_A^2 + \exp(\gamma_0) + \exp(\beta_0)$, the ACE model with A×E and A×C interaction effects leads to a narrow-sense heritability estimate h^2 of 0.54 (HPD: [0.34;0.51]) with a standard deviation of 0.04.

Analysis of sum scores

Figure 4.6 shows the distribution of sum scores in twins. The results of the sum score approach can be found in Table 4.4.

Table 4.4: Sum score analysis: Posterior mean (standard deviation) and HPD of all parameters of the ACE model with A×E and A×C interaction effects.

	σ_A^2	$\exp(\gamma_0)$	$\exp(\beta_0)$	β_1	γ_1
Mean (SD)	0.64 (0.03)	0.35 (0.03)	0.53 (0.02)	0.10 (0.05)	-2.21 (0.09)
HPD	[0.58;0.70]	[0.29;0.41]	[0.49;0.56]	[0.00;0.21]	[-2.38;-2.05]

The sum score approach leads to a narrow-sense heritability h^2 of 0.42 (HPD: [0.38;0.46]) with a standard deviation of 0.02. Note that, conversely to the results gained by the new methodology, the sum score approach found a non-significant *positive* A×E interaction effect and a significant *negative* A×C interaction effect.

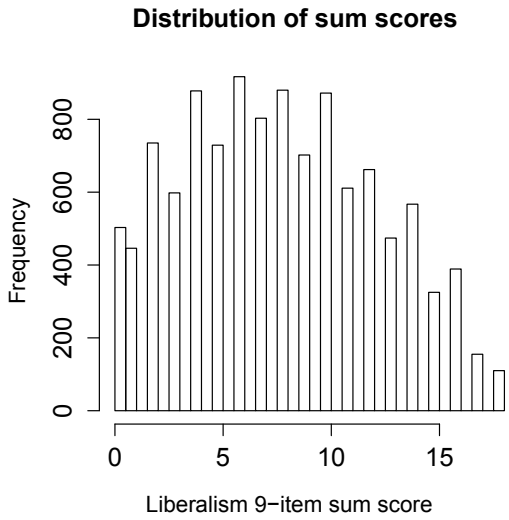


Figure 4.6: Distribution of sum scores of all MZ and DZ twins on the 9-item liberalism scale, with item "School Prayer" reverse-coded. Item categories were coded as 0,1 and 2 respectively.

4.4 Discussion

In this paper, we evaluated the Wilson-Patterson conservatism scale psychometrically before using a shorter version of the scale for a biometric analysis including genotype-environment interaction.

A psychometric evaluation of the 28-item conservatism scale (Eaves et al., 1999) showed that this scale actually measures two different aspects in people: while one set of items distinguished between people's agreement with either conservative or liberal catch-phrases, another set of items distinguished mainly between people who answered "?" or either "yes" or "no". Earlier political research (e.g. Campbell et al., 1960; Converse, 1964) has shown that most Americans are uninformed about politics and are not consistent in their agree- or disagreement with political statements. Arguably, it is likely that the dimension that differentiated mainly between respondents who answered "?" or either "yes" or "no", mainly distinguished between individuals with and without an opinion. Considering the item content of the two sets of items, this seems plausible. While this dimension mainly consists of politically-loaded items (e.g., "Property Tax", "Pacifism") and seems to measure economic liberalism, the conservatism-liberalism dimension is composed of more approachable items (e.g., "Gay Rights", "Abortion"),

likely measuring social liberalism. Exceptions are the item "Liberals" on the conservatism-liberalism dimension and the items "Astrology" and "Death Penalty" on the dimension that distinguished between individuals who answered "?" or either "yes" or "no".

To handle this multidimensionality, we decided to use a shorter version of the scale, consisting of only 9 items with a high loading on the liberalism-conservatism dimension. Ignoring the multidimensionality of the Wilson-Patterson conservatism scale can threaten validity of future or existing studies that use this scale. Therefore, we advise researchers to use the 9-item subscale as presented here rather than the full 28 items scale. Furthermore, results from the homogeneity analysis suggest that the item schoolprayer should be reverse-coded. An IRT analysis of this 9-item subscale showed good fit with a Generalized Partial Credit model and the reliability of the new scale was sufficient. The 9-item subscale is, however, constrained in the sense that the content of the remaining items reflects a measure of social liberalism rather than economic liberalism or security attitudes.

Comparing model fit and parsimony of different biometric models, an ACE model with both A×E and A×C was chosen as the best model for the data of this study. A biometric analysis that included an IRT model to correct for bias due to category response frequencies suggested a *negative* A×E and a *positive* A×C: a higher genetic propensity towards liberalism was associated with *less* unique-environmental and *more* common-environmental variance. The finding of a negative A×E effect means that the non-shared environment plays a more important role in explaining differences in individuals with a genetic tendency towards favouring conservative ideas than explaining differences in individuals genetically predisposed towards favouring liberal ideas. Conversely, the finding of a positive A×C effect means that the shared environment seems to be more important in explaining differences in individuals predisposed towards liberalism than in explaining differences in individuals predisposed towards conservatism. Arguably, genetic effects important for the expression of conservatism do not work in isolation, but instead influence the extent to which individuals are sensitive to environmental influences, favouring an interactionist framework for the study of conservatism as a personality trait.

These findings suggest that there are unique-environmental factors that affect attitudes in the conservative genotype but much less affect attitudes in the liberal genotype. Likewise, the familial environment seems to be more important in forming political attitudes in families with a genetic tendency for liberalism than in families with a genetic tendency for conservatism. These results are surprising. Conservative people are generally seen as people who do not like change; they generally favour the safety of the known over the unknown (Wilson, 1973). Research by Carney, Jost, Gosling and Potter (2008) showed that two Big Five personality traits differentiate between liberal and conservative individuals: Openness to new experiences and conscientiousness. In general, conservative participants score higher

on conscientiousness (e.g. being more conventional, orderly and better organized) whereas liberals score higher on openness to new experiences (e.g. being more curious, novelty-seeking and creative). The differences in personalities were even reflected in personal possessions and the characteristics of living and working spaces: Liberal participants collected more CDs, books, movie tickets, and travel paraphernalia, whereas conservative participants showed more sports decor, U.S. flags, cleaning supplies, calendars, and uncomfortable furniture. Based on these trait differences, one could expect that family environmental influences would be more important for individuals with a genetic tendency for conservatism than for individuals with a genetic tendency for liberalism. Likewise, it could be expected that unique-environmental influences would be more important for individuals genetically predisposed towards liberalism - with a tendency for novelty-seeking behaviour. Liberalism has been shown to be associated with higher IQ scores (Kanazawa, 2010), which predicts that conservative people generally end up in different environmental circumstances than liberal-minded people and perhaps different amounts of variation of those environmental factors that act on political views and personality. The finding of a negative A×E suggests that conservatives might come into contact with people and ideas outside of their shared environment that might be more reflective of their genetic preference. Eaves et al. (1997) indeed showed that genetic expression of conservatism-liberalism only occurs after individuals have left their parental home. Individuals with a genetic tendency for conservatism then seem to be influenced by unique-environmental influences that might affect their thinking about political issues, while, surprisingly, individuals with a genetic tendency towards liberalism, are still influenced by their family environment. Future research on genotype-environment interaction in conservatism should focus on the exact nature of both, common and unique, influences by including specific, environmental moderators, measured at the family and individual level. This can be done, for example, by using the genotype-environment parametrization introduced by Purcell (2002) by regressing moderators directly on the genotypic value.

In order to compare results gained by the new methodology with the sum score approach, the same biometric model was estimated using sum scores instead of item scores. As the sum score approach does not take into account measurement unreliability, estimated average environmental variance was much higher. Furthermore, the sum score approach suggested a positive A×E and a negative A×C interaction effect, meaning that people with a genetic tendency towards liberalism show more residual variance and less common-environmental variance than people with a genetic tendency towards conservatism. However, since the distribution of sum scores was skewed, this may be an artefact of item characteristics (see e.g. Schwabe & van den Berg, 2014; Molenaar & Dolan, 2014).

To our knowledge, this is the first study that used the Wilson-Patterson scale to investigate genotype-environment interaction in case of unmeasured

environmental variables. Regarding the testing of genotype-environment interaction in future research, we advise researchers to use the same IRT model (i.e., the GPCM) to make results concerning any interaction effects comparable. Results regarding genotype-environment interaction replicate only when the same underlying scale is used, as every transformation leads to a different result (see also Schwabe & van den Berg, 2014; Molenaar & Dolan, 2014).

In this research, the psychometric evaluation of the scale was done on all available data, of parents and offspring, enhancing statistical power. For the biometric modelling, however, only twin data was used. Unfortunately it was not possible to use a parent-offspring model for this paper, since methodology for the inclusion of a genotype-environment interaction effect in parent-offspring design with significant spouse correlation is still lacking. In future research, the method that was used in this paper will be extended to the parent-offspring design.

CHAPTER 5

Moderating Variance Decomposition at Item Level

Based on:
Inga Schwabe and Stéphanie M. van den Berg,
Manuscript in preparation [1]

Abstract

In this paper, a method is introduced that incorporates an item response theory (IRT) model into the modelling of variance decomposition moderation (ACE×M) which makes it possible to analyse raw (phenotypic) item data. The method is based on an alternative ACE×M parametrization which is uniquely identified and therefore to be preferred over the parametrization introduced by Purcell (2002). The IRT model is estimated simultaneously with the biometric model, which prevents the finding of spurious interaction effects due to scale issues such as heterogeneous measurement error. Simulation studies suggest that a large sample size is required to detect A×M and C×M interaction effects. The new method is illustrated by applying it to the data of 2110 12-year-old Dutch twin pairs to test moderating effects of a family's socio-economic status (SES) on individual differences in mathematical ability.

5.1 Introduction

Using the ACE model, the classical twin study divides total phenotypic variance in a trait into components due to additive genetic (A) influences,

[1] The twin data used in the application study comes from the Netherlands Twin Register (NTR, Boomsma et al., 2002)

common-environmental (C) influences that are shared by family members and unique-environmental (E) influences (Jinks & Fulker, 1970). This approach however ignores the possible existence of genotype-environment interaction: different genotypes might respond differently to the same environment, or conversely, some genotypes may be more sensitive to changes in the environment than others. Genotype by environment interaction can be assessed in the case that the common and unique environment feature as latent (i.e., unmeasured) variables. This provides a powerful single omnibus test to assess whether there is any statistically significant interaction. Beyond that, however, no conclusions can be drawn on the nature and importance of *specific* environmental influences. Alternatively, genotype by environment interaction can be detected using one or more *measured* moderator variable(s), which can make results very informative. A well-known finding for example is that the heritability of cognitive ability varies with socioeconomic status (Turkheimer et al., 2003; Harden et al., 2007). Having collected one ore more moderator variable(s), one can test not only for interaction effects with additive genetic influences (A×M), but also with common-environmental influences (C×M) or unique-environmental influences (E×M) - that is, moderation of variance components (henceforth referred to as ACE×M).

5.1.1 Purcell's moderation models

Purcell (2002) proposed a general method to investigate ACE×M. In the *univariate moderation model*, interaction effects are modelled directly on the path loadings of the ACE model components. That is, the variances of A, C and E are fixed to unity, but path coefficients are parametrized as $(a+\beta_a M_{ij})$, $(c+\beta_c M_{ij})$ and $(e+\beta_e M_{ij})$ respectively, where M_{ij} represents a moderator variable for individual j from twin family i. a, c and e are intercepts estimating influences of all components (A, C and E) independent of M and β_a, β_c and β_e represent regression coefficients that express the respective interaction effects (A×M, C×M and E×M). A graphical representation of this model in structural equation modelling (SEM) notation can be found in Figure 5.1.

By regressing out main effects of the moderator variable, this model is limited in the sense that any genetic or environmental effects that operate through or are common with this variable are partialled out. The moderator variable may itself be correlated with the phenotype via A, C, or E (i.e., ACE by (measured) environment *correlation*, henceforth referred to as rACE-M). The presence of rACE-M can bias estimation of heritability and mask ACE×M interaction effects. This directly motivated Purcell (2002) to introduce the *bivariate moderation model* by extending the Cholesky parametrization of the bivariate biometric model to include A×M, C×M and E×M interaction effects. So, interaction effects are considered not only

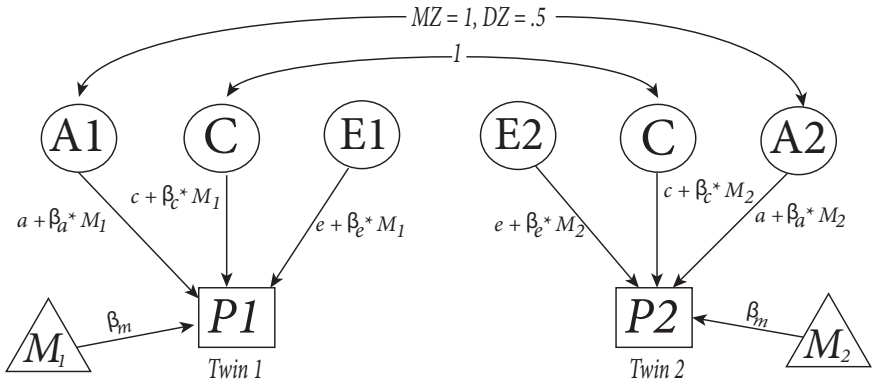


Figure 5.1: Univariate moderation model as proposed by Purcell (2002) for one twin pair displayed as structural equation model (SEM). P denotes the phenotypic value of the first ($P1$) and second twin ($P2$) and M represents the moderator value of the first ($M1$) and second ($M2$) twin. A refers to additive genetic influences for the first ($A1$) and second ($A2$) twin, correlated 1 in MZ twins and .5 in DZ twins. C represents common-environmental influences, which are the same for all twins from one family. E denotes unique-environmental influences of the first ($E1$) and second ($E2$) twin, which are parametrized as uncorrelated within one twin pair. Path loadings of the ACE model components are all divided into an intercept (a, c and e), independent of M, and a part that is dependent on M and represents the interaction effects (i.e., β_a represents A×M, β_c C×M and β_e E×M).

on influences that are unique to the phenotype, but also on influences that are common to the phenotype and the moderator variable.

This paper is concerned with the univariate modelling of ACE×M which focuses on the question whether the decomposition of variance *unique* to the phenotypic variable depends on a moderator variable. Although the univariate modelling of ACE×M is limited in the sense that it does not incorporate the possibility of rACE-M, it has some favourable statistical properties such as being more parsimonious and computationally more feasible (van der Sluis, Posthuma & Dolan, 2012).

The aim of this paper was to incorporate a measurement model into the analysis of ACE×M. This is important since statistical findings regarding non-linear effects such as ACE×M interaction effects are dependent on the scale at which the analysis takes place - a simple non-linear transformation can obscure or reveal interaction effects (see e.g. Eaves et al., 1977; van der Sluis et al., 2006; Schwabe & van den Berg, 2014; Molenaar & Dolan, 2014). The incorporation of a measurement model can overcome potential bias due to scale properties (explained in further detail below). For this extension, an alternative ACE×M parametrization was used, which will be introduced in the following.

5.1.2 Alternative ACE×M parametrization

As described in more detail above, Purcell (2002) models ACE×M by moderating the regression of the phenotype on the latent A, C and E variables such that regression paths have to be squared in order to produce a variance expectation (e.g., $\sigma^2_A = (\beta_{0a} + \beta_{1a} M_{ij})^2$ for twin j from family i). This parametrization is used in both the univariate as well as bivariate moderation model. It can however be shown that this parametrization is ill determined in the sense that there is no unique Maximum Likelihood solution for a given data set (see Appendix C for a detailed proof). This makes the interpretation of results non-trivial. For instance, confidence intervals for the moderation effects cannot be interpreted meaningfully nor the sign of the a, c, e and $\beta_a, \beta_c, \beta_e$ parameters. This is an undesirable statistical property; in particular for researchers that are new to the field of behaviour genetics and unaware of this behaviour.

Alternatively, we can parametrize ACE×M by specifying the moderator variable to modify the log-transformed variances of A,C and E (see also Tucker-Drob, Harden & Turkheimer, 2009; Turkheimer & Horn, 2013). In case of a moderator variable that is the same for every twin family i (e.g., socioeconomic status), this makes variance components different for every twin family i. For example, to model C×M, we have for every twin family i:

$$\sigma^2_{Ci} = \exp(\beta_{0c} + \beta_{1c} \boldsymbol{M}_i) \tag{5.1}$$

where $\boldsymbol{M}$ denotes a vector consisting of the moderator values for every twin family i. β_{0c} represents the intercept (estimating common-environmental variance when $\boldsymbol{M}_i = 0$) and β_{1c} represents the C×M interaction effect. The intuition behind this parametrization is that the variance cannot be negative. Contrary to Purcell's (2002) parametrization, this parametrization is uniquely identified in that there is only one maximum in the likelihood function of the model. Here, a measurement model is integrated into the biometric model including this alternative ACE×M parametrization.

5.1.3 Integration of a measurement model

Existing tests or questionnaires are usually not evenly reliable across the entire range of sum scores - that is, measurement error is *heterogeneous* across the trait continuum. It has been shown that this can lead to the finding of spurious interaction effects in the case that environmental variables are unmeasured (for more detail see Schwabe & van den Berg, 2014; Molenaar & Dolan, 2014). The finding of spurious interaction effects can also be expected in case of ACE×M where environmental variables are measured (see e.g. Tucker-Drob et al., 2009).

Schwabe & van den Berg (2014; see also Molenaar & Dolan, 2014) have shown that the incorporation of an item response theory (IRT) measurement

model into the biometric analysis can overcome potential bias due to scale properties; results regarding interaction effects are free of artefacts due to heterogeneous measurement error across the trait continuum. Further advantages of the IRT approach include the flexible handling of missing data and the harmonization of traits measured on different measurement scales (see e.g. van den Berg et al., 2014). For example, when different twin registers have used different IQ tests and these are not comparable in difficulty, IRT can be used to set the scores on the same scale. The IRT approach is introduced in the following.

Item response theory

The IRT approach is model-based measurement in which a twin's latent trait (e.g., mathematical ability) is estimated using not only trait levels (e.g., performance on a mathematics test) but also test item properties such as the difficulty of each item are used as information. The simplest IRT model is the so called Rasch model, also known as the one-parameter logistic model (1PLM). In the Rasch model, the probability of a correct answer to item k (e.g. on a mathematics test) by twin j from family i, $P(Y_{ijk} = 1)$, is modelled as a logistic function of the difference between the twin's latent trait score (e.g. mathematical ability) and the difficulty of the item:

$$P(Y_{ijk} = 1) = \frac{\exp(\theta_{ij} - b_k)}{1 + \exp(\theta_{ij} - b_k)} \tag{5.2}$$

where θ_{ij} denotes the latent trait score of individual twin j from family i. b_k estimates the difficulty of item k which represents the trait level associated with a 50% chance of answering an item correctly. This IRT model is suitable for dichotomous data (e.g., scored as correct = 1 and false = 0), as for example collected from ability tests (e.g. mathematical ability). An underlying assumption of the Rasch model is that all items discriminate equally well between varying abilities. An extension of the Rasch model, the two-parameter model (2PLM), estimates discrimination parameters (comparable to factor-loadings) that differ across items (ee e.g. Embretson and Reise, 2009). There are several IRT models that can be used for non-dichotomous data such as ordered categories (e.g. Likert scale data). In this paper, the Rasch model was used, but extensions to the 2PLM or ordinal IRT models are straightforward.

5.1.4 Earlier research

Tucker-Drob et al. (2009) proposed to, next to the biometric ACE×M model, explicitly model the factor structure of the phenotype at the psychometric level. By means of a simulation study, they show how ignoring non-linearity in the factor structure can lead to the finding of spurious interaction

effects, further highlighting the need for a new method that integrates a measurement model into the ACE×M biometric model. However, they do not validate their approach using a simulation study, but only demonstrate it by re-analysing ACE×M on IQ data previously used by Turkheimer et al. (2003).

As sum score data were available for twelve separate cognitive tests, it was tested whether there is a single factor common to all twelve tests. For this psychometric analysis, the authors used Mplus. Subsequently, the ACE×M analysis and a non-linear factor structure were modelled using the Markov chain Monte Carlo (MCMC) estimation program WinBUGS (Lunn et al., 2000), where factorloadings were set to the values that had been retained from the Mplus output. As only sum scores were available, they were unable to perform item level analyses on the IQ data to for example determine test reliability, check for ceiling and floor effects or use IRT modelling. Furthermore, by estimating factorloadings separately, error might be introduced in the factorloadings, and consequently, uncertainty on the latent scores is not taken into account. In order to take full advantage of the measurement model (e.g. IRT model), both, biometric and measurement model, have to be estimated *simultaneously* (van den Berg et al., 2007). This has been done by Schwabe & van den Berg (2014, see also Molenaar & Dolan 2014) to investigate genotype by environment interaction in case of unmeasured variables. By using an item response theory (IRT) model on phenotypic item data, the interaction is modelled at the level of the latent phenotypic construct. It remains, however, unclear how this solution performs in case of measured moderators (i.e., ACE×M).

5.1.5 This research

In this paper, we introduce an MCMC method that fits the biometric ACE×M model and the IRT model simultaneously, taking full advantage of the IRT approach. A simulation study was conducted to evaluate the performance of the new method. The method is illustrated by applying it to the data of 2110 12-year-old Dutch twin pairs to test moderator effects of a family's socio-economic status (SES) on the etiology of mathematical ability. In the following, the full model, consisting of a biometric and a psychometric part, is presented in more detail for monozygotic (MZ) and dizygotic (DZ) twins separately. This model is restricted to the case where the moderator variable is the same for every twin pair. Extending the model to include moderator variables that are measured separately for every individual twin leads to a slightly different parametrization (for more details see the discussion and for a script that can be used to fit the model see Appendix F).

5.2 Full model

MZ twins

For each MZ twin pair i, we assumed that the effect of common-environmental influences is perfectly correlated within the pair and normally distributed with an expected value consisting of the phenotypic population mean, μ, and the main effect of the moderator variable $\boldsymbol{M}$. Furthermore, familial influences F, consisting of additive genetic influences as well as common-environmental influences, were assumed to be normally distributed for every family i:

$$C_i \sim (\mu + \beta_{1m}\boldsymbol{M}_i, \sigma^2_{Ci}) \tag{5.3}$$
$$F_i \mid C_i \sim N(C_i, \sigma^2_{Ai}) \tag{5.4}$$

where μ represents the phenotypic population mean and β_{1m} represents a regression coefficient that expresses the estimated main effect of the moderator variable. In order to model A×M and C×M, for every MZ family i, variance components were divided into an intercept (representing variance components when $\boldsymbol{M}_i = 0$) and a linear interaction term (denoting A×M and C×M respectively):

$$\sigma^2_{Ai} = \exp(\beta_{0a} + \beta_{1a}\boldsymbol{M}_i) \tag{5.5}$$
$$\sigma^2_{Ci} = \exp(\beta_{0c} + \beta_{1c}\boldsymbol{M}_i) \tag{5.6}$$

where β_{0c} (β_{0a}) represents common-environmental (additive genetic) variance when $\boldsymbol{M}_i = 0$ and β_{1c} and β_{1a} represent A×M and C×M respectively.

The expected value of the phenotypic trait θ_{ij} of individual twin j from family i then consisted of the familial effect:

$$\theta_{ij} \sim N(F_i, \sigma^2_{Ei}) \tag{5.7}$$

In order to model E×M, variance due to unique-environmental influences was different for every MZ family i and divided into an intercept (unique-environmental variance when $\boldsymbol{M}_i = 0$) and an interaction term:

$$\sigma^2_{Ei} = \exp(\beta_{0e} + \beta_{1e}\boldsymbol{M}_i) \tag{5.8}$$

In the psychometric part of the model, the probabilities for correct item responses, P_{ijk}, were modelled in the Rasch model, conditional on θ_{ij}:

$$\ln(P_{ijk}/(1 - P_{ijk})) = \theta_{ij} - b_k \tag{5.9}$$
$$Y_{ijk} \sim Bernoulli(P_{ijk}) \tag{5.10}$$

where b_k refers to the difficulty parameter of item k. In the simulation study, it was assumed that item parameters were known.

DZ twins

Similar to MZ twin pairs, a normal distribution was used to model a common-environmental effect for every DZ family i that is specified to be perfectly correlated within a twin pair with an expected value consisting of the phenotypic mean and the main effect of the moderator variable $\boldsymbol{M}$:

$$C_i \sim N(\mu + \beta_{1m}\boldsymbol{M}_i, \sigma^2_{Ci}) \tag{5.11}$$

Similar to MZ twin pairs, variance due to common-environmental influences was divided into an intercept and a part that represents C×M ($\sigma^2_{Ci} = \exp(\beta_{0c} + \beta_{1c}\boldsymbol{M}_i)$, see Equation 6).

While the total genetic variance is assumed to be the same for DZ and MZ twins, the genetic covariance in MZ twins is twice as large as in DZ twins, as DZ twin pairs share on average only 50% of their genomic sequence. To model a genetic correlation of .5 for DZ twins, first a normal distributed additive genotypic value was assumed for each DZ family i. Then, for each individual twin j from family i, a normally distributed additive genotypic value was assumed, representing the Mendelian sampling term:

$$A1_i \sim N(0, \frac{1}{2}) \tag{5.12}$$

$$A2_{ij} \sim N(A1_i, \frac{1}{2}) \tag{5.13}$$

Then, the genotypic value was scaled by multiplying it with the standard deviation σ_{Ai}, where $\sigma^2_{Ai} = \exp(\beta_{0a} + \beta_{1a}\boldsymbol{M}_i)$. This yielded a genetic effect $A3_{ij}$ that was unique for every individual twin j from DZ family i:

$$A3_{ij} = A2_{ij}\sqrt{\exp(\beta_{0a} + \beta_{1a}\boldsymbol{M}_i)} \tag{5.14}$$

The expectation of the phenotypic variable, θ_{ij}, then consisted of the common-environmental effect for every DZ family i and the additive genetic effect for every individual twin j:

$$\theta_{ij} \sim N(C_i + A3_{ij}, \sigma^2_{Ei}) \tag{5.15}$$

where, as for MZ families, σ^2_{Ei} was divided into an intercept (representing unique-environmental variance when $\boldsymbol{M}_i = 0$) and an interaction term ($\sigma^2_{Ei} = \exp(\beta_{0e} + \beta_{1e}\boldsymbol{M}_i)$, see also Equation 5.8). Furthermore, in the psychometric part of the model, a Rasch model was applied (see Equation 5.9 and 5.10) of which the item parameters were assumed to be known.

5.2.1 Estimation of the model

We used Bayesian statistical modelling to estimate both, the psychometric and the biometric models, simultaneously. In the Bayesian framework,

statistical inference is based on the so-called joint *posterior* density of all model parameters. The posterior density is proportional to the product of the likelihood function and a prior probability density for unknown parameters (for an introduction to Bayesian statistics, see e.g. Bolstad, 2007). We applied Gibbs sampling (Geman & Geman, 1984; Gelfand & Smith, 1990; Gelman et al., 2004), a Markov chain Monte Carlo (MCMC) algorithm. Gibbs sampling works by iteratively drawing samples from the full conditional distributions of all unobserved parameters of a model. The full conditional distribution refers to the distribution of a parameter given the current or known values of all other relevant parameters in the model (see e.g. Gilks et al., 1996). A sample from the full conditional distribution is taken in every iteration of the Gibbs sampling. After a number of so-called "burn-in" iterations, subsequent draws can be regarded as draws from the joint posterior distribution. For the MCMC estimation, we used the freely available software package JAGS (Plummer, 2003). The JAGS script that was used for above described model can be found in Appendix D.

5.2.2 Prior distributions

As the above described model was fitted using Bayesian statistics, prior distributions had to be specified. We used independent normal distribution for all intercepts (β_{0a}, β_{0c} and $\beta_{0e} \sim N(-1, 2)$) and interaction effects (β_{1a}, β_{1c} and $\beta_{1e} \sim N(0, 10)$). Also for the phenotypic population mean and the regression coefficient that expresses the main effect of the moderator variable, independent normal distributions were chosen as prior distributions ($\mu \sim N(0, 10)$ and $\beta_{1m} \sim N(0, 10)$). As non-informative priors were used in the biometric part of the model, the Bayesian approach will yield comparable results as the maximum likelihood framework.

5.3 Simulation study

A simulation study was conducted to evaluate the feasibility of the above described method. In each condition, 250 datasets were simulated and the ratio of MZ (28% of all pairs) and DZ (72% of all pairs) twin pairs was the same, reflecting the usual ratio of MZ and DZ twin pairs in European twin registers. Furthermore, the phenotypic population mean, μ, was fixed to zero. Using similar values as Purcell (2002), $\exp(\beta_{0a})$ was set to 0.25, $\exp(\beta_{0c})$ was 0.25 and $\exp(\beta_{0e})$ was fixed to 0.5. All interaction effects (β_{1a}, β_{1c} and β_{1e}) were set to 0. For every MZ and DZ family, a dichotomous moderator value was simulated such that the moderator explained approximately 11% of the total phenotypic variance (with β_{1m} fixed to 0.7). Item difficulty parameters were simulated equally spaced within the interval [-3.2;3.2]. The minimum and maximum value of this interval were based on (minus) three times the standard deviation of all phenotypic values. The Rasch model

was used to simulate responses to phenotypic dichotomous items. The simulated datasets were then analysed using the above described method. Item difficulty parameters were assumed to be known.

In order to assess the impact of the number of twin pairs and items, first we varied the number of twin pairs (1000, 2000, 4000 and 80000 twin pairs) with a fixed number of 60 dichotomous items. Next, we varied the number of items (20, 100 and 250 items) while fixing the number of twin pairs to 1000. Cronbach's alpha was 0.90 when responses to 60 items were simulated and 0.75 in case of 20 items.

All simulations were carried out using the software package R (development core team, 2013), an open-source language for statistical computing. As an interface from R to JAGS, the R package rjags was used (Plummer, 2013). After an adaption phase of 5,000 iterations and a burn-in phase of 50,000 iterations, the characterisation of the posterior distribution for the model parameters was based on an additional 25,000 iterations from 1 Markov chain. The average posterior means of all model parameters as well as the standard deviation of posterior means and the means of all posterior standard deviations were calculated. The mean of posterior standard deviations can be seen as the Bayesian version of the standard error.

5.3.1 Results

Part I: Varying number of twin pairs

The results of the first part of the simulation study can be found in Table 5.1.

It can be seen that, with only 1000 twin pairs, β_{1m} and $\exp(\beta_{0e})$ were estimated at the true value (β_{0e}) or very close to the true value (β_{1m}). Average posterior means of $\exp(\beta_{0a}), \exp(\beta_{0c})$ and β_{1e} were close to their true values with only a slight bias and reasonably small standard deviations and standard errors. Although β_{1a} and β_{1c} were close to their true values, their standard deviations as well as standard errors were relatively large. Furthermore, standard deviations and standard errors for β_{1a} and β_{1c} were different, which might be an indication that the Markov chain did not mix well. Average posterior means were closer to their true values when sample size increased. Also, standard deviations as well as standard errors decreased with larger sample size. Although all other parameters were estimated at or very close to their true value when the sample consisted of 8000 twin pairs, $\exp(\beta_{0a})$ was biased even in this condition. Furthermore, with 8000 twin pairs, standard deviations and errors of β_{1a} and β_{1c} were still large, which suggests that large sample sizes are needed to have sufficient power to detect A×M and C×M interaction effects.

Table 5.1: Results of the simulation study, part I (differing number of twin pairs). Posterior means (SD) averaged over 250 replications. Second line: Mean of posterior standard deviations. N refers to the number of twin pairs. The number of phenotypic items was fixed to 60.

	β_{1m}	$\exp(\beta_{0a})$	$\exp(\beta_{0c})$	$\exp(\beta_{0e})$	β_{1a}	β_{1c}	β_{1e}
True value	0.70	0.25	0.25	0.50	0.00	0.00	0.00
N = 1000	0.69 (0.06)	0.23 (0.09)	0.21 (0.07)	0.50 (0.04)	-0.02 (0.98)	-0.05 (0.91)	0.01 (0.18)
	0.06	0.11	0.08	0.04	1.35	1.09	0.18
N = 2000	0.70 (0.04)	0.21 (0.08)	0.23 (0.06)	0.50 (0.03)	0.00 (0.87)	-0.02 (0.68)	0.00 (0.12)
	0.04	0.09	0.07	0.03	1.13	0.76	0.14
N = 4000	0.70 (0.03)	0.22 (0.07)	0.24 (0.05)	0.50 (0.03)	-0.01 (0.72)	-0.02 (0.46)	0.00 (0.11)
	0.03	0.07	0.05	0.02	0.79	0.46	0.10
N = 8000	0.70 (0.02)	0.23 (0.05)	0.25 (0.04)	0.50 (0.02)	0.00 (0.47)	-0.03 (0.28)	0.00 (0.07)
	0.02	0.05	0.03	0.02	0.48	0.27	0.07

Part II: Varying number of items

The results of the second part of the simulation study can be found in Table 5.2.

Table 5.2: Results of the simulation study, part II (differing number of items). Posterior means (SD) averaged over 250 replications. Second line: Mean of posterior standard deviations. Ni refers to the number of items. The number of twin pairs was fixed to 1000.

	β_{1m}	$\exp(\beta_{0a})$	$\exp(\beta_{0c})$	$\exp(\beta_{0e})$	β_{1a}	β_{1c}	β_{1e}
True value	0.70	0.25	0.25	0.50	0.00	0.00	0.00
Ni = 20	0.70 (0.06)	0.24 (0.09)	0.20 (0.07)	0.49 (0.06)	0.02 (0.99)	-0.01 (0.90)	-0.02 (0.22)
	0.06	0.12	0.09	0.06	1.74	1.24	0.25
Ni = 100	0.70 (0.05)	0.23 (0.09)	0.21 (0.07)	0.50 (0.04)	-0.03 (0.95)	-0.02 (0.83)	0.01 (0.17)
	0.05	0.10	0.08	0.04	1.92	1.07	0.17
Ni = 250	0.70 (0.05)	0.21 (0.08)	0.22 (0.06)	0.50 (0.04)	-0.02 (0.92)	0.02 (0.80)	0.01 (0.14)
	0.05	0.10	0.08	0.04	1.32	0.97	0.16

It can be seen that, with only 20 items, average posterior means of most parameters were close to their true values with bias in $\exp(\beta_{0c})$. This precision is comparable to the results of the first simulation study (1000 twin pairs) where the number of items was fixed to 60. There was only a

small decrease in standard deviations and standard errors with increasing number of items. Also the increase in precision with increasing sample size was small, suggesting that as much as 20 items are sufficient to find interaction effects.

5.4 Application

The model above described was used to investigate ACE×SES in mathematical ability of 2110 12-year-old Dutch twin pairs. Mathematical ability was assessed using the 60 items from the mathematics subscale of a Dutch national educational achievement test, the *Eindtoets Basisonderwijs*, which is administered in the last year of primary education.

5.4.1 Data

The data of 12-year-old twins from the Netherlands Twin Register (NTR, Boomsma et al., 2002) from birth cohorts 1998-2000 were used to link individual twins to their dichotomous item scores (coded as 0=incorrect, 1=correct) on the mathematics subscale of the *Eindtoets Basisonderwijs* test. The linking procedure and harmonization of different test versions is described in detail in Chapter 3. This led to a total of 4220 individual twins, forming 2110 twin pairs of which 581 were MZ pairs and 1529 DZ pairs. Of the MZ twins, 299 were female and 282 pairs were male. The DZ twins consisted of 360 male pairs, 309 female pairs and 860 were of opposite sex. For 711 twins, item scores were unknown. The reasons that the scores were missing were either that the child had not reached final grade yet (N twins = 52), the child was attending special education (N twins = 34), a different test was used at the school the twin was attending (N twins = 13), the child (N twins =2) or the whole school (N twins = 1) did not attend the test or the reason was unknown (N twins = 609).

Family SES

The NTR collects longitudinal data from all registered twins by mailed surveys. Among other information, parents are asked for the family SES operationalized as highest parental education. For this application, family SES scores measured at ages three, seven and ten were used. Family SES was scored in five different categories that approximately translate to: 1) *"Unskilled labor"*, 2) *"Job for which lower vocational education is required"*, 3) *"Job at medium level"*, 4) *"Job at college level"* and 5) *"Job at university level"*.

In order to gain statistical power, the data of all ages were combined into one measure. The lowest correlation between the different SES scores was 0.72 between SES at age three and SES at age ten. Family SES scores were available for a total of 1708 families (81%). Of the complete data, the

scores of 334 families (16%) were measured at age ten, the scores of 831 (39%) families at age seven and of 543 (26%) families at age three.

5.4.2 Analysis

For an easier interpretation, SES categories were summarized into one dichotomous dummy variable, coded as 0 (all cases with a score of or lower than three on family SES) and 1 (all cases with a score of or higher than four on parental SES). We interpret these two categories as "families with average or low SES" (coded as 0) and "families with high SES" (coded as 1). 720 (34%) families had a high family SES and 988 (47%) an average or low SES (see Table 5.3 for more detail).

Table 5.3: Total number (percentage) of twin pairs with high SES, average or low SES or missing data.

	Total N of pairs (%)		
	High SES	Average or low SES	Missing data
MZ twin pairs	202 (35%)	282 (48%)	97 (17%)
DZ twin pairs	518 (34%)	706 (46%)	305 (20%)
All pairs (MZ + DZ)	720 (34%)	988 (47%)	402 (19%)

As measurement model for the mathematics item scores, the one parameter logisitc model (OPLM, Verhelst, Glas, & Verstralen, 1995) version of an IRT model was used. In the OPLM, item difficulty parameters, b_k, are estimated and item discrimination parameters, a_k are imputed as known constants. Item parameters were known beforehand and imputed as known parameters (for more details see Chapter 3).

As there were twins with missing SES data, independent Bernoulli distributed prior distributions were defined ($\boldsymbol{SES}_i \sim Bernoulli(\pi)$). On the probability, π, a Beta distributed hyperprior was used, which was different for MZ and DZ twins ($\pi_{mz} \sim Beta(1,1)$ and $\pi_{dz} \sim Beta(1,1)$). After a burn-in phase of 50,000 iterations, the characterisation of the posterior distribution for the model parameters was based on a total of 175,000 iterations from five different Markov chains. The posterior means and standard deviations were calculated for each parameter as was the 95% highest posterior density (HPD, see e.g. Box & Tiao, 1973) interval. The HPD can be interpreted as the Bayesian analogue of a confidence interval (CI). When the HPD does not contain zero, the influence of a parameter can be regarded as significant. This however does not hold for the variance components of this particular application, as these are bounded at zero due to a very low phenotypic variance.

5.4.3 Results

The posterior means for the intercepts, interaction effects and estimated heritability, h^2, can be found in Table 5.4. Histograms of the posterior distributions of all interaction effects can be seen in Figure 5.2.

Table 5.4: Posterior means (SD) of all relevant parameters. HPD refers to the *95% highest posterior density interval*.

	Posterior point estimate (SD)	HPD
β_{1m}	0.1078 (0.0126)	[0.0833;0.1329]
$\exp(\beta_{0a})$	0.0511 (0.0048)	[0.0410;0.0598]
$\exp(\beta_{0c})$	0.0051 (0.0029)	[0.0008;0.0107]
$\exp(\beta_{0e})$	0.0159 (0.0023)	[0.0117;0.0204]
β_{1a}	0.1132 (0.1263)	[-0.1364;0.3627]
β_{1c}	-3.0777 (1.9280)	[-7.3886;-0.0300]
β_{1e}	-0.1958 (0.2426)	[-0.6769;0.2759]
h^2	0.7088 (0.0565)	[0.5800;0.7995]

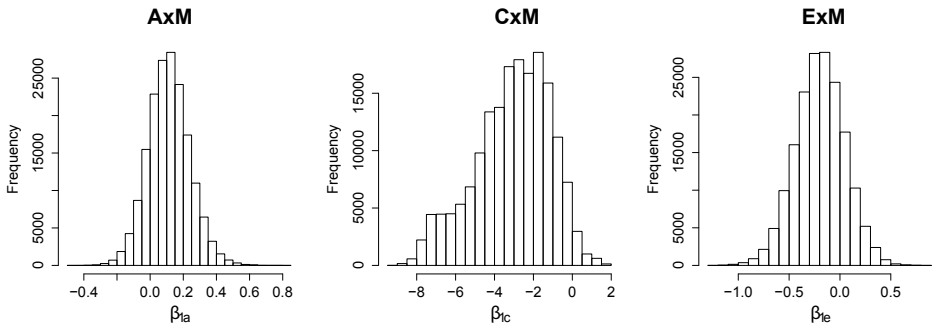


Figure 5.2: Application: Moderating effects of a family's SES on individual differences in mathematical abilty. Histograms of the posterior distributions of β_{1a} (A×SES, left), β_{1c} (C×SES, middle) and β_{1e} (E×SES, right).

Heritability was defined as $\frac{\exp(\beta_{0a})}{\sigma_P^2}$, where $\sigma_P^2 = \exp(\beta_{0a}) + \exp(\beta_{0c}) + \exp(\beta_{0e})$. The results suggest that the largest part of phenotypic variance could be explained by genetic influences, a substantial part by unique-environmental influences and a negligibly small part by common-environmental

influences. Furthermore, a significant and negative C×SES interaction effect was found, meaning that common-environmental influences are less important in creating individual differences in mathematical ability in families with a high SES than in creating individual differences in mathematical ability in twin pairs with a low or average SES.

5.5 Discussion

In this paper, the biometric modelling of ACE×M was extended with an IRT model at the phenotypic level. This prevents the spurious finding of interaction effects due to scale issue such as heterogeneous measurement error. For this extension, a different ACE×M parametrization than introduced by Purcell (2002) was used.

A simulation study was conducted to evaluate the new method. First, the sample size was varied with fixed number of phenotypic items and second, the number of items was changed while the sample size was fixed. Results show that large sample sizes are required to have sufficient power to find A×M and C×M interaction effects while only a small number of phenotypic items (20) is required. The low power might be related to the well known problem that A and C are less well resolvable compared to A and E, or C and E (Martin, Eaves, Kearsey & Davies, 1978). A possible approach to improve the distinction of A and C proposed by Jinks and Fulker (1970) is to include data of MZ twins who are reared apart. Another way to increase power would be to fit submodels consisting of only one interaction term (i.e., an ACE model including either an A×M or C×M interaction effect). However, an earlier application of the method described by Schwabe & van den Berg (2014) to model genotype-environment interaction in case of unmeasured environmental influences has shown that the direction of interaction effects can change when not all interactions are modelled. In the application, a submodel consisting of only an A×C interaction effect resulted in a negative effect, while, when both interaction effects were modelled, the A×C effect was positive. Similar results can also be expected in case of ACE×M and therefore it is not advised to fit submodels. Due to the often very large sample sizes available in twin registers, the low power is a drawback, but the method can still be used. Furthermore, the IRT approach can be used to link item data from different inventories, cohorts or twin register, which increases sample size and therefore enhances statistical power to detect interaction effects.

In this paper, a different ACE×M parametrization than introduced by Purcell (2002) was used. The parametrization of Purcell (2002) is not fully identified and therefore, results are difficult to interpret. Here, we model ACE×M on the log-transformed variances. This parametrization is uniquely identified and therefore to be preferred over the parametrization introduced by Purcell (2002). While the parametrization of Purcell (2002) results in

a parabolic form of interaction effects, the alternative parametrization is either monotonically increasing or decreasing with respect to the moderator, leading to an easier (biological) interpretation (see also Turkheimer & Horn, 2013).

The method described in this paper can be used to model ACE×M when moderator variable values are the same for every twin family. Having separate values for both twins slightly changes the ACE×M parametrization. Furthermore, van der Sluis et al. (2012) showed that, in case of a moderator variable that is measured for both members of a twin pair, spurious interaction effects can be expected when M and P are correlated and moderator values of one family are correlated as well. van der Sluis et al. (2012) showed that this potential bias can be prevented by extending the expectation of the phenotype such that the trait value of an individual twin is corrected for his or her own moderator value but also for the moderator value of the co-twin. Furthermore, regression parameters as well as the phenotypic mean should be estimated separately in MZ and DZ twin pairs, allowing their values to differ across zygosity (see van der Sluis et al., 2012). A JAGS script that extends the ACE×M model such that separate moderator values can be used can be found in Appendix E.

An IRT measurement model was used to model ACE×M at the latent phenotypic level. That is, ACE×M is modelled on phenotypic trait scores that are corrected for measurement error. However, measurement error might not only appear at the level of the phenotype but also in the measurement of the *moderator variable* as often self-reports (e.g. questionnaires) are used to gather information on the environment of a family or an individual twin. To correct for measurement error in the moderator level, in future research, the method will be extended to include an IRT model also at the level of the moderator variable.

A drawback of the method is that it is computationally intensive and, depending on the number of items and twin pairs, can take several hours to complete. In future research, more efficient sampling algorithms will be applied to lighten the computation burden.

CHAPTER 6

A New Approach to Handle Missing Covariate Data in Twin Research

With an Application to Educational Achievement Data

Based on:
Inga Schwabe, Dorret I. Boomsma, Eveline L. de Zeeuw and Stéphanie M. van den Berg, *Behavior Genetics*, *in press*

Abstract

The often-used ACE model which decomposes phenotypic variance into additive genetic (A), common-environmental (C) and unique-environmental (E) parts can be extended to include covariates. Collection of these variables however often leads to a large amount of missing data, for example when self-reports (e.g. questionnaires) are not fully completed. The usual approach to handle missing covariate data in twin research results in reduced power to detect statistical effects, as only phenotypic and covariate data of individual twins with complete data can be used. Here we present a *full information* approach to handle missing covariate data that makes it possible to use *all* available data. A simulation study shows that, independent of missingness scenario, number of covariates or amount of missingness, the full information approach is more powerful than the usual approach. To illustrate the new method, we applied it to test scores on a Dutch national school achievement test (*Eindtoets Basisonderwijs*) in the final grade of primary school of 990 twin pairs. The effects of school-aggregated measures (e.g. school denomination, pedagogical philosophy, school size) and the

effect of the sex of a twin on these test scores were tested. None of the covariates had a significant effect on individual differences in test scores.

6.1 Introduction

In the genomics era, twin studies remain useful to estimate the relative importance of genetic and environmental influences on individual differences. In the often-used ACE model, the total variance of a trait (e.g. mathematical ability) is decomposed into components due to additive genetic (A) influences, common-environmental (C) influences that are shared by family members and unique-environmental (E) influences (Jinks & Fulker, 1970). This model can be extended to include covariates. Figure 6.1 is an example of the structural equation model (SEM) for a basic univariate twin analysis extended with three covariates (denoted as x_{11}, x_{12} and x_{13} for the first twin and x_{21}, x_{22} and x_{23} for the second twin of one family). The path coefficients β_1, β_2 and β_3 represent regression coefficients that express the estimated effect of the respective covariate. This model implies that the ACE variance decomposition takes place on the residuals of the phenotypic scores, after the effects of the covariates have been partialled out.

6.1.1 Missing covariate data

The collection of covariate data however often leads to a high amount of missingness. For example, when self-reports (e.g. questionnaires) are used to gather information on the environment of a family or an individual twin, they are often not fully completed (e.g. the last items are skipped) or items on sensitive topics (e.g. alcohol or drug use) are not answered. Likewise, the linkage of two datasets may lead to missing data. A twin researcher might want to link twin data from a twin registry to data from the same twins from another (external) source. For example, an environmental variable such as the socio-economic status of a neighbourhood might not be available in the twin registry, but there is a publicly available dataset from a governmental or local organisation which includes the desired variable. For the linking of the two datasets, usually, a common identifier such as the name or address of a family or individual twin can be used. However, this potentially leads to a lot of missing data, for example when entities cannot be (uniquely) linked to the common identifier, as may be the case due to differences in record shape or choice of identification variables.

We can distinguish between three different mechanisms that describe relationships between measured variables and the probability of missing data (Rubin, 1976; Little & Rubin, 2002). Data are said to be *missing completely at random* (MCAR) when the probability that a value is missing is unrelated to both observed and unobserved data. For example, a respondent might flip a coin to decide whether to answer a questionnaire item or not. Note that

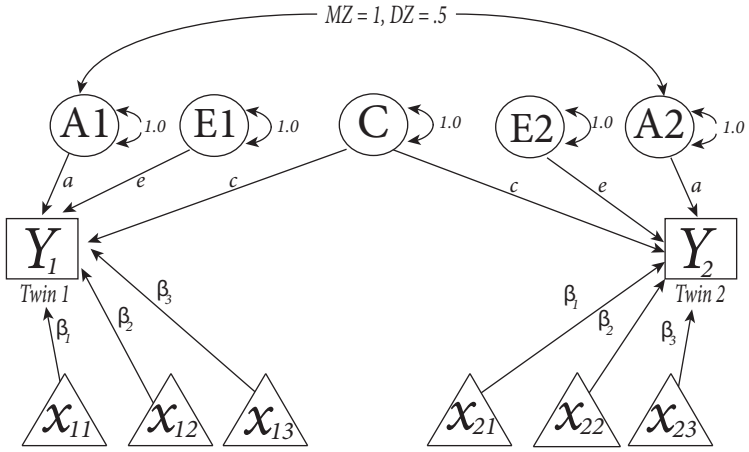


Figure 6.1: Structural equation model (SEM) for a basic univariate twin analysis (ACE model) extended with three covariates (denoted as x_{11}, x_{12} and x_{13} for the first and as x_{21}, x_{22} and x_{23} for the second twin of a family). Y denotes the phenotypic values of the first ($Y1$) and second ($Y2$) twin and A refers to additive genetic influences for the first ($A1$) and second ($A2$) twin, which are correlated 0.5 in dizygotic twins and 1 in monozygotic twins. $E1$ and $E2$ denote unique-environmental influences of the first and second twin respectively and are assumed to be uncorrelated. C, common-environmental influences, are the same for all family members. Double-headed arrows denote (co-)variances. The path coefficients, β_1, β_2, β_3, a, c and e represent regression coefficients that express the estimated effect of the respective influences.

this is a rather strong assumption. A weaker assumption is that covariates are *missing at random* (MAR), that is, the probability that a covariate value is missing is unrelated to its unobserved value, after controlling for other variables in the analysis. For example, female twins might be more likely to not give information on their income, but this might be unrelated to the amount of income once one controls for gender. Lastly, a covariate value can be *missing not at random* (MNAR), that is, the probability that it is missing is related to its unobserved value. In this case, for example, twins with a lower income might be more or less likely to reveal this information.

When there is (partly) missing data, complete-cases analysis (also referred to as listwise deletion) can be used, meaning that only twin pairs with complete data (e.g. data of twin pairs with known values for all covariates) enter the analysis. This certainly leads to reduced statistical power, but might also introduce bias or affect the representativeness of the results (Allison, 2001). Using OpenMx (Boker et al., 2011), a SEM program often used to fit twin models, twin researchers usually apply a strategy that

minimizes the loss of information by excluding phenotypic data of *individual* twins with at least one missing covariate value. So, when an individual twin has a missing covariate value, the phenotypic (and covariate) data of his or her co-twin can still be used for statistical inference (provided that the co-twin does not have any missing data). This results in twin-wise rather than twin pair-wise deletion of incomplete cases.

In this paper, we present a *full information* approach to handle missing covariate data. The new approach involves including covariates in the expected covariance matrix. While in the usual approach, the phenotypic and covariate values of a twin with (at least) one missing covariate value are completely ignored, the full information approach models *all* data that are observed - including observed phenotypic data as well as observed covariate data. The new approach will be described in more detail in the following.

6.1.2 Full information approach

In the traditional univariate ACE model, the phenotypic variance is decomposed into variance due to additive genetic influences, σ_A^2, variance explained by common-environmental influences, σ_C^2, and variance due to unique-environmental influences, σ_E^2. Conditioning on the covariate data, phenotypic data are assumed to be multivariate normally distributed:

$$\begin{matrix} y_{i1} \\ y_{i2} \end{matrix} \Bigg| \begin{matrix} \boldsymbol{x}_{i1} \\ \boldsymbol{x}_{i2} \end{matrix} \sim MVN\left(\begin{pmatrix} \mu_y \\ \mu_y \end{pmatrix} + \begin{pmatrix} \boldsymbol{x}_{i1}^T \\ \boldsymbol{x}_{i2}^T \end{pmatrix} \boldsymbol{\beta}, \Sigma_{ACE} \right) \tag{6.1}$$

where

$$\Sigma_{ACE} = \begin{bmatrix} \sigma_A^2 + \sigma_C^2 + \sigma_E^2 & \rho\sigma_A^2 + \sigma_C^2 \\ \rho\sigma_A^2 + \sigma_C^2 & \sigma_A^2 + \sigma_C^2 + \sigma_E^2 \end{bmatrix} \tag{6.2}$$

and μ_y refers to the phenotypic mean. y_{i1} denotes the phenotypic value of the first twin of family i and y_{i2} represents the phenotypic value of the second twin. $\boldsymbol{x}_{i1}$ and $\boldsymbol{x}_{i2}$ are covariate data vectors that include the values of the covariates of the first and second twin respectively and the vector $\boldsymbol{\beta}$ consists of the regression coefficients of the covariates. Σ_{ACE} refers to total phenotypic covariance and ρ is the correlation between the twins' additive polygenic factors, which is unity in monozygotic (MZ) twins and $\frac{1}{2}$ in dizygotic (DZ) twins. The phenotypic variance decomposition takes place after the effects of the covariates have been partialled out, but other than that the covariate data are not part of the covariance model.

Here, we propose to model the covariance between *all* observed variables - consisting of phenotypic data but also covariate data. In twin data, it is reasonable to assume not only covariance among the covariates of one twin (e.g. correlations between covariates), but also covariance among the values of the covariates of one twin and the covariates of the co-twin. To incorporate this dependence structure into the biometric model we

decompose the covariance structure of the values on the covariates of both twins into covariance shared by twins from the same pair and non-shared twin covariance. Covariate data were then assumed to be multivariate normally distributed:

$$\begin{matrix} \boldsymbol{x}_{i1} \\ \boldsymbol{x}_{i2} \end{matrix} \sim MVN\left(\begin{pmatrix} \boldsymbol{\mu}_x \\ \boldsymbol{\mu}_x \end{pmatrix}, \Sigma_{\text{cov}} \right) \tag{6.3}$$

where $\boldsymbol{\mu}_x$ is a vector that contains of the means of the covariates and

$$\Sigma_{\text{cov}} = \begin{bmatrix} \Sigma_{\text{twin1}} & \Sigma_{\text{b}} \\ \Sigma_{\text{b}} & \Sigma_{\text{twin2}} \end{bmatrix} = \begin{bmatrix} \Sigma_{\text{w}} + \Sigma_{\text{b}} & \Sigma_{\text{b}} \\ \Sigma_{\text{b}} & \Sigma_{\text{w}} + \Sigma_{\text{b}} \end{bmatrix} \tag{6.4}$$

Σ_{cov} denotes total covariate covariance. Σ_{b} denotes between twin pair variance and Σ_{w} within twin pair variance. Thus, the covariance matrix for covariates of an individual twin is decomposed into covariance shared with the co-twin, Σ_{b}, and covariance not shared with the co-twin, Σ_{w}.

By including covariate data in the expected covariance matrix, the joint distribution of phenotypes and covariates, $(y_{i1}, y_{i2}, \boldsymbol{x}_{i1}, \boldsymbol{x}_{i2})^T$, is multivariate normal with the following covariance structure:

$$\Sigma = \begin{bmatrix} ACE+\boldsymbol{\beta}^T(\Sigma_{\text{w}}+\Sigma_{\text{b}})\boldsymbol{\beta} & AC+\boldsymbol{\beta}^T\Sigma_{\text{b}}\boldsymbol{\beta} & \boldsymbol{\beta}^T(\Sigma_{\text{w}}+\Sigma_{\text{b}}) & \boldsymbol{\beta}^T\Sigma_{\text{b}} \\ AC+\boldsymbol{\beta}^T\Sigma_{\text{b}}\boldsymbol{\beta} & ACE+\boldsymbol{\beta}^T(\Sigma_{\text{w}}+\Sigma_{\text{b}})\boldsymbol{\beta} & \boldsymbol{\beta}^T\Sigma_{\text{b}} & \boldsymbol{\beta}^T(\Sigma_{\text{w}}+\Sigma_{\text{b}}) \\ (\Sigma_{\text{w}}+\Sigma_{\text{b}})\boldsymbol{\beta} & \Sigma_{\text{b}}\boldsymbol{\beta} & \Sigma_{\text{w}}+\Sigma_{\text{b}} & \Sigma_{\text{b}} \\ \Sigma_{\text{b}}\boldsymbol{\beta} & (\Sigma_{\text{w}}+\Sigma_{\text{b}})\boldsymbol{\beta} & \Sigma_{\text{b}} & \Sigma_{\text{w}}+\Sigma_{\text{b}} \end{bmatrix} \tag{6.5}$$

where phenotypic variances are represented on the first two elements of the diagonal and covariate data variances on the remaining elements of the diagonal. Cross-phenotypic and cross-covariates covariances and within twin covariate covariances are contained on the off-diagonal elements. ACE denotes $\sigma_A^2+\sigma_C^2+\sigma_E^2$, AC refers to $\rho\sigma_A^2+\sigma_C^2$ and $\boldsymbol{\beta}$ is a vector that includes the regression coefficients of the covariates. A graphical representation of this model, including ACE decomposition and the model for covariate data, can be found in Figure 6.2 (SEM notation). In the example, answers to three different covariates are modelled for the first (x_{11}, x_{12} and x_{13}) and second (x_{21}, x_{22} and x_{23}) twin of one family. To model between twin pair variance (i.e., Σ_{b}, covariance between the values of the first and second twin on the same covariate but also cross-covariance), we model latent variables for every covariate, ψ_1, ψ_2 and ψ_3. To model within twin pair variance (i.e., Σ_{w}), we use different latent variables for the first (γ_{11}, γ_{12} and γ_{13}) and second (γ_{21}, γ_{22} and γ_{23}) twin.

6.1.3 Benefits of the new approach

In the usual approach, the phenotypic score as well as the covariate data of a twin with (at least) one missing value is not used for statistical inference. Adopting the full information approach, all observed data (including

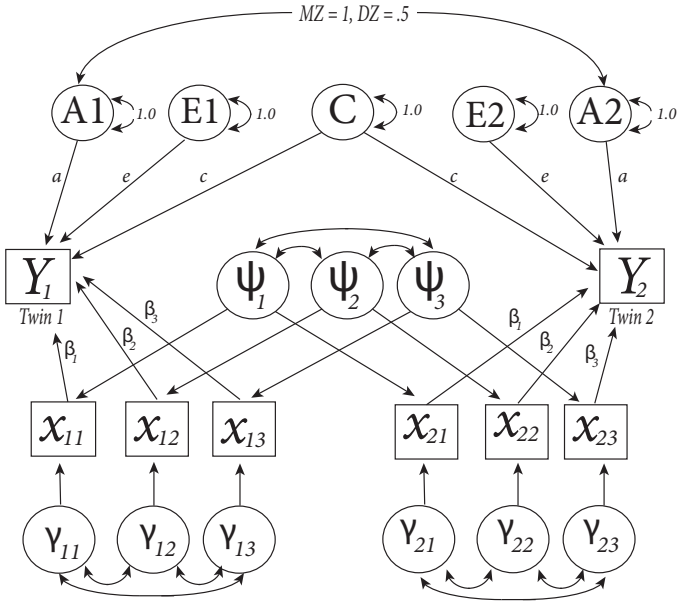


Figure 6.2: Structural equation model (SEM) of the full information approach. Answers to three different covariates are displayed for the first (x_{11}, x_{12}, x_{13}) and second (x_{21}, x_{22}, x_{23}) twin of one family. ψ is a latent variable that is estimated for every covariate (ψ_1, ψ_2, ψ_3) and models covariance within families. The different latent variables for the first (γ_{11}, γ_{12} and γ_{13}) and second (γ_{21}, γ_{22} and γ_{23}) twin model within twin covariance.

phenotypic scores) can be used. This fact alone makes the full information approach more powerful.

Furthermore, the usual approach may result in biased estimates when covariate data were missing not completely at random. Imagine for example that the probability that a covariate value is missing depends on the phenotypic value of an individual twin. For example, twins with a high score on a depression assessment are less likely to give information on their income. Using the usual approach, the phenotypic values of these twins do not enter the analysis - therefore, phenotypic variance is underestimated, which might lead to biased estimates of variance components and under- or overestimation of heritability.

The third advantage is that, by modelling the relationships between data that are unobserved and data that are observed, information on model parameters can be statistically borrowed by information on data that are observed and correlated with unobserved data - a principle that is often referred to as *borrowing strength*. Borrowing strength means that

the inference of a parameter of interest or unobserved data point can be improved by borrowing from information on other related data also included in the model. For example, imagine that we have measured one environmental covariate, separately for the first and second twin of every family. In one of the families, the covariate value of the second twin is known whereas the value for the first twin is missing. Based on covariance between the observed values, our model can then borrow information from the covariate value of the co-twin but also from the phenotypic value of the twin with missing value and the phenotypic value of his or her co-twin, which leads to lower standard errors. Note that this is especially true when the data are highly correlated, for example with high twin correlations for a covariate, high correlations among the covariates, or when there is a strong relationship between phenotypic and covariate data.

In a simulation study, it is shown that the full information approach is more powerful than the usual approach, independent of missingness scenario, number of covariates and amount of missingness. To illustrate the new approach, it is applied to test scores on a Dutch national school achievement test. Syntax to apply the full information approach using the R package OpenMx (Boker et al., 2011) can be found in Appendix F.

6.2 Simulation study

In order to show that the full information approach retrieves parameters reliably and is more powerful than the usual approach, a simulation study was conducted with a fixed number of twin pairs and different number of covariates (two, three, four and five) and percent of missing observations (2%, 6% and 10%). In each combination of these conditions, 1000 datasets were generated consisting of 280 MZ (28% of all pairs) and 720 DZ pairs (72% of all pairs). This ratio reflects the usual ratio of MZ and DZ twin pairs in European twin registers. The amount of missing observations for the different conditions (2%, 6% and 10%) was based on the total number of covariate answers (e.g., in case of five covariates: five $\times$ 2000 individual twins = 10,000). In all conditions, additive genetic variance was assumed 0.5, common-environmental variance was set to 0.3 and unique-environmental variance was assumed 0.2. The data were simulated with a phenotypic population mean of zero for all twins ($\mu_y = 0$). In every condition, regression coefficients, $\boldsymbol{\beta}$, were chosen such that covariates explained 39% of total phenotypic variance, leading to a total variance of 1.64. A multivariate normal distribution was used to simulate the covariate data. The expectation of the multivariate distribution was set to zero ($\boldsymbol{\mu}_x = 0$) and the covariance matrix was based on Σ_{w} and Σ_{b}. The same values were used for the diagonals and off-diagonals of Σ_{w} and Σ_{b} in every condition. For example for five covariates, Σ_{w} was equal to $\begin{pmatrix} 1 & .1 & .1 & .1 & .1 \\ .1 & 1 & .1 & .1 & .1 \\ .1 & .1 & 1 & .1 & .1 \\ .1 & .1 & .1 & 1 & .1 \\ .1 & .1 & .1 & .1 & 1 \end{pmatrix}$ and Σ_{b} was equal to

$\begin{pmatrix} 1 & .5 & .5 & .5 & .5 \\ .5 & 1 & .5 & .5 & .5 \\ .5 & .5 & 1 & .5 & .5 \\ .5 & .5 & .5 & 1 & .5 \\ .5 & .5 & .5 & .5 & 1 \end{pmatrix}$. This led to following covariance matrix:

$$\Sigma_{\text{tot}} = \begin{bmatrix} 2 & .6 & .6 & .6 & .6 & 1 & .5 & .5 & .5 & .5 \\ .6 & 2 & .6 & .6 & .6 & .5 & 1 & .5 & .5 & .5 \\ .6 & .6 & 2 & .6 & .6 & .5 & .5 & 1 & .5 & .5 \\ .6 & .6 & .6 & 2 & .6 & .5 & .5 & .5 & 1 & .5 \\ .6 & .6 & .6 & .6 & 2 & .5 & .5 & .5 & .5 & 1 \\ & & & & & & & & & \\ 1 & .5 & .5 & .5 & .5 & 2 & .6 & .6 & .6 & .6 \\ .5 & 1 & .5 & .5 & .5 & .6 & 2 & .6 & .6 & .6 \\ .5 & .5 & 1 & .5 & .5 & .6 & .6 & 2 & .6 & .6 \\ .5 & .5 & .5 & 1 & .5 & .6 & .6 & .6 & 2 & .6 \\ .5 & .5 & .5 & .5 & 1 & .6 & .6 & .6 & .6 & 2 \end{bmatrix} \tag{6.6}$$

The pattern of the missingness was generated under three different scenarios. In the first setting, covariate data were simulated to be missing completely at random (MCAR). That is, every covariate value had the same probability of being missing, independent of unobserved or observed data. In the second scenario, it was assumed that the data were missing at random (MAR). Here, the probability that a covariate value was missing was dependent on the (observed) phenotypic value of an individual twin. We modelled the probability of missingness for every covariate x_{ijk} as a logistic function of the respective phenotypic value of every individual twin j from family i:

$$p(x_{ijk} \text{ is missing}) = \frac{1}{1 + \exp(2 + 1.7\ y_{ij})} \tag{6.7}$$

The resulting probabilities were then used in the R in-built function *sample()* to control the overall proportion of missing values. By using Equation 6.7 to model missingness, the probability that a covariate value was missing was higher with decreasing phenotypic value. In the last scenario, covariate data were assumed to be missing not at random (MNAR). Here, the probability that a covariate value was missing was dependent on its observed (simulated) value. As the range of phenotypic values was similar to the range of covariates values, the same logistic function was used as in the MAR scenario, but the probability was dependent on the observed value of the covariate (i.e. $p(x_{ijk} \text{ is missing}) = \frac{1}{1+\exp(2+1.7\ x_{ijk})}$). As in the MAR setting, the resulting probabilities were used in the R in-built function *sample()* to control the overall proportion of missing values.

In every scenario, the remaining data were analysed using 1) the usual approach and 2) the full information approach. For the simulations, the software package R (development core team, 2013) was used. The models were fit using the R package OpenMx (Boker et al., 2011). The point estimates of the variance components and regression coefficients were determined as

were their standard errors. Furthermore, narrow-sense heritability, h^2, was determined, which we defined here as $\frac{\sigma_A^2}{\sigma_P^2}$, where $\sigma_P^2 = \sigma_A^2 + \sigma_C^2 + \sigma_E^2$.

6.2.1 Results

As estimates of regression coefficients were close to their true values and very similar for both approaches under all conditions, results are not displayed here but can be obtained from the first author.

MCAR

Standard errors for σ_A^2, σ_C^2 and σ_E^2 can be found in Figure 6.3. The standard errors were generally lower when the full information approach was used compared to the usual approach. Furthermore, while standard errors were very similar under different amounts of missingness and number of covariates when the full information approach was applied, they increased with increasing number of covariates when the usual approach was used. This effect was the largest for the 10% missingness condition. Compared to the other variance components, standard errors of σ_E^2 were, in general, small and only increased slightly with increasing number of covariates when the usual approach was used. For both approaches, estimates of σ_A^2, σ_C^2, σ_E^2 and h^2 were all very close to their true values and are therefore not displayed here.

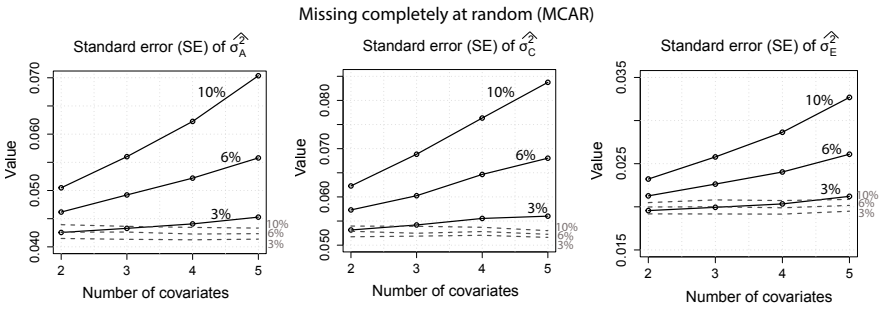


Figure 6.3: MCAR: Standard errors for σ_A^2, σ_C^2 and σ_E^2 of both approaches when 2%, 6% and 10% of the covariate data were missing. *Dotted lines*: Full information approach.

MAR

The standard errors for σ_A^2, σ_C^2 and σ_E^2 can be found in Figure 6.4.

The same pattern as for the MCAR condition can be observed: Standard errors of the full information approach were generally lower than the

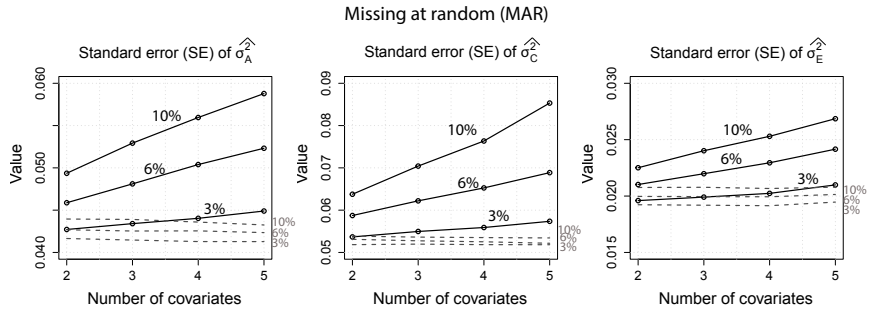


Figure 6.4: MAR: Standard errors for σ^2_A, σ^2_C and σ^2_E of both approaches when 2%, 6% and 10% of the covariate data were missing. *Dotted lines*: Full information approach.

standard errors obtained with the usual approach. Furthermore, while the standard errors of the usual approach increased with increasing number of covariates, standard errors of the full information approach were very similar across all missingness conditions. As in the first scenario, standard errors for σ^2_E were generally low and increased only slightly with increasing number of covariates when the usual approach was used.

Figure 6.5 displays the estimates of σ^2_A, σ^2_C and h^2 for both approaches.

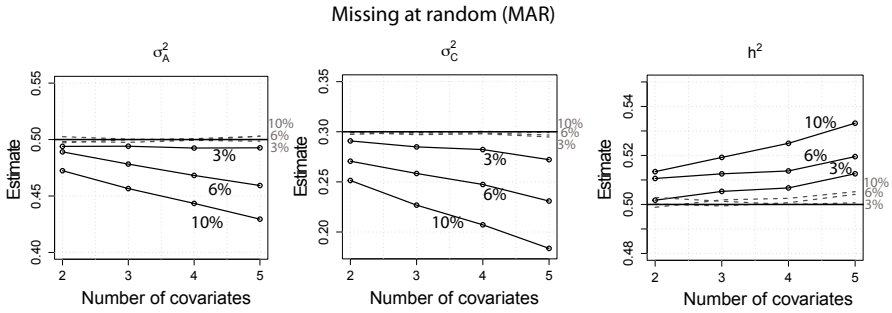


Figure 6.5: MAR: Estimates of σ^2_A, σ^2_C and h^2 of both approaches when 2%, 6% and 10% of the covariate data were missing. *Dotted lines*: Full information approach. The horizontal lines denote the true values of σ^2_A, σ^2_C and h^2.

It can be seen that the variance components estimates of the full information approach were all very close to their true values, independent of amount of missingness and missingness scenario. Using the usual approach, estimates were close to their true values in the 3% missingness condition, but

variance components were underestimated when the amount of missingness increased. This bias further increased with increasing number of covariates. Furthermore, the bias was generally more severe for estimates of σ^2_C than for estimates of σ^2_A. This is also reflected in the heritability estimates. h^2 was overestimated and this bias systematically increased with increasing number of covariates. However, we can also see that this bias was negligible for the 3% and 6% missingness condition. Estimates of σ^2_E were unbiased for the full information approach as well as the usual approach and are therefore not displayed.

MNAR

Figure 6.6 shows the standard errors for both approaches for all variance components. Again, we can observe the same pattern: the full information approach had lower standard errors which were very similar under different conditions while standard errors increased with increasing number of covariates when the usual approach was used. Similar to the MCAR and MAR scenario, the standard errors of σ^2_E only increased slightly with increasing number of covariates and were generally low also when the usual approach was used. Estimates of σ^2_A, σ^2_C and σ^2_E were all very close to their true values for both approaches and are therefore not displayed here. Using the full information approach, there was a negligible bias in the 10% missingness condition with estimates of σ^2_E closer to 0.21 instead of the true value 0.20.

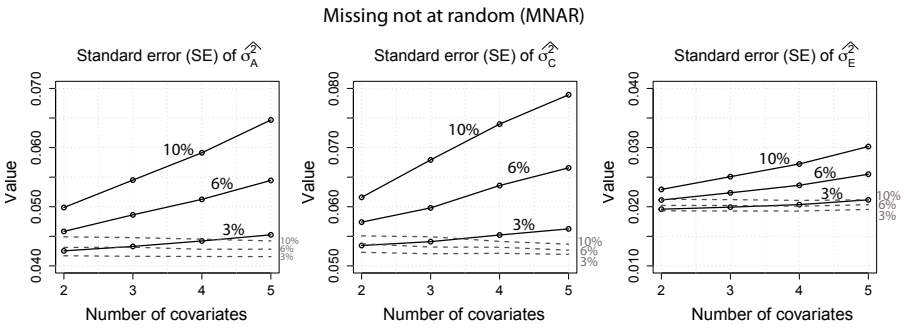


Figure 6.6: MNAR: Standard errors for σ^2_A, σ^2_C and σ^2_E of both approaches when 2%, 6% and 10% of the covariate data were missing. *Dotted lines*: Full information approach.

6.3 Application

To illustrate the full information approach, we applied it to test scores on a Dutch national school achievement test in the final grade of primary

school. The effect of school-aggregated measures (e.g. school denomination, pedagogical philosophy, school size) and the effect of the sex of a twin on these test scores was tested. These covariates were a mix of continuous and categorical variables. Therefore, due to its flexibility, it was chosen to use a Bayesian parametrization of the model for this application. In Bayesian analysis, statistical inference is based on the joint *posterior density* of the model parameters, which is proportional to the product of a prior probability distribution and the likelihood function (for a general introduction to Bayesian statistics see e.g. Bolstad, 2007 and for Bayesian analysis of twin models see e.g. Eaves & Erkanli, 2003 or van den Berg et al., 2006). A prior probability distribution represents information about an uncertain parameter before any data have been observed. In this application, uninformative prior distributions were chosen. That is, they expressed only vague information about the parameters of our model and therefore, posterior point estimates presented here are close to maximum likelihood estimates as would be obtained by using for example OpenMx (Boker et al., 2011).

6.3.1 Sample

The sample of this study originated in the Netherlands Twin Register (NTR, Boomsma et al., 2002), which includes approximately 40 per cent of all multiple births in the Netherlands. If parents give their consent, teachers of the children are approached with a survey when the twins are 7, 9 and 12 years old. In 2000, the NTR started collecting the results of a national test of educational achievement (*Eindtoets Basisonderwijs*) from the parents of all 12-year-old twins. The *Eindtoets Basisonderwijs* test is yearly administered in the final grade of primary school.

The present study analyzed data of 12-year-old twins from birth cohorts 1997-2000 to determine the importance of measured covariates for individual differences in *Eindtoets Basisonderwijs* test scores. The sample included data of children from 990 twin pairs, consisting of 340 MZ twin pairs and 650 DZ twin pairs. Of the MZ twin pairs, 175 pairs were male and 165 female. 159 of the DZ twin pairs were male, 167 female and 324 twin pairs were of opposite sex. For 120 individual twins, the score on the *Eindtoets Basisonderwijs* test was unknown. The reason that the score was missing was either that the child had not reached final grade yet (N twins = 66), the child was attending special education (N twins = 33), a different test was used at the school the twin was attending (N twins = 6), the child did not attend the test (N twins = 2) or the reason was unknown (N twins = 23).

6.3.2 Measures

The items on the teacher report form that were used in this paper are: The name, postal code, denomination and pedagogical philosophy of the school.

The reported names and postal codes were used to link the twin data from the NTR with school-aggregated environmental measures obtained from external sources such as official authorities. This was only done with data for twins for which parents had given written permission to link databases. Reported denomination and pedagogical philosophy of a school on the teacher report were used to complement the retrieved data. The *Eindtoets Basisonderwijs* test consists of 290 multiple choice items in four different subjects (language, world studies [optional], arithmetic and study skills). We used the total score on the *Eindtoets Basisonderwijs* test, a standardized measure that ranges from 500 to 550. As administration of the questions concerning world studies is optional, they were not included in the total score. Information on the denomination of a specific school was retrieved from the Dutch ministry of education (Dienst Uitvoering Onderwijs, DUO). This information was supplemented with information available from answers of the teachers on the teacher report form. The variable was measured in seven categories: *Collaboration of Protestant-Christian and Roman Catholic*, *Protestant-Christian*, *Reformed*, *Reformed liberated*, *Roman Catholic*, *Special* and *State* schools. Information on pedagogical philosophy was retrieved online from a database that provides basic information about Dutch primary education schools (`http://www.scholenopdekaart.nl`). Again, this information was supplemented with information available from answers on the teacher report form. The variable was categorized into five different categories: *Regular education*, *Dalton plan education*, *Jenaplan edcation*, *Montessori education*, *Specialised regular education* and *Specialised education.* Data on school size, measured in 2011, were retrieved from the Dutch ministry of education (Dienst Uitvoering Onderwijs, DUO). The data were linked to the postal codes of the schools, retrieved from the teacher report form. An overview of all covariates that were used in this paper can be found in Table 6.1.

6.3.3 Analysis

The analysis was done in the Markov chain Monte Carlo (MCMC) sampling program JAGS (Plummer, 2003). R (R development core team, 2013) was used for further data handling and as an interface from R to JAGS, the rjags package (Plummer, 2013) was used. The syntax that was used can be found in Appendix G.

Prior to the analysis, $c - 1$ dummy variables were created for the categorical variables *Sex*, *School denomination* and *Pedagogical philosophy* with c being the number of categories and the largest category serving as the reference group. For these covariates, a multivariate normal distribution (see Equation 6.3) was used to model liabilities. The built-in function *step()* of JAGS was then used to create a Boolean variable $V = \text{step}(x_{ijk} - t)$ that equals one if $(x_{ijk} - t) \geq 0$ and equals zero if $(x_{ijk} - t) < 0$ where t is a threshold that was fixed to zero for identification purposes. The phenotypic variable (*Eindtoets Basisonderwijs* test scores) as well as the

Table 6.1: Overview of covariates for educational achievement (*Eindtoets Basisonderwijs* test scores) that were used in the application. N = total number of individual twins.

	N
Sex	
Boy	992 (50.10%)
Girl	988 (49.90%)
School size	1447 (73.08%)
Missing	533 (26.92%)
Pedagogical philosophy	
Regular education	1467 (74.09%)
Dalton	41 (2.10%)
Jenaplan	12 (0.61%)
Montessori	21 (1.06%)
Specialised regular education	16 (0.81%)
Specialised education	6 (0.30%)
Missing	417 (21.10%)
Denomination of school	
Protestant-Christian (PC)	332 (16.77%)
Reformed	22 (1.11%)
Reformed liberated	8 (0.40%)
Roman-Catholic (RC)	606 (30.61%)
Collaboration of PC & RC	10 (0.51%)
Special	59 (2.98%)
State	461 (23.28%)
Missing	482 (24.34%)

numeric covariate *School size* were standardized to have an expected value of zero and a variance of one. The missing *Eindtoets Basisonderwijs* test scores (N twins = 120) were assumed missing at random.

The mean and standard deviation of the posterior distribution were calculated for each parameter as was the the 95% highest posterior density (HPD, see e.g. Box & Tiao, 1973) interval for variance components and the 99.6% HPD interval for covariates. The HPD can be interpreted as the Bayesian analog of a confidence interval (CI). When the HPD does not contain zero, the influence of a parameter can be regarded as significant.

6.3.4 Results

The posterior means for the variance components σ^2_A, σ^2_C and σ^2_E as well as the estimated heritability (h^2, defined as $\frac{\sigma^2_A}{\sigma^2_P}$, where $\sigma^2_P = \sigma^2_A + \sigma^2_C + \sigma^2_E$) for the fitted model are displayed in Table 6.2. The results suggest that the largest part of the variance could be explained by genetic influences. A substantial part of the phenotypic variance could be explained by unique-environmental influences and a small part by common-environmental influences.

Table 6.2: Educational achievement (*Eindtoets Basisonderwijs* test scores): Posterior means (SD) of the variance components. HPD refers to the 95% highest posterior density interval.

	Posterior mean (SD)	HPD
σ^2_A	0.66 (0.05)	[0.57;0.76]
σ^2_C	0.16 (0.04)	[0.09;0.24]
σ^2_E	0.20 (0.02)	[0.17;0.23]
h^2	0.64 (0.04)	[0.56;0.72]

The posterior means and HPD intervals of the regression coefficients are displayed in Table 6.3. There was no covariate that had a significant effect on individual differences in *Eindtoets Basisonderwijs* test scores.

Table 6.3: Educational achievement (*Eindtoets Basisonderwijs* test scores): Regression coefficients for the estimated model. For the covariates Sex, School denomination and Pedagogical philosophy, dummy variables were used. The reference categories are: *Female*, *Roman-Catholic* and *Regular education*. HPD refers to the *99.6% highest posterior density interval*.

	Sex	School denomination					School size		Pedagogical philosophy			
		Collaboration PC & RC	Protestant-Christian	Reformed	Reformed liberal	Special	State	Dalton	Jenaplan	Montessori	Specialised regular	Specialised
Posterior mean	-0.11 (0.04)	-0.03 (0.14)	0.02 (0.06)	-0.22 (0.13)	-0.11 (0.13)	0.21 (0.11)	0.07 (0.06)	0.04 (0.03)	-0.09 (0.14)	-0.20 (0.16)	0.04 (0.15)	0.27 (0.16)	0.02 (0.16)
HPD	[-0.23;0.00]	[-0.44;0.36]	[-0.17;0.19]	[-0.58;0.12]	[-0.49;0.27]	[-0.12;0.50]	[-0.10;0.23]	[-0.07;0.13]	[-0.49;0.30]	[-0.66;0.24]	[-0.38;0.47]	[-0.18;0.71]	[-0.46;0.48]

6.4 Discussion

The often-used ACE model can be extended to include covariates. However, problems in the data collection or the linking of two different datasets often leads to a high amount of missing data. The usual approach to handle missing data results in reduced power, because phenotypic and covariate data of a twin with at least one missing value cannot be used for statistical inference.

In this paper, we present a full information approach that entails modelling covariance of all data, both phenotypic data as well as covariate data. The covariance structure for the covariates in a twin pair was decomposed into between and within family covariance. This makes it possible, to use *all* available data, which makes the new approach more powerful than the usual approach.

In a simulation study, the performance of the full information approach was compared to the usual approach under different conditions. Independent of missingness scenario (MCAR, MAR and MNAR), amount of missingness (2%, 6% or 10% of the total number of covariate values) and number of covariates (two, three, four and five), standard errors for all variance components were lower for the new approach than for the usual approach. Furthermore, standard errors of the full information approach were constant while the power of the usual approach rapidly decreased with increasing number of covariates. Note that this pattern has to do with the fact that, in the usual approach, twins with at least one missing value do not enter the analysis. Therefore, although the percentage of missingness remained constant, the probability that a twin ends up with at least one missing value was higher with increasing number of covariates. This reduced the number of twins that entered the analysis, resulting in decreasing power with increasing number of covariates.

Note that we used the same covariance structure for MZ and DZ twin pairs to model covariate covariance between families (i.e., Σ_b) and covariance within families (i.e., Σ_w). Therefore, the exact results of the simulation study are restricted to this situation. Also, different covariance structures for MZ and DZ twin pairs can be specified. Simulating and analysing covariate data with different covariance structures among MZ and DZ twins might lead to slightly different effects on power, as the probability that covariate data are missing in both twins might be different for MZ and DZ twins respectively. Generally though the effects on power will be the same: the full information approach is more powerful than the usual approach. Depending on the particular application, it might be more appropriate to use different covariance structures for MZ and DZ twins when differences in the covariance structure are expected a priori or when the data seem to suggest it (i.e. twin correlations on the covariate data might be higher in MZ twin pairs than in DZ twin pairs).

To illustrate the new approach, the effects of specific covariates on

the test scores of 990 12-year-old twin pairs on a national Dutch educational achievement test (*Eindtoets Basisonderwijs*) in primary school were investigated. We used school-aggregated measures (school denomination, pedagogical philosophy, school size) and the sex of a twin as covariates. Similar to earlier findings on *Eindtoets Basisonderwijs* scores of a Dutch sample (Bartels, Rietveld, Van Baal & Boomsma, 2002), the results suggest that differences in test scores are mainly due to genetic influences. There was no covariate that had a significant effect on individual differences in test scores. As, however, a substantial part of the phenotypic variance could be explained by environmental variance, this suggests that there are environmental influences that were not investigated in this study that cause individual differences in *Eindtoets Basisonderwijs* test scores. Variables that might be important but were not examined in this paper might for example be resources the twins have (e.g. libraries in the neighbourhood or books at home) or the composition of their class (e.g. the IQ of their classmates). Another explanation for the non-significant result could be that environmental influences on students test scores are highly multifactorial, meaning that there are a lot of influences that each have small effects and contribute to variance in test scores when they are combined.

In the application, 18% of the total number of covariate answers was missing (i.e., 1432/(four $\times$ 1980 individual twins)). This shows that even the most extreme missingness condition (i.e., 10%) of the simulation study is realistic and to be expected in real data applications. When the usual approach would be applied to the same data, this would result in the loss of the phenotypic as well as covariate data of in total 496 individual twins, reducing the twin sample from 1980 individual twins to 1484 individual twins (75% of the original sample size). This highlights the added value of the full information approach over the usual approach in practical situations.

In conclusion, as it could be shown that the full information approach is more powerful than the usual approach and can be easily applied in OpenMx (Boker et al., 2011) by using the syntax we provide here, we advise researchers to use the new approach whenever a) more than 3% of the total covariate data are missing and b) when more than two covariates are used in the analysis.

CHAPTER 7

Summary and Discussion

7.1 Summary

This dissertation discusses a number of psychometric issues that have to be taken into account in the analysis of genetically-informative data. These include heterogeneous measurement error, arbitrary scaling and harmonization of phenotypes. It is shown how ignoring these issues can result in bias such as the spurious finding of genotype-environment interaction and how item response theory (IRT) models can help to solve these problems.

Chapter 2 is concerned with the modelling of genotype-environment interaction in the case that the environment features as a latent (i.e., unmeasured) variable. While, often, sum scores are used in the biometric model, it is proposed to model raw item data instead by incorporating an explicit measurement model into the analysis. By means of a simulation study, it is shown that the use of sum scores can lead to the spurious finding of genotype-environment interaction due to properties of the given measurement scale while the proposed method is unbiased. A second simulation study illustrates the power of the new approach.

In Chapter 3, this new approach is applied to the item test scores of 2110 Dutch twin pairs on the mathematics subscale of a national educational achievement test (*Eindtoets Basisonderwijs*). The results of international comparisons are often taken to indicate that mathematical education in Dutch schools is more appropriate for the weakest than for the mathematically gifted students. In this chapter, this hypothesis was tested by investigating whether the importance of environmental influences differs depending on genotype (innate talent for mathematics performance). As hypothesized, results suggest that environmental influences are relatively more important in explaining individual differences in the high-ability students than in the low-ability students.

In Chapter 4, the Wilson-Patterson conservatism scale is psychometrically evaluated using homogeneity and IRT models. The results suggest that the scale actually measures two different aspects in people: on the one hand people vary in their agreement with either conservative or liberal catch-phrases and on the other hand people vary in their use of the "?" response category of the scale. Based on these results, a new scale is devised and used in a biometric analysis that includes genotype-environment interaction, extending the method introduced in Chapter 2 to ordinal data. Biometric results showed significant genetic and shared environmental influences, and significant genotype-environment interaction effects, suggesting that individuals with a genetic predisposition for conservatism show more non-shared variance but less shared variance than individuals with a genetic predisposition for liberalism.

In Chapter 5, a method is introduced that incorporates an IRT model into the modelling of variance decomposition moderation (ACE×M). For the biometric modelling of ACE×M, an alternative parametrization is used, which is uniquely identified and therefore to be preferred over the parametrization introduced by Purcell (2002). Biometric and IRT model are estimated simultaneously, which prevents the finding of spurious interaction effects due to scale issues. Simulation studies suggest that a large sample size is required to detect A×M and C×M interaction effects. The method is illustrated by applying it to the data of 2110 12-year-old Dutch twin pairs to test moderating effects of a family's socio-economic status on individual differences in mathematical ability.

In Chapter 6, a new approach to handle missing covariate data in twin data is proposed. The usual approach to handle missing covariate data in twin research results in reduced statistical power, because only phenotypic and covariate data of individual twins with complete data can be used. By including covariates in the expected covariance matrix, the new approach makes it possible to use all observed data. A simulation study shows that, independent of missingness scenario (MCAR/MAR/MNAR), number of covariates (two, three and four) or amount of missingness (2%, 6% and 10% of the total number of covariate answers), the new approach is more powerful than the usual approach. To illustrate the new method, it is applied to test scores on the *Eindtoets Basisonderwijs* test of 990 twin pairs to investigate the effects of school-aggregated measures and the sex of a twin on these test scores.

7.2 Discussion

The roots of behaviour genetics go back a long way, with Francis Galton in the nineteenth-century being the first to study family resemblance based on family relatedness. Studying human ability in relatives, Galton noted that similar eminence was more likely to be found in close relatives such as

parents or offspring than in more distant relatives (Galton, 1869). More than a century after Galton's work, the field of behaviour genetics has experienced an immense growth. From the simple method of calculating twin correlations, the field has gone to structural equation models and Markov chain Monte Carlo algorithms. There has been progress also in the estimation of genetic influences. The revolution in molecular genetics has provided more effective tools in describing the genome, making it possible to for example estimate the effects of specific genes. Consequently, the focus of behaviour genetics studies has shifted from twin and adoption studies to genome-wide association studies (GWAS, see e.g. Neale, Ferreira, Medland & Posthuma, 2008) and genome-wide complex trait analysis (GCTA, Yang, Lee, Goddard & Visscher, 2011). One part of behaviour genetics, however, has not experienced as much progress: The measurement of the trait of interest, the *phenotype*.

Measuring directly observable attributes such as height or weight is relatively simple. However, the measurement of *unobserved* psychological traits such as personality or cognitive abilities is more complicated. Results can be biased due to psychometric issues such as heterogeneous measurement error, arbitrary scaling or test versions that are not comparable across samples. Only a small body of work so far has concentrated on the development of methodology to solve these issues (e.g. van den Berg et al., 2007; van der Sluis et al. 2010; van den Berg & Service, 2012; Molenaar et al., 2012; van der Sluis et al., 2013; van den Berg et al., 2014; Molenaar & Dolan, 2014). More recently, the emerging transdiscipline of *phenomics*, dedicated to the systematic study of phenotypes on a genome-wide scale, has led to increased attention to the measurement of the phenotype. Still, however, our ability to solve psychometric issues in the analysis of genetically-informative data lags behind our ability to characterize genomes. This dissertation contributes to the psychometric toolkit of behaviour genetics by providing theoretical as well as applied studies that show the potential that item response theory (IRT) has to offer to behaviour geneticists.

7.2.1 A psychometric approach to behaviour genetics

In a typical twin study, the scores on a test or questionnaire are added up in order to obtain a *sum score* and this measure is then used in the biometric analysis to estimate the relative importance of genetic and environmental influences. In this dissertation, instead, item scores were modelled by estimating an item response theory (IRT) measurement model while simultaneously modelling the biometric model. This was done to address a number of psychometric issues, summarized in the following.

Ability or personality tests are usually more suitable in discriminating between individuals scoring in the average range of the ability (or personality) continuum than discriminating between individuals that score in the left or right tail of the distribution. In other words, measurement error is not the same across all twins, but *heterogeneous*. Chapter 2 shows that, in a biometric analysis including genotype-environment interaction, this can lead to the finding of a spurious interaction effects. It is furthermore shown that the integration of an IRT measurement model solves this problem. Chapter 2 concentrates on genotype-environment interaction in the case that environmental variables and genotypes are unmeasured. The finding of spurious interaction effects is however not limited to this situation, but can also arise when a measured variable serves as a moderator on the variance components (i.e., ACE×M). Therefore, in Chapter 5, a method is introduced that incorporates an IRT model into the biometric ACE×M model.

In Chapters 3 and 4, the methodology introduced in Chapter 2 was applied to item test data that measured ability (mathematics performance, Chapter 3) and a personality trait (conservatism, Chapter 4). The sum scores distribution of the data of Chapter 3 as well as Chapter 4 was skewed, which highlights the relevance of the simulation study conducted in Chapter 2. In the simulation study, skewed test data was simulated to illustrate the finding of spurious genotype-environment interaction due to properties of the measurement scale. The sum score distribution of Chapter 3 and 4 show that the severity of skewness that was chosen for this simulation study is to be expected also in real data, confirming that it is realistic to find spurious effects with the magnitude observed in the simulated data. Furthermore, in order to directly compare results gained by the new methodology with the traditional sum score approach, in Chapter 4 (conservatism data), the same biometric model was estimated using sum scores instead of item scores. As the sum score approach does not take into account measurement unreliability, the estimated average environmental variance that also includes measurement error was much higher. Furthermore, results of the sum score approach suggested a positive interaction between additive genetic influences and unique-environment variance (i.e., A×E) and a negative interaction between additive genetic effects and common-environment variance (i.e., A×C). That is, the direction of the interaction effects changed compared to the results of the new methodology (i.e., negative A×E and positive A×C). Therefore, in this particular application, the interpretation of the data would have been completely opposite.

In general, a wrong conclusion concerning genotype-environment interaction does not only bias the results of a particular genotype-environment interaction study, but potentially also future research results. That is, the genotype-environment interaction method where the environment features as a latent variable (as in Chapters 2, 3 and 4) can be used to guide future

research: Depending on the results (e.g., significant A×E but non-significant A×C), either specific unique-environmental measures at the individual or common-environmental measures at the family level can be collected to investigate the exact nature of the interaction. As data collection is expensive and time consuming, collection of the "wrong" variables based on spurious interaction effects can lead to waste of resources.

While the integration of an IRT model does address the problem of heterogeneous measurement error, it does not solve the issues that come with poor scaling such as for example having little information of the right tail when there is a large ceiling effect. For example, in the study discussed in Chapter 3, a relatively low percentage of twins scored higher on the mathematics scale than the mode, leading to a skewed performance distribution. In other words, the test does not discriminate very well in the right tail of the distribution and estimates of genotype-environment interaction effects are based foremost on the information on the rest of the population. Therefore, we have to be cautious in drawing conclusions on children with extreme high ability. To draw reliable conclusions, a larger sample of mathematically talented twins who took a test with very difficult items would be required.

It should therefore be stressed here that it is still important to pursue proper scaling of the measurement by aiming at reliable and valid scales. The process of designing, testing and conducting a questionnaire or test should be given care and effort. The methods introduced in this dissertation to investigate A×E, A×C and ACE×M can correct for the spurious finding of interaction effects due to scale issues, but a thoroughly validated scale is a necessary perquisite to draw reliable conclusions. No amount of statistical analysis can compensate for methodological failures such as poor scaling.

In the methods described in this dissertation, IRT measurement models are used to conduct biometric analyses at the latent phenotypic level. That is, the biometric model (e.g., ACE model including ACE×M or A×E and/or A×C) is modelled on trait scores that are corrected for measurement error in the phenotypic trait. It is important to note that it is only controlled for error introduced by the limited amount of items that are available to assess the phenotypic trait. The methods introduced in this dissertation do not correct for systematic bias typical for self-reports (e.g. questionnaires) such as for example that items on sensitive topics (e.g. alcohol or drug use) might not be answered truthfully because twins wish to present themselves in a socially acceptable manner.

The (remaining) arbitrariness of scaling

In psychology, measures often do not have an absolute scale of measurement. Results regarding genotype-environment interaction replicate only when the same underlying scale is used, as every non-linear transformation leads to a different result. In this dissertation, an IRT model was used to define the

scale for a particular trait, which makes statistical inference independent of the choice of items used for a scale. Provided that the same IRT model is used and is identified in the same way, the methodology introduced in Chapter 3 and Chapter 5 can be used to make results comparable across studies (given that the same population was studied).

However, the scaling remains arbitrary. In this dissertation, it was chosen to use an IRT model, where a logistic function is used to model the relationship between ability, item characteristics and the probability of scoring a particular value on any particular item. This is an arbitrary choice, since other measurement models could be used as well which could potentially lead to different results. For example, instead of using the logistic function, another function (e.g. the step function from the Guttman scale), or no function at all (non-parametric IRT) or methods based on multidimensional scaling (see e.g. Borg & Groenen, 2005) could be used for the scaling of test performance. Using the logistic function has some benefits. First of all, the interpretation of IRT models is easy and intuitive: The probability that an individual endorses an item increases with the ability of that individual, expressed as the difference $(\theta - \beta)$ where θ denotes the ability and β the difficulty of the item. For example, if the individual is more able than the item is difficult (i.e., $\theta > \beta$ such that $\theta - \beta > 0$), the probability is large that the individual answers the item correctly. If the individual however is less able than the item is difficult (i.e., $\theta < \beta$ such that $\theta - \beta < 0$), then the probability that the individual endorses the item is low (Strobl, 2012). The logistic function can take values that are between zero and one, which makes it a well suited function to express probabilities (i.e., in case of test performance, the probability of answering an item correctly). Lastly, IRT models are flexible. There is no requirement that tests or questionnaires have the same length or contain exactly the same items in order to compare performance across individuals. This is related to a property that is unique to the Rasch family of IRT models, the principle of *specific objectivity*. This property implies that comparisons between individuals are independent of the items that are used to compare their abilities. That is, if individual B has a higher ability than individual A, the comparison of these two individuals will always be the same: Independent of the choice of item, individual B has a higher probability to endorse an item and therefore a higher ability (Strobl, 2012).

7.2.2 Beyond psychometrics

The genetics of educational achievement

A large part of this thesis was devoted to genetic analyses of educational achievement data on the *Eindtoets Basisonderwijs*, a Dutch national test that is administered every year in the last year of primary school. Using twins' item scores on the mathematics subscale of this test, in Chapter

3, it was tested whether there is genotype-environment interaction where the environment featured as latent variable. In the application study of Chapter 5, it was tested whether a family's socio-economic status (SES) had any moderating effects on individual differences in mathematical ability. Furthermore, in the application study of Chapter 6, the effect of school-aggregated measures (i.e., school denomination, school size and pedagogical philosophy) on individual differences was tested on twins' sum scores on the *Eindtoets Basisonderwijs*.

Genotype-environment interaction in mathematics ability

The results of Chapter 3 suggest that most of the variance in mathematics performance could be explained by genetic influences. A substantial part of phenotypic variance could be explained by unique-environmental influences while variance due to common-environmental influences was negligibly small. Furthermore, environmental influences were relatively more important in explaining individual differences in the high-ability students than in the low-ability students. The finding of a significant interaction with variance explained by unique-environmental influences (i.e., A×E) compared to a non-significant interaction with variance due to common-environmental influences (i.e., A×C) implies that factors that are not shared in children from the same family are the most important source for an interaction with genetic influences. So, a higher genetic predisposition for mathematical high ability was associated with more residual variance in mathematical ability.

What exactly does this mean? In the following, the result of a positive A×E will be explained more thoroughly using two theoretical scenarios. Imagine that we have measured mathematics performance in a large twin sample and conducted a variance decomposition to estimate the relative importance of additive genetic, common-environmental and unique-environmental influences in explaining individual differences in test scores. Our results show that, similar to the results of Chapter 3, variance due to common-environmental influences is negligibly small (i.e., $\sigma_C^2 \approx 0$). That is, similarities in a twin pair can only be explained by additive genetic effects, A_1 for the first twin of a twin pair and A_2 for the second twin. In the first scenario, variance explained by unique-environmental influences is zero and phenotypic variance is explained solely by genetic influences (i.e., $\sigma_A^2 \approx 1, \sigma_E^2 \approx 0$). In this scenario, the correlation between the phenotype and the genotype of a twin is at its maximum. If we have for example a DZ twin pair with different genotypes, the twin with a higher genotypic value (e.g., $A_1 > A_2$) will most certainly also have a higher phenotypic value such as for example a higher mathematics test score. In the second scenario, unique-environmental influences explain a large proportion (e.g., 50%) of individual differences (i.e., $\sigma_A^2 \approx .5, \sigma_E^2 \approx .5$). This introduces random noise in the sense that in some families, the twin with the lower genotype might

perform better on the mathematics test than his or her co-twin with a lower genetic value (i.e., $A_1 > A_2$, but $P_1 < P_2$). For example, the second twin might be ill on the day of the teacher's lesson on finding a common denomination in fraction addition. The weeks before the mathematics test, the teacher pays attention to new material and has no time to help the twin catching up on the missed lesson. Consequently, he or she is unable to answer all test questions related to the missed material and gets a lower test score than the co-twin. As genetic influences are still important, in this scenario, there are also families where this is not the case and the twin with the higher genotypic value will score higher on the mathematics test. In other words: There is still a positive correlation between phenotype and genotype, but this correlation is much lower than in the first scenario. A graphical representation of these two scenarios can be found in Figure 7.1.

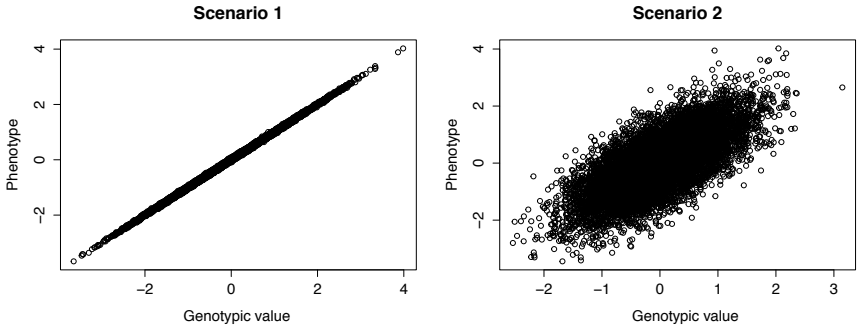


Figure 7.1: Graphical representation of the two theoretical scenarios (based on simulated phenotypes of 10,000 individual twins).

Together, these two scenarios can be seen as an extreme example of a positive A×E effect: In the first scenario, unique-environmental influences are not important and mathematics performance is mostly determined by a twin's genotype. This scenario is similar to the situation of the low-ability twins from Chapter 3. Their performance is better predicted by their genotype and there is only a small amount of random noise. In the second scenario, the performance of twins is not only determined by their genotypic value but also by random noise (e.g., a mathematics teacher that pays attention to one but not the other twin from the same family). This scenario is similar to the situation of the high-ability twins from Chapter 3. Different to the the low-ability twins, their performance is for a large part determined by random noise - residual variance was larger. Note that the two scenarios are only used to illustrate the results of Chapter 3. The interaction effect that was found is subtler than in the simplified scenarios. That is, with decreasing genotypic value variance due to unique-

environmental influences (residual variance) decreases and is very small for twins with a very low genotypic value.

In Chapter 3, no conclusions could be drawn on the nature and importance of *specific* unique-environmental influences that are more important for students genetically predisposed towards high mathematical ability than for students with a genetic predisposition towards low or average mathematical ability. Still, one important implication for future research can be drawn: future research should aim to explain within-family variance rather than between-family variance.

One factor that could be important is of course the school environment, but many twin pairs shared the same classroom which makes it unlikely that school influences contributed a lot to unique-environmental variance. Still, however, the school environment might be important. Although children in the same family appear to receive the "same" school environment, this does not necessarily mean that twins from the same family perceive this environment similarly (Plomin & Daniels, 1987). A process referred to as *genotype-environment correlation* in the behaviour genetics literature might be important. One can distinguish between three different types of genotype-environment correlation. The correlation can be *passive*, referring to an association between the genotype a twin inherits from the parents and the environment in which the twin is raised: The family environment that is created by the parents is also influenced by their own heritable characteristics. The fact that variance due to common-environmental influences was very small suggest however that it is not likely that this process was very important. The nature of genotype-environment correlation can furthermore be *evocative*, meaning that an individual's (heritable) behaviour evokes an environmental response. A mathematics teacher might recognize the mathematical talent of one twin and therefore pay more attention to him or her than to the co-twin who is not as talented. Lastly, genotype-environment correlation can be *active* − referring to a heritable propensity to actively seek out environmental exposure. A scenario that could potentially explain the pattern in Chapter 3 could be the following: twins with mathematical talent ask more questions, leading to amplification of effects (e.g., becoming even more talented because of a better understanding of the material), which would result in A×E.

In general, there is a broad range of influences that can contribute to differences in twins, ranging for example from the birth order to different perceptions of environments to subtle differences in brain structures. Future research should first focus on moderator variables that index influences that have proven to be important for talented students, such as for example peer influences (Austin & Draper, 1981), personality characteristics (Ackerman, 1997) and motivation (Vallerand, 1994). The methodology in Chapter 5 can be used to investigate whether these influences are more important for the high ability students.

Moderating effects of SES

The results of the application study in Chapter 5 suggest that common-environmental influences are less important in creating individual differences in mathematical ability in families with high SES than in creating individual differences in families with low or average SES. The increase in common-environmental variance for low SES families is in line with the findings of earlier studies that focused on SES interaction on intelligence (IQ), a trait that is closely related to mathematical ability and shows considerable genetic overlap (i.e., the same genes affect both IQ and mathematical ability, see e.g. Davis et al., 2009; Plomin & Deary, 2015). Analysing the scores of 623 7-year-old twin pairs on the Wechsler Intelligence Scale (WISC) for children, Turkheimer et al. (2003) found that, among children of low SES, differences in IQ scores were mainly accounted for by common-environmental influences. Additionally, Turkheimer et al. (2003) found that the heritability of IQ was much higher in children of above-average SES than it was for children of low SES.

We have, however, to be cautious in drawing conclusions. The methodology introduced in Chapter 5 is concerned with the univariate modelling of ACE×M which focuses on the moderation of variance components *unique* to the phenotypic variable. By regressing out effects on the moderator variable, any genetic or environmental effects that operate through or are common with this variable are partialled out; phenotypic scores are corrected for the influence of family SES. This makes the interpretation very difficult given the fact that SES has consistently been associated with higher academic achievement and cognitive performance throughout childhood and adolescence (e.g. White, 1982; Sirin, 2005), which suggests that there is also a (genetic) correlation between SES and mathematics ability. Due to the genetic correlation between SES and mathematical ability, mathematical ability might be underestimated for twins from high SES families, because their phenotypic latent scores are corrected for a measure that actually correlates with the trait itself.

Effects of school influences

The results of the application study in Chapter 6 show that none of the investigated school-aggregated measures had a significant effect on individual differences in educational achievement. However, a substantial part of the variance in *Eindtoets Basisonderwijs* sum scores could be explained by environmental influences, indicating that there are environmental influences that were not investigated that cause individual differences in *Eindtoets Basisonderwijs* test scores. Another explanation for some of the non-significant results might be that there was simply not enough variance to find effects on individual differences. For example, 74% of all individual twins went to a school with regular education while only a few went to schools with a different pedagogical philosophy. Likewise, most of the twins

visited either a roman-catholic school (31%), a state school (23%) or a protestant-christian school (17%), while only a small subset of twins went to a school with different denomination.

7.2.3 Future statistical developments

The method introduced in Chapter 2 estimates genotype-environment interaction in the case that the environment and genotypes feature as latent (i.e., unmeasured) variables, which makes an interpretation of results difficult. For a better interpretation, a follow-up study that collects moderator variables that index specific environmental influences can be conducted, using the methodology introduced in Chapter 5. The results of genotype-environment interaction in case of unmeasured variables and genotypes can then guide the data collection - depending on the significance of A×C and A×E, either common-environmental variables at the family level or unique-environmental measures at the individual twin level can be collected. The interpretation of a significant interaction effect however remains difficult. In a statistical sense, an interaction is very specific - the effect of one variable cannot be understood without taking into account the other variable. In case of genotype-environment interaction, the effect of the genes depends on the environment and/or the effect of the environment depends on the genes (Dick, 2011). These two alternative conceptualisations are however indistinguishable statistically which makes it impossible to determine causality.

The most straightforward method to detect genotype-environment interaction while also investigating the causal agent can be found in animal experimentation where different genetic strains of animals can be subjected to different environments. So, environmental exposure can be made random while the genotype is fixed, which eliminates the complicating factor of a genotype by environment correlation. However, this strategy is not possible to apply in human genetics due to ethical constraints. Therefore, in human traits, interpretations of genotype-environment interaction are complicated because environmental factors such as for example school and teacher effects are also *correlated* with the phenotypic trait of interest and therefore correlated with genetic influences on the trait. Twin as well as adoption studies have provided evidence for the presence of such genotype-environment correlations (Kendler & Baker, 2007; Plomin & Bergeman, 1991).

The presence of genotype-environment correlations complicates not only the interpretation of genotype-environment interaction results, but also the interpretation of biometric models extended with (environmental) covariates. In the application study of Chapter 6, measured aggregated school influences were included in the biometric model as covariates. These measures can also be (genetically) correlated to the phenotypic trait of interest. For example,

educational achievement and IQ are predictors for SES at a later age and therefore likely to be (genetically) correlated with each other.

The models presented here should therefore be extended to 1) multivariate models where measured (environmental) covariates can be modelled not only as covariates but as extra phenotypes (latent traits), 2) model genotype-environment interaction in the presence of genotype-environment correlation. This would not only improve the reliability of results, but make it possible to investigate more interesting research questions. Related to the genetic analyses of educational achievement data that were performed in this dissertation (as described above), it would be interesting for example to study how a genetic predisposition for high IQ interacts with a partly environmental but also partly genetic variable like school motivation to achieve high mathematics scores.

7.2.4 Conclusion

The use of IRT models has become more popular in many research fields, including for example educational science, public health or political science. IRT based analyses are, however, still scarce in behaviour genetics. Using sum scores, the relationship between the (theoretical) true score and the observed score is described in a linear fashion - that is, a twin's observed score is the total score and is different from the true score by only one error term. The IRT approach on the other hand is more difficult to understand and its application is often difficult and time consuming.

The aim of this dissertation was to show the potential that IRT has to offer to the field of behaviour genetics. Theoretical as well as applied studies demonstrate why behaviour geneticists should take the effort and replace sum score based analyses with IRT based analyses. It is for example shown that sum score based analyses can result in the finding of spurious genotype-environment interactions.

In this dissertation, scripts are provided which makes the application of the new methodology easy and approachable for twin researchers (see `http://github.com/ingaschwabe` for a digital version of all scripts). Furthermore, the IRT based methods can be easily applied by using the R package *BayesTwin* that includes most of the models described in this thesis (see `http://github.com/ingaschwabe` for a developmental version).

Nederlandse samenvatting

Dit proefschrift, getiteld "*Nature, Nature en Item Respons Theorie – Een Psychometrische Benadering van de Gedragsgenetica*" bespreekt een aantal psychometrische problemen waarmee rekening moet worden gehouden bij het analyseren van genetisch-informatieve data. Behandelde onderwerpen zijn de heterogeniteit van meetfouten, het gebruik van schalen met arbitraire eenheden, en het vergelijkbaar maken van fenotypes die met verschillende item sets zijn gemeten. Dit proefschrift laat zien dat het negeren van deze psychometrische problemen kan leiden tot vertekende resultaten, zoals bijvoorbeeld een spurieuze vondst van een interactie tussen genetische- en omgevingsinvloeden (i.e., gen-omgeving interactie). Tevens wordt besproken hoe het toepassen van modellen uit de item respons theorie (IRT) kan helpen om een potentiële vertekening van onderzoeksresultaten te voorkomen.

In hoofdstuk 2 wordt het modelleren van gen-omgeving interactie besproken. Hierbij zijn omgevingsinvloeden niet direct geobserveerd maar worden als latente variabele in het genetische model opgenomen. Terwijl traditioneel in een dergelijke analyse vaak som scores (i.e., bij elkaar opgetelde scores op de items van een test) worden gebruikt, wordt in dit hoofdstuk een nieuwe methode voorgesteld. Hierbij worden de ruwe item gegevens gebruikt door een expliciet meetmodel in de genetische analyse te integreren. Door middel van een simulatiestudie wordt aangetoond dat het gebruik van somscores kan leiden tot het ten onrechte vinden van gen-omgeving interactie ten gevolge van eigenschappen van de gebruikte meetschaal. De nieuwe methode is daarentegen niet gevoelig voor deze bias. Een tweede simulatie studie illustreert de statistische power van de nieuwe methode.

In hoofdstuk 3 wordt de methodologie uit hoofdstuk 2 toegepast. De resultaten van internationale studies wijzen uit dat de beste studenten van Nederland slechter presteren dan de beste studenten uit andere Westerse en Aziatische landen. Een verklaring voor deze ogenschijnlijk slechte prestatie wordt vaak gezocht in een tekortkoming van het huidige onderwijs voor deze Nederlandse leerlingen. Om ondersteuning te vinden voor deze hypothese, is in dit hoofdstuk onderzocht of variatie van omgevingsinvloeden afhankelijk is van de genotypische waarde van een leerling (i.e., genetische aanleg voor een hoge vaardigheid in wiskunde). Hiervoor zijn de ruwe item test scores van 2110 Nederlandse tweelingparen op het wiskunde deel van de *Eind-*

toets Basisonderwijs (voormalig: Cito eindtoets) gebruikt. De resultaten suggereren dat unieke omgevingsinvloeden inderdaad relatief belangrijker zijn voor het verklaren van verschillen in kinderen met aanleg voor een hoge vaardigheid in wiskunde dan voor kinderen met weinig aanleg voor wiskunde.

In hoofdstuk 4 zijn homogeniteits- en IRT-modellen gebruikt om de psychometrische kwaliteit van de Wilson-Patterson schaal te beoordelen. De resultaten laten zien dat de schaal feitelijk twee verschillende constructen meet: aan de ene kant verschillen respondenten in hoeverre zij instemmen met conservatief danwel liberaal gerichte stellingen en aan de andere kant verschillen zij in hoeverre zij geneigd zijn om de stellingen met een "?" danwel met een "Ja" of "Nee" te antwoorden. Op basis van deze resultaten is een nieuwe schaal ontwikkeld en de antwoorden op de items zijn vervolgens gebruikt voor een genetische gen-omgeving interactie analyse. Hiervoor is de methode geïntroduceerd in hoofdstuk 2 uitgebreid naar het gebruik van ordinale data. De resultaten suggereren dat individuele verschillen in het antwoordpatroon op deze schaal vooral zijn toe te wijzen aan genetische invloeden en gedeelde omgevingsinvloeden. Verder zijn er significante gen-omgeving interactie effecten gevonden die suggereren dat de variantie verklaard door niet-gedeelde omgevingsinvloeden groter is bij individuen met een genetische aanleg voor conservatisme, terwijl variantie verklaard door gedeelde omgevingsinvloeden kleiner was voor individuen met een genetische aanleg voor liberalisme.

In hoofdstuk 5 wordt een methode geïntroduceerd die een IRT model integreert in het modelleren van een variantie decompositie die gemodereerd wordt door een covariaat M (ACE×M). Hierbij is voor het biometrische ACE×M model een alternatieve parametrisatie gebruikt. Deze parametrisatie is, in tegenstelling tot de parametrisatie geïntroduceerd door Purcell (2002), volledig geïdentificeerd en daarom de betere keuze voor het analyseren van ACE×M modellen. Het biometrische model en het IRT-model worden tegelijkertijd geschat waardoor het vinden van een vals interactie-effect op basis van schaalproblemen wordt voorkomen. Simulatiestudies laten zien dat grote steekproeven nodig zijn om A×M en C×M interactie effecten te kunnen detecteren. De methode wordt geïllustreerd aan de hand van de modererende effecten van de socio-economische status van een familie op de etiologie van individuele verschillen in wiskundevaardigheid. Hiervoor zijn de eerder in hoofdstuk 2 gebruikte *Eindtoets Basisonderwijs* gegevens gebruikt. Er is een significante en negatieve C×SES interactie effect gevonden. Dit suggereert dat gedeelde omgevingsinvloeden minder belangrijk zijn voor het ontstaan van individuele verschillen in wiskundevaardigheid in families met een hoge sociaaleconomische status dan in families met een lage of gemiddelde sociaaleconomische status.

In hoofdstuk 6 wordt een nieuwe methode beschreven voor het omgaan met gedeeltelijk ontbrekende covariaatgegevens in tweelingdata. De traditionele manier om met de ontbrekende gegevens om te gaan resulteert vaak

in lage statistische power, omdat alleen fenotypische en covariaatgegevens van individuele tweelingen met volledige data meegenomen kunnen worden bij het schatten van een genetisch model. Door covariaten expliciet in de verwachte covariantiematrix op te nemen wordt het mogelijk om alle beschikbare gegevens te gebruiken. Een simulatiestudie laat zien dat deze nieuwe methode, onafhankelijk van aannames over de achtergrond van het ontbreken van data (i.e., MCAR/MAR/MNAR), het aantal covariaten (twee, drie en vier), of de hoeveelheid ontbrekende gegevens (2%, 6% en 10% van het totale aantal van covariaten antwoorden), een grotere power heeft. De nieuwe methode is toegepast om de effecten van school-geaggregeerde variabelen en het geslacht van een tweeling op test scores op de *Eindtoets Basisonderwijs* test van 990 tweelingparen te onderzoeken.

Bibliography

Ackerman, C. (1997). Identifying gifted adolescents using personality characteristics: Dabrowski's overexcitabilities. *Roeper Review*, *19*(4), 229–236.

Alarcon, M., Knopik, V. & DeFries, J. (2000). Covariation of mathematics achievement and general cognitive ability in twins. *Journal of School Psychology*, *28*, 63–77.

Allison, P. (2001). *Missing data*. Thousand Oaks, CA: SAGE Publications Inc.

Austin, A. & Draper, D. (1981). Peer relationships of the academically gifted: A review. *Gifted Child Quarterly*, *25*, 129-133.

Bartels, M., Rietveld, M., Van Baal, G. & Boomsma, D. (2002). Heritability of educational achievement in 12-year-olds and the overlap with cognitive ability. *Twin Research and Human Genetics*, *5*(6), 544-553.

Bartholomew, D., Steele, F., Galbraith, J. & Moustaki, I. (2008). *Analysis of multivariate social science data*. NY: Taylor.

Bauer, D. & Hussong, A. (2009). Psychometric approaches for developing commensurate measures across independent studies: traditional and new models. *Psychological Methods*, *2*, 101–125.

Béguin, A. & Glas, C. (2001). Mcmc estimation of multidimensional irt models. *Psychometrika*, *66*, 541–562.

Boada, R., Willcutt, E., Tunick, R., Chabildas, N., Olson, R., DeFries, J. & Pennington, B. (2002). A twin study of the etiology of high reading ability. *Reading and Writing*, *15*, 683–707.

Boer, G. D., Minnaert, A. & Kamphof, G. (2013). Gifted education in the netherlands. *Journal of the Education of the Gifted*, *36*(1), 133–150.

Boker, S., Neale, M., Maes, H., Wilde, M., Spiegel, M., Brick, T., ... Fox, J. (2011). Openmx: An open source extended structural equation modeling framework. *Psychometrika*, *76*(2), 306–317.

Bolstad, W. (2007). *Introduction to bayesian statistics*. Hoboken, New Jersey: John Wiley & Sons.

Boomsma, D., Vink, J., Beijsterveldt, C., de Geus, E., Beem, A., Mulder, E., ... van Baal, G. (2002). Netherlands twin register: A focus on longitudinal research. *Twin Research and Human Genetics*, *5*, 401–406.

Borg, I. & Groenen, P. (2005). *Modern multidimensional scaling.* New York: Springer.

Bouchard, T., Segal, N., Tellegen, A., McGue, M., Keyes, M. & Krueger, R. (2003). Evidence for the construct validity and heritability of the wilson-patterson conservatism scale: a reared-apart twins study of social attitudes. *Personality and Individual Differences*, *34*, 959–969.

Box, G. & Tiao, G. (1973). *Bayesian inference in statistical analysis.* Reading, Mass: Wiley-Interscience.

Bukowski, W., Dionne, G., Tremblay, R., Pe, D., Brendgen, M., Vitaro, F., ... Perusse, D. (2009). Gene-environment interplay between peer rejection and depressive behavior in children. *Journal of Child Psychology and Psychiatry*, *50*(8), 1009–1017.

Cadoret, R., Cain, C. & Crowe, R. (1983). Evidence for gene-environment interaction in the development of adolescent antiscocial behavior. *Behavior Genetics*, *13*(3), 301–310.

Cameron, N. (1993). Methodologies for estimation of genotype with environment interaction. *Livestock Production Science*, *35*, 237–249.

Campbell, A., Converse, P., Miller, W. & Stokes, D. (1960). *The american voter.* New York: John Wiley & Sons, Inc.

Carney, D., Jost, J., Gosling, S. & Potter, J. (2008). The secret lives of liberals and conservatives: Personality profiles, interaction styles, and the things they leave behind. *Political Psychology*, *29*(6), 807–840.

Caspi, A., McClay, J., Moffitt, T., Mill, J., Martin, J., Craig, I., ... Poulton, R. (2002). Role of genotype in the cycle of violence in maltreated children. *Science*, *297*(5582), 851–854.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29. Verkregen van `http://www.jstatsoft.org/v48/i06/`

Converse, P. (1964). The nature of belief systems in mass publics (1964). *Critical Review: A Journal of Politics and Society*, *18*(1–3), 1–74.

Davis, O., Band, G., Pirinen, M., Haworth, C., Meaburn, E., Kovas, Y., ... Hunt, S. (2014). The correlation between reading and mathematics ability at age twelve has a substantial genetic component. *Nature communications*, *5*.

Davis, O., Haworth, C. & Plomin, R. (2009). Learning abilities and disabilities: Generalist genes in early adolescence. *Cognitive Neuropsychiatry*, *14*(4-5), 312–331.

Dekker, S. (2014). *Plan van aanpak toptalenten 2014-2018.* Den Haag: Ministerie van Onderwijs, Cultuur en Wetenschap.

de Leeuw, J. & Mair, P. (2009). Gifi methods for optimal scaling in R: The package homals. *Journal of Statistical Software*, *31*(4), 1–20. Verkregen van `http://www.jstatsoft.org/v31/i04/`

development core team, R. (2013). R: A language and environment for statistical computing [Handleiding van computersoftware]. Austria, Vienna. Verkregen van `http://www.R-project.org`

Dick, D. (2011). Gene-environment interaction in psychological traits and disorders. *Annual Review Clinical Psychology*, *7*, 383–409.

Eaves, L. (1983). Errors of inference in the detection of major gene effects on psychological test scores. *The American Journal of Human Genetics*, *35*, 1189–1983.

Eaves, L. (2006). Genotype x environment interaction in psychopathology: Fact or artifact? *Twin Research and Human Genetics*, *9*(1), 1–8.

Eaves, L. & Erkanli, A. (2003). Markov chain monte carlo approaches to analysis of genetic and environmental change and g x e interaction. *Behavior Genetics*, *33*(3).

Eaves, L., Heath, A., Martin, N., Maes, H., Neale, M., Kendler, K., ... Corey, L. (1999). Comparing the biological and cultural inheritance of personailty and social attitudes in the virgina 30,000 study of twins and their relatives. *Twin Research and Human Genetics*, *2*, 62–80.

Eaves, L., Last, K., Martin, N. & Jinks, J. (1977). A progressive appraoch to non-additivity and genotype-environmental covariance in the analysis of human differences. *British Journal of Mathematical and Statistical Psychology*, *30*, 1–42.

Eaves, L., Martin, N., Heath, A., Schieken, R., Meyer, J., Silberg, J., ... Corey, L. (1997). Age changes in the causes of individual differences in conservatism. *Behavior Genetics*, *27*, 121–124.

Embretson, S. & Reise, S. (2009). *Item response theory for psychologists*. New Jersey: Psychology Press.

Faith, M., Berkowitz, R., Stallings, V., Kerns, J., Storey, M. & Stunkard, A. (2004). Analysis of a gene-environment interaction parental feeding attitudes and styes and child body mass index: Prospective analysis of gene-environment interaction. *Pediatrics*, *114*(4), e429 -e436.

Falconer, D. & MacKay, T. (1995). *Introduction to quantitative genetics*. Essex, UK: Pearson Education Limited.

Fox, J. & Glas, C. (2003). Bayesian modeling of measurement error in predictor variables. *Psychometrika*, *68*, 169–191.

Friend, A., DeFries, J., Olson, R., Pennington, B., Harlaar, N., Byrne, B., ... Keenan, J. (2009). Heritability of high reading ability and its interaction with parental education. *Behavior Genetics*, *39*(4), 427–436.

Galton, F. (1869). *Hereditary genius*. London, UK: Macmillan.

Gelfand, A. & Smith, A. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, *85*, 398–409.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, *1*(3), 515–533.

Gelman, A., Carlin, J., Stern, H. & Rubin, D. (2004). *Bayesian data analysis* (2nd dr.). London: Chapman and Hall.

Gelman, A. & Rubin, D. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*, 457–511.

Geman, S. & Geman, D. (1984). Stochastic relaxation, gibbs distributions and the bayesian restoration o images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *6*, 721–741.

Gifi, A. (1990). *Nonlinear multivariate analysis*. New York: Wiley.

Gilks, W., Richardson, S. & Spiegelhalter, D. (1996). *Markov chain monte carlo in practice*. Boca Raton, FL: Chapman & Hall/CRC.

Harden, K. P., Turkheimer, E. & Loehlin, J. (2007). Genotype by environment interaction in adolescent's cognitive aptitude. *Behavior Genetics*, *37*(2), 273–283.

Hart, S., Petrill, S., Thompson, L. & Plomin, R. (2009). The abcs of math: A genetic analysis of mathematics and its links with reading ability and general cognitive ability. *Journal of Educational Psychology*, *101*(2), 388–340.

Hatemi, P., Funk, C., Medland, S., Maes, H., Silberg, J. & Eaves, N. M. L. (2009). Genetic and environmental transmission of political attitudes over a life time. *Journal of Politics*, *71*(3), 1141–1156.

Hatemi, P., Klemmensen, R., Medland, S., Oskarsson, S., Littvay, L., Dawes, C., ... Martin, N. (2014). Genetic influences on political ideologies: Genome-wide findings on three populations, and a mega-twin analysis of 19 measures of political ideologies from five western democracies. *Behavior Genetics*.

Heath, A., Eaves, L. & Martin, N. (1998). Interaction of marital status and genetic risk for symptoms of depression. *Twin Research and Human Genetics*, *1*(3), 119–122.

Heiser, W. & Meulman, J. (1994). Homogeneity analysis: Exploring the distribution of variables and their nonlinear relationship. In M. Greenacre, J. Blasius & W. Kristof (red.), *Corrspondence analysis in the social sciences: Recent developments and applications*. London, UK: Harcourt Brace & Co. Publishers.

Henningham, J. (1996). A 12-item scale of social conservatism. *Personality and Inivididual Differences*, *20*(4), 517–519.

Hessen, D. & Dolan, C. (2009). Heteroscedastic one-factor models and marginal maximum likelihood estimation. *British Journal of Mathematical and Statistical Psychology*, *62*, 57–77.

Hibbing, J. R., Smith, K. & Alford, J. (2014). Differences in negativity bias underlie variations in political ideology. *Behavioral and Brain Science*, *37*(3), 297–350.

Hicks, B., DiRago, A., Iacono, W. & McGue, M. (2009). Gene-environment interplay in internalizing discorders: consistent findings across six environmental risk factors. *Journal of Child Psychology and Psychiatry*, *50*(10), 1309–1317.

IBM. (2013). Released 2013. ibm spss statistics for windows, version 22.0 [Handleiding van computersoftware]. Armonk, NY. (3-900051-07-0)

Inbar, Y., Pizarro, D. & Bloom, P. (2009). Conservatives are more easily disgusted than liberals. *Cognition and Emotion*, *23*, 714–725.

Jinks, J. & Fulker, D. (1970). Comparison of the biometrical genetical, mava, and classical approaches to the analysis of human behavior. *Psychological Bulletin*, *73*, 311–349.

Joanes, D. & Gill, C. (1998). Comparing measures of sample skewness and kurtosis. *The Statistician*, *47*, 183–189.

Johnson, W. & Krueger, R. (2005). Higher perceived life control decreases genetic variance in physical health: evidence from a national twin study. *Personality and Social Psychology*, *88*, 165–173.

Kanazawa, S. (2010). Why liberals and atheists are more intelligent. *Social Psychology Quarterly*, *73*(1), 33–57.

Kendler, K. & Baker, J. (2007). Genetic influences on measures of the environment: a systematic review. *Psycholigal Medicine*, *37*, 615–626.

Kim-Cohen, J., Caspi, A., Taylor, A., Williams, B., Newcombe, R., Craig, I. & Moffitt, T. (2006). Maoa, maltreatment, and gene-environment interaction predicting children's mental health: new evidence and a meta-analysis. *Molecular Psychiatry*, *11*, 903–913.

Kovas, Y., Haworth, C., Petrill, S. & Plomin, R. (2007). Mathematical ability of 10-year-old boys and girls: genetic and environmental etiology of normal and low performance. *Journal of Learning Disabilities*, *40*, 554–567.

Lau, J. & Eley, T. (2008). Disentangling gene environment correlations and interactions on adolescent depressive symptoms. *Journal of Child Psycholgoy and Psychiatry*, *49*, 142–150.

Lews-Beck, M., Bryman, A. & Liao, T. (2004). *The sage encyclopedia of social science research methods.* Thousand Oaks, CA: SAGE Publications.

Li, Y. & Baser, R. (2012). Using r and winbugs to fit a generalized partial credit model for developing and evaluating patient-reported outcomes assessments. *Statistical Medicine*, *31*(18).

Little, R. & Rubin, D. (2002). *Statistical analysis with missing data.* NY, NY: John Wiley & Sons.

Loehlin, J. & Nichols, P. (1976). *Heredity, environment, and personality: a set of 850 twins.* Austin: University of Texas Press.

Lord, F. (1980). *Applications of item response theory to practical testing problems.* Hilsdale, NJ: Erlbaum.

Lunn, D., Thomas, A., Best, N. & Spiegelhalter, D. (2000). A bayesian modeling framework: concepts, structure, and extensibility. *Statistics and Computing*, *10*, 325–337.

Markowitz, E., Willemsen, G., Trumbetta, S., van Beijsterveldt, T. & Boomsma, D. (2005). The etiology of mathematical and reading (dis)ability covariation in a sample of dutch twins. *Twin Research and Human Genetics*, *8*(6), 585–593.

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

Martin, N., Eaves, L., Heath, A., Jardine, R., Feingold, L. & Eysenck, H. (1986). Transmission of social attitudes. *Proceedings of the National Academy of Sciences USA*, *83*, 4364–4368.

Martin, N., Eaves, L., Kearsey, M. & Davies, P. (1978). The power of the classical twin study. *Heredity*, *40*, 97–116.

Masters, G. (1982). A rasch model for partial credit scoring. *Psychometrika*, *47*, 149–174.

Meelissen, M., Netten, A., Drent, M., Punter, R., Droop, M. & Verhoeven, L. (2012). *Pirls en timss 2011. trends in leerprestaties in lezen, rekenen en natuuronderwijs*. Nijmegen & Enschede, Netherlands: Radboud Universiteit & Universiteit Twente.

Minne, B., Rensman, M., Vroomen, B. & Webbink, D. (2007). *Excellence for productivity? bijzondere publicatie 69*. Den Haag, Netherlands: Centraal Planbureau.

Molenaar, D. & Dolan, C. (2014). Testing systematic genotype by environment interactions using item level data. *Behavior genetics*, *44*, 212–231.

Molenaar, D., van der Sluis, S., Boomsma, D. I. & Dolan, C. (2012). Detecting specific genotype by environment interactions using marginal maximum likelihood estimation in the classical twin design. *Behavior Genetics*, *42*, 483–499.

Muraki, E. (1992). A generalized partial credit model: Application of an em algorithm. *Applied Psychological Measurement*, *16*(2), 159–176.

Neale, B., Ferreira, M., Medland, S. & Posthuma, D. (2008). *Statistical genetics*. New York: Taylor & Francis.

Oliver, B., Harlaar, N., Thomas, M., Kovas, Y., Walker, S., Petrill, S., ... Plomin, R. (2004). A twin study of teacher-reported mathematics performance and low performance in 7-year-olds. *Journal of Educational Psychology*, *96*, 504–517.

Pedhazur, E. & Schmelkin, L. (1991). *Measurement, design, and analysis: An integrated approach*. East Sussex, UK: Psychology Press.

Petrill, S., Kovas, Y., Hart, S., Thompson, L. & Plomin, R. (2009). The genetic and environmental etiology of high math performance in 10-year-old twins. *Behavior Genetics*, *39*(4), 371–379.

Petrill, S., Saudino, K., Cherny, S., Emde, R., Fulker, D., Hewitt, J. & Plomin, R. (1998). Exploring the genetic and environmental etiology of high general cognitive ability in fourteen- to thirty-six month-old twins. *Child Development*, *69*, 68–74.

Plomin, R. & Bergeman, C. (1991). The nature of nurture: genetic influence on environmental measures. *Behavioral and Brain Sciences*, *14*, 373–427.

Plomin, R. & Daniels, D. (1987). Why are children in the same family so different from each other? *Behavioral and Brain Sciences*, *10*, 1–16.

Plomin, R. & Deary, I. (2015). Genetics and intelligence differences: five special findings. *Molecular Psychiatry*, *20*, 98–108.

Plummer, M. (2003). *Jags: A program for analysis of bayesian graphical models using gibbs sampling.*

Plummer, M. (2013). rjags: Bayesian graphical models using mcmc [Handleiding van computersoftware]. Verkregen van `http://CRAN.R-project.org/package=rjags` (R package version 3-10)

Purcell, S. (2002). Variance components models for gene-environment interaction in twin analysis. *Twin Research and Human Genetics*, *5*(6), 554–571.

Ronald, A., Spinath, F. & Plomin, R. (2002). The etiology of high cognitive ability in early childhood. *High Ability Studies*, *13*, 103–114.

Rubin, D. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592.

Samejima, F. (1969). Estimation of latent ability using a pattern of graded scores. *Psychometrika, Monograph Supplement*(17).

SanChristobal-Gaudy, M., Elsen, J., Bodin, L. & Chevalet, C. (1998). Prediciton of the response to a selection for canalisation of a continous trait in animal breeding. *Genetics, Selection, Evolution*, *30*, 423–451.

Saudino, K., Plomin, R., Pedersen, N. & McClearn, G. (1994). The etiology of high and low cognitive ability during the second half of the life span. *Intelligence*, *19*, 359–371.

Schreiber, D., Fonzo, G., Simmons, A., Dawes, C., Flagan, T., Fowler, J. & Paulus, M. (2013). Red brain, blue brain: Evaluative processes differ in democrats and republicans. *Plos One*, *8*(2).

Schwabe, I. & van den Berg, S. (2014). Assessing genotype by environment interaction in case of heterogeneous measurement error. *Behavior Genetics*, *44*(4), 394–406.

Shakeshaft, N., Trzaskowski, M., McMilan, A., rimfeld, K., Krapohl, E., Haworth, C., ... Plomin, R. (2013). Strong genetic influence on a uk nationwide test of educational achievementat the end of compulsory education at age 16. *Plos One*, *8*(12).

Sirin, S. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research*, *75*(3), 417–453.

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

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

Strobl, C. (2012). *Das rasch model.* Muenchen, Germany: Rainer Hampp Verlag.

Tucker-Drob, E., Harden, K. & Turkheimer, E. (2009). Combining nonlinear biometric and psychometric models of cognitive abilities. *Behavior Genetics*, *39*, 461-471.

Turkheimer, E. (2000). The three laws of behavior genetics and what they mean. *Current Directions in Psychological Science*, *9*(5).

Turkheimer, E., Haley, A., Waldron, M., D'Onofrio, B. & Gottesman, I. I. (2003). Socioeconomic status modifies heritability of iq in young children. *Psychological Science*, *14*(6), 623–628.

Turkheimer, E. & Horn, E. (2013). Interactions between socioeconomic status and components of variation in cognitive ability. In *Behavior genetics of cognition across the lifespan* (pp. 41–68). New York: Springer.

Turkheimer, E. & Waldron, M. (2000). Nonshared environment: theoretical, methodological, and quantitative review. *Psychological Bulletin*, *1*, 78–108.

Tuvblad, C., Grann, M. & Lichtenstein, P. (2006). Heritability for adolescent antisocial behavior differs with socioeconomic status: gene-environment interaction. *Journal of Child Psychology and Psychiatry*, *47*(7), 734–743.

Vallerand, R. (1994). A comparison of the school intrinsic motivation and perceived competence of gifted and regular students. *Gifted Child Quarterly*, *38*(4), 172–175.

van den Berg, S., Beem, L. & Boomsma, D. (2006). Fitting genetic models using winbugs. *Twin Research and Human Genetics*, *9*, 334–342.

van den Berg, S., de Moor, M., McGue, M., Pettersson, E., Terracciano, A., Verweij, K., ... Derringer, J. (2014). Harmonization of neuroticism and extraversion phenotypes across inventories and cohorts in the genetics of personality consortium: an application of item response theory. *Behavior Genetics*, *44*(4), 295–313.

van den Berg, S., Glas, C. & Boomsma, D. (2007). Variance decomposition using an irt measurement model. *Behavior Genetics*, *37*, 604–616.

van den Berg, S. & Service, S. (2012). Power of irt in gwas: Successful qtl mapping of sum score phenotypes depends on interplay between risk allele frequency, variance explained by the risk allele, and test characteristics. *Genetic Epidemiology*, *36*(8), 882–889.

van der Kloot, W. (1997). *Meerdimensionele schaaltechnieken voor gelijkenis- en keuzedata*. Utrecht, the Netherlands: Uitgeverij Lemma BV.

van der Sluis, S., Dolan, C., Neale, M., Boomsma, D. & Posthuma, D. (2006). Detecting genotype-environment interaction in monozygotic twin data: comparing the jinks and fullker test and a new test based on marginal maximum likelihood estimation. *Twin Research and Human Genetics*, *9*(3), 377–392.

van der Sluis, S., Posthuma, D. & Dolan, C. (2012). A note on false positives and power in gxe modelling of twin data. *Behavior Genetics*, *42*(1), 170–186.

van der Sluis, S., Posthuma, D., Nivard, M., Verhage, M. & Dolan, C. (2013). Power in gwas: lifting the curse of the clinical cut-off. *Molecular Psychiatry*, *18*, 2–3.

van der Sluis, S., Verhage, M., Posthuma, D. & Dolan, C. (2010). Phenotypic complexity, measurement bias, and poor phenotypic resolution contribute to the missing heritability problem in genetic association studies. *Plos One*.

van der Steeg, M., Vermeer, N. & Lanser, D. (2011). *Nederlandse onderwijsprestaties in perspectief*. Den Haag: Centraal Planbureau.

Veldkamp, B. & Paap, M. (2013). *Robust automated test assembly for testlet based tests: an illustration with the analytical reasoning section of the lsat (lsac research report* (Rapport). Newton: Law School Admission Council.

Verhelst, N., Glas, C. & Verstralen, H. (1995). *One-parameter logistic model*. Arnhem, Netherlands: CITO.

White, K. (1982). The relation between socioeconomic status and academic achievement. *Psychological Bulletin*, *91*(3), 461–481.

Wilson, G. (1973). *The psychology of conservatism*. New York: Academic Press.

Wilson, G. & Patterson, J. (1968). A new measure of conservatism. *British Journal of Social and Clinical Psychology*, *7*(4), 264–269.

Yang, J., Lee, H., Goddard, M. & Visscher, P. (2011). Gcta: A tool for genome-wide complex trait analysis. *American Journal of Human Genetics*, *88*(1), 76–82.

Acknowledgements

This dissertation is the result of three years of work carried out at the department of Research Methodology, Measurement and Data Analysis of the University of Twente. I would like to take this opportunity to thank all people who supported and encouraged me during this time. First of all, I would like to thank Stéphanie van den Berg for her dedication, our countless discussions and for pushing me beyond my own limits. My gratitude extends to Cees Glas for providing me with support and advice throughout the years and giving me the possibility to work on an extra project.

Also, I would like to thank my colleagues for the pleasant working environment and the many lunch walks. Special thanks go to Sukaesi Marianti for being a wonderful office mate in three (and a half) different offices.

This dissertation consisted of a collaboration with the Netherlands Twin Register at the VU University in Amsterdam and the psychometric group of Cito in Arnhem. I would like to express my gratitude to Stéphanie van den Berg, Anton Béguin and Dorret Boomsma for making this research possible in the first place by initiating the collaboration and writing the research proposal. I would also like to thank Ron Engelen and Wilco Budding for the time and effort they have put into the linking of the twin data to the Eindtoets Basisonderwijs data.

I am also very grateful for the help and support of my family and friends. A big thank you goes to my paranymphs, Sybren Wijnia and Birte Böning. I would also like to thank Rosanne Andriessen for taking the time to help me with the design of my cover. Finally, I would like to thank Papa, Mama and Arne Schwabe.

Appendix

A digital version of all scripts printed here can be found at
`http://github.com/ingaschwabe`

Appendix A

Following JAGS script incorporates an ACE model with an 1PL IRT model while modelling genotype-environment (A×E) interaction

```
#y_dz = Item responses of DZ twins (matrix)
#y_mz = Item responses of MZ twins (matrix)
#n_mz = Number of MZ twin pairs,
#n_dz = Number of DZ twin pairs,
#n_items = Number of items administered
#b = item parameters, assumed known in the analysis

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

#This results in a matrix of n_mz (or, in
#case of y_dz, n_dz) rows and 2*n_items columns
#(e.g., y_mz[1,22] is the response of
#MZ twin 1 from family 1 to item 22

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

#for (i in 1:n.items){
#  b[i] ~ dnorm(0,.1)
#  }
#Mu then has to be set to zero to identify the scale

#JAGS uses precision parameters for the variance
#parameters. Therefore, after running the script,
```

```
27 #these precision parameters should be inverted.
28 #e.g.: var_c = 1/outputAnalysis$tau_c[,,1]
29 #with the rjags package
30
31 model{
32
33 ##MZ twins
34 for (fam in 1:n_mz){
35   c_mz[fam] ~ dnorm(mu, tau_c)
36   f_mz[fam] ~ dnorm(c_mz[fam], tau_a)
37   a_mz[fam] <- f_mz[fam] - c_mz[fam]
38   tau_e[fam] <- 1/(exp(beta0 + (beta1*a_mz[fam])))
39
40  for (twin in 1:2){
41    pheno_mz[fam,twin]~ dnorm(f_mz[fam],tau_e[fam])
42   }
43
44 #1pl model twin1
45  for (k in 1:n_items){
46    logit(p[fam,k]) <- pheno_mz[fam,1] - b[k]
47    y_mz[fam,k] ~ dbern(p[fam,k])
48   }
49
50 #1pl model twin2
51  for (k in (n_items+1):(2*n_items)){
52          logit(p[fam,k]) <- pheno_mz[fam,2] - b[k-n_items]
53          y_mz[fam,k] ~ dbern(p[fam,k])
54  }
55 }
56
57 ##DZ twins
58 for (fam in 1:ndz){
59   c_dz[fam] ~ dnorm(mu, tau_c)
60   f0_dz[fam] ~ dnorm(c_dz[fam], doubletau_a)
61
62  for (twin in 1:2){
63          f_dz[fam,twin] ~ dnorm(f0_dz[fam], doubletau_a)
64          a_dz[fam,twin] <- f_dz[fam,twin] - c_dz[fam]
65          tau_e_dz[fam,twin] <- 1/(exp(beta0 +
66                                  (beta1*a_dz[fam,twin])))
67          pheno_dz[fam,twin] ~ dnorm(f_dz[fam,twin],
68                                  tau_e_dz[fam,twin])
69  }
70
71 #1pl model twin1 (DZ)
```

```
72   for (k in 1:n.items){
73    logit(p2[fam,k]) <- pheno_dz[fam,1] - b[k]
74     Ydz[fam,k] ~ dbern(p2[fam,k])
75   }
76
77  #1pl model twin2 (DZ)
78   for (k in (n_items+1):(2*n_items)){
79            logit(p2[fam,k]) <- pheno_dz[fam,2] - b[k-n_items]
80            Ydz[fam,k] ~ dbern(p2[fam,k])
81           }
82
83  }
84
85  #Prior distributions
86  mu ~ dnorm(0,.1)
87  beta1 ~ dnorm(0,.1)
88  beta0 ~ dnorm(0,1)
89
90  doubletau_a <- 2*tau_a
91  tau_a ~ dgamma(1,1)
92  tau_c ~ dgamma(1,1)
93  }
```

Appendix B

Following JAGS script incorporates an ACE model with a GPCM IRT model while modelling genotype-environment (A×E and A×C) interaction


```
#y_dz = Item responses of DZ twins (matrix)
#y_mz = Item responses of MZ twins (matrix)
#n_mz = Number of MZ twin pairs
#n_dz = Number of DZ twin pairs
#n_items = Number of items administered

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

#This results in a matrix of n_mz (or, in case of y_dz, n_dz)
#rows and 2*n_items columns. e.g. y_mz[1,22] is the response
#of MZ twin 1 from family 1 to item 22 if n_items = 22

#JAGS uses precision parameters for the variance parameters.
#Therefore, after running the script, these precision parameters
#have to be inverted. For example:
#var_a <-  1/outputAnalysis$tau_a[,,1] with the rjags package

model{
#MZ twins
for (i in 1:n_mz){
  c_mz[i] ~ dnorm(0, tau_c_mz[i])
  a_mz[i] ~ dnorm(0, tau_a)

  tau_c_mz[i] <- 1/(exp(gamma0 + gamma1*a_mz[i]))
  tau_e_mz[i] <- 1/(exp(beta0 + beta1*a_mz[i]))

  #Phenotypic values:
  mz[i,1] ~ dnorm(a_mz[i] + c_mz[i], tau_e_mz[i])
  mz[i,2] ~ dnorm(a_mz[i] + c_mz[i], tau_e_mz[i])

  for (j in 1:n_items){
    for (k in 1:3){
          eta[i,j,k] <- alpha[j] *
                        (mz[i,1]-beta[j,k])
          psum[i,j,k] <- sum(eta[i,j,1:k])
          exp_psum[i,j,k] <- exp(psum[i,j,k])
          prob[i,j,k] <- exp_psum[i,j,k]/sum(exp_psum[i,j,1:3])
     }
```

```
41      }
42
43    for (j in (n_items+1):(2*n_items)){
44      for (k in 1:3){                                                    135
45            eta[i,j,k] <- alpha[j-n_items] *
46                        (mz[i,2]-beta[j-n_items,k])
47            psum [i,j,k] <- sum(eta[i,j,1: k])
48            exp_psum [i,j,k] <- exp( psum [i , j , k])
49            prob [i,j,k] <- exp_psum[i,j,k]/sum(exp_psum [i,j,1:3])
50      }
51    }
52
53    for (j in 1:(2*n_items)){
54      y_mz[i,j] ~ dcat (prob[i,j,1:3])
55    }
56
57  } #end MZ twins
58
59  #DZ twins
60  for (i in 1:n_dz){
61    c_dz[i] ~ dnorm(0, 1)
62
63    a1_dz[i] ~ dnorm(0, doubletau_a)
64    a2_dz[i,1] ~ dnorm(a1_dz[i], doubletau_a)
65    a2_dz[i,2] ~ dnorm(a1_dz[i], doubletau_a)
66
67    tau_c_dz[i,1] <- exp(gamma0 + gamma1*a2_dz[i,1])
68    tau_c_dz[i,2] <- exp(gamma0 + gamma1*a2_dz[i,2])
69
70    c_dz_twin1[i] <- c_dz[i] * sqrt(tau_c_dz[i,1])
71    c_dz_twin2[i] <- c_dz[i] * sqrt(tau_c_dz[i,2])
72
73    tau_e_dz[i,1] <- 1/(exp(beta0 + beta1*a2_dz[i,1]))
74    tau_e_dz[i,2] <- 1/(exp(beta0 + beta1*a2_dz[i,2]))
75
76    dz[i,1] ~ dnorm(a2_dz[i,1] + c_dz_twin1[i], tau_e_dz[i,1])
77    dz[i,2] ~ dnorm(a2_dz[i,2] + c_dz_twin2[i], tau_e_dz[i,2])
78
79    for (j in 1:n_items){
80      for (k in 1:3){
81            etadz[i,j,k] <- alpha [j] * (dz[i,1]-beta [j,k])
82            psumdz [i,j,k] <- sum(etadz[i,j,1: k])
83            exp_psumdz[i,j,k] <- exp(psumdz [i,j,k])
84            probdz [i,j,k] <- exp_psumdz [i,j,k]/
85                          sum(exp_psumdz[i,j,1:3])
```

```
86        }
87      }
88
89       for (j in (n_items+1):(2*n_items)){
90         for (k in 1:3){
91               etadz[i,j,k] <- alpha[j-n_items] *
92                           (dz[i,2]-beta [j-n_items,k])
93               psumdz [i,j,k] <- sum(etadz[i,j,1:k])
94               exp_psumdz [i,j,k] <- exp(psumdz [i,j,k])
95               probdz [i,j,k] <- exp_psumdz [i,j,k]/
96                                 sum(exp_psumdz [i,j,1:3])
97         }
98      }
99
100     for (j in 1:(2*n_items)){
101       y_dz[i,j] ~ dcat (probdz[i,j,1:3])
102      }
103   } #end DZ twins
104
105   #DZ twins genetic correlation 0.5:
106   doubletau_a <- 2*tau_a
107
108   #Set alpha of item 3 to 1 to identify the scale
109   alpha[3] <- 1
110
111   #for the rest of the alpha parameters:
112   #lognormal prior with expectation of 0
113   #and variance of 10
114   alpha[1] ~ dlnorm(0, .1)
115   alpha[2] ~ dlnorm(0, .1)
116
117   for (j in 4:n_items){
118     alpha[j] ~ dlnorm(0, .1)
119   }
120
121   #Beta IRT parameters:
122   for (j in 1:n_items){
123     beta [j , 1] <- 0.0
124     for (k in 2:3){
125       beta [j , k] ~ dnorm (0, .1)
126     }
127   }
128
129   #Priors distributions:
130   tau_a ~ dgamma(1,1)
```

```
131  beta0 ~ dnorm(-1,.5)
132  beta1 ~ dnorm(0,.1)
133  gamma0 ~ dnorm(-1,.5)
134  gamma1 ~ dnorm(0,.1)
135  }
```

Appendix C

On the indeterminacy of Purcell's ACE×M parametrization

Purcell's univariate model for genotype-environment interaction resembles the general ANOVA model with a two-way interaction deceptively close:

$$P_{ij} = \beta_0 + \beta_1 M_{ij} + \beta_2 A_{ij} + \beta_3 A_{ij} M_{ij} + eE_{ij} \quad (1)$$

$$E_{ij} \sim N(0,1) \quad (2)$$

$$A_{ij} \sim N(0,1) \quad (3)$$

We have an intercept β_0, a main effect of the measured covariate M_{ij}, β_1, a main effect of the unmeasured genotypic value A_{ij}, β_2, an interaction effect of the measured covariate and the genotypic value, β_3, and a residual term with variance e^2. This model can be extended with a main effect of the shared environment, $\beta_4 C_{ij}$, plus an additional interaction effect, $\beta_5 C_{ij} M_{ij}$, but for our purpose here it suffices to discuss the model with additive genetic effects alone.

The two things that are different from the general ANOVA model is that variable A is unobserved, and that we have data on twin pairs, where phenotypes are correlated. In case of additive genetic effects, the genetic correlation equals 1 for monozygotic twin pairs and $\frac{1}{2}$ for dizygotic twin pairs, that is, for MZ twin pairs we have $Cov(A_{i1}, A_{i2}) = 1$ and for DZ twin pairs we have $Cov(A_{i1}, A_{i2}) = \frac{1}{2}$.

If we assume that M is a dichotomously scored covariate, the sufficient statistics for our model are the variance of phenotype P, the observed covariance in MZ twin pairs and the observed phenotypic covariance in DZ twin pairs, under the two conditions $M = 0$ and $M = 1$, where we assume that M has equal values for the two twins in each pair. This consists of 6 different statistics. Thus, the following set of equations needs to be solved for the regression coefficients:

M=0

$$Cov_{MZ}(P_{i1}, P_{i2}) = \beta_2^2 \quad (4)$$

$$Cov_{DZ}(P_{i1}, P_{i2}) = \frac{1}{2}\beta_2^2 \quad (5)$$

$$Var(P_{i1}) = Var(P_{i2}) = \beta_2^2 + e^2 \quad (6)$$

M=1

$$Cov_{MZ}(P_{i1}, P_{i2}) = (\beta_2 + \beta_3)^2 \quad (7)$$

$$Cov_{DZ}(P_{i1}, P_{i2}) = \frac{1}{2}(\beta_2 + \beta_3)^2 \quad (8)$$

$$Var(P_{i1}) = Var(P_{i2}) = (\beta_2 + \beta_3)^2 + e^2 \quad (9)$$

Since β_2 is only linked to observed statistics through a quadratic function, it is easily seen that the probability of the data under $M = 0$ given a value $\beta_2 = a$ is equal to the probability of the data given $\beta_2 = -a$.

Under $M = 1$, we immediately see that combinations of values for $(\beta_2 + \beta_3) = b + c$ are equally likely as $(\beta_2 + \beta_3) = -b - c$. Thus, any combination of values for β_2 and β_3 is equally likely as the combination of their negatives. Taking the negative of the value for β_2 does, as we saw, not affect the probability of the data under $M = 0$, so that for any data set, there are always two combinations that have the exact same likelihood.

This problem cannot be easily solved by constraining β_2 to be positive, $\beta_2 > 0$. This is because for any positive value of β_2, there are two values for β_3 that result in the same expected variances and covariances for $M = 1$, since using the square root formula for quadratic functions we get:

$$Var(P) = (\beta_2 + \beta_3)^2 = \beta_2^2 + 2\beta_2\beta_3 + \beta_3^2 + e^2 \quad (10)$$

$$\beta_3^2 + 2\beta_2\beta_3 + (\beta_2^2 + e^2 - Var(P)) = 0 \quad (11)$$

$$\beta_3 = -\beta_2 \pm \sqrt{\beta_2^2 - (\beta_2^2 + e^2 - Var(P))} = -\beta_2 \pm \sqrt{Var(P) - e^2} \quad (12)$$

$$Cov_{MZ}(P_1, P_2) = (\beta_2 + \beta_3)^2 = \beta_2^2 + 2\beta_2\beta_3 + \beta_3^2 \quad (13)$$

$$\beta_3^2 + 2\beta_2\beta_3 + (\beta_2^2 - Cov_{MZ}(P, P)) = 0 \quad (14)$$

$$\beta_3 = -\beta_2 \pm \sqrt{\beta_2^2 - (\beta_2^2 - Cov_{MZ}(P_1, P_2))} = -\beta_2 \pm \sqrt{Cov_{MZ}(P_1, P_2)} \quad (15)$$

Thus, there is no unique Maximum Likelihood solution for a given data set. The problem lies in the fact that the genotypic value is unobserved (so that there is not observed covariance between A and P), and that therefore all observed statistics are related only quadratically with the parameters that need to be estimated. The proof can be extended to ACE models and continuous measured covariates M in a similar manner.

Appendix D

Following JAGS script incorporates an ACE model with a 1 PL IRT model while modelling ACE×M (same moderator value for every twin pair)

```
#y_dz = Item responses of DZ twins (matrix)
#y_mz = Item responses of MZ twins (matrix)
#x_MZ = Values on moderator variable for all MZ twin pairs
#x_DZ = Values on moderator variable for all DZ twin pairs
#n_mz = Number of MZ twin pairs
#n_dz = Number of DZ twin pairs
#n_items = Number of phenotypic items administered
#b = Vector with item difficulty parameters, assumed known here

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

#This results in a matrix of n_mz (or, in case of y_dz, n_dz)
#rows and 2*n_items columns. e.g. y_mz[1,22] is the response
#of MZ twin 1 from family 1 to item 22 if n_items = 22

#JAGS uses precision parameters for the variance parameters.
#Therefore, after running the script, these precision parameters
#have to be inverted. For example:
#var_a <-  1/outputAnalysis$tau_a[,,1] with the rjags package

model{
##MZ twins
for (fam in 1:n_mz){
  c_mz[fam] ~ dnorm(mu + beta_1m * x_MZ[fam], tau_c_mz[fam])
  f_mz[fam] ~ dnorm(c_mz[fam], tau_a_mz[fam])

  tau_c_mz[fam] <- 1/exp(beta_0c + beta_1c * x_MZ[fam])
  tau_a_mz[fam] <- 1/exp(beta_0a + beta_1a * x_MZ[fam])
  tau_e_mz[fam] <- 1/exp(beta_0e + beta_1e * x_MZ[fam])

  pheno_mz[fam,1] ~ dnorm(f_mz[fam], tau_e_mz[fam])
  pheno_mz[fam,2] ~ dnorm(f_mz[fam], tau_e_mz[fam])

  #1pl model twin1
  for (k in 1:n_items){
    logit(p[fam,k]) <- pheno_mz[fam,1] - b[k]
    y_mz[fam,k] ~ dbern(p[fam,k])
  }
```

```
41
42    #1pl model twin2
43    for (k in (n_items+1):(2*n_items)){
44      logit(p[fam,k]) <- pheno_mz[fam,2] - b[k-n_items]
45      y_mz[fam,k] ~ dbern(p[fam,k])
46    }
47  } #end MZ twins
48
49  ##DZ twins
50  for (fam in 1:n_dz){
51    c_dz[fam] ~ dnorm(0,tau_c_dz[fam])
52    a1_dz[fam] ~ dnorm(0,2)
53    a2_dz[fam,1] ~ dnorm(a1_dz[fam], 2)
54    a2_dz[fam,2] ~ dnorm(a1_dz[fam], 2)
55
56    tau_c_dz[fam] <- 1/(exp(beta_0c + beta_1c * x_DZ[fam]))
57    var_a_dz[fam] <- exp(beta_0a + beta_1a * x_DZ[fam])
58    tau_e_dz[fam] <- 1/(exp(beta_0e + beta_1e * x_DZ[fam]))
59
60    a_dz_twin1[fam] <- a2_dz[fam,1] * sqrt(var_a_dz[fam])
61    a_dz_twin2[fam] <- a2_dz[fam,2] * sqrt(var_a_dz[fam])
62
63    pheno_dz[fam,1] ~ dnorm(mu + beta_1m * x_DZ[fam] +
64                            c_dz[fam] + a_dz_twin1[fam],
65                            tau_e_dz[fam])
66    pheno_dz[fam,2] ~ dnorm(mu + beta_1m * x_DZ[fam] +
67                            c_dz[fam] + a_dz_twin2[fam],
68                            tau_e_dz[fam])
69
70    #1pl model twin1
71    for (k in 1:n_items){
72      logit(p_dz[fam,k]) <- pheno_dz[fam,1] - b[k]
73      y_dz[fam,k] ~ dbern(p_dz[fam,k])
74    }
75
76    #1pl model twin2
77    for (k in (n_items+1):(2*n_items)){
78      logit(p_dz[fam,k]) <- pheno_dz[fam,2] - b[k-n_items]
79      y_dz[fam,k] ~ dbern(p_dz[fam,k])
80    }
81  } #end DZ twins
82
83  #Priors
84  mu ~ dnorm(0, .1)
85  beta_1a ~ dnorm(0, .1)
```

```
86  beta_1c ~ dnorm(0, .1)
87  beta_1e ~ dnorm(0, .1)
88  beta_1m ~ dnorm(0, .1)
89
90  beta_0a ~ dnorm(-1, .5)
91  beta_0c ~ dnorm(-1, .5)
92  beta_0e ~ dnorm(-1, .5)
93  }
```

Appendix E

Following JAGS script incorporates an ACE model with a 1 PL IRT model while modelling ACE×M (separate moderator values)


```
#y_dz = Item responses of DZ twins (matrix)
#y_mz = Item responses of MZ twins (matrix)
#n_mz = Number of MZ twin pairs
#n_dz = Number of DZ twin pairs
#n_items = Number of phenotypic items administered
#b = Vector with item difficulty parameters, assumed known here
#x_MZ = moderator variable values for every MZ twin (matrix)
#x_DZ = moderator variable values for every DZ twin  (matrix)

#Required structure of the y_mz/y_dz data matrix:
#y_dz[i,k] = kth datapoint from the ith DZ twin pair
#y_mz[i,k] = kth datapoint from the ith MZ twin pair
#This results in a matrix of n_mz (or, in case of y_dz, n_dz)
#rows and 2*n_items columns. e.g. y_mz[1,22] is the response
#of MZ twin 1 from family 1 to item 22 if n_items = 22

#Required structure of the x_MZ/x_DZ data matrix:
#x_MZ[i,1:2] = Vector with moderator variable values
#for the first (x_MZ[i,1]) and second (x_MZ[i,2])
#twin of every ith MZ twin pair.
#x_DZ[i,1:2] = Vector with moderator variable values
#for the first (x_MZ[i,1]) and second (x_MZ[i,2])
#twin of every ith MZ twin pair.

#JAGS uses precision parameters for the variance parameters.
#Therefore, after running the script, these precision parameters
#have to be inverted. For example:
#var_a <-  1/outputAnalysis$tau_a[,,1] with the rjags package

model{
##MZ twins
for (fam in 1:n_mz){
  c_mz[fam] ~ dnorm(0, 1)
  a_mz[fam] ~ dnorm(0, 1)

  tau_c_mz[fam,1] <- exp(beta_0c + beta_1c*x_MZ[fam,1])
  tau_c_mz[fam,2] <- exp(beta_0c + beta_1c*x_MZ[fam,2])

  tau_a_mz[fam,1] <- exp(beta_0a + beta_1a*x_MZ[fam,1])
  tau_a_mz[fam,2] <- exp(beta_0a + beta_1a*x_MZ[fam,2])
```

```
41
42    tau_e_mz[fam,1] <- 1/(exp(beta_0e + beta_1e*x_MZ[fam,1] ))
43    tau_e_mz[fam,2] <- 1/(exp(beta_0e + beta_1e*x_MZ[fam,2] ))
44
45    c_mz_twin1[fam] <- c_mz[fam] * sqrt(tau_c_mz[fam,1])
46    c_mz_twin2[fam] <- c_mz[fam] * sqrt(tau_c_mz[fam,2])
47
48    a_mz_twin1[fam] <- a_mz[fam] * sqrt(tau_a_mz[fam,1])
49    a_mz_twin2[fam] <- a_mz[fam] * sqrt(tau_a_mz[fam,2])
50
51    pheno_mz[fam,1] ~ dnorm(mu_mz + beta_1m_mz * x_MZ[fam,1] +
52                            beta_1mc_mz * x_MZ[fam,2] +
53                            c_mz_twin1[fam] + a_mz_twin1[fam],
54                            tau_e_mz[fam,1])
55    pheno_mz[fam,2] ~ dnorm(mu_mz + beta_1m_mz * x_MZ[fam,2] +
56                            beta_1mc_mz * x_MZ[fam,1] +
57                            c_mz_twin2[fam] + a_mz_twin2[fam],
58                            tau_e_mz[fam,2])
59
60    #### Measurement model for phenotypic item data:
61    for (k in 1:n_items){
62      logit(p[fam,k]) <- pheno_mz[fam,1] - b[k]
63      y_mz[fam,k] ~ dbern(p[fam,k])
64    }
65
66    for (k in (n_items+1):(2*n_items)){
67      logit(p[fam,k]) <- pheno_mz[fam,2] - b[k-n_items]
68      y_mz[fam,k] ~ dbern(p[fam,k])
69    }
70  } #end MZ twins
71
72  ##DZ twins
73  for (fam in 1:n_dz){
74    c_dz[fam] ~ dnorm(0,1)
75    a1_dz[fam] ~ dnorm(0,2)
76    a2_dz[fam,1] ~ dnorm(a1_dz[fam], 2)
77    a2_dz[fam,2] ~ dnorm(a1_dz[fam], 2)
78
79    tau_c_dz[fam,1] <- exp(beta_0c + beta_1c*x_DZ[fam,1])
80    tau_c_dz[fam,2] <- exp(beta_0c + beta_1c*x_DZ[fam,2])
81
82    tau_a_dz[fam,1] <- exp(beta_0a + beta_1a*x_DZ[fam,1])
83    tau_a_dz[fam,2] <- exp(beta_0a + beta_1a*x_DZ[fam,2])
84
85    tau_e_dz[fam,1] <- 1/(exp(beta_0e + beta_1e*x_DZ[fam,1] ))
```

```
86    tau_e_dz[fam,2] <- 1/(exp(beta_0e + beta_1e*x_DZ[fam,2] ))
87
88    c_dz_twin1[fam] <- c_dz[fam] * sqrt(tau_c_dz[fam,1])
89    c_dz_twin2[fam] <- c_dz[fam] * sqrt(tau_c_dz[fam,2])
90
91    a_dz_twin1[fam] <- a2_dz[fam,1] * sqrt(tau_a_dz[fam,1])
92    a_dz_twin2[fam] <- a2_dz[fam,2] * sqrt(tau_a_dz[fam,2])
93
94    pheno_dz[fam,1] ~ dnorm(mu_dz + beta_1m_dz * x_DZ[fam,1] +
95                            beta_1mc_dz * x_DZ[fam,2] +
96                            c_dz_twin1[fam] + a_dz_twin1[fam],
97                            tau_e_dz[fam,1])
98    pheno_dz[fam,2] ~ dnorm(mu_dz + beta_1m_dz * x_DZ[fam,2] +
99                            beta_1mc_dz * x_DZ[fam,1] +
100                           c_dz_twin2[fam] + a_dz_twin2[fam],
101                           tau_e_dz[fam,2])
102
103   #### Measurement model for phenotypic item data:
104   for (k in 1:n_items){
105     logit(p_dz[fam,k]) <- pheno_dz[fam,1] - b[k]
106     y_dz[fam,k] ~ dbern(p_dz[fam,k])
107   }
108
109   for (k in (n_items+1):(2*n_items)){
110     logit(p_dz[fam,k]) <- pheno_dz[fam,2] - b[k-n_items]
111     y_dz[fam,k] ~ dbern(p_dz[fam,k])
112   }
113
114 } #end DZ twins
115
116 #Priors
117 mu_mz ~ dnorm(0, .1)
118 mu_dz ~ dnorm(0, .1)
119
120 beta_1m_mz ~ dnorm(0, .1)
121 beta_1m_dz ~ dnorm(0, .1)
122 beta_1mc_mz ~ dnorm(0, .1)
123 beta_1mc_dz ~ dnorm(0, .1)
124
125 beta_1a ~ dnorm(0, .1)
126 beta_1c ~ dnorm(0, .1)
127 beta_1e ~ dnorm(0, .1)
128
129 beta_0a ~ dnorm(-1, .5)
130 beta_0c ~ dnorm(-1, .5)
```

```
131   beta_0e ~ dnorm(-1, .5)
132   }
```

Appendix F

Following R script can be used to apply the full information approach to handle missing covariate data, using the R package OpenMx (Boker et al., 2011)


```
#Install OpenMx
source('http://openmx.psyc.virginia.edu/getOpenMx.R')
install.packages("OpenMx") #Install OpenMx package

#Load package
library(OpenMx)

#OpenMx analysis
twin_ACE_cov <- mxModel("twinACE",
 #Matrices X,Y,Z to store a,c,e path coefficients
 mxMatrix(type="Full", nrow=1, ncol=1, free=TRUE,
          values=.6, label="a", name="X" ),
 mxMatrix(type="Full", nrow=1, ncol=1, free=TRUE,
          values=.6, label="c", name="Y" ),
 mxMatrix(type="Full", nrow=1, ncol=1, free=TRUE,
          values=.6, label="e", name="Z" ),
 mxMatrix(type="Full", nrow=1, ncol=(2+Nvar*2),
          free=TRUE, values= 0,
          label=c('meanfeno','meanfeno',
                  rep('mean',Nvar*2)), name="expMean"),
 mxMatrix(type ='Lower', nrow=Nvar , ncol=Nvar,
          values=0.5, free=TRUE, name="CholCovW"),
 mxMatrix(type ='Lower', nrow=Nvar , ncol=Nvar,
          values=0.5, free=TRUE, name="CholCovB"),

 mxAlgebra(expression=CholCovW %*% t(CholCovW),
           name="CovW" ),
 mxAlgebra(expression=CholCovB %*% t(CholCovB),
           name="CovB" ),
 mxAlgebra(expression=CovB + CovW,
           name="CovWplusB"),

 #Matrices A, C,E + compute variance components
 mxAlgebra(expression=X %*% t(X), name="A"),
 mxAlgebra(expression=Y %*% t(Y), name="C"),
 mxAlgebra(expression=Z %*% t(Z), name="E"),

 #Declare a vector for the regression parameters
 mxMatrix(type="Full", nrow=Nvar, ncol=1, free=TRUE,
```

```
40              values= 0,
41              label=c("beta1","beta2","beta3",
42                      "beta4","beta5"),
43              name="beta"),
44
45    #Algebra for expected variance/covariance matrix
46    #in MZ twins
47    mxAlgebra(
48    expression=rbind(cbind(
49                           A+C+E+t(beta)%*%CovWplusB%*%beta,
50                           A+C+t(beta)%*%CovB%*%beta,
51                           t(beta)%*%CovWplusB,
52                           t(beta)%*% CovB),
53                         cbind(
54                           A+C+t(beta)%*%CovB%*%beta,
55                           A+C+E+t(beta)%*%CovWplusB%*%beta,
56                           t(beta)%*%CovB,
57                           t(beta)%*%CovWplusB),
58                         cbind(
59                           CovWplusB%*%beta,
60                           CovB%*%beta,
61                           CovWplusB, CovB),
62                         cbind(
63                           CovB%*%beta,
64                           CovWplusB%*%beta,
65                           CovB, CovWplusB)),
66                         name="expCovMZ"),
67
68    #Algebra for expected variance/covariance matrix
69    #in DZ twins
70    mxAlgebra(expression=rbind(
71                               cbind(
72                                    A+C+E+t(beta)%*%CovWplusB%*%beta,
73                                    0.5%x%A+C+t(beta)%*%CovB%*%beta,
74                                    t(beta)%*%CovWplusB,
75                                    t(beta)%*%CovB),
76                               cbind(
77                                    0.5%x%A+C+t(beta)%*%CovB%*%beta,
78                                    A+C+E+ t(beta)%*%CovWplusB%*%beta,
79                                    t(beta)%*%CovB,
80                                    t(beta)%*%CovWplusB),
81                               cbind(
82                                    CovWplusB%*%beta,
83                                    CovB%*%beta,
84                                    CovWplusB, CovB),
```

```
85                                  cbind(
86                                      CovB%*%beta,
87                                      CovWplusB%*%beta,
88                                      CovB, CovWplusB)),
89                                  name="expCovDZ"),
90
91    mxModel("MZ",
92    mxData( observed=mzData, type="raw" ),
93
94    #Algebra for making the means a function
95    #of the definition variables
96    mxFIMLObjective(covariance="twinACE.expCovMZ",
97                        means="twinACE.expMean",
98                        dimnames=names(mzData))),
99
100   mxModel("DZ",
101       mxData( observed=dzData, type="raw"),
102       mxFIMLObjective(covariance="twinACE.expCovDZ",
103                           means="twinACE.expMean",
104                           dimnames=names(dzData))),
105       mxAlgebra(expression=MZ.objective + DZ.objective,
106                   name="twin" ),
107       mxAlgebraObjective("twin")
108   )
```

Appendix G

Following R script can be used to apply the full information approach to handle missing covariate data, using a Bayesian approach. The script was used in the application study of Chapter 6.


```
## Following script was used in the application
## study of Chapter 6 and consists of a Bayesian
## estimation of the full information approach

# X_mz_twin1[i, k] is a matrix of the values on k
# covariates for twin 1 of family i (e.g. family 1)
# X_mz_twin2[i,k] is a matrix of the values on k
# covariates for twin 2 of the samee family i.
# The same logic applies to X_dz_twin1 & X_dz_twin2.

# N = number of covariates (without intercept)
# n_mz = total number of MZ twin pairs
# n_dz = total number of DZ twin pairs
# y_dz = Matrix of phenotypic variable, DZ twins
# y_mz = Matrix of phenotypic variable, MZ twins

# Required structure of the y_dz matrix:
# y_dz[i,1] = Answer of twin 1 of DZ family i
# y_dz[i,2] = Answer of twin 2 of DZ family i
# The same logic applies to the y_mz matrix.

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

# The prior values for mu_b, tau_b, omega_tau_w
# and omega_tau_b have to be given as input to the
# JAGS program. For a relatively flat prior you
# can choose e.g. (R syntax):
# mu_b = rep(0, N); tau_b = diag(1, N)
# omega_tau_w = diag(1, N); omega_tau_b = diag(1, N)

model{
##MZ twins
for (fam in 1:n_mz){ #for each MZ family:
        c_mz[fam] ~ dnorm(mu, tau_c)
```

```
40      f_mz[fam] ~ dnorm(c_mz[fam], tau_a)
41
42          #Response variables:
43      y_mz[fam,1] ~ dnorm(f_mz[fam] +                                  151
44                              inprod(X_mz_twin1[fam,], b[1:N]),
45                              tau_e)
46      y_mz[fam,2] ~ dnorm(f_mz[fam] +
47                              inprod(X_mz_twin2[fam,], b[1:N]),
48                              tau_e)
49      }
50
51  ##DZ twins
52  for (fam in 1:n_dz){ #for each DZ family:
53          c_dz[fam] ~ dnorm(mu, tau_c)
54          f1_dz[fam] ~ dnorm(c_dz[fam], doubletau_a)
55          f2_dz[fam,1] ~ dnorm(f1_dz[fam], doubletau_a)
56      f2_dz[fam,2] ~ dnorm(f1_dz[fam], doubletau_a)
57
58      #Response variables:
59      y_dz[fam,1] ~ dnorm(f2_dz[fam,1] +
60                              inprod(X_dz_twin1[fam,], b[1:N]),
61                              tau_e)
62      y_dz[fam,2] ~ dnorm(f2_dz[fam,2] +
63                              inprod(X_dz_twin2[fam,], b[1:N]),
64                              tau_e)
65  }
66
67  #For DZ twins (share half of their genetic material):
68  doubletau_a <- 2*tau_a
69
70  #Priors:
71  tau_a ~ dgamma(1,1)
72  tau_e ~ dgamma(1,1)
73  tau_c ~ dgamma(1,1)
74
75  #Value of population mean was set to zero as the
76  #phenotypic variable was standardized to have
77  #a mean equal to zero.
78  #If this is  not the case, you can use:
79  #mu ~ dnorm(0, .1)
80  mu <- 0
81
82  #Prior for regression coefficients:
83  #Multivariate normal
84  b[1:N] ~ dmnorm(mu_b[1:N], tau_b[,])
```

```
85
86  #Input needed for mu_b & tau_b
87  #For example in R: mu_b = rep(0,N), tau_b = diag(1, N)
88
89  ##Latent variable psi for MZ and DZ twins
90  for(fam in 1:n_mz){
91          psi_mz[fam, 1:N] ~ dmnorm(gamma_cov[1:N],tau_b_mz)
92   }
93
94  for(fam in 1:n_dz){
95          psi_dz[fam, 1:N] ~ dmnorm(gamma_cov[1:N],tau_b_dz)
96  }
97
98  #For the continuous covariate (school size)
99  #that was standardized the expected value for the
100 #average is set to 0.
101 #For the rest, we use a normal distribution as prior
102 gamma_cov[2] <- 0
103 gamma_cov[1] ~ dnorm(0, .1)
104
105 for (i in 3:N){
106         gamma_cov[i] ~ dnorm(0, .1)
107 }
108
109 #As a prior distribution for tau b, we use a
110 #wishart distribution.
111 tau_b_mz[1:N, 1:N] ~ dwish(omega_tau_b[,], N)
112 tau_b_dz[1:N, 1:N] ~ dwish(omega_tau_b[,], N)
113 #Input needed for omega_tau_b.
114 #For example in R: omega_tau_b = diag(1, N)
115
116 #Final prior distribution for the covariates:
117 #(Equation 11 and 12 in the paper)
118 for (fam in 1:n_mz){
119         exp_mz_twin1[fam,1:N] ~ dmnorm(psi_mz[fam,],
120                                        tau_w_mz)
121         exp_mz_twin2[fam,1:N] ~ dmnorm(psi_mz[fam,],
122                                        tau_w_mz)
123 }
124
125 for (fam in 1:n_dz){
126         exp_dz_twin1[fam,1:N] ~ dmnorm(psi_dz[fam,],
127                                        tau_w_dz)
128     exp_dz_twin2[fam,1:N] ~ dmnorm(psi_dz[fam,],
129                                        tau_w_dz)
```

```
}

#As a prior distribution for tau w, we use a
#wishart distribution.
tau_w_mz[1:N, 1:N] ~ dwish(omega_tau_w[,], N)
tau_w_dz[1:N, 1:N] ~ dwish(omega_tau_w[,], N)
#Input needed for omega_tau_w.
#For example in R: omega_tau_w = diag(1, N)

#For the dummy variables that are coded as 0
#and 1, we use exp_mz_twin1, exp_mz_twin2,
#exp_dz_twin1 and exp_dz_twin2 as liabilites
#and use a bernoulli distribution as prior:

#In the application, the columns for the dummy
#variables are 1 and 3:N

#To identify the model, the threshold t is set
#to 0:
t <- 0

#Dummy variable sex (first column):
for (fam in 1:n_mz){
        V_mz_twin1[fam,1]<-step(exp_mz_twin1[fam,1] - t)
        V_mz_twin2[fam,1]<-step(exp_mz_twin2[fam,1] - t)
        X_mz_twin1[fam,1] ~ dbern(ifelse(
                V_mz_twin1[fam,1]==1,0.999,0.001))
        X_mz_twin2[fam,1] ~ dbern(ifelse(
                V_mz_twin2[fam,1]==1,0.999,0.001))
}

for (fam in 1:n_dz){
        V_dz_twin1[fam,1]<-step(exp_dz_twin1[fam,1] - t)
        V_dz_twin2[fam,1]<-step(exp_dz_twin2[fam,1] - t)
        X_dz_twin1[fam,1] ~ dbern(ifelse(
                V_dz_twin1[fam,1]==1,0.999,0.001))
        X_dz_twin2[fam,1] ~ dbern(ifelse(
                V_dz_twin2[fam,1]==1,0.999,0.001))
        }

#Rest of the dummy variables (3:N):
for(fam in 1:n_mz){
        for (covariate in 3:N){
          V_mz_twin1[fam,covariate]<-step(
                exp_mz_twin1[fam,covariate] - t)
```

```
175             V_mz_twin2[fam,covariate]<-step(
176               exp_mz_twin2[fam,covariate] - t)
177             X_mz_twin1[fam,covariate] ~ dbern(ifelse(
178               V_mz_twin1[fam,covariate]==1,0.999,0.001))
179             X_mz_twin2[fam,covariate] ~ dbern(ifelse(
180               V_mz_twin2[fam,covariate]==1,0.999,0.001))
181           }
182 }
183
184 for(fam in 1:n_dz){
185           for (covariate in 3:N){
186             V_dz_twin1[fam,covariate]<-step(
187               exp_dz_twin1[fam,covariate] - t)
188             V_dz_twin2[fam,covariate]<-step(
189               exp_dz_twin2[fam,covariate] - t)
190             X_dz_twin1[fam,covariate] ~ dbern(ifelse(
191               V_dz_twin1[fam,covariate]==1,0.999,0.001))
192             X_dz_twin2[fam,covariate] ~ dbern(ifelse(
193               V_dz_twin2[fam,covariate]==1,0.999,0.001))
194           }
195 }
196
197 #As JAGS does not allow to define anything other
198 #than probability distributions for observed data,
199 #we place a normally distributed prior distribution
200 #on the continous covariates. We use a very small
201 #residual variance of 0.01:
202
203 #In the application, the columns of the continous
204 #covariate is 2.
205 for (fam in 1:n_mz){
206    #for total number of students:
207    X_mz_twin1[fam,2] ~ dnorm(exp_mz_twin1[fam,2],100)
208    X_mz_twin2[fam,2] ~ dnorm(exp_mz_twin2[fam,2],100)
209  }
210
211 for (fam in 1:n_dz){
212    #for total number of students:
213    X_dz_twin1[fam,2] ~ dnorm(exp_dz_twin1[fam,2],100)
214    X_dz_twin2[fam,2] ~ dnorm(exp_dz_twin2[fam,2],100)
215  }
216 }
```

This dissertation discusses a number of psychometric issues that require special attention in the analysis of genetically-informative data, such as data on twins. These include heterogeneous measurement error, scaling and scale transformations, and harmonization of phenotypes. It is shown how ignoring these issues can result in spurious findings of genotype by environment interaction. Multilevel item response theory models are proposed that can help solve these problems.