Testing distributional assumptions in psychometric measurement models with substantive applications in psychology

Molenaar, D.

Publication date
2012

1
Introduction

1.1 Latent Variables in Psychology

In the behavioral sciences and psychology in particular, hypotheses generally include statements about unobservable psychological constructs like perceptual organization, attachment, depression, extraversion, and working memory. As we can not observe these so called latent variables directly, we need to focus on observable indicators of these variables to be able to test hypotheses about them. For example, in case of the latent variable ‘perceptual organization’, we can consider the performance of a sample of subjects on the Block Design and Matrix Reasoning subtests of the Wechsler Adult Intelligence Scale III (WAIS-III; Wechsler, 1997). As solving the items of these tests is purported to involve perceptual organization, the scores can be taken as indicators of the latent variable ‘perceptual organization’.

1.1.1 The Generalized Linear Item Response Model

To operationalize a latent variable in terms of the observed variables, a measurement model is formulated in which the scores on the observed variables are linked to the latent variable (Borsboom, 2008). Within the family of parametric measurement models, the exact choice depends on the distribution of the observed and latent variables. A wide range of popular measurement models are included within the more general family of measurement models referred to by the Generalized Linear Item Response Model (GLIRM; Mellenbergh, 1994; see also Borsboom, 2008). The GLIRM is given by

$$z_i = \upsilon_i + \lambda_i \eta + \varepsilon_{i.} \qquad (1)$$

where z_i is a continuously distributed variable associated with the *i*-th observed variable, υ_i is an intercept, λ_i is a regression coefficient or factor loading, η is the latent variable, and ε_i is a residual. Special cases of the GLIRM arise by specifying a distribution for the observed variables, y_i, and linking this distribution to the continuous distribution of z_i in Equation 1 (see Mellenbergh, 1994).

Although the GLIRM also includes models for nominal latent variables, we focus on cases in which the latent variable is continuously distributed, as these models are appropriate measurement model for many psychological constructs, e.g., cognitive abilities like working memory, and personality traits like extraversion. As a consequence, nominal constructs like for instance attachment style (Bowlby, 1969) or Piagetian conservation (Piaget, 1952) are not covered. Specifically, this dissertation focuses on two important special cases of the GLIRM: the linear factor model (Spearman; 1904; 1927) and the graded response model (Samejima, 1969). The linear factor model is an important measurement model as it dominated psychometric studies into cognitive ability for the past 100 years. In addition, it is considered an appropriate measurement model for observed variables with 7 or more levels (Dolan, 1994; Mellenbergh, 1994) which is common in psychology. For instance, observed variables are often scores on questionnaires that use a 7 point Likert scale, or the observations are subtest scores of scales with dichotomous items. From Equation 1, the linear factor model arises by specifying

$$f(y_i) = h(z_i) \tag{2}$$

i.e., the probability density of y_i is equal to the probability density of z_i. At this point, the exact probability density of z_i is unspecified; we will specify it later, implying a specific distribution for the observed data.

The graded response model is the second special case of the GLIRM that we consider in this dissertation. This model is appropriate for ordinal observed variables. The graded response model is especially useful in cases where the observed variables have less than 7 levels, as then, the linear factor model does not suffice (Dolan, 1994). Such observed variables are generally encountered in the field of personality research where many questionnaires consist of Likert scales with less than 7 answer categories. If the discrete data, y_i, is coded as $y_i = 0, \ldots, C$, the graded response model arise from Equation 1 by

$$p(y_i = c) = \int_{\tau_{ic}}^{\tau_{ic+1}} h(z_i)dz_i \qquad \text{for c} = 0, \ldots, C \tag{3}$$

i.e., the probability distribution of y_i is given by categorizing the continuously distributed z_i at given thresholds, τ_{ic}, where $\tau_{i0} = -\infty$, and $\tau_{iC} = \infty$ (see Wirth & Edwards, 2007). Other special cases of the GLIRM, which are not covered in this dissertation, include the nominal response model (Bock, 1972) and the partial credit model (Masters, 1982), see Mellenbergh (1994).

1.1.2 The Structural Model

When an appropriate measurement model is specified, hypothesis about the latent variable can be tested in the structural model (Verhelst & Verstralen, 2002; Zwinderman, 1991). The latent variable could be regressed on certain background variables including observed variables like age and gender, or other latent variables like depression and working memory. In doing so it could for instance be tested whether males and females differ in their level of perceptual organization or whether working memory is related to perceptual organization.

To enable structural modeling, a distribution for z_i should be specified. Commonly a normal distribution is used, i.e.,

$$h(z_i) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left[-\frac{1}{2}\frac{(z_i-\mu_i)^2}{\sigma_i^2}\right]$$

where μ_i and σ_i^2 are respectively the mean and variance of z_i. Specifying this specific distribution for z_i implies that in case of the linear factor model, the observed data, y_i, is normally distributed because of Equation 2. In addition, in the graded response model it is assumed that a normally distributed variable underlies the ordinal observed data because of Equation 3. Even though a normal distribution can undoubtedly be a reasonable approximation for the true distribution of z_i in some areas of application, in the behavioral sciences a non-normal distribution for the data is not exceptional. For instance, Micceri (1989) collected 440 datasets containing psychometric and achievement measures and found all measures to be non-normally distributed.

1.2 Testing for normality

As we will point out in this dissertation, within the GLIRM, non-normality in z_i can have multiple causes which are not necessarily mutually exclusive. Testing the specific loci of non-normality within the GLIRM is important as hypotheses in the behavioral sciences often have distributional implications. In this dissertation we discuss Ability Differentiation (Spearman, 1927; Deary et al, 1996), Schematicity (Markus, 1977; Rogers, Kuiper & Kirker, 1977; Tellegen, 1988), and Genotype by Environment interaction (Jinks & Fulker, 1970; Molenaar & Boomsma, 1987; van der Sluis et al., 2006). All these phenomena imply specific departures from normality in z_i. It is important to disentangle the different sources of non-normality within the GLIRM, as for instance heteroscedasticity of the residuals in Equation 1 can not be taken as evidence for Ability Differentiation for instance, whereas non-linear factor loadings can. In addition, besides these substantive reasons to investigate specific departures from normality in z_i, there are some statistical reasons to take non-normality into account. For instance, Curran, West, & Finch (1996) show that in case of the linear factor model, non-normal observed data results in biased parameter estimates and biased goodness-of-fit measures of models like Equation 1 when using the traditional method of Maximum Likelihood estimation (ML; Lawley, 1943). There exist alternatives to ML estimation that appear to be more robust to violations of normality, i.e., the Asymptotic Distribution Free estimation procedure (ADF; Browne, 1984) and the rescaled χ^2 procedure (Satorra & Bentler; 1988). However these methods require large sample sizes and may not be appropriate in case of models with many indicators (see Curran et al., 1996). More optimistic results are obtained in case of ordinal data (Flora & Curran; 2004; Stone, 1992), but still non-normality can bias results under specific circumstances (e.g., small sample sizes and small model size; Kirisci, Ksu, & Yu;2001). In addition, other studies show that non-normality can result in bias (e.g., Azevedo, Bolfarine, & Andrade, 2011; Swaminathan & Gifford, 1983).

Without proper statistical tests it remains unclear when the assumption of a normal distribution for z_i is not reasonable. It would thus be valuable to have explicit and specific tests at our disposal to test the normality assumption of z_i. When we do not find any departures, we can confidently interpret the results of the traditional latent variable analyses (i.e., using a normal distribution). However, if we find departures from normality, we can investigate whether it makes a large difference when we use the normal distribution nevertheless. If large parameter bias is found, it is safer to take the non-normality in z_i into account in the statistical model.

1.3 Outline

In this dissertation we present statistical approaches to test for specific departures from normality in z_i. We will focus on the models as statistical devices to test the normality assumption of z_i per se. In addition, we discuss substantive applications of the models. The outline of this dissertation is as follows: In *Chapter 2,* we show that within the linear factor model, we can test for normality in z_i by testing in Equation 1 for

#1: heteroscedasticity of ε_i,
#2: non-normality of η, and
#3: level dependency of λ_i.

We show that in reasonable sample sizes, we can statistically distinguish between #1 and #2, and we can distinguish between #1 and #3, but #2 and #3 are not resolvable. Using these results in *Chapter 3,* we propose an extension of the graded response model to include #1 and #2. We apply this model to test the Schematicity hypothesis. As the previous chapters concerned a one factor model, we propose an extension to the second-order factor model in *Chapter 4*. We use the Schmid-Leiman decomposition (Schmid & Leiman, 1957) to include #1, #2, and/or #3 in the second-order factor model. This extension is conducted specifically with an application to Ability Differentiation in mind. In *Chapter 5*, it is shown how a proxy for η in Equation 1 can be used to conduct tests on #1, #2, and #3. These tests are more flexible, as the proxy could be replaced by any other observed variable (e.g., age) to test for moderation of the parameters in Equation 1 (see Bauer & Hussong, 2009). We illustrate this advantage by means of an application to age differentiation (Garrett, 1946). In *Chapter 6* the methodology concerning tests on #1 is extended to incorporate multiple factors and multiple groups. These developments are conducted with the specific aim to test for Gene by Environment (GxE) interactions in twin data. *Chapter 7* involves an application of these twin models including #1, to a large dataset comprising over 11,000 twin pairs. In *Chapter 8,* we discuss how interpreting statistical effects in terms of substantive hypothesis (like ability differentiation and GxE) can be risky, and what can be done about it. In addition, we present some concrete ideas for applications and further model development within the framework outlined in this dissertation. As nearly all Chapters in this dissertation include power analyses, we added an Appendix that contains an elaborate illustration of how power can be calculated in models like Equation (1). The illustration focuses around the issue of power to detect sex differences in intelligence test scores.