Factor Analysis with Categorical Indicators: A Comparison Between Traditional and Latent Class Approaches

Jeroen K. Vermunt

Tilburg University

Jay Magidson

Statistical Innovations Inc.

1 INTRODUCTION

The linear factor analysis (FA) model is a popular tool for exploratory data analysis or, more precisely, for assessing the dimensionality of sets of items. Although it is well known that it is meant for continuous observed indicators, it is often used with dichotomous, ordinal, and other types of discrete variables, yielding results that might be incorrect. Not only parameter estimates may be biased, but also goodness-of-fit indices cannot be trusted. Magidson and Vermunt (2001) presented a nonlinear factor-analytic model based on latent class (LC) analysis that is especially suited for dealing with categorical indicators, such as dichotomous, ordinal, and nominal variables,

and counts. The approach is called latent class factor analysis (LCFA) because it combines elements from LC and traditional FA. This LCFA model is one of the LC models implemented in the Latent GOLD program (Vermunt & Magidson, 2000, 2003).

A disadvantage of the LCFA model is, however, that its parameters may be somewhat more difficult to interpret than the typical factor-analytic coefficients – factor loadings , factor-item correlations, factor correlations, and communalities . In order to overcome this problem, we propose using a linear approximation of the maximum likelihood estimates obtained with a LCFA model. This makes it possible to provide the same type of output measures as in standard FA, while retaining the fact that the underlying factor structure is identified by the more reliable nonlinear factor-analytic model.

Bartholomew and Knott (1999) gave a four-fold classification of latent variable models based on the scale types of the latent and observed variables; i.e., factor analysis , latent trait (LT) analysis, latent profile (LP) analysis , and latent class analysis . As shown in Table 1, in FA and LT models, the latent variables are treated as continuous normally distributed variables. In LP and LC models on the other hand, the latent variable is assumed to be discrete and to come from a multinomial distribution. The manifest variables in FA and LP model are continuous. In most cases, their conditional

distribution given the latent variables is assumed to be normal. In LT and LC analysis, the indicators are dichotomous, ordinal, or nominal categorical variables, and their conditional distributions are assumed to be binomial or multinomial.

[INSERT TABLE 1 ABOUT HERE]

The distinction between models for continuous and discrete indicators is not a fundamental one since the choice between the two should simply depend on the type of data. The specification of the conditional distributions of the indicators follows naturally from their scale types. A recent development in latent variable modeling is to allow for a different distributional form for each indicator. This can, for example, be a normal, student, log-normal, gamma, or exponential distribution for continuous variables, binomial for dichotomous variables, multinomial for ordinal and nominal variables, and Poisson, binomial, or negative-binomial for counts. Depending on whether the latent variable is treated as continuous or discrete, one obtains a generalized LT (Moustaki & Knott, 2000) or LC (Vermunt & Magidson, 2001) model.

The more fundamental distinction in Bartholomew's typology is the one between continuous and discrete latent variables. A researcher has to decide whether to treat the underlying latent variable(s) as continuous or discrete.

However, Heinen (1996) demonstrated that the distribution of a continuous latent variable can be approximated by a discrete distribution, and that such a discrete approximation may even be superior[1] to a misspecified continuous (usually normal) model. More precisely, Heinen (1996; also, see Vermunt, 2001) showed that constrained LC models can be used to approximate well-known unidimensional LT or item response theory (IRT) models [2], such as the Rasch, Birnbaum, nominal-response, and partial credit model. This suggests that the distinction between continuous and discrete latent variables is less fundamental than one might initially think, especially if the number of latent classes is increased. More precisely, as shown by Aitkin (1999; also, see Vermunt and Van Dijk, 2001; Vermunt, 2004), a continuous latent distribution can be approximated using a nonparametric specification; that is, by a finite mixture model with the maximum number of identifiable latent classes. An advantage of such a nonparametric approach is that it is not necessary to introduce possibly inappropriate and unverifiable assumptions about the distribution of the random effects.

[1] With superior we refer to the fact that mispecification of the distribution of the continuous latent variables may cause bias in the item parameter estimates. In a discrete or nonparametric specificaton, on the other hand, no assumptions are made about the latent distribution and, as a result, parameters cannot be biased because of mispecification of the latent distribution.

[2] We will use the terms latent trait (LT) and item response theory (IRT) interchangeably.

The proposed LCFA model is based on a multidimensional generalization of Heinen's (1996) idea: it is a restricted LC model with several latent variables. As exploratory FA, the LCFA model can be used to determine which items measure the same dimension. The idea of defining an LC model with several latent variables is not new: Goodman (1974) and Hagenaars (1990) proposed such a model and showed that it can be derived from a standard LC model by specifying a set of equality constraints on the item conditional probabilities. What is new is that we use IRT-like regression-type constraints on the item conditional means/probabilities[3] in order to be able to use the LC model with several latent variables as an exploratory factor-analytic tool. Our approach is also somewhat more general than Heinen's in the sense that it cannot only deal with dichotomous, ordinal, and nominal observed variables, but also with counts and continuous indicators, as well as any combination of these.

Using a general latent variable model as the starting point, it will be shown that several important special cases are obtained by varying the model assumptions. In particular, assuming 1) that the latent variables are dichoto-

[3]With regression-type constraints on the item conditional probabilities we mean that the probability of giving a particular response given the latent traits is restricted by means of a logistic regression model, or another type of regression model. In the case of continuous responses, the means are restricted by linear regression models, as in standard factor analysis.

mous or ordinal, and 2) that the effects of these latent variables on the transformed means are additive, yields the proposed LCFA model. We show how the results of this LCFA model can be approximated using a linear FA model , which yields the well-known standard FA output. Special attention is given to the meaning of the part that is ignored by the linear approximation and to the handling of nominal variables. Several real life examples are presented to illustrate our approach.

2 THE LATENT CLASS FACTOR MODEL

Let $\boldsymbol{\theta}$ denote a vector of L latent variables and $\mathbf{y}$ a vector of K observed variables. Indices ℓ and k are used when referring to a specific latent and observed variable, respectively. A basic latent variable model has the following form:

$$f(\boldsymbol{\theta}, \mathbf{y}) = f(\boldsymbol{\theta}) f(\mathbf{y}|\boldsymbol{\theta}) = f(\boldsymbol{\theta}) \prod_{k=1}^{K} f(y_k|\boldsymbol{\theta}),$$

where the primary model assumption is that the K observed variables are independent of one another given the latent variables $\boldsymbol{\theta}$, usually referred to as the local independence assumption (Bartholomew and Knott, 1999). The various types of latent variable models are obtained by specifying the distribution of the latent variables $f(\boldsymbol{\theta})$ and the K conditional item distributions $f(y_k|\boldsymbol{\theta})$. The two most popular choices for $f(\boldsymbol{\theta})$ are continuous multivari-

ate normal and discrete nominal. The specification for the error functions $f(y_k|\boldsymbol{\theta})$ will depend on the scale type of indicator k.[4] Besides the distributional form of $f(y_k|\boldsymbol{\theta})$, an appropriate link or transformation function $g(\cdot)$ is defined for the expectation of y_k given $\boldsymbol{\theta}$, $E(y_k|\boldsymbol{\theta})$. With continuous $\boldsymbol{\theta}$ (FA or LT), the effects of the latent variables are assumed to be additive in $g(\cdot)$; that is,

$$g[E(y_k|\boldsymbol{\theta})] = \beta_{0k} + \sum_{\ell=1}^{L} \beta_{\ell k}\theta_\ell, \tag{1}$$

where the regression intercepts β_{0k} can be interpreted as "difficulty" parameters and the slopes $\beta_{\ell k}$ as "discrimination" parameters. With a discrete $\boldsymbol{\theta}$ (LC or LP), usually no constraints are imposed on $g[E(y_k|\boldsymbol{\theta})]$.

The new element of the LCFA model is that a set of discrete latent variables is explicitly treated as multidimensional, and that the same additivity of their effects is assumed as in Equation 1. In the simplest specification, the latent variables are specified to be dichotomous and mutually independent, yielding what we call the *basic LCFA model*. An LCFA model with L dichotomous latent variables is, actually, a restricted LC model with 2^L latent classes (Magidson & Vermunt, 2001) . Our approach is an extension of Heinen's work to the multidimensional case. Heinen (1996) showed that LC models with

[4]The term error function is jargon from the generalized linear modeling framework. Here, it refers to the distribution of the unexplained or unique part (the error) of y_k.

certain log-linear constraints yield discretized versions of unidimensional LT models. The proposed LCFA model is a discretized multidimensional LT or IRT model . With dichotomous observed variables, for instance, we obtain a discretized version of the multidimensional two-parameter logistic model (Reckase, 1997).

A disadvantage of the (standard) LC model compared to the LT and LCFA models is that it does not explicitly distinguish different dimensions, which makes it less suited for dimensionality detection. Disadvantages of the LT model compared to the other two models are that it makes stronger assumptions about the latent distribution and that its estimation is computationally much more intensive, especially with more than a few dimensions. Estimation of LT models via maximum likelihood requires numerical integration: for example, with 3 dimensions and 10 quadrature points per dimension, computation of the log-likelihood function involves summation over 1000 ($=10^3$) quadrature points. The LCFA model shares the advantages of the LT model, but is much easier to estimate, which is a very important feature if one wishes to use the method for exploratory purposes. Note that a LCFA model with 3 dimensions requires summation over no more than 8 ($=2^3$) discrete nodes. Of course, the number of nodes becomes larger with more than two categories per latent dimension, but will still be much smaller

than in the corresponding LT model.

Let us first consider the situation in which all indicators are dichotomous. In that case, the most natural choices for $f(y_k|\boldsymbol{\theta})$ and $g(\cdot)$ are a binomial distribution function and a logistic transformation function. Alternatives to the logistic transformation are probit, log-log, and complementary log-log transformations. Depending on the specification of $f(\boldsymbol{\theta})$ and model for $g[E(y_k|\boldsymbol{\theta})]$, we obtain a LT, LC, or LCFA model. In the LCFA model,

$$\begin{aligned} f(\boldsymbol{\theta}) &= \pi(\boldsymbol{\theta}) = \prod_{\ell=1}^{L} \pi(\theta_\ell) \\ g[E(y_k|\boldsymbol{\theta})] &= \log\left[\frac{\pi(y_k|\boldsymbol{\theta})}{1-\pi(y_k|\boldsymbol{\theta})}\right] = \beta_{0k} + \sum_{\ell=1}^{L} \beta_{\ell k}\theta_\ell. \end{aligned} \tag{2}$$

The parameters to be estimated are the probabilities $\pi(\theta_\ell)$ and the coefficients β_{0k} and $\beta_{\ell k}$. The number of categories of each of the L discrete latent variables is at least 2, and θ_ℓ are the fixed category scores assumed to be equally spaced between 0 and 1. The assumption of mutual independence between the latent variables θ_ℓ can be relaxed by incorporation two-variable associations in the model for $\pi(\boldsymbol{\theta})$. Furthermore, the number of categories of the factors can be specified to be larger than two: A two-level factor has category scores 0 and 1 for the factor levels, a three-level factor scores 0, 0.5, and 1, etc.

The above LCFA model for dichotomous indicators can easily be extended

to other types of indicators. For indicators of other scale types, other distributional assumption are made and other link functions are used. Some of the possibilities are described in Table 2. For example, the restricted logit model we use for ordinal variables is an adjacent-category logit model. Letting s denote one of the S_k categories of variable y_k, it can be defined as

$$\log\left[\frac{\pi(y_k = s|\boldsymbol{\theta})}{\pi(y_k = s-1|\boldsymbol{\theta})}\right] = \beta_{0ks} + \sum_{\ell=1}^{L} \beta_{\ell k}\theta_\ell,$$

for $2 \leq s \leq S_k$.

[INSERT TABLE 2 ABOUT HERE]

Extensions of the basic LCFA model are among others that local dependencies can be included between indicators and that covariates may influence the latent variables and the indicators (Magidson & Vermunt, 2001, 2004). These are similar to extensions proposed for the standard latent class model (for example, see Dayton & McReady, 1988; Hagenaars, 1988, Van der Heijden, Dessens & Böckenholt, 1996) .

Similarly to standard LC models and IRT models, the parameters of a LCFA model can be estimated by means of maximum likelihood (ML) . We solved this ML estimation problem by means of a combination of an EM and a Newton-Raphson algorithm . More specifically, we start with EM and

switch to Newton-Raphson when close to the maximum likelihood solution. The interested reader is referred to Vermunt and Magidson (2000: Appendix).

3 LINEAR APPROXIMATION

As mentioned above, the proposed nonlinear LCFA model is estimated by means of ML. However, as a result of the scale transformations $g(\cdot)$, the parameters of the LCFA model are more difficult to interpret than the parameters of the traditional FA model. In order to facilitate the interpretation of the results, we propose approximating the maximum likelihood solution for the conditional means $\widehat{E}(y_k|\boldsymbol{\theta})$ by a linear model, yielding the same type of output as in traditional FA . While the original model for item k may, for example, be a logistic model, we approximate the logistic response function by means of a linear function.

The ML estimates $\widehat{E}(y_k|\boldsymbol{\theta})$ are approximated by the following linear function:

$$\widehat{E}(y_k|\boldsymbol{\theta}) = b_{0k} + \sum_{\ell=1}^{L} b_{\ell k}\theta_\ell + e_{k|\boldsymbol{\theta}}. \tag{3}$$

The parameters of the K linear regression models are simply estimated by means of ordinary least squares (OLS). The residual term $e_{k|\boldsymbol{\theta}}$ is needed because the linear approximation will generally not be perfect.

With 2 dichotomous factors, a perfect description by a linear model is

obtained by

$$\widehat{E}(y_k|\theta_1, \theta_2) = b_{k0} + b_{k1}\theta_1 + b_{k2}\theta_2 + b_{k12}\theta_1\theta_2;$$

that is, by the inclusion of the interaction between the two factors. Because the similarity with standard FA would otherwise get lost, interaction terms such as b_{k12} are omitted in the approximation.

Special provisions have to be made for ordinal and nominal variables. Because of the adjacent-category logit model specification

indexlogistic transformation, for ordinal variables, it is most natural to define $\widehat{E}(y_k|\boldsymbol{\theta}) = \sum_{s=1}^{S} s\,\widehat{\pi}(y_k = s|\boldsymbol{\theta})$.[5] With nominal variables, analogous to the Goodman and Kruskal tau-b (GK-τ_b), each category is treated as a separate dichotomous variable, yielding one coefficient per category. For category, we model the probability of being in the category concerned. These category-specific coefficients are combined into a single measure in exactly the same way as is done in the computation of the GK-τ_b coefficient. As is shown below, overall measures for nominal variables are defined as weighted averages of the category-specific coefficients.

The coefficients reported in traditional linear FA are factor loadings ($p_{\theta_\ell y_k}$), factor correlations ($r_{\theta_\ell \theta_{\ell'}}$), communalities or proportion explained item vari-

[5]The same would apply with other link functions for ordinal variables, such as with a cumulative logit link.

ances ($R^2_{y_k}$), factor-item correlations ($r_{\theta_\ell y_k}$), and in the case that there are local dependencies, also residual item correlations ($r_{e_k e_{k'}}$). The correlations $r_{\theta_\ell \theta_{\ell'}}$, $r_{\theta_\ell y_k}$, and $r_{y_k y_{k'}}$ can be computed from $\widehat{\pi}(\boldsymbol{\theta})$, $\widehat{E}(y_k|\boldsymbol{\theta})$, and the observed item distributions using elementary statistics computation. For example, the $r_{\theta_\ell \theta_{\ell'}}$ is obtained by dividing the covariance between θ_ℓ and $\theta_{\ell'}$ by the product of their standard deviations; that is,

$$r_{\theta_\ell \theta_{\ell'}} = \frac{\sigma_{\theta_\ell \theta_{\ell'}}}{\sigma_{\theta_{\ell'}} \sigma_{\theta_{\ell'}}} = \frac{\sum_{\theta_\ell} \sum_{\theta_{\ell'}} [\theta_\ell - \widehat{E}(\theta_\ell)][\theta_{\ell'} - \widehat{E}(\theta_{\ell'})]\, \widehat{\pi}(\theta_\ell \theta_{\ell'})}{\sqrt{\sum_{\theta_\ell} [\theta_\ell - \widehat{E}(\theta_\ell)]^2\, \widehat{\pi}(\theta_\ell)} \sqrt{\sum_{\theta_{\ell'}} [\theta_{\ell'} - \widehat{E}(\theta_{\ell'})]^2\, \widehat{\pi}(\theta_{\ell'})}},$$

where $\widehat{E}(\theta_\ell) = \sum_{\theta_\ell} \theta_\ell\ \widehat{\pi}(\theta_\ell)$.

The factor-factor ($r_{\theta_\ell \theta_{\ell'}}$) and the factor-item ($r_{\theta_\ell y_k}$) correlations can be used to compute OLS estimates for the factor loadings ($p_{\theta_\ell y_k}$), which are standardized versions of the regression coefficients appearing in Equation 3. The communalities or R^2 values ($R^2_{y_k}$) corresponding to the linear approximation are obtained with $r_{\theta_\ell y_k}$ and $p_{\theta_\ell y_k}$: $R^2_{y_k} = \sum_{\ell=1}^{L} r_{\theta_\ell y_k}\, p_{\theta_\ell y_k}$. The residual correlations ($r_{e_k e_{k'}}$) are defined as the difference between $r_{y_k y_{k'}}$ and the total correlation (not only the linear part) induced by the factors, denoted by $r^{\boldsymbol{\theta}}_{y_k y_{k'}}$.

The linear approximation of $\widehat{E}(y_k|\boldsymbol{\theta})$ is, of course, not perfect. One error source is caused by the fact that the approximation excludes higher-order interaction effects of the factors. More specifically, in the LCFA model pre-

sented in Equation **??**, higher-order interactions are excluded, but this does not mean that no higher-order interactions are needed to get a perfect linear approximation. On the other hand, with all interaction included, the linear approximation would be perfect. For factors having more than two ordered levels, there is an additional error source caused by the fact that linear effects on the transformed scale are nonlinear on the nontransformed scale. In order to get insight in the quality of the linear approximation, we also compute the R^2 treating the joint latent variable as a set of dummies; that is, as a single nominal latent variable.

As was mentioned above, for nominal variables, we have a separate set of coefficients for each of the S_k categories because each category is treated as a separate dichotomous indicator. If s denotes one of the S_k categories of y_k, the category-specific R^2 ($R^2_{y_k^s}$) equals

$$R^2_{y_k^s} = \frac{\sigma^2_{\widehat{E}(y_k=s|\boldsymbol{\theta})}}{\sigma^2_{y_k^s}},$$

where $\sigma^2_{\widehat{E}(y_k=s|\boldsymbol{\theta})}$ is the explained variance of the dummy variable corresponding to category s of item k, and $\sigma^2_{y_k^s}$ is its total variance defined as $\pi(y_k = s)[1 - \pi(y_k = s)]$. The overall $R^2_{y_k}$ for item k is obtained as a weighted sum of the S_k category-specific R^2 values, where the weights $w_{y_k^s}$

are proportional to the total variances $\sigma^2_{y_k^s}$; that is,

$$R^2_{y_k} = \sum_s^{S_k} \frac{\sigma^2_{y_k^s}}{\sum_t^{S_k} \sigma^2_{y_k^t}} R^2_{y_k^s} = \sum_s^{S} w_{y_k^s} R^2_{y_k^s}.$$

This weighting method is equivalent to what is done in the computation of the GK-τ_b, an asymmetric association measure for nominal dependent variables.

We propose using the same weighting in the computation of $p_{\theta_\ell y_k}$, $r_{\theta_\ell y_k}$, and $r_{e_k e_{k'}}$ from their category-specific counterpart. This yields

$$\begin{aligned}
p_{\theta_\ell y_k} &= \sqrt{\sum_{s=1}^{S_k} w_{y_k^s} \left(p_{\theta_\ell y_k^s}\right)^2} \\
r_{\theta_\ell y_k} &= \sqrt{\sum_{s=1}^{S_k} w_{y_k^s} \left(r_{\theta_\ell y_k^s}\right)^2} \\
r_{e_k e_{k'}} &= \sqrt{\sum_{s=1}^{S_k} \sum_{t=1}^{S_{k'}} w_{y_k^s} w_{y_{k'}^t} \left(r_{e_k^s, e_{k'}^t}\right)^2}.
\end{aligned}$$

As can be seen the signs are lost, but that is, of course, not a problem for a nominal variable.

4 EMPIRICAL EXAMPLES

4.1 Rater Agreement

For our first example we factor analyze dichotomous ratings made by 7 pathologists, each of whom classified 118 slides as to the presence or absence of carcinoma in the uterine cervix (Landis & Koch, 1977). This is an example of an inter-rater agreement analysis. We want to know whether the

ratings of the seven raters are similar or not, and if not, in what sense the ratings deviate from each other.

Agresti (2002), using standard LC models to analyze these data found that a two-class solution does not provide an adequate fit to these data. Using the LCFA framework, Magidson and Vermunt (2004) confirmed that a single dichotomous factor (equivalent to a two-class LC model) did not fit the data. They found that a basic two-factor LCFA model provides a good fit.

Table 3 presents the results of the two-factor model in terms of the conditional probabilities. These results suggest that Factor 1 distinguishes between slides that are "true negative" or "true positive" for cancer. The first class ($\theta_1 = 0$) is the "true negative" group because it has lower probabilities of a "+" rating for each of the raters than class two ($\theta_1 = 1$), the "true positive" group. Factor 2 is a bias factor, which suggests that some pathologists bias their ratings in the direction of a "false +" error ($\theta_2 = 1$) while others exhibit a bias towards "false –" error ($\theta_2 = 0$). More precisily, for some raters we see a too high probability of a "+" rating if $\theta_1 = 0$ and $\theta_2 = 1$ (raters A, G, E, and B) and for others we see a too high probability of a "–" rating if $\theta_1 = 1$ and $\theta_2 = 0$ (raters F and D). These results demonstrate the richness of the LCFA model to extract meaningful information from these data. Valuable

information includes an indication of which slides are positive for carcinoma,[6] as well as estimates of "false +" and "false –" error for each rater.

[INSERT TABLE 3 ABOUT HERE]

The left-most columns of Table 4 lists the estimates of the logit coefficients for these data. Although the probability estimates in Table 3 are derived from these quantities (recall Equation 2), the logit parameters are not as easy to interpret as the probabilities. For example, the logit effect of θ_1 on A, a measure of the validity of the ratings of pathologist A, is a single quantity, exp(7.74)=2,298. This means that among those slides at $\theta_2 = 0$, the odds of rater A classifying a "true +" slide as "+" is 2,298 times as high as classifying a "true –" slide as "+". Similarly, among those slides at $\theta_2 = 1$, the corresponding odds ratio is also 2,298.

[INSERT TABLE 4 ABOUT HERE]

We could instead express the effect of Factor 1 in terms of differences between probabilities. Such a linear effect is easier to interpret, but is not the same for both types of slides. For slides at $\theta_2 = 0$, the probability of classifying a "true +" slide as "+" is .94 higher than classifying a "true –"

[6]For each patient, we can obtain the posterior distribution for the first factor. This posterior distribution can be used determine whether a patient has carcinoma or not, corrected for rater bias (the second factor).

slide as "+"(.99-.06=.94) , while for slides at $\theta_2 = 1$, it is .59 higher (1.00 - .41=.59), markedly different quantities. This illustrates that for the linear model, a large interaction term is needed to reproduce the results obtained from the logistic LC model.

Given that a substantial interaction must be added to the linear model to capture the differential biases among the raters, it might be expected that the traditional (linear) FA model also fails to capture this bias. This turns out to be the case, as the traditional rule of choosing the number of factors to be equal to the number of eigenvalues greater than 1 yields only a single factor: The largest eigenvalue was 4.57, followed by 0.89 for the second largest. Despite this result, for purposes of comparison with the LCFA solution, we fitted a two-factor model anyway, using maximum likelihood for estimation.

Table 5 shows that the results obtained from Varimax (orthogonal) and Quartimax (oblique) rotations differ substantially. Hence, without theoretical justification for one rotation over another, FA produces arbitrary results in this example.

[INSERT TABLE 5 ABOUT HERE]

The three right-most columns of Table 4 present results from a linearization of the LCFA model using the following equation to obtain "linearized

loadings" for each variable y_k:

$$\widehat{E}(y_k|\theta_1, \theta_2) = b_{k0} + b_{k1}\theta_1 + b_{k2}\theta_2 + b_{k12}\theta_1\theta_2.$$

These 3 loadings have clear meanings in terms of the magnitude of validity and bias for each rater. They have been used to sort the raters according to the magnitude and direction of bias. The logit loadings do not provide such clear information .

The loading on θ_1 corresponds to a measure of validity of the ratings. Raters C, A, and G who have the highest loadings on the first linearized factor show the highest level of agreement among all raters. The loading on θ_2 relates to the magnitude of bias and the loading on $\theta_1\theta_2$ indicates the direction of the bias. For example, from Table 3 we saw that raters F and B show the most bias, F in the direction of "false –" ratings and B in the direction of "false +". This is exactly what is picked up by the nonlinear term: the magnitude of the loadings on the nonlinear term (Table 4) is highest for these 2 raters, one occurring as "+", the other as "–".

Table 4 also lists the communalities ($R^2_{y_k}$ values) for each rater, and decomposes these into linear and nonlinear portions (the "Total" column includes the sum of the linear and nonlinear portions). The linear portion is the part accounted for by $b_{k1}\theta_1 + b_{k2}\theta_2$, and the nonlinear part concerns the

factor interaction $b_{k12}\theta_1\theta_2$. Note the substantial amount of nonlinear variation that is picked up by the LCFA model. For comparison, the left-most column of Table 5 provides the communalities obtained from the FA model, which are quite different from the ones obtained with the LCFA model.

4.2 MBTI Personality Items

In our second example we analyzed 19 dichotomous items from the Myers-Briggs Type Indicator (MBTI) test – 7 indicators of the sensing-intuition (S-N) dimension, and 12 indicators of the thinking-feeling (T-F) personality dimension.[7] The total sample size was 8,344. These items were designed to measure 2 hypothetical personality dimensions, which were posited by Carl Jung to be latent dichotomies. The purpose of the presented analysis was to investigate whether the LCFA model was able to identify these two theoretical dimensions and whether results differed from the ones obtained with a traditional factor analysis.

We fitted 0-, 1-, 2-, and 3-factor models for this data set. Strict adherence to a fit measure like BIC or a similar criterion suggest that more than 2 latent factors are required to fit these data due to violations of the local independence assumption. This is due to similar wording used in several

[7]Each questionnaire item involves making a choice between two categories, such as, for example, between thinking and feeling, convincing and touching, or analyze and sympathize.

of the S-N items and similar wording used in some of the T-F items. For example, in a three-factor solution, all loadings on the third factor are small except those for S-N items S09 and S73. Both items ask the respondent to express a preference between "practical" and a second alternative (for item S09, 'ingenious'; for item S73, "innovative"). In such cases, additional association between these items exists which is not explainable by the general S-N (T-F) factor. For our current purpose, we ignore these local dependencies and present results of the two-factor model.

In contrast to our first example, the decomposition of communalities ($R^2_{y_k}$ values) in the right-most columns of Table 6 shows that a linear model can approximate the LCFA model here quite well. Only for a couple of items (T35, T49, and T70) is the total communality not explained to 2 decimal places by the linear terms only. The left-most columns of Table 6 compares the logit and linearized "loadings" ($p_{\theta_\ell y_k}$) for each variable. The fact that the latter numbers are bounded between -1 and +1 offers easier interpretation.

[INSERT TABLE 6 ABOUT HERE]

The traditional FA model also does better here than the first example. The first four eigenvalues are 4.4, 2.8, 1.1 and 0.9. For comparability to the LC solution, Table 7 presents the loadings for the two-factor solution under

Varimax (orthogonal) and Quartimax (oblique) rotations. Unlike the first example where the corresponding loadings showed considerable differences, these two sets of loadings are quite similar. The results are also similar to the linearized loadings obtained from the LCFA solution.

[INSERT TABLE 7 ABOUT HERE]

The right-most column of Table 7 shows that the communalities obtained from FA are quite similar to those obtained from LCFA. In general, these communalities are somewhat higher than those for LCFA, especially for items S27, S44, and S67.

Figure 1 displays the two-factor LCFA bi-plot for these data (see Magidson & Vermunt, 2001, 2004). The plot shows how clearly differentiated the S-N items are from the T-F items on both factors. The seven S-N items are displayed along the vertical dimension of the plot which is associated with Factor 2, while the T-F items are displayed along the horizontal dimension, which is associated with Factor 1. This display turns out to be very similar to the traditional FA loadings plot for these data. The advantage of this type of display becomes especially evident when nominal variables are included among the items.

[INSERT FIGURE 1 ABOUT HERE]

4.3 Types of Survey Respondents

We will now consider an example that illustrates how these tools are used with nominal variables. It is based on the analysis of 4 variables from the 1982 General Social Survey (white respondents) given by McCutcheon (1987) to illustrate how standard LC modeling can be used to identify different types of survey respondents.

Two of the variables ascertain the 1202 respondent's opinion regarding (A) the purpose of surveys (good, depends, or waste of time and money) and (B) how accurate survey are (mostly true or not true), and the others are evaluations made by the interviewer of (C) the respondent's levels of understanding of the survey questions (good, fair/poor) and (D) cooperation shown in answering the questions (interested, cooperative, or impatient/hostile). McCutcheon initially assumed the existence of 2 latent classes corresponding to 'ideal' and 'less than ideal' types . The purpose of the present analysis is to show how to apply the LCFA model with nominal indicators; that is, to answer the question as to whether these four items measure a single dimension as hypothesized by McCutcheon or whether there is two underlying dimensions. Note that it is not possible to use traditional factor analytic techniques with nominal indicators. A more elaborate analysis of this data

set is presented in Magidson and Vermunt (2004).

The two-class LC model – or, equivalently, the one-factor LC model – does not provide a satisfactory description of this data set. Two options for proceeding are to increase the number of classes or to increase the number of factors. The two-factor LC model fits very well, and also much better than the unrestricted three-class model that was selected as the final model by McCutcheon.

The logit parameter estimates obtained from the two-factor LC model are given in Table 8 and the linearized parameters are given in Table 9. The factor loadings ($p_{\theta_\ell y_k}$) show much clearer than the logit parameters the magnitude of the relationship between the observed variables and the two factors. As can be seen, the interviewers' evaluations of respondents and the respondents' evaluations of surveys are clearly different factors. The communalities ($R^2_{y_k}$) reported in the two right-most columns of Table 9 show that the linear approximation is accurate for each of the four items.

[INSERT TABLES 8 and 9 ABOUT HERE]

Figure 2 depicts the bi-plot containing the category scores of the four indicators. The plot shows that the first dimension differentiates between the categories of understanding and cooperation and the second between

the categories of purpose and accuracy. This display is similar to the plot obtained in multiple correspondence analysis (Van der Heijden, Gilula & Van der Ark, 1999).

[INSERT FIGURE 2 ABOUT HERE]

5 CONCLUSION

In this study, we compared LCFA with FA in 3 example applications where the assumptions of FA were violated. In the MBTI example, the resulting linear factor model obtained from standard FA provided results that were quite similar to those obtained with LCFA, even though the factors were taken to be dichotomous in the LCFA model. In this case, decomposition of the LCFA solution into linear and nonlinear portion suggested that the systematic portion of the results was primarily linear, and the linearized LCFA solution was quite similar to the FA solution. However, the LCFA model was able to identify pairs and small groups of items that have similar wording because of some violations of the assumption of local independence.

In the rater-agreement example, LCFA results suggested that the model contained a sizeable nonlinear component, and in this case the standard FA was unable to capture differential biases between the raters. Even when a second factor was included in the model, no meaningful interpretation of this

second factor was possible, and the loadings from 2 different rotations yielded very different solutions.

The third example illustrated the used of LCFA with nominal indicators, a situation for which standard FA techniques cannot be used at all. For this example, the factor-analytic loadings and communalities obtained with the proposed linear approximation provided much easier interpretation than the original logit parameters.

Overall, the results suggest improved interpretations from the LCFA approach, especially in cases where the nonlinear terms represent a significant source of variation. This is due to the increased sensitivity of the LCFA approach to all kinds of associations among the variables, not being limited as the standard linear FA model to the explanation of simple correlations.

The linearized LCFA parameters produced improved interpretation , but in the rater agreement example, a third (nonlinear) component model was needed in order to extract all of the meaning from the results. This current investigation was limited to two dichotomous factors. With three or more dichotomous factors, in addition to each two-way interaction, additional loadings associated with components for each higher-order interaction might also be necessary. Moreover, for factors containing three or more levels, additional terms are required. Further research is needed to explore these issues

in practice.

REFERENCES

Aitkin (1999). A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics*, *55*, 218-234.

Agresti, A. (2002). *Categorical data analysis*. Second Edition. New York: Wiley.

Bartholomew, D. J., & Knott, M. (1999). *Latent variable models and factor analysis.* London: Arnold.

Dayton, C. M., & Macready, G. B. (1988). Concomitant-variable latent-class models. *Journal of the American Statistical Association*, *83*, 173-178.

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

Hagenaars, J. A. (1988). Latent structure models with direct effects between indicators: local dependence models. *Sociological Methods and Research*, *16*, 379-405.

Hagenaars, J. A. (1990). *Categorical longitudinal data: Log-linear analysis of panel, trend and cohort data*. Newbury Park, CA: Sage.

Heinen, T. (1996). *Latent class and discrete latent trait models: Similarities and differences*. Thousand Oakes, CA: Sage Publications.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data, *Biometrics*, *33*, 159-174.

Magidson, J., & Vermunt, J. K. (2001). Latent class factor and cluster models, bi-plots and related graphical displays, *Sociological Methodology*, *31*, 223-264.

Magidson, J., & Vermunt, J. K. (2004). Latent class models. In D. Kaplan (Ed.), *Handbook of quantitative methods in social science research* (in press), Newbury Park, CA: Sage.

McCutcheon, A.L. (1987). *Latent class analysis*, Sage University Paper. Newbury Park, CA: Sage.

Moustaki, I., & Knott. M (2000). Generalized latent trait models. *Psychometrika*, *65*, 391-412.

Reckase, M. D. (1997). A linear logistic multidimensional model for dichotomous item response data. In W. J. van der Linden & R. K. Hambleton

(Eds.), *Handbook of modern item response theory* (pp. 271-286). New York, NJ: Springer.

Van der Heijden, P. G. M., Dessens, J., & Böckenholt, U. (1996). Estimating the concomitant-variable latent class model with the EM algorithm. *Journal of Educational and Behavioral Statistics*, *5*, 215-229.

Van der Heijden P. G. M., Gilula, Z., & Van der Ark, L. A. (1999). On a relationship between joint correspondence analysis and latent class analysis. *Sociological Methodology*, *29*, 147-186.

Vermunt, J. K. (2001). The use restricted latent class models for defining and testing nonparametric and parametric IRT models. *Applied Psychological Measurement*, *25*, 283-294.

Vermunt, J. K. & Van Dijk, L. (2001) A nonparametric random-coefficients approach: the latent class regression model. *Multilevel Modelling Newsletter*, 13, 6-13

Vermunt, J. K. (2004). An EM algorithm for the estimation of parametric and nonparametric hierarchical nonlinear models. *Statistica Neerlandica*, in press.

Vermunt, J. K., & Magidson, J. (2000). Latent GOLD 2.0 user's guide

[Software manual]. Belmont, MA: Statistical Innovations.

Vermunt, J. K., & Magidson, J. (2002). Latent class cluster analysis. In J. A. Hagenaars & A. L. McCutcheon (Eds.), *Applied latent class analysis* (pp 89-106). Cambridge, UK: Cambridge University Press.

Vermunt, J. K., & Magidson, J. (2003). Addendum to the Latent GOLD user's guide: Upgrade manual for version 3.0 [Software Manual]. Belmont, MA: Statistical Innovations.

Table 1. Four-fold Classification of Latent Variable Models

	Latent variable(s)	
Manifest variables	Continuous	Categorical
Continuous	Factor analysis	Latent profile analysis
Categorical	Latent trait analysis	Latent class analysis

Table 2. Distribution and Transformation Functions From Generalized Linear Modeling Family

Scale type y_k	Distribution $f(y_k\|\boldsymbol{\theta})$	Transformation $g\left[E(y_k\|\boldsymbol{\theta})\right]$
Dichotomous	Binomial	Logit
Nominal	Multinomial	Logit
Ordinal	Multinomial	Restricted logit
Count	Poisson	Log
Continuous	Normal	Identity

Table 3: Estimates of the Unconditional Latent Class Probabilities and the Conditional Item Probabilities Obtained from the two-factor LC Model with the Rater Agreement Data

		$\theta_1 = 0$ (True –)		$\theta_1 = 1$ (True +)	
		$\theta_2 = 0$	$\theta_2 = 1$	$\theta_2 = 0$	$\theta_2 = 1$
Class size		0.35	0.18	0.31	0.16
Rater F	–	1.00	0.99	0.80	0.11
	+	0.00	0.01	0.20	0.89
Rater D	–	1.00	0.98	0.61	0.11
	+	0.00	0.02	0.39	0.89
Rater C	–	1.00	1.00	0.22	0.14
	+	0.00	0.00	0.78	0.86
Rater A	–	0.94	0.59	0.01	0.00
	+	0.06	0.41	0.99	1.00
Rater G	–	0.99	0.46	0.01	0.00
	+	0.01	0.54	0.99	1.00
Rater E	–	0.94	0.28	0.03	0.00
	+	0.06	0.72	0.97	1.00
Rater B	–	0.87	0.01	0.03	0.00
	+	0.13	0.99	0.97	1.00

Table 4. Logit and Linearized Parameter Estimates for the Two-Factor LC Model Applied to the Rater Agreement Data

	Logit		Communality		Linearized		
Rater	θ_1	θ_2	Linear	Total	θ_1	θ_2	$\theta_1\theta_2$
F	7.2	3.4	0.45	0.60	0.53	0.38	0.40
D	6.0	2.6	0.47	0.54	0.62	0.26	0.26
C	7.2	0.5	0.68	0.68	0.82	0.04	0.04
A	7.7	2.4	0.72	0.75	0.82	0.18	-0.16
G	10.1	5.2	0.76	0.82	0.82	0.27	-0.25
E	6.4	3.8	0.65	0.75	0.72	0.35	-0.31
B	5.3	6.3	0.59	0.76	0.60	0.47	-0.42

Table 5. Loadings and Communalities Obtained when Applying a Traditional Two-Factor Model to the Rater Agreement Data

Rater	Commu-nality	Varimax		Quartimax	
		θ_1	θ_2	θ_1	θ_2
F	0.49	0.23	0.66	0.55	0.43
D	0.60	0.29	0.72	0.63	0.45
C	0.62	0.55	0.56	0.77	0.18
A	0.73	0.71	0.48	0.85	0.03
G	0.86	0.83	0.42	0.92	-0.09
E	0.78	0.82	0.31	0.86	-0.18
B	0.69	0.80	0.24	0.80	-0.22

Table 6. Logit and Linearized Parameter Estimates and Communalities for the Two-Factor LC Model as Applied to 19 MBTI Items

	Logit		Linear		Communality	
Item	θ_1	θ_2	θ_1	θ_2	Linear	Total
S02	-0.03	-1.51	-0.01	-0.61	0.37	0.37
S09	-0.01	-1.16	0.00	-0.50	0.25	0.25
S27	0.03	1.46	0.01	0.55	0.30	0.30
S34	-0.07	-1.08	-0.03	-0.45	0.21	0.21
S44	-0.11	1.13	-0.04	0.47	0.22	0.22
S67	-0.06	1.54	-0.02	0.53	0.28	0.28
S73	-0.01	-1.05	0.00	-0.46	0.21	0.21
T06	1.01	0.53	0.43	0.19	0.22	0.22
T29	1.03	0.59	0.44	0.20	0.23	0.23
T31	-1.23	-0.47	-0.52	-0.15	0.29	0.29
T35	-1.42	-0.29	-0.55	-0.09	0.31	0.32
T49	1.05	0.65	0.44	0.22	0.24	0.25
T51	1.32	0.30	0.53	0.09	0.29	0.29
T53	1.40	0.77	0.56	0.22	0.36	0.36
T58	-1.46	-0.12	-0.62	-0.03	0.38	0.38
T66	-1.23	-0.27	-0.54	-0.09	0.30	0.30
T70	1.07	0.61	0.43	0.19	0.22	0.23
T75	-1.01	-0.39	-0.45	-0.14	0.22	0.22
T87	-1.17	-0.45	-0.50	-0.15	0.28	0.28

Table 7. Loadings and Communalites from Traditional Factor Analysis of the 19 MBTI Items

	Quartimax		Varimax		Commu-
	θ_1	θ_2	θ_1	θ_2	nality
S02	0.08	-0.63	0.06	-0.63	0.40
S09	0.07	-0.50	0.06	-0.50	0.26
S27	-0.06	0.62	-0.05	0.62	0.38
S34	0.07	-0.46	0.06	-0.46	0.22
S44	-0.02	0.55	0.00	0.55	0.30
S67	-0.02	0.64	-0.01	0.64	0.41
S73	0.06	-0.46	0.05	-0.46	0.21
T06	-0.49	0.09	-0.49	0.10	0.25
T29	-0.49	0.10	-0.49	0.11	0.25
T31	0.56	-0.04	0.56	-0.05	0.32
T35	0.58	0.05	0.58	0.04	0.34
T49	-0.50	0.13	-0.50	0.15	0.27
T51	-0.57	-0.03	-0.57	-0.02	0.33
T53	-0.61	0.09	-0.61	0.10	0.38
T58	0.64	0.11	0.64	0.10	0.42
T66	0.58	0.05	0.58	0.03	0.33
T70	-0.49	0.10	-0.49	0.11	0.25
T75	0.50	-0.03	0.50	-0.04	0.25
T87	0.55	-0.04	0.55	-0.05	0.30

Table 8. Logit Parameter Estimates for the Two-Factor LC Model as Applied to the GSS'82 Respondent-Type Item

Item	Category	θ_1	θ_2
Purpose	Good	-1.12	2.86
	Depends	0.26	-0.82
	Waste	0.86	3.68
Accuracy	Mostly true	-0.52	-1.32
	Not true	0.52	1.32
Understanding	Good	-1.61	0.58
	Fair/poor	1.61	-0.58
Cooperation	Interested	-2.96	-0.57
	Cooperative	-0.60	-0.12
	Impatient/hostile	3.56	0.69

Table 9. Linearized Parameter Estimates and Communalities for the Two-Factor LC Model as Applied to the GSS'82 Respondent-Type Items

	Loadings		Communalities	
	θ_1	θ_2	Linear	Total
Purpose	0.14	0.45	0.24	0.26
Accuracy	0.15	0.55	0.33	0.33
Understanding	0.57	0.14	0.35	0.36
Cooperation	0.42	0.07	0.18	0.19

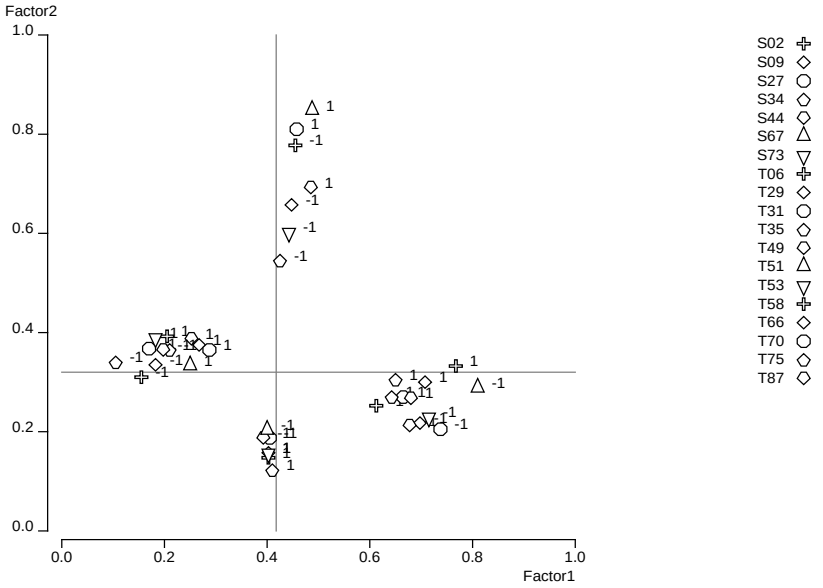

Figure 1. Bi-plot of Two-Factor LC Model as Applied to the 19 MBTI Items

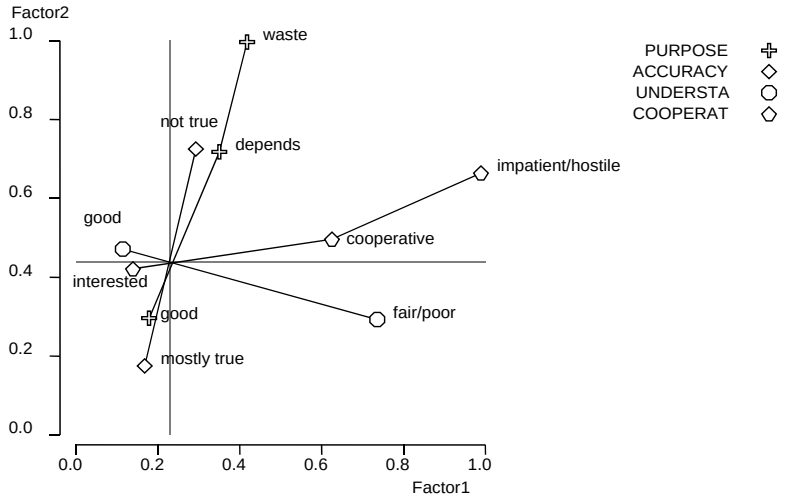

Figure 2. Bi-plot of Two-Factor LC Model as Applied to the GSS'82 Respondent-Type Items