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

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;

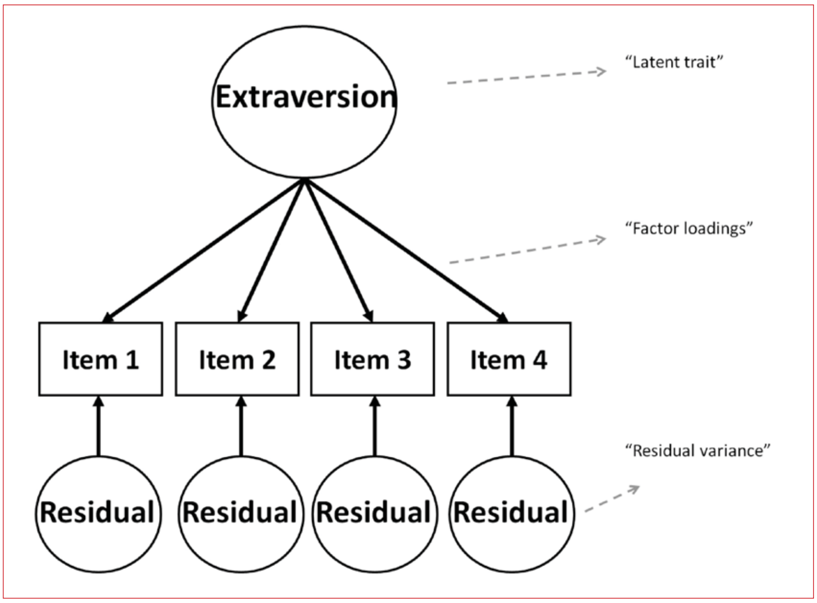

Figure 1 Example of a latent trait model for the psychological construct 'extraversion'

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.

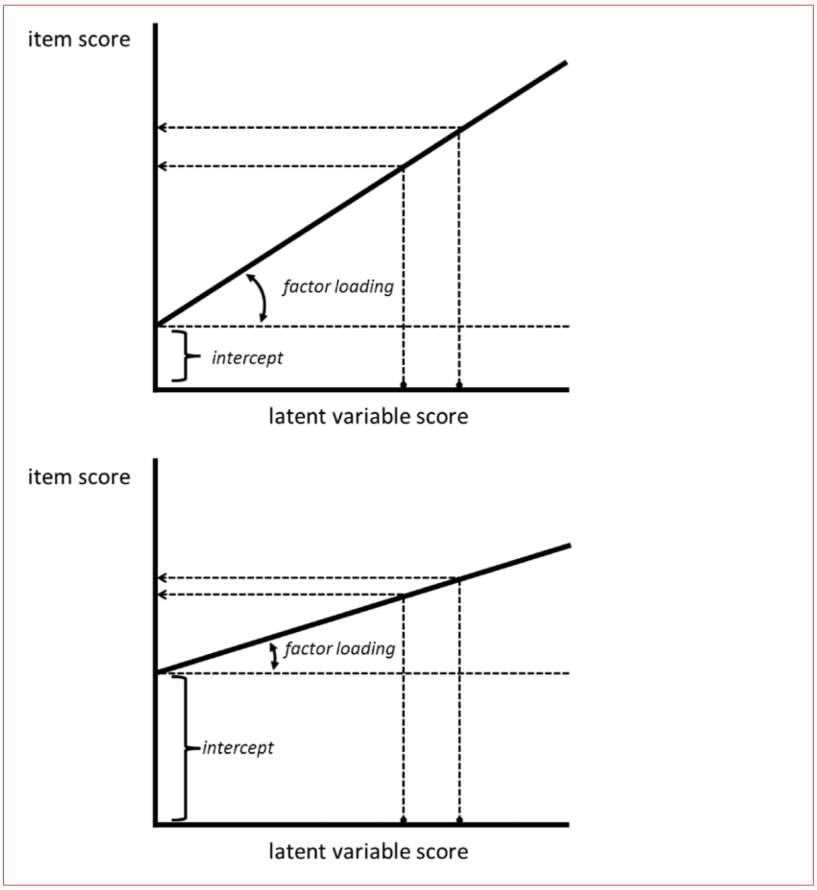

Figure 2 Illustration of how factor loadings account for the degree to which variability in the latent trait is captured by the item scores (in case of –approximately– continuous item scores)

Table 1 Overview of the different types of latent trait models

Item scale	Latent trait scale	
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.

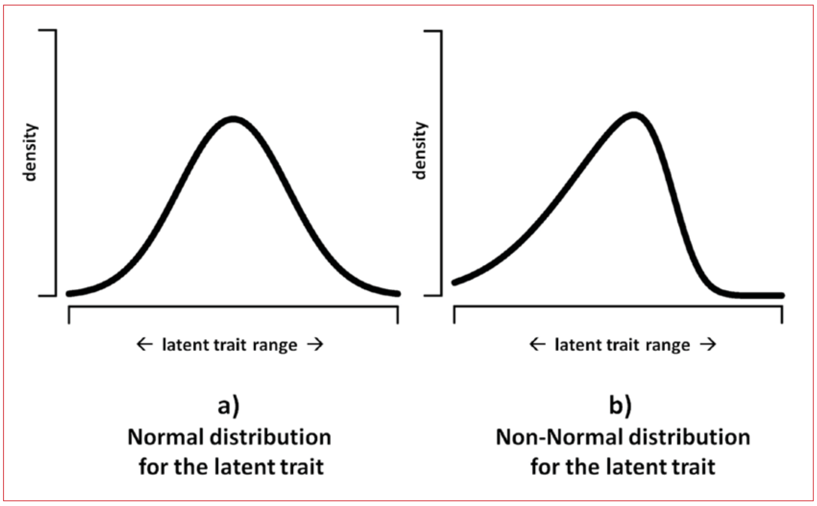

Figure 3. A) The assumption of a normal distribution for the latent trait; B) a violation of this assumption

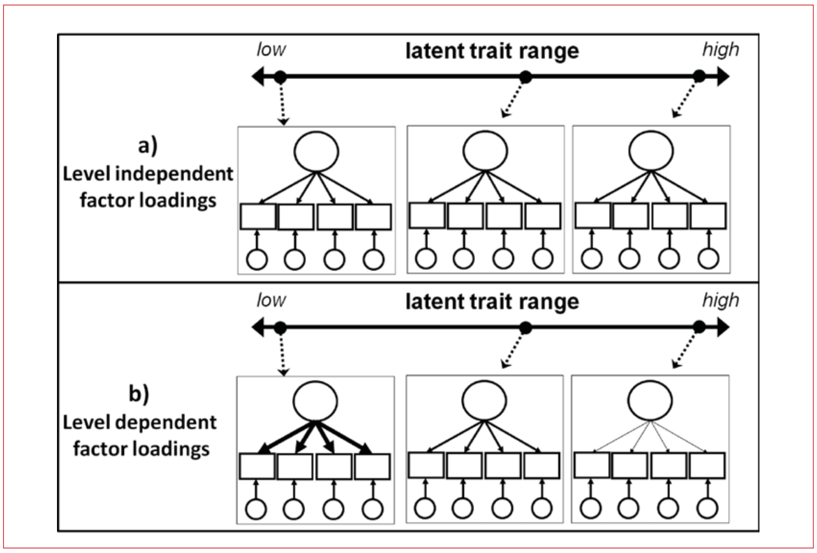

Figure 4. A) The assumption of level independent factor loadings; B) a violation of this assumption, the factor loadings are larger for increasing levels of the latent trait

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.

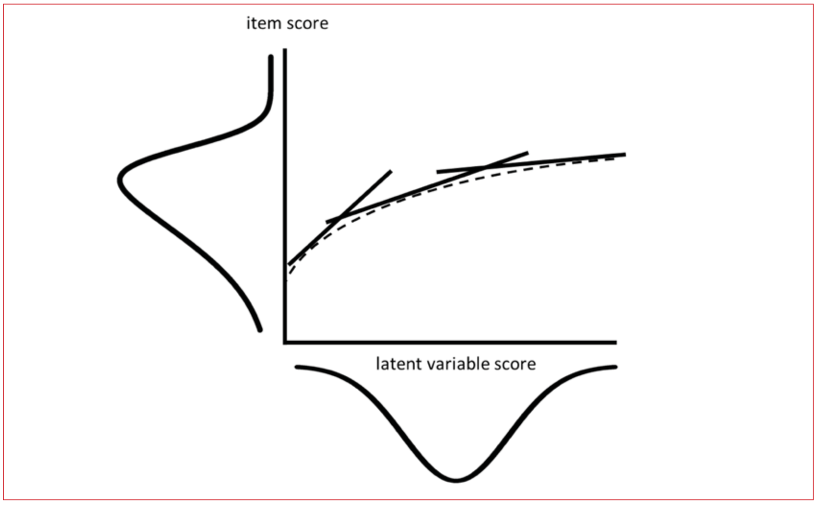

Figure 5. Illustration of how level dependent factor loadings will result in a non-normal distribution of the item scores (in case of –approximately– continuous item scores)

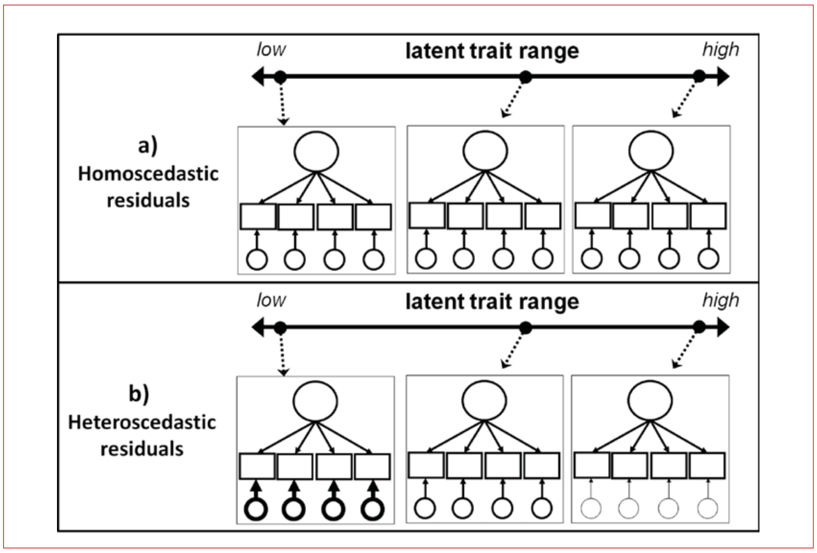

Figure 6. A) The assumption of homoscedastic residuals; B) a violation of this assumption, the residuals are heteroscedastic, i.e., they increase across the latent trait

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

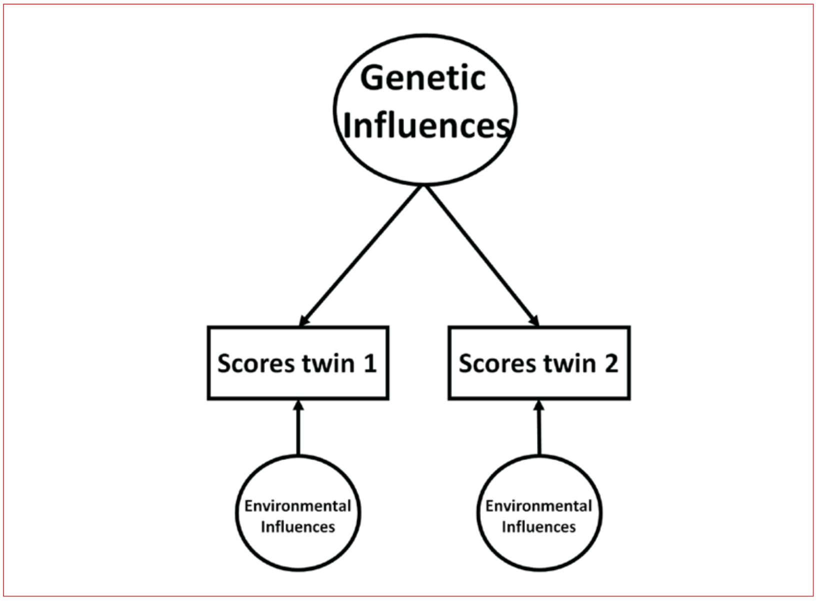

Figure 7. The classical twin design as a latent trait model. Note that this model is for monozygotic twins only

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.