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

Molenaar, D.

Publication date
2012

8

Discussion

When non-normality is detected statistically in the observed scores, a conclusion like "a non-normal distribution underlies these data" is straightforward and will elicit little criticism. However, it is much less straightforward to draw conclusions on possible substantive effects underlying this non-normality, as statistical effects are not necessarily amenable to a meaningful psychological interpretation. In present concluding Chapter, we discuss under what circumstances substantive conclusions on non-normality are risky, and what can be done about it. In addition we argue that a substantive effect is ideally also investigated within an appropriate psychological process model. We end with some concrete ideas of how the methods in this dissertation could be used in future statistical, psychometric, and/or substantive work.

8.1 Drawing Substantive Conclusions about Statistical Effects

In this dissertation we discussed how hypotheses in psychology can have distributional implications within the Generalized Linear Item Response Model (GLIRM; Mellenbergh, 1994). Specifically, we showed how 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) can result in non-normality. Testing these hypotheses at specific loci within the GLIRM has turned out to be valuable in terms of power and specificity. However, substantive conclusions should be drawn with care, as there are more sources of non-normality, which could be plausible alternative explanations for the effects in the data. Here, we discuss 3 common causes of non-normality: censoring, bad scaling, and the use of unrepresentative samples. In addition, we discuss possible solutions to rule out these alternative explanations.

8.1.1 Censoring

When the difficulties of the items of a test do not match the ability of the sample to which it is administered, censoring may occur. When a test is too difficult for a given sample, data may be left censored (floor effect), or when a test is too difficult, the data may be right censored (ceiling effect). If the linear factor model is applied to censored data, heteroscedastic residual variances will result. Luckily, it is relatively easy to find

out whether heteroscedastic residual variances are due to censoring, as floor and ceiling effects will be noticeable in e.g., a QQ-plot or a histogram of the data. Statistically, censoring poses no problem as the heteroscedasticity parameters take the censoring effects into account. However, from a substantive point of view, censoring can distort substantive interpretations of specific departures from normality in the data. For instance, the schematicity hypothesis (Chapter 2), predicts that residuals are heteroscedastic with less variance towards the lower end of a personality scale. Within the linear factor model, this specific effect could easily arise when a given observed personality measure (item score or subtest score) is associated with a floor effect. Similar could happen in case of testing for a Gene by Environment effect (GxE; see Chapter 5 and Chapter 6), i.e., the GxE effect can be due to censoring.

8.1.2 Bad Scaling

Another source of non-normality is bad scaling. When item difficulties of a test are not distributed uniformly across the underlying latent variable, the sum score of these items is likely to be badly scaled. Depending on the distribution of the item difficulties, information included in the sum score about the latent variable can differ across the levels of that variable. For instance, if a test contains a disproportional number of easy items, information in the sum score about the latent variable is higher for lower levels of that variable. In this case, heteroscedastic residuals will arise, with more residual variance at the upper end of the latent variable. As in case of censoring (see above), this is no problem from a statistical point of view, as the heteroscedasticity effect can be used to indicate and correct bad scaling. The exact form of the heteroscedasticity that should be modeled, i.e., linear and/or curvilinear effects, depends on the distribution of the item difficulties. In Figure 8.1, a graphical representation is given of the form of the heteroscedasticity for common situations. As appears from the Figure, if items are uniformly distributed across the latent variable, their sum score will be homoscedastic in the linear factor model. If items are disproportionate easy, the sum score will be associated with residuals that increase across the latent variable. A linear model on the logarithm of the residual variances (see Chapter 1) can be used to model this kind of bad scaling. For cut-off tests (i.e., tests in which information at a specific level of the latent variable is maximized to make cut-off decisions) the residuals are larger at the extremes and smaller near the cut-off. In this case, heteroscedasticity can be modeled using a curvilinear function for the residual variances. Finally, if the sum score consists of a disproportionate number of easy and difficult items, either a quadratic or a curvilinear model should be used. The exact choice depends on whether the number of easy/difficult items is approximately equal. If so, a quadratic model will suffice. If not, a curvilinear model is needed. Note that in case of ordinal data, bad scaling can also

occur on item level. If the item category location parameters are not well distributed across the latent variable, the items scores will show heteroscedasticity in the linear factor model in a comparable way to the situations outlined in Figure 8.1 (then the vertical lines in the top panel represent item category parameters).

As opposed to censoring, bad scaling may not be clear by visualizing the data like in a QQ-plot or histogram. This is most problematic when an effect is interpreted substantively, e.g., in case of GxE research. It could be unclear whether a given set of subtest scores are heteroscedastic due to bad scaling or due to a GxE effect.

8.1. 3 Unrepresentative Samples

When a sample of subjects is not representative of the population, non-normality could arise in the data. For instance, in intelligence research, subjects low on general intelligence could be less willing to visit the laboratory to complete an IQ test battery. In principle, this will result in a skewed latent variable distribution in the sample (or level dependent factor loadings, see Chapter 1). The observed score distribution is then not an adequate reflection of the true population distribution. This complicates tests on ability differentiation (Chapter 3 and Chapter 4) which predicts that the general intelligence factor is distributed with a thinner upper tail (i.e., negatively skewed). In unrepresentative samples, it could be hard to detect this effect as the lower tail could also be thinner due to the biased sampling. On the contrary, a differentiation effect can be an artifact because subjects in the upper tail of the general intelligence distribution are less well represented (because they are harder to recruit for instance).

8.1.4 Solution

In case of substantive interpretations of specific departures from normality (e.g., heteroscedastic residuals), alternative explanations like censoring, bad scaling, and unrepresentative samples should be ruled out. To start with the latter, unrepresentative samples are a problem for the validity of almost any statistical test in psychology. For instance, in testing for sex differences in intelligence test scores (e.g., Dolan, et al., 2006), it is important to have representative samples of both males and females. A reviewer of the article that is in Appendix A of this dissertation suggested that females are more willing to participate in IQ studies than males (which may suggest that the males who do participate are brighter than the females). If the reviewer is right, comparing the intelligence test scores of the male and female sample is meaningless whatever advanced statistical method is used. A standard statistical tool to correct for biased samples is to put more weight on the scores of underrepresented subjects. This procedure is common in marketing research, but it is also developed within latent

variable modeling (Asparouhov, 2005). However, the method should be used with care, as validity of the results depends highly on the chosen weights.

Solutions to the problem of censoring and bad scaling depend on the data that are available to the researcher (i.e., item level, or subtest level data) and the measurement properties of the data (i.e., dichotomous or ordinal; uni-dimensional or multi-dimensional). If there are item level data available, and these data are ordinal and uni-dimensional, best solution to censoring and bad scaling is to consider the graded response model and test for heteroscedasticity within this model using the methods outlined in Chapter 2. Alternatively, in case of multi-dimensional data and/or dichotomous data, appropriate methods are not available yet (but see the discussion on this matter later). In these cases, we propose to estimate the ability parameters of the subjects in the sample for all dimensions, using an appropriate standard item response model (e.g., a 2 or 3 parameter item response model). Next, these estimates could be analyzed using the linear factor model for non-normal data as outlined in Chapter 1. In both of these procedures, a possible censoring effect is absorbed by the item difficulty parameters, making the ability estimates free of such an effect.

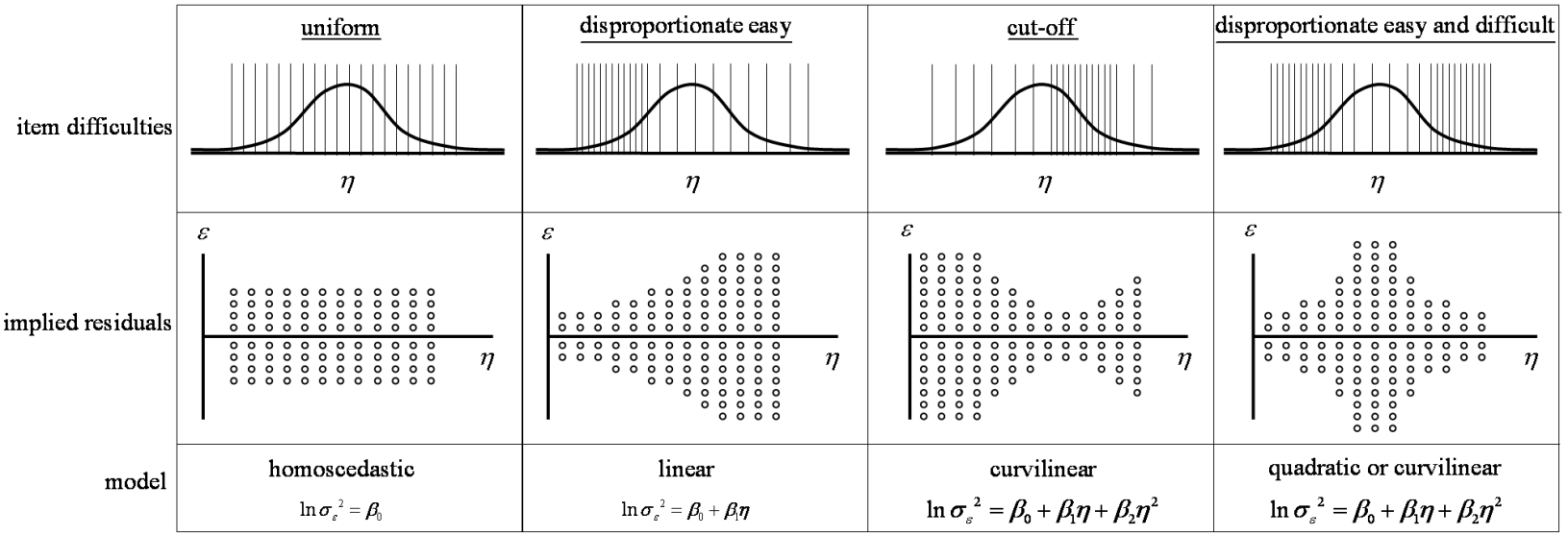

Figure 8.1. Effects of the item difficulties distribution on the residuals, ε, when the sum score of the items is analyzed in the linear factor model. Vertical lines in the top plots are the location of the item difficulties on the η scale. 'model' refers to the model for heteroscedasticity that could be used in the corresponding situations.

When no item level data are available, it is difficult to evade the censoring problem. If the censoring effect is only limited to a few subtests, a practical solution might be to omit these subtests from the data. Results could then be considered with and without these subtests in the analysis. If the absence of these subtests does not influence the results for the other subtests, heteroscedasticity can be carefully interpreted in terms of the substantive effect (e.g., in terms of schematicity or GxE). When there are only subtests scores available to the researcher, bad scaling could always be an alternative explanation for an effect on the residuals. Sometimes, information on the scaling of the items could be tracked, e.g., in case of intelligence test scores, the manual of the specific test could be consulted. However, if there appears to be an under or over representation of easy items, no solution remains. In that case, it is best to be really conservative in drawing substantive conclusions about the statistical effects.

8.2 Beyond statistical testing: Process models for psychological phenomena

In this dissertation we focused exclusively on the *detection* of substantively interpretable effects, e.g., ability differentiation was operationalized as an effect on the distribution of the general intelligence factor and schematicity was operationalized as an effect on the residual variances. A next step in studying these kinds of hypotheses might be to consider the psychological process behind these phenomena. Ability differentiation for instance, implies decreasing intelligence subtest correlations across the general intelligence factor. When this effect is empirically detected, question raises what (psychological) process underlies such a diminishing general intelligence factor. This question enjoyed little attention in the literature. Some explanations of ability differentiation are proposed among which: the 'minimal cognitive architecture' explanation (Anderson, 1992) and the 'economic investment' explanation (Brand, 1984). The minimal cognitive architecture explanation assumes Basic Processing Mechanisms (BPM) to be involved in intelligent behavior. The faster your BPM, the more complex algorithms you can implement. These algorithms are then used to solve problems like those in intelligence tests. Subjects with a fast BPM (i.e., high level of *g*) have dissimilar responses to intelligence tests, depending on the implemented algorithm. As a consequence, subjects with a slow BPM have highly similar responses as implemented algorithms do not differ majorly. An alternative but related theory, the economic investment theory, explains the differentiation by comparing the level of *g* to the amount of money that someone has to spend. Poor people do not differ importantly in where they spend their money on, most of their money is used to pay rent and food. Rich people differ relatively more as they spend their money on highly differing goods (e.g., one person spends a lot of money on going to theatres and concerts, while someone else spends a lot of money on city trips and vacations). As appears from the

above, both explanations of ability differentiation are metaphoric which makes it hard to test specific predictions. An important step could be taken in studying phenomena like ability differentiation and schematicity, by formulation the effects explicitly within a psychological process model.

For ability differentiation a candidate process model is the mutualism model (van der Maas et al., 2006) which is mathematically well developed. In the mutualism model, the general intelligence factor is assumed to have risen because of mutually interacting basic processes during development (e.g., mutual interactions between memory, perceptual, and reasoning processes). A possibility in which ability differentiation can arise in this model is that a higher level on the processes[38] is associated with smaller interactions among them. When differentiation is formulated in such an explicit mathematical model as the mutualism model, it could be studied using computer simulations, and hopefully tested on empirical data in the near future (as at present, the mutualism model is not yet implemented). An alternative process model that could be used to explain ability differentiation is the Q-diffusion model (van der Maas, Molenaar, Maris, Kievit, & Borsboom, 2011). This model describes a specific information accumulation process that connects a latent variable to the observed response. Currently, the Q-diffusion model is only applicable to single abilities only (e.g., working memory). Within the simplest form of the Q-diffusion model (i.e., no response times are modeled), a response is elicited when a certain internal information accumulation process reaches a boundary. The latent variable form psychometric theory is interpreted as the efficiency or speed in which information is accumulated, and the item difficulty is interpreted as the height of the boundary. Now within this model, ability differentiation can be explained when higher efficiency of information accumulation is associated with a smaller boundary. If the full Q-diffusion model is considered (i.e., including reaction times), this formulation can have implications for the reaction time distribution of a given ability test. This is interesting, as to our knowledge, implications of ability differentiation on reaction times have not yet been investigated.

For personality constructs, appropriate process models are not readily available. However, as argued by van der Maas et al (2011), the 'traditional' diffusion model can be considered an adequate process model for attitude and personality items (see Tuerlinckx & de Boeck, 2005). However, this model concerns dichotomous item scores only, and personality questionnaires often have multiple answer categories. A new and promising line of research into process models for personality is the so-called network approach (Schmittman et al., in press), which is already successfully applied to

[38] We speak here of a 'higher level on the process', but this could also be read as 'more efficient processes' or 'better developed processes'.

comorbidity (Cramer et al., 2010). How the network approach could help personality research is a topic of current investigation (see Cramer et al., 2011).

8.3 Future statistical developments

Below we discuss some possible future developments of the models presented in this dissertation. In Chapter 2, we outlined a graded response model with heteroscedastic residuals and a non-normal latent variable distribution. Main idea was that ordinal item scores arise by categorization of an underlying continuously distributed variable at specific thresholds. To be able to introduce the heteroscedastic residuals, we identified the graded response model by fixing two adjacent thresholds. This identification constraint makes it unfeasible to generalize the model to models for dichotomous data like the Rasch model (Rasch 1960; Wright & Stone, 1979) or the 2 parameter item response model (Lord, 1952; Birnbaum, 1968), as these models only have 1 threshold. We think that this is a good topic for further research, as items from performance tests in educational measurement, or intelligence tests in psychology are commonly scored correct (1) and false (0). There are two possibilities: The first possibility remains in the framework of Chapter 2, as shortly described above. Within this framework, an alternative identification constraint could be thought of that does not exclude dichotomous data. A possibility is to fix the variance of the underlying continuous variable to 1 at a specific value of the latent variable. A similar constraint was earlier proposed by Medsland, Neale, Eaves, & Neale (2009) in an extension of the moderation model of Purcell (2002). However, in that case, the latent variable is an observed moderator. This makes it unclear whether their constraint can be used for the heteroscedastic models from Chapter 2. This could however be investigated. The second possibility to include heteroscedastic residuals in dichotomous data is to consider an alternative framework than that of Chapter 2. For instance, in logistic regression of observed dependent and independent variables, overdispersion is accounted for by

$$E(y) = \pi, \text{ and}$$
$$Var(y) = \varphi \times \pi \times (1- \pi)$$

where y is the observed dichotomous variable, and φ is an overdispersion parameter (see Agresti, 2002). Whether this approach is possible within the 2 parameter item response model and/or the Rasch model remains to be investigated.

In Chapter 3 we outlined the method of moderated factor analysis which is based on work of Bauer & Hussong (2009) and Purcell (2002). Idea is that the

parameters of the GLIRM are moderated by an observed variable e.g., for the factor loadings:

$$\lambda = \lambda_0 + \lambda_1 \times M$$

These models have important applications as models for measurement invariance with respect to a continuous background variable (see Bauer & Hussong, 2009), and to investigate GxE (Purcell, 2002). The deterministic nature of the function between the factor loadings and the moderator is a strong assumption. If the factor loadings are not constant across subjects, it is implausible that the moderator can account for all this variability. For instance, using the Purcell model, it has been shown that heritability of IQ is moderated by SES (Turkheimer, Haley, Waldron, D'Onofrio, & Gottesman, 2003) but also educational attainment (Johnson, Deary & Iacono, 2009). Thus, focusing only on e.g., SES in the moderated factor model, leaves some variability in the factor loadings due to educational attainment (assuming that SES and educational attainment explain at least some unique variance in the factor loadings). Neglecting this residual variability can distort tests on moderation and bias parameters. Extending the model for the factor loadings with a residual term results in:

$$y_i = \upsilon_i + \lambda_i \eta + \varepsilon_i, \text{ with}$$
$$\lambda_i = \lambda_{i0} + \lambda_{i1} M + \delta_i.$$
$$\rightarrow y_i = \nu_i + \lambda_{i0} \eta + \lambda_{i1} M \eta + \delta_i \eta + \varepsilon_i$$

where y_i is the i-th observed variable, υ_i is an intercept, η is the latent variable, ε_i is the residual of the observed variable and δ_i is the residual of the factor loading. As can be seen, the model contains a random factor loading and an interaction between the moderator and the latent variable. This model readily be fitted in Mplus (Muthén & Muthén, 2007) and WinBUGS (Lunn, Thomas, Best, & Spiegelhalter, 2000). However, the exact benefits in terms parameter accuracy and power for e.g., investigating GxE remains to be investigated.

Other possibilities are extensions of the uni-dimensional heteroscedastic graded response model with a skew-normal trait (Chapter 2) to multiple dimensions (see for a discussion on multi-dimensional item response model, Reckase, 2003). These extensions are valuable as psychological tests are often multi-dimensional (e.g., intelligence tests like the WAIS; Wechseler, 1997, and personality tests like the NEO-PI, Costa & McCrae, 1992). In Chapter 2 we analyzed the Bermond-Vorst alexithymia questionnaire (Vorst & Bermond, 2001) which comprises 6 dimensions. We considered each dimension in isolation which is more or less common practice when sample size is too small for a multi-dimensional analysis. However, by doing so, we neglected the

inter-dimension correlations which raises the question how this affects heteroscedasticity and non-normality of the dimensions. In principle an extension of the Chapter 2 model to include multiple dimensions is straightforward: The multivariate skew-normal distribution is well developed (see Azzalini & Capatanio, 1999), and Chapter 5 already outlines how to handle heteroscedasticity in case of 2 dimensions. As already noted in Chapter 1, challenging to this undertaking will be the numerical feasibility of such an extension, as with increasing dimensionality of a psychometric measurement model, computation becomes more and more demanding. In the framework of maximum likelihood estimation as used throughout this dissertation, the maximum number of practically feasible dimensions equals 5 (Wood et al., 2002). For more dimensions, a Bayesian approach can possibly be considered.