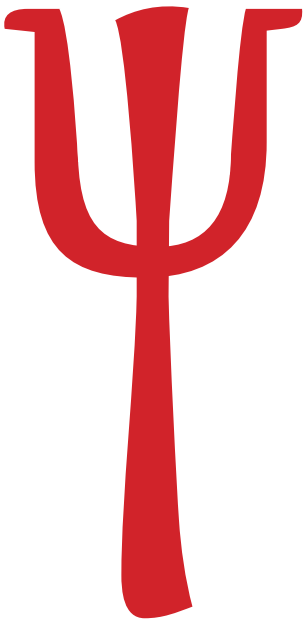

NETHERLANDS JOURNAL OF PSYCHOLOGY

VOLUME 67, NUMBER 3, DECEMBER 2012

SPECIAL ISSUE – PART 1

Non-standard structural equation modelling

GUEST EDITORS

Suzanne Jak

Annemarie Zand Scholten

Frans J. Oort

Volume 67/number 3/December 2012
Netherlands Journal of Psychology

Editorial Policy

The *Netherlands Journal of Psychology* publishes original articles of high quality, including empirical articles, review essays on selected books, theoretical and methodolological papers in any area of psychology, as well as ongoing commentaries and discussion, and short reports on the validation of psychodiagnostic instruments and related methodological tools.

The *Journal's* focus is on empirical and conceptual studies that contribute to the theoretical explanation of human behaviour and experience. Manuscripts can deal with human development, social processes and the social perspective of human behaviour, psychopathology, forensic psychology and psychiatry neuroscience, psychophysiology, philosophy of mind, the computational approach, emotion, cognition (including attention, perception, and memory), decision-making, human performance, educational psychology, health, selection and assessment and human behaviour in organisations, selection and assessment. Any other manuscripts that may be of importance to those involved with psychological research are welcomed. Manuscripts should be written to the professional and academic readership.

It is the *Netherlands Journal of Psychology's* policy to publish about recent developments in psychology and related fields. It is also the editors' policy to regularly focus on recent developments in psychology in the Netherlands. In particular, authors may be invited to contribute to special issues.

The *Netherlands Journal of Psychology* will help to focus the interest of psychologists, and others professionally interested in the field, on information about developments in a wide area of specialist research activities and results of empirical and theoretical work in the field of psychology.

The *Netherlands Journal of Psychology* offers specialists the opportunity to publish their findings to a broad audience that is scientifically interested in psychology.

The *Netherlands Journal of Psychology* aims at offering a continuous forum for intellectual discussion and transfer of knowledge and information about developments in psychology.

Categories of articles to be published and manuscripts to be welcomed are:

Review articles

Manuscripts should be limited to a maximum of 8000 words. Authors planning contributions that exceed this maximum must contact the editor prior to submitting their manuscript.

Original reports of empirical research

Manuscripts should be limited to a maximum of 5000 words. Authors planning contributions that exceed this maximum must contact the editor prior to submitting their manuscript.

Thematic issues

Authors with plans for planning a thematic issue are encouraged to contact the editor in an early stage. Consultation will be necessary about the number of manuscripts, number of words per manuscript, review procedure, covering of the field, etc.

Developments in the field (short)

Manuscripts should be limited to a maximum of 1000 words. Authors planning contributions that exceed this maximum must contact the editor prior to submitting their manuscript.

Book review essays (short)

Manuscripts with a maximum of 1000 words need no preliminary contact with the editor. Authors planning larger contributions must contact the editor.

Letters to the editor

Manuscripts should be limited to a maximum of 500 words. Authors planning contributions that exceed this maximum must contact the editor prior to submitting their manuscript.

Technical notes

Manuscripts should be limited to a maximum of 2000 words. Authors planning contributions that exceed this maximum must contact the editor prior to submitting their manuscript. Authors reporting (in English) on, e.g. validation of tests and other instruments in the Dutch language or in the Netherlands are encouraged to report in *The Netherlands Journal of Psychology*. The editors want to inform non-Dutch researchers of developments and of the use of instruments in the Dutch language.

Editor's address

Professor René van Hezewijk
Faculty of Psychology, Open University of the Netherlands, PO Box 2960, 6401 DL Heerlen, the Netherlands
tel + 31 45 576 23 99
All correspondence by e-mail: rene.vanhezewijk@ou.nl

Publisher: Nederlands Instituut van Psychologen, PO Box 2085, 3500 GB Utrecht, the Netherlands

Subscription rates

Personal rate: € 125.–
Institutional rate: € 226.–
Student rate: € 62.50
All prices are per calendar year and include value added tax (VAT)
Price per issue: € 31.95 (incl. VAT)

Subscription administration

Performis Media, PO Box 2396, 5202 CJ 's Hertogenbosch, the Netherlands, tel: + 31 73 689 58 89. For information and orders, please consult www.performis.nl

Change of address

Please notify any changes in addressee and/or address via www.performis.nl.

Payment

Please use the payment/accept giro form if possible as this simplifies the administrative process.

Advertisements

Performis Media, PO Box 2396, 5202 CJ 's Hertogenbosch, the Netherlands, tel: + 31 73 689 58 89.
The *Netherlands Journal of Psychology* is published four times a year.

ISSN 1872-552x

Contents

VOLUME 67, NUMBER 3, DECEMBER 2012

Non-standard structural equation modelling

The well-known technique of structural equation modelling (SEM) has roots in two very different techniques developed in two very different fields. Path analysis with its graphical representations of effects and effect decomposition comes from genetics research, where Sewall Wright (1920, 1921) proposed a method to predict heritability of the piebald pattern of guinea-pigs. Factor analysis with its latent variables is even older, with an early paper by Spearman in 1904, and was developed in research on intelligence, to explain correlations between various ability tests (Spearman, 1928). Karl Jöreskog (1973) coined the name LISREL (LInear Structural RELations) for the framework that integrates the techniques of path analysis and factor analysis, as well as for the computer program that made the technique available to researchers in psychology and related fields. They embraced the technique, now generally referred to as SEM, for its sophistication of the underlying theory, the suitability to address substantive hypotheses, and the availability and simplicity of the related software. SEM is constantly developing, as researchers are extending SEM to non-standard situations, so that it can be used with both observed and latent variables, both continuous and discrete variables, with normal and non-normal distributions, to model linear and non-linear relationships. SEM estimation methods are extended for applications of, for example, interaction effects, non-linear growth, multilevel data, meta-analysis, and so on.

This issue of the *Netherlands Journal of Psychology* is Part 1 of two special issues that give account of ten papers that illustrate non-standard applications of SEM, as presented at the 2012 Meeting of the Working Group SEM (Amsterdam, 22-23 March 2012). Since the foundation of the Working Group Structural Equation Modelling in 1986, advanced structural equation modelling has been discussed in annual meetings at various locations in Germany, the Netherlands, and Switzerland. The presentation and discussion of methodological problems and developments in structural equation modelling are the main objective of the Working Group.

The first article in the present issue, by Molenaar and Dolan (2012), introduces the traditional latent trait models commonly used in psychology to make inferences about constructs such as extraversion. The authors show that these traditional models are not in line with some substantive hypotheses from the psychological literature and they present specific extensions of the traditional models to account for these violations. The other four articles in this issue involve longitudinal structural equation models. Verdam, Oort, Visser, and Sprangers (2012) use SEM to compare two approaches to the detection of true change and response shift, and to test the assumptions that underlie the use of retrospective pretests. They illustrate the methods with quality of life data from cancer patients undergoing invasive surgery. Delsing and Oud (2012) propose an alternative to two existing models (cross-lagged panel models and latent growth curve models) to investigate development over time, in which the time variable is considered continuous instead of discrete. They use data of adolescents' externalising and internalising problem behaviour to demonstrate how continuous time analysis of the cross-lagged panel model does not have some of the problems of the other two models. The article by De Kort, Dolan, and Boomsma (2012) deals with genetics research in longitudinal twin designs. In the classical twin study, genetic and environmental influences on a phenotype are usually assumed to be independent. De Kort et al. explain why this may be an unrealistic assumption and they propose a way of incorporating covariance between genetic and environmental factors into the classical twin model. In the final article in this issue, King-Kallimanis, Oort, Tishelman, and Sprangers (2012) compare the results of two procedures that can be used in the specification search that is part of testing measurement invariance in longitudinal data with SEM: One procedure involves modification indices and expected parameter change, while the other procedure is based on global tests and observed parameter change.

The second special issue with non-standard applications of SEM, presented at the 2012 meeting, will cover meta-analytical SEM, extensions of growth curve models, quadratic effects, exploratory factor analysis of multilevel discrete data, and ROC analysis. We think that these ten non-standard SEM papers, five papers in each of two issues of the Netherlands Journal of Psychology, together give a fine impression, not only of the state of the art of advanced SEM, but also of the variety of substantive research questions that can benefit from a SEM approach.

Suzanne Jak
Annemarie Zand Scholten
Frans J. Oort
University of Amsterdam, Amsterdam, the Netherlands
E-mail: S.Jak@uva.nl

References

De Kort J. M., Dolan, C. V., & Boomsma, D. I. (2012). Accommodation of genotype-environment covariance in a longitudinal twin design. *Netherlands Journal of Psychology 2012; 67,* 81-90.

Delsing, M. J. M. H., & Oud, J. H. L. (2012). Causal directions between adolescents' externalising and internalising problems: A continuous-time analysis. *Netherlands Journal of Psychology 2012; 67*, 68-80.

Jöreskog, K. G. (1973). A general method for estimating a linear structural equation system. In: A. S. Goldberger, & O. D. Duncan, (eds.). *Structural Equation Models in the Social Sciences*. New York: Academic Press.

King-Kallimanis B. L., Oort, F. J., Tishelman, C., & Sprangers, M. A. G. (2012). Comparison of procedures used to test measurement invariance in longitudinal factor analysis. *Netherlands Journal of Psychology 2012*; 67, 91-100.

Molenaar D., & Dolan, C. V. (2012). Substantively motivated extensions of the traditional latent trait model. *Netherlands Journal of Psychology 2012*; 67, 48-57.

Spearman, C. (1904). General intelligence, objectively determined and measured. *American Journal of Psychology, 15,* 201-293.

Spearman, C. (1928). The sub-structure of the mind. *British Journal of Psychology, 18*, 249-261.

Verdam, M. G. E., Oort, F. J., Visser , M. R. M., & Sprangers, M. A. G. (2012) Response shift detection through then-test and structural equation modelling: Decomposing observed change and testing tacit assumptions. *Netherlands Journal of Psychology 2012; 67*, 58-67.

Wright, S. (1920). The relative importance of heredity and environment in determining the piebald pattern of guinea–pigs. *Proceedings of the National Academy of Sciences, 6,* 320-332.

Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research, 20*, 557-585.

SUZANNE JAK
PhD candidate at the Methods and Statistics Department of the Research Institute of Child Development and Education, University of Amsterdam. Her research interests include structural equation modelling, measurement bias, multilevel structural equation modelling and meta-analytical structural equation modelling.

ANNEMARIE ZAND SCHOLTEN
Assistant professor with the Methods and Statistics research group in the Department of Child Development and Educational Sciences at the University of Amsterdam. Her research interests include the quantitative structure of psychological measurement and the meaningfulness and stability of inference based on statistical and modelling techniques under different types of measurement level assumption violations.

FRANS J. OORT
Full professor of Methods and Statistics of Educational Research and Director of the Research Institute of Child Development and Education, University of Amsterdam. His research interests include measurement and measurement bias, and statistical modelling, especially non-standard applications of structural equation modelling

Substantively motivated extensions of the traditional latent trait model

In this paper we advocate the use of latent trait models to make inferences about psychological construct as measured by psychological tests and questionnaires. Latent trait models have the advantage that measurement error is isolated, that items are weighted according to how well they measure the construct, and that explicit tests concerning the underlying construct are feasible. However, latent trait models come with the requirement of distributional assumptions concerning the item scores. We show in this paper that these assumptions may conflict with specific psychological phenomena. We discuss a substantively motivated latent trait model that can accommodate these phenomena.

Where: Netherlands Journal of Psychology, Volume 67, 48-57

Received: 27 June 2012; Accepted: 20 November 2012

Keywords: Latent trait model; Intelligence; Personality; Gene-by-environment interactions; Censoring, Bad scaling

Authors: Dylan Molenaar and Conor V. Dolan

Department of Psychology,
University of Amsterdam

Correspondence to:
Dylan Molenaar,
Psychological Methods,
Department of Psychology,
University of Amsterdam,
Weesperplein 4,
1018 XA, Amsterdam,
the Netherlands
E-mail: D.Molenaar@uva.nl.

In psychology, the dominant approach to measuring psychological constructs is by means of tests and questionnaires. Psychological tests are administered to measure cognitive abilities such as working memory, arithmetic ability, and general knowledge, while psychological questionnaires are administered to measure personality traits or mood and affect. Tests and questionnaires differ importantly in the nature of their items. A typical test consists of a number of tasks that need to be completed. For instance, in the subtest Picture Completion of the Wechsler Adult Intelligence Scale (Wechsler, 1997), a set of pictures displaying an event (e.g., a carpenter building a house) need to be placed in the correct chronological order. Or, in the 'Intelligenz Struktur Test' (IST; Amthauer, Brocke, Liepmann, & Beauducel, 2001), the subtest Arithmetic involves traditional arithmetic problems that need to be solved. On the contrary, a questionnaire item typically involves a statement about one's behaviour, attitudes, and/or feelings. The respondent indicates on a fixed scale the extent to which the statement applies to him or her. For instance, the Bermond-Vorst Alexitymia Questionnaire (Vorst & Bermond, 2001) includes items such as 'If I see someone cry, I start feeling sad' and 'If something totally unexpected happens, I stay calm and unaffected'. Or, in the Positive Affect and Negative Affect scale (Guadagnoli & Mor, 1989) the items are words that describe a particular affect (e.g., desperate, or happy) to which the respondents need to report how much they experienced the affect during past week.

Administration of a test or questionnaire to a sample of respondents results in observed item scores, which are regarded as measures of the underlying psychological construct (Borsboom, 2008). This means that we have multiple measures of the same construct available, as we have multiple items. However, these multiple measures should be combined into a single score to enable inferences about the construct. In practice, researchers often rely on taking the sum or average of the item scores and use this sum score as the construct score (Borsboom, 2006). However, this approach is suboptimal as 1) all items are weighted equally, while some items can be a better measure of the construct than others; 2) all items are assumed to be a perfectly reliable measure of the construct as measurement error is not extracted from the item;

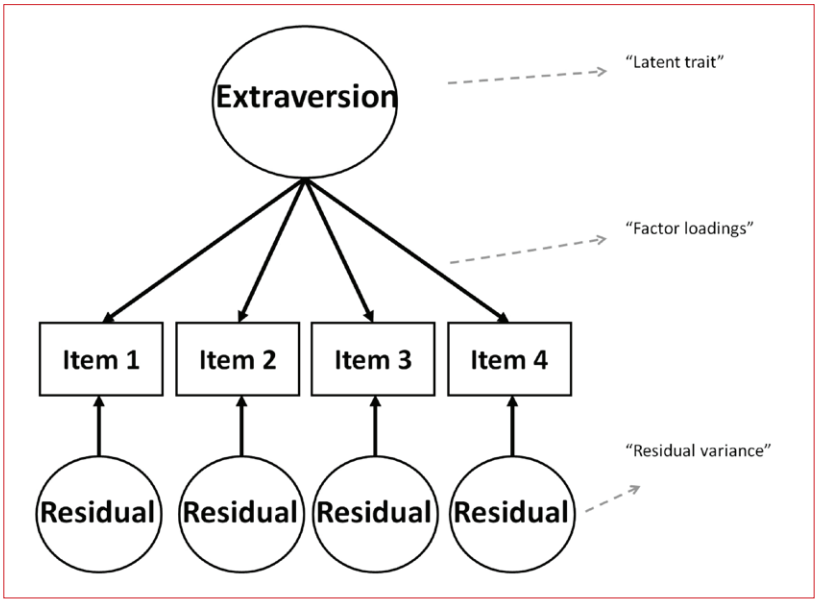

Figure 1 Example of a latent trait model for the psychological construct 'extraversion'

3) it is not tested whether it is justified to add the item scores, i.e., it is not checked whether the test or questionnaire truly measures a single construct (and not two or three); 4) the sum score is instrument-dependent, i.e., it cannot easily be compared with other sum scores obtained using a different test or questionnaire that measures the same construct; 5) the sum score is sample-dependent, e.g., a respondent can have the highest extraversion score in one sample, while in another sample, the same respondent scores relatively low.

A more rigorous alternative to obtain a score on the psychological construct is to use *latent trait models* (Mellenbergh, 1994). In these statistical models, the psychological construct is represented by an unobserved or latent trait and the items of the test or questionnaire are indicators of this trait. By doing so, one can easily estimate the scores of the respondents on the latent trait using software such as Mplus (Muthén & Muthén, 2007), LISREL (Jöreskog, & Sörbom , 1993), Amos (Arbuckle, 1997), and OpenMX (Boker et al., 2010). This latent trait score can then be regarded as the psychological construct score. Advantages are that 1) items are weighted according to how well they measure the construct; 2) measurement error is taken into account in the item scores; 3) it can be tested whether a single construct is measured by the test or questionnaire; 4) the latent trait does not in principle depend on the exact items that are used; and 5) the latent trait is not in principle sample-dependent. Despite these advantages, there are a number of challenges to the latent trait model: 1) large sample sizes are necessary to enable estimation of the latent trait; and 2) latent trait models can get so complex that it is numerically difficult to apply the appropriate model (e.g., for data from multidimensional intelligence tests); and 3) commonly, a multivariate normal distribution for the observed data needs to be imposed, which results in assumptions that are not necessarily psychologically meaningful[2].

The first challenge, the large sample size requirement, is an important reason why most researchers prefer working with ANOVA-based methods that require only a handful of respondents (Borsboom, 2006). Indeed, the large sample size requirement is a drawback; however, due to the upcoming facilities of online data archiving, more and more data are becoming available, hopefully benefitting the use of latent trait models. In addition, due to advances in computer technology, the numerical challenge to the latent trait model is also becoming progressively less problematic. This paper focuses on the third disadvantage concerning the assumptions in the latent trait that are not necessarily psychologically meaningful. Specifically, in this paper we show how the traditional latent trait model includes assumptions that are not necessarily in line with specific hypotheses from the psychological literature. We argue that, to test these hypotheses, these assumptions need to be relaxed. The outline is as follows: First we conceptually present the traditional latent trait model including its parameters. Then, we present three substantive hypotheses from the literature that predict specific violations of the assumptions of the latent trait model. These are: ability differentiation, schematicity, and gene-by-environment interactions. Next, we show conceptually how these hypotheses can be included in the traditional latent trait model to arrive at a substantively motivated latent trait model. Finally, we discuss some challenging aspects of the present approach.

Latent trait models in psychology

A common way to visualise a latent trait model is illustrated in Figure 1. In the figure, the latent trait is the psychological construct 'extraversion' and it is measured by four items. These items could include for instance:

At parties, I like to talk with people I don't know, and

Talking to people gives me energy.

[1] A latent trait can also be referred to as a *latent variable*.

[2] Strictly, multivariate normality is imposed on data that are (approximately) continuous (see below). For ordinal data, it is assumed that the data arose from categorisation of an underlying multivariate normally distributed variable. Thus, in case of ordinal data, the normal distribution for the data is also imposed but in a slightly different manner.

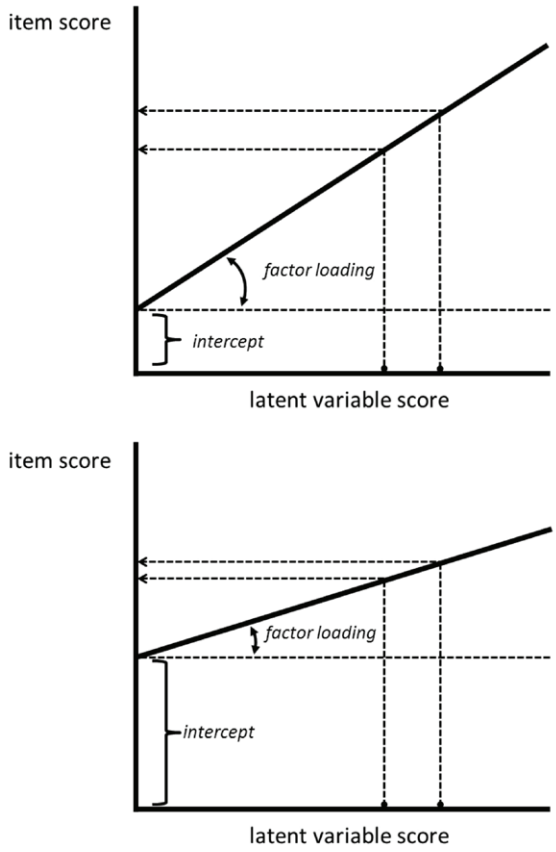

Figure 2 Illustration of how factor loadings account for the degree to which variability in the latent trait is captured by the item scores (in case of –approximately– continuous item scores)

Table 1 Overview of the different types of latent trait models

Item scale	Latent trait scale	Caregivers
Categorical	*Categorical* Latent Class Model (Goodman, 1974; McCutcheon, 1987)	*Continuous* Item Response Model (Rasch, 1960; Birnbaum, 1968; Lord, 1952)
Continuous	Latent Profile Model (Gibson, 1959; Lazarsfeld & Henry, 1968)	Linear Factor Model (Spearman, 1904; Lawley & Maxwell, 1963; Mellenbergh, 1994)

The latent trait is visualised by a circle to indicate that the trait is unobserved. The items are visualised as squares to indicate that these scores are observed. From the latent extraversion trait, arrows go down to each of the items indicating that the position on the latent trait of a given respondent will result in a specific expected score on each of the items.

The degree to which variation in the latent trait is captured by the item is quantified by a parameter called *factor loadings* (Figure 2). The higher a factor loading, the better an item is at measuring the latent trait. In addition to the extraversion latent trait that is common to all items, each item is associated with a unique latent trait. These are the residuals that contain the measurement error of the item.[3] The strength of the influence of the residuals on the items is quantified by the parameter called residual variances. The higher a *residual variance*, the more 'noisily' the item is measuring the latent trait. In addition to the factor loadings and the residual variances, the items are characterised by an *intercept* (not depicted), which reflects the mean of the item when applied to a single group of respondents, or the baseline level when applied to multi groups (e.g., males and females, or experimental conditions). Furthermore, the amount in which respondents differ on the latent trait is quantified by the *factor variance* (not depicted), which is simply the variance of the latent trait scores. In case of the extraversion example, a large factor variance suggests that there are large individual differences on extraversion in the population. When the factor variance equals 0, the population is homogenous with respect to extraversion, i.e., all subjects in the sample have the same level of extraversion.

Different kinds of latent trait models

The model in Figure 1 is general in that it can handle different kinds of data. By specifying the structure of the item and the latent trait scales, different latent trait models arise that go by different names in the literature, see Table 1. As can be seen in the table, the nature of the items can be categorical or continuous. Categorical item scales include items with either 2 or more unordered categories, e.g., 'male/female', or items with 2 to 6 ordered categories, e.g., Likert answer scales or item scores that are scored correct (1) and false (0). Continuous item scales include items that require responses to a continuous line segment (see Samejima, 1973; Mellenbergh, 2012) or response times. In addition, ordered categorical items with at least 7 categories can also be considered continuous (see Dolan, 1994).

Like the scale of items, latent trait scales can also be continuous or categorical. Examples of continuous latent traits in psychology include working memory, depression, verbal comprehension, and neuroticism.

[3] Technically, the residuals contain both a random component (measurement error) and a systematic component (due to unmodelled latent variables; Bollen, 1989). Here we assume that the model in Figure 1 is the true model, i.e., there is no systematic component in the residuals.

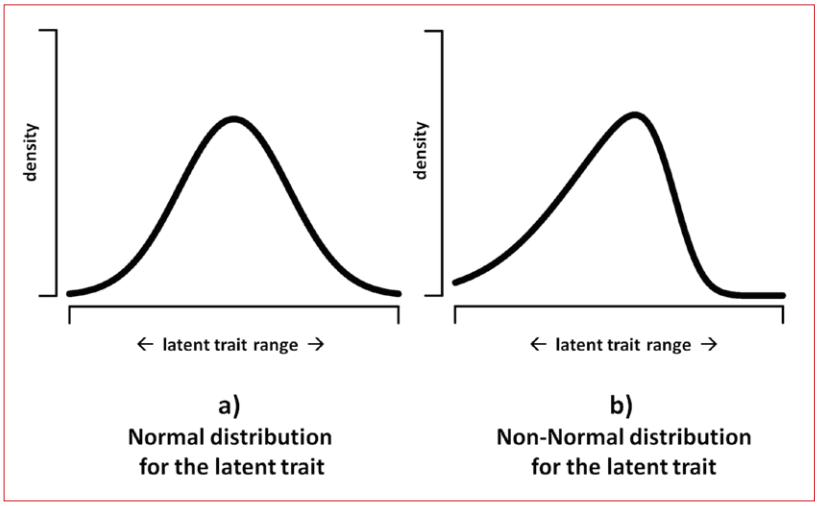

Figure 3. A) The assumption of a normal distribution for the latent trait; B) a violation of this assumption

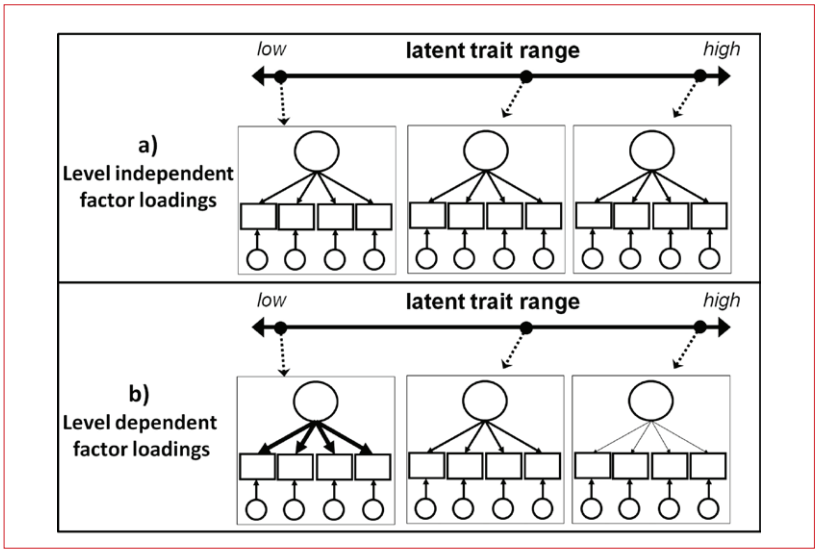

Figure 4. A) The assumption of level independent factor loadings; B) a violation of this assumption, the factor loadings are larger for increasing levels of the latent trait

Examples of categorical latent traits include attachment style ('secure', 'avoidant', or 'anxious') and Piagetian stage of development ('sensorimotor stage', 'preoperational stage', 'concrete operational stage', and 'formal operational stage'). As can be seen from the table, four basic models arise: the Latent Class Model, the Item Response Model, the Latent Profile Model, and the Linear Factor Model. To name just a few applications in psychology: the Latent Class Model has been used to infer what cognitive strategies children use in solving arithmetic problems (Jansen & Van der Maas, 2002), the Item Response Model has been used to study liability to substance use disorders (Vanyukov, 2003) and to identify type D-personality (which is associated with increased cardiovascular disease; Emons, Meijer, & Denollet, 2006), the Latent Profile Model has been used to study eating disorders (Wade, Crosby, & Martin, 2006), and the Linear Factor Model has been used to study group differences in intelligence (Dolan, 2000) and personality (Smits, Dolan, Vorst, Wicherts, & Timmerman, 2011).

Assumptions

As the focus of this paper is on the analysis of psychological tests and questionnaires, the remainder of this paper will be on Item Response Models and Linear Factor Models (see Table 1). The most popular method that is used to apply these models to data (i.e., maximum likelihood; Lawley, 1943) requires a normal distribution for the item scores.[4] Note that some estimation procedures such as the asymptotic distribution free method for approximately continuous data (Browne, 1984) and non-parametric methods for ordinal data (e.g., Mokken, 1971) do not require a normal distribution. However, such methods are not suitable to explicitly model psychological hypotheses like those we are considering in this paper.

As pointed out in Molenaar, Dolan, and Verhelst (2010) for approximate continuous items, and in Molenaar, Dolan, & De Boeck (2012) for ordered categorical items, the assumption of a multivariate normal distribution implies three characteristics of the model in Figure 1. First, the latent trait scores should be normally distributed (Figure 3a), as opposed to a non-normal distribution (Figure 3b). Second, the factor loadings should not depend on the level of the latent trait (Figure 4a). That is, the factor loadings should be equal for every respondent irrespective of his or her position on the latent trait. If the factor loadings do depend on the level of the latent trait, as depicted in Figure 4b, the assumption of a normal distribution for the items will be violated. See Figure 5 for an illustration. Third, similarly, the residual variances should not depend on the level of latent trait, see Figure 6a. This notion is called *homoscedasticity* of the residual variances. If the residual variances depend on the level of the latent trait, as depicted in Figure 6b, this is referred to as *heteroscedasticity* of the residual variances. Note that the residuals should also be normally distributed. This is not the same as the assumption of normal data, as the distribution of the data also depends on the distribution of the latent trait (as discussed above).

[4] Note again from footnote 2 that for ordinal data, the normality assumption is also imposed but in a different manner.

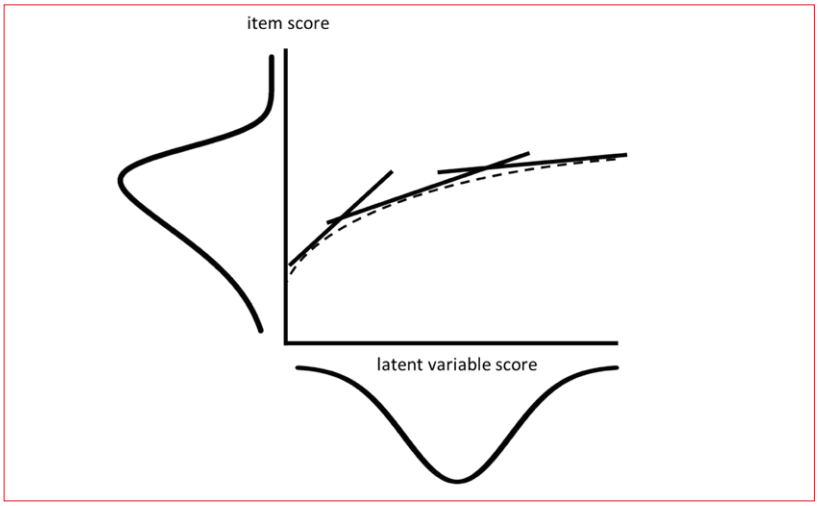

Figure 5. Illustration of how level dependent factor loadings will result in a non-normal distribution of the item scores (in case of –approximately– continuous item scores)

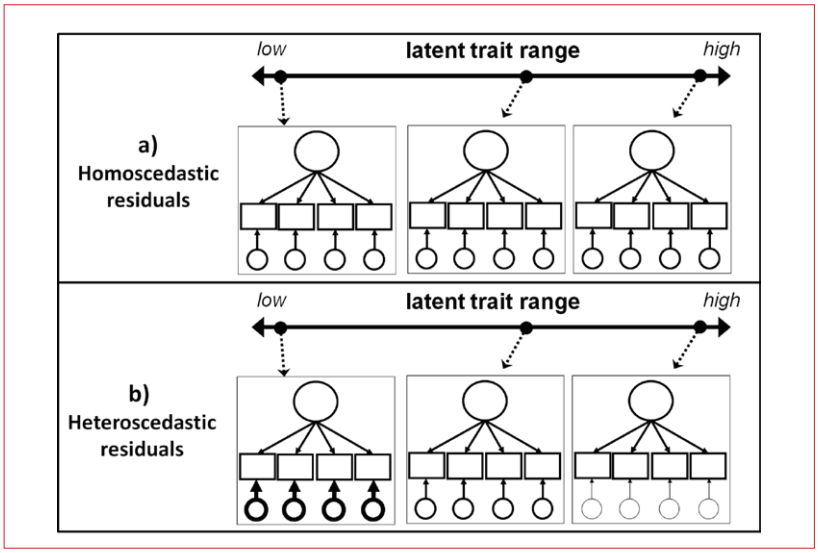

Figure 6. A) The assumption of homoscedastic residuals; B) a violation of this assumption, the residuals are heteroscedastic, i.e., they increase across the latent trait

These three characteristics of the latent trait model, normality of the latent trait, level independency of the factor loadings, and homoscedasticity of the residual variances, result in a normal distribution for the item scores. If one of these characteristics does not hold for a given dataset, e.g., the factor loadings are level dependent, the distribution of the items will not follow a normal distribution. Thus, all three characteristics should hold for a given dataset to enable application of latent traits models.

Substantive hypothesis in psychology that implies non-normality

A normal distribution for the item scores is a reasonable approximation in many applications in psychology. However, in the psychological literature, there are substantive hypotheses that predict specific departures from normality of the item scores. Thus, these hypotheses give a substantive reason why a normal distribution could not be assumed. Below we discuss three of these psychological phenomena, ability differentiation, schematicity, and gene-by-environment interactions (see Molenaar, et al., 2012).

Ability differentiation

In the intelligence literature, *positive manifold* refers to the well-replicated phenomenon that all subtests of a given IQ test (e.g., the WAIS-III; Wechsler, 1997) are all positively inter-correlated notwithstanding the fact that they all concern different cognitive abilities (such as working memory, perceptual speed, etc). This phenomenon was explained by Spearman (1904) by postulating that a single common factor underlies all subtest scores. He referred to this factor as the general intelligence factor, or *g*. Although the notion of a single common factor appeared to be untenable, g remains the most dominant dimension of individual differences in intelligence test scores as a single higher-order factor. In 1927, Spearman discovered that correlations between intelligence subtests were generally higher among a sample of 'mentally defective' children (average correlation: .782) as compared with a sample of 'normal' children (average correlation: .466).[5] This observation led Spearman to formulate the hypothesis that is now called 'the ability differentiation hypothesis', i.e., the g factor is not an equally strong source of individual differences across its range. Specifically, the *g* factor is stronger in people in the low end of the *g* distribution (e.g., the 'defective' children) as compared with people in the high end of the g distribution (e.g., the 'normal' children). As pointed out by Tucker-Drob (2009), Reynolds and Keith (2007), and Molenaar, Dolan, Wicherts, and Van der Maas (2010), a stronger g factor at the lower *g*-range implies 1) non-normality of *g*; 2) larger factor loadings for people low on *g*; and/or 3) smaller residual variances for people low on g. That is, ability differentiation implies at least one (and possibly more) of the violations that are depicted in Figure 3b, 4b and 6b.

[5] The terms 'defective' and 'normal' are from Spearman (1924)
[6] and measurement error

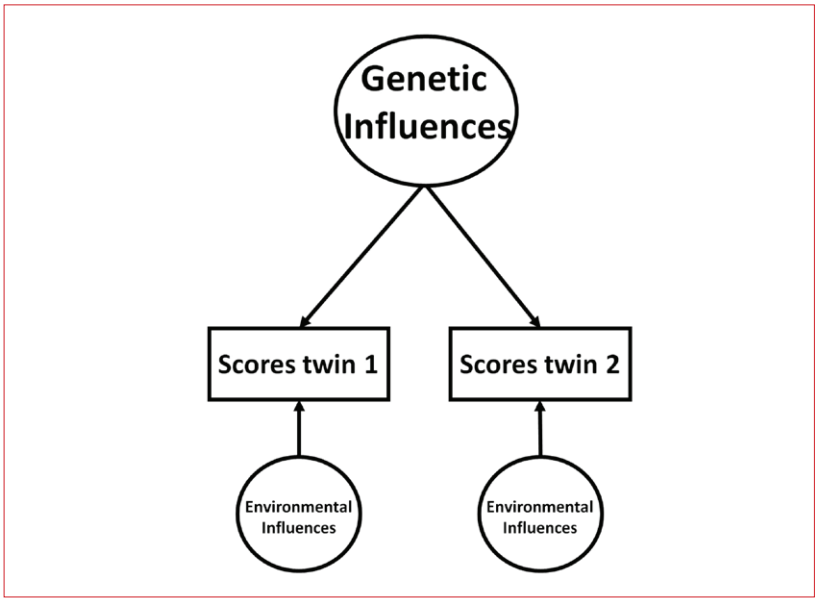

Figure 7. The classical twin design as a latent trait model. Note that this model is for monozygotic twins only

Schematicity

Psychological questionnaires differ from psychological tests in that the former concern an evaluation of the self (in personality research) or an object (organisational and educational psychology) while the latter involve some kind of problem solving. In case of self-evaluations, this difference makes psychological questionnaires relatively vulnerable to 'systematic noise factors' because evaluations of the self may be distorted due to social desirability or an inaccurate self image. With respect to the latter, researchers formulated the schematicity hypothesis. This hypothesis predicts that people differ in the accuracy with which they rate themselves on personality characteristics because of differences in their cognitive structures that are concerned with processing information about the self (Markus, 1977; Rogers, Kuiper & Kirker, 1977; Tellegen, 1988). Research indicated that high schematicity (i.e., having strong cognitive structures about the self) is associated with an extreme position on the construct (Markus, 1977). A possible explanation for this phenomenon could be that personality dimensions do not apply equally well to everybody (Allport, 1937; and Baumeister & Tice, 1988). This implies that toward the extreme of a personality dimension, less individual differences are present. This causes non-normality of the personality dimension as in Figure 3b, lower factor loadings on the extreme of the personality dimension as in Figure 4b, and/or larger residual variances towards the extreme of the personality dimension as in Figure 6b. That is, like ability differentiation, schematicity implies violation of the normality assumption in the traditional latent trait model.

Genotype-by-environment interaction

Psychological constructs are often found to be heritable, i.e., individual differences on for instance working memory can be explained to some degree by individual differences in genes. The remaining unexplained part is accounted for by environmental effects. The effect of both genes and environment is commonly studied within the classical twin design (e.g., Martin & Eaves, 1977; Eaves, Last, Martin, & Jinks, 1977). In this design, twins are tested on a (psychological) trait of interest. In the most simple version of the design, only monozygotic twins are considered, e.g., twins that share 100% of their genes. In this case, the similarities one observes between the two members of a twin on, for instance, a measure of verbal comprehension are due to genes. In addition, the differences one observes between twins are due to environmental effects. This instance of the twin design can be specified as a latent trait model (see Figure 7). As can be seen in the figure, the genetic effects are operationalised as the common latent trait underlying the scores of two members of a twin (comparable to the extraversion latent trait from Figure 1). The environmental effects are operationalised as the residual variances, i.e., the item specific effects. Within the twin design in Figure 7, factor loadings and residual variances are equal for the two items, as both items are from the same twin. Now –given standardisation of the item scores– the squared factor loading will equal the heritability of the measures in Figure 7 (assuming of course that the model is correctly specified). Thus, the squared factor loading represents the proportion of variance in the item that is due to genes. In addition, the residual variance will equal the proportion of variance that is due to the environment (i.e., one minus heritability).

In the standard model in Figure 7, the effects of genes and environment are additive, i.e., they do not interact (see Eaves, et al., 1977). It is, however, possible that sensitivity to environmental influences depends on genes, or vice versa. For instance, Turkheimer, Haley, Waldorn, D'Onofrio, and Gottesman (2003) found that genes are more expressed in measures of cognitive ability in environments of high SES as compared with environments low in SES. This is indicative of genotype-by-environment interaction, i.e., the effect of the environment depends on the specific genetic makeup of a sample of subjects. In the classical twin design in Figure 7, a genotype-by-environment interaction will be apparent if the effect of the environment (i.e., the residual variances) increases or decreases across the genetic factor. That is, a genotype-by-environment interaction will arise as heteroscedastic residuals similarly to in Figure 6b. Thus, genotype-by-environment interaction is again a phenomenon that is not in line with the standard assumption of normality in the latent trait model.

Table 2 Overview of all parameters in the extended model

	Parameter	Interpretation
TRADITIONAL PARAMETERS	Factor variance	Amount of individual differences on the psychological construct (e.g., extraversion)
	Factor loading	Degree to which a given item measures the psychological construct
	Residual variance Intercept	Amount of noise on measure of the construct In single group applications: The mean of the item. In multiple group applications: The baseline level for group comparison
NEW PARAMETERS	Shape parameter	Degree of skewness in the latent trait distribution. When zero, the latent trait is normally distributed.
	Non-linearity parameter	Degree of level dependency in the factor loadings. The larger this parameter, the more the factor loadings differ across the latent trait. When zero, the factor loadings are the same for everyone
	Heteroscedasticity parameter	Degree of heteroscedasticity in the residual variances. The larger this parameter, the more the residual variances differ across the latent trait. When zero, the residual variances are homoscedastic

The substantively motivated extended latent trait model

As argued above, some hypotheses in psychology predict specific violations of the normality assumption of the standard latent trait model. Therefore, extensions of the standard latent trait model from Figure 1 are needed to accommodate these hypotheses into the model. Specifically, as discussed above, the model needs to be extended so that it incorporates non-normality in the latent trait, level dependency in the factor loadings, and heteroscedasticity in the residual variances. We discuss these extensions below.

First, as illustrated by Azevedo, Bolfarine, and Andrade (2011), Verhelst (2008) and Molenaar et al. (2010), the distribution of the latent trait can be submitted to a so-called skew-normal density (Azzalini, 1985; 1986; Azzalini & Capatanio, 1999). This is a more flexible alternative to the normal distribution, as it can take a skewed shape. The skew-normal density introduces an additional parameter in the model, the *shape parameter*. This parameter quantifies the degree of skewness in the distribution of the latent trait. When the parameter is zero, the distribution of the latent trait is normal.

Second, Molenaar et al. (2010) showed that level dependent factor loadings as in Figure 4b result in non-linear factor loadings. This result is useful as there is a large body of literature on non-linear factor models (e.g., McDonald, 1965; Kenny & Judd, 1984; Klein & Moosbrugger, 2000; Bauer, 2005). Thus, these models can readily be used to model violations of normality. This introduces an additional parameter in the model, the non-linearity parameter. This parameter quantifies the degree to which the factor loadings depend on the level of the latent trait. For instance, for large values of this parameter, the factor loadings are highly different for people who are in the high regions of the latent trait distribution compared with those who are in the low region. When the parameter is zero, the factor loadings do not differ across the latent trait.

Third, Hessen and Dolan (2009) and Molenaar et al. (2010) propose factor models that incorporate heteroscedastic residuals. Specifically, the residual variance from Figure 1 is made an exponential function of the score on the latent trait. By doing so, an additional parameter arises in the model, the heteroscedasticity parameter. This parameter quantifies the degree of heteroscedasticity in the data. When the parameter is large, residual variances are clearly different for people who differ on the latent trait. In addition, when the parameter is zero, the residual variances are homoscedastic.

All three effects can be introduced into the latent trait model in Figure 1.[7] See Table 2 for an overview of the parameters of the resulting model. As can be seen, the model consists of the parameters from the traditional model and the new parameters that are discussed above. As this extended model does not assume normality of the item scores, it can be used to model the substantive psychological phenomena that are discussed in this paper. For instance, from applications of the extended model, it has become clear that ability differentiation is mainly due to non-linear factor loadings (Tucker-Drob, 2009) and a non-normal g factor (Molenaar, Dolan, & Van der Maas, 2011). In addition, using Item Response Theory, a systematic effect of schematicity was found in a dataset on alexythimia (Molenaar et al., 2012),

[7] Note that some effects can be combined (non-linear factor loadings with heteroscedastic residuals, and a skew-normal distribution for the latent trait with heteroscedastic residuals), and some of them cannot be combined (a skew-normal distribution for the latent trait with non-linear factor loadings), see Molenaar et al. (2010).

and a gene-by-environment interaction was found on cognitive ability (the effect of the environment increased with the genetic factor; Molenaar, Van der Sluis, Boomsma, & Dolan, in press).

Discussion

As all the hypotheses discussed in this paper predict non-normality of the item scores, one should be careful in drawing conclusions concerning the existence of phenomena such as ability differentiation and schematicity. Non-normality can have different causes that are not necessarily in line with the hypothesis under consideration. Below we discuss three of them: poor scaling, censoring, and unrepresentative samples.

First, poor scaling results from adding individual items that differ disproportionately in how difficult they are (e.g., adding item scores of 20 easy items and five difficult items), and from Likert scales in which one or more of the categories are disproportionately little used (e.g., a five-point scale in which nobody uses the middle category). We use the term poor scaling only within a measurement context in which a test or questionnaire is administered to assess a given psychological construct. In case of a classification context in which individuals are assigned to certain categories (e.g., depressed or not-depressed), item difficulties are commonly distributed around the cutoff point (see Hambleton, Swaminathan, & Rogers (1990; Chapter 7). In the Linear Factor Model (see Table 1), poor scaling can result in heteroscedastic residuals (Van der Sluis, Dolan, Neale, Boomsma, & Posthuma, 2006). It is therefore important to check for poor scaling in the data before testing the schematicity hypothesis for instance to avoid spurious effects. Items that show poor scaling should be omitted from the analysis, or the analyses should be done using Item Response Models. In these models, poor scaling is not a problem as each answer category of an item is modelled separately.

Another alternative source of non-normality is censoring. Censoring occurs when the majority of a sample obtains the highest or lowest possible score for a Likert scale item, or a sum score. This is also referred to as a ceiling or floor effect, respectively. When Likert scales are analysed as continuous using the Linear Factor Model, censoring of the item scores can result in heteroscedastic residuals (Van der Sluis et al., 2006). Again this is problematic as this heteroscedasticity might wrongfully be taken as evidence for a gene-by-environment interaction, for instance. As with poor scaling, it is thus wise to first check the items in the analyses for possible floor and/or ceiling effects. Items that show censoring could be omitted from the analyses, or the analyses can be conducted using Item Response Model. As with poor scaling, censoring is not a problem in these models as each answer category of an item is modelled separately.

Third, unrepresentative samples can bias conclusion concerning non-normality. For instance, in a study in which an intelligence test is administered to a sample of subjects, less bright people could be less willing to participate in the study as they know they will do badly. This causes a skewed intelligence distribution in the sample that might wrongfully be interpreted in terms of ability differentiation. Therefore, it is of importance that the data are not subject to sampling bias. The problem of unrepresentative samples might be the most difficult problem of the ones discussed above, as no clear post hoc solution exists. There are some statistical possibilities, e.g., the subjects in the sample could be weighted on important background variables. However, appropriate procedures are not yet available within the model as discussed in this paper. This could be interesting work for future research.

References

Allport, G. W. (1937). *Personality. A psychological interpretation*. New York: Henry Holt.

Amthauer R., Brocke B., Liepmann D., Beauducel A. (2001). I-S-T 2000 R. *Intelligenz-Struktur-Test 2000 R*. Hogrefe Verlag; Göttingen.

Arbuckle, J. L. (1997). Amos (version 3.61) [Computer software]. Chicago, IL: Small Waters. SAS Institute Inc. (2011). *SAS/STAT software: Release 9.3*. Cary, NC: SAS Institute, Inc.

Azevedo, C. L. N., Bolfarine, H., & Andrade, D. F. (2011). Bayesian inference for a skew-normal IRT model under the centred parameterization. *Computational Statistics and Data Analysis*, 55, 353-365.

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, 12, 171-178.

Azzalini, A. (1986). Further results on a class of distributions which includes the normal ones. *Statistica*, 46, 199-208.

Azzalini, A., & Capatanio, A.(1999). Statistical applications of the multivariate skew normal distribution. *Journal of the Royal Statistical Society. Series B*, 61, 579-602.

Bauer, D. J. (2005). The Role of Nonlinear Factor-to-Indicator Relationships in Tests of Measurement Equivalence. *Psychological Methods*, 10, 305–316.

Baumeister, R. E., & Tice, T. M. (1988). Metatraits. *Journal of Personality*, 56, 571-598.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In E. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (chap. 17-20), Reading, MA: Addison Wesley.

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

Bollen, K. A. (1989). *Structural equations with latent variables*. John Wiley, New York.

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

Borsboom, D. (2008). Latent variable theory. *Measurement, 6*, 25-53.

Browne, M. W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology, 37*, 62-83.

Dolan, C. V. (1994). Factor analysis of variables with 2, 3, 5 and 7 response categories: A comparison of categorical variable estimators using simulated data. *British Journal of Mathematical and Statistical Psychology, 47*, 309-326.

Dolan, C. V. (2000). Investigating Spearman's hypothesis by means of multi-group confirmatory factor analysis. *Multivariate Behavioral Research, 35*, 21-50.

Eaves, L. J., Last, K., Martin, N. G., & Jinks, J. L. (1977). A progressive approach to non additivity and genotype-environmental covariance in the analysis of human differences. *British Journal of Mathematical and Statistical Psychology, 39*, 1-42.

Emons, W. H., Meijer R. R., & Denollet. J. (2007). Negative affectivity and social inhibition in cardiovascular disease: Evaluating type-D personality and its assessment using item response theory. *Journal of Psychosomatic Research, 63,* 27-39.

Gibson, W.A. (1959). Three multivariate models: factor analysis, latent structure analysis, and latent profile analysis. *Psychometrika, 24*, 229-252.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika, 61*, 215-231.

Guadagnoli, E., & Mor, V. (1989). Measuring cancer patients' affect: Revision and psychometric properties of the Profile of Mood States (POMS). *Psychological Assessment, 1*, 150-154.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage: Newbury Park, CA.

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

Jansen, B. R. J., & Van der Maas, H. L. J. (2002). The development of children's rule use on the balance scale task. *Journal of Experimental Child Psychology, 81*, 383-416.

Jöreskog, K. G. & Sörbom, D. (1993). LISREL 8: *Structural equation modelling with the SIMPLIS command language*. Chicago: Scientific Software International.

Kenny, D. A., & Judd, C. M. (1984). Estimating the nonlinear and interactive effects of latent variables. *Psychological Bulletin, 96*, 201-210.

Klein, A., & Moosbrugger, H. (2000). Maximum likelihood estimation of latent interaction effects with the LMS method. *Psychometrika, 65*, 457-474.

Lawley, D. N. (1943). The application of the maximum likelihood method to factor analysis. *British Journal of Psychology, 33*, 172–175.

Lawley, D. N., & Maxwell, A. E. (1971). *Factor analysis as a statistical method*. New York: American Elsevier.

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis*. Boston, MA: Houghton Mifflin.

Lord, F. M. (1952). A theory of test scores. New York, NY: Psychometric Society.

Markus, H. (1977). Self-schemata and processing information about the self. *Journal of Personality and Social Psychology, 35*, 63-78.

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

McCutcheon, A. L. (1987). *Latent Class Analysis*. Newbury Park, Calif: Sage Publications.

Mellenbergh, G.J. (1994). Generalized linear item response theory. *Psychological Bulletin, 115*, 300-307.

Mellenbergh, G. J. (2012). Models for continuous responses. Manuscript submitted for publication.

McDonald, R. P. (1965). Difficulty factors and non-linear factor analysis. *British Journal of Mathematical and Statistical Psychology, 18*, 11-23.

Mokken, R. J. (1971). *A theory and procedure of scale analysis*. Berlin: De Gruyter.

Molenaar, D., Dolan, C. V., & Van der Maas, H. L. J. (2011). Modeling ability differentiation in the second-order factor model. *Structural Equation Modeling, 18*, 578-594.

Molenaar, D., Dolan, C. V., & Verhelst, N. D. (2010). Testing and modeling non-normality within the one factor model. *British Journal of Mathematical and Statistical Psychology, 63*, 293-317.

Molenaar, D., Dolan, C. V., Wicherts, J. M., & Van der Maas, H. L. J. (2010). Modeling differentiation of cognitive abilities within the higher-order factor model using moderated factor analysis. *Intelligence, 38*, 611-624.

Molenaar, D., Van der Sluis, S., Boomsma, D. I., & Dolan, C. V. (in press). Detecting specific genotype by environment interaction using marginal maximum likelihood estimation in the classical twin design. *Behavior Genetics*.

Molenaar, D., Dolan, C. V., & de Boeck, P. (2012). The Heteroscedastic Graded Response Model with a Skewed Latent Trait: Testing Statistical and Substantive Hypotheses related to Skewed Item Category Functions. *Psychometrika, 77*, 455-478.

Muthén, L. K., & Muthén, B. O. (2007). *Mplus User's Guide. Fifth Edition*. Los Angeles, CA.

Rasch, G. (1960). Probabilistic Models for Some Intelligence and Attainment Tests. Copenhagen, Denmark: Danmarks Paedogogiske Institut.

Reynolds, M. R., & Keith, T. Z. (2007). Spearman's law of diminishing returns in hierarchical models of intelligence for children and adolescents. *Intelligence, 35*, 267-281.

Rogers, T. B., Kuiper, N. A., & Kirker, W. S. (1977). Self-reference and the encoding of personal information. *Journal of Personality and Social Psychology, 35*, 677-688.

Samejima, F. (1973). Homogeneous Case of the Continuous Response Level. *Psychometrika, 38*, 203-219.

Smits, I. A. M., Dolan, C. V., Vorst, H. C. M., Wicherts, J. M., & Timmerman, M. E. (2011). Cohort differences in big five personality factors over a period of 25 years. *Journal of Personality and Social Psychology, 100*, 1124-1138.

Spearman, C. E. (1904). 'General intelligence' objectively determined and measured. *American Journal of Psychology, 15*, 201-293.

Spearman, C. E. (1927). *The abilities of man: Their nature and measurement*. New York: Macmillan.

Tellegen, A. (1988). The analysis of consistency in personality assessment. *Journal of Personality, 56*, 621-663.

Tucker-Drob, E. M. (2009). Differentiation of cognitive abilities across the life span. *Developmental Psychology, 45*, 1097-1118.

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

Van der Sluis, S., Dolan, C. V., Neale, M. C., Boomsma, D. I., & Posthuma, D. (2006). Detecting Genotype Environment Interaction in Monozygotic Twin Data: Comparing the Jinks and Fulker Test and a New Test Based on Marginal Maximum Likelihood Estimation. *Twin Research and Human Genetics, 9*, 377-392.

Vanyukov, M. M., Kirisci, L., Tarter, R. E., Simkevitz, H. F., Kirillova, G. P., Maher, B. S., et al. (2003). Liability to substance use disorders: 2. A measurement approach. *Neuroscience Biobehavioral Reviews, 27,* 517–526.

Verhelst, N. D. (2008). *Latent variable analysis with skew distributions*. Internal report. Arnhem: Cito.

Vorst, H. C. M., & Bermond, B. (2001). Validity and reliability of the Bermond-Vorst alexithymia questionnaire. *Personality and Individual Differences, 30*, 413-434.

Wade, T. D., Crosby, R. D., & Martin, N. G. (2006). Use of latent profile analysis to identify eating disorder phenotypes in an adult Australian twin cohort. *Archives of General Psychiatry, 63*,1377–1384.

Wechsler, D. (1997). *Wechsler Adult Intelligence Scale- III (WAISIII)*. San Antonio, TX: Psychological Corp.

DYLAN MOLENAAR

Assistant professor at the department of psychological methods, University of Amsterdam. He graduated cum laude in 2012. Title of his thesis was 'testing distributional assumptions in psychometric measurement models with substantive applications in psychology'. His research interests include: item response theory, factor analysis, and response time modelling.

CONOR DOLAN

Associate professor at the department of psychological methods, University of Amsterdam, and full professor at the VU University, Amsterdam. His research interests include: covariance structure modelling, mixture analyses, modelling of multivariate intelligence test scores, and the detection of genotype by environment interactions.

Response shift detection through then-test and structural equation modelling: Decomposing observed change and testing tacit assumptions

Assessment of change in patient-reported outcomes may be invalidated by the occurrence of response shift. Response shift refers to a change in a respondent's internal standards that may cause changes in observed variables that are not directly related to change in the construct of interest. An established approach for detecting response shift in the area of health-related quality of life (HRQL) is to administer a retrospective pre-test (then-test). In this study, the then-test was incorporated in the structural equation modelling (SEM) approach to (1) compare the then-test approach and the SEM approach in their decomposition of observed change and (2) to test the underlying assumptions of the then-test approach. In an application to HRQL data of 170 cancer patients undergoing invasive surgery, we found that both approaches revealed a similar pattern of decomposition, although there were some differences in the size and direction of change. With regard to the underlying assumptions of the then-test approach, results showed: (1) no evidence for recall-bias (Recall Assumption supported for all scales), (2) that internal standards of measurement were not invariant across post- and then-test measures (Consistency Assumption rejected for four out of nine scales), and (3) that internal standards were not only affected by the recalibration type of response shift (Recalibration Assumption rejected for three out of nine scales). Valid approaches for detecting response shift and the consequences assessing changes in HRQL should be further investigated.

Where: Netherlands Journal of Psychology, Volume 67, 58-67

Received: 2 August 2012; Accepted: 4 December 2012

Keywords: Structural equation modelling; Response shift; Then-test; Health-related quality of life

Authors: Mathilde G. E. Verdam*,**, Frans J. Oort*,**, Mechteld R. M. Visser** and Mirjam A. G. Sprangers**

* Research Institute of Child Development and Education, University of Amsterdam, the Netherlands
** Department of Medical Psychology, Academic Medical Centre, University of Amsterdam, the Netherlands
Correspondence to:
M.G.E. Verdam, Department of Child Development and Education, University of Amsterdam, Nieuwe Prinsengracht 130, 1018 VZ Amsterdam, the Netherlands, email: m.g.e.verdam@uva.nl

Patient-reported outcomes of health-related quality of life (HRQL) are becoming increasingly more important in evaluating treatment effects in clinical settings. However, there is a well-known disparity between patient-reported and clinical measures of function. One explanation for this disparity is related to the dynamic nature of the HRQL construct (Allison, Locker, & Feine, 1997). The dynamic nature of the construct entails that the standards by which individuals assess their HRQL can differ between subjects and can change within subjects over time. Such a change in standards (or frame of reference) may cause changes in observed variables that are not directly related to change in the construct of interest. It is therefore important to detect possible changes in respondent's internal standards. Change in internal standards is also referred to as 'response shift'. The term response shift was first used in research on educational training interventions (Howard et al., 1979) and was also investigated in the field of organisational change where they used the terminology of 'alpha', 'beta' and 'gamma' change (Golembiewski et al, 1976). In the area of HRQL research, Schwartz & Sprangers (1999) proposed a theoretical model of response shift that distinguishes three types of response shift: (1) recalibration, which refers to a change in the respondent's internal standards of measurement, (2) reprioritisation, that refers to a change in respondent's values regarding the relative importance of component domains of the target construct, and (3) reconceptualisation, referring to a change in definition of the target construct. Response shift causes comparison of measurements over time

Table 1 Decomposition of observed change according to the then-test approach and the SEM approach

Then-test approach

Observed change = True change + Recalibration

$(X_{post} - X_{pre}) = (X_{post} - X_{then}) + (X_{then} - X_{pre})$

SEM approach

Observed change = True change + Recalibration + Reprioritisation & Reconceptualisation

$(\mu_{post} - \mu_{pre}) = \Lambda_{pre} * \alpha_{post} + (\tau_{post} - \tau_{pre}) + (\Lambda_{post} - \Lambda_{pre}) * \alpha_{post}$

In the then-test approach scores for the different measurements are denoted with 'X' to reflect the observed nature of the scores. In the SEM approach Greek symbols reflect the parameter estimates of observed factor means (μ), factor loadings (Λ), common factor means (α) and intercepts (τ)

to be incomparable. Therefore, when investigating changes in HRQL, it is important to also investigate – and account for – response shift effects.

Several methodological approaches are available to investigate response shift in longitudinal HRQL research (Schwartz & Sprangers, 1999; Schwartz et al., 2011). The 'then-test' approach is most commonly used, and includes a retrospective pre-test measure in addition to the usual pre and post measures. This retrospective pre-test is administered on the post-test occasion and asks respondents to re-evaluate their HRQL at the time of pre-test. As the then-test and post-test are administered at the same time, it is assumed that both measurements are completed with the same internal standard, thus avoiding response shift effects. Comparison of the post-test and then-test scores would yield an unbiased indication of the treatment effect ('true change', see Table 1). Furthermore, differences between the then-test and pre-test scores could be used as an assessment of changes in subjects' internal standards (response shift). The then-test approach thus allows a decomposition of observed change (differences between pre-test and post-test scores) into true change and response shift. However, these interpretations are only valid when the following assumptions are met:

1) Recall Assumption: At then-test occasion respondents are able to recall their state at pre-test. The validity of the then-test depends on the underlying assumption that memory (the recall of the pre-test state) is accurate and alternative cognitive explanations (e.g. social desirability, cognitive dissonance, implicit theory of change, expectancy or experimenter effects) do not play a role.
2) Consistency Assumption: Post- and then-test are completed with the same internal standard. A valid comparison of then-test and post-test scores depends on the underlying assumption that the respondent's internal standards of measurement are invariant across these assessments.
3) Recalibration Assumption: All response shift is of the recalibration type. As the then-test approach aims to assess only recalibration – not reprioritisation and reconceptualisation – the comparison of then-test and pre-test scores in assessing response shift is only accurate if all response shift is of the recalibration type.

An alternative method to detecting response shift is the structural equation modelling (SEM) approach (Oort, 2005). Similar to the then-test approach, the SEM approach provides a way to decompose observed change into true change and response shift (Oort, 2005, p. 495), based on the estimates of the factor model parameters (Table 1). An advantage of the SEM approach is that it allows for the statistical comparison of separate components of the measurement model over time, enabling operationalisation of the different types of response shift.

The SEM approach can therefore be used not only as a technique for the detection of response shift, but also for a substantive analysis of the decomposition of change. Moreover, the characteristics of the SEM approach provide a unique opportunity to test the underlying assumptions of the then-test approach. Incorporating the then-test into the SEM approach allows for testing the validity (and consistency) of the measurement model for post- and then-test (Consistency Assumption) and assessing not only the occurrence of recalibration, but also reprioritisation and reconceptualisation (Recalibration Assumption). Moreover, recall bias can be investigated by examining effects on the underlying constructs instead of the observed variables (Recall Assumption).

Therefore, the aim of this study is to illustrate how incorporation of the then-test into the SEM approach enables: 1) a substantive comparison of both approaches in their decomposition of observed change into true change and (types of) response shift, and 2) testing the underlying assumptions of the then-test approach.

Method

Cancer patients' health-related quality of life was assessed prior to surgery (pre-test) and three months following surgery (post-test and then-test). These

data have been used before to investigate response shift with the then-test and the SEM approach (Visser, Oort & Sprangers, 2005).

Patients

A consecutive series of 170 newly diagnosed cancer patients were enrolled, including 29 lung cancer patients undergoing either lobectomy or pneumectomy, 43 pancreatic cancer patients undergoing pylorus-preserving pancreaticoduodenectomy, 46 oesophageal cancer patients undergoing either transhiatal or transthoracic resection and 52 cervical cancer patients undergoing hysterectomy. Exclusion criteria were being under the age of 18, having a life expectancy less than 9 months, or not being able to complete a (Dutch) questionnaire. The sample consisted of 87 men and 83 women, with ages ranging from 27 to 83 (mean 57.5, standard deviation 14.1).

Measures

Generic health-related quality of life was assessed with the Dutch language version (Aaronson et al., 1998) of the SF-36 health survey (Ware, Snow, Kosinski, & Gandek, 1993), encompassing eight scales: physical functioning (PF), role limitations due to physical health (role-physical, RP), bodily pain (BP), general health perceptions (GH), vitality (VT), social functioning (SF), role limitations due to emotional problems (role-emotional, RE), and mental health (MH). Fatigue (FT) was measured with a six-item short form of the multidimensional fatigue inventory (MFI; Smets, Garssen, Bonke, & De Haes, 1995), to cover effects on patients' fatigue more thoroughly. For computational convenience the original scale scores of the SF-36 scales and the short form of the MFI were transformed to scales ranging from 0 to 5, with higher scores indicating better health. There were no missing data, as completion of the self-administered questionnaires was checked by an interviewer.

Structural equation modelling

The SEM procedure (Oort, 2005) was applied to the data of pre-, post- and then-tests to detect response shift and includes: 1) establishing an appropriate measurement model, 2) fitting a model of no response shift, 3) detection of response shift, and 4) assessment of true change. The measurement model was established on the basis of published results of principal components analyses of the SF-36 (Ware et al., 1993), results of exploratory factor analyses of the present data, and substantive considerations. The measurement model has no across measurement constraints. To test for the occurrence of response shift the second step in the SEM procedure is to fit a model of no response shift (where all model parameters that are associated with response shift are constrained to be equal across measurements). To test the presence of response shift, the no response shift model is compared with the model with no across measurement constraints. The third step in the SEM procedure begins with the no response shift model and uses step-by-step modification to arrive at the response shift model where all apparent response shifts are accounted for. Response shift is operationalised as across-measurement differences between patterns of common factor loadings (reconceptualisation), values of common factor loadings (reprioritisation), differences between intercepts (uniform recalibration), and differences between residual variances (nonuniform recalibration). In the fourth step of the SEM procedure, true change is assessed in the model where response shift is accounted for.

Structural equation models were fitted to the means, variances and covariances of the SF-36 and MFI scale scores of pre-, post- and then-test, using standard statistical computer programs (Jöreskog & Sorbom, 1996; Neale, Boker, Xie & Maes, 1999) (LISREL provides modification indices and Mx provides likelihood-based confidence intervals). To achieve identification of all model parameters, scales and origins of the common factors were established by fixing the factor means at zero and the factor variances at one. In Steps 2 and 3 of the procedure, factor means and variances are only fixed for first occasion (pre-test); post-test and then-test factor means and variances are then identified by constraining intercepts and factor loadings to be equal across assessments (Oort, 2005).

Goodness-of-fit was evaluated with the χ^2 test of exact fit, where a significant χ^2 indicates a significant difference between data and model. However, in the practice of structural equation modelling, exact fit is rare, and with large sample sizes the χ^2 test generally turns out to be significant. An alternative measure of overall goodness-of-fit is the root mean square of approximation (RMSEA). According to a generally accepted rule of thumb, an RMSEA value below .08 indicates 'reasonable' fit and one below .05 'close' fit (Browne & Cudeck, 1992). In addition, the comparative fit index (CFI; Bentler, 1990) gives an indication of model fit based on model comparison (compared with the model with no across measurement constraints), where CFI of .97 or higher is indicative of good fit and CFI between .95 and .97 of acceptable fit. Yet another fit index is the expected cross validation index (ECVI; Browne & Cudeck, 1989) which is a measure of the discrepancy between the model-implied covariance matrix in the analysed sample ('calibration' sample), and the covariance matrix that would be expected in another sample of the same size ('validation'

Table 2 Means and standard deviations for SF-36 and MFI scales before surgery (pre-test) and three months after surgery (post-test and then-test)

	Pre-test		Post-test		Then-test	
Scale	Mean	SD	Mean	SD	Mean	SD
PF	3.96	1.22	3.18	1.32	4.05	1.37
RP	2.73	2.09	2.13	2.02	2.99	2.14
BP	3.94	1.19	3.68	1.21	4.20	1.27
SF	3.81	1.32	3.62	1.47	3.72	1.32
MH	3.25	1.08	3.69	1.05	3.26	1.14
RE	3.00	2.12	3.55	1.93	2.84	2.13
VT	3.14	1.26	2.77	1.23	3.18	1.32
GH	2.96	0.95	2.96	1.06	2.76	1.08
FT	3.30	1.10	2.92	1.18	3.24	1.17

n = 170; SF-36 and MFI scale scores range from 0 to 5

sample). The ECVI can be used to compare different models for the same data, where the model with the smallest ECVI indicates the model with the best fit.

The χ^2 difference test (Bollen, 1989) was used to compare the fit of nested models, where a significant χ^2 indicates that the addition of model parameters significantly improves the model fit. Significant modification indices (Jöreskog & Sorbom, 1996) and standardised residuals > .10 were assumed to indicate response shift. The specification search was consistently guided by substantive consideration in order to retain a theoretical sensible model. Each modification was tested with the χ^2 difference test (Bollen, 1989).

Objective 1: Decomposition of change

The equations in Table 1 give the decomposition of observed change into true change and response shift for both the then-test approach and the SEM approach. For the then-test approach the standard deviations of the observed change scores are used to calculate standardised mean differences (as effect size indices d) for the components of observed change. For the SEM approach the parameter estimates of the final model (in which all response shift is accounted for) were used to calculate standardised mean differences (as effect size indices *d*) for the components of observed change (Oort, 2005). Effect-size values of d = .2, .5 and .8 are considered 'small', 'medium', and 'large' (Cohen, 1988).

Objective 2: Testing the assumptions of the then-test approach

The Recall Assumption can be tested by testing the equality of pre-test and then-test common factor means because the common factor means of the response shift model should refer to the same state (of pre-test). The Recall Assumption would be supported when the equality constraint across pre- and then-test common factor means is tenable (indicated by the χ^2 difference test).

The Consistency Assumption can be tested by imposing equality constraints across post- and then-test factor loadings (reconceptualisation and reprioritisation), intercepts (uniform recalibration) and residual variances (nonuniform recalibration). When response shift detection (using the χ^2 difference test) is invariant across assessments, the Consistency Assumption is supported.

The Recalibration Assumption can be tested by examining recalibration, reprioritisation and reconceptualisation types of response shift. When all response shifts detected (using the χ^2 difference test) are of the recalibration type, the Recalibration Assumption is supported.

Results

Table 2 gives pre-, post- and then-test means and standard deviations for all SF-36 and MFI scales.

Measurement model

Results from exploratory factor analyses and substantive considerations gave rise to the measurement model in Figure 1 (see Oort, Visser & Sprangers, 2005 for more information on selection of this measurement model). The circles represent unobserved, latent variables and the squares represent the observed variables. Three latent variables are the common factors general physical health (GenPhys), general mental health (GenMent), and general fitness (GenFitn). GenPhys is measured by PF, RP, BP and SF, GenMent is measured by MH, RE, and again SF, and GenFitn is measured by VT, GH, and FT. Other latent variables are the residual factors ResPF, ResRP, ResBP, etc. The residual factors represent all that is specific to PF, RP, BP, etc., plus random error variation. In addition, Figure 1 shows the measurement model of the model in which all response shift is accounted for (dotted lines represent factor loadings that were present at post- and/or then-test only). Numbers in Figure 1 are maximum likelihood estimates of common factor loadings, common factor correlations, residual

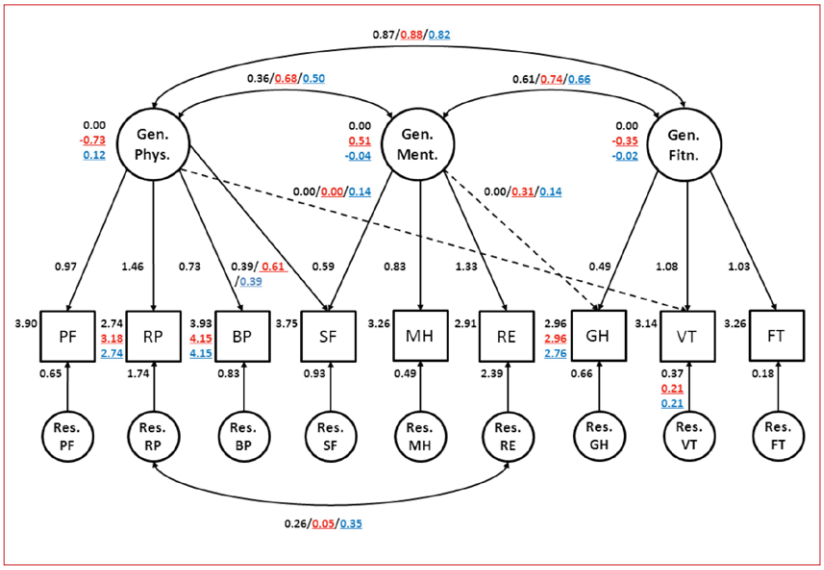

Figure 1 The measurement model used in response shift detection
Circles represent latent variables (common and residual factors) and squares represent observed variables (the SF-36 and MFI scales). Numbers are maximum likelihood estimates of the response shift model parameters: common factor loadings, common factor correlations, residual variances, and residual correlations. Single values represent estimates that were constrained to be equal across time, whereas multiple values represent different pre-test (black), post-test (red) and then-test (blue) estimates.

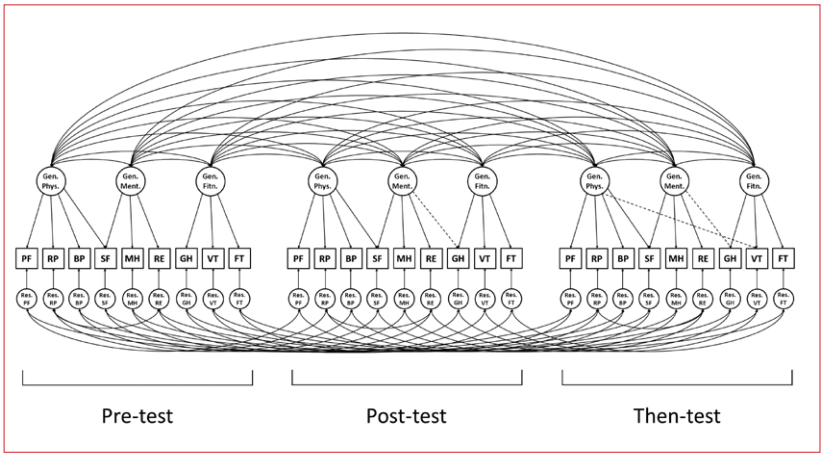

Figure 2 The longitudinal structural equation model fitted to the data
Circles represent latent variables (common and residual factors) and squares represent observed variables (the SF-36 and MFI scales). Dotted lines represent factor-loadings unique for post- or then-test assessment.and then-test (blue) estimates.

variances, and three residual correlations (single values represent estimates that are constrained to be equal across pre-, post- and then-test, whereas multiple values represent separate estimates for pre-test (black), post-test (red), and then-test (blue)). Figure 2 gives a visual representation of the full longitudinal model that was fitted to the data.

The measurement model of Figure 1 was the basis for a structural equation model for pre-, post and then-test with no across measurement constraints. The $\chi 2$ test of exact fit was significant (CHISQ(255) = 349.13, p<.001) but the RMSEA measure indicated close fit (RMSEA = .041, see Table 3).

Detection of response shift

To test for the occurrence of response shift, all model parameters that are associated with response shift were held invariant across measurements. This means that all across measurement invariance constraints on factor loadings, intercepts, and residual variance were imposed. The fit of the no response shift model, although still satisfactory (RMSEA = .049, see Table 3), was significantly worse than the fit of model with no across measurement constraints (χ^2 difference test: CHISQ(56) = 99.26, p < .001), indicating the presence of response shift.

Inspection of modification indices and standardised residuals indicated which of the equality constraints were not tenable. Step by step modifications yielded the response shift model, which showed several cases of response shift, as will be explained below. The fit of the response shift model was good (RMSEA = .035, see Table 3), and significantly better than the fit of the no response shift model (χ^2 difference test: CHISQ(8) = 74.12, p < .001). All estimates of the response shift model parameters are given in Table 3.

Evaluation of response shifts and true change

Reconceptualisation: A change in the pattern of factor loadings across assessments is indicative of reconceptualisation. Comparison of the common factor loadings of the pre-test with those of the post-test and then-test (Table 4, top rows) showed that at both the post- and then-test GH became an indicator for GenMent, indicating reconceptualisation of GH. The VT scale became an indicator for GenPhys at the then-test, indicating reconceptualisation of VT at the then-test only.

Reprioritisation: The values of the factor loadings contain information about reprioritisation. The common factor loading of SF on GenPhys became larger at the post-test, indicating reprioritisation of SF at the post-test only.

Recalibration: Intercepts and residual variances contain information about uniform and nonuniform recalibration. For RP, we found differences between the pre- and post-test intercepts, indicating uniform recalibration of RP at the post-test only. For GH, we found differences between the pre- and then-test intercepts, indicating uniform recalibration of GH at the then-test only. For BP, we found differences between the pre-test and both the post- and then-test intercepts, indicating uniform calibration of BP that equally affects both the post- and then-test. We also

Table 3 Goodness of overall fit of models in the three-step response shift detection procedure

Model	Description	DF	CHISQ	RMSEA	ECVI	CFI
Model 1	Measurement model (no across measurement constraints)	255	349.13	.041 (.026; .053)	3.39 (3.14; 3.69)	.99
Model 2	No response shift model	299	448.39	.049 (.037; .059)	3.73 (3.28; 3.92)	.98
Model 3	Response shift model	291	374.27	.035 (.019; .048)	3.43 (3.01; 3.58)	.99

n = 170; Numbers between parentheses represent 90% confidence intervals

found a change in the variance of the residual factor ResVT, indicating nonuniform recalibration of VT that affects both the post- and then-test equally.

True change: Common factor means were fixed at zero for the pre-test (because of identification requirements), so that the post-test estimates serve as direct representations of true change. The differences between the pre- and post-test common factor means were significant ($p < .001$) for each of the common factors. GenPhys (-0.73) and GenFitn (-0.35) deteriorated, and GenMent (+0.51) improved, with effect sizes that can be considered 'medium' (d = -.66, -.33, .55 respectively).

Table 4 Parameter estimates in the response shift model

	Pre-test			Post-test			Then-test		
	GenPhys	**GenMent**	**GenFitn**	**GenPhys**	**GenMent**	**GenFitn**	**GenPhys**	**GenMent**	**GenFitn**
Factor loadings (Λ)									
PF	0.97			0.97			0.97		
RP	1.46			1.46			1.46		
BP	0.73			0.73			0.73		
SF	0.39	0.59		**0.61**	0.59		0.39	0.59	
MH		0.83			0.83			0.83	
RE		1.33			1.33			1.33	
VT			1.08			1.08	**0.14**		1.08
GH			0.49		**0.31**	0.49		**0.14**	0.49
FT			1.03			1.03			1.03
Intercepts (τ)	**PF**	**RP**	**BP**	**SF**	**MH**	**RE**	**VT**	**GH**	**FT**
Pre-test	3.90	2.74	3.93	3.75	3.26	2.91	3.14	2.96	3.26
Post-test	3.90	**3.18**	**4.15**	3.75	3.26	2.91	3.14	2.96	3.26
Then-test	3.90	2.74	**4.15**	3.75	3.26	2.91	3.14	**2.76**	3.26
Residual variance (Diag(Θ))	**ResPF**	**ResRP**	**ResBP**	**ResSF**	**ResMH**	**ResRE**	**ResVT**	**ResGH**	**ResFT**
Pre-test	0.65	1.74	0.83	0.93	0.49	2.39	0.37	0.66	0.18
Post-test	0.65	1.74	0.83	0.93	0.49	2.39	**0.21**	0.66	0.18
Then-test	0.65	1.74	0.83	0.93	0.49	2.39	**0.21**	0.66	0.18
Residual correlations (Θ *)									
Pre x Post	0.28	0.13	0.35	0.05	0.43	0.00	0.27	0.32	0.15
Pre x Then	0.62	0.22	0.42	0.26	0.54	- 0.06	0.26	0.27	-0.02
Post x Then	0.41	0.04	0.19	0.18	0.58	0.26	0.26	0.27	0.22
Common factor variances (Diag(Φ))	**Pre-test**			**Post-test**			**Then-test**		
	GenPhys	**GenMent**	**GenFitn**	**GenPhys**	**GenMent**	**GenFitn**	**GenPhys**	**GenMent**	**GenFitn**
	1.00	1.00	1.00	1.23	0.86	1.13	1.33	1.19	1.08
Common factor correlations (Φ*)									
Pre-test									
Gen-Phys	1								
Gen-Ment	0.36	1							
Gen-Fitn	0.87	0.61	1						
Post-test									
Gen-Phys	0.55	0.35	0.53	1					
Gen-Ment	0.38	0.41	0.43	0.68	1				
Gen-Fitn	0.47	0.43	0.59	0.88	0.74	1			
Then-test									
Gen-Phys	0.82	0.37	0.73	0.41	0.25	0.32	1		
Gen-Ment	0.40	0.59	0.45	0.20	0.32	0.18	0.50	1	
Gen-Fitn	0.76	0.50	0.83	0.35	0.32	0.38	0.82	0.66	1
Common factor means (α)	0.00	0.00	0.00	**-0.73**	**0.51**	**-0.35**	0.12	-0.04	-0.02

n = 170; Results indicating across-measurement variance are printed in bold. Factor loadings are unstandardised, but covariances are decomposed into variances and correlations

Table 5 The decomposition of observed change into true change and response shift (displayed as standardised differences), for both the then-test approach and the SEM approach

	Then-test approach			SEM approach		
Scale	Observed change	True Change	Response shift [a]	Observed change	True change	Response shift
PF	-0.59**	-0.66**	0.07	-0.51**	-0.51**	-
RP	0.27**	-0.38**	0.12	-0.28**	-0.47**	0.19a**
BP	-0.20**	-0.40**	0.20**	-0.25**	-0.42**	0.17a**
SF	-0.11	-0.06	-0.05	-0.09	0.01	-0.10b*
MH	0.40**	0.39**	0.01	0.37**	0.37**	-
RE	0.21**	0.27**	-0.06	0.26**	0.26**	-
VT	-0.30**	-0.33**	0.03	-0.31**	-0.31**	-
GH	-0.00	0.18*	-0.18*	-0.01	-0.15**	0.14b**
FT	-0.35**	-0.30**	-0.05	-0.32**	-0.32**	-

*n = 170; standardised mean differences of 0.2, 0.5, and 0.8 indicate small, medium, and large differences (Cohen, 1988); *p < 0.05, **p < 0.01 in paired t-test (then-test approach) or inspection of confidence intervals (SEM approach); a = recalibration response shift, b = reprioritisation response shift*

Objective 1: Comparison of then-test approach and SEM approach in the decomposition of observed change

The results of the decomposition of observed change for both the then-test approach and the SEM approach are presented in Table 5.

Observed change: The results of the observed change indicate deteriorations that are considered 'small' effects for RP, BP, VT and FT, deterioration that is considered a 'medium' effect on PF and improvements that are considered 'small' effects for MH and RE. The pattern of observed change is found to be similar for both approaches, with only small differences in the standardised mean differences.

True change: Both the then-test approach and the SEM approach also revealed a similar pattern of change for true change, except for GH. While the observed change for GH was not significant, both approaches revealed a significant true change for GH, albeit in the opposite direction. The then-test approach showed significant improvement of GH, while the SEM approach showed significant deterioration of GH.

Response shift: Both approaches revealed a significant positive response shift for BP, indicating that the true change of BP is larger than the observed change in BP. Only the SEM approach revealed a significant positive response shift for RP (resulting in a larger true change), and a significant negative response shift for SF (resulting in a smaller true change), while for the then-test approach these response shifts did not reach statistical significance. The response shifts detected for GH are in the opposite direction, with a negative response shift for GH according to the then-test approach and a positive response shift for GH according to the SEM approach, although the latter reaches a higher level of significance.

Objective 2: Tenability of assumptions underlying the then-test approach

Recall assumption: Results indicate that the differences between pre-test and then-test common factor means were non-significant ($p > .05$) for all common factors: GenPhys (0.12), GenMent (-0.04), and GenFitn (-0.02). This indicates that the assumption that respondents are able to recall their state at pre-test has been met for all SF-36 and MFI scales.

Consistency assumption: Results indicate that there are some parameters of the measurement model that are not invariant across post- and then-test measures: uniform recalibration of RP and reprioritisation of SF on GenPhys affect only the post-test, while uniform calibration of GH and reconceptualisation of VT for Genphys affect only the then-test. Also, the factor loading of GH on GenMent differs between the post- and then-test. Therefore, the second assumption is rejected for GH, RP, SF and VT.

Recalibration assumption: Results show that indeed some response shifts of the recalibration type were found (uniform recalibration of RP, BP and GH and nonuniform recalibration of VT), but that reprioritisation (of SF) and reconceptualisation (of GH on GenMent and VT on GenPhys) were also found. Therefore, the third assumption is rejected for GH, SF and VT.

Discussion

In this study a comparison was made between the then-test approach and the SEM approach in the detection of response shift in HRQL data from cancer patients undergoing invasive surgery. Results indicate that the decomposition of observed change is similar for both approaches, in that the size of true change is equal except for the direction of change in GH. The assessment of response shift differs somewhat, as only the SEM approach reveals

response shifts for RP and SF, and the response shift detected in GH reaches a higher level of significance (see Visser et al., 2005 for a substantive explanation of these differences). In a study by Ahmed, Mayo, Wood-Dauphinee, Hanley, and Cohen (2005) the then-test approach was also compared with a method that also uses SEM. They did not detect any response shift using the SEM technique, while the then-test approach did reveal several response shifts. However, an explanation for this discrepancy could be that the measurement model used in the study by Ahmed et al. was suboptimal (Borsboom, Korfage, Essink-Bot, & Duivenvoorden, 2007) and that their SEM method is not as sensitive in detecting response shift effects as our SEM approach (Ahmed, Bourbeau, Maltais, & Mansour, 2009). In the present study, we showed that it is possible to use the SEM approach to make a substantive comparison between different methodologies for the detection of response shift by looking at the decomposition of observed change into true change and response shifts.

The second objective of this study was to test the underlying assumptions of the then-test approach. Our results supported the Recall Assumption for all scales (indicating no evidence of recall bias or alternative cognitive explanations), but failed to support the assumption that internal standards of measurement are invariant across post-and then-test (Consistency Assumption rejected for four scales), and indicated that not all response shifts found were of the recalibration type (Recalibration Assumption rejected for three scales). These results are in line with a study by Nolte, Elsworth, Sinclair, and Osborne (2009) who applied SEM to assess psychometric properties of the then-test (using the Health Education Impact Questionnaire (heiQ)). They tested measurement invariance for the pre- and post-test factor model and then- and post-test factor model. They found different types of response shift for the post- and then-test, thus rejecting the Consistency Assumption for several scales, and concluded that the application of the then-test is not supported. Although their SEM approach used two models to test the underlying assumptions of the then-test approach, whereas our SEM approach consisted of a single combined model, both studies are illustrative of how the then-test can be incorporated into the SEM approach so that the underlying assumptions of the then-test approach can be evaluated. Testing the underlying assumptions of the then-test approach through SEM is useful for determining the validity of the then-test approach in assessing changes in HRQL.

If we combine our findings, we can assess the consequences of rejection of the assumptions underlying the then-test approach for the decomposition of observed change. For example, the rejection of the Recalibration Assumption (required for a valid assessment of response shift) for GH and SF coincides with a difference in assessment of response shift between the then-test approach and the SEM approach for these scales. Also, the rejection of the Consistency Assumption (required for a valid assessment of true change) for GH goes together with a difference in the assessment of true change between approaches, in that the then-test approach reveals a change in the opposite direction compared with the change detected in the SEM approach. However, the rejections of the Consistency Assumption for RP and SF do not seem to affect the assessment of true change and rejection of assumptions for VT was inconsequential for the decomposition of observed change. Concluding, the rejection of underlying assumptions is reflected in the decomposition of observed change, but the pattern is not fully consistent.

The then-test approach and SEM approach use different methods for the detection of response shift. An advantage of the then-test approach is the relatively simple analysis for detecting response shift (e.g. t-tests). However, a valid assessment of change depends on the Recall, Recalibration, and Consistency Assumption. Also, the then-test approach requires an additional assessment, which can be an extra burden to patients. Using the SEM approach, there is no need for an additional assessment and a valid assessment of change does not depend on the Recall, Recalibration, and Consistency Assumption. However, the statistical analysis for the detection of response shift is relatively complicated. Moreover, decisions on which parameters are freed (e.g. which variable shows which type of response shift) are guided not only by statistical procedures or thresholds, but also by substantive considerations. This is necessary because relying on statistics alone could lead to freeing parameters that might not be theoretically sensible (e.g. an observed variable at post-test that is an indicator of a latent variable at pre-test). Consequently, this decision involves a subjective judgment by the researcher. For example, it could be that freeing the factor loading of either one of two indicators of a latent variable yields the same result, and renders it unnecessary to free the other factor loading. This means that the researchers have to decide which of those factor loadings would be justified to free. The advantage is that response shift detection in the SEM approach is not only statistically but also theoretically driven, and will therefore lead to more logical and probable models. A disadvantage is that results depend – partly – on these subjective decisions; others may make different choices. Therefore, it should be noted that it was

not the objective of this study to draw substantive (clinical) conclusions about the response shift found. The results in this study serve for illustrative purposes only.

In conclusion, incorporating the then-test into the SEM approach: 1) allows for a comparison of the then-test approach and the SEM approach in their decomposition of observed change; 2) provides the possibility to test the underlying assumptions of the then-test approach; and 3) gives an idea of the consequences of rejection of underlying assumptions on the decomposition of observed change. To be able to draw valid conclusions in the assessment of HRQL, we need to be aware of the limitations of HRQL measurement. Quantifying the existence and size of response shift and true change will help to better understand the observed change of HRQL. Future research should focus not only on validating the measurements of HRQL, but also on investigating the (clinical) consequences of violating the validity on the change assessments.

Author's note

This research was supported by the Dutch Cancer Society (KWF grant 2011-4985).

References

Aaronson, N. K., Muller, M, Cohen, P. D. A., et al. (1998). Translation, validation, and norming of the Dutch language version of the SF-36 health survey in community and chronic disease populations. *Journal of Clinical Epidemiology, 51*, 1055-1068.

Ahmed, S., Mayo, N. E., Wood-Dauphinee, S., Hanley, J. A., & Cohen, S. R. (2005). The structural equation modeling technique did not show a response shift, contrary to the results of the then test and the individualized approaches. *Journal of Clinical Epidemiology, 58*, 1125-1133.

Ahmed, S., Bourbeau, J., Maltais, F., & Mansour, A. (2009). The Oort structural equation modeling approach detected a response shift after a COPD self-management program not detected by the Schmitt technique. *Journal of Clinical Epidemiology, 62*, 1165-1172.

Allison, P. J., Locker, D., & Feine, J. S. (1997). Quality of life: A dynamic construct. *Social Science & Medicine, 45*, 221-230.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*, 238-246.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Borsboom, G. J. J. M, Korfage, I. J., Essink-Bot, M., & Duivenvoorden, H. J. (2007). The structural equation modeling technique did not show a response shift, contrary to the results of the then test and the individualized approaches. *Journal of Clinical Epidemiology, 60*, 426-427.

Browne, M. W., & Cudeck, R. (1989). Single sample cross-validation indices for covariance structures. *Multivariate Behavioral Research, 24*, 445-455

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods Research, 21*, 230-258.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Erlbaum: Hillsdale, NJ.

Golembiewski, R. T., Billingsley, K., & Yeager, S. (1976). Measuring change and persistence in human affairs: Types of change generated by OD designs. *Journal of Applied Behavioral Science, 12*, 133-157.

Howard, G. S., Ralph, K. M., Gulanick, N. A., Maxwell, S. E., Nance, S. W., & Gerber, S. K. (1979). Internal invalidity in pretest-posttest self-report evaluations and reevaluation of retrospective pretests. *Applied Psychological Measurement, 3*, 1-23.

Jöreskog, K. G., & Sorbom, D. (1996). *LISREL 8 users' guide* (2nd ed.). Chcago, IL: Scientific software international, Inc.

Neale, M. C., Boker, S. M., Xie, G., & Maes, H. H. (1999). *Mx: Statistical Modeling* (5th ed.). Richmond, VA: Department of Psychiatry.

Nolte, S., Elsworth, G. R., Sinclair, A. J., & Osborne, R. H. (2009). Tests of measurement invariance failed to support the application of the 'then-test'. *Journal of Clinical Epidemiology, 62*, 1173-1180.

Oort, F. J. (2005). Using structural equation modeling to detect response shifts and true change. *Quality of Life Research, 14*, 587-598.

Oort, F. J., Visser, M. R. M., & Sprangers, M. A. G. (2005). An application of structural equation modeling to detect response shifts and true change in quality of life data from cancer patients undergoing invasive surgery. *Quality of Life Research, 14*, 599-609.

Schwartz, C. E., & Sprangers, M. A. G. (1999). Methodological approaches for assessing response shift in longitudinal health-related quality-of-life research. *Social Science & Medicine, 48*, 1531-1548.

Schwartz, C. E., Sprangers, M. A. G., Oort, F. J., Ahmed, S., Bode, R., Li, Y., & Vollmer, T. (2011). Response shift in patients with multiple sclerosis: an application of three statistical techniques. *Quality of Life Research, 20*, 1561–1572.

Smets, E. M. A., Garssen, B., Bonke B., & De Haes, J. C. J. M. (1995). The multidimensional fatigue inventory (MFI): Psychometric qualities of an instrument to assess fatigue. *Journal of Psychosomatic Research, 39* (3), 315-325.

Sprangers, M. A. G., & Schwartz, C. E. (1999). Integrating response shift into health-related quality of life research: a theoretical model. *Social Science & Medicine, 48*, 1507-1515.

Visser, M. R. M., Oort, F. J., & Sprangers, M. A. G. (2005). Methods to detect response shift in quality of life data: A convergent validity study. *Quality of Life Research, 14*, 629-639.

Ware, J. E., Snow, K. K., Kosinski M., & Gandek, B. (1993). *SF-36 health survey: Manual and interpretation guide*. Boston, MA: The Health Institute, New England Medical Center.

MATHILDE G. E. VERDAM
PhD Candidate at the Department of Child Development and Education and the Academic Medical Centre, University of Amsterdam. Research interests: Structural Equation Modelling, longitudinal measurement invariance, response shift, quality of life.

PROF. DR. FRANS J. OORT
Professor of Methods and Statistics in Educational Research, director of the Graduate School of Child Development and Education, and program director of the Research Master Educational Sciences at the University of Amsterdam.

DR MECHTELD R.M. VISSER
Researcher and quality manager at the Academic Medical Centre, University of Amsterdam.

PROF. DR. MIRJAM A. G. SPRANGERS
Professor in Medical Psychology. Her research interests include patient-reported outcomes (e.g., quality of life, health care needs), response shift (i.e. shifts in patients' perspective over time), and genetic disposition of quality of life.

Causal directions between adolescents' externalising and internalising problems: A continuous-time analysis

It is generally acknowledged that internalising and externalising problems are closely related and often co-occur. This comorbidity may result from various patterns of influence between internalising and externalising symptoms. The cross-lagged panel model and the latent growth curve model have become popular tools to assess the causal pathways between both types of problem behaviour in a non-experimental context. The present study shows, however, that both methods have serious limitations that do not allow for definite conclusions concerning the causal direction between externalising and internalising problems. The present article describes a continuous-time procedure for analysing cross-lagged panel data by means of structural equation modelling (SEM) that circumvents these limitations: The ADM/SEM procedure using the approximate discrete model (ADM). The procedure is applied to the analysis of three waves of annually collected self-report data on adolescents' externalising and internalising problems. Results suggest a unidirectional effect from externalising problems to internalising problems, thus providing support for the failure perspective. Implications for theory, research, and practice are discussed.

Where: Netherlands Journal of Psychology, Volume 67, 68-80

Received: 6 July 2012; Accepted: 20 November 2012

Keywords: Continuous-time modelling; Cross-lagged panel model; Approximate discrete model; Latent growth curve model; Stochastic differential equations; Structural equation modeling

Authors: Marc J. M. H. Delsing* and Johan H. L. Oud**

*Praktikon, Radboud University Nijmegen, the Netherlands
**Behavioural Science Institute, Radboud University Nijmegen, the Netherlands

Correspondence to:
Marc J. M. H. Delsing, Praktikon, Radboud University Nijmegen, PO Box 6909, 6503 GK Nijmegen, the Netherlands, e-mail: m.delsing@acsw.ru.nl.

A vast body of research has shown that externalising and internalising problems are closely related and often co-occur (Beyers & Loeber, 2003; Gilliom & Shaw, 2004; Lilienfeld, 2003; Overbeek et al., 2006). Although comorbidity has been reported in many studies, the nature of the association between externalising and internalising symptoms is not yet fully understood (Lee & Bukowski, 2012). It is important to know what causal mechanisms account for the comorbidity of externalising and internalising problems in order to develop effective interventions.

Several possible pathways have been proposed that may explain the association between externalising and internalising problems. In attempts to find empirical evidence for these pathways, researchers have generally employed two types of longitudinal models: cross-lagged panel models (e.g., Overbeek, Vollebergh, Meeus, Luijpers, & Engels, 2001) and latent growth curve models (Lee & Bukowski, 2012). Below, we will review findings from research using either type of model and point out that, although these longitudinal approaches clearly have important advantages over cross-sectional designs, they are still problematic in several ways. Next we will demonstrate how continuous-time analysis can solve most of these problems. Our general aim is to encourage applied researchers to use this approach.

Explanations for the co-occurrence of externalising and internalising problems

Directional explanations of the association between internalising and externalising argue that

this association may be the result of three causal pathways. According to the failure perspective (see e.g., Burke, Loeber, Lahey, & Rathouz, 2005; Capaldi, 1992), conduct problems may lead to internalising problems. Conduct problems are assumed to lead to failures in social situations that, in turn, lead to depression and anxiety. There is also literature suggesting effects in the opposite direction, that is from internalising problems to externalising problems. The theory of masked depression, for example, suggests that depressive symptoms lead to acting out behaviours, as children express their underlying depression by acting out (Glaser, 1967). Depression may impair individuals' concern about the adverse consequences of their actions, thereby increasing the risk for certain forms of antisocial behaviour (Capaldi, 1991). Finally, according to the mutual influence perspective, externalising problems lead to internalising problems and vice versa. In addition to these directional models, common vulnerability models (Jackson & Sher, 2003) assume no influences between externalising and internalising problems, but believe that the impact of non-specific (i.e., shared or overlapping) risk factors accounts for the co-occurrence of externalising and internalising problems (Overbeek et al., 2001). Below, we give a short description of the cross-lagged panel model and the latent growth curve model, as well as a short review of the findings emerging from their application to the analysis of reciprocal relations between externalising and internalising problems.

Cross-lagged panel models

Cross-lagged panel models circumvent the difficult problem of assessing causal direction in cross-sectional research as the causal direction in cross-lagged panel models is not based on instantaneous relations between simultaneously measured variables x and y. Instead, different variables are used for opposite directions, in this case externalising problems at time point 1 affecting internalising problems at time point 2, and internalising problems at time point 1 affecting externalising problems at time point 2. Within the cross-lagged panel model, both problem behaviour variables at one measurement time point are regressed on their own lagged score plus the lagged score of the other problem behaviour variable at the previous measurement time point (Delsing & Oud, 2008). The resulting cross-lagged coefficients inform about the causal direction between both problem behaviours. Although it is not impossible to add the mean structure to the analysis, usually only the covariance structure is analysed in cross-lagged panel analyses and not the mean structure.

Mixed results have emerged from studies using cross-lagged panel models. Using data from a four-wave longitudinal study with six-month intervals, Wiesner (2003) found a relatively small unidirectional effect from delinquency to depression for middle adolescent boys, whereas bidirectional effects were found for girls. Using a cross-lagged design with a two-year interval in a sample of middle adolescents, Ritakallio et al. (2008) found that depression predicted subsequent antisocial behaviour among girls, but conversely, antisocial behaviour did not predict subsequent depression. Surprisingly, a negative effect was found from depression to antisocial behaviour among boys, suggesting that depression protects from subsequent antisocial behaviour. Hipwell et al. (2011) used a cross-lagged design with nine waves of annually collected data (ages 8 through 16 years) and found that conduct disorder often preceded depression across this developmental period, although the effect sizes were small. There was less consistent prediction from depression to conduct disorder. Vieno, Kiesner, Pastore, and Santinello (2008) used a cross-lagged panel model with a ten-month interval in a sample of early adolescents and found that depressive symptoms predicted later antisocial behaviour, but that antisocial behaviour did not predict later depression. In one of their models, Vieno et al. (2008) estimated both a cross-lagged and instantaneous (i.e., going from one problem behaviour variable to the other at the same time point) effect from depression to antisocial behaviour. Only the instantaneous effect was found to be significant. Curran and Bollen (2001) used a four-wave cross-lagged panel model with two-year intervals and found antisocial behaviour to positively predict later depressive symptoms, but earlier depressive symptoms did not predict later antisocial behaviour. Finally, using cross-lagged panel analyses with two-year intervals in a community sample of adolescents and young adults, Overbeek et al. (2001) found that a stability model with no cross-lagged relations between emotional disturbance and delinquency fits best for the total sample, as well as across age and gender categories. This led them to conclude that the co-occurrence of emotional disturbance and delinquency during adolescence and young adulthood seems to result from associated but separate psychopathological processes.

Latent growth curve models

Reciprocal associations between externalising and internalising problems have also been analysed within a latent growth curve framework (see e.g., Gilliom & Shaw, 2004; Lee & Bukowski, 2012). In contrast to the cross-lagged panel models in which

prior values of variables determine the current value of the same or other variables, the latent growth curve model specifies separate trajectories over time for separate variables and separate cases. Each case in the sample can have a different time trend as marked by a different intercept or slope when tracked over time (Rao, 1958; Tucker, 1958; Meredith & Tisak, 1990). The latent intercept corresponds to the $t = 0$ value of the individual longitudinal curve, while the latent slope reflects the individual change rate over time. Latent growth curve models also inform about the variances of the latent intercepts and slopes, indicating the amount of inter-individual differences in the $t = 0$ values and in the longitudinal change process, respectively. In contrast with the typical cross-lagged panel model, the latent growth curve model also analyses the mean structure.

Important in this respect is that latent growth curve models enable researchers to test for correlations between the latent intercept and the latent slope. Thus, one may find out whether the change in a given construct (e.g., externalising problems) is related to its initial value (if $t = 0$ is located at the initial time point). Reciprocal associations between externalising and internalising problems are investigated by means of the simultaneous specification and estimation of latent growth curve models for each of the problem behaviour variables. Latent growth factors (intercept and slope) for both problem behaviour variables and the relations between the intercepts of the problem behaviour variables, between the slopes, and between the intercepts and slopes are evaluated. Somewhat similar to the cross-lagged design in which one problem behaviour variable at one point in time is used to predict the other problem behaviour variable at a later point in time, authors have used the intercept of externalising problems as a predictor of the change rate (slope) in internalising problems, and the intercept of internalising problems as a predictor of the change rate in externalising problems (Oud, 2010).

Applying latent growth curve analysis, Gilliom and Shaw (2004) found that initial values of mother-reported externalising problems were related to steeper increases in mother-reported internalising problems over time with boys followed from age 2 to 6. Similarly, Keiley, Bates, Dodge, and Pettit (2000) found that children with a relatively high initial status on teacher-reported externalising behaviours in kindergarten were seen by later teachers as becoming increasingly internalising. Using a sample of South Korean fourth graders, Lee and Bukowski (2012) also found that higher initial levels of externalising problems were related to steeper increases in internalising problems over time, but such a unidirectional relationship was only found for girls. Among boys, initial levels of externalising and internalising problems were found to be related to the developmental pattern of the other domain (i.e., internalising and externalising problems, respectively), thus suggesting a bidirectional pattern of influence. A similar bidirectional pattern was reported by Measelle, Stice, and Hogansen (2006), who found that initial depressive symptoms predicted future increases in antisocial behaviour and that initial antisocial symptoms predicted future increases in depressive symptoms in a community sample of adolescent females who were followed annually from early to late adolescence. In the Curran and Bollen (2001) study referred to above, the authors also applied a latent growth curve model to the analysis of their data. A significant positive relation was found between the depressive symptoms intercept and the antisocial slope, suggesting that individual differences in depressive symptoms are positively associated with antisocial behaviour increases over time. In addition to the cross-lagged panel model and the latent growth curve model, Curran and Bollen also used a model combining both cross-lagged associations with associations at the level of the growth factors. Findings from this hybrid so-called autoregressive latent trajectory (ALT) model again revealed earlier levels of antisocial behaviour to predict later levels of depressive symptoms as well as a positive relation between the depressive symptoms intercept and the antisocial slope.

Altogether, empirical support for externalising problems leading to internalising problems and internalising problems leading to externalising problems is inconsistent. Accordingly, we do not know which causal direction is more likely to occur. As we will demonstrate below, these inconsistencies may partly be due to methodological limitations of the cross-lagged panel model as it is typically used, namely in its discrete-time version. We will also see that substantive interpretations of the associations between the intercept and slope factors in latent growth curve models leave much to be desired. Below, we will first describe shortcomings of the discrete-time cross-lagged panel model and the latent growth curve model, respectively, for establishing the causal direction between externalising and internalising problems. Next, we present a continuous-time version of the cross-lagged model that circumvents these shortcomings. This procedure is applied to the analysis of reciprocal relations between externalising and internalising problems.

Limitations of discrete-time cross-lagged panel models

Although externalising and internalising problems influence themselves and each other continuously over time, researchers are forced to use 'snapshots' of this developmental process in order to learn something about the underlying continuous-time process (Voelkle, Oud, Davidov, & Schmidt, 2012). Longitudinal designs, in which the same subjects are repeatedly observed across time, are typical examples of such 'snapshots.' In such designs, measurements are typically taken not more than once or twice a year, resulting in relatively large observation intervals. The challenge for applied researchers is then to obtain an estimate of the underlying continuous-time effect that adequately reflects the underlying continuous-time process. Below, we will see that discrete-time cross-lagged effects do a rather poor job in this respect. As a consequence, discrete-time modelling is an oversimplification and often a distortion of reality.

A serious limitation of discrete-time cross-lagged effects is that they are highly dependent on the length of the discrete-time observation intervals (Delsing, Oud, & De Bruyn, 2005; Delsing & Oud, 2008; Oud, 2007; Oud & Delsing, 2010; Voelkle et al., 2012). In general, cross-lagged effects have a value of 0 over a zero time interval (different variables cannot yet have any influence on each other over a zero time interval), increase more or less rapidly with increasing intervals until a maximum is reached, and eventually return to 0 with further increasing intervals. Most of the studies referred to above used different measurement intervals (e.g., 6 months, 10 months, 1 year), which makes their outcomes incomparable and often contradictory. Even conclusions regarding cross-lagged effects across equal measurement intervals within or across studies are problematic since these effects may be totally different, and may even change sign, across other measurement intervals. They may also provide a totally different picture than the underlying continuous-time effects (Oud, 2007). This makes discrete-time analysis useless for establishing the causal directions between externalising and internalising problem behaviour.

The final problem concerns the possibility to analyse cross-lagged effects (e.g., externalising problems at time point 1 affecting internalising problems at time point 2, internalising problems at time point 1 affecting externalising problems at time point 2) as well as instantaneous cross-effects (i.e., externalising problems at time point 2 affecting internalising problems at time point 2 and vice versa) in the cross-lagged panel model for the same data. The study by Vieno et al. (2008) referred to above, for example, which simultaneously estimated a cross-lagged effect and instantaneous effect from depression to antisocial behaviour, found that the results for both kinds of effects differed and only the instantaneous effect was significant. In discrete time, there is no way to relate the two different sets or to combine them in a unitary, unequivocal measure for the underlying causal effects.

Limitations of latent growth curve models

In studies using latent growth curve models, conclusions regarding reciprocal associations between externalising and internalising problems are based upon the associations between the intercept factor of one problem behaviour variable and the slope factor of the other problem behaviour variable. Such conclusions are highly problematic, however, since both the intercept and slope are time-unspecific (Bollen & Curran, 2004; Delsing & Oud, 2008). Note the fundamental difference in this respect with the cross-lagged parameters in the cross-lagged panel model, which reflect the connection between specific temporally ordered moments in time. Causal mechanisms can be characterised as 'time-specific' or, as it is called in the state-space literature, 'nonanticipative' (Oud, 2010). Because of their time-unspecific character, causal interpretations of intercept-slope associations are impossible. Therefore, these associations cannot provide any 'support' for causal connections. Another factor complicating the interpretation of the association between the intercept and slope factor is that this association is highly dependent on the coding of time, as reflected by the factor loadings of the slope factor. This has long been recognised for the association between the intercept and slope factor in the same variable (Biesanz, Deeb-Sossa, Papadakis, Bollen, & Curran, 2004; Bollen & Curran, 2006; Mehta & West, 2000). Recently, however, Oud (2010) has proven that this also holds for associations between the intercept in one variable (e.g., externalising problems) and the slope in a different variable (e.g., internalising problems). In fact, Oud (2010) showed that by shifting the time scale almost any covariance and correlation value can be reached. This dependence on the choice of the zero time point in the time scale renders any substantive interpretation of the intercept-slope covariances doubtful.

Continuous-time analysis

The problems discussed above associated with the use of discrete-time cross-lagged panel models and latent growth curve models can be solved by using a continuous-time version of the cross-lagged panel

model[1]. Oud (2007) discusses two continuous-time procedures using structural equation modelling (SEM): the exact discrete model (EDM/SEM) and the approximate discrete model (ADM/SEM). The EDM, introduced in 1961-1962 by Bergstrom (1988), links in an exact way the discrete-time model parameters to the underlying continuous-time model parameters by means of nonlinear restrictions. The link is made by solving the stochastic differential equation for the discrete-time interval. Crucial for the EDM is the exact solution, that is, the unique solution satisfying the differential equation. Mathematical details and derivations for the EDM/SEM procedure can be found in Oud and Jansen (2000) and in Oud and Singer (2008). A general way to estimate a stochastic differential equation model is by applying the nonlinear constraints of the EDM during estimation by means of an appropriate nonlinear SEM software package such as Mx (Neale, Boker, Xie, & Maes, 1999) or OpenMx (Boker et al., 2011). This method can be applied in models with either equal or unequal measurement intervals as well as in models with time-varying parameters.

The solution of the problem referred to above regarding the two sets of coefficients in a cross-lagged analysis (i.e., instantaneous and cross-lagged coefficients) gives rise to an alternative model: Bergstrom's (1966, 1984) approximate discrete model (ADM). Whereas the exact nonlinear constraints of the EDM require SEM programs possessing the exponential function (e.g., Mx and OpenMx), the ADM can also be applied in less nonlinearly oriented SEM programs such as LISREL (Jöreskog & Sörbom, 1996), AMOS (Arbuckle, 2007), EQS (Bentler, 2006), or Mplus (Muthén & Muthén 1998). Like the EDM/SEM procedure, the ADM/SEM procedure is applicable for models with unequal observation intervals but not for models with continuously time-varying parameters.

According to Bergstrom's ADM (Bergstrom, 1966; 1984, pp. 1172–1173), the simple linear constraints

$$\mathbf{A}_{ins} = .5\ \tilde{\mathbf{A}}\,\Delta t$$

$$\mathbf{A}_{lag} = \mathrm{I} + .5\ \tilde{\mathbf{A}}\,\Delta t$$

lead to reasonable, so-called 'trapezoid' (Gard, 1988, pp. 192), approximations of time-invariant continuous-time parameters. $\mathbf{A}_{ins}$ refers to the instantaneous effects matrix, $\mathbf{A}_{lag}$ refers to the lagged effects matrix, I refers to the identity matrix of which diagonal elements equal 1 and off-diagonal elements equal zero, and $\tilde{\mathbf{A}}$ refers to the approximate continuous-time effects matrix. Suppose the observation intervals Δt are equal and for convenience set at $\Delta t = 1$, then one needs only to constrain the off-diagonal elements in $\mathbf{A}_{ins}$ (i.e., instantaneous cross-effects) and $\mathbf{A}_{lag}$ (i.e., cross-lagged effects) to be equal and each diagonal element in $\mathbf{A}_{lag}$ (i.e., autoregression effects) to be 1 plus the corresponding diagonal element in $\mathbf{A}_{ins}$ (i.e., self-loop effects), while one computes $\tilde{\mathbf{A}}$ (i.e., approximate continuous-time effects) as $2\mathbf{A}_{ins}$. The easy implementation of the linear constraints makes the ADM a feasible and attractive alternative to the EDM for comparing the results of different observation intervals between and within studies and for solving the problem of 'contradictory' results between instantaneous and lagged coefficients (e.g., Vieto et al., 2008). Interestingly, the ADM is one of the few cases in SEM where the self-loop coefficients (diagonal element in $\mathbf{A}_{ins}$) are estimated instead of being specified to be zero. Because of its simplicity and easy implementation, the present study applies the ADM procedure. By doing so, we want to encourage applied researchers to start using this approach in their own work.

Method

Participants and Procedure

The data were taken from a more comprehensive Dutch study of family relationships and adolescent problem behaviour (the Nijmegen Family and Personality Study; Haselager & Van Aken, 1999). The participants were 280 adolescents (140 boys, 140 girls) who were 14.5 years old on average (ranging from 11.4 to 16.0) at the first measurement wave. Further details regarding sample characteristics and procedure can be found in Delsing, Van Aken, Oud, De Bruyn, & Scholte (2005).

Measures

To assess adolescents' externalising and internalising problem behaviour, four scales of the Nijmegen Problem Behaviour List (NPBL; Scholte, Vermulst, & De Bruyn, 2001) were used at each of the three annual measurement waves: Aggressive and Delinquent Behaviour problems (together Externalising), and Withdrawn and Anxious/Depressed Behaviour problems (together Internalising). Each scale consists of five items. The structure of the NPBL was modelled according to

[1] The problems discussed with regard to the discrete-time cross-lagged panel model and the latent growth curve model equally apply to the ALT model (Curran & Bollen, 2001), which is a synthesis of both models. Delsing and Oud (2008) and Oud (2010) demonstrated how these and other problems associated with the ALT model can be solved by, respectively, a continuous-time autoregressive latent trajectory (CALT) model and a second-order stochastic differential equation model.

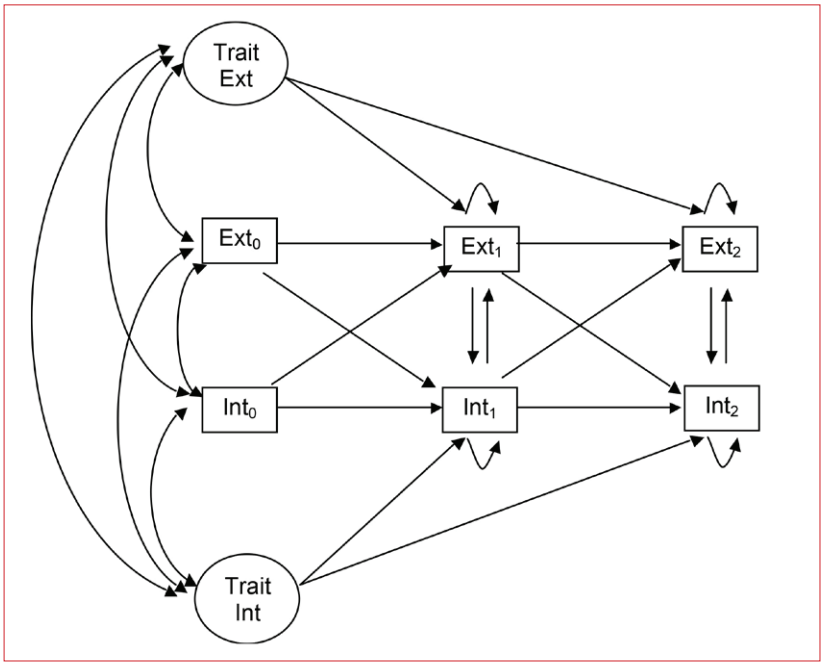

Figure 1 The three-wave ADM/SEM cross-lagged panel model for adolescents' externalising and internalising problem behaviours, including random subject effects
Ext = Externalising problem behaviour, Int = Internalising problem behaviour

the Child Behaviour Checklist (CBCL; Verhulst, Van der Ende, & Koot, 1996). However, in contrast with the CBCL, the NPBL focuses on subclinical instead of clinical problem behaviour. The items represent the most common problems in adolescence that cause some concern, but are not serious enough for referral. Examples of items are: 'This person does things that could get him/her into trouble with the law' for Delinquency; 'This person fights a lot' for Aggressive Behaviour; 'This person would rather be alone than with other people' for Withdrawn; 'This person feels sad and unhappy' for Anxious/Depressed Behaviour. Adolescents were asked to indicate on five-point Likert scales ranging from 1 (not at all true) to 5 (completely true) the extent to which each item was true. Internal consistencies were .80 and .83 with regard to externalising problems and internalising problems, respectively.

Data analysis

Because applied researchers are generally more familiar with linearly oriented SEM programs such as LISREL (Jöreskog & Sörbom, 1996) or Mplus (Muthén & Muthén, 1998) which, however, do not provide the exponential function with matrix algebraic means necessary for applying the EDM/SEM procedure, we applied the ADM/SEM procedure to establish the reciprocal effects between adolescents' externalising and internalising problem behaviours. In addition to the ADM continuous-time restrictions described above, two other adaptations were made to the 'standard' cross-lagged panel model, namely, the addition of intercepts and random subject effects. Intercepts accommodate for the frequently observed nonzero and nonconstant mean trajectories. By means of the specification of random subject effects, subject specific conditional mean trajectories are obtained, each keeping a subject specific distance from the sample mean trajectory. The zero mean normally distributed random subject effects can be viewed as a special kind of (unobserved and constant over time) state variables, sometimes called 'trait' variables. The additions of nonconstant means and trait variables to the 'standard' cross-lagged panel model are highly relevant in behavioural science. The very concept of development implies nonconstant means and developmental curves of different subjects rarely can be assumed to coincide or even to follow parallel paths (Oud, 2007).

Figure 1 shows the model that was estimated. The model contains the state variables Ext and Int and corresponding constant trait variables Trait-Ext and Trait-Int which, because of the number of time points being 3, leads to a total of 8 variables in the structural equation model. In total, 21 parameters had to be estimated:

- 4 continuous-time drift coefficients
- 2 initial latent means
- 3 initial state variances and covariances
- 2 intercepts feeding changes in mean development
- 3 trait variances and covariances
- 4 covariances between traits and initial states
- 3 state variable disturbance variances and covariances.

There are 6 observed variables or 6 observed means and 21 (distinct) elements in the observed covariance matrix, resulting in (21 + 6) – 21 = 6 degrees of freedom for the SEM model. The latent means for the trait variables are 0 by definition. Trait variables, constant over time but varying over subjects, accommodate for deviations of subject specific developmental curves from the mean curve. Effects from the trait variables to Ext and Int at waves 2 and 3 are fixed at 1. Measurement error variances were fixed at zero. Parameter estimates pertaining to the first measurement interval (t_0 to t_1) were set equal to corresponding estimates pertaining to the second measurement interval (t_1 to t_2). See the Appendix for the LISREL script for ADM/SEM procedure.

Results

In Table 1, the estimation results for the ADM are given. With regard to the four drift coefficients (auto- and cross-effects), an important difference in interpretability exists between the auto-effects

Table 1 Parameter estimates and model fit information for the approximate discrete model (ADM)

Parameter	ADM
$Ext_1 \rightarrow Ext_1$ (self-loop)	-0.780**
$Ext_1 \rightarrow Int_1$	0.791**
$Int_1 \rightarrow Ext_1$	0.347
$Int_1 \rightarrow Int_1$ (self-loop)	-1.058**
Variance Ext_0	27.281**
Variance Int_0	38.195**
Covariance Ext_0-Int_0	11.007**
Error variance Ext_1	24.071**
Error variance Int_1	40.463**
Error covariance Ext_1-Int_1	6.590**
Mean Ext_0	17.943**
Mean Int_0	21.103**
Mean change coefficient Ext	7.632
Mean change coefficient Int	5.581
Variance Trait-Ext	10.721
Variance Trait-Int	26.237
Covariance Trait-Ext Trait-Int	-15.029
Covariance Trait-Ext Ext_0	7.022
Covariance Trait-Ext Int_0	-5.495
Covariance Trait-Int Int_0	14.777
Covariance Trait-Int Ext_0	-12.769
χ^2	5.389
Df	6
RMSEA	.000

*$p \leq .05$; **$p \leq .01$

of externalising problems ($Ext_1 \rightarrow Ext_1$) and internalising problems ($Int_1 \rightarrow Int_1$) on the one hand and the cross-effects from externalising problems to internalising problems ($Ext_1 \rightarrow Int_1$) and from internalising problems to externalising problems ($Int_1 \rightarrow Ext_1$) on the other hand. The auto-effects are scale free in the sense that they do not change under arbitrary linear transformations of Ext and Int and so are directly interpretable. In particular, both Ext and Int show negative feedback (-0.780 and -1.058), implying stability or a tendency for individuals to converge to the subject specific mean trajectories. The continuous-time auto-effects of -0.780 and -1.058 in the ADM transform for $\Delta t = 1$ to autoregressive coefficients of 0.458 and 0.347 for externalising and internalising problems, respectively. To become comparable, the cross-effects, not being scale free, have been standardised by multiplying by the ratios of the initial standard deviations. The standardised values of 0.791 ($p < .01$) and 0.347 ($p > .05$) in Table 1 seem to reveal the existence of a unidirectional effect from externalising problems to internalising problems. The longitudinal influence from externalising problems to internalising problems goes together with a moderate cross-sectional correlation of 0.341 between the two variables at the first time point (implied by covariance 11.007 in Table 1 and variances 27.281 and 38.195 for externalising and internalising problems, respectively).

Autoregression and cross-lagged coefficient functions, based on the differential equation model

An interesting feature of the continuous-time modelling approach is that, on the basis of the continuous-time effects, one can assess the discrete-time effects as a function of the measurement interval. The autoregressive functions in Figure 2 and cross-lagged effect functions in Figure 3 are based on the estimates of the continuous-time auto-effects and cross-effects resulting from an EDM analysis of the cross-lagged panel model for externalising and internalising problem behaviours. The autoregression functions show the autoregression values (for $\Delta t = 1$ being 0.458 for Ext and 0.347 for Int) as part of an ongoing process. Starting from 1 (autoregression between a variable and its lag, when the interval length $\Delta t = 0$), they display the autoregression over increasing intervals. It turns out that both Ext and Int go down rather rapidly, and after $\Delta t = 1$ (the one-year period between the first and second wave) less than half of the autoregression at the start is left. The autonomous decrease in Ext appears to be somewhat slower than that of Int. So, Ext turns out to be a somewhat more persistent property across time than Int.

Figure 3 shows the standardised cross-lagged effects between externalising and internalising problems as a function of the time interval. It is clear that the standardised effect of Ext on Int (standard deviation unit increase in Int as a result of an isolated increase of 1 standard deviation in Ext) is a) much stronger than in the opposite direction over quite a long period of time, b) builds up rather rapidly until it reaches its maximum value of 0.386 shortly after 1 year and c) goes down rather rapidly afterwards, having a standardised effect of less than one third of its maximum level after 5 years.

Interpolated and predicted mean development of Ext and Int autoregression

One may wonder what the interplay between Ext and Int leads to in terms of interpolated and predicted mean development. This is shown in Figure 4. The model implies a rather flat pattern, indicating an extremely slow, virtually linear, decrease in both Ext and Int. The interpolated decrease in Ext over the five-year period is from 17.943 to 17.687. The interpolated decrease in Int over the five-year period is even smaller, and goes from 21.103 to 20.945.

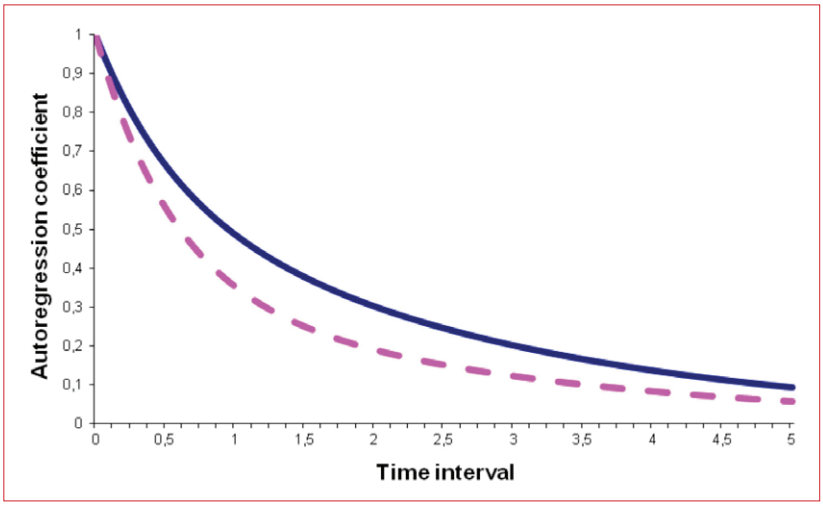

Figure 2 Autoregression functions of Externalising problem behaviour (solid line) and Internalising problem behaviour (dotted line)

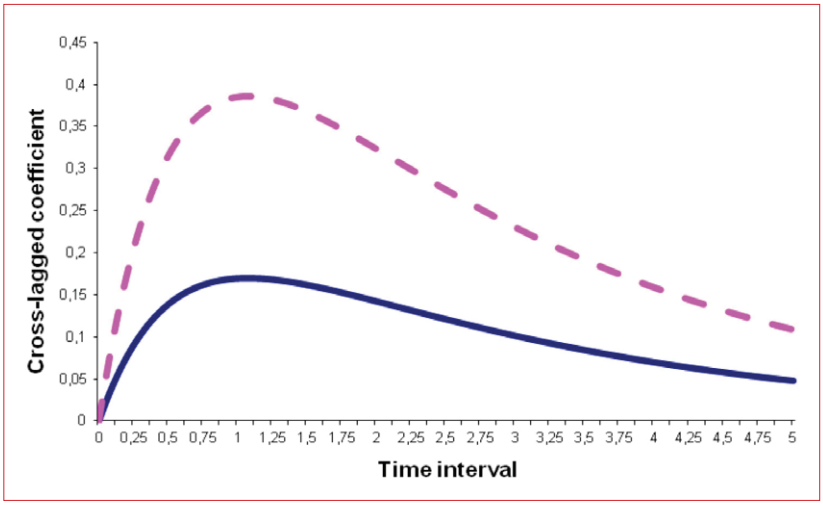

Figure 3 Cross-lagged effect functions of Internalising problem behaviour on Externalising problem behaviour (solid line) and of Externalising problem behaviour on Internalising problem behaviour (dotted line)

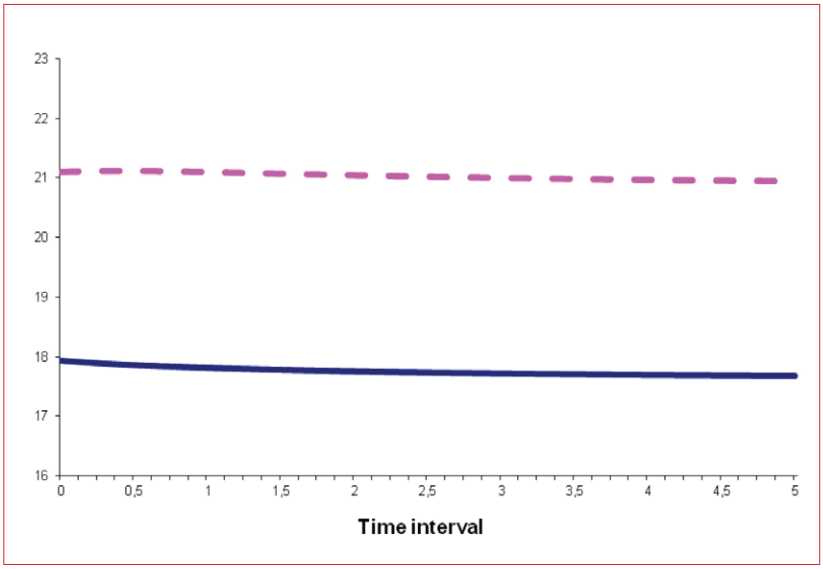

Figure 4 Interpolated and predicted mean development of Externalising problem behaviour (solid line) and Internalising problem behaviour (dotted line) over 5 years

Discussion

The purpose of the present study was to investigate, in continuous time, the causal influences between adolescents' externalising and internalising problem behaviour. We started by pointing out several problems associated with the methods typically used to shed light on this issue, namely, the discrete-time cross-lagged panel model and the latent growth curve model. These problems arise mainly from the effects in the cross-lagged panel model being highly dependent on the length of the measurement interval, and from the time-unspecific character of the intercept and slope factors in the latent growth curve model. Moreover, the association between the intercept and slope factor is an artifact of the time scale used and contains little empirical value. We demonstrated how continuous-time analysis of the cross-lagged panel model circumvents these problems.

Application of the ADM/SEM revealed a unidirectional effect from externalising problems to internalising problems, thus providing support for the failure perspective (see e.g., Burke et al., 2005, Capaldi, 1992). A specific cascade model that has been proposed to account for the cross-over effect of externalising problems to internalising problems is the so-called dual failure model (Patterson & Capaldi, 1990; Patterson, Reid, & Dishion, 1992). According to this model, externalising problems hamper successful development of two key domains of competence, namely peer relations and academic performance. In turn, difficult peer relationships and poor academic development lead to an increase in internalising problems. Empirical support for this model was provided by Van Lier et al. (2012). In their recent cross-lagged panel study in which externalising and internalising problems, peer victimisation, and school achievement were assessed annually in a sample of early elementary school aged children, externalising problems were found to lead to academic underachievement and experiences of peer victimisation. Academic underachievement and peer victimisation, in turn, predicted increases in externalising problems. These findings applied equally to boys and girls. In accordance with the present study, no links from internalising to externalising problems were found. These findings suggest that interventions should target adolescents' externalising problems in order to prevent spillover to academic achievement and peer functioning and, eventually, internalising problems. Interventions could also focus directly at academic achievement and peer functioning in order to 'eliminate' the mediating link between externalising and internalising problems.

Because no studies to date have applied a continuous-time approach to the analysis of the bidirectional influences between externalising and internalising problems, it is difficult to compare our results to previous findings. The discrete-time effects reported in previous studies do not inform directly about the underlying continuous-time effects. As noted before, contradictory results may be obtained when translating discrete-time effects to their underlying continuous-time effects.

In spite of its innovative methodology to unravel the causal processes behind the co-morbidity of adolescents' externalising and internalising problems, the present study has several limitations. First, we investigated bidirectional associations between externalising and internalising problems in a community sample of adolescents. We do not know to what extent our findings generalise to other age groups. The failure experiences that are supposed to mediate the link between externalising and internalising problems may be age-graded and may be more or less important across varying developmental stages (Wiesner, 2003). The results of Van Lier et al.'s (2012) study reported above, however, suggest that failure experiences may also be relevant as a mediator in younger children. Furthermore, causal influences may be different in referred adolescents. Our findings regarding the co-occurrence of sub-clinical levels of externalising and internalising problems cannot automatically be generalised to psychiatric disorders. Therefore, future studies should attempt to corroborate our findings in other age groups and clinical samples.

Second, rather than the EDM, we applied the ADM procedure by means of which we obtained approximate parameter estimates of the continuous-time cross-effects. The EDM, introduced in 1961–1962 by Bergstrom (1988), links in an exact way the discrete-time model parameters to the underlying continuous-time model parameters by means of nonlinear restrictions. An advantage of the ADM, however, is that it allows less nonlinearly oriented SEM programs such as LISREL (Jöreskog and Sörbom, 1996) or Mplus (Muthén & Muthén, 1998) to be used in parameter estimation. As we have seen, the ADM utilises only simple linear restrictions to approximate the differential equation model. We chose to use the ADM because we want to encourage applied researchers to use this procedure, and anticipated most of them would be familiar with popular programs such LISREL or Mplus. In a simulation study comparing the EDM and ADM procedure, Oud (2007) demonstrated that the ADM performed about equally well in overall quality as the EDM, and even defeated it in models with trait variables and samples $N \leq 400$, as was the case in the present study.

Acknowledging these limitations, the present study has demonstrated a clear alternative for the methods currently used in practice. These methods have serious limitations that prevent the accumulation of knowledge regarding the nature of the influences between externalising and internalising symptoms. By showing how these problems can be solved by means of continuous-time analysis, we want to encourage applied researchers to apply the continuous-time procedure in their investigations of the causal effects between externalising and internalising problems. For this specific purpose we have included an Appendix with the LISREL script for the ADM/SEM procedure. Application of continuous-time methods facilitates the comparability of results across studies and may thus help to build a knowledge base that leads to a better understanding of the causal processes between adolescents' externalising and internalising problems. In the long run, this knowledge base may facilitate the development of more specifically tailored interventions aimed at improving adolescents' behavioural and emotional functioning.

References

Arbuckle, J. L. (2007). AMOS (Version 7) [Computer software]. Chicago: SPSS.

Bentler, P. M. (2006). EQS (Version 6.1) [Computer software]. Encino, CA: Multivariate Software.

Bergstrom, A. R. (1966). Nonrecursive models as discrete approximations to systems of stochastic differential equations. *Econometrica, 34*, 173-182.

Bergstrom, A.R. (1984). Continuous time stochastic models and issues of aggregation over time. In Z. Griliches & M.D. Intriligator (Eds.), *Handbook of econometrics* (Vol. 2, pp. 1145-1212). Amsterdam: North-Holland.

Bergstrom, A. R. (1988). The history of continuous-time econometric models. *Econometric Theory, 4*, 365-383.

Beyers, J. M., & Loeber, R. (2003). Untangling developmental relations between depressed mood and delinquency in male adolescents. *Journal of Abnormal Child Psychology, 31*, 247-266.

Biesanz, J. C., Deeb-Sossa, N., Papadakis, A. A., Bollen, K. A., Curran, P. J. (2004). The role of coding time in estimating and interpreting growth curve models. *Psychological Methods, 9*, 30-52.

Boker, S., Neale, M., Maes, H., Wilde, M., Spiegel, M., Brick, T. R., Spies, J., Estabrook, R., Kenny, S., Bates, T. C., Mehta, P., & Fox, J. (2011). OpenMx: *Multipurpose software for statistical modeling. (Version R package version 1.0.4)*. Virginia.

Bollen, K. A., & Curran, P. J. (2004). Autoregressive latent trajectory (ALT) models: A synthesis of two traditions. *Sociological Methods & Research, 32*, 336-383.

Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation perspective*. Hoboken, NJ: Wiley.

Burke, J. D., Loeber, R., Lahey, B. B., & Rathouz, P. J. (2005). Developmental transitions among affective and behavioral disorders in adolescent boys. *Journal of Child Psychology and Psychiatry, 46*, 1200-1210.

Capaldi, D. M. (1991). The co-occurrence of conduct problems and depressive symptoms in early adolescent boys: I. Familial factors and general adjustment at grade 6. *Development and Psychopathology, 3,* 277-300.

Capaldi, D. M. (1992). Co-occurrence of conduct problems and depressive symptoms in early adolescent boys: II. A 2-year follow-up at grade 8. *Development and Psychopathology, 4*, 125-144.

Curran, P. J., & Bollen, K. (2001). The best of both worlds: combining autoregressive and latent curve models. In: A. Sayer and L. Collins (eds.), *New methods for the analysis of change*, American Psychological Association, Washington, DC, 107-135.

Delsing, M. J. M. H., van Aken, M. A. G., Oud, J. H. L., De Bruyn, E. E. J., & Scholte, R. J. H. (2005). Family loyalty and adolescent problem behavior: The validity of the family group effect. *Journal of Research on Adolescence, 15*, 127-150.

Delsing, M. J. M. H., & Oud, J. H. L. (2008). Analyzing reciprocal relationships by means of the continuous-time autoregressive latent trajectory model. *Statistica Neerlandica, 62*, 58-82.

Delsing, M. J. M. H., Oud, J. H. L., & De Bruyn (2005). Assessment of bidirectional influences between family relationships and adolescent problem behavior: Discrete versus continuous time analysis. *European Journal of Psychological Assessment, 21*, 226-231.

Gard, T. C. (1988). I*ntroduction to stochastic differential equations*. New York: Marcel Dekker.

Gilliom, M., & Shaw, D. S. (2004). Codevelopment of externalizing and internalizing problems in early childhood. *Development and Psychopathology, 16*, 313-333.

Glaser, K. (1967). Masked depression in children and adolescents. *American Journal of Psychotherapy, 21*, 565-574.

Haselager, G. J. T., & van Aken, M. A. G. (1999). *Codebook of the research project Family and Personality: Volume 1. First measurement wave*. Nijmegen, the Netherlands: University of Nijmegen, Faculty of Social Science.

Hipwell, H. E., Stepp, S., Feng, X., Burke, J., Battista, D. R., Loeber, R., & Keenan, K. (2011). Impact of oppositional defiant disorder dimensions on the temporal ordering of conduct problems and depression across childhood and adolescence in girls. *Journal of Child Psychology and Psychiatry, 52*, 1099-1108.

Jackson, K. M., & Sher, K. J. (2003). Alcohol use disorders and psychological distress: a prospective state-trait analysis. *Journal of Abnormal Psychology, 112*, 599-613.

Jöreskog, K. G., & Sörbom, D. (1996). *LISREL 8: User's reference guide*. Chicago: Scientific Software International.

Keiley, M. K., Bates, J. E., Dodge, K. A., & Pettit, G. S. (2000). A cross-domain growth analysis: externalizing and internalizing behaviors during 8 years of childhood. *Journal of Abnormal Child Psychology, 28*, 161-179.

Lee, E. J., & Bukowski, W. M. (2012). Co-development of internalizing and externalizing problem behaviors: Causal direction and common vulnerability. *Journal of Adolescence, 35*, 713-729.

Lilienfeld, S. O. (2003). Comorbidity between and within childhood externalizing and internalizing disorders: reflections and directions. *Journal of Abnormal Child Psychology, 31*, 285-291.

Measelle, J. R., Stice, E., & Hogansen, J. M. (2006). Developmental trajectories of co-occurring depressive, eating, antisocial, and substance abuse problems in female adolescents. *Journal of Abnormal Psychology, 115*, 524-538.

Mehta, P. D., & West, S. G. (2000). Putting the individual back in individual growth curves. *Psychological Methods, 5*, 23-43.

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

Muthén, L. K., & Muthén, B. O. (1998). *Mplus user's guide*. Los Angeles, CA: Muthèn & Muthèn.

Neale, M. C., Boker, S. M., Xie, G., & Maes, H. H. (1999). *Mx: Statistical Modeling* (4th ed.). Richmond, VA: Department of Psychiatry.

Oud, J. H. L. (2007). Continuous time modeling of reciprocal effects in the cross-lagged panel design. In S.M. Boker & M.J. Wenger (Eds.), *Data analytic techniques for dynamical systems in the social and behavioral sciences*. Mahwah, NJ: Lawrence Erlbaum Associates.

Oud, J. H. L. (2010). Second-order stochastic differential equation model as an alternative for the ALT and CALT models. *AStA Advances in Statistical Analysis, 94*, 203-215.

Oud, J. H. L., & Delsing, M. J. M. H. (2010). Continuous time modeling of panel data by means of SEM. In K. van Montfort, J. H. L. Oud, & A. Satorra (Eds.), *Longitudinal research with latent variables* (pp. 201-244). New York, NY: Springer.

Oud, J. H. L., & Jansen, R. A. R. G. (2000). Continuous time state space modeling of panel data by means of SEM. *Psychometrika, 65*, 199-215.

Oud, J. H. L., & Singer, H. (2008). Continuous time modeling of panel data: SEM versus filter techniques. *Statistica Neerlandica, 62*, 4-28.

Overbeek, G., Biesecker, G., Kerr, M., Stattin, H., Meeus, W., & Engels, R. (2006). Co-occurrence of depressive moods and delinquency in early adolescence: the role of failure expectations, manipulativeness, and social contexts. *International Journal of Behavioral Development, 50*, 433-443.

Overbeek, G. J., Vollebergh, W. A. M., Meeus, W. H. J., Luijpers, E. T .H. & Engels, R. C. M. E. (2001). Course, co-occurrence and longitudinal associations between emotional disturbance and delinquency from adolescence to young adulthood: A six-year three-wave study. *Journal of Youth and Adolescence, 30*, 401-426.

Patterson, G. R., & Capaldi, D. M. (1990). A meditational model for boys' depressed mood. In J. Rolf, A. S. Masten, D. Cicchetti, K. H. Nuechterlein, & S. Weintraub (Eds.), *Risk and protective factors in the development of psychopathology* (pp. 141–163). New York: Cambridge University Press.

Patterson, G. R., Reid, J. B., & Dishion, T. J. (1992). *Antisocial boys* (Vol. 4). Eugene, OR: Castina.

Rao, C. R. (1958). Some statistical models for comparison of growth curves. *Biometrics 14*, 1-17.

Ritakallio, M., Koivisto, A., Von der Pahlen, B., Pelkonen, M., Marttunen, M., & Kaltiala-Heino, R. (2008). Continuity, comorbidity and longitudinal associationsbetween depression and antisocial behavior in middle adolescence: a 2-year prospective follow-up study. *Journal of Adolescence, 31*, 355-370.

Scholte, R. H. J., Vermulst, A. D., & De Bruyn, E. E. J. (2001). The Nijmegen Problem Behavior List: Construction and validation. Presentation at the 6th Conference of the European Association of Psychological Assessment, Aachen, Germany.

Tucker, L. R. (1958). Determination of parameters of a functional relation by factor analysis. *Psychometrika 23*, 19-23.

Van Lier, P. A. C., Vitaro, F., Barker, E. D., Brendgen, M., Tremblay, R. E., & Boivin, M. (2012). Peer victimization, poor academic achievement, and the link between childhood externalizing and internalizing problems. *Child Development*. Article first published online: 20 JUN 2012, DOI: 10.1111/j.1467-8624.2012.01802.x.

Verhulst, F. C., Van der Ende, J., & Koot, H. M. (1996). *Handleiding voor de CBCL/4-18 [Manual for the CBCL/4-18]*. Rotterdam: Afdeling Kinder-en Jeugdpsychiatrie, Sophia Kinderziekenhuis/ Academisch Ziekenhuis Rotterdam/Erasmus Universiteit Rotterdam.

Vieno, A., Kiesner, J., Pastore, M., & Santinello, M. (2008). Antisocial behavior and depressive symptoms: longitudinal and concurrent relations. *Adolescence, 43*, 649-660.

Voelkle, M. C., Oud, J. H. L, Davidov, E., & Schmidt, P. (2012). An SEM approach to continuous time modeling of panel data: Relating authoritarianism and anomia. *Psychological Methods, 17*, 176-192.

Wiesner, M. (2003). A longitudinal latent variable analysis of reciprocal relations between depressive symptoms and delinquency during adolescence. *Journal of Abnormal Psychology, 112*, 633-645.

MARC J. M. H. DELSING

A researcher at Praktikon Nijmegen. He received his PhD from Radboud University Nijmegen in 2004. His research interests include adolescent problem behaviour, the effectiveness of youth care interventions, youth culture, and longitudinal methodology. He is co-developer of BergOp, a web-based software program for monitoring treatment progress.

JOHAN H. L. OUD

Associate professor at the Behavioural Science Institute of the Radboud University Nijmegen. His research interests are in monitoring system construction, structural equation modelling (SEM), longitudinal research, and recently continuous time analysis by means of SEM. He has published a large number of papers, book chapters and edited several books in these fields.

Appendix.

LISREL script for ADM/SEM procedure

See the model specified in Figure 1 and the ADM restrictions. Appropriate observation intervals should be substituted for 'delta1' and 'delta2' in lines 14-31 (in our application, delta1 and delta2 both equal 1).
Last variable in data matrix 'filename' on line 2 should be the unit variable.

```
1   DA NI=7 NO=280 MA=MM
2   RA FI=filename
3   MO NY=7 NE=9 LY=FU,FI BE=FU,FI PS=SY,FI TE=SY,FI
    AP=4

    !Measurement model
4   VA 1 LY 1 1 LY 2 2 LY 3 3 LY 4 4 LY 5 5 LY 6 6

    !Mean structure
5   VA 1 LY 7 7

    !Lagged and instantaneous and auto- and cross-effects
6   FR BE 3 1 BE 3 2 BE 4 1 BE 4 2 BE 3 3 BE 3 4 BE 4 3 BE 4 4
7   FR BE 5 3 BE 5 4 BE 6 3 BE 6 4 BE 5 5 BE 5 6 BE 6 5 BE 6 6

    !Intercepts
8   FR BE 1 7 BE 2 7 BE 3 7 BE 4 7 BE 5 7 BE 6 7

    !Process error (co)variances
9   FR PS 1 1 PS 2 2 PS 3 3 PS 4 4 PS 5 5 PS 6 6
10  FR PS 2 1 PS 4 3 PS 6 5

    !Trait (co)variances
11  FR PS 8 8 PS 9 9 FR PS 9 8

    !Initial state-trait covariances
12  FR PS 8 1 PS 8 2 PS 9 1 PS 9 2

    !Unit moment
13  FR PS 7 7

    !Continuous-time restrictions
14  CO BE 3 3 = .5* delta1*par(1)
15  CO BE 3 4 = .5* delta1*par(2)
16  CO BE 4 3 = .5* delta1*par(3)
17  CO BE 4 4 = .5* delta1*par(4)
18  CO BE 5 5 = .5*delta2*par(1)
19  CO BE 5 6 = .5*delta2*par(2)
20  CO BE 6 5 = .5*delta2*par(3)
21  CO BE 6 6 = .5*delta2*par(4)

22  CO BE 3 1 = .5* delta1*par(1) +1
    !BE 3 1 = BE 3 3 + 1
23  CO BE 3 2 = .5* delta1*par(2)
    !BE 3 2 = BE 3 4
24  CO BE 4 1 = .5* delta1*par(3)
    !BE 4 1 = BE 4 3
25  CO BE 4 2 = .5* delta1*par(4) +1
    !BE 4 2 = BE 4 4 + 1
26  CO BE 5 3 = .5* delta2*par(1) +1
    !BE 5 3 = BE 5 5 + 1
27  CO BE 5 4 = .5* delta2*par(2)
    !BE 3 4 = BE 5 6
28  CO BE 6 3 = .5* delta2*par(3)
    !BE 6 3 = BE 6 5
29  CO BE 6 4 = .5* delta2*par(4) +1
    !BE 6 4 = BE 6 6 + 1

    !Trait state effects
30  VA delta1 BE 3 8 BE 4 9
31  VA delta2 BE 5 8 BE 6 9

32  OU ADD=OFF
```

Comments:

Line 1-3
As seen in Figure 1, there are 3 time points and 2 observed variables at each time point which makes the number of variables 6, but the extra unit variable (1 for all sample units) to accommodate for the mean structure in the model makes the total number of input variables NI=7. The sample size specified is NO=280 and, as we include the mean structure in the model, the matrix to be analysed becomes the moment matrix MA=MM. A raw data file (280x7) with name 'filename' is specified, as the model has as many observed variables as the data file, NY=7. The number of latent variables is NE=9, because there are 2 variables per time point (see Figure 1), 2 trait variables, and the unit variable is added to the model. The model is specified by means of 4 matrices: LY (factor matrix), BE (effect matrix), PS (error covariance matrix) and TE (measurement error covariance matrix). To start with, all 4 matrices are specified to be fixed (FI); parameters to be estimated will be specified free (FR) below. While the matrices LY and BE are specified full (FU), the covariance matrices PS and TE are specified symmetric (SY). Finally AP=4 stands for 4 extra continuous-time parameters which will be defined and used in the ADM continuous-time restrictions (lines 14-29).

Line 4
This line specifies the fixed nonzero coefficients in the measurement model matrix LY.

Line 5
The fixed 1 in line 5 makes sure that the observed unit variable (observed variable 7) becomes variable 7 of the latent part.

Line 6-7
Line 6 frees the 4 elements of lagged effects matrix $\mathbf{A}_{lag}$ and the 4 elements of instantaneous effects matrix $\mathbf{A}_{ins}$ (see ADM continuous-time restrictions) at time point t_1 and line 7 at time point t_2. These coefficients will be restricted according to the ADM in lines 14-29.

Line 8
Line 8 frees first the initial means and then the intercepts at time points t_1 and t_2, respectively.

Line 9-12
Line 9 frees first the initial variances and then the pairs of error variances at time points t_1 and t_2. Line 10 frees the covariances for the pairs of variables specified in line 9. Next, line 11 and 12 free trait variances and covariances and the covariances of the traits with the initial states.

Line 13
This frees the moment of the unit variable, which is 1 and should be estimated at this value in a good solution.

Line 14-31
In this part, the 4 approximate continuous-time parameters of the ADM are defined by means of the ADM constraints (CO is the LISREL command for specifying this type of constraints). Par(1), par(2), par(3), and par(4) are the 4 drift parameters in $\tilde{\mathbf{A}}$. The lines 14-21 implement $\mathbf{A}_{ins} = .5\ \tilde{\mathbf{A}}\Delta t$. Lines 22-29 implement $\mathbf{A}_{lag} = \mathrm{I} + .5\ \tilde{\mathbf{A}}\Delta t$. Because LISREL does not allow recursive constraining, parameters like BE 3 3 in line 22, which were constrained earlier, are replaced by the original constraint. Finally, in lines 30-31, the right coefficients for ‘delta1’ and ‘delta2’ (in this case 1) are inserted in BE for the trait variables.

Line 32
To avoid the program stops running before a solution is found, the admissibility check is put off.

Accommodation of genotype-environment covariance in a longitudinal twin design

In the classical twin study, genetic and environmental influences on a phenotype are usually estimated under the assumption that genotype-environment covariance (GE covariance) is absent. We explore possibilities to accommodate GE covariance in longitudinal data using the genetic simplex model. First, the genetic simplex model is presented, accompanied by a brief summary of results found in cognitive developmental studies. Second, GE covariance is specified via niche picking and sibling effects. Third, numerical and analytical identification is established, and the statistical power to detect GE covariance is examined. In a simplex model comprising four time points, GE covariance can be accommodated by introducing phenotype to environment cross-lagged pathways, either within or between twins. By using different parameter constraints within the genetic simplex, the extended models are numerically and analytically identified. The power to detect GE covariance is relatively low and therefore large sample sizes are needed.

Where: Netherlands Journal of Psychology, Volume 67, 81-90

Received: 16 July 2012; Accepted: 20 November 2012

Keywords: Quantitative genetics; Longitudinal models; Genotype – environment covariance; Cognitive abilities

Authors: Johanna M. de Kort* , Conor V. Dolan*,** and Dorret I. Boomsma**

*Department of Psychology, University of Amsterdam
**Department of Biological Psychology, Free University Amsterdam

Correspondence to:
Johanna M. de Kort,
University of Amsterdam,
Department of Psychology,
Weesperplein 4,
1018 XA Amsterdam,
the Netherlands,
e-mail: jjmdekort@gmail.com

Quantitative genetics is concerned with determining the genetic and environmental influences on behavioural variance within a well-defined population. To determine which portion of the phenotypic variance is due to genetic and environmental influences, researchers often use the classical twin design (Figure 1), i.e., the comparison of monozygotic (MZ) and dizygotic (DZ) twins growing up together (Eaves, Last, Martin, & Jinks, 1977; Van Dongen, Slagboom, Draisma, Martin, & Boomsma, 2012). Using this design, quantitative genetic studies have produced a wealth of results concerning the genetic and environmental influences to the observed multivariate and longitudinal covariance structure of different complex traits, such as cognitive abilities (see Plomin, DeFries, McClearn, & McGuffin, 2008). The application of the classical twin design to longitudinal data has shown that genetic and environmental influences are present throughout the lifespan. For a range of traits, the contribution of genetic influences to the phenotypic variance tends to increase, while the contribution of the environmental influences decreases from childhood into adulthood. For traits such as IQ, common environmental influences (i.e., environmental influences shared by twins that contribute to their similarity) are present prior to adolescence, but decrease in importance later in adolescence, while unique environmental influences (i.e. environmental effects unique to each twin which contribute to the dissimilarity of twins) are present throughout (e.g., Bartels, Rietveld, Van Baal, & Boomsma, 2002; Boomsma et al., 2002; Cardon, Fulker, & DeFries, 1992; Hoekstra, Bartels, & Boomsma, 2007; Petrill et al., 2004; Rietveld, Dolan, Van Baal, & Boomsma, 2000).

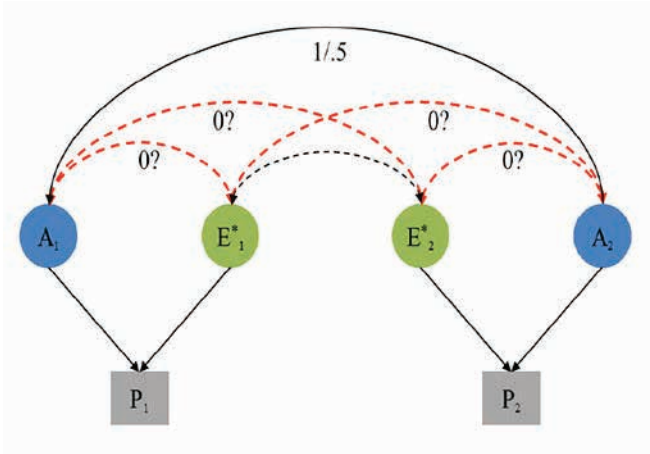

Figure 1 Classical twin-design model. This is the simplified representation of the classical twin model. In this simplified representation, (A) represents the additive genetic influences, (E) represents the total environmental influences, and (P) the phenotype. The correlation between the environmental effects (E1* and E2*) is due to common environmental effects (C). Correlation between A1 and A2 equals 1 in monozygotic, and .5 in dizygotic twins, due to genetic resemblance. In this model GE covariance (dotted arrows) and GxE interaction are assumed to be absent*

Most applications of the classical twin design come with well-known model assumptions, including absence of genotype by environment interaction (GxE interaction), zero genotype-environment covariance (GE covariance) (see dotted arrows in Figure 1), and random mating (i.e., zero spousal correlation; Eaves et al., 1977). These assumptions are known, or are suspected, to be violated to some degree in different complex traits. For instance, it is well known that assortative mating plays a role in intelligence (i.e., a positive correlation between IQ test scores of spouses; Eaves, 1973). For example, if data on parents of twins are available this information can be accommodated in the twin model (e.g., Martin, Eaves, Heath, Jardine, Feingoldt, & Eysenck, 1986). GxE interaction, i.e., moderation of genetic effects by environmental variables, or a dependence of environmental exposures on genotype, has been assessed thanks to advances in statistical modelling, enabling researchers to incorporate measured moderators, such as SES, into the twin model (Purcell, 2002; Harden, Turkheimer, & Loehlin, 2006; Boomsma & Martin, 2002). GE covariance has generally received less attention, although theoretically GE covariance is probably important, and has been hypothesised to explain increased heritability with age (Kan, Wicherts, Dolan, & Van der Maas, under revision).

The absence of GE covariance is certainly a strong assumption for many complex traits (Plomin et al., 2008). This assumption is often made pragmatically; in a design that includes only MZ and DZ twins applied to univariate data obtained at a single occasion, the covariance between genetic and environmental influences is not identified, and therefore cannot always be estimated. Here we explore whether GE covariance can be estimated from longitudinal data in the classical twin design using the genetic simplex (Boomsma & Molenaar, 1987). The genetic simplex provides a decomposition of phenotypic variance into genetic and environmental components at each measurement occasion (Figure 2). In addition, the genetic simplex expresses the phenotypic stability, i.e., the phenotypic correlation of a trait over time, in terms of genetic and environmental stability. In the standard genetic simplex, GE covariance is assumed to be absent as there is no direct or indirect pathway between genotypic and environmental components (Figure 2).

Developmental psychologists and behaviour geneticists, however, have long recognised definite processes giving rise to GE covariance (Carey, 1986; Eaves et al., 1977; Loehlin & DeFries, 1987; Plomin, DeFries, & Loehlin, 1977; Scarr, 1992; Scarr & McCartney, 1983). An important theoretical distinction is made between *passive*, *reactive*, and *active* GE covariance (Loehlin & DeFries, 1987; Plomin et al., 1977; Scarr, 1992; Scarr & McCartney, 1983): passive GE covariance arises when parents supply both genes and environment during the development of their offspring (i.e., smart parents transmit 'smart' genes and provide a 'smart' environment); reactive GE covariance arises when certain genotypes evoke certain reactions in the environment (e.g. 'smart' individuals evoke 'smart' reactions from their environment); and active GE covariance arises when individuals actively seek out environments consistent with their phenotype (i.e., 'smart' children seeking out a 'smart' environment). Provided that individual differences in the phenotype are at least partially due to genetic factors, these processes give rise to GE covariance.

Two conceptualisations of GE covariance are *niche picking* and *sibling effects*. Niche picking gives rise to within-individual GE covariance, as it involves an individual's choice or preference for certain environments, based on personal interest, talent, and personality (Scarr, 1992, Scarr & McCartney, 1983). This process thus implies a pathway between the individual's genotype and his or her environment, possibly mediated via the phenotype. Sibling effects give rise to between-individual GE covariance, as one sibling might directly or indirectly influence the other sibling's environment (Eaves, 1976; Carey, 1986), creating a pathway from one individual's genotype toward another individual's environment.

The aim of the present paper is to consider the specification of GE covariance processes in the genetic simplex and to explore the possibility to incorporate GE covariance in longitudinal twin models. We limit ourselves to the processes of niche picking and sibling effects in the simplex model including additive genetic (*A*) and unique environmental effect (*E*), and in a special case of a model with A, E, and common environmental effects (*C*) (see below). The setup of this paper is as follows: First, we represent the genetic simplex model, which we use as our starting model in which we incorporate GE covariance. Second, we consider the specification of GE covariance as arising through the processes of niche picking and sibling effects in three models: the within twin member's *niche picking model*, the between twin members' *sibling effects model*, and the combination of these two in the *combined model*. Third, we investigate the identification and resolution of these extended models, and compute the power to detect the parameters which give rise to GE covariance. We conclude with a brief discussion.

The genetic simplex

The genetic simplex model (Boomsma & Molenaar, 1987) has been used extensively to model longitudinal data in the classical twin design (e.g., Bartels et al., 2002; Bishop et al., 2003; Cardon et al., 1992; Petrill et al., 2004; Rietveld et al., 2000). The genetic simplex involves the regression of the phenotype measure at time point t, P_{tij}, on the additive genetic (A_{tij}), common (C_{tij}), and unique environmental variables (E_{tij}):

$$P_{tij} = A_{tij} + C_{tij} + E_{tij} + e_{tij} \quad (1)$$

where t denotes the measurement occasion (t=1...T), i denotes the twin pair, and *j* denotes the twin member. The term e_{tij} represents an occasion-specific residual, which may include genetic and environmental influences along with measurement error. Assuming the variables *A*, *C*, and *E* are uncorrelated, and given a correction for the occasion specific variance var(e_t) (e.g., if var(e_t) is a pure measurement error, this would be a correction for attenuation), the implied decomposition of variance at occasion *t* is

$$\text{var}(P_t) = \text{var}(A_t) + \text{var}(C_t) + \text{var}(E_t), \quad (2)$$

and the narrow sense heritability is $h^2_t = \text{var}(A_t)/[\text{var}(A_t)+\text{var}(C_t)+\text{var}(E_t)]$. The phenotypic stability is modelled by specifying autoregressive processes for A_t, C_t, and E_t. Limiting the equations to the additive genetic process, this entails the regression of A_{t+1} on A_t:

$$A_{t+1ij} = \beta_{At+1} A_{tij} + \zeta_{At+1}, \quad (3)$$

where β_{At+1} is the autoregressive coefficient and ζ_{At+1} is the residual, or innovation term. The implied variance decomposition is $\text{var}(A_{t+1}) = \beta_{At+1}^2 \text{var}(A_t) + \text{var}(\zeta_{At+1})$, where $\text{var}(\zeta_{At+1})$ is the residual or innovation variance. The covariance between *A* at *t* and *t+1* equals $\text{cov}(A_t A_{t+1}) = \beta_{At+1}\text{var}(A_t)$. We may also consider the percentage of explained variance in this regression, i.e., $R_{At+1}^2 = \beta A_{t+1}^2 \text{var}(A_t) / [\beta A_{t+1}^2 \text{var(At)} + \text{var}(\zeta A_{t+1})]$. Note that this percentage depends on the relative magnitudes of the autoregressive coefficient, β_{At+1}, and the residual variance, $\text{var}(\zeta_{At+1})$. The regression model applies to C_{tij} and E_{tij} as well, so that the phenotypic covariance of the phenotype at t and t+1 is decomposed as follows:

$$\text{cov}(P_t P_{t+1}) = \beta_{At+1}\text{var(A}_t) + \beta_{Ct+1}\text{var}(C_t) + \beta_{Et+1}\text{var}(E_t). \quad (4)$$

The genetic simplex provides an informative decomposition of the phenotypic variance at each occasion and of the contribution of genetic and environmental effect to the stability and change over time. Note that in the simplex (i.e., excluding the parameters giving rise to GE covariance), the first and the last occasion specific variances (var(e_1) & var(e_2)) are not identified. Identification can be achieved by setting these terms to zero, or by the imposition of the constraints var(e_1) = var(e_2) and var(e_{T-1}) = var(e_T). During our model evaluation, we imposed these latter equality constraints. Also note that the model includes several special cases. For instance, if the parameters of β_A approach zero this implies that genetic effects do not contribute to stability. If var(ζ_A) approaches zero (given β_A are not equal to zero), the genetic stability is perfect (i.e., R_A^2 approach 1). If this is the case throughout the time period considered, the (genetic part of the) autoregressive model tends towards a single common factor model (Bishop, et al., 2003).

Twin studies based on the genetic simplex have provided detailed information on the contributions of genetic and environmental factors to the longitudinal covariance structure of complex traits, such as cognitive abilities. During the development of cognitive abilities during early childhood, the influences of additive genetic components (*A*) follow a simplex pattern (i.e., both β_A and var(ζ_C) greater than zero; Bishop et al., 2003; Cardon et al., 1992; Rietveld et al., 2000, Petrill et al., 2004). As such, additive genetic influences are both a source of stability and change. Unique environmental influences (*E*) mostly contribute to instability, as β_E are relatively low and var(ζ_E) are non-zero (Bartels et al., 2002; Cardon et al., 1992; Petrill et al., 2004; Rietveld et al., 2000). Common environmental influences (*C*) mostly contribute to stability during early development, as β_C tends to approach unity and var(ζ_C) tend to zero (Bartels et

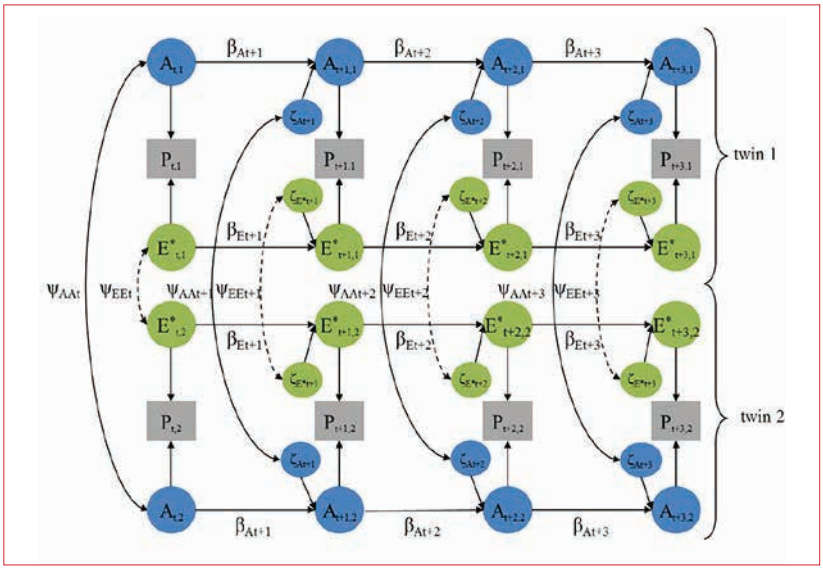

Figure 2 Simplified basic genetic simplex model, depicted with the minimum set of time points necessary to be able to identify the model. Note that E represents total environmental influences, which are correlated between twins due to common environmental influences (C) depicted by the dotted pathways (AE *model). If this pathway is dropped, the model reduces to the AE model, in which only additive genetic variance (A) and unique environment (E) influence the phenotype (P)*

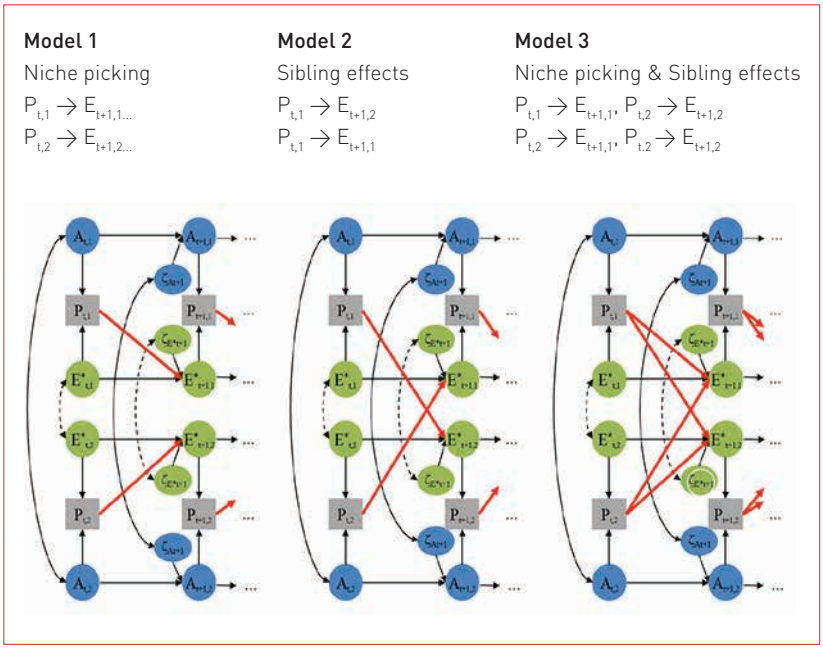

Figure 3 Path diagrams of three extensions of the basic genetic simplex model. To avoid clutter, only two time points (t=t, t+1) are depicted. The correlations between twins for var(A) (at t=t) and var(ζA) (at t=t+1,...) equals 1 and .5 in MZ and DZ twins, respectively. The covariance between the total environmental effects var(E) (at t=t) and var(ζE) (at t=t+1,...) are estimated, to accommodate common environmental effects

al., 2002; Bishop et al., 2003; Cardon et al., 1992; Petrill et al., 2004; Rietveld et al., 2000). Common environmental influences decrease in magnitude later in life, disappearing altogether in late adolescence. In addition, it has been established that the relative contribution of *A* increases, and that of *E* decreases over time (i.e. heritability increases over time, e.g., Bartels et al., 2002; Bishop et al., 2003; Boomsma et al., 2002; Haworth et al., 2010; Petrill et al., 2004,). Although the contributions of heritability and environment are robust and well established, the role of GE covariance has not been taken into account in these longitudinal studies.

Methods

Introducing GE covariance processes

We took the genetic simplex (Figure 2) as our starting model to introduce parameters giving rise to GE covariance. The simplex, as shown, accommodates common environmental influences (*C*) by the specification of correlated environmental influences (dotted arrows) rather than by the specification of a separate simplex process for C. By assessing the total environmental effects ($E^*=C+E$), instead of estimating each component separately, the specification and investigation of GE covariance originating in sibling effects and niche picking is greatly simplified[1]. So we considered two different models namely; 1) the A*E* model in which only additive genetic variance and unique environmental variance influence the phenotypic variance (i.e. the pathway between $E_{,,1}$, and $E_{,,2}$ is not included; 2) the AE^* model in which the unique environmental effects and the common environmental effects are captured in one term namely E^*.

By introducing crossed lagged phenotype to environment pathways within the two longitudinal models, we accommodated GE covariance within (i.e., niche picking) and between twins (i.e., sibling effects). Specifically, we viewed niche picking as the influence of phenotypic variable at occasion *t* on the environmental variable at time point *t+1* within each individual (Figure 3, Model 1). We accommodated sibling effects by introducing a cross lagged pathway from the phenotypic variable of one twin member at occasion *t* on the environment of the other twin member at time point *t+1* (Figure 3, Model 2)[2]. Finally, these two models can be combined (Figure 3,

[1] Note that our AE* simplex model is nested under the standard ACE simplex model, i.e., the standard ACE simplex will fit data generated with our AE* simplex model. The AE* simplex implies that in the standard ACE simplex the autoregressive coefficients of the E simplex equal those of the C simplex. This nesting is amenable to statistical testing.

[2] Note that this parameterisation of sibling effects deviates from previous methods such as that of Carey (1986) in which a polynomial was used to estimate GE covariance.

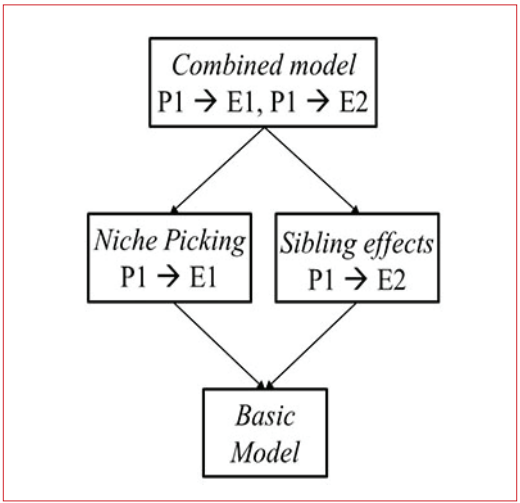

Figure 4 Nesting of the models of interest

Model 3), incorporating niche picking and sibling effects simultaneously. Note that we did not consider the direct path from *A* to *E*. Therefore we worked under the assumption that any effect of *A* on *E* must be mediated by the phenotype *P*. Still the pathway from *P* to *E* does imply GE covariance, as with this path in place, *A* and *E* are connected indirectly. For instance, in model 1 of Figure 3, the covariance between $A_{t,1}$ and E_{t+1},1 is due to the path from $A_{t,1}$ to $P_{t,1}$, and from $P_{t,1}$ to $E_{t,1}$.

Model evaluation

To establish whether extending the simplex model (i.e., the proposed cross lagged pathways) is practically feasible, we evaluated the extended models with respect to local identification, resolution, and power. First, we established model identification, which concerns the question whether the unknown parameters in the model can be estimated uniquely given appropriate longitudinal twin data. We distinguished between numerical and analytical identification. We considered both, because analytical identification does not rule out empirical under-identification. Empirical identification implies that fitting the true model to exact population MZ and DZ matrices produces a zero χ^2 value and perfect recovery of the parameter estimates regardless of variation in the starting values. A model is analytically identified if the Jacobian matrix of the model is of full column rank (Bekker, Merckens, & Wansbeek, 1993). The elements in the Jacobian matrix are the derivatives of each element in the population (MZ and DZ) covariance matrices to the unknown parameters (Derks, Dolan, & Boomsma, 2006; see also Bollen & Bauldry, 2010). We established analytical identification using the Maple program (Heck, 1993). To specify GE covariance, we added the cross lagged parameters (i.e., the additional parameters in the models depicted in Figure 3) to the basic simplex, without giving any additional constraints. If this model is not identified, we proceeded by imposing constraints on the parameters underlying GE covariance or on the other parameters in the model (Maple input is available on request). Second, we determined the resolution to see how well the competing models can distinguish between different effects. It is necessary to establish that two models (say model 1 and 2, as depicted in Figure 3), while both being identified, are not equivalent (i.e., they should not fit the data containing different effects equally well). To establish this, we fitted data generated according to one model and fitted all other models, which should result in misfit expressed in χ^2 values greater than zero. Third, we computed the power of each model to detect the parameters underlying the GE covariance, given an α of .05. To calculate the power, we first constructed MZ and DZ population covariance matrices according to the model of interest, i.e., giving the parameters underlying the GE covariance a certain value. Fitting the true model will then produce a χ^2 statistic of zero. Dropping the parameter of interest, i.e., those associated with niche picking and/or sibling effects, will result in a positive χ^2 statistic. This statistic can be used to calculate the power to detect the parameters underlying the GE covariance (Satorra & Saris, 1985). We computed the power for all nested models (Figure 4) using sample sizes up to 3000 twins and a fixed α of .05 (R scripts are available on request).

Calculation of covariance matrices

The numerical population MZ and DZ covariance matrices are calculated in four different scenarios (no GE covariance; GE covariance in the form of niche picking, GE covariance in the form of sibling effect, GE covariance in the form of a combined effect; see Figures 2 and 3), the two different models (AE and AE*), four time points and 1000 MZ and 1000 DZ twin pairs (see Table 1 for parameter values). In the AE* models, we included common environmental variance as the covariance between the environmental variables. To accommodate increasing heritability, the genetic innovations terms var(ζ_A) and the autoregressive coefficients β_A increase with time, while the values for the environmental innovations terms var(ζ_E) and the autoregressive coefficients β_E decrease. The strength of the niche picking effect (β_{PE} i.e. the GE covariance due to paths from Pt,1 to Et+1,1, and from $P_{t,2}$ to $E_{t+1,2}$, see Figure 3) is set to equal .1 for the first time point *t*, adding a value of .01 for each additional time point. The strength of the sibling effects (β_{PE^*} i.e. GE covariance due the path from $P_{t,1}$ to $E_{t+1,2}$ and $P_{t,2}$ to $E_{t+1,1}$, see Figure 3) is set to .05 at time point one, again adding a value of .01 for each additional time point (* indicates these parameters concern the sibling effects). We chose the

Table 1 Overview of the parameter values used to calculate the MZ and DZ covariance matrices

	Value given at time point			
Parameter	t	*t+1*	*t+2*	*t+3*
Ψ_A	10	2	3	4
$\Psi_{AA(MZ/DZ)}$	10/5	2/1	3/1.5	4/2
Ψ_E	10	3	2.5	2
Ψ_{EE}	2	1	1	1
var(e)	3	3	2	2
var(z_A)		2	3	4
var(z_E)		3	2.5	2
β_A		0.6	0.7	0.8
β_E		0.2	0.25	0.3
β_{PE}		0.1	0.11	0.12
β_{PE*}		0.05	0.06	0.07

Table 2 Overview of constraints, the number of parameters used to estimate cross lagged GE covariance, and analytical identification. The same results are found for both the AE and the AE*models

		Is the model identified?		
Identifying constraint	Max # of parameters for GE covariance	Niche picking	Sibling effects	Combined Model
$\beta_{Pet+1}=\beta_{PEt+2}=\beta_{Pet+3}$	1	Yes	-	No
$\beta_{Pet+1}^*=\beta_{PEt+2}^*=\beta_{PEt+3}^*$	1	-	Yes	No
$\beta_{Pet+1}=\beta_{PEt+2}=\beta_{PEt+3}$ & $\beta_{PEt+1}^*=\beta_{PEt+2}^*=\beta_{PEt+3}^*$	1	-	-	Yes
$\beta_{PE}=\delta_{00}+\delta_{01}(t-2)$	2	Yes	-	No
$\beta_{PE}^*=\delta_{00}^*+\delta_{01}^*(t-2)$	2	-	Yes	No
$\beta_{PE}=\delta_{00}+\delta_{01}(t-2)$ & $\beta_{Pet}^*=\delta_{00}^*+\delta_{01}^*(t-2)$	2	-	-	Yes
$\beta_{At+1}=\beta_{At+2}=\beta_{At+3}$	3	Yes	Yes	Yes
$\beta_{Et+1}=\beta_{Et+2}=\beta_{Et+3}$	3	Yes	Yes	Yes
$\beta_{At+1}=\beta_{At+2}=\beta_{At+3}$&$\beta_{Et+1}=\beta_{Et+2}=\beta_{Et+3}$	3	Yes	Yes	Yes
var(z_{At+1})=var(z_{At+2})=var(z_{At+3})	3	No	Yes	No
var(z_{Et+1})=var(z_{Et+2})=var(z_{Et+3})	3	No	Yes	No
var(z_{At+1})=var(z_{At+2})=var(z_{At+3}) & var(z_{Et+1})=var(z_{Et+2})=var(z_{Et+3})	3	No	Yes	No
$\Psi_{Et+1}=\Psi_{Et+2}=\Psi_{Et+3}$	3	No	Yes	No
var(e_{t+1})=var(e_{t+2})=var(e_{t+3})=var(e_{t+3})	3	No	Yes	No

occasion specific variance (var(e_t)) to approach 20% of the phenotypic variance. While the parameter values chosen here are somewhat arbitrary, the parameters do give rise to summary statistics that resemble those reported in the literature on cognitive abilities. That is, given the present parameter values, heritability increases over time (h^2 = .50, .622, .679, & .774). We performed numerical analyses using R (R Development Core Team, 2012) and LISREL 8.80 (Jöreskog & Sörbom, 2006).

Results

Model identification

We first established analytic identification of the three GE covariance extensions (Figure 3) in both the AE and the AE* models. To establish which constraints are identifying, we started with the most unconstraint model, a model without any equality constraints on the parameters, except for the standard equality constraints on the occasion specific variance mentioned above (i.e. var(e_t) = var(e_{t+1}) and var(e_{t+2}) = var(e_{t+3})), and worked through different constraints to see if the models were identified (see Table 2). Note that the basic model is identified if the parameters underlying GE covariance are fixed to zero.

The analytical identification procedures indicated that none of the extended models are identified without additional constraints. This was true in both the AE and the AE* models. For each extension (i.e., either for niche picking, sibling effects, or these effects combined), we determined which restrictions rendered the models identified. To this end, we first explored the possibilities within the parameters used to model GE covariance. One way to restrict the GE covariance parameters is by constraining the GE covariance parameters to be equal over time (i.e., for niche picking model: $\beta_{PEt+1}=\beta_{PEt+2}=\beta_{PEt+3}$, for sibling effects model: $\beta_{PEt+1}^*=\beta_{PEt+2}^*=\beta_{PEt+3}^*$, and for the combined model: $\beta_{PEt+1}=\beta_{PEt+2}=\beta_{PEt+3}$ & $\beta_{PEt+1}^*=\beta_{PEt+2}^*=\beta_{PEt+3}^*$). These equality constraints resulted in identification of the models in both the AE and the AE* models. A less restrictive identifying constraint is the use of a two parameter model (i.e., $\beta_{PE}=\delta_{00}+\delta_{01}(t-2)$), in which δ_{00} resembles the intercept of the regression (i.e. the initial influence of GE covariance) and δ_{01} the direction coefficient of the regression slope coefficient (i.e. the change in the influence of GE covariance with time). By using the two parameter model we allowed linear changes in the GE covariance estimates over time. We used the following parameters for the niche picking model: $\beta_{PE}=\delta_{00}+\delta_{01}(t-2)$, the sibling effects model $\beta_{PE}^*=\delta_{00}^*+\delta_{01}^*(t-2)$, and for the combined model $\beta_{PE}=\delta_{00}+\delta_{01}(t-2)$ & $\beta_{PEt}^*=\delta_{00}^*+\delta_{01}^*(t-2)$). Again this identifying constraint resulted in model identification. By using different constraints for the GE covariance parameter, the extended models are thus identified.

Second, we explored constraints on other parameters in the model to determine if these constraints rendered the parameters underlying GE covariance identified (without imposing any constraints on these parameters). As can be seen in Table 2, in the sibling effects model, many different constraints render the sibling effect parameters (i.e., model 2 in Figure 3)

Table 3 Overview of χ^2 values obtained when fitting different models to different data sets

χ^2 values obtained when fitting different models in AE model

Fitted model	Data generating model			
	Basic	*Niche picking*	*Sibling effects*	*Combined*
Basic	-	1.14	17.02+	22.62+
Niche picking	Perfect	-13.8+	16.75	
Sibling effects	Perfect	.77	- 1.71	
Combined model	Perfect	Perfect	Perfect	-

χ^2 values obtained when fitting different models in AE* model

Fitted model	Data generating model			
	Basic	*Niche picking*	*Sibling effects*	*Combined*
Basic	-	.67	21.74	28.14
Niche picking	Perfect	-16.16	20.33	
Sibling effects	Perfect	.55	- .47	
Combined model	Perfect	Perfect	Perfect	-

+ Models that experience computational problems when certain parameter values are used to calculate the MZ and DZ covariance matrices

identified. Within the niche picking model and combined model (models 1 and 3 in Figure 3), only constraints on the autoregressive coefficients (i.e., either $\beta_{At+1} = \beta_{At+2} = \beta_{At+3}$ *or* $\beta_{Et+1} = \beta_{Et+2} = \beta_{Et+3}$ or $\beta A_{t+1} = \beta_{At+2} = \beta_{At+3}$ & $\beta_{Et+1} = \beta_{Et+2} = \beta_{Et+3}$) rendered the GE covariance parameters identified.

Lastly, we established numerical identification for each of the three GE covariance extensions (see Figure 3) in both the AE and the AE* models using the two parameter model to estimate GE covariance. To do so, we first calculated the population MZ and DZ covariance matrices, to which we fitted the data generating model, i.e., the true model under which the covariance matrix is calculated, in LISREL 8.80 (Jöreskog & Sörbom, 2006). Although our results are limited to the parameter values chosen, we had no trouble fitting these models in LISREL. This suggests that, given the chosen parameter values, empirical under-identification was not a problem.

Resolution of the models

To determine whether the models were distinguishable, we generated MZ and DZ covariance matrices for all different effects (no effect, niche picking effect, sibling effect, combined effect) for both AE and the AE* models, and fitted various competing models to these covariance matrices (see Table 3). For instance, we fitted the niche picking model to covariance matrices generated with the sibling effects.

Our analyses led to several noteworthy observations (Table 3). First, in both AE and AE* models, fitting the basic model (i.e., no GE covariance) to the covariance matrices including GE covariance parameter leads to deviations from the zero χ^2 value. This shows the possibility to distinguish our proposed GE covariance models from the basic genetic simplex model. The low χ^2 value obtained when fitting the basic model to niche picking data, indicates low power given any reasonable α. Thus given the chosen parameters values, niche picking (i.e., within individual GE covariance) has a relatively weak effect on the phenotypic covariance structure. The higher χ^2 values, obtained when fitting the basic model to the sibling effects (i.e. between twin GE covariance) and combined model, indicate greater power, and thus a stronger effect on the phenotypic covariance structure. Second, when fitting the different GE covariance models to covariance matrices generated under the basic genetic simplex (i.e., fitting models with GE covariance parameters to data where GE covariance is absent) led to perfect model fit, as expected. This shows that the GE covariance parameters are correctly estimated to be zero when a GE covariance effect is absent. Third, when fitting the sibling effects model to niche picking or combined data, the model fit is almost perfect. This again indicates that the niche picking effect is hard to detect and to distinguish from the sibling effect. Fourth, fitting the niche picking model to the sibling effects and combined data led to large χ^2 values, which indicates that when the sibling effect is present, the niche picking model will not fit well.

Statistical power

The statistical power to detect different forms of GE covariance depends on the sample size and on α. Table 4 and Figure 5 give an overview of the number of twins needed to attain certain power given an α of .05. It can be concluded that, in terms of power, detecting niche picking is more difficult than detecting sibling effects. This conclusion is in line with the results presented earlier, where we found that the χ^2 values were lower when fitting the basic simplex to data including the niche picking effect than to data including the sibling effects. The greatest power is found for the detection of the combined effects. It should be noted that this is an omnibus test, in which the power to detect sibling effects and niche picking effects are combined. When computing the power to detect these effects separately, it is clear that the sibling effects are easier to detect (Figure 5).

Table 4 The power, non-centrality parameter (*italic*), and degrees of freedom, given an α of .05, for different sample sizes for both the AE models and the AE* models

			AE models			AE* models		
Data generating model	Fitted model	df	2x 500	2x 1000	2x 1500	2x 500	2x 1000	2x 1500
Niche picking	*Basic model*	2	.10	.15	.20	.08	.10	.13
			.57	*1.14*	*1.71*	*.34*	*.67*	*1.01*
Sibling effects	*Basic model*	2	.75	.97	1.00	.85	.99	1.00
			8.51	*17.02*	*25.53*	*10.87*	*21.74*	*32.61*
Combined model	*Basic model*	4	.78	.98	1.00	.87	1.00	1.00
			11.31	*22.62*	*33.93*	*14.07*	*28.14*	*42.21*
Combined model	*Niche picking*	2	.74	.96	1.00	.82	.99	1.00
			8.38	*16.75*	*25.12*	*10.16*	*20.33*	*30.49*
Combined model	*Sibling effects*	2	.12	.20	.28	.07	.09	.11
			.86	*1.71*	*2.56*	*.24*	*.47*	*.70*

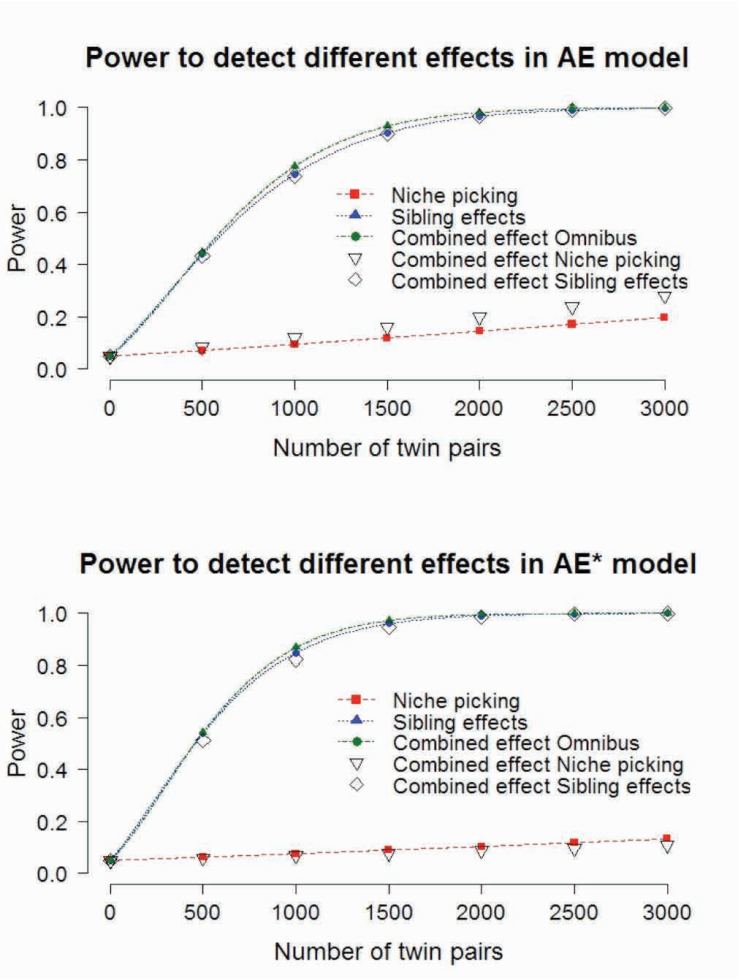

Figure 5 Graphical representation of ratio between sample size and power, given an α of .05, for the different models

Discussion

The aim of this paper was to specify processes giving rise to GE covariance within the genetic simplex model. To model GE covariance in the genetic simplex, we introduced phenotype to environment cross lagged relationships, representing niche picking effects, sibling effects, and the combined effects. We considered two models: one model with additive genetics and unique environmental influences (AE), and one model in which we accommodated the common environmental influences by covariance between E of each twin (AE*). First, we demonstrated the possibility to accommodate GE covariance in both the AE and AE* simplex models. The additional GE covariance parameters are identified under various identifying constraints. Identifying constraints may be imposed on the parameters accounting for the GE covariance. For instance, equality constraints and the use of a two parameter model (constraining the change in the parameters to be linear) rendered the model identified. Identification can also be achieved by imposing constraints on the standard parameters in the genetic simplex (e.g., the autoregressive coefficients). Given such constraints the parameter used to model GE covariance can be estimated freely at each time point. Second, we showed that it is possible, in principle, to determine whether an effect of GE covariance is present or not, as fitting a different model than the data generating model leads to non-zero χ^2 values. Third, we showed that relatively large sample sizes are needed to reach sufficient power to detect GE covariance effects, given our present parameter values. It turns out that the power to detect GE covariance depends on the type of effect. Larger sample sizes are needed to detect the niche picking effects than the sibling effects or combined effects. As power depends on the number of observations, we expect that adding time points to the models will lead to greater power in addition to simply increasing the sample size.

We emphasise that the present study is a first step towards establishing viable twin models including processes giving rise to GE covariance. Our present results are limited in the following respect. First, our results are limited to the scenarios considered, both in terms of measurement occasions (T=4) and of our choice of parameter values in our numerical results. Increasing the number of occasions is not likely to given rise to problems of identification. However, fewer occasions (say, 2 or 3) requires further study.

Second, our results concerning power and resolution depend wholly on our choice of parameter values, and are limited accordingly. More extensive power analyses were beyond the present scope, but we

note that such analyses pose no great problem to carry out, and can be tailored to the researcher's specific expectations. Our explorations of other parameter values showed that identification did not depend on the exact values (as expected). However, we did find that certain choices of parameters resulted in computational problems in fitting the basic (i.e., excluding parameters giving rise to GE covariance) genetic simplex. Notably, low values of the environmental autoregressive coefficient (e.g., $\beta_{Et+1} = .1, \beta_{Et+2} = .15, \beta_{Et+3} = .2$) in the sibling effects and the combined model rendered the basic simplex model computationally hard to fit as the occasion specific residual variances assumed negative values. This problem can be resolved by fixing these variances to zero.

Finally, we have only considered the AE model and the AE* model. The AE model is standard in the absence of common environmental influences (C). The AE* model treats common and unique environmental influences as 'total environmental effects', rather than explicitly modelling separate E and C processes. The AE* model is nested under the ACE model (as the ACE model with equal autoregressive C and E parameters implies the AE* model). In our current exploration of GE covariance, we only considered processes giving rise to AE covariance or AE* covariance. We have not addressed other sources of covariance, such as AC covariance, which are distinct from AE* and AE covariance, as these forms were beyond the scope of this article. We hope to extend our present results to the ACE model in the near future.

We conclude that sibling interaction and niche picking, conceptualised as the regression of environmental influences (E or E*) on the phenotypic variable, can be accommodated in the genetic simplex models considered here. While these models are identifiable given appropriate constraints, the issue of power requires attention, as does the generalisation to the standard ACE model. The application of these models, given adequate sample sizes, will ultimately allow one to establish whether these sources of GE covariance play any role in complex phenotypes, as is often suggested (e.g., in discussions of cognitive abilities).

References

Bartels, M., Rietveld, M. J. H., Van Baal, G. C. M., & Boomsma, D. I. (2002). Genetic and environmental influences on development of intelligence. *Behavior Genetics, 32*, 237-249.

Bekker, P. A., Merkens, A., & Wansbeek, T. J. (1993). *Identification, equivalent models, and computer algebra*. Boston, MA: Academic Press.

Bishop, E. G., Cherny, S. S., Corley, R., Plomin, R., DeFries, J. C., & Hewitt, J. K. (2003). Development genetic analysis of general cognitive ability from 1 to 12 years in a sample of adoptees, biological siblings and twins. *Intelligence, 31*, 31-49.

Bollen, K. A., & Bauldry, S. (2010). Model identification and computer algebra. *Sociological Methods & Research, 39*, 127-156.

Boomsma, D. I., Martin, N. G. (2002). Gene-environment interactions. In H. D'haenen, J. A. den Boer, P. Wilner (Eds), *Biological Psychiatry* (181-187). John Wiley & Sons Ltd.

Boomsma, D. I., & Molenaar, P. C. M. (1987). The genetic analysis of repeated measures. I. Simplex models. *Behavior Genetics, 17*, 111-123.

Boomsma, D. I., Vink, J. M., van Beijsterveldt, T. C., de Geus, E. J., Beem, A. L., Mulder, E. J., Derks, E. M., Riese, H., Willemsen, G. A., Bartels, M., van den Berg, M., Kupper, N. H., Polderman, T. J., Posthuma, D., Rietveld, M. J., Stubbe, J. H., Knol, L. I., Stroet, T., & Van Baal G. C. (2002). Netherlands twin register: A focus on longitudinal research. *Twin Research, 5,* 401-406.

Cardon, L. R., Fulker, D. W., & DeFries, J. C. (1992). Continuity and change in general cognitive ability from 1 to 7 years of age. *Developmental Psychology, 28*, 64-73.

Carey, G. (1986). Sibling imitation and contras effects. *Behavior Genetics, 16*, 319-341.

Derks, E. M., Dolan, C. V., & Boomsma D. I. (2006). A test of the equal environment assumption (EEA) in multivariate twin studies. *Twin Research and Human Genetics, 9*, 403-11.

Van Dongen, J., Slagboom, P. E., Draisma, H. H., Martin, N. G., Boomsma, D. I. (2012). The continuing value of twin studies in the omics era. *Nature Reviews Genetics, 13*, 640-653.

Eaves, L. J. (1973). Assortative mating and intelligence: An analysis of pedigree data. *Heredity, 30*, 199-210.

Eaves, L. J. (1976). A model for sibling effects in man. *Heredity, 36*, 205-214.

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

Harden K. P., Turkheimer, E., & Loehlin, J. C. (2006). Genotype by environment interaction in adolescents' cognitive aptitude, *Behavior Genetics*, DOI 10.1007/s10519-006-9113-4.

Haworth C. M. A., Wright, M. J., Luciano, M., Martin, N. G., de Geus, E. J. C., van Beijsterveldt, C. E. M., Bartels, M., Posthuma, D., Boomsma, D. I., Davis, O. S. P., Kovas, Y., Corley, R. P., DeFries, J. C., Hewitt, J. K., Olson, R. K., Rhea, S-A., Wadsworth, S. J., Iacono, W.G., McGue, M., Thompson, L. A., Hart. S. A., Petrill, S. A., Lubinski, D., & Plomin, R. (2010). The heritability of general cognitive ability increases linearly from childhood to young adulthood, *Molecular Psychiatry, 15*, 1112–1120.

Heck, A. (1993). *Introduction to Maple, a computer algebra system*. New York; Springer-Verlag.

Hoekstra, R. A., Bartels, M., Boomsma, D. I. (2007). Longitudinal genetic study of verbal and nonverbal IQ from early childhood to young adulthood. *Learning and Individual Differences, 17*, 97-114.

Jöreskog, K. G., & Sörbom, D. (2006). *LISREL 8.80 for Windows [Computer Software]*. Lincolnwood, IL: Scientific Software International, Inc.

Kan, K-J, Wicherts, J. M., Dolan, C. V., & Van der Maas, H. L. J. (2012). On the nature and nurture of intelligence and specific cognitive abilities: the more heritable, the more culture dependent. [Under revision].

Loehlin, J. C., & DeFries, J. C. (1987). Genotype-environment correlation and IQ. *Behavior Genetics, 17*, 263-277.

Martin, N. G., Eaves, L. J., Heath, A. C., Jardine, R., Feingoldt, L. M., & Eysenck, H. J. (1986). Transmission of social attitudes. *Proceedings of the National Academy of Sciences of the United States of America, 83*, 4364-4368.

Petrill, S. A., Hewitt, J. K., Cherny, S. S., Lipton, P. A., Plomin, R., Corley, R., & DeFries, J. C. (2004). Genetic and environmental contributions to general cognitive ability through the first 16 years of life. *Developmental Psychology, 40*, 805-812.

Plomin, R., DeFries, J. C., & Loehlin, J. C. (1977). Genotype-environment interaction and correlation in the analysis of human behavior. *Psychological Bulletin, 84*, 309-322.

Plomin, R., DeFries J. C., McClearn, G. E., & McGuffin, P. (2008). *Behavioral Genetics*. New York; Worth Publishers

Plomin, R., Loehlin, J. C., & DeFries, J. C. (1985). Genetic and environmental components of 'environmental' influences. *Developmental Psychology, 21*, 391-402.

Purcell, S. (2002). Variance components models for gene-environment interaction in twin analysis. *Twin Research, 5*, 554–771.

Rietveld, M. J. H., van Baal, G. C. M., Dolan, C. V., & Boomsma, D. I. (2000). Genetic factor analyses of specific cognitive abilities in 5-year-old Dutch children. *Behavioral Genetics, 30*, 29-40.

R Development Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Satorra, A., & Saris, W. E. (1985). Power of the likelihood ratio test in covariance structure analysis. *Psychometrika, 50*, 83-90.

Scarr, S. (1992). Developmental theories for the 1990s: Development and Individual differences. *Child Development, 63*, 1-19.

Scarr, S., & McCartney, K. (1983). How people make their own environments: A theory of genotype à environment effects. *Child Development, 54*, 424-435.

JOHANNA M. DE KORT
Following the Research Master Psychology at the UvA. She is interested in structural equation modelling of longitudinal data, with applications in the fields of psychometrics, quantitative genetics, developmental psychopathology, and clinical psychology.

CONOR V. DOLAN
Interested in the theory of structural equation modelling, with applications in the areas of psychometrics, quantitative genetics, and cognitive abilities.

DORRET I. BOOMSMA
She established the Netherlands Twin Register which is used as a resource for research projects that aim to quantify the influence of genes on phenotypic differences between individuals and identify the responsible variants.

Comparison of procedures used to test measurement invariance in longitudinal factor analysis

In this paper we hope to further advance the use of structural equation modelling to test longitudinal measurement invariance. To achieve this we discuss two different procedures to test invariance. We illustrate the differences by applying both procedures to an example of longitudinal data from lung cancer patients. One procedure relies on the modification indices (MI) and expected parameter changes (EPC) to assess the tenability of the equality constraints imposed on parameters across two measurement occasions. However, as Saris, Satorra and Van der Veld (2009) have suggested that this procedure can be improved upon by taking the power of the MI into account, our first procedure will include MI, EPC, and power. In the second procedure, we rely on global tests and standardised observed parameter changes (SOPC) rather than expected changes. Both procedures guard against chance findings, though they do so in very different ways that can lead to different results.

Where: Netherlands Journal of Psychology, Volume 67, 91-100

Received: 23 May 2012; Accepted: 4 December 2012

Keywords: Measurement invariance, Structural equation modelling, Modification indices, Longitudinal factor analysis

Authors: Bellinda L. King-Kallimanis*, Frans J. Oort**, Carol Tishelman*** and Mirjam A.G. Sprangers*

*Department of Medical Psychology, Academic Medical Center, University of Amsterdam, the Netherlands
**Methods and Statistics, Department of Education, University of Amsterdam, the Netherlands
***Department of Learning, Informatics, Management and Ethics, Karolinska Institutet, Stockholm, Sweden

Correspondence to:
B.L. King-Kallimanis, Department of Medical Psychology, Academic Medical Center, University of Amsterdam, Meibergdreef 15, 1105 AZ Amsterdam, the Netherlands
E-mail: kingkalb@tcd.ie

When we use structural equation modelling (SEM) to investigate measurement invariance the aim is to investigate whether the relationships between the observed items and the latent attribute remain constant across measurement occasions or groups. Violations of measurement invariance (measurement bias) can distort conclusions about common factor mean differences either among diverse groups or over time. Testing measurement invariance has become prevalent; therefore it is important that valid and reliable procedures are used to detect measurement bias. Formally, measurement invariance is expressed as

$$f_1(X|A = a, V = v) = f_2(X|A = a)$$

where X refers to a set of observed variables, A is the latent attribute measured by X, and V can represent anything that has the potential to affect the relationship between X and A. In longitudinal research, and in this paper, V represents time, but could also represent group membership, such as sex or race, or another attribute not represented by A. The function f_1 is the conditional distribution function of X given a and v; the function f_2 is the conditional distribution function of X given a. In the above equation conditional independence holds. However, if $f_1 \neq f_2$, then it can be said that the measurement of A by X is biased with respect to V and measurement invariance has been violated (Mellenbergh, 1989).

To test measurement invariance over time we rely on confirmatory factor analysis (CFA) of the mean and covariance structures. There are three levels of invariance essential for unbiased comparison of the common factor means: 1) *Configural invariance* - tests whether the same measurement model holds over time, i.e., with the same factor pattern, 2) *Metric invariance* - tests the invariance of the factor loadings over time by constraining them to equality, and 3) *Scalar invariance* - tests the invariance of the intercepts by adding equality constraints. It is these three tests that are critical to making valid

conclusions regarding change because they may affect assessment of change in the common factor means (Meredith & Horn, 2001; Oort, 2005; Sayer & Cumsille, 2001). Additional measurement invariance hypotheses can be tested, see Vandenberg and Lance, (2000) for a comprehensive review.

There are a number of procedures that have been proposed to detect measurement bias, or to test metric and scalar invariance. The problem that affects all procedures is chance findings. This is due to the large number of tests being considered. In this paper, we highlight why this is particularly problematic when relying solely on modification indices (MI), the most frequently used procedure, to test metric and scalar invariance, and discuss in detail two alternative procedures. For illustrative purposes, both alternative procedures are applied to an empirical example. We will discuss any differences in the detection of measurement invariance identified in the illustrative example and why these occur. Advantages and disadvantages of the procedures will also be discussed.

Procedures for identifying biased parameters

If the assumption of measurement invariance is violated, overall model fit statistics cannot be used to locate the biased item and identify which parameter the bias is associated with. To locate and identify such bias, most researchers turn to the MIs. The MIs can be helpful in this situation as they provide an indication of the size of the improvement in the overall chi-square statistic if the parameter in question were freed to be estimated. In general there are three important points to consider when using the MIs; 1) the number of possible modifications, 2) the interpretability of the modification, and 3) the power of the MI and size of the sample. These points are important so as to prevent changes made to the model that are solely data driven. MacCallum, Roznowski and Necowitz (1992) argue that whenever a model is modified using a data-driven strategy there is a strong possibility that some of the re-specifications made will be due to chance, and this possibility must be addressed and dealt with.

Number of possible modifications

The researcher can re-adjust the critical value the MI is assessed with, using a Bonferroni correction. This is achieved by calculating the number of plausible parameters to be investigated in regards to bias and re-adjusting the critical value to maintain a family wise Type I error rate of 5%. For example, if we have 80 parameters to consider, then only MI (with its chi-square distribution with one degree of freedom) larger than 11.7 (associated with a probability of 0.05/80) would be considered as an indication of bias. This classical approach has a detrimental effect on power. To overcome this, power-increasing adaptations to the Bonferroni procedure have been suggested; for example, when the number of tests under consideration changes, then the size of the critical value is re-adjusted (Hochberg & Benjamini, 1990; Holm, 1979).

Interpretability of the modification

Despite the warnings regarding the use of MIs, they are still the main and often only tool used to identify parameter misspecification. Even when precautions are taken, it may be the case that an MI suggests freeing a parameter that results in a significant decrease in the chi-square statistic, but the parameter estimate may change very little. This suggests that the modification to the model was not substantively meaningful and when testing equality constraints, it suggests that the equality constraint was in fact tenable. This can occur because 1) large sample size leads to high power to detect small but significant changes in the parameter, or 2) the model is poorly specified (Kaplan, 1990). As a result, decisions regarding the tenability of equality constraints should not be made solely on the size of the MIs. In a review of the techniques for evaluating misspecifications Kaplan (1990) discusses going beyond the MI, and also looking at the power associated with the MI, and expected parameter change (EPC). When using power and EPCs in conjunction with the MIs, the assessment of the tenability of the equality constraints will have greater reliability and chance findings will be reduced (Saris, Satorra, & Sörbom, 1987; Kaplan, 1990). Each of these established additional statistics are briefly discussed below.

When testing invariance, power is the probability of rejecting an equality constraint when in reality the equality constraint does not hold. Knowing the size of the MI and the power of the MI is not enough additional information, therefore the EPC can also be investigated (Saris, et al., 1987) to assess equality constraints. The EPC indicates the size of the expected change in the parameter estimate were it to be freed. This information helps to determine whether the parameter changes should be considered substantially valuable. The scale, however, of the EPC in CFA is arbitrary, which leads to difficulties in making comparisons across different estimate values and determining whether the value is ‘large’ or ‘small’ (Kaplan, 1989). There are a number of ways to overcome this; both Kaplan (1989) and Chou and Bentler (1993) suggested standardising the expected parameter change, these values are available in most standard structural equation modelling software.

Saris et al. (1987) propose a slightly different solution that will be further discussed below under Procedure 1.
How the MI, power of the MI, and EPC are applied, as suggested by Saris et al. (1987), has the potential to lead to different conclusions with respect to which parameters show measurement bias when testing measurement invariance. In the following sections we will describe how these statistics can be combined and we will also introduce an alternative procedure.

Procedure 1 – Modification indices and expected parameter changes

In this procedure MI, power of the MI and EPCs are considered to detect measurement invariance once the across occasion equality constraints have been placed on the factor loadings and intercepts. Rather than directly standardising the EPC, it is possible to choose a standardised misspecification size and convert this value to the scale of each EPC, and investigate whether the EPC is less than or greater than this value. To achieve this, the value chosen to represent misspecification is divided by the standard deviation of the observed score associated with the particular EPC in question. Substantial misspecification is indicated when the EPC is greater than the un-standardised cutoff value, with the cutoff point being determined by the researcher. As calculating all these cutoff values is time consuming, Saris, Satorra and Van der Veld (2009) developed the JRule program to aid in these calculations. JRule reads the output generated by two popular SEM programs (LISREL and Mplus), and then produces its own output file with the additional calculations for these cutoffs and power and the MIs and EPCs. As we will rely on Jrule to test measurement invariance, we will use the standardisation of the EPCs suggested by Saris et al. (1987).

Using the MI, EPC and power, Saris et al. (1987) argue that there are four possible outcomes. These four outcomes are: 1) no bias – the equality constraint is tenable, as the MI is not significant and the power is high; 2) bias – the equality constraint is not tenable as the MI is significant and the power is low; 3) possible bias – the MI is significant and the power is high. Saris et al. (2009) propose in this situation that the researcher checks whether the EPC is greater than the cutoff value. If the EPC is greater than the cutoff, bias is considered present and; 4) undetermined – there is not enough information to determine if the equality constraint is tenable as the power is low (<.80) and the MI is small (not significant). These four outcomes are also reported in JRule.

Advantages and disadvantages

The advantage of using this procedure is that it is time efficient as only the relevant models (i.e., models where a biased parameter has been freed) need to be investigated and with the aid of JRule no additional hand calculations are required. It can be argued that the four outcomes provide a clear course of action. However, in the fourth outcome, it is not possible to determine whether bias is present, leaving the researcher with no clear course of action, which is disadvantageous. Finally, while the procedure was proposed to reduce chance findings, we must face the fact that MI, power of MI and EPC are all related to each other, so chance findings may still occur. (see Saris et al. (2009)).

Procedure 2 – Global tests and observed parameter change

An alternative procedure to guard against chance findings is to apply global tests and calculate the observed parameter change when testing the invariance of each observed variable in a series of nested models. In global testing (King-Kallimanis, Oort & Garst, 2010) we rely on the chi-square difference test. We do this in an attempt to guard against chance findings in three ways: 1) by directly testing specific hypotheses, 2) by using the global test we simultaneously consider the invariance of multiple parameters (both factor loadings and intercepts), and 3) by conducting all tests at Bonferroni adjusted levels of significance (Holm, 1979). However, the limitation of using only global tests is that we rely solely on statistical testing to detect measurement bias. This is problematic because as power increases, small, yet significant differences are detected as bias, which in fact are substantively irrelevant.

To overcome this shortcoming, we also calculate the standardised observed parameter change (SOPC) for both the factor loadings and intercepts: $SOPC_{\lambda} = \lambda_{iA}^{*} - \lambda_{iR}^{*}$ where λ_{iA}^{*} refers to the standardised estimated factor loadings ($\lambda_{iA}^{*} = \lambda_{iA} / \sigma_{i}$ where λ_{iA} is the unstandardised factor loading and σi is the standard deviation) in the alternative model for item i when the equality constraints are removed and λ_{iR}^{*} refers to the standardised estimated factor loadings in the fully restricted model where those same parameters are constrained to equality over time. Similarly, $SOPC_{\tau} = \tau_{iA}^{*} - \tau_{iR}^{*}$ where τ_{iA}^{*} refers to the rescaled estimated intercepts ($\tau_{iA}^{*} = \tau_{iA} / \sigma_{i}$ where τ_{iA} is the unstandardised intercept estimate and σi is the standard deviation) in the alternative model for item i when the equality constraints are removed and τ_{iR}^{*} refers to the rescaled intercepts in the fully restricted model where those same parameters

are constrained to equality over time. Using the SOPC we are able to simultaneously test the size of multiple parameters, whereas the standardised EPC considers only single parameters.

To investigate measurement invariance, the fully restricted model is fit with all across occasion equality constraints on the factor loadings and intercepts included. Next, a series of alternative models are fit. There is an alternative model for each observed variable included in the measurement model. In each alternative model the equality constraints on both the factor loadings and the intercepts associated with the particular observed variable are removed. The alternative and fully restricted models are compared using the global test, which is a multiple degree of freedom test (degrees of freedom are dependent on the number of measurement occasions). An SOPC is calculated for each parameter where the equality constraint was removed. When the global test is significant and in conjunction with an SOPC that meets a pre-defined cutoff value, then we consider the associated observed variable as biased. If this occurs, the factor loadings and intercepts of the biased observed item remain free, and a new series of alternative models are fit, this is repeated with adjusted levels of significance (Holm, 1979) until no global tests and SOPCs meet the criteria. As with Procedure 1, the cutoff value is determined by the researcher.

Advantages and disadvantages

The main advantage of using global tests and SOPCs is that in testing all alternative models, the impact of freeing a small number of parameters associated with an observed variable can be seen and it reduces chance findings as multiple parameters are considered simultaneously. The main disadvantage is that all possible alternative models are tested; therefore the procedure is time intensive.

Illustrative example

Data

To illustrate the procedures outlined above, we used data from a longitudinal study that investigated the health-related quality-of-life (HRQoL) of patients with primary inoperable lung cancer. The data utilised in this example came from 216 patients who completed the baseline questionnaire that was on average about two weeks after diagnosis and the first follow-up questionnaire that was approximately two weeks following baseline (Tishelman, et al., 2005).

To measure HRQoL the European Organisation for Research and Treatment of Cancer (EORTC) QLQ-C30 questionnaire was used (Aaronson, et al., 1993). The questionnaire includes 30 items on a seven-point response scale that cover nine multi-item domains: Physical Functioning (PF: 5 items), Role Functioning (RF: 2 items), Fatigue (FA: 3 items), Nausea and Vomiting (NV: 2 items), Pain (PA: 2 items), Emotional Functioning (EF: 4 items), Cognitive Functioning (CF: 2 items), Social Functioning (SF: 2 items) and Global Health Status (GH: 2 items). An additional six domains are measured with single items. Also included was the lung cancer specific module of the EORTC QLQ-C30, the EORTC-LC13 (Bergman, Aaronson, Ahmedzai, Kaasa & Sullivan, 1994). The questionnaire has three items scored on a four-point response scale that cover one multi-item domain, symptoms of dyspnoea (DY), and 10 single item domains. Only the multi-item domains from both the EORTC QLQ-C30 and the EORTC-LC13 (ten in total) are included in the current analysis and items are simply summed to create continuous scores. Symptom related scores were reversed, so that higher scores indicate less symptoms. The range of scores is scaled such that all sub-scales range from 0 to 100, with higher scores, for all scales, are indicative of better HRQoL.

Analysis strategy

To test invariance we carried out three steps. The goal of Step 1 was to identify a factor model for the EORTC QLQ-C30 that met the requirements for configural invariance and had a clear interpretation and good fit. Fit was assessed using the chi-square test of exact fit, and the approximate fit indices, root mean square error of approximation (RMSEA) and expected cross-validation index (ECVI). When using CFA it is necessary to provide scale and origin to the common factors. There are three possible ways to achieve this (for details see Reise, Widaman and Pugh (1993)). As we provide scale and origin via constraints on the factor loadings and intercepts the goal of Step 2 was to identify invariant indicators using the List and Delete procedure outlined by Rensvold and Cheung (2001). In this procedure, a series of models are run in order to identify which unbiased indicators could be used to provide scale and origin. In Step 3 we constrain all factor loading and intercepts to equality across measurement occasions and investigate measurement invariance using Procedures 1 and 2. In Procedure 1 we investigated measurement invariance using the MIs and EPCs and in Procedure 2 we used Global Tests and SOPCs. All analyses were conducted using LISREL 8.54 with maximum likelihood estimation. For additional calculations Mx (Neale, 2010) and JRule 3.0.4 (van der Veld, Saris & Satorra, 2008) were used.

Table 1 Results from Procedure 1 for factor loadings from fully constrained model (Model 2)

Scale	Modification index	EPC	Cutoff†	Power	Decision
DY_1 – Phys HRQoL	3.28	- 0.01	0.09	0.99	No bias
PF_1 – Phys HRQoL	4.03	0.02	0.09	0.99	EPC -> no bias
RF_1 – Phys HRQoL	NA				
FA_1 – Phys HRQoL	0.05	0.00	0.10	0.99	No bias
NV_1 – Phys HRQoL	6.36	0.02	0.07	0.99	EPC -> no bias
PA_1 – Phys HRQoL	0.07	0.00	0.12	0.99	No bias
GH_1 – Phys HRQoL	0.36	0.00	0.09	0.99	No bias
GH_1 – Ment HRQoL	0.12	- 0.01	0.12	0.98	No bias
EF_1 – Ment HRQoL	15.78	- 0.09	0.12	0.99	EPC -> no bias
CF_1 – Ment HRQoL	5.00	0.04	0.12	0.99	EPC -> no bias
SF_1 – Ment HRQoL	NA				
DY_2 – Phys HRQoL	3.28	0.03	0.09	0.99	No bias
PF_2 – Phys HRQoL	4.03	- 0.02	0.08	0.99	EPC -> no bias
RF_2 – Phys HRQoL	NA				
FA_2 – Phys HRQoL	0.05	0.00	0.10	0.99	No bias
NV_2 – Phys HRQoL	6.36	- 0.05	0.08	0.98	EPC -> no bias
PA_2 – Phys HRQoL	0.07	- 0.01	0.11	0.99	No bias
GH_2 – Phys HRQoL	0.36	0.02	0.09	0.98	No bias
GH_2 – Ment HRQoL	0.12	0.00	0.12	0.99	No bias
EF_2 – Ment HRQoL	15.78	0.00	0.11	0.99	EPC -> no bias
CF_2 – Ment HRQoL	5.00	0.00	0.12	0.99	EPC -> no bias
SF_2 – Ment HRQoL	NA				

† These cutoffs correspond to a difference of 0.10 for standardised factor loadings; NA refers to parameter used to provide scale to the common factors; EPC = expected parameter change; DY = dyspnoea; PF = physical functioning; RF = role functioning; FA = fatigue; NV = nausea; PA = pain; GH = general health; EF = emotional functioning; CF = cognitive functioning; SF = social functioning; Phys HRQoL = physical health-related quality-of-life; Ment HRQoL = mental health-related quality-of-life

Results

Step 1 – Establish a measurement model

The construct of HRQoL is often represented by two factors: Physical HRQoL and Mental HRQoL. Therefore, we specified a model where, at both measurement occasions, seven domains loaded on the Physical HRQoL common factor and four domains loaded on the Mental HRQoL common factor (see Figure 1). Though the chi-square test was significant, the RMSEA indicated satisfactory fit (Model 1, χ^2 302.08 (152), p <0.001, RMSEA =0.068, 90% CI (0.056 ; 0.079)). In addition, the standardised residuals did not suggest any substantive misspecification. To provide scale and origin to this model (factor loading and intercepts set to a constant), the first observed variable of Physical HRQoL and the last observed of Mental HRQoL at both measurement occasions were used.

Step 2 – Providing unbiased scale and origin to the common factors

To ensure that we chose an invariant indicator, we used the List and Delete procedure (Rensvold & Cheung, 2001). This required fitting a series of models where there was a null model (Model 1, Step 1, no equality constraints), and an alternative model where one factor loading and one intercept were set to a constant and another factor loading and intercept were set to equality over time and the fit of the alternative model was compared with the null model. We tested all possible pairs of constraints for Physical HRQoL and then did the same for Mental HRQoL. After testing all possible pairs, Role Functioning was designated to provide scale and origin for Physical HRQoL and Social Functioning for Mental HRQoL. These variables were chosen because when tested there was the least deterioration in model fit.

Step 3 –Procedure 1: Investigate measurement invariance using modification indices and expected parameter changes

In this step, all factor loadings and intercepts from Model 1 were constrained to be equal across time (Model 2, χ^2 (169) = 359.98, p = <0.001, RMSEA =0.071, 90% CI (0.061 ; 0.081)). We chose to consider misspecifications that minimally corresponded to small effect sizes (Cohen, 1988) as indicative of measurement bias. This corresponded to a factor loading difference of ≥. 10 and an intercept difference of ≥ .20. Using the MI, power of the MI and EPC, the results indicated that ten out of 16 plausible factor loading constraints were tenable (no bias) and six out of 16 plausible intercept constraints were tenable (no bias). The results for the remaining factor loadings and intercepts indicated that the power was high (≥ .80) and the MIs were significant, therefore we investigated the EPCs (see Tables 1 and 2). The EPCs indicated that the misspecifications were generally very small and did not meet our criteria. There was one exception, where the EPC associated with the intercept of Emotional Functioning at the second measurement occasion met our criteria (Table 2). Equality constraints were removed for the intercept of Emotional Functioning, and the MIs, power of the MIs and EPCs for this new model were investigated. No additional misspecifications were identified. The chi-square for this final model was significant, but the RMSEA was satisfactory (χ^2 (168) = 341.94, p = <0.001, RMSEA =0.069, 90% CI (0.058 ; 0.079)) (See Figure 1, for parameter estimates).

Table 2 Results from Procedure 1 for intercepts from fully constrained model (Model 2)

	Modification	EPC	Cutoff†	Power	Decision
DY_1	10.10	- 0.17	0.47	0.99	EPC -> no bias
PF_1	6.44	0.16	0.44	0.99	EPC -> no bias
RF_1	NA				
FA_1	0.13	- 0.03	0.54	0.99	No bias
NV_1	12.01	0.20	0.35	0.99	EPC -> no bias
PA_1	0.19	- 0.05	0.61	0.99	No bias
GH_1	0.06	0.00	0.44	0.99	No bias
EF_1	17.58	- 0.11	0.46	0.99	EPC -> no bias
CF_1	7.25	0.05	0.45	0.99	EPC -> no bias
SF_1	NA				
DY_2	10.10	0.19	0.48	0.99	EPC -> no bias
PF_2	6.44	- 0.09	0.45	0.99	EPC -> no bias
RF_2	NA				
FA_2	0.13	0.02	0.54	0.99	No bias
NV_2	12.01	- 0.30	0.44	0.99	EPC -> no bias
PA_2	0.19	0.03	0.58	0.99	No bias
GH_2	0.06	0.03	0.47	0.99	No bias
EF_2	**17.58**	**0.46**	**0.44**	**0.98**	EPC -> bias
CF_2	7.25	- 0.31	0.46	0.98	EPC -> no bias
SF_2	NA				

† These cutoffs correspond to a difference of 0.20 for standardised intercepts; NA refers to parameter used to provide origin to the common factors; EPC = expected parameter change; DY = dyspnoea; PF = physical functioning; RF = role functioning; FA = fatigue; NV = nausea; PA = pain; GH = general health; EF = emotional functioning; CF = cognitive functioning; SF = social functioning; Phys HRQoL = physical health-related quality-of-life; Ment HRQoL = mental health-related quality-of-life

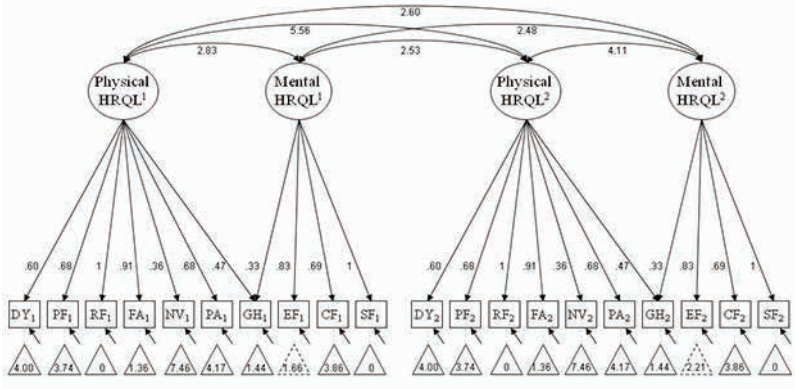

Figure 1 EORTC QLQ-C30 measurement model for Procedure 1. Factor loadings and intercepts of Model 3.1a
DY = dyspnoea; PF = physical functioning; RF = role functioning; FA = fatigue; NV = nausea; PA = pain; GH = general health status; EF = emotional functioning; CF = cognitive functioning; SF = social functioning

In regards to the bias identified, the estimate of the intercept of Emotional Functioning was lower at baseline than at follow-up. Apparently it was more difficult to give a positive response to the Emotional Functioning scale shortly after diagnosis, relative to the respondents Mental HRQoL (see Figure 3b). At follow-up, patients scored higher on the Emotional Functioning scale (better Emotional Functioning) relative to their Mental HRQoL. Perhaps when patients were initially diagnosed they felt overwhelmed, but after starting treatment some of the anxiety was relieved because something was being done to treat their lung cancer, and they also had the support of family and friends.

Step 3 – Procedure 2: Investigate measurement invariance using Global Tests and SOPCs

Initially, all factor loadings and intercepts were simultaneously constrained to equality over time (Model 2, Table 3). All alternative models were tested and their fit assessed. Equality constraints were considered not tenable when there was a significant global test ($\alpha^* = 0.05/10 = 0.005$) and large SOPCs. The SOPCs were considered to represent substantial misspecification when the factor loading SOPCs were ≥.10 and the intercept SOPCs ≥.20 (the same sizes that were used in Procedure 1). In the first iteration, the removal of the equality constraints associated with Dyspnoea parameters lead to a significant chi-square difference test and large SOPCS (Table 3, Model 3). Leaving all parameters associated with Dyspnoea free to be estimated and readjusting the significance ($\alpha^* = 0.05/9 = 0.005$), the model with the parameters of Nausea free also met our criteria. In the next series of models, no model met our criteria; therefore Model 4 (Table 3) was considered as the final model (See Figure 2, for parameter estimates).

The two violations of invariance were associated with the intercepts for the Dyspnoea and Nausea scales (see Table 3). Apparently it was more difficult for patients to give a positive (less symptoms) response to the Dyspnoea scale shortly after their diagnosis relative to the respondents Physical HRQoL (see Figure 3a). At follow-up, patients scored higher (less symptoms) relative to their overall Physical HRQoL. In general, symptoms of Dyspnoea did not worsen as much as expected given that Physical HRQoL significantly decreased over time (Figure 3a). Perhaps when patients were initially diagnosed, they conceptualised Dyspnoea as a primary symptom of lung cancer, but after beginning treatment they believed that the treatment was reducing this symptom. In regards to Nausea, the opposite was seen. Apparently it was more difficult for patients to give a negative response (more symptoms) to Nausea items shortly after diagnosis. At follow-up, patients scored lower (more symptoms) relative to their overall Physical HRQoL. Therefore, Nausea worsened more than we would have expected given the general decrease in Physical HRQoL as can be seen in Figure 3b. It is possible

Table 3 Results from procedure 2 – overall goodness-of-fit, SOPCs and chi-square difference tests

	Model	CHISQ (df)	SOPC	RMSEA (90 % CI)	Comparison models	CHISQ DIFF (df)	P -value	ECVI (90% CI)
1	Measurement model 1	302.08 (152)	NA	0.068 (0.056 ; 0.079)	NA	NA	NA	2.131 (1.825 ; 2.286)
2	Addition of equality constraints	359.98 (169)	NA	0.071 (0.061 ; 0.081)	1 vs. 2	57.90 (17)	<0.001	2.208 (1.883 ; 2.384)
3	Dyspnoea equality constraints freed	345.91 (167)	λ_1 -0.015 λ_2 0.010 τ_1 0.110 τ_2 -0.160	0.069 (0.058 ; 0.079)	2 vs. 3	14.07 (2)	0.001	2.152 (1.833 ; 2.321)
4	Nausea equality constraints freed	334.34 (165)	λ_1 0.010 λ_2 -0.009 τ_1 -0.179 τ_2 0.140	0.068 (0.057 ; 0.079)	3 vs. 4	11.57 (2)	0.003	2.139 (1.824 ; 2.305)

Note: λ_1 = time 1 factor loading, λ_2 = time 2 factor loading, τ_1 = time 1 intercept, τ_2 = time 2 intercept

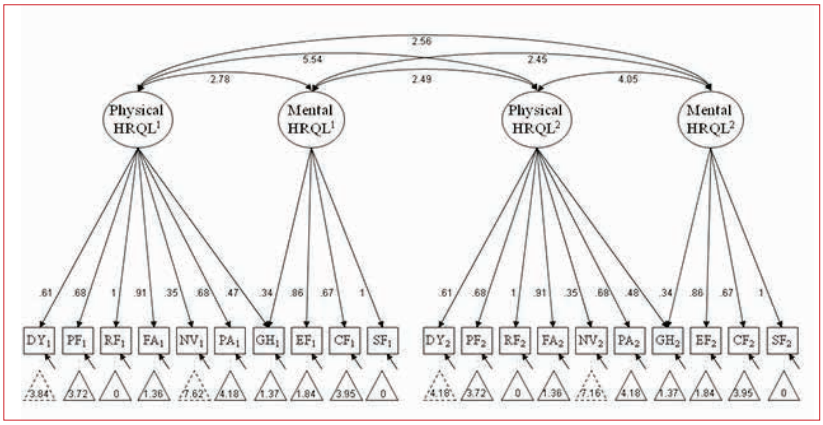

Figure 2 EORTC QLQ-C30 measurement model for Procedure 2. Factor loadings and intercepts of Model 3.2b

DY = dyspnoea; PF = physical functioning; RF = role functioning; FA = fatigue; NV = nausea; PA = pain; GH = general health status; EF = emotional functioning; CF = cognitive functioning; SF = social functioning

that this bias occurred as a result of treatment side effects. It seems likely that once treatment is completed, we would see the intercept for Nausea to increase to its pre-treatment value.

Comparison of results

Different results were obtained in Procedure 1 to those found in Procedure 2. We believe the primary reason the results differ is because in Procedure 1 each equality constraint is treated as a single parameter that may or may not be misspecified, whereas in Procedure 2, the equality constraints are treated as a set of multiple parameters. For example, in Table 3, we can see that the SOPCs for the intercepts of Dyspnoea were not larger than 0.20 individually; however the combined difference between the two estimates was greater than a small effect. A secondary reason is because Procedure 1 relies on expected changes in the model, whereas in Procedure 2, the SOPCs are observed changes. When the OPC was calculated for the intercept at follow-up for Emotional Functioning, the value was smaller than the EPC (OPCτ2 = 0.037) and below the cutoff for this parameter. This indicates that there was little actual difference in the intercept.

Before bias was accounted for, there was no significant change in either the Physical or Mental HRQoL latent means. After accounting for bias in Procedure 1, this result remained the same. However, the latent means for Mental HRQoL before accounting for bias had a slight upward trend (not significant) and after accounting for bias, this trend became negative (not significant). Once bias was accounted for in Procedure 2, we concluded that that there was a small significant decrease in Physical HRQoL; the conclusion for Mental HRQoL remained the same (See Figures 3a and 3b).

Discussion/ Conclusion

In this paper, we illustrate two procedures for testing invariance with SEM. Ultimately we came to different conclusions regarding which parameters were associated with measurement bias. As noted in the results, in Procedure 1 we tested single parameters, whereas in Procedure 2 we considered multiple parameters simultaneously. In both procedures, misspecification was considered substantial when EPCs and SOPCs were associated with a standardised change of ≥.10 for factor loadings and ≥.20 for intercepts. However, because of the single/multiple parameter distinction we are not actually testing for the same size of misspecification. In an attempt to overcome this, we also investigated what would happen if we reduced the size of the misspecification tested in

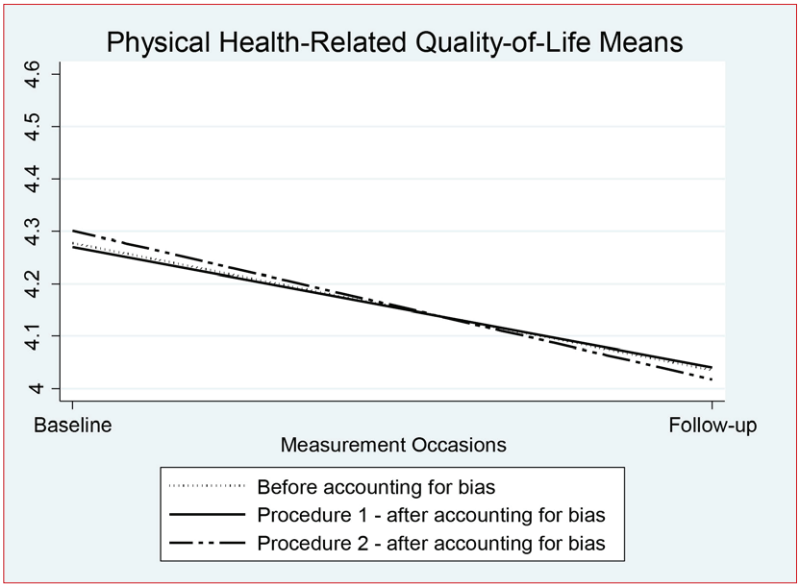

Figure 3a Physical HRQoL mean change

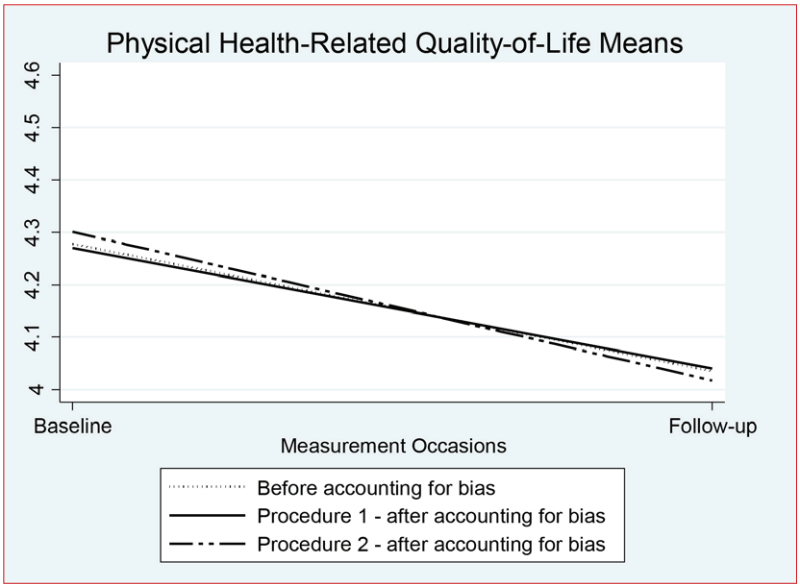

Figure 3b Mental HRQoL mean change

Procedure 1. This resulted in identifying different results than those reported in the results section for both Procedure 1 and Procedure 2. For example, when misspecification associated with the intercepts was reduced to ≥.10 rather than ≥.20. The results suggested equality constraints on the intercepts of Nausea over measurement occasions and the intercept of Cognitive Functioning at follow-up were not tenable. This was because the MI was significant, the power was moderate and the EPC was large. While the finding regarding Nausea is consistent with the results of Procedure 2, the finding regarding Cognitive Functioning intercepts was not. Therefore, simply reducing the misspecification size does not resolve the differences found in the results for both procedures. In this illustrative example, we are limited because we do not know where the true bias exists. To resolve this limitation, future work is needed where data are simulated and both procedures are applied to identify biased parameters.

Both procedures presented in this paper offer the flexibility to allow the researcher to choose what size of misspecification they believe to be acceptable. However, it appears that the decision to test single or multiple parameters needs to be taken into consideration as it has the potential to affect the results. Also, using Cohen small effect sizes to indicate substantive measurement bias in both procedures may or may not be ideal, (though was initially suggested by Kaplan (1990). This is because limited work has been conducted that investigates appropriate misspecification sizes (Whittaker, 2012). Further research that investigates the impact of decisions with respect to the size of the misspecification under different circumstances (i.e., sample size, number of parameter under consideration) is needed to guide applied researchers wishing to apply either procedure.

Both procedures presented in this paper aim to guard against chance findings, and both procedures are an improvement on relying on MIs alone; however, there are limitations. In Procedure 1, power was calculated for each MI, however, knowing the power in our illustrative example did not aid the decisions regarding the tenability of the equality constraints. This is because the power was very high (> .97) for all parameters investigated. Therefore, all decisions regarding the tenability of the equality constraints when the MI was significant were based on the size of the EPC. However, the EPC should be informative regardless of the power and when investigating both MIs and EPCs the Type II error rate should be reduced. This may be especially true in the presence of multivariate non-normality. This is because the MI may be affected by the multivariate non-normality, but the EPC should be relatively stable as they are associated with the parameter estimates (Hoogland & Boomsma, 1998). Parameter estimates are less affected than the standard errors and MIs, though no research, to our knowledge, has been conducted regarding the conditions under which the EPC remains stable. In this paper, the EPC predicted a larger change in the intercept associated with Emotional Functioning than was actually observed. This is an area in need of further research.

In Procedure 2, unlike Procedure 1 where the calculations are easily computed with the aid of JRule, all alternative models were run and the SOPCs were calculated separately. This is a very time-intensive process, and while this aspect of

the procedure is disadvantageous, it can also be considered to be advantageous. This is because each constraint is given substantial consideration before it is determined whether the equality constraint is tenable. As the process of testing equality constraints is essentially a data-driven process, by testing each alternative model, the researcher is able to observe the impact of equality constraint removal on all estimated parameters. This in turn should improve substantively based decisions, especially when the bias is marginal.

In this paper we relied on the factor loadings and intercepts to provide scale and origin to the latent constructs. While this is the most frequently used approach, it is not the best choice when testing invariance. This is because by setting a factor loading and intercept to a constant to provide scale and origin we assume that these parameters are invariant. While procedures have been proposed for ensuring that invariant parameters are used to provide scale and origin (Byrne, Shavelson & Muthen, 1989; Rensvold & Cheung, 2001) these procedures are not frequently used and are time consuming to carry out. An alternative for setting scale and origin is to constrain the factor variances of the first measurement occasion to equal unity, and to include at least one factor loading that is constrained to equality over measurement occasions (Oort, 2001; Reise et al., 1993; Yoon & Millsap, 2007). In this paper, we scale via factor loadings and intercepts, to highlight the most frequently used form of scaling and because the computations calculations for Procedure 1 required this form of scaling. The global tests and SOPCs can be calculated regardless of scaling, and thus provide the researcher with the flexibility to proceed using methods that are suited to their research needs.

In conclusion, both procedures can be used to detect measurement bias. However, we believe that additional step of transferring the output to the JRule program in Procedure 1 does not provide any additional information when the sample size is adequate as the power will be high. Had we disregarded the power of MI then we would not have made any different decisions than those we came to by using Jrule. Procedure 2 requires additional calculations; however, we believe if the primary research question revolves around detecting measurement invariance then Procedure 2 is exhaustive and provides empirical evidence with respect to biased parameters.

References

Aaronson, N. K., Ahmedzai, S., Bergman, B., Bullinger, M., Cull, A., Duez, N. J., Filiberti A., Flechtner H., Fleishman S. B., de Haes, J.C. (1993). The European Organization for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology. *Journal of the National Cancer Institute, 85,* 365-376.

Bergman, B., Aaronson, N. K., Ahmedzai, S., Kaasa, S., & Sullivan, M. (1994). The EORTC QLQ-LC13 - A modular supplement to the EORTC core quality-of-life questionnaire (QLQ-C30) for use in lung-cancer clinical-trials. *European Journal of Cancer, 30A*, 635-642.

Byrne, B. M., Shavelson, R. J., & Muthen, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105*, 456-466.

Chou, C. P., & Bentler, P. M. (1993). Invariant standardized estimated parameter change for model modification in covariance structure-analysis. *Multivariate Behavioral Research, 28*, 97-110.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawerence Erlbaum associates.

Hochberg, Y., & Benjamini, Y. (1990). More powerful procedures for multiple significance testing. *Statistics in Medicine, 9*, 811-818.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*, 65-70.

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

Kaplan, D. (1989). Model modification in covariance structure-analysis: Application of the expected parameter change statistic. *Multivariate Behavioral Research, 24*, 285-305.

Kaplan, D. (1990). Evaluating and modifying covariance structure models: A review and recommendation. *Multivariate Behavioral Research, 25*, 137-155.

King-Kallimanis, B. L., Oort, F. J., & Garst, G.J.A. (2010). Using structural equation modelling to detect measurement bias and response shift in longitudinal data. *Asta-Advances in Statistical Analysis, 94*, 139-156.

MacCallum, R.C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure-analysis: The problem of capitalization on chance. *Psychological Bulletin, 111*, 490-504.

Mellenbergh, G.J. (1989). Item bias and item response theory. *International Journal of Educational Research, 13*, 127-143.

Meredith, W., & Horn, J. (2001). The role of factorial invariance in modeling growth and change. In L. M. Collins, & A. G. Sayer (Eds.), *New methods for the analysis of change. Decade of behavior* (pp. 201-240). Washington DC, US: American Psychological Association.

Neale, M. C.(2009). MxGui (Version 32) [Software]. Available from http://www.vcu.edu/mx/.

Oort, F. J. (2001). Three-mode models for multivariate longitudinal data. *British Journal of Mathematical & Statistical Psychology, 54*, 49-78.

Oort, F. J . (2005). Using structural equation modeling to detect response shifts and true change. *Quality of Life Research, 14*, 587-598.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psycholocial Bulletin, 114*, 552-556.

Rensvold, R. B., & Cheung, G. W. (2001). Testing for metric invariance using structural equation models: Solving the standardization problem. In C. A. Schriesheim, & L. L. Neider (Eds.), *Equivalence in measurement. Research in management series (Vol. 1* (pp. 25-50). Greenwich, Connecticut: Information Age Publishing.

Saris, W. E., Satorra, A., & Sörbom, D. (1987). The detection and correction of specification errors in structural equaltion models. *Sociological Methodology, 17*, 105-129.

Saris, W. E., Satorra, A., & Van der Veld, W. M. (2009). Testing structural equation models or detection of misspecifications? *Structural Equation Modeling-A Multidisciplinary Journal, 16*, 561-582.

Sayer, A. G., & Cumsille, P. E. (2001). Second-order latent growth models. New methods for the analysis of change. In L. M. Collins, & A. G. Sayer (Eds.), *New methods for the analysis of change. Decade of behavior* (pp. 179-200). Washington DC, US: American Psychological Association.

Tishelman, C., Degner, L. F., Rudman, A., Bertilsson, K., Bond, R., Broberger, E., Doukkali, E., & Levealahti, H. (2005). Symptoms in patients with lung carcinoma: Distinguishing distress from intensity. *Cancer, 104*, 2013-2021.

van der Veld, W. M., Saris, W. E., & Satorra, A. (2008). JRule 2.0: User manual. 3.0.4.

Vandenberg, R.J., & Lance, C.E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*, 4-70.

Whittaker, T. A. (2012). Using the modification index and standardized parameter change for model modification. *Journal of Experimental Education, 80*, 26-44

Yoon, M., & Millsap, R. E. (2007). Detecting violations of factorial invariance using data-based specification searches: A Monte Carlo study. *Structural Equation Modeling-A Multidisciplinary Journal, 14*, 435-463.

BELLINDA KING-KALLIMANIS

Research fellow working on the Irish Longitudinal Study on Ageing at Trinity College Dublin, Ireland. Her PhD was completed at the University of Amsterdam and her dissertation project, titled 'Unbiased measurement of health-related quality-of-life', focused on issues of measurement invariance and response shift using structural equation modelling.

FRANS OORT

Full professor of Methods and Statistics in Educational Research, director of the Graduate School of Child Development and Education, and program director of the Research Master Educational Sciences at the University of Amsterdam.

CAROL TISHELMAN

Professor of Care Sciences at the Department of Learning, Informatics, Management and Ethics Nursing at the Karolinska Institutet, Sweden. Her research aims to demonstrate new ways for people to cope with their situation. She is studying the symptoms of people with cancer, both before and after diagnosis and aims to bring together nursing research, training and clinical activity.

MIRJAM SPRANGERS

Professor in Medical Psychology and one of the Academic Medical Center's principal investigators. Her focus of research is on patient-reported outcomes (e.g., quality of life, health care needs), response shift (i.e. shifts in patients' perspectives over time) and genetic dispositions.