UNIVERSITAT ROVIRA I VIRGILI

Unrestricted item factor analysis and some relations with Item Response Theory

Pere J. Ferrando

Urbano Lorenzo-Seva

Tarragona 2013

Please reference this document as:

Ferrando, P.J. & Lorenzo-Seva, U. (2013). *Unrestricted item factor analysis and some relations with item response theory. Technical Report*. Department of Psychology, Universitat Rovira i Virgili, Tarragona.

Document available at: http://psico.fcep.urv.cat/utilitats/factor/

Contents

1. The unrestricted Factor Analysis model: Some basic results

The Factor Analysis (FA) model can be viewed as a regression model in which the independent variables (i.e., the factors) are latent variables, and the predicted variables (for example, the answers of a person to a particular set of items in a questionnaire) are observed variables. The simplest case is Spearman's unidimensional model. If X_{ij} represents the score of individual i in the observed variable X_j, the unidimensional model can be written as:

$$X_{ij} = \mu_j + \lambda_j \theta_i + \varepsilon_{ij}, \qquad (1)$$

where μ_j is the mean of observed variable X_j, λ_j is the regression weight of observed variable X_j on the latent variable, θ_i is the non-observable level of individual i on the latent variable, and ε_{ij} is the measurement error associated to individual i in the observed variable X_j. The latent variable (that could also be referred to as *factor* θ) is considered to be a continuous unbounded variable. In addition, as the observed variable X_j is a linear combination of the latent variable θ, the observed variable X_j is also expected to be a continuous unbounded variable. The model makes the usual assumptions in linear regression: linearity and homoscedasticity,

$$E(X_j \mid \theta_i) = \mu_j + \lambda_j \theta_i; \qquad Var(X_j \mid \theta_i) = \sigma_{\varepsilon_j}^2, \qquad (2)$$

where σ_{ε_j} is the residual or error variance of observed variable X_j. Because factor θ is non-observable, some additional restrictions are needed to identify the model. A common restriction is to scale factor θ as a standard variable. Furthermore, in most of the procedures we shall discuss here, the observed variables are also scaled as standard variables (so that μ_j=0 and σ_j=1). With this scaling, for a set of m observed variables (i.e., j=1…m) in a sample of N individuals (i.e., i=1…N), model (1) gives rise to the correlation structure:

$$\mathbf{R} = \mathbf{\Lambda\Lambda'} + \mathbf{\Psi} \qquad (3)$$

where **R** is the correlation matrix between the *m* observed variables, **Λ** is the vector of *m* regression weights (also referred to in the context of factor analysis as *factor loadings*), and **Ψ** is a diagonal matrix that contains in the diagonal the *m* error variances associated with the observed variables.

In statistical applications of FA, in which model-data fit is assessed by using inferential procedures and standard errors of the estimates are obtained, it is generally assumed that the conditional distribution of the observed variables for a fixed value θ_i is normal. A stronger additional assumption is that the distribution of factor θ is also normal. If both these sets of assumptions hold, the joint distribution the joint distribution of the observed variables is normal. As in any statistical model, these assumptions are unattainable ideal conditions. Experience, however, suggests that in this type of applications the model works acceptably well when the observed variables are approximately continuous, with unimodal and symmetric distributions (e.g. Ferrando, 1999, Hofstee *et al.*, 1998).

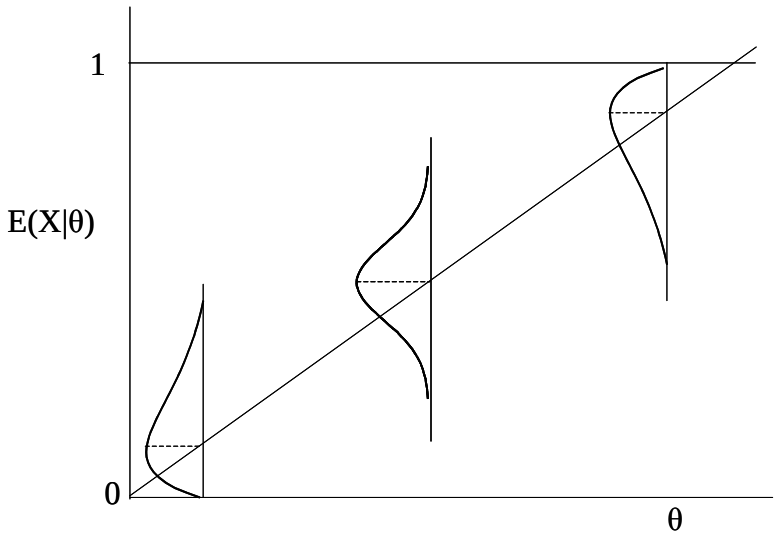
To close this section, we shall summarize the main assumptions of FA as a basic regression model. These assumptions are the main sources of *difficulties* when the dependent (i.e., the observed variables) are item scores. The assumptions are:

- The regression of observed variable X_j on factor θ is linear.
- The conditional variance of observed variable X_j is the same for any value of factor θ (homoscedasticity).
- The conditional distribution of observed variable X_j for any value of factor θ is normal.

2. Problems in the Factor Analysis model when the observed variables are test items

In the type of FA applications that we shall consider in this document, the observed variables are item responses based on a binary or a graded format. These variables cannot be considered to be continuous-unbounded not even approximately. So, strictly speaking, FA is inappropriate. The inappropriateness can clearly be seen if we consider the simple case of binary responses that are fitted with the unidimensional model. For the sake of simplicity we shall consider here raw item scores with assigned 0 and 1 values (e.g. fail and pass).

Consider again equation (2). Because factor θ is continuous and unlimited the expected item score X_j is also unlimited according to the model. However, if the response is binary, its expected value is necessarily bounded between 0 and 1. So, the item-factor regression cannot be linear, as the model assumes. Second, and for the same reasons (θ is unlimited but X_j is not), the conditional distribution of the residuals can be neither normal nor homoscedastic. At the upper end of the trait value, the residuals can only be negative (i.e. there is a ceiling effect). So, when going to the upper end the conditional distribution becomes negatively skewed with reduced variance. Conversely, when going to the lower end, the conditional distribution becomes positively skewed with reduced variances (i.e. there is a floor effect). Clearly, the conditional variance decreases when going to the extremes. So, none of the three basic assumptions of the FA model – linearity, homoscedasticity and conditional normality – is met if the dependent variables are discrete and bounded item responses. The accompanying figure shows these results in more detail.



Regression cannot be linear

$$P(X_{ij}=1|\theta_i)=\mu_j+\lambda_j\theta_i;\quad 0\le P\le 1,\quad -\infty<\theta_i<\infty$$

Regression cannot be homoscedastic:

$$Var\left(X_{ij}=1|\theta_i\right)=P\left(X_{ij}=1|\theta_i\right)\left(1-P\left(X_{ij}=1|\theta_i\right)\right)$$

Figure 1. Problems of linear FA with discrete and bounded item responses

Given the results just discussed, it seems of interest to conjecture what a *plausible* item-factor regression for binary responses should be like. Lord (1953) discussed some conditions which are shown in figure 2 together with some theoretical curves that fulfil them.

1. For low enough trait levels, the expected score should be close to 0.
2. For high enough trait levels, the expected score should be close to 1.
3. The expected score increases with trait level.
4. The curve is smooth and has a single inflexion point.

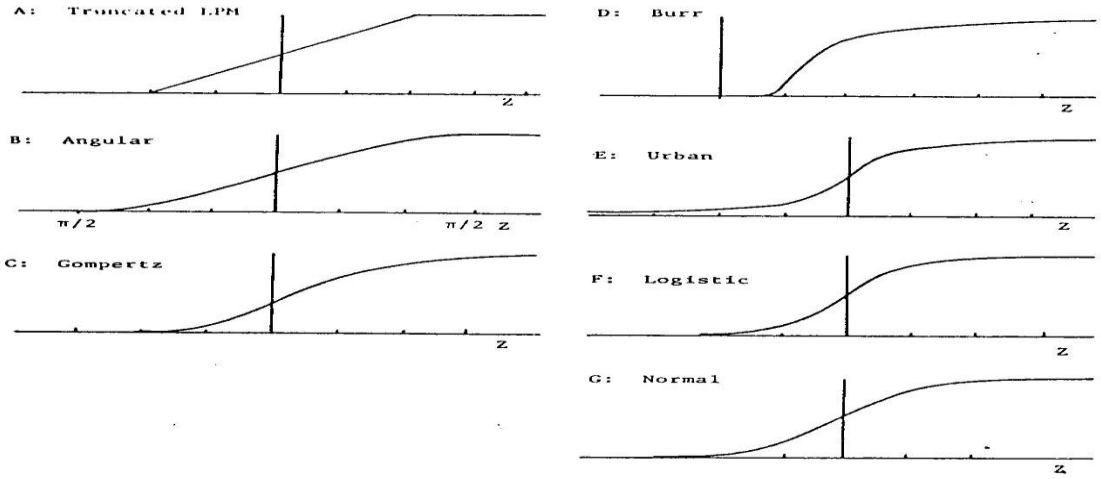


Figure 2. A plausible item-factor regression for binary responses (Lord, 1953)

To summarize, when the observed variables are bounded and discrete, their regression on factor θ is non-linear (possibly S-shaped according to Lord's conjectures) and heteroscedastic. So, the standard FA model is not a correct model, and (at best) it should be taken as an approximation. A key point is, then, to assess in which conditions the simple linear model can be a reasonably good approximation, mainly as far as purposes of item analysis is concerned.

3. The unidimensional (congeneric) model

3.1. The underlying-variables approach: item factor analysis and the two-parameter Item Response Theory model

Because the assumptions of the linear model are not tenable for discrete item scores, a reasonable alternative is to propose a more plausible model. In psychometrics, this proposal was initially made for binary items in which the inadequacies of the FA model were most evident. The proposal, which we shall call the *Underlying-Variables Approach* (UVA), gives rise to a curve that fulfils Lord's conditions.

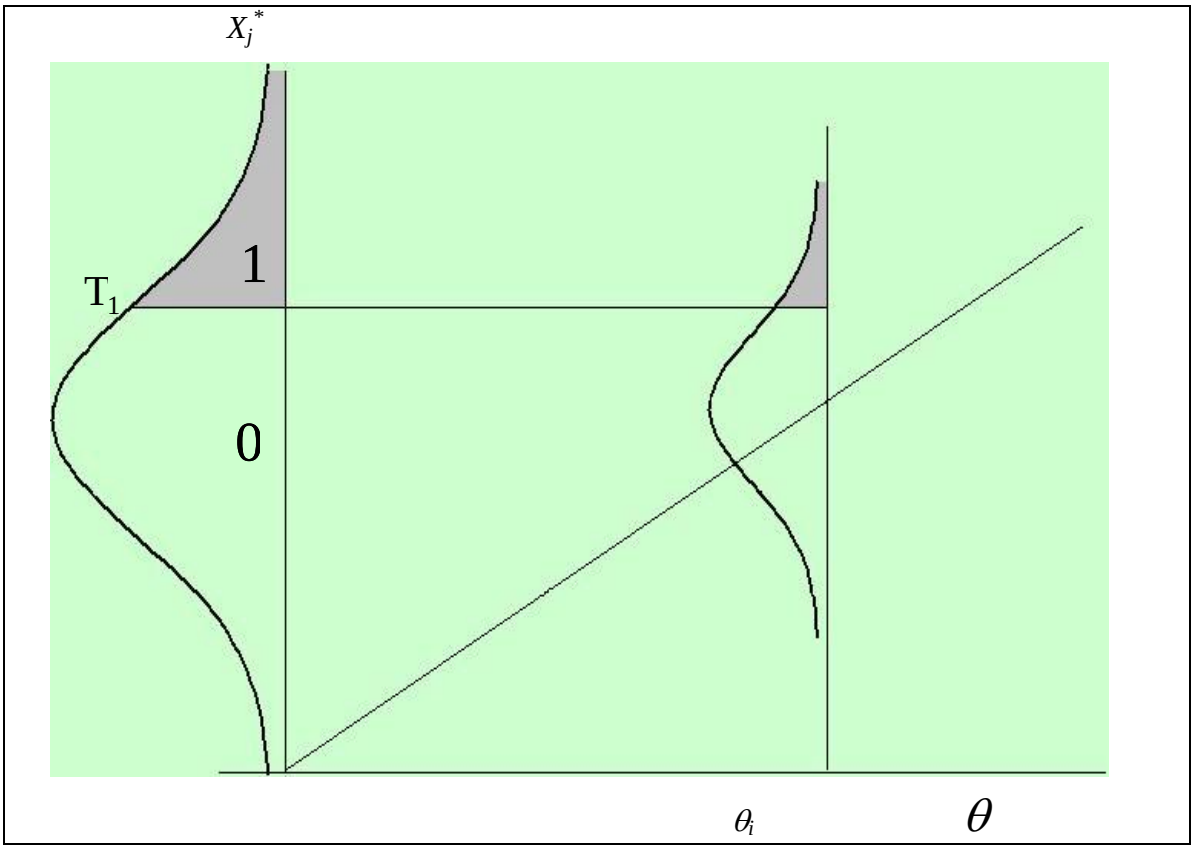


Figure 3.UVA model for binary responses

In the UVA model, each observed score X_{ij} (0 and 1) is considered to arise from the dichotomization of an underlying response variable X^*_j which is both continuous and unbounded and which is usually scaled as a standard variable (mean zero and unit variance). These assumptions (together with the one mentioned below) establish the first-level model. Then, the hypothetical latent response variables, which were assumed in the first-level model, are assumed to behave according to the linear FA model in equations 1 to 3 (so, strictly speaking, the FA is the second level model). As we shall see, it is important that this two-level structure is taken into account .

In the first-level model, the mechanism that relates the observed binary item score to the underlying response variable is a step process. It is assumed that the observed scores *0*

and *1* arise as a result of an arbitrary dichotomization of the corresponding underlying response variable at a given threshold τ_{j0}, such that:

$$X_{ij} = 0 \quad if \quad X^{*}_{ij} < \tau_{j0}$$
$$X_{ij} = 1 \quad if \quad \tau_{j0} \leq X^{*}_{ij} \tag{4}$$

The first-level model described so far is the *tetrachoric model*. The second level model is the congeneric FA model which is assumed to hold for X^{*}_{j},

$$X^{*}_{ij} = \alpha_j \theta_i + \omega_{ij} \tag{5}$$

where X^{*}_{ij} is the underlying response variable, and ω_{ij} is its associated model error.
Now, if the FA assumption of normality in the conditional distributions discussed above is used, it follows that the conditional probabilities of scoring *0* or *1* in item *j* are:

$$P(X_{ij} = 0 \mid \theta_i) = \Phi\left(\frac{\tau_{j0} - \lambda_j \theta_i}{\sigma_j}\right)$$
$$P(X_{ij} = 1 \mid \theta) = \Phi\left(\frac{\lambda_j \theta_i - \tau_{j0}}{\sigma_j}\right) = 1 - P(X_{ij} = 0 \mid \theta_i) \tag{6}$$

It then follows that the regression of the observed raw score on θ is the conditional probability of scoring 1. If we now make the familiar transformations,

$$a_j = \frac{\alpha_j}{\sigma_{\omega j}}; \quad b_j = \frac{\tau_{j0}}{\alpha_j}, \tag{7}$$

where $\sigma_{\omega j}$ is the error variance associated with the underlying response variable X_j, and a_j and b_j are the discrimination and location/difficulty indices, respectively. Now, the regression becomes the basic item response theory *two-parameter normal ogive item characteristic curve* (ICC, see e.g. Lord, 1980). This curve can be approximated by a simpler logistic ogive by using the constant D=1.702 in which case the two curves become virtually indistinguishable.

$$P(X_{ij} = 1 \mid \theta) = \Phi(a_j(\theta_i - b_j)) \cong \frac{\exp(Da_j(\theta_i - b_j))}{1 + \exp(Da_j(\theta_i - b_j))}. \tag{8}$$

From the UVA-FA approach, the item discrimination a_j in (7) is a signal-to-noise ratio: the numerator is the factor weight (signal) and the denominator is the residual standard

deviation (noise). High values of a_j indicate that the scores of this item cleanly and sensitively reflect the trait levels, so the item is a good indicator or measure of the trait. On the other hand, low values indicate that the scores are affected by determinants other than the trait level (i.e., the item is a noisy or poor indicator of the trait). The range of a_j values in many tests is about 0.2 to 3. As for the item location/difficulty b_j, we interpret it in a threshold sense: it is the trait level at which the probability of giving a correct response (or endorsing the item) is .5. In typical-response measurement, b_j is the trait level which marks the transition from the tendency to respond *no* to the tendency to respond *yes*.

The binary item-trait regression is illustrated below for different values of the discrimination parameter. Clearly, the normal- ogive curve meets Lord's requirements. Note the role of the discriminating parameter in determining the sensitivity of the item to detect variations in the trait level.

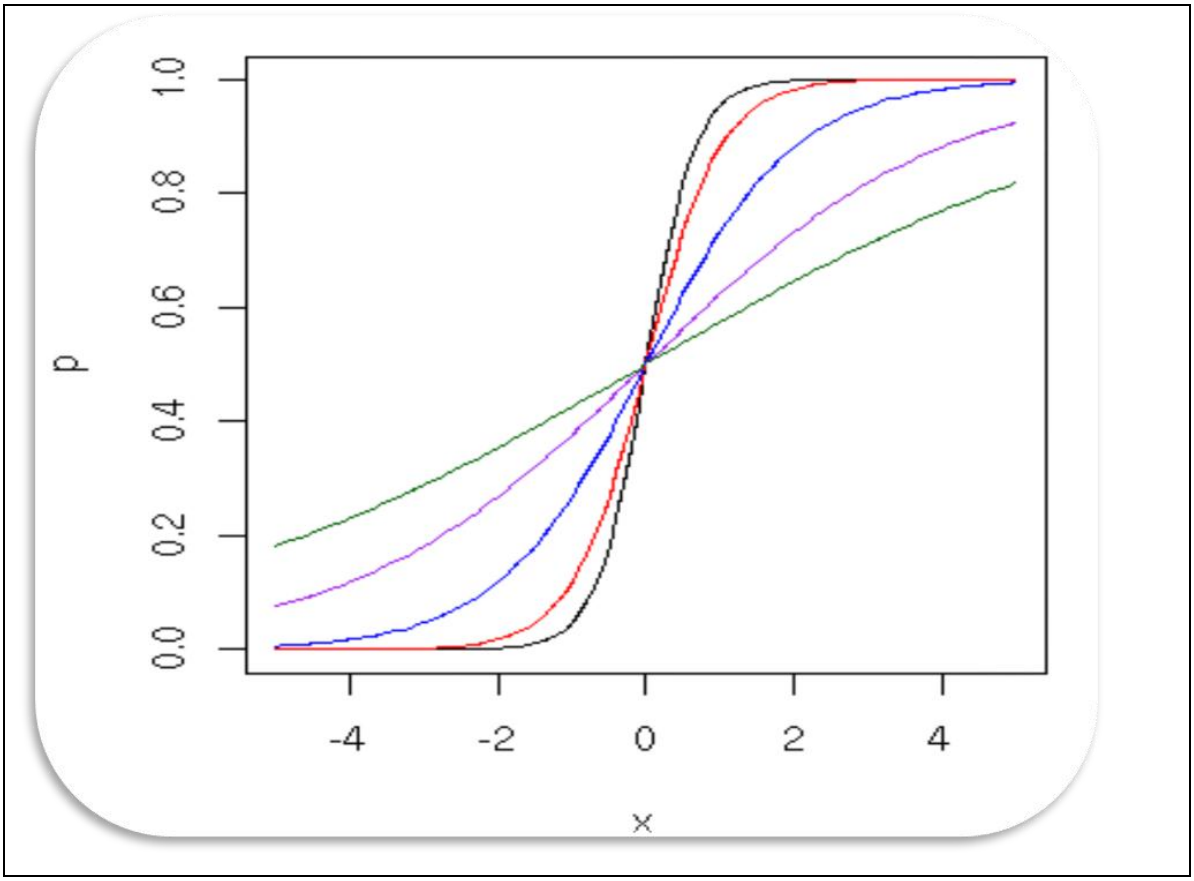


Figure 4. Item-factor regression related to a binary response

3.2. The underlying-variables approach: Item Factor analysis and the graded response model

The specific model we shall consider in this section is Samejima's (1969) normal ogive version of the graded response model (GRM). Of the general GRM family, the normal ogive version (or its logistic counterpart) has a series of desirable features and is the most commonly used in practical applications (Baker, 1992, Samejima, 1969, 1997).

The UVA uses the same two-level approach as in the binary case. In the first level we assume that the observed item response arises as a result of a categorization of an underlying response variable. In the second level we assume that the congeneric model holds for these underlying responses. The main difference with the binary model is that a graded response item with *k* categories is now characterized by *k*–1 thresholds. So, instead of the first-level tetrachoric model we now have the first-level polychoric model.

For example, Figure 5 shows a graphical representation of the corresponding categorization process in a 4-point Likert item.

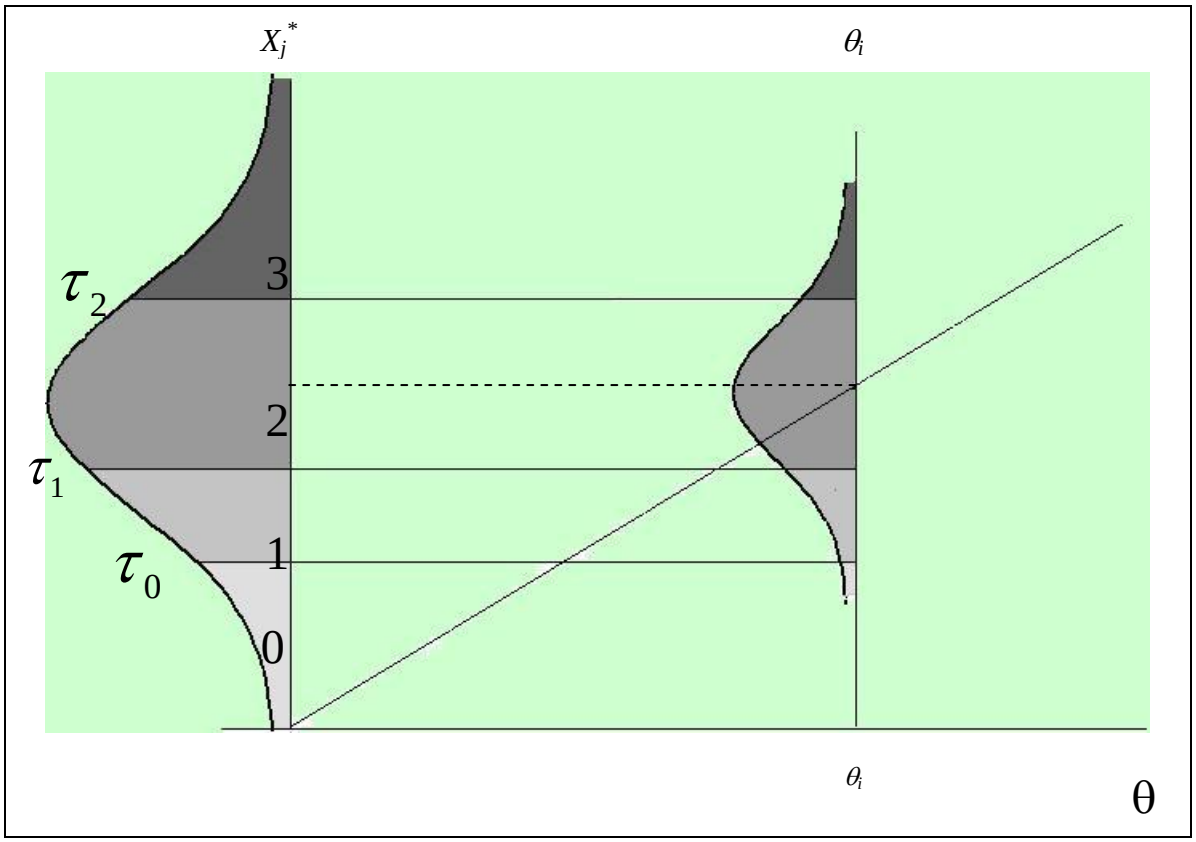


Figure 5.UVA model for graded responses

For a typical 5-point Likert item, the categorization process is:

$$\begin{aligned} X_{ij} &= 0 \quad if \quad X_{ij}^{*} < \tau_{j1} \\ X_{ij} &= 1 \quad if \quad \tau_{j1} \le X_{ij}^{*} < \tau_{j2} \\ X_{ij} &= 2 \quad if \quad \tau_{j2} \le X_{ij}^{*} < \tau_{j3} \\ X_{ij} &= 3 \quad if \quad \tau_{j3} \le X_{ij}^{*} < \tau_{j4} \\ X_{ij} &= 4 \quad if \quad \tau_{j4} \le X_{ij}^{*} \end{aligned} \tag{9}$$

The FA model in the second level is the same as in the binary case:

$$X_{ij}^{*} = \alpha_j \theta_i + \omega_{ij}, \tag{10}$$

and the parameter transformation is a direct extension of the binary model:

$$a_j = \frac{\alpha_j}{\sigma_{\omega j}}; \quad b_{jk} = \frac{\tau_{jk}}{\alpha_j}. \tag{11}$$

Note that each item is characterized by a single discrimination parameter and k–1 location parameters. The discrimination has the same interpretation as in the binary case. However, the locations do not. The location b_{jk} is the point on the θ continuum at which the probability of scoring in category k or higher is 0.5. This is not a simple interpretation, and we believe it is better to view these locations as boundaries that mark the transitions between the successive categories. The advantage of using the IRT locations (instead of the FA thresholds) is that the IRT locations are on the same scale as θ, so they indicate how the item is located on the trait continuum. To obtain information about this issue we must assess: (a) the distances between locations and their spread, and (b) its central location.

Finally, the probability of obtaining a score X_k for a given trait level is now

$$P(X_j = k \mid \theta_i) = \Phi(a_j(\theta_i - b_{jk})) - \Phi(a_j(\theta_i - b_{jk+1})) \tag{12}$$

where Φ is the standard normal c.d.f. which can be approximated by the virtually indistinguishable logistic function. Equation (12) is the usual IRT expression of the GRM (Samejima, 1969 Baker, 1992).

The binary ICC can be generalized in different ways to the graded response case. Here we shall make a generalization based on the regression of the raw item score (k=0,1,2..) on θ (see for example, Chang & Mazzeo, 1994). In the GRM this regression is:

$$E(X \mid \theta_i) = \sum_{k=0}^{nk-1} kP(X_j = k \mid \theta_i) \qquad (13)$$

where X is the raw score, and nk is the number of response categories. The regression (13) is illustrated below in figure 6. As in the binary case, it is nonlinear (S-shaped) and heteroscedastic: the conditional distributions become skewed at the ends of the scale with reduced variance. This agrees well with the plausibility conditions that Lord established in the binary case.

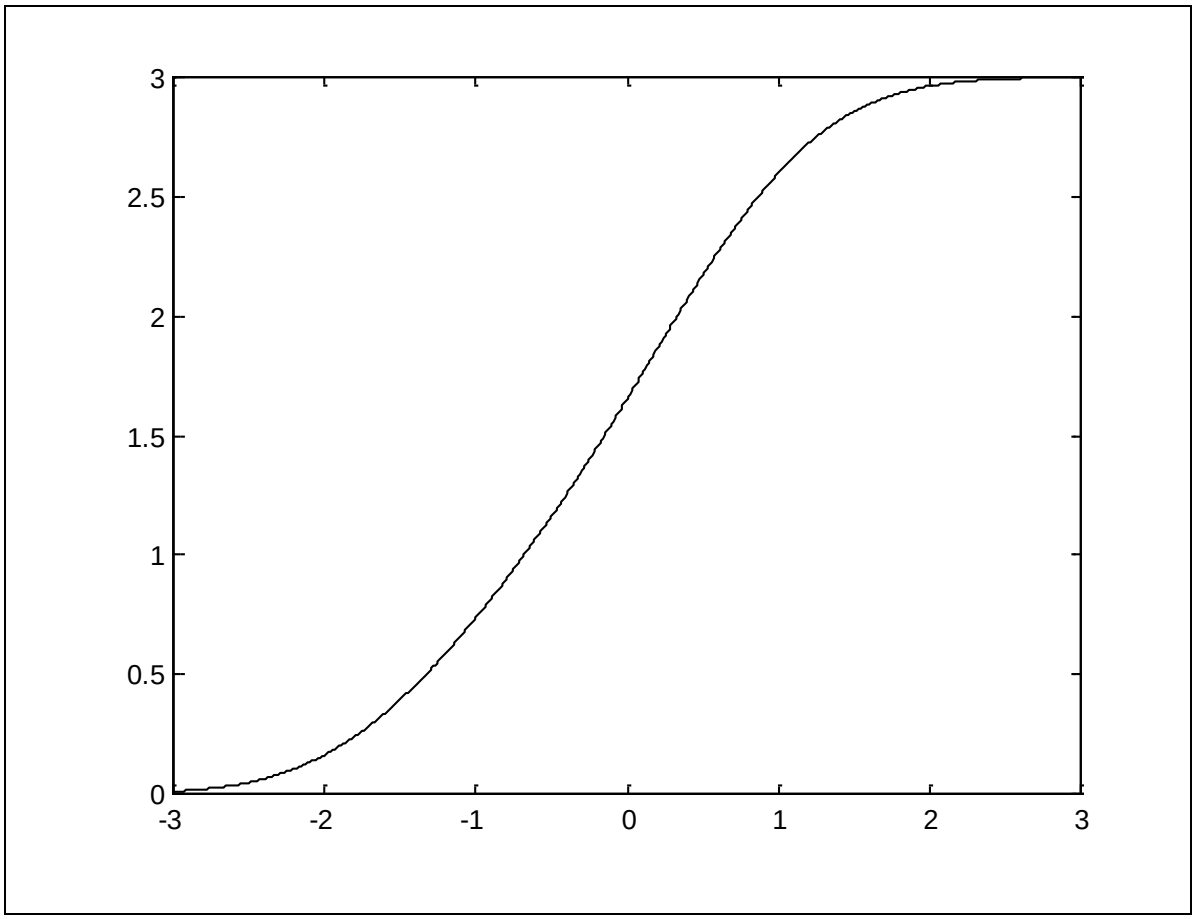


Figure 6. Item-factor regression related to a graded response

3.3. The direct approach: basic results and some comparisons

In the direct approach we treat the discrete and bounded item scores as if they were continuous-unlimited variables. So, the linear homoscedastic model in equations (1) and (2) is fitted directly to the observed responses (and no underlying response variables are

modelled at the first level). Using the correlational FA approach we mostly discuss in this report, the differences between the UVA and the direct approach can be summarized as follows. In the UVA approach, the correlational FA model in (3) is fitted to the tetrachoric or polychoric correlation matrix, which is the matrix derived from the first-level model. In contrast, in the direct approach no first-level model is considered, and the FA model (3) is directly fitted to the product-moment inter-item correlation matrix.

As discussed above, direct modelling is, at best, as an approximation. Therefore, the conditions in which the linear approximation can be expected to be reasonably correct should be assessed. This point has been addressed in theoretical research (Lord, 1952, 1953), empirical research by using simulation (Olsson, 1979, Muthén & Kaplan, 1985), and applied research (Ferrando, 1999, Hofstee et al. 1998), and the results are consistent. The linear model tends to work rather well when (a) the discriminating power of the items is moderate or low, and (b) the items have no extreme locations. In the graded-response case, condition (b) is obtained when the thresholds are centred about the population mean of θ and the distances between thresholds are similar. If this is so, the marginal distributions of the items are unimodal and approximately symmetrical (see Muthén & Kaplan, 1985). Overall, if conditions (a) and (b) are met, it follows that the item-trait regressions are essentially linear and homoscedastic for the range of θ that contains most of the respondents, and this is the reason why the linear model is expected to be a good approximation. Conversely, the UVA is expected to outperform the direct approach when the items are both extreme and highly discriminating. Extreme items, especially with skewed distributions that have opposite signs (e.g. very easy and very difficult items), lead to differential attenuations of the product-moment correlations with respect to the polychoric correlations, and this result leads to biased loading estimates that reflect more the extremeness of the item distribution than the strength of the item-factor relation. If, in addition, the items are highly discriminating, the item-factor relations become markedly nonlinear. Overall, in the case of extreme and highly discriminating items additional *curvature factors* are needed to explain the nonlinear item-factor relations (see McDonald & Alhawat, 1974; in the traditional FA literature these factors were known as *difficulty factors*).

To sum up, if we factor-analyze an essentially unidimensional item set in which the items are strongly skewed in opposite directions and highly discriminating, two types of distortions are expected to arise: (a) spurious evidence of multidimensionality, due to the

need of additional *curvature factors* that have no substantive interpretation; and (b) differential attenuation of the loadings on the *content factor*.

A second discussing is whether using the *wrong* linear model is of any interest even if it provides a reasonable good approximation to the data. In other words, why not always use the (a priori) more correct UVA-based IRT model. Two points must be considered here. First, the ogive curves implied by the UVA, although more plausible than the straight line, might still differ from the 'true' item traces (as estimated, for example, using nonparametric smoothed regression). This point might explain some empirical results in which the *wrong* linear model fitted the smoothed ICC better than the theoretically superior UVA model (Ferrando, 2002, 2004). As for the second point, assume that the linear and non-linear models provide similar results in terms of model–data fit. Then the linear model has some advantages derived from its simplicity: (a) it is more easily interpretable, and (b) it is more likely to lead to more stable solutions. As far as this issue is concerned we note again a basic result: in the linear approach there is only one level of analysis. So the basic product-moment matrix is expected to be more stable than the tetrachoric/polychoric matrix, and this greater stability in turn is expected to lead to more stable item parameter estimates.

In the UVA case, the first level model (i.e. the tetrachoric/polychoric model) is potentially problematic. First, it makes strong assumptions (normal response variables underlying the observed scores) that might be incorrect. Second, the tetrachoric/polychoric correlations are far less reliable and stable than the product-moment correlations obtained from continuous data and, the more extreme the variables and the smaller the sample, the more unstable they are (McNemar, 1969). In the binary case Guilford and Fruchter (1973), and McNemar (1969) noted that, at the very least, twice the sample size must be used for the tetrachoric estimate to be as accurate as the corresponding product-moment, and they advise using samples of at least 200-300 observations. To sum up, we note that if the first level model is incorrect, the second level model (i.e. the item FA in the strict sense) cannot be correct. Furthermore, if the tetrachoric/polychoric estimates at the first level are unreliable, the factor loading estimates obtained from the first level estimates would be even less reliable.

To summarise the material in this section, in figures 7 and 8 we provide some guidelines/suggestions regarding the use of the linear or the UVA model.

1. Linear FA is likely to be a good approximation if:
   - Items are distributed symmetrically.
   - Items are not highly discriminative.
2. These points can be checked by conventional item analysis, using the standard difficulty and discrimination indices.
3. If many items are extreme and highly discriminating, it is more convenient to use non-linear FA (if data permits).

Figure 7. Linear vs. non-linear (UVA) Factor Analysis

1. Check whether your sample is large enough to obtain accurate tetrachoric/polychoric estimates.
2. Assess the distribution of the item scores, particularly skewness, and the item-total correlations (i.e. the classical item discriminations).
3. If possible, fit both the linear and the UVA models and compare the results in terms of:
   - Goodness of model-data fit
   - Estimated factor loadings
4. Decide which approach is more appropriate for your data.

Figure 8. A suggested approach

3.4. Parameter estimation and assessment of model-data fit

In the direct approach based on the linear model (equations 1 to 3), the parameters of most interest are the factor loadings λ_j and residual variances σ^2_j which, as discussed above, are estimated from the sample inter-item product-moment correlation matrix according to the structure in equation (3).

In the nonlinear UVA model, parameters are estimated from the bivariate tetrachoric/polychoric tables between pairs of item scores. In the simplest and most usual approach that we shall consider here (see e.g. Mislevy, 1986), the item thresholds (τ_{jk}) are

estimated from the marginals of the table, and the polychoric correlations are estimated from the joint frequency cells. Next, the usual FA of the polychoric correlation matrix according to the structure (3) provides the estimates of the loadings (α_j) and of the residual variances ($\sigma^2_{\omega j}$). Once the estimates have been obtained, they can be re-parameterized using equations (7) or (11) to put the model in the most usual IRT form. This approach is called the heuristic solution (Bock & Aitkin, 1981).

For both approaches, we shall now discuss two estimation procedures that are implemented in FACTOR to fit the factor model in equation (3): Maximum Likelihood (ML) and Unweighted least squares (ULS).

For item scores that can be treated as (approximately) continuous and which are fitted with the direct approach, ML estimation of the item parameters is efficient and can be regarded as theoretically optimal. If the ML procedure converges properly, and the joint distribution of the variables is (approximately) normal, then the measures of fit based on the chi-squared exact-fit test and the indices of fit derived from it are correctly interpretable. However, some points must be taken into account when interpreting ML estimates and model-data fit results. For these results to be approximately correct, two basic conditions are needed. First, the item scores must have a sizable number of points and symmetric, normal-like distributions. Second, the model must be a good approximation so that the misspecification error is not larger than the random sampling error. Most real data does not fulfil these conditions.

Under the UVA model, ML estimation can be used in some cases but it is important to know that some results will not be correct (see Mislevy, 1986). The tetrachoric/polychoric correlations are pairwise ML estimates of the population correlations. However, the polychoric matrix is not a 'true' product-moment correlation matrix. Furthermore, the pairwise estimation in many cases produces a matrix that is not positive definite and, when it does, ML estimation is not feasible. If estimation is feasible and converges properly, the parameter estimates are expected to be unbiased (provided that the degree of misspecification is small). However, the standard errors are unreliable and the chi-square and derived indices are expected to be upwardly biased.

Pros	Cons
• Optimal in the statistical sense (efficient) • Provides standard errors and inferential procedures for assessing model data fit	• Requires continuous-unbounded variables • Inferential interpretation is based on the normality assumption • The degree of model misspecification must be small • The procedure can be unstable

Figure 9. Pros and contras of ML estimation

The simpler ULS estimation procedure is not theoretically optimal (as is ML) but has important advantages. First, it does not require any distributional assumptions. Second, it is quite robust: usually the ULS solution converges when the ML solution does not. Third, in complex solutions, which are not exact but only approximately correct (and models are never correct with real data), ULS tends to provide less biased estimates of the *true* parameter values (e.g. Briggs & MacCallum, 2003). Given these advantages, ULS seems to be an appropriate choice especially for the case of large models and not too large samples (and this choice refers to both the direct approach and the UVA). Furthermore, as we discussed above, the statistical assumptions which ML makes are not correct in UVA analysis. For this reason, ULS is our recommended choice for the UVA analysis of tetrachoric and polychoric matrices in FACTOR. When used with these matrices it tends to provide accurate estimates even when models are large (Lee, Zhang & Edwards, 2012). In fact, some simulation studies suggest that ULS provides better estimates than far more complex and theoretically superior procedures (Knol & Berger 1991, Parry & McArdle, 1991).

The shortcomings of ULS are that (a) it uses limited information and (b) provides neither standard errors for the loadings nor a rigorous test of exact and/or approximate fit based on the chi-square distribution. As for the second point, however, approximate closed-form standard errors can be obtained for both the direct approach and the UVA (Lee et al., 2012). As for the test of fit, if the variables are treated as continuous, then a ULS-based chi-square test and derived indices can be obtained by using normality assumptions (see Harman, 1976).

Pros	Cons
• Simple and robust • Does not make distributional assumptions • ULS estimates tend to outperform ML estimates when the model is misspecified (e.g. minor factors)	• Statistically non-optimal • Provides neither standard errors nor inferential procedures for assessing model data fit

Figure 10. Pros and contras of ULS estimation

We turn now to model-data fit assessment, and we shall discuss two general measures of fit that are implemented in FACTOR. These measures are perfectly valid for any estimation procedure, including indeed ULS estimation. The first is the gamma-GFI index initially proposed by Tanaka and Huba (1985) and latter implemented in the LISREL programs. The gamma-GFI measures the relative amount of variance-covariance (correlation in our case) in the empirical correlation matrix that is accounted for by the fitted model. It can be regarded as an absolute index of fit with an interpretation similar to that of the multiple R^2 in regression analysis.

$$\gamma - GFI = \frac{\hat{\mathbf{r}}'\hat{\mathbf{r}}}{\mathbf{r}'\mathbf{r}} \tag{14}$$

where $\hat{\mathbf{r}}$ and $\mathbf{r}$ are two vectors that contain the reproduced and the observed correlations, respectively.

The second index is the root mean square of the standardized residuals (RMSR-z), an absolute descriptive index that measures the average size of the residual correlations once the prescribed model has been fitted.

$$RMSR = \sqrt{\frac{\sum_{i=1}^{m}\sum_{j=1}^{i}\left(\hat{r}_{ij} - r_{ij}\right)^2}{m(m+1)/2}} \tag{15}$$

where $\hat{r}_{ij}$ are the reproduced correlations, and r_{ij} are the observed correlations. The usual ad-hoc RMSR-z reference value for considering model-data fit to be acceptable is 0.05. In FACTOR, however, we propose to use instead a reference value that has a stronger rationale and which was first proposed by Kelley and Thurstone in 1935 (see Harman, 1976). This value is $1/\sqrt{N}$, which is (approximately) the standard error of a zero

correlation for a sample size of *N*. So, for a sample of *N*=300, the standard error of a zero correlation is 0.06. Now if the RMSR-z is about 0.06 or less, we can consider that, on average, the residual correlations once the model has been fitted do not significantly depart from zero, so the prescribed model is acceptably correct. Indeed this is a very crude criterion for deciding whether the model is appropriate and does not take into account such other determinants as the size of the model. Experience, however, suggests that in practical applications Kelley's criterion works as well (or even better) as more rigorous and exact criteria.

In a practical situation, independently of the factorial model that is chosen (direct approach or UVA) and the method of estimation (ML or ULS), we believe that it is very important to inspect the distribution of the residuals (preferably the standardized residuals when assessing model-data fit). As McDonald and Ho (2002) noted, a given degree of misfit can arise because (a) there are a limited number of misspecifications that give rise to a few large discrepancies; or (b) there is a general scatter of discrepancies not associated with any particular misspecification. This information, which is provided by FACTOR, is important for possible model modifications and is not provided by the overall scalar-valued indices. In a well-fitting model, the distribution of the standardized residuals is expected to be symmetrical, bell-shaped and centred about zero, with no definite clusters that would suggest large misspecifications.

4) The multidimensional model

In the rest of the report, we shall consider the situation in which the item responses are (partly) determined by more that one trait or common factor. Because the resulting model is the same for any number of common factors equal to or greater than two, for illustrative purposes we shall mostly discuss the simplest case of two factors.

Most of the results discussed in the sections above apply to both the unidimensional and the multidimensional model. More specifically, the results concerned with the following topics are general results that do not depend on whether the analysis is unidimensional or multidimensional: (a) the problems that arise when discrete and bounded item responses are analysed; (b) the distinction between the direct linear approach and the nonlinear UVA and the conditions in which one or another are more appropriate; and (c) the discussion about estimation procedures and estimation or model-data fit. For this reason, in the rest of the report we shall only discuss some extensions to the parameterization and

interpretation of the models and several multidimensional solutions that are useful in item analysis.

4.1 The direct approach

The linear model with multiple common factors, also known as Thurstone's model, is a direct extension of Spearman's congeneric model in equation (1). For example, a bidimensional model (i.e., two factors are estimated from the mean vector and the variance/covariance matrix), is expressed as

$$X_{ij} = \mu_j + \lambda_{j1}\theta_{i1} + \lambda_{j2}\theta_{i2} + \varepsilon_{ij} \tag{16}$$

where λ_{j1} and λ_{j2} are the regression weights of observed variable X_j on each latent variable (also known as factor loadings), and θ_{i1} and θ_{i2} are the non-observable level of individual i on each latent variable. Note that the assumptions of linearity, homoscedasticity and normality in the conditional distributions that were made in the unidimensional model still hold. The scaling that we shall consider from now on is also the same in both cases: both the dependent variables and the common factors are scaled as standard variables.

The key point in the multidimensional extension (16) is whether the common factors are related or not. If the factors are uncorrelated (orthogonal solution), the standardized factor loadings are still interpreted as variable-factor correlations. When the factors are correlated, the solution is oblique. In this case the loadings are no longer item-factor correlations but rather standardized regression weights (i.e. Beta weights). Conceptually each loading now measures the impact of the corresponding factor on the item score when the remaining factors are still constant. The orthogonal vs. oblique distinction is clearly seen in the correlation structures derived from equation (16). They are:

$$\begin{aligned} &\textit{Oblique}: \mathbf{R} = \mathbf{\Lambda\Phi\Lambda'} + \mathbf{\Psi}; \qquad \mathbf{\Psi}\ \textit{diagonal} \\ &\textit{Orthogonal}: \mathbf{R} = \mathbf{\Lambda\Lambda'} + \mathbf{\Psi}; \qquad \mathbf{\Psi}\ \textit{diagonal} \end{aligned} \tag{17}$$

where $\mathbf{\Phi}$ is the inter-factor correlation matrix, and $\mathbf{\Lambda}$ is the factor pattern. In the unidimensional case, $\mathbf{\Lambda}$ is a column vector of dimension items × 1. In the multidimensional case it is a matrix of dimension items × factors. Figure 11 provides more details regarding the correlational structure derived from the multidimensional model.

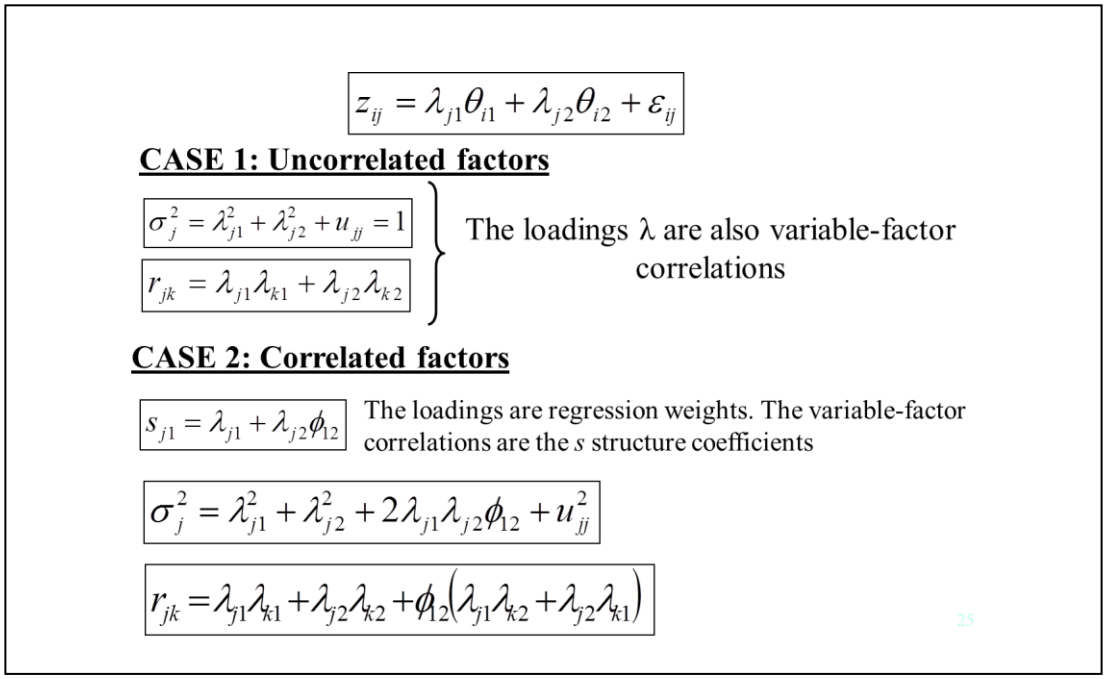


Figure 11. Multidimensional factor analysis model

4.2 The UVA approach: the multidimensional two-parameter model

As in the unidimensional case, the multidimensional two-parameter model (M2PM) allows an FA parameterization and an IRT parameterization. Under the FA parameterization and for *p* common factors, each binary item is characterized by a single threshold and *p* loadings or weights. The threshold has the same interpretation as in the unidimensional case. However, as in the direct approach, the loadings can only be interpreted as correlations in the case of uncorrelated factors. In the correlated-factors case, the loadings are interpreted standardized regression (i.e. Beta) weights. We note that, under the UVA approach, the correlations or the Beta weights refer to relations between the factors and the hypothetical latent response variables that underlie the observed item responses.

The most usual IRT parameterization of the M2PM is:

$$a_{jl} = \frac{\alpha_{jl}}{\sigma_{\omega j}}; \quad d_j = \frac{\tau_{j0}}{\sigma_{\omega j}}, \tag{18}$$

where $\sigma_{\omega j}$ is the error variance associated to the underlying response variable X_j, and a_{jl} and d_j are the discrimination and location indices, respectively. So, each item is characterized by a single location d_j and p discriminations a_{jl}. The d_j location is usually termed as the intercept, and is related to the difficulty of item X_j although it is not strictly

a difficulty parameter and has a complex interpretation. The discrimination a_{jl} has a similar interpretation as in the unidimensional case: it is related to the slope of the item response surface in the direction of the l axis. More conceptually, it indicates the discriminating power of the item in the direction of the l factor (see Reckase, 2009).

Reckase (2009) has proposed two additional item parameters – multidimensional difficulty and multidimensional discrimination – that attempt to clarify the IRT interpretation of the item estimates. They are defined as

$$MDIF_j = -\frac{d_j}{\sqrt{\sum_l^m a_{jl}^2}}; MDISC_j = \sum_l^r a_{jl}^2 \ . \tag{19}$$

The multidimensional difficulty (*MDIF*) of item X_j is the distance in the space of the p factors from the origin to the point of the steepest slope, and the direction is taken as the direction of greatest slope. It has essentially the same interpretation as the difficulty parameter in the unidimensional case (i.e. the same units: *easy* items have large negative values and *difficult* items have large positive values).

The multidimensional discrimination (*MDISC*) of item X_j reflects the overall discriminating power of the item for the best combination of the factors, where *best* is understood as *the combination which provides the maximum discrimination*. Except for items that are factorially pure, the overall discrimination will be larger than any of the single discriminations a_{jl}.

4.3 The UVA approach: the multidimensional GRM

The IRT parameterization of this model has not been developed to the same extent as in the binary model (see Reckase, 2009). So, we shall use here only the FA (threshold/loading) formulation already discussed. The thresholds are the same as in the unidimensional case (see equation 9). The interpretation of the loadings is the same as that given at the beginning of section 4.2.

4.4. Some useful multidimensional semi-confirmatory solutions

4.4.1. The canonical solution and applications

The canonical solution (Harman, 1976) is a mathematically well defined multidimensional FA solution in which the factors are orthogonal, and each successive factor accounts for as much common variance as possible. More specifically, in each successive factor the quantity that is maximized is the sum of the squared weights on this factor. The unrotated principal-axes solution, for example, is a canonical solution. In FACTOR, all the unrotated solutions are in the canonical form. So, a canonical solution in *p* factors is obtained by specifying *rotation: none* in FACTOR.

The canonical solution is generally considered to be arbitrary and uninterpretable. However, we do not agree with this view. The canonical solution can be very useful for constructing a unidimensional test. The starting point for this to be so is a strong core of unidimensional items which serves as basis for making the process of item selection feasible (Lumsden, 1961). Suppose that a two-factor canonical solution is fitted to a complete set of items which are intended to measure a single solution. Furthermore, assume that the items are scored in the same direction of the trait intended to be measured.

If the item set (a) behaves as essentially unidimensional, and (b) reliably measures the trait, then the first canonical factor will exhibit a positive manifold pattern (all the loadings on this first factor will have the same sign) in which the magnitude of the loadings will be substantial for all of the items. Furthermore, all the loadings on the second canonical factor will be close to zero. These expected conditions allow us to discard poorly functioning items: those items that have non-substantial loadings on the first factor (i.e. they poorly measure the intended trait) and/or those items that have substantial loadings on the second factor (i.e. they are impacted by determinants other than the ones the trait intended to measure). In particular, item doublets or triplets or extreme items (if a direct-approach solution is fitted) are usually very well identified by their substantial loadings on the second canonical factor.

- It is an orthogonal solution in which each successive factor accounts for as much variance as possible
- It is equivalent to an unrotated principal-axes solution
- It is particularly useful for item selection and construction of a unidimensional test.

Figure 12. Definition of a canonical solution

In a canonical solution, after the general first factor has been estimated, the orthogonality constraint implies that all the items must have both positive and negative loadings on the subsequent factors. The resulting bipolarity allows the differences between item groups to be interpreted. In this way, the column patterns of the successive factors allows group factors (i.e. groups of factors that share specificities) to be identified.

4.4.2. The Bi-factor solution

As discussed at the end of the section above, in many item sets that are intended to measure a single trait, there are often sub-sets of items that share specificities (e.g. parcels, common expressions, similar required answers). These specificities might explain response variance beyond the common factor that all of the items intend to measure. Thus, from an FA perspective, the data is multidimensional even when all of the items clearly measure a common dimension.

The bi-factor solution (see e.g. Reise, Morizot & Hays, 2007) is a particular FA specification which is multidimensional but which is able to reflect the essential unidimensionality of the data. This solution prescribes a common, general factor which reflects what is common in all of the items. In addition, a series of orthogonal group factors model the item specificities. Frequently, these additional orthogonal factors are usually interpreted as nuisance dimensions. However, in a clear bi-factor solution in which the loadings on the general factor exhibit a positive manifold, the goal of measuring a single common dimension can still be retained while the unwanted variance

due to the specificities is controlled. Figure 13 describes a hypothetical bi-factor pattern with a general first factor and two group factors: the asterisks indicate loadings that are to be freely estimated, whereas the remaining loadings are expected to be zero.

F1	F2	F3
*	0	*
*	0	*
*	0	*
*	0	*
*	0	*
*	*	0
*	*	0
*	*	0
*	*	0
*	*	0
*	*	0

Figure 13. A hypothetical target matrix for a bi-factor solution (A general factor and two group factors)

The bi-factor solution can be viewed as a more confirmatory evolution of the canonical solution described in the section above. Indeed, in the canonical solution the group factors are identified by the bipolar patterns that emerge in the successive (more exploratory) Factors. In contrast, in the bi-factor solution, the group factors are prescribed in advance (more confirmatory).

The bi-factor solution can be restricted or semi-restricted (Reise, Moore & Haviland, 2010). Here we shall only consider the semi-restricted approach, which is the one available in FACTOR. It consists of two steps: (a) defining a target matrix, and (b) rotating the initial FA solution to the position which minimizes the sum of squared discrepancies between the prescribed target and the rotated solution (i.e., Procrustean rotation). As Browne (1982) explains, the target matrix reflects only partial knowledge of what the pattern should be, so each element in the pattern can be treated as specified (zero loading) or unspecified (*). In the bi-factor case all the loadings on the first factor are unspecified whereas zero loadings are set on the remaining columns to identify the group factors. Figure 13 shows a target matrix that describes the expected bi-factor solution in a set of 11 items: the 11 items are mainly unidimensional (Factor 1); the first 5 items share

a common specificity (Factor 3); and (c) the last 6 items share a common specificity (Factor 2).

4.4.3. The independent-cluster basis

Most prescribed solutions in confirmatory factor analysis are, in fact, *independent-cluster* (I-C) solutions (see McDonald, 1999), defined by the property that each variable has a large loading on just one factor and zero loadings on the remaining factors. In FA terminology this type of variable is termed *factorially simple* or *marker* (because it defines only one factor). So, the I-C solution (which is usually prescribed in confirmatory factor analysis) is one in which all the variables are factorially simple.

The I-C solution described above is generally too restrictive for real data, especially if the variables are test items (see Ferrando & Lorenzo-Seva, 2000). More specifically, many personality and attitude items are actually factorially complex. So, an *ideal* I-C solution in which none of the items in the test is factorially complex is unrealistic. On the other hand, it is reasonable to assume that, in a multidimensional test, for each factor there will be a *nuclear* subset of factorially simple items that can clearly define the factor (i.e. markers). McDonald (2000, 2005) considered a specific requirement of this type in which there are at least 3 markers per factor (uncorrelated factors) or 2 markers per factor (correlated factors). He called this solution the *independent-cluster basis* (I-CB). An I-CB condition is sufficient to identify a solution with no rotational indeterminacies, and is very advantageous in terms of interpretation.

As in the bi-factor case, the I-CB solution can be restricted or semi-restricted. The semi-restricted procedure for fitting an I-CB solution is the same as the one used in the bi-factor case. For each marker, the target matrix in this case specifies zeros in the loadings corresponding to the factors not defined by the marker item. The remaining loadings are left unspecified. Figure 14 shows a target matrix for fitting an I-CB solution in a set of 11 items that are expected to be determined by three correlated factors: items 1 and 2 are defined as markers of the first factor (F1); items 5 and 6 are defined as markers of the second factor (F2); items 9 and 10 are defined as markers of the third factor (F3); finally, items 3, 4, 7, 8, and 11 are the items that are not defined as markers of any particular factor and so they are allowed to be factorially complex.

F1	F2	F3
*	0	0
*	0	0
*	*	*
*	*	*
0	*	0
0	*	0
*	*	*
*	*	*
0	0	*
0	0	*
*	*	*

Figure 14. A hypothetical target matrix for a semi-restricted independent-cluster-basis solution (two markers per factor)

The semi-restricted approach for obtaining a bi-factor or an I-CB solution are summarized in Figure 15.

- Define a target matrix reflecting partial knowledge of what the pattern should be.
- Rotate an arbitrary solution to the position that minimizes squared discrepancies between the target and the rotated solution.

Figure 15. Semi-restricted approach for obtaining a bi-factor or an ICB solution

Recommended further readings

McDonald (1999, in the references list) for sections 2, 3.1, 3.2. 3.3, 3.4, 4.1, 4.2, 4.3 and 4.4.

Mislevy (1986, in the references list) for sections 2, 3.1, 3.2. 3.3, 3.4, 4.1, 4.2, and 4.3.

Muthén (1993, in the references list) for sections 3.1, 3.2, and 3.4.

Reckase (2009, in the references list) for sections 4.2 and 4.3.

Reise, Morizot & Hays (2007, in the references list) for sections 4.4.

References

Baker, F. B. (1992). *Item response theory. Parameter estimation techniques*. New York: Marcel Dekker.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: an application of the EM algorithm. *Psychometrika.* 46, 443-459.

Briggs, N. E., & MacCallum, R.C. (2003). Recovery of weak common factors by maximum likelihood and ordinary least squares estimation. *Multivariate Behavioral Research, 38,* 25-56.

Browne, M.W. (1982) Covariance structures. In D.M. Hawkins (Ed.) *Topics in applied multivariate analysis* (pp 72-141). Cambridge: Cambridge University Press.

Chang, H., & Mazzeo, J. (1994). The unique correspondence of the item response function and item category response functions in polytomously scored item response models. *Psychometrika, 59,* 391-404.

Ferrando, P. J. (1999). Likert scaling using continuous, censored and graded response models: effects on criterion related validity. *Applied Psychological Measurement*, *23*, 161-175.

Ferrando, P. J. (2002). Theoretical and empirical comparisons between two models for continuous item responses. *Multivariate Behavioral Research, 37,* 521-542.

Ferrando, P. J. (2004). Kernel smoothing estimation of item characteristics functions for continuous personality items: An empirical comparison with the linear and the continuous response model. *Applied Psychological Measurement.* 28 , 95-109.

Ferrando, P. J., & Lorenzo-Seva, U. (2000). Unrestricted versus restricted factor analysis of multidimensional test items: some aspects of the problem and some suggestions. *Psicológica*. 21, 301-323.

Guilford, J. P., & Fruchter, B. (1973). *Fundamental statistics in psychology and education*. New York: McGraw-Hill.

Harman, H. H. (1976). *Modern factor analysis.* Chicago: Univ. of Chicago press.

Hofstee, W. K. B., Ten Berge, J.M.F., & Hendriks, A.A.J. (1998). How to score questionnaires. *Personality and Individual Differences, 25,* 897-909.

Knol, D. L. & Berger, M. P. F. (1991). Empirical comparisons between factor analysis and multidimensional item response models. *Multivariate Behavioral Research, 26,* 457-477.

Lee, Ch., Zhang, G. & Edwards, M.C. (2012). Ordinary least squares estimation of parameters in exploratory factor analysis with ordinal data. *Multivariate Behavioral Research, 47,* 314-339.

Lord, F. M. (1952). *A theory of test scores*. Psychometrika Monograph. No 7.

Lord, F. M. (1953). The relation of test score to the trait underlying the test. *Educational and Psychological Measurement, 13*, 517-548.

Lord, F. M. (1980). *Applications of Item Response Theory*. Hillsdale, New Jersey: LEA.

Lumsen, J. (1961). The construction of unidimensional tests. *Psychological Bulletin*, 58, 122-131

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah (NJ): LEA.

McDonald, R. P. (2000). A basis for multidimensional item response theory. *Applied Psychological Measurement.* 24 , 99-114.

McDonald, R. P. (2005). Semiconfirmatory factor analysis: The example of anxiety and depression. *Structural Equation Modeling,* 12, 163-172.

McDonald, R.P. & Ahlawat, K.S. (1974). Difficulty factors in binary data. *British Journal of Mathematical and Statistical Psychology. 27,* 82-99

McDonald, R. P., & Ho, R. M. (2002). Principles and practice in reporting structural equation analyses *Psychological Methods. 7,* 64-82.

McNemar, Q. (1969). *Psychological statistics.* New York: Wiley.

Mislevy, R. J. (1986) Recent developments in the factor analysis of categorical variables. *Journal of educational statistics.* 11, 3-31.

Muthén, B. (1993). Goodness of fit with categorical and other nonnormal variables. In K.A. Bollen & J.S. Long (Eds.) *Testing structural equation models* (pp. 205-234). Newbury Park: Sage.

Muthén, B., & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal Likert variables. *British Journal of Mathematical and Statistical Psychology, 38,* 171-189.

Olsson, U. (1979). On the robustness of factor analysis against crude classification of the observations. *Multivariate Behavioral Research, 14,* 485-500.

Parry, C. D. H. & McArdle, J. J. (1991). An applied comparison of methods for least-squares factor analysis of dichotomous variables. *Applied Psychological Measurement, 15*, 35-46.

Reckase, M. (2009). *Multidimensional Item Response Theory*. New York: Springer.

Reise, S. P., Morizot, J., & Hays, R.D. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Quality of Life Research*, 16, 19-31.

Reise, S. P., Moore, T. M., & Haviland, M.G. (2010). Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. *Journal of Personality Assessment*, 92, 544-559.

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores.* (Psychometrika Monograph No. 17). Iowa City: Psychometric Society.

Samejima, F. (1997). Graded response model. In W.J. van der Linden and R.K. Hambleton (Eds.) *Handbook of modern item response theory* (pp. 85-100). New York: Springer.

Tanaka, J. S., & Huba, G.J. (1985). A fit index for covariance structure models under arbitrary GLS estimation. *British Journal of Mathematical and Statistical Psychology*, *38*, 197-201.