

Tilburg University

Latent class factor and cluster models, bi-plots and tri-plots and related graphical displays

Magidson, J.; Vermunt, J.K.

Published in:
Sociological Methodology

Publication date:
2001

Document Version
Peer reviewed version

[Link to publication in Tilburg University Research Portal](#)

Citation for published version (APA):
Magidson, J., & Vermunt, J. K. (2001). Latent class factor and cluster models, bi-plots and tri-plots and related graphical displays. *Sociological Methodology*, 31(1), 223-264.

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

Take down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

LATENT CLASS FACTOR AND CLUSTER MODELS, BI-PLOTS AND RELATED
GRAPHICAL DISPLAYS

Jay Magidson,
Statistical Innovations

Jeroen K. Vermunt,
Tilburg University

Address of correspondence:

Jay Magidson
Statistical Innovations Inc.
Suite 007
375 Concord Ave.
Belmont MA. 02478

phone: (617) 489-4490 fax: (617) 489-4499

E-mail: jay@statisticalinnovations.com

LATENT CLASS FACTOR AND CLUSTER MODELS, BI-PLOTS AND RELATED GRAPHICAL DISPLAYS

We propose an alternative method of conducting exploratory latent class analysis that utilizes latent class factor models, and compare it to the more traditional approach based on latent class cluster models. We show that when formulated in terms of R mutually independent, dichotomous latent factors, the LC factor model has the same number of distinct parameters as an LC cluster model with $R+1$ clusters. Analyses over several data sets suggest that LC factor models typically fit data better and provide results that are easier to interpret than the corresponding LC cluster models. We also introduce a new graphical “bi-plot” display for LC factor models and compare it to similar plots used in correspondence analysis and to a “tri-plot” display for LC cluster models. New results on identification of LC models are also presented. We conclude by describing various model extensions and an approach for eliminating boundary solutions in identified and unidentified LC models, that we have implemented in a new computer program.

Acknowledgments:

The authors wish to thank Jeremy F. Magland, Leo A Goodman and Peter G.M. van der Heijden for helpful comments.

LATENT CLASS FACTOR AND CLUSTER MODELS, BI-PLOTS AND TRI-PLOTS

1. INTRODUCTION

Latent class (LC) analysis is becoming one of the standard data analysis tools in social, biomedical, and marketing research. While the traditional LC model described by Lazarsfeld and Henry (1968) and Goodman (1974a, 1974b) contains only nominal indicator variables, variants have been proposed for ordinal (Clogg 1988; Uebersax 1993; Heinen 1996) and continuous indicators (Wolfe 1970; McLachlan and Basford 1988; Fraley and Raftery 1998), as well as for combinations of variables of different scale types (Lawrence and Krzanowski 1996; Moustaki 1996; Hunt and Jorgensen 1999; Vermunt and Magidson 2001). This paper concentrates on exploratory LC analysis with nominal and ordinal indicators.

In an exploratory LC analysis, the usual approach is to begin by fitting a 1-class (independence) model to the data, followed by a 2-class model, a 3-class model, etc., and continuing until a model is found that provides an adequate fit (Goodman 1974a, 1974b; McCutcheon 1987). We refer to such models as LC cluster models since the T nominal categories of the latent variable serve the same function as the T clusters desired in cluster analysis (McLachlan and Basford 1988; Hunt and Jorgensen 1999; Vermunt and Magidson 2001).

Van der Ark and Van der Heijden (1998) and Van der Heijden, Gilula and Van der Ark (1999) showed that exploratory LC analysis can be used to determine the number of dimensions underlying the responses on a set of nominal items. A LC model with three classes, for example, can be seen as a two-dimensional model similar to a two-dimensional joint correspondence analysis (JCA). However, within the context of LC analysis, a more natural manner of specifying the existence of two underlying dimensions for a set of items is to specify a model containing two latent variables.

Goodman (1974b), Haberman (1979), and Hagenars (1990, 1993) proposed restricted 4-class LC models yielding confirmatory LC models with two latent variables. Their approach is confirmatory since, as in confirmatory factor analysis, it requires a priori

knowledge on which items are related to which latent variables. In *exploratory* data analysis settings, we do not know beforehand which items load on the same latent variable. Hence, in exploratory analyses with several latent variables, this approach has limited practical applicability.

In this paper, we propose combining the exploratory model fitting strategy of the traditional latent class model with the possibility of increasing the number of latent variables to study the dimensionality of a set of items. Our alternative model fitting sequence involves increasing the number of latent variables (factors) rather than the number of classes (clusters). We call the latter sequence the LC factor approach because of the natural analogy to standard factor analysis. The basic LC factor model contains R mutually independent, dichotomous latent variables. To exclude higher-order interactions, logit models are specified on the response probabilities. An interesting feature of the basic R -factor model is that it has exactly the same number of parameters as an LC cluster model with $T = R+1$ clusters. In section 2, we describe the two types of exploratory LC models using the log-linear formulation introduced by Haberman (1979).

Section 3 compares the use of LC cluster and factor models in several examples and describes various graphical displays that facilitate the interpretation of the results obtained from these models. In particular, we consider some variations (called “tri-plots”) of the ternary diagram originally proposed by Van der Ark and Van der Heijden (1998) for LC cluster models, and introduce a new display (called a “bi-plot”) for LC factor models to represent various kinds of information in a 2-dimensional factor space. These two graphs are compared to each other and to similar displays used in correspondence analysis.

Section 4 describes some important extensions of the basic LC factor model, such as various model modifications needed for a more confirmatory analysis and for the inclusion of covariates. In section 5, we discuss identification issues. The paper ends with some final remarks regarding the applicability of these models.

2. TWO APPROACHES FOR EXPLORATORY LATENT CLASS ANALYSIS

In this section we describe and compare two competing alternative approaches for exploratory LC analysis. The traditional approach utilizes LC cluster models, while the alternative is based on LC factor models. For the sake of simplicity of exposition, below we use the log-linear formulation of LC models introduced by Haberman (1979). In Appendix A, we give the alternative probability formulation of the two types of LC models, as well as the relationship between the two formulations.

2.1 The Latent Class Cluster Model

For concreteness, consider 4 nominal variables denoted A, B, C, and D. Let X represent a nominal latent variable with T categories. The log-linear representation of the LC cluster model with T classes is:

$$\ln(F_{ijklt}) = \lambda + \lambda_t^X + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_l^D + \lambda_{it}^{AX} + \lambda_{jt}^{BX} + \lambda_{kt}^{CX} + \lambda_{lt}^{DX} \quad (1)$$

where $i = 1, 2, \dots, I$; $j = 1, 2, \dots, J$; $k = 1, 2, \dots, K$; $l = 1, 2, \dots, L$; and $t = 1, 2, \dots, T$.

For convenience in counting distinct parameters and without loss of generality, we choose the following “dummy coding” restrictions to identify the parameters¹:

$$\lambda_1^X = \lambda_1^A = \lambda_1^B = \lambda_1^C = \lambda_1^D = 0$$

$$\lambda_{i1}^{AX} = \lambda_{j1}^{BX} = \lambda_{k1}^{CX} = \lambda_{l1}^{DX} = 0 \text{ for } i = 1, 2, \dots, I; j = 1, 2, \dots, J; k = 1, 2, \dots, K; l = 1, 2, \dots, L;$$

$$\text{and } \lambda_{1t}^{AX} = \lambda_{1t}^{BX} = \lambda_{1t}^{CX} = \lambda_{1t}^{DX} = 0 \text{ for } t = 2, 3, \dots, T.$$

¹ See Haberman (1979) for an alternative set of identifying restrictions based on ANOVA effects coding.

As can be seen, the LC model described in equation (1) has the form of a log-linear model for the five-way frequency table cross-classifying the 4 observed variables and the latent variable; that is, the table with cell entries F_{ijkl} . The assumed model contains one-variable terms (“main effects”) associated with the latent variable X and the four observed indicators A, B, C, and D, as well as all two-variable “interaction” terms that involve X which pertain to the association between X and each of the observed indicators. The one-variable effects are included because we do not wish to impose constraints on the univariate marginal distributions. The assumption that the observed responses to A, B, C, and D are mutually independent given $X = t$ (“local independence”) is imposed by the omission of all interaction terms pertaining to the associations between the indicators. As shown in Appendix A, this set of conditional independence assumptions can also be formulated in another way, yielding the probability formulation for the LC model.

Note that for the 1-class model, since $T=1$, the model described in equation (1) reduces to the usual log-linear model of mutual independence between the 4 observed variables:

$$\ln(F_{ijkl}) = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_l^D . \quad (2)$$

More generally, for models with any number of variables, we will denote the model of mutual independence as H_0 , and use it as a baseline to assess the improvement in fit to the data of various LC models. The number of distinct parameters² in the model of independence as described in equation (2) is:

$$\text{NPAR}(\text{indep}) = (I-1) + (J-1) + (K-1) + (L-1)$$

Expressing the number of distinct parameters in the model described in equation (1) as a function of $\text{NPAR}(\text{indep})$, yields:

$$\begin{aligned} \text{NPAR}(T) &= (T-1) + \text{NPAR}(\text{indep}) \times [1 + (T-1)] \\ &= (T-1) + \text{NPAR}(\text{indep}) \times T \end{aligned}$$

² By convention, we do not count λ as a distinct parameter because of the redundancy to the overall sample size, and we subtract 1 from the number of cells when computing degrees of freedom.

The number of degrees of freedom (DF) associated with the test of model fit is directly related to the number of distinct parameters in the model tested³.

$$\begin{aligned} \text{DF}(T) &= \text{IJKL} - \text{NPAR}(T) - 1 \\ &= \text{IJKL} - [1 + \text{NPAR}(\text{indep})] \times T \end{aligned}$$

Beginning with this baseline model (T=1), each time the number of latent classes (T) is incremented by 1 the number of distinct parameters increases by 1 + NPAR(indep), and, as a consequence, the degrees of freedom are reduced by 1 + NPAR(indep). The first additional parameter is the main effect for the additional latent class, and the NPAR(indep) further parameters correspond to the effects of each observed (manifest) variable on this additional latent class.

2.2 The Latent Class Factor Model

Certain LC models can be interpreted in terms of 2 or more component latent variables by treating those components as a joint variable (Goodman 1974b; McCutcheon 1987; Hagenaars 1990). For example, a 4-category latent variable $X = \{1, 2, 3, 4\}$ can be re-expressed in terms of 2 dichotomous latent variables $V = \{1, 2\}$ and $W = \{1, 2\}$ using the following correspondence:

	W=1	W=2
V=1	X = 1	X = 2
V=2	X = 3	X = 4

Thus, X=1 corresponds with V=1 and W=1, X=2 with V=1 and W=2, X=3 with V=2 and W=1, and X=4 with V=2 and W=2.

The LC cluster model given in (1) with T = 4 classes can be re-parameterized as an *unrestricted* LC factor model with two dichotomous latent variables V and W as follows:

³ It is customary when one or more distinct parameters are unidentified or not estimable (a boundary solution), to adjust the DF, increasing it by the number of such unidentified or not estimable parameters.

$$\begin{aligned} \ln(F_{ijklrs}) = & \lambda + \lambda_r^V + \lambda_s^W + \lambda_{rs}^{VW} + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_l^D + \lambda_{ir}^{AV} + \lambda_{jr}^{BV} + \lambda_{kr}^{CV} + \lambda_{lr}^{DV} \\ & + \lambda_{is}^{AW} + \lambda_{js}^{BW} + \lambda_{ks}^{CW} + \lambda_{ls}^{DW} + \lambda_{irs}^{AVW} + \lambda_{jrs}^{BVW} + \lambda_{krs}^{CVW} + \lambda_{lrs}^{DVW}, \end{aligned} \quad (3)$$

The correspondence between the two representations is that the one-variable terms pertaining to X are now written as $\lambda_{2(r-1)+s}^X = \lambda_r^V + \lambda_s^W + \lambda_{rs}^{VW}$, and the two-variable terms involving X as $\lambda_{i,2(r-1)+s}^{AX} = \lambda_{ir}^{AV} + \lambda_{is}^{AW} + \lambda_{irs}^{AVW}$, $\lambda_{j,2(r-1)+s}^{BX} = \lambda_{jr}^{BV} + \lambda_{js}^{BW} + \lambda_{jrs}^{BVW}$, etc. It is easy to verify that this re-parameterization does not alter the *number* of distinct parameters in the model.

We define the *basic* R-factor LC model as a *restricted* factor model that contains R mutually independent, dichotomous latent variables, containing parameters (“factor loadings”) that measure the association of each latent variable on each indicator. Specifically, the basic R-factor model is defined by placing two sets of restrictions on the unrestricted LC factor model. The resulting 2-factor LC model is a restricted form of the 4-class LC cluster model. Without these restrictions, the 2-factor model would be unconstrained and would be equivalent to a 4-cluster model.

The first set of restrictions sets to zero each of the 3-way and higher-order interaction terms. For the basic 2-factor model, we have $\lambda_{irs}^{AVW} = \lambda_{irs}^{BVW} = \lambda_{irs}^{CVW} = \lambda_{irs}^{DVW} = 0$. After imposing these restrictions, the 2-variable terms in the basic 2-factor model become

$$\lambda_{i,2(r-1)+s}^{AX} = \lambda_{ir}^{AV} + \lambda_{is}^{AW}, \quad \lambda_{j,2(r-1)+s}^{BX} = \lambda_{jr}^{BV} + \lambda_{js}^{BW}, \quad \text{etc.}$$

For variable A, λ_{ir}^{AV} represents the loading of A on factor V and λ_{is}^{AW} denotes the loading of A on factor W, etc. The second set of restrictions imposes mutual independence between the latent variables. For the 2-factor model, this latter restriction imposes independence in the 2-way table <VW>.

Although the basic R-factor model is a special case of an LC *cluster* model containing 2^R classes, we show in Appendix A that because of the restrictions of the type given above, the basic R-factor LC model is actually comparable to an LC cluster model with only $T = R+1$ clusters in terms of parsimony. This large reduction in number of

parameters will be sufficient to achieve model identification in many situations. That is, in practice, it will frequently be the case that the basic R-factor will be identified when the LC cluster model with 2^R classes is not.

[INSERT TABLE 1 ABOUT HERE]

Table 1 verifies the equivalence in number of parameters (and the associated degrees of freedom) between the various identified LC cluster models and the corresponding basic LC factor models in the case of 5 dichotomous indicator variables. From this table we can also calculate, for example, that the basic LC 2-factor model requires $23 - 17 = 6$ fewer parameters than the 4-class LC cluster model. This reduction corresponds to the 5 restrictions $\lambda_{irs}^{AVW} = \lambda_{irs}^{BVW} = \lambda_{irs}^{CVW} = \lambda_{irs}^{DVW} = \lambda_{irs}^{EVW} = 0$, plus the restriction that V and W are independent. (See Appendix A for a simple formula for calculating the number of such restrictions in the more general case.)

[INSERT TABLE 1 ABOUT HERE]

We conclude this section by noting an important difference between our LC factor model and the LC models with several latent variables proposed by Goodman (1974b), Haberman (1979), McCutcheon (1987), and Hagenaars (1990, 1993). The basic LC factor model described above includes all factor loadings between the latent variables and the indicators. This means that no assumptions need be made about which indicators are related to which latent variables. This makes this LC factor model better suited for exploratory data analysis than the LC models with several latent variables described in the literature.

Thus far we have described two alternative approaches for exploratory LC analysis, one involving the fitting of LC cluster models, the other fitting basic LC factor models. In the next section we consider some examples to illustrate and compare their performance on real data and introduce graphical displays that facilitate the interpretation of the obtained results.

3. EXAMPLES AND GRAPHICAL DISPLAYS

Comparison of the two approaches for exploratory LC analysis across several data sets found that the factor approach resulted in a more parsimonious and easier to interpret model almost every time. Since our selection of data sets was not random, we do not present those results here. Rather, for purposes of illustration, this section considers the analysis from two data sets where a basic 2-factor model fits the data. In the first example, the comparable cluster model also provides an acceptable (but not as good) fit to the data; in the second example, the comparable cluster model provides a *much* worse fit, one that is not acceptable for these data.

This section also introduces graphical displays useful in displaying results from LC cluster and factor models. Details on the computation of the conditional probabilities appearing in the plots are given in Appendix B.

3.1. Example 1: 1982 General Social Survey Data

Our first example, taken from McCutcheon (1987) and reanalyzed by Van der Heijden, Gilula, and Van der Ark (1999) involves four categorical variables from the 1982 General Social Survey. Two items are evaluations of surveys by white respondents and the other two are evaluations of these respondents by the interviewer (see Table 2). A summary of various LC models fit to these data is given in Table 3.

[INSERT TABLE 2 ABOUT HERE]

[INSERT TABLE 3 ABOUT HERE]

Model H_0 is the baseline model given in equation (2) which specifies mutual independence between all four variables. Model H_0 is a 1-class LC model (a 1-cluster model) which can also be interpreted as the equivalent 0-factor LC model. Since $L^2 = 257.26$ with $DF = 29$, this model is rejected. Next, consider the 2-class model (H_1) that can be interpreted as either a 2-cluster model or the equivalent 1-factor model where the factor

is dichotomous. The L^2 is now reduced to 79.34, a 69.1% reduction from the baseline model, but too high to be acceptable with $DF = 22$.

Next, consider the two 15-DF models⁴ -- H_{2C} , the 3-cluster model and H_{2F} , the basic 2-factor model. Each of these models provide an adequate fit to the data, although the factor model fits better, the L^2 being half that of the comparable cluster model. For comparison, Table 3 also provides results for the 4-cluster model (H_3). Among the first 5 models listed in Table 3, H_{2F} is preferred according to the BIC criteria. The last 2 models in Table 3 are extended models that will be discussed in the next section.

[INSERT TABLE 4 ABOUT HERE]

Table 4 compares results obtained from the 3-cluster Model (H_{2C}) with that from the basic 2-factor model (H_{2F}). The cell entries in the left-most columns are “rescaled parameter estimates” suggested by Van der Heijden, Gilula, and Van der Ark (1999), and represent the estimated *conditional* probabilities of being a member of one of the three clusters. The right-most columns contain corresponding quantities for the basic 2-factor model, representing the estimated probabilities of being at level 1 for each of the 2 factors. *Unconditional* membership probabilities for the clusters and for level 1 of the factors are given in the last row of the table.

Graphical displays of the conditional probabilities reported in Table 4 are useful in comparing results between the two models. For the 3-cluster model H_2 , Van der Heijden, Gilula, and Van der Ark (1999, Figure 4) present a ternary diagram for visualizing the results and show the close relationship to 2-dimensional plots produced by joint correspondence analysis (JCA). A slightly modified graphic, referred to as the “tri-plot” display by Vermunt and Magidson (2000) is given in Figure 1 for the 3-cluster model H_{2C} . The shaded triangle in Figure 1 with lines emanating to the sides represents

⁴ For both models H_{2C} and H_{2F} , the maximum likelihood solution contains 2 boundary solutions and hence, by convention (see note 3) we increased the DF by 2. For model H_{2C} , McCutcheon (1987) reported an adjusted DF of 16, increasing the usual DF by only 1 because the solution reported was not fully converged and contained, therefore, only 1 boundary solution. The solution presented in Van der Heijden et. al. (1999) is the same solution as that presented here (containing 2 boundary solutions) but they also misreport the DF to be 16 instead of 17.

the overall sample which is plotted at the point corresponding to the unconditional membership probabilities for the clusters.

[INSERT FIGURE 1 ABOUT HERE]

[INSERT FIGURE 2 ABOUT HERE]

A different display for LC factor-models called the “bi-plot”⁵ (Vermunt and Magidson, 2000) is given in Figure 2 for the 2-factor model H_{2F} . For comparability to the tri-plot where cluster 1 is assigned to the top vertex, we take factor 1 to be the *vertical* axis and factor 2 the horizontal. By comparing these plots we can see the large degree of similarity between the models, the primary difference being the relative positioning of COOPERATION = Impatient/ Hostile and UNDERSTANDING = Fair, poor.

[INSERT FIGURE 3 ABOUT HERE]

Lines connecting the categories of a variable can make it easier to see to which factor the variables are most related. For example, Figure 3 shows that separation between the categories of the two respondent evaluation variables, PURPOSE and ACCURACY occurs primarily along Factor 2 (the horizontal axis in Figure 3) while for the two interviewer evaluation variables, UNDERSTANDING and COOPERATION separation occurs primarily along Factor 1 (the vertical axis). This makes clear that Factor 1 pertains primarily to the interviewer valuation while Factor 2 pertains primarily to the respondent valuation. These two factors are not only distinct (i.e., the 1-factor model H_1 does not fit these data) but according to model H_{2F} , they are mutually independent.

⁵ In the context of correspondence analysis, the term “biplot” refers to a particular joint display of points representing both the rows and columns of a frequency table (Greenacre, 1993). On the other hand, Gower and Hand (1996) stress that the “bi” in biplots arises from the fact that cases and variables are presented in the same plots. In Vermunt and Magidson (2000), we chose the term “bi-plot” because of the similarity of our plots to the plots used in correspondence analysis. However, despite the fact that in most of our examples we depict only variable categories, it is also possible to depict cases (or answer patterns) in our plots as we illustrated in our Figures 4, 6 and 8. For more detail about our plots see Appendix B.

Since our models yield estimated membership probabilities for each individual case, both displays can easily be extended to include points for individual cases and covariate levels as well as any other desired groupings of the cases (see Appendix B). Our methodology is unified in the sense that the same methods and models that yield our tri-plots for LC cluster models also yield the bi-plots for the LC factor models. Our tri-plot display can be more easily extended in this manner than the methods proposed by Van der Heijden, Gilula, and Van der Ark (1999) with the ternary diagram. In our next example we will illustrate the inclusion in our plots of cases by including specific cases with selected response patterns. Then in section 4, we show how the display of *all* response patterns can be used to identify a natural ordering between the classes (when such an ordering exists), and we describe two different approaches for overlaying covariate values (levels) onto the displays.

The bi-plots offer several advantages over the related plots produced in correspondence analysis (CA) even when the data justifies a 2-dimensional CA solution. That is because the 2-dimensional CA solution is closely related to the 3-cluster solution (Gilula and Haberman 1986; De Leeuw and Van der Heijden, 1991) which we have found typically does not fit the data as well as the 2-factor solution. As suggested in this paper, the LC factor models generally provide simpler explanations of data than LC cluster models and the related canonical models used in CA and principal components analysis.

Our LC factor model is more closely related to traditional factor analysis than to CA. Advantages over traditional factor analysis include 1) the variables can include different scale types – nominal, ordinal, continuous and/or counts, 2) solutions are typically uniquely identified and interpretable without the need for a rotation – there is no rotational indeterminacy, and 3) factor scores can be obtained for each case without the need for additional assumptions. Like traditional factor analysis, LC factor analysis can be used as a first step in a more confirmatory analysis. Later in this paper (section 4) we describe a more confirmatory analysis of the data analyzed above.

3.2. Example 2: Rheumatoid Arthritis Data

Our second example consists of 5 dichotomous responses obtained from a mail survey regarding various musculo-skeletal symptoms (see Table 5). Specifically, persons were asked whether they had any of the following symptoms today: back pain, neck pain, joint pain, joint swelling, and joint stiffness. For further details see Wasmus, et al. (1989).

[INSERT TABLE 5 ABOUT HERE]

The traditional LC cluster approach, as applied by Kohlmann and Formann (1997) to these data, rejects the 1-, 2-, and 3-class models in favor of the 4-class model which provides an acceptable fit to the data ($L^2 = 8.4$ with 8 degrees of freedom; $p = .39$). The BIC statistic also selects the 4-class model as the one to be preferred among the LC cluster models listed in Table 6.

[INSERT TABLE 6 ABOUT HERE]

The close relationship between the latent class *cluster* model and the canonical model (Gilula and Haberman 1986; De Leeuw and Van der Heijden, 1991) justifies a 2-dimensional display such as that produced in joint correspondence analysis (JCA) when the 3-cluster model is true (Van der Heijden, Gilula, and Van der Ark 1999). On the other hand, when the 3-class model must be *rejected* as not providing an adequate fit to data, as in the present example, the 2-dimensional JCA display can not provide a complete description of these data because a third dimension is also needed. However, as we show below, a *different* 2-dimensional display obtained from the LC factor model *does* provide a complete description of these data.

[INSERT TABLE 7 ABOUT HERE]

Table 7 provides a closer look at the differences between the 3- and 4-class solutions to these data. We see that for the most part, the 4-class solution maintains classes 1 and 2 from the 3-class solution, but splits class 3 into two separate clusters. One way to visualize the close relationship between these two solutions is to combine classes 3

and 4 of the 4-class solution and compare the resulting tri-plot (displayed in Figure 5) with the original tri-plot from the 3-cluster model (Figure 4). As can be seen, these plots are almost identical, adding visual support to our conclusion (based on inspection of Table 7) that the primary difference between the two solutions is the splitting of class 3 into separate clusters. However, these plots do not describe the significant differences that exist between clusters 3 and 4 of the 4-cluster solution.

[INSERT FIGURES 4 and 5 ABOUT HERE]

Results from fitting various basic factor models to these data are also included in Table 6. In particular, we see that despite the fact that the 3-cluster model H_{2C} does *not* provide an adequate fit to these data, the comparable LC factor model H_{2F} which posits two dichotomous factors, provides an excellent fit. While the traditional exploratory approach yields the 4-class LC cluster model H_{3C} , this model requires 3 dimensions for a display of the results. On the other hand, the alternative approach yields factor model H_{2F} , which justifies a valid 2-dimensional display without the necessity of collapsing or otherwise reducing the variables in the model. The resulting bi-plot presented in Figure 6 shows that JOINT, SWELL and STIFF are more strongly related to factor 1 (the arthritis factor), and BACK and NECK to factor 2 (the pain factor).

[INSERT FIGURE 6 ABOUT HERE]

In most cases where models suggest that at least 2 dimensions are needed to provide an adequate fit to the data, it seems reasonable to expect there to be 2 underlying factors and hence at least 4 different classes to take into account both the ‘low’ and ‘high’ levels of each factor – i.e., (low, low), (high, low), (low, high) and (high, high). If this speculation is true, it would explain why the factor approach typically provides a better fit to real data. Closer inspection of the results of the 4-cluster model parameters reported in Table 7 shows that, actually, the 4-cluster model also suggests a two-dimensional solution: the 4 clusters correspond to the (low, low), (high, low), (low, high) and (high, high) combinations of the same two dimensions encountered in the 2-factor model.

[INSERT TABLE 8 ABOUT HERE]

Using BACK and NECK as proxies for factor 2 and the other variables for factor 1, we selected 4 response patterns as proxies for the 4 classes. Table 8 compares the estimates of the expected frequency counts obtained from models H_{2C} , H_{3C} , and H_{2F} for these 4 selected response patterns. We see that the 3-class cluster model fails to provide a good estimate for respondents who reported having all 5 pain symptoms – the (high, high) group.

Overall, the expected frequencies estimated under the 3-cluster model differ significantly from the observed frequencies for 7 of the 32 response patterns, while the other 2 models provide good estimates for *all* response patterns. The 4 selected response patterns (or cases) are plotted in Figures 4 and 6 using the symbols ①, ②, ③, and ④. The symbol ④ appears in reverse shading as ④ in Figure 4 to indicate the lack of fit. Figure 6 shows that these 4 response patterns appear in the 4 corners of the bi-plot, suggesting that they are in fact good indicators of the (low, low) ... (high, high) levels of the joint factor. Figure 4 on the other hand shows that 3 clusters are inadequate to separate cases with response patterns 3 and 4, and indicates that the estimate of the expected count for response pattern 4 is poor.

4. SOME EXTENSIONS OF THE BASIC LC FACTOR MODEL

In this section we consider some modifications and extensions of the basic LC factor model that may be of interest in certain situations. First, although in example 1 we treated the trichotomous variables COOPERATE (A) and PURPOSE (C) as nominal, they can be treated as ordinal in several different ways. The most straight-forward approach is to assume the middle category to be equidistant from the others and modify the log-linear model described in equation (3) by using the uniform scores v_i^A and v_k^C

$$v_i^A = \{0 \text{ if } i = 1, 0.5 \text{ if } i=2, 1 \text{ if } i = 3\}$$

$$v_k^C = \{0 \text{ if } k = 1, 0.5 \text{ if } k=2, 1 \text{ if } k = 3\}$$

for the categories of variables A and C. Secondly, analogous to confirmatory factor analysis, we may wish to allow the two factors V and W to be correlated (with association parameter γ_{rs}^{VW}) and restrict the variables COOPERATION (A) and UNDERSTANDING (B) to load only on factor 1 and PURPOSE (C) and ACCURACY (D) to load only on factor 2. The log-linear representation for a confirmatory model of this type as compared to the basic 2-factor model in Appendix A is as follows:

$$\begin{aligned} \gamma_{rs}^{VW} &\neq 0; \\ \lambda_{ir}^{AV} &= \lambda_r^{AV} v_i^A; \lambda_{ks}^{CW} = \lambda_s^{CW} v_k^C; \quad \text{where} \quad i,k = 1,2,3; \quad j,l,r,s = 1,2; \\ \lambda_{is}^{AW} &= \lambda_{js}^{BW} = \lambda_{jr}^{CV} = \lambda_{ks}^{DV} = 0. \end{aligned}$$

The results of fitting this restricted 2-factor model (H_{R2F}) are reported in Table 3. These suggest that this model fits the data very well ($L^2 = 22.17$, $DF=23$; $p = .51$). The corresponding bi-plot is shown in Figure 7.

[INSERT FIGURE 7 ABOUT HERE]

Our examples thus far utilized only dichotomous factors. To extend the factor model so that any factor may contain more than 2 ordered levels, we assign equidistant numeric scores between 0 and 1 to the levels of the factor. Clogg (1988) and Heinen (1996) used the same strategy for defining LC models that are similar to certain latent trait models. The use of fixed scores for the factor levels in the various two-way interaction terms guarantees that each factor captures a single dimension. For factors with more than two levels, in the bi-plot we display conditional means rather than conditional probabilities (see Appendix B). Note that if we assign the score of 0 to the first level and 1 to the last level (or vice versa), for dichotomous factors the conditional mean equals the conditional probability of being at level 2 (or level 1).

Finally, the extension to include covariates in a log-linear LC model is straightforward. To illustrate the use of covariates and the extension to a 3-level factor, we will use the depression scale data for white respondents from the “Problems of Everyday Life” study conducted in 1972 by Pearlin (Pearlin and Johnson 1977) as reported

separately for males and females (Schaeffer,1988). Persons who reported having the symptom during the previous week were coded 1, all others 0. The symptoms measured were lack of enthusiasm, low energy, sleeping problem, poor appetite and feeling hopeless.

Gender was included in the model as an *active* covariate (see the discussion in Appendix B on ‘active vs. inactive covariates’). Note that in the case of a single covariate, the log-linear LC model is identical whether GENDER is treated as a covariate or as another indicator (Clogg 1981; Hagenaars 1990).

[INSERT TABLE 9 ABOUT HERE]

[INSERT TABLE 10 ABOUT HERE]

Table 9 shows the results from fitting various LC models to these data. The traditional strategy required 3 classes as neither the 1- or 2-class models provided adequate solutions. We see again that the basic 2-factor model fits the data better than the comparable 3-cluster model. The results for the 3-cluster solution are shown in Table 10 in terms of conditional response probabilities. Notice that those probabilities conditional on cluster 2 are ordered between the corresponding probabilities conditional on clusters 1 and 3, a pattern that is consistent with the depression scale being uni-dimensional, and suggests that we consider fitting a 3-level 1-factor model to these data.

Table 9 shows that the 3-level factor solution is very similar to that given by the 3-class solution. Both suggest that 10% of the population are in the “depressed” group (cluster 3 in the cluster model and level 3 in the factor model), and the rest are about equally distributed among the “healthy” (cluster 1) and the “troubled” cluster 2. The 3-level model provides an acceptable fit to these data and only contains one parameter more than the 2-class model (see Table 9). Unlike the 3-class extension to the 2-class model which requires 7 additional parameters, the 3-level model provides an attractive alternative to the 3- (unordered) class model. The BIC suggests that the 3-level 1-factor model should be preferred over all models including the basic 2-factor model.

In our experience with various data, increasing the number of levels in a factor does often provide a significant improvement in fit. This is, however, not always the case. For

example, with our first data set we found that 2 distinct factors were required to provide an adequate fit to the data. In that situation, increasing the number of levels from 2 to 3 in the single factor solution provides no benefit. Table 3 shows only a slight, non-significant reduction in the L^2 due to the inclusion of the additional parameter -- from 79.34 for the 1-factor 2-level solution to 77.25 for the 1-factor 3-level solution. On the other hand, in the present example, the addition of this single parameter causes a reduction of the L^2 from 138.5 for the 1-factor 2-level solution to 67.0 under the 1-factor 3-level model (see Table 9).

[INSERT FIGURE 8 ABOUT HERE]

An informative graph can provide an attractive alternative to a table (such as TABLE 10) when the goal is to determine whether a natural ordering exists among a set of clusters. For example, a standard profile plot will show immediately that the conditional probabilities associated with cluster 2 always fall between the corresponding conditional probabilities associated with clusters 1 and 3.

As an alternative to the profile plot, we will now examine the implications obtained from a tri-plot (FIGURE 8) of the 3-cluster solution which includes a point for each observation (i.e., each observed response pattern). Note the obvious pattern that the points appear primarily along the left and right sides of the triangle, and not along the base. This visual pattern can be interpreted as follows -- among persons who are likely to be “troubled“ (those with response patterns plotted near the top vertex, associated with cluster 2), there is a substantial amount of overlap with the other clusters. Some of these cases also have a substantial probability of belonging to the “healthy” cluster and some have a substantial probability of belonging to the “depressed” cluster. However, there is virtually *no* overlap between those likely to be “healthy” and those likely to be “depressed” (the inner part of the base of the triangle contains no points). This asymmetric pattern suggests that cluster 2 (“troubled”) is the middle cluster.

[INSERT FIGURE 9 ABOUT HERE]

In both the 3-cluster model and the 3-level 1-factor model, we find that GENDER has a significant relationship with the latent variable, females being more likely to be in the depressed group. Figure 9 displays 2 uni-plots resulting from the 3-level factor model (the bi-plot reduces to the uni-plot in the case of a single factor). The top uni-plot was obtained using GENDER as an active covariate. For comparison, the uni-plot at the bottom of Figure 9 was obtained using GENDER as an inactive covariate (it's effect is not included in the model). Being "inactive" implies that if the 'male' and 'female' symbols were removed from the latter, it would be equivalent to the uni-plot that would be obtained using a 3-level model that *excludes* GENDER from the model (see Appendix B). The lesser distance between the 'male' and 'female' symbols in the latter uni-plot (as compared to that displayed at the top of Figure 9) reflects the reduced association between GENDER and the latent variable, which is the result of the well-known attenuation phenomenon. In general, inclusion of covariates in a model can provide useful descriptive information on the latent variable(s). The decision to treat a covariate as active or inactive is largely a matter of personal preference.

5. IDENTIFICATION ISSUES

In some situations, LC models may not be identified. Two well-documented examples of LC models that are not identified without further constraints are the unrestricted 3-class model for 4 dichotomous items (Goodman 1974a) and the unrestricted 2- and 3-class models for 2 polytomous items (Gilula and Haberman 1986; De Leeuw and Van der Heijden, 1991; Clogg 1995; Van der Ark, Van der Heijden and Sikkel 1999).

The formal method to check for identification of a LC model is by means of the expected information matrix (Formann 1992).⁶ If all model parameters are identified, this information matrix will be full rank; that is, all its eigenvalues will be larger than zero. On

⁶ The expected information matrix is the negative of the expected value of the matrix of second-order derivatives to all model parameters.

the other hand, if k model parameters are not identified, k eigenvalues will be equal to zero. To get more insight in the identifiability of the LC factor model, we determined the rank of the information matrix for various hypothesized LC cluster and LC factor models.⁷ In particular, we studied 3 situations in which there might be identification problems; that is, tables of 4 dichotomous items, of 5 dichotomous items, and of 2 polytomous items with 4 and 5 categories. The results are reported in Table 11.

[INSERT TABLE 11 ABOUT HERE]

As can be seen, in all situations in which the LC cluster model with $R+1$ clusters is identified, the LC factor model with R factors is also identified. However, in two situations, we see that the LC factor model has *fewer* unidentified parameters than the corresponding LC cluster model having the same number of distinct parameters. For example, we see that while the 3-cluster model for 4 dichotomous items is *not* identified (it has one unidentified parameter), the 2-factor model is exactly identified and hence requires no identifying restrictions. Also, we see that while the 3-cluster model for a 4-by-5 table has 6 unidentified parameters, the 2-factor model has only 4. These results on identification show that all models presented in the foregoing examples are identified.

[INSERT TABLE 12 ABOUT HERE]

[INSERT TABLE 13 ABOUT HERE]

Consider the classic 4x5 table given by Fisher (1940) classifying school children in Caithness according to their hair and eye colors (Table 12). Table 13 provides results from various LC models. Gilula and Haberman (1986) showed that the 1-component canonical model does not fit these data but a 2-component model does ($L^2 = 4.73$ with $DF = 2$). They also showed that this model is equivalent to the 3-class LC model (H_{2C} in Table 13), with

⁷ As an extra check we estimated the models of interest using the assumed (constructed) population distributions as observed data. For models that are identified, the parameter estimates should perfectly reproduce the population parameters. This result was obtained in all situations.

the same DF if we take into account the fact that there are 6 unidentified parameters⁸ (see Table 11). From the test results reported in Table 13, it can be seen that the basic 2-factor model (H_{2F}) is saturated for these data ($DF=0$), and hence provides a perfect fit ($L^2 = 0$).

[INSERT FIGURE 10 ABOUT HERE]

[INSERT FIGURE 11 ABOUT HERE]

The tri-plot and bi-plot obtained from the 3-class LC model and the basic 2-factor LC model are not unique since the posterior classification (membership) probabilities are dependent upon the particular identifying restrictions used to identify the parameters (4 distinct restrictions are needed for the basic 2-factor model). However, the specification of restrictions is typical of a confirmatory rather than exploratory analysis. Rather than specifying restrictions (or using a particular set of boundary or other nonunique parameter estimates) to obtain a unique solution, one can alternatively apply some prior information to the parameters. Table 13 provides the results of fitting the LC cluster and factor models (H_{2C+} and H_{2F+}), and Figures 10 and 11 present the associated displays that are when a slight departure from non-informative Dirichlet prior distributions are assumed for the model probabilities.⁹

From the bi-plot (Figure 11) we see that factor 1, the more prominent factor, is a “lightness-darkness” dimension. Factor 2 serves primarily as a contrast of black hair and dark eyes with medium and red hair color and lighter eye colors, (with fair and dark hair and blue eyes somewhere in between).

6. FINAL REMARKS

⁸ We assume that 6 identifying restrictions are made to identify these parameters. These restrictions need not be the same as those used to identify the 2-component canonical model.

⁹ The influence of the priors is equivalent to adding one fictitious observation for which the independence model holds to the data. As a result, the priors will smooth the estimates slightly to the independence model. For more details on the use of priors to prevent boundary solution and to obtain identifiability, see Vermunt and Magidson (2000).

This paper presented a new method for performing exploratory LC analysis. Rather than increasing the number of classes, we proposed increasing the number of latent variables. We showed that because of the imposed constraints, the basic LC factor model with R latent variables has the same number of parameters as the LC cluster model with $R+1$ classes. This is an important result because it shows that in terms of parsimony, increasing the number of factors is equivalent to increasing the number of clusters.

The examples showed that in most cases the LC factor model provides a more parsimonious and easier to interpret description of the data. There is a simple explanation for this phenomenon. When applying a LC cluster model it is not known how many dimensions the solution will capture: A 3-cluster model may describe either one or two dimensions, while a 4-cluster model may describe either one, two, or three dimensions. When a 3-cluster model describes *one* dimension, it is very probable that a 1-factor model with 3 or more levels will describe the data almost as well (see the depression example). When a 3-cluster model describes *two* dimensions, it has the disadvantage that it can not capture all four basic combinations – (low, low), (high, low), (low, high) and (high, high) – of the two latent dimensions. Therefore, the 2- factor model will fit better than the 3-cluster model in these cases. In situations in which the 4-cluster model gives a 2-dimensional solution (as in the rheumatic arthritis data set where the 4 clusters represent the 4 possible latent combinations), it can be expected that a restricted 4-cluster model (the 2-factor model) will fit the data almost as well (and may be better in terms of BIC or p-value).

The above explanation yields strong arguments for using the two approaches in combination with one another, as we have been doing in the examples. There are two things that can happen when switching from the cluster to the factor model. First, the factor model may give a simpler description of the data than the cluster model. This occurs when the 3-cluster solution is one dimensional or when the 4-cluster solution is two dimensional, both of which are situations where the LC cluster model is overparametrized. Second, the factor model may give a better fit. We saw that this occurs when the three-cluster model is two-dimensional.

APPENDIX A: The LC Cluster and Factor Models Formulated
Using Conditional Probabilities

In this paper we used Haberman's (1979) log-linear formulation of the LC model because that made it easy to explain the similarities and differences between LC cluster and unrestricted LC factor models. However, in the case of the restricted 2-factor model, a more general formulation is required. This appendix describes these two types of LC models using the more general probability formulation, and explains the relationship between the two formulations.

An alternative expression for the LC cluster model described in equation (1) is

$$\pi_{ijklt} = \pi_t^X \pi_{it}^{A|X} \pi_{jt}^{B|X} \pi_{kt}^{C|X} \pi_{lt}^{D|X},$$

which is the formulation used by Lazarsfeld and Henry (1968), Goodman (1974a, 1974b) and Clogg (1981, 1995). As was shown by several authors (see, for instance, Haberman 1979; Formann 1992; and Heinen 1996), there is a simple relationship between the conditional response probabilities appearing in the above equation and the log-linear parameters of equation (1), i.e.,

$$\pi_{it}^{A|X} = \frac{F_{i++++t}}{F_{++++t}} = \frac{\exp(\lambda_i^A + \lambda_{it}^{AX})}{\sum_{i'=1}^I \exp(\lambda_{i'}^A + \lambda_{i't}^{AX})}.$$

Similar expressions apply to the other three indicators. The probability of being in class t , π_t^X , can, however, not be written in terms of the log-linear parameters λ_t^X appearing in equation (1). These latent probabilities can be obtained by

$$\pi_t^X = \frac{F_{++++t}}{F_{++++}} = \frac{\exp(\gamma_t^X)}{\sum_{t'=1}^T \exp(\gamma_{t'}^X)},$$

where the symbol γ is used to denote a log-linear parameter of the marginal distribution of the latent variable(s).

The 2-factor LC model can be written as

$$\pi_{ijklrs} = \pi_{rs}^{VW} \pi_{ijklrs}^{ABCD|VW} = \pi_{rs}^{VW} \pi_{irs}^{A|VW} \pi_{jrs}^{B|VW} \pi_{krs}^{C|VW} \pi_{lrs}^{D|VW} \quad (4)$$

whereas, in the case of the *unrestricted* model we have

$$\pi_{rs}^{VW} = \frac{F_{++++rs}}{F_{++++++}} = \frac{\exp(\gamma_r^V + \gamma_s^W + \gamma_{rs}^{VW})}{\sum_{r'=1}^R \sum_{s'=1}^S \exp(\gamma_{r'}^V + \gamma_{s'}^W + \gamma_{r's'}^{VW})}$$

$$\pi_{rst}^{A|VW} = \frac{F_{i+++rs}}{F_{++++rs}} = \frac{\exp(\lambda_i^A + \lambda_{ir}^{AV} + \lambda_{is}^{AW} + \lambda_{irs}^{AVW})}{\sum_{i'=1}^I \exp(\lambda_{i'}^A + \lambda_{i'r}^{AV} + \lambda_{i's}^{AW} + \lambda_{i'rs}^{AVW})}, \text{ etc.,}$$

while, for the *basic* 2-factor model, the conditional response probabilities in (4) are restricted by the following logit models

$$\pi_{rs}^{VW} = \frac{\exp(\gamma_r^V + \gamma_s^W)}{\sum_{r'=1}^R \sum_{s'=1}^S \exp(\gamma_{r'}^V + \gamma_{s'}^W)}$$

$$\pi_{irs}^{A|VW} = \frac{\exp(\lambda_i^A + \lambda_{ir}^{AV} + \lambda_{is}^{AW})}{\sum_{i'=1}^I \exp(\lambda_{i'}^A + \lambda_{i'r}^{AV} + \lambda_{i's}^{AW})}, \text{ etc.}$$

Note that this latter formulation excludes the marginal association between the latent variables, as well as the higher-order interaction terms.

The number of distinct parameters in the basic R-factor model is:

$$\text{NPAR}(\text{basic R-factor}) = R + \text{NPAR}(\text{indep}) \times (1 + R)$$

$$= R + (R+1) \times \text{NPAR}(\text{indep}),$$

while the number of distinct parameters in the LC cluster model was shown to be

$$\begin{aligned} \text{NPAR}(\text{T-cluster}) &= (\text{T}-1) + \text{NPAR}(\text{indep}) (1 + (\text{T}-1)) \\ &= (\text{T}-1) + \text{T} \times \text{NPAR}(\text{indep}). \end{aligned}$$

Hence, it is seen that the degree of parsimony in the LC R-factor model is the same as that of a cluster model with $T = R+1$ classes.

As shown in this paper, the *unrestricted* LC 2-factor model is equivalent to the LC cluster model with 4 classes. Hence, the number of restrictions that are placed by the basic 2-factor model given above can be computed as the difference between the number of distinct parameters in the LC cluster model with $T = 4$ classes and the number in the basic LC 2-factor model. More generally, the number of restrictions placed by the R -factor model can be computed as the difference between the number of distinct parameters in the LC cluster model with $T=2^R$ classes and the basic LC R-factor model as follows:

$$\begin{aligned} \text{NRES} &= \text{NPAR}(2^R\text{-cluster}) - \text{NPAR}(\text{basic R-factor}) \\ &= [2^R - R - 1] \times [\text{NPAR}(\text{indep}) + 1]. \end{aligned}$$

APPENDIX B: Functions of Class-membership Probabilities Appearing in the Plots

The quantities depicted in the various plots presented in this paper are functions of class-membership probabilities. This appendix explains how these quantities are computed. For the types of LC models considered by Van der Ark and Van der Heijden (1998) and Van der Heijden, Gilula, and Van der Ark (1999), our measures coincide with the rescaled parameters which they plotted, but for more general LC models this need not be the case.

Let us take the basic two-factor model with four indicators described in equations (3) and (4) as an example. The estimated probability of being in level r of the first factor V given a person's observed scores on the 4 indicators $A, B, C,$ and D is defined as

$$\hat{\pi}_{ijkl}^{V|ABCD} = \frac{\hat{\pi}_{ijklr+}}{\hat{\pi}_{ijkl++}}.$$

Once the latent class model of interest is estimated, these class-membership probabilities can be computed for each individual in the sample or, equivalently, for each observed response pattern.

A common quantity that we use to position each point in each of our plots is the conditional probability of being at a certain level of a latent variable given a certain response to one or more items. In the bi-plot associated with the LC factor model, we will, for instance, use the estimated conditional probability of being at level r of factor V given that $A=i$, denoted as $\hat{\pi}_{ri}^{V|A}$. Note that the more these probabilities differ between levels of A , the stronger A is related to factor V .

The probabilities $\hat{\pi}_{ri}^{V|A}$ can be obtained by aggregating the estimated class-membership probabilities $\hat{\pi}_{ijkl}^{V|ABCD}$. There are, however, two possible ways to perform the aggregation. Method 1 utilizes the *observed* cell probabilities p_{ijkl}^{ABCD} as weights. This yields

$$\hat{\pi}_{ri}^{V|A} (1) = \frac{\sum_{j=1}^J \sum_{k=1}^K \sum_{l=1}^L \hat{\pi}_{rijkl}^{V|ABCD} P_{ijkl}^{ABCD}}{\sum_{r=1}^R \sum_{j=1}^J \sum_{k=1}^K \sum_{l=1}^L \hat{\pi}_{rijkl}^{V|ABCD} P_{ijkl}^{ABCD}} .$$

Alternatively, method 2 utilizes the *estimated* cell probabilities as weights; that is,

$$\hat{\pi}_{ri}^{V|A} (2) = \frac{\sum_{j=1}^J \sum_{k=1}^K \sum_{l=1}^L \hat{\pi}_{rijkl}^{V|ABCD} \hat{\pi}_{ijkl}^{ABCD}}{\sum_{r=1}^R \sum_{j=1}^J \sum_{k=1}^K \sum_{l=1}^L \hat{\pi}_{rijkl}^{V|ABCD} \hat{\pi}_{ijkl}^{ABCD}}$$

Method 1 was used to obtain the plots presented in Figures 1-11. In Figures 4, 6, and 8 we also included the individual response patterns, including those not observed in the sample.

In the case of *unrestricted* models, if the model provides a good fit to the data the estimated proportions should provide good approximations to the observed proportions so that both methods will yield very similar plots. However, for certain *restricted* models, where the estimated proportions satisfy the restrictions exactly but the observed proportions do not, the alternative displays may contain clear discernable differences, even when the model provides a good fit to the data.

For example, the restrictions for model H_{R2F} imply that the basic bi-plot should consist of two intersecting straight lines, one formed by connecting the points corresponding to the categories of the variables (C) PURPOSE and (D) ACCURACY, and the second formed by connecting the points corresponding to the categories of (A) COOPERATION and (B) UNDERSTANDING..

[INSERT FIGURE 12 ABOUT HERE]

Figure 12 shows the resulting bi-plot for model H_{R2F} when method 2 is used to compute the conditional probabilities. We see the two straight lines with an acute angle between them suggesting positive correlation between the latent variables V and W

(labeled Factor 1 and Factor 2 in Figure 12).¹⁰ On the other hand, the plot obtained in Figure 7 showed only the *approximation* of two straight lines since the observed proportions for these data do not satisfy exactly the restrictions imposed by the model.

In LC factor models with factors having more than 2 levels¹¹ such as Model H_{1F3}, the results of which were displayed in Figure 9, we plot the factor means

$$\hat{E}_i^{V|A} = \sum_{r=i}^R \hat{\pi}_{ri}^{V|A} \cdot v_r^V ,$$

where R is the number of levels of factor V, and v_r^V denotes the fixed score assigned to level r of factor V.

In the case of a LC cluster model, we would plot $\hat{\pi}_{ii}^{X|A}$, which is the estimated conditional probability of being in a certain category of the single latent variable X. Van der Ark and Van der Heijden (1998), who called these quantities *rescaled parameters*, proposed computing them as follows:

$$\hat{\pi}_{ii}^{X|A} (3) = \frac{\hat{\pi}_{it}^{AX}}{\hat{\pi}_i^A} = \frac{\hat{\pi}_t^X \hat{\pi}_{it}^{A|X}}{\sum_{t=1}^T \hat{\pi}_t^X \hat{\pi}_{it}^{A|X}} .$$

It can easily be shown that in a standard LC model with a single latent variable and no restrictions on the model probabilities, all three methods yield the same results; that is,

$$\hat{\pi}_{ii}^{X|A} (1) = \hat{\pi}_{ii}^{X|A} (2) = \hat{\pi}_{ii}^{X|A} (3) .$$

¹⁰ In a companion paper (Magidson and Vermunt, 2000), we show how to derive the equations for the straight lines. Moreover, in it we demonstrate that the angle between these lines has meaning – for example, to the extent to which this angle is less than 90°, the two latent variables V and W exhibit positive correlation – and show how the magnitude of the correlation can be determined from the plot.

¹¹ In the case of dichotomous latent variables, the relationship between the expected value and the conditional probability provides a “true score regression” interpretation of the lines plotted in Figure 12.

The difference between our method and that of Van der Ark and Van der Heijden is that we derive and plot quantities that are defined for each individual in the sample; namely, the probability $\hat{\pi}_{rikl}^{V|ABCD}$. A category-specific marginal conditional probability like $\hat{\pi}_{ri}^{V|A}$ is, therefore, just one of the several types of measures that can be depicted in the same plot. Other possibilities are depicting the location of specific response patterns (as in Figure 4 and Figure 6 of this paper),¹² the marginal probabilities for a subset of observed variables (for instance, $\hat{\pi}_{rij}^{V|AB}$), or the marginal probabilities for categories of variables that are not included in the LC model. We labeled the latter application the inactive-covariate method (Vermunt and Magidson, 2000) since it yields information on the association of a covariate with each of the latent variables without including the covariate concerned in the LC model.¹³

To illustrate the inactive-covariate method assume that there is a variable E whose levels are indexed by m. The probability of being in level r of latent variable V given that E equals m, $\hat{\pi}_{rm}^{V|E}$, is obtained as follows:

$$\hat{\pi}_{rm}^{V|E} (1) = \frac{\sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K \sum_{l=1}^L \hat{\pi}_{rikl}^{V|ABCD} P_{ijklm}^{ABCDE}}{\sum_{r=1}^R \sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K \sum_{l=1}^L \hat{\pi}_{rikl}^{V|ABCD} P_{ijklm}^{ABCDE}}$$

Note that in this case we must use the observed cell probabilities as weights (method 1) because we do not have estimated probabilities for the joint distribution including E.

Another important advantage of our way of computing the plotted measures is that it can easily be extended to variables of other scale types, such as continuous dependent or independent variables. This is something that is used in the new computer program

¹² It should be noted that Van der Heijden, Gilula, and Van der Ark, (1999) already mentioned the possibility of incorporating information on individual cases in their ternary plots. They, however, did not explicitly discuss the relationship between the individual posterior membership probabilities and the rescaled probabilities nor the possibility of collapsing the individual posterior membership probabilities in ways other than to form categories of individual variables.

¹³ In correspondence analysis it is quite common to plot levels of inactive covariates. There they are called passive variables.

LatentGOLD[®] (Vermunt and Magidson, 2000), which implements the uni-, bi-, and tri-plots discussed in this paper.

APPENDIX C: Estimation of the LC Cluster and LC Factor Models

The standard estimation method for LC models is the Maximum Likelihood (ML) method under the assumption that the data come from a multinomial distribution. ML estimation of the model parameters of the LC Factor model described in equation (4) involves finding the parameters values that maximize the following likelihood function:

$$L \propto \prod_{ijkl} \left(\sum_{rs} \pi_r^V \pi_s^W \pi_{irs}^{A|VW} \pi_{jrs}^{B|VW} \pi_{krs}^{C|VW} \pi_{lrs}^{D|VW} \right)^{Np_{ijkl}},$$

where N denotes the sample size and p_{ijkl}^{ABCD} the proportion of the sample belonging to the cell entry concerned. Maximization of the likelihood is a quite standard task that can be performed with an EM or a Newton-Raphson algorithm, or some combination of the two. Software packages that can be used to obtain ML estimates of the parameters of LC factor models are Newton (Haberman 1988), LEM (Vermunt 1997), and LatentGOLD (Vermunt and Magidson 2000).

REFERENCES

- Clogg, C.C. 1981. "New developments in latent structure analysis." Pp. 215-246 *Factor analysis and measurement in sociological research*, edited by D.J. Jackson and E.F. Borgotta. Beverly Hills: Sage Publications.
- Clogg, C.C. 1988. "Latent class models for measuring." In *Latent trait and latent class models*, edited by R. Langeheine and J. Rost, New York, London: Plenum Press.
- Clogg, C.C. 1995. "Latent class models." Pp. 311-359 in *Handbook of statistical modeling for the social and behavioral sciences*, edited by G. Arminger, C.C. Clogg, and M.E. Sobel. New York: Plenum Press.
- De Leeuw, J., and P.G.M. Van der Heijden. 1991. "Reduced rank models for contingency tables." *Biometrika* 78:229-232.
- Fisher, R.A. 1940. "The precision of discriminant functions," *Annals of Eugenics*, London, 10: 422-429.
- Formann, A.K. 1992. "Linear logistic latent class analysis for polytomous data." *Journal of the American Statistical Association* 87:476-486.
- Fraley, C., and Raftery, A.E. 1998. *How many clusters? Which clustering method? - Answers via model-based cluster analysis*. Department of Statistics, University of Washington: Technical Report no. 329.
- Gilula, Z., and S.J. Haberman. 1986. "Canonical analysis of contingency tables by maximum likelihood." *Journal of the American Statistical Association* 81:780-788.
- Goodman, L.A. 1974a. "Exploratory latent structure analysis using both identifiable and unidentifiable models." *Biometrika* 61: 215-231.
- Goodman, L.A. 1974b. "The Analysis of systems of qualitative Variables when some of the Variables are Unobservable. Part I: A Modified Latent Structure Approach", *American Journal of Sociology* 79: 1179-1259.
- Gower, J.C., and D.J. Hand. 1996. *Biplots*. London: Chapman & Hall.
- Greenacre, M.J. 1993, *Correspondence analysis*. London: Academic Press.
- Haberman, S.J. 1979. *Analysis of qualitative data, Vol 2, New developments*. New York: Academic Press.

Haberman, S.J. 1988. "A stabilized Newton-Raphson algorithm for log-linear models for frequency tables derived by indirect observations", Pp. 193-211 in *Sociological Methodology 1988*, edited by C. Clogg. Washington DC: American Sociological Association.

Hagenaars, J.A. 1990. *Categorical longitudinal data – loglinear analysis of panel, trend and cohort data*. Newbury Park: Sage.

Hagenaars, J.A. 1993. *Loglinear models with latent variables*. Newbury Park: Sage.

Hunt, L, and M. Jorgensen. 1999. "Mixture model clustering using the MULTIMIX program." *Australian and New Zealand Journal of Statistics* 41:153-172.

Kohlmann, T. and A.K. Formann. 1997. "Using Latent Class Models to Analyze Response Patterns in Epidemiologic Mail Surveys", Chapter 33 in *Applications of Latent Trait and Latent Class Models in the Social Sciences*, edited by J. Rost and R. Langeheine. New York: Waxmann.

Lawrence C.J., W.J. Krzanowski. 1996. "Mixture separation for mixed-mode data. " *Statistics and Computing* 6:85-92.

Lazarsfeld, P.F., and N.W. Henry. 1968. *Latent structure analysis*. Boston: Houghton Mill.

Magidson, J. and Vermunt, J.K. and 2000. "Bi-plots and Related Graphical Displays based on Latent Class Factor and Cluster Models", Proceedings of the RC33 Conference, University of Cologne, Cologne Germany.

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

McLachlan, G.J., and K.E. Basford. 1988. *Mixture models: inference and application to clustering*. New York: Marcel Dekker.

Moustaki, I. 1996. "A latent trait and a latent class model for mixed observed variables. " *The British Journal of Mathematical and Statistical Psychology* 49:313-334.

Pearlin, L.I. and J.S. Johnson. 1977. "Marital status, life-strains, and depression." *American Sociological Review* 42:104-115.

Schaeffer, N.C. 1988. "An application of item response theory to the measurement of depression", Pp. 271-308 in *Sociological Methodology 1988*, edited by C. Clogg. Washington DC: American Sociological Association.

Van der Ark, L.A. and P.G.M. Van der Heijden. 1998. "Graphical display of latent budget and latent class analysis." Pp. 489-509 in *Visualization of categorical data*, edited by J. Blasius and M. Greenacre. Boston: Academic Press.

Van der Ark, L.A. Van der Heijden, P.G.M., and D. Sikkel. 1999. "On the identifiability in the latent budget model." *Journal of Classification* 16:117-137.

Van der Heijden, P.G.M. Gilula, Z. and L.A. Van der Ark. 1999 "On a Relation Between Joint Correspondence Analysis and Latent Class Analysis." Pp. 147-186 in *Sociological Methodology 1999*, edited by M. Sobel and M. Becker.

Vermunt, J.K. 1997. *LEM: A general program for the analysis of categorical data. User's manual*. Tilburg University, The Netherlands.

Vermunt, J.K. and J. Magidson. 2000. *Latent GOLD 2.0 User's Guide*. Belmont, MA: Statistical Innovations Inc.

Vermunt, J.K. and J. Magidson. 2001. "Latent Class Cluster Analysis", Chapter 3 in *Applied Latent Class Analysis*. edited by J.A. Hagenaars and A.L. McCutcheon , Cambridge University Press.

Wasmus, A., Kindel, P., Mattussek, S. and H.H, Raspe. 1989. "Activity and severity of rheumatoid arthritis in Hannover/FRG and in one regional referral center." *Scandinavian Journal of Rheumatology*, Suppl. 79: 33-44.

Wolfe, J.H. 1970. "Pattern clustering by multivariate cluster analysis." *Multivariate Behavioral Research* 5:329-350.

TABLE 1
 Equivalency Relationship between LC Cluster and Basic LC Factor Models
 (Example with 5 Dichotomous Variables)

LC Cluster Models			Basic LC Factor Models		
Number of Latent Classes	Number of Parameters	Degrees of Freedom	Number of Factors	Number of Parameters	Degrees of Freedom
1	5	26	0	5	26
2	11	20	1	11	20
3	17	14	2	17	14
4	23	8	3	23	8
5	29	2	4	29	2

TABLE 2
 Cross-tabulation of Observed Variables for White Respondents
 to the 1982 General Social Survey

(C) PURPOSE	(D) ACCURACY	(B) UNDERSTANDING	(A) COOPERATION		
			Interested	Cooperative	Impatient/ Hostile
Good	Mostly true	Good	419	35	2
		Fair, poor	71	25	5
	Not true	Good	270	25	4
		Fair, poor	42	16	5
Depends	Mostly true	Good	23	4	1
		Fair, poor	6	2	0
	Not true	Good	43	9	2
		Fair, poor	9	3	2
Waste	Mostly true	Good	26	3	0
		Fair, poor	1	2	0
	Not true	Good	85	23	6
		Fair, poor	13	12	8

TABLE 3: Results from Various LC Models Fit to Data in Table 2

Model	Model Description	BIC	L²	DF	p-value	% Reduction in L²(H₀)
H ₀	1-class	51.6	257.26	29	2.0x10 ⁻³⁸	0 %
H ₁	2-class	-76.7	79.34	22	2.1x10 ⁻⁸	69.1%
H _{2C}	3-class	-98.7	21.89	15+2 [†]	0.19	91.5%
H _{2F}	basic 2-factor	-109.6	10.93	15+2 [†]	0.86	95.7%
H ₃	4-class	-72.0	6.04	8+3 [†]	0.87	97.7%
H _{R2F}	restricted 2-factor	-140.9	22.17	22+1 [†]	0.51	91.4%
H _{1F3}	1-factor (3 levels)	-71.7	77.25	21	2.3x10 ⁻⁸	70.0%

† DF is increased by these boundary solutions

TABLE 4

**Comparison of results from the 3-Cluster Model with the Basic 2-Factor Model
Conditional Membership Probability of being in Cluster $j=1,2,3$ (for Model H_{2C})
or level 1 of Factor $k=1,2$ (for Model H_{2F})**

Indicators	Model H_{2C}			Model H_{2F}	
	Cluster 1	Cluster 2	Cluster 3	Factor1(1)	Factor2(1)
PURPOSE					
Good	0.72	0.25	0.03	0.83	0.71
Depends	0.38	0.17	0.45	0.65	0.28
Waste	0.24	0.02	0.73	0.59	0 †
ACCURACY					
Mostly True	0.73	0.26	0.01	0.83	0.83
Not True	0.50	0.15	0.35	0.71	0.28
UNDERSTAND					
good	0.76	0.08	0.16	0.89	0.53
Fair, poor	0 †	0.77	0.23	0.28	0.71
COOPERATE					
Interested	0.70	0.17	0.13	0.86	0.58
Cooperative	0.27	0.40	0.33	0.38	0.51
Impatient/ Hostile	0 †	0.39	0.61	0 †	0.35
Overall Probability	0.62	0.21	0.17	0.78	0.57

† indicates a boundary solution

Table 5

Rheumatoid Arthritis Mail Survey Data

BACK	NECK	JOINT	SWELL	STIFF	Frequency
no	no	no	no	no	3,634
no	no	no	no	yes	73
no	no	no	yes	no	87
no	no	no	yes	yes	10
no	no	yes	no	no	440
no	no	yes	no	yes	89
no	no	yes	yes	no	106
no	no	yes	yes	yes	75
no	yes	no	no	no	295
no	yes	no	no	yes	25
no	yes	no	yes	no	15
no	yes	no	yes	yes	5
no	yes	yes	no	no	137
no	yes	yes	no	yes	42
no	yes	yes	yes	no	35
no	yes	yes	yes	yes	39
yes	no	no	no	no	489
yes	no	no	no	yes	37
yes	no	no	yes	no	23
yes	no	no	yes	yes	7
yes	no	yes	no	no	255
yes	no	yes	no	yes	116
yes	no	yes	yes	no	71
yes	no	yes	yes	yes	65
yes	yes	no	No	no	306
yes	yes	no	No	yes	48
yes	yes	no	Yes	no	16
yes	yes	no	Yes	yes	11
yes	yes	yes	No	no	229
yes	yes	yes	No	yes	162
yes	yes	yes	Yes	no	44
yes	yes	yes	Yes	yes	176
Total					<hr/> 7,162

TABLE 6: Results from Various LC Models Fit to Data in Table 5

H_m	Model Description	BIC	L²	DF	p-value	% Reduction in L²(H₀)
H ₀	1-class	4592.8	4823.6	26	3.0x10 ⁻¹⁰¹	0%
H ₁	2-class	376.6	554.2	20	1.3x10 ⁻¹⁰⁴	88.5%
H _{2C}	3-class	38.2	162.4	14	2.3x10 ⁻²⁷	96.6%
H _{2F}	basic 2-factor	-110.5	13.7	14	0.5	99.7%
H _{3C}	4-class	-62.6	8.4	8	0.4	99.8%
H _{3F}	basic 3-factor	-85.1	3.7	8+2 [†]	1.0	99.9%

† DF is increased by these boundary solutions

TABLE 7
Comparison of Results obtained under Models H_{2C} and H_{3C}
Conditional Membership Probabilities

Variables	3-Class Solution (H _{2C})			4-Class Solution (H _{3C})			
	Class 1	Class 2	Class 3	Class 1	Class 2	Class 3	Class 4
BACK							
No	0.94	0.32	0.37	0.93	0.31	0.60	0.09
Yes	0.06	0.68	0.63	0.07	0.69	0.40	0.91
NECK							
No	0.96	0.48	0.50	0.96	0.44	0.77	0.15
Yes	0.04	0.52	0.50	0.04	0.56	0.23	0.85
JOINT							
No	0.91	0.63	0.07	0.93	0.60	0.10	0.05
Yes	0.09	0.37	0.93	0.07	0.40	0.90	0.95
SWELL							
No	0.97	0.96	0.49	0.98	0.96	0.55	0.44
Yes	0.03	0.04	0.51	0.02	0.04	0.45	0.56
STIFF							
No	0.98	0.89	0.39	0.99	0.88	0.58	0.08
Yes	0.02	0.11	0.61	0.01	0.12	0.42	0.92
Overall Probabilities	0.62	0.21	0.17	0.61	0.21	0.12	0.06

TABLE 8
Comparison between Models H_{2C}, H_{3C}, and H_{2F}
Observed vs. Expected Frequencies for 4 Response Patterns

Response Pattern						Observed	Frequency Counts		
	Back	Neck	Joint	Swell	Stiff		Expected		
							H _{2C}	H _{3C}	H _{2F}
1	No	No	No	No	No	3,634	3,621.4	3,633.8	3,630.2
2	Yes	Yes	No	No	No	306	304.5	304.8	307.6
3	No	No	Yes	Yes	Yes	75	65.4	70.8	73.0
4	Yes	Yes	Yes	Yes	Yes	176	112.0*	173.7	174.9

*Significantly different from observed

TABLE 9
Results from Various LC Models Fit to the Depression Data

Model	Model Description	BIC	L²	DF	p-value	% Reduction in L²(H₀)
H ₀	1-class	672.8	1097.1	57	2.3x10 ⁻¹⁹²	0
H ₁	2-class	-233.7	138.5	50	3.1x10 ⁻¹⁰	87.4%
H _{2C}	3-class	-260.5	59.6	43	0.05	94.6%
H _{2F}	basic 2-factor	-274.6	45.5	43+1†	0.37	95.9%
H _{1F3}	1-factor (3-levels)	-297.8	67.0	49	0.05	93.9%

† df is increased by these boundary solutions

Table 10

**Conditional Probabilities Estimated under the 3-Cluster model
and the 1-Factor 3-level model**

	3-Cluster Model			1-Factor 3-level Model		
	Cluster1	Cluster2	Cluster3	Level1	Level2	Level3
Cluster Size	0.46	0.44	0.10	0.45	0.45	0.10
ENTHUS						
Lack of enthusiasm	0.26	0.82	0.96	0.26	0.81	0.98
No	0.74	0.18	0.04	0.74	0.19	0.02
ENERGY						
Low energy	0.03	0.63	0.95	0.03	0.61	0.99
No	0.97	0.37	0.05	0.97	0.39	0.01
SLEEP						
sleeping problem	0.10	0.37	0.78	0.09	0.38	0.79
No	0.90	0.63	0.22	0.91	0.62	0.21
APPETITE						
poor appetite	0.04	0.22	0.73	0.04	0.24	0.72
No	0.96	0.78	0.27	0.96	0.76	0.28
HOPELESS						
hopeless	0.03	0.10	0.67	0.02	0.13	0.61
no	0.97	0.90	0.33	0.98	0.87	0.39

TABLE 11: Number of unidentified parameters in various LC cluster and factor models

Model	2x2x2x2 table	2x2x2x2x2 table	4x5 table
2 clusters/1 factor	0	0	2
3 clusters	1	0	6
4 clusters	†	0	†
5 clusters	†	0	†
2 factors	0	0	4
3 factors	†	0	†
4 factors	†	0	†

†: situations that we did not consider because they are not very relevant

TABLE 12
Fisher (1940) Data

EYE COLOR	HAIR COLOR				
	fair	red	medium	dark	black
blue	326	38	241	110	3
light	688	116	584	188	4
medium	343	84	909	412	26
dark	98	48	403	681	85

TABLE 13: Results from Various LC Models Fit to Fisher Data

Model	Model Description	L²	DF[†]	p-value	% Reduction in L²(H₀)
H ₀	1-class	1218.31	12	2.0x10 ⁻²⁵³	0
H ₁	2-class	166.91	6	4.8x10 ⁻³⁵	86.3%
H _{2C}	3-class	4.73	2	.094	99.6%
H _{2F}	basic 2-factor	0.00	0		100.0%
H _{2C+}	3-class (alpha=1)	4.73	2	.094	99.6%
H _{2F+}	basic 2-factor (alpha=1)	0.35	0		100.0%

† DF is increased by the number of unidentified parameters (see TABLE 11)

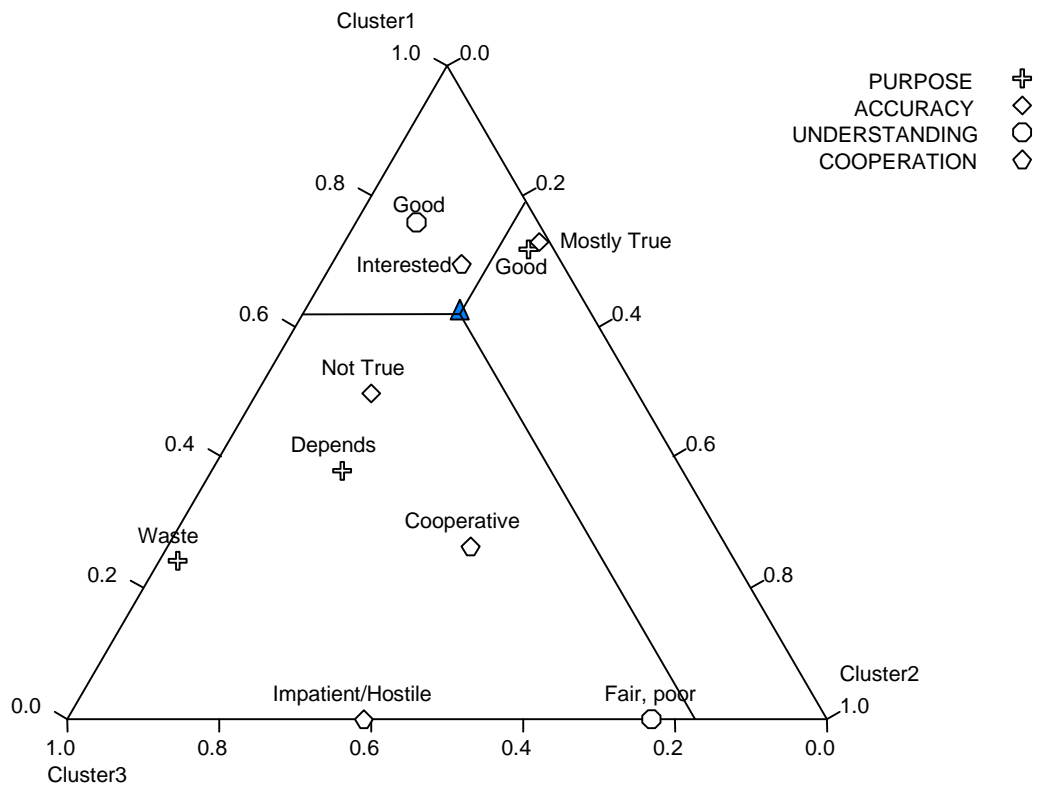


FIGURE 1. Tri-plot of Results Reported in Table 4 for Model H_{2c}

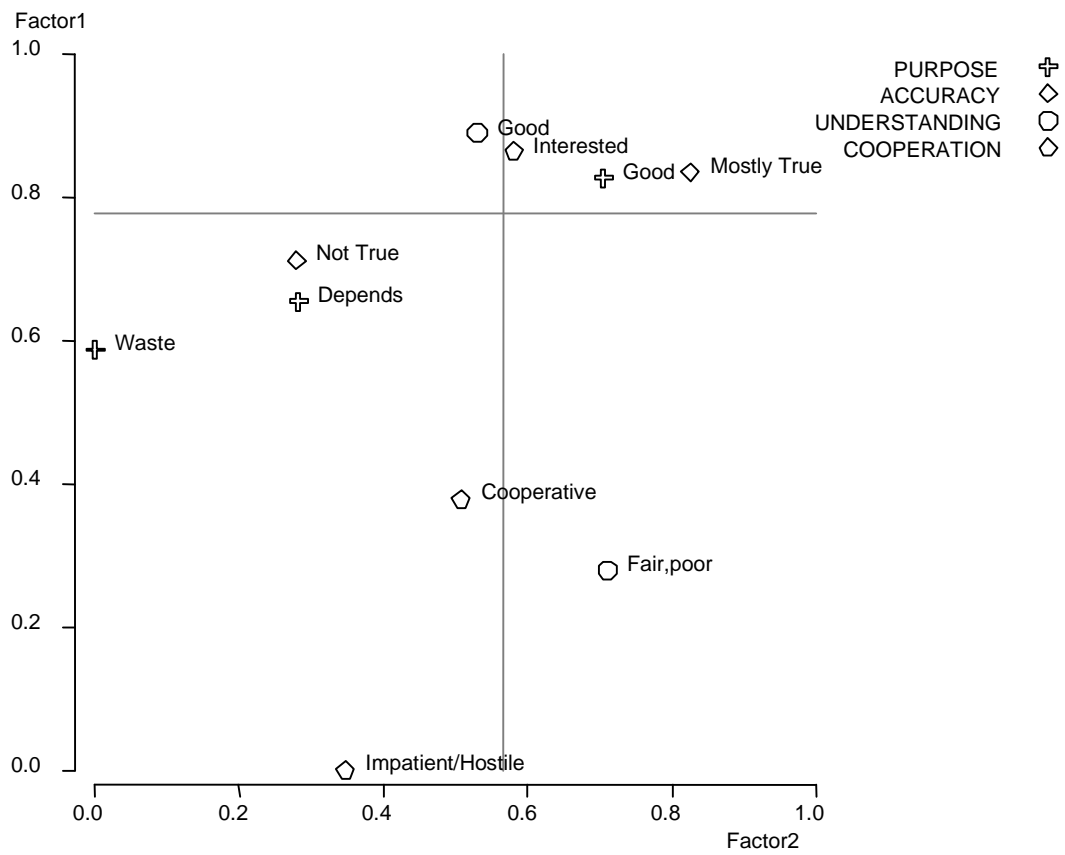


FIGURE 2. Bi-plot of Results Reported in Table 4 for Model H_{2F}

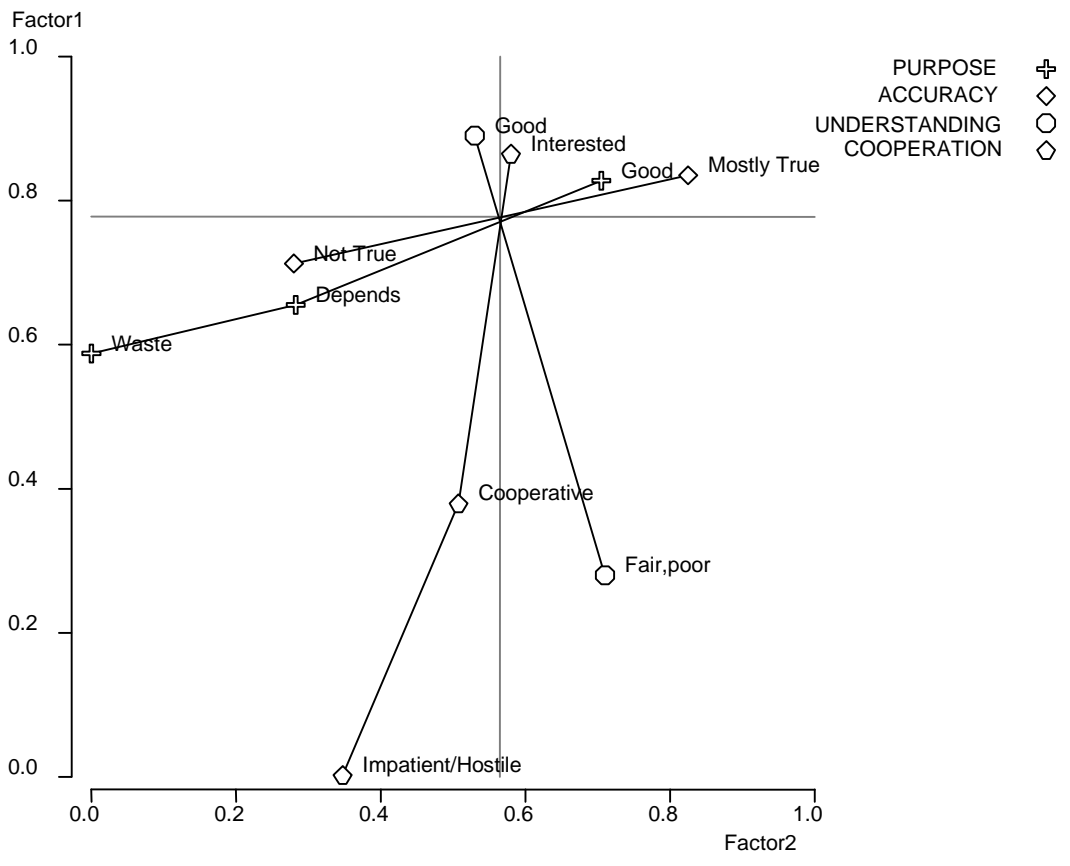


FIGURE 3. Bi-plot for Model H_{2F} with Lines connecting categories of a Variable

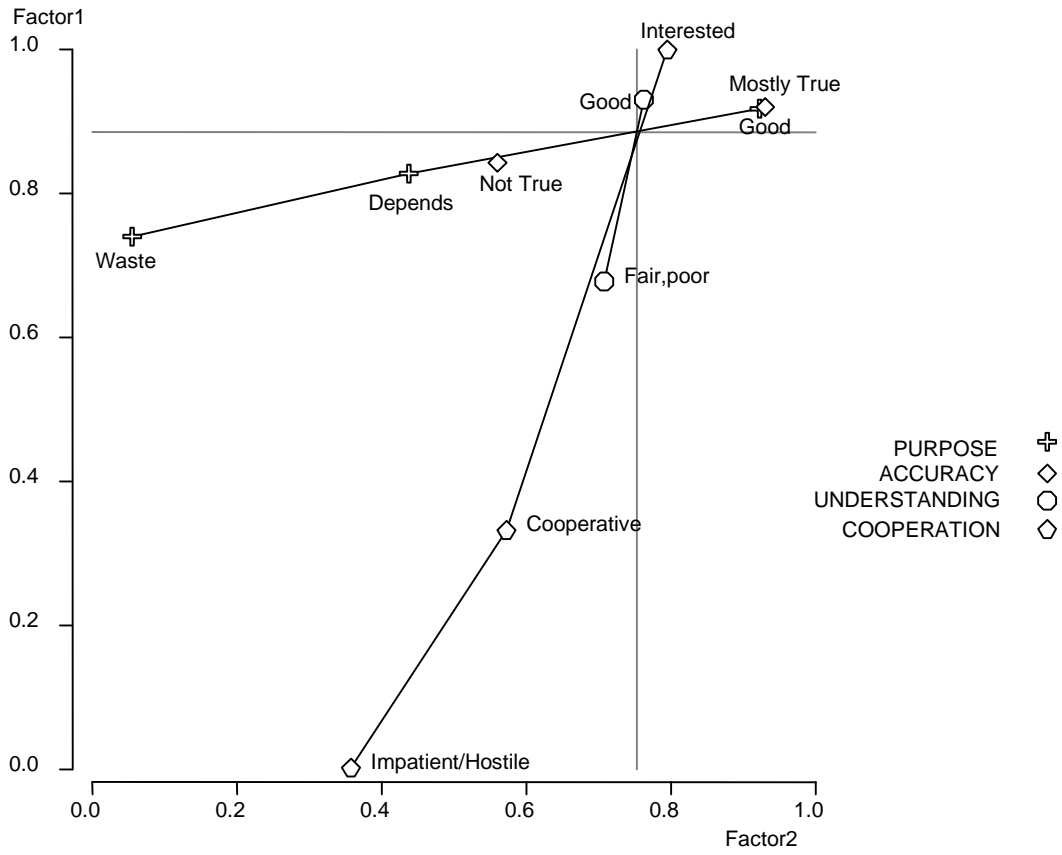


FIGURE 7. Bi-plot for Model H_{R2F} with Lines connecting the categories of a Variable

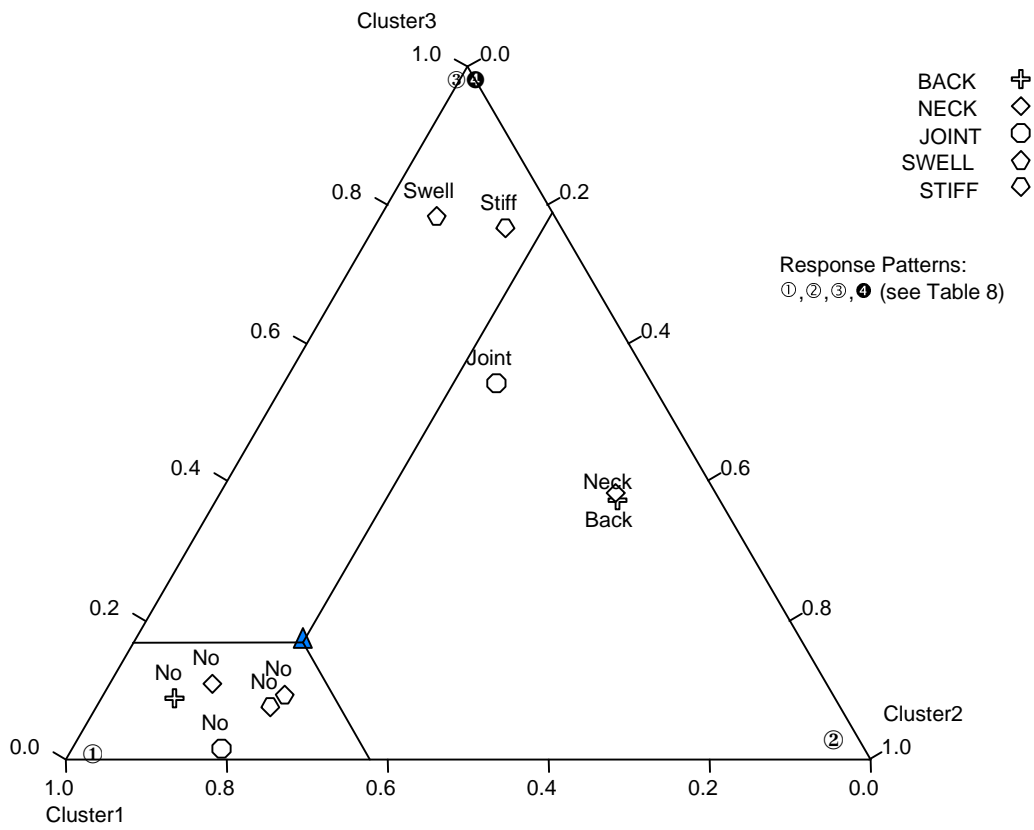


FIGURE 4. Tri-plot for Model H_{2C} and 4 Selected Response Patterns

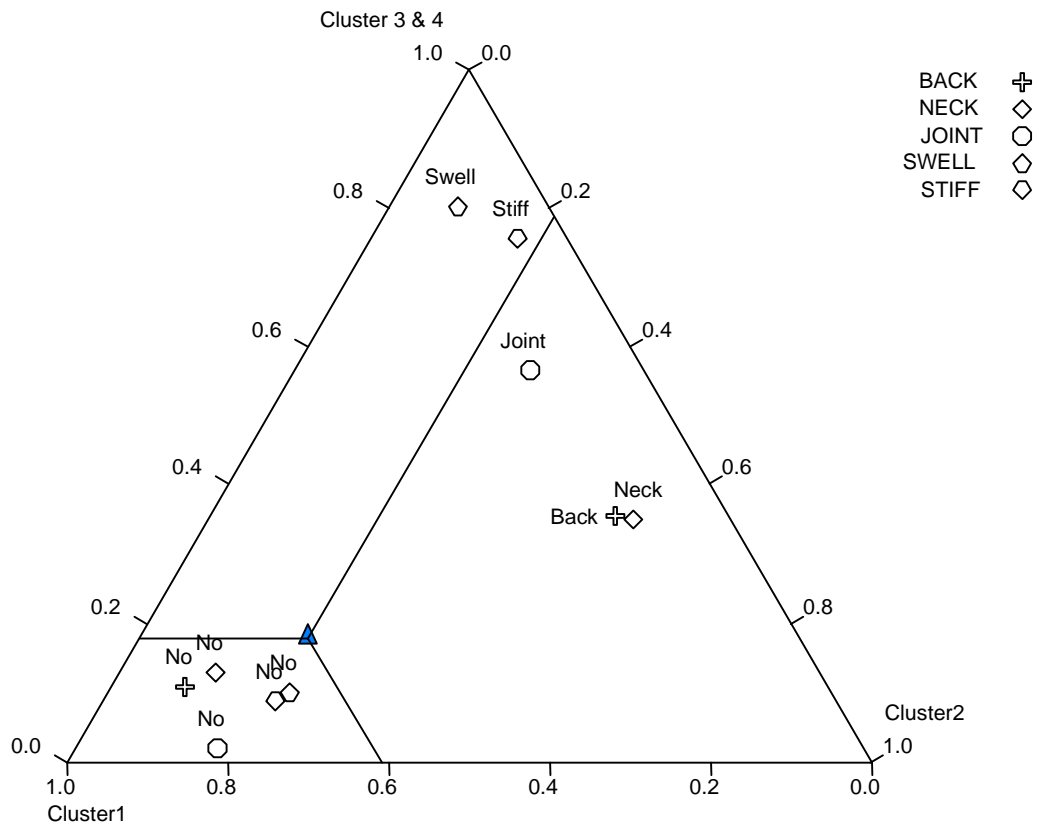


FIGURE 5. Tri-plot for Model H_{3C} where Clusters 3 and 4 are Combined

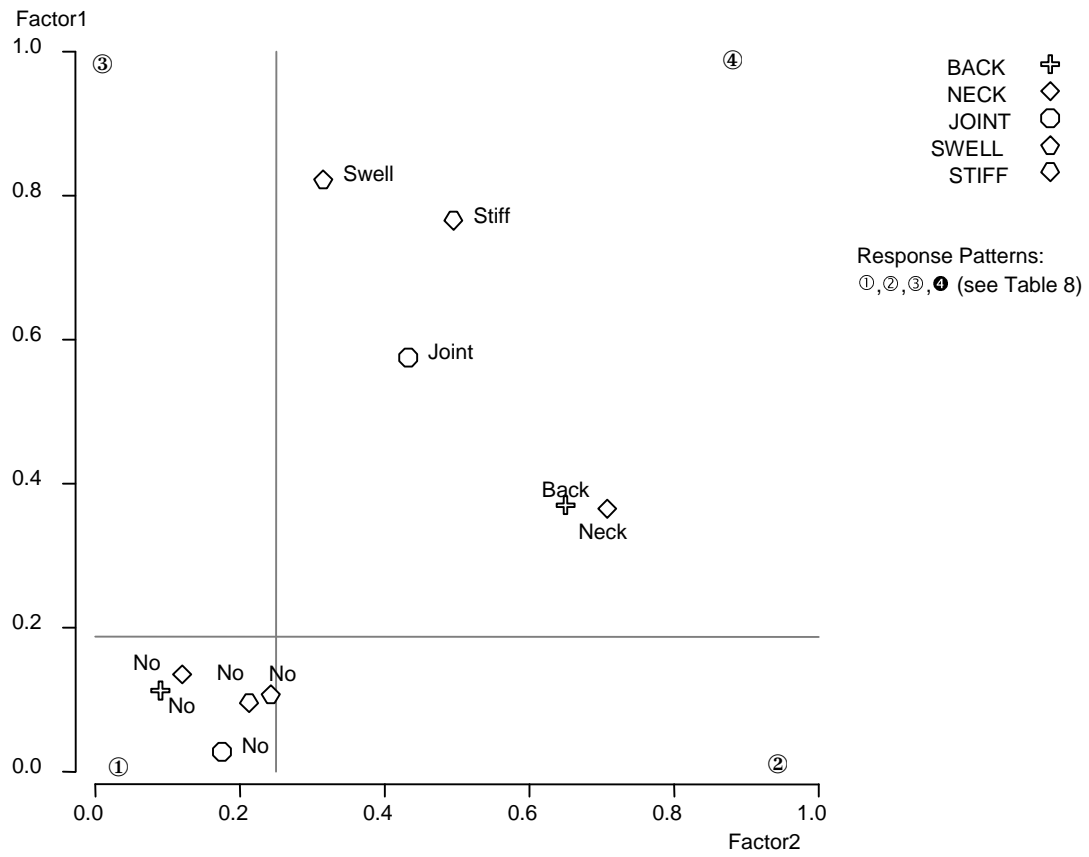


FIGURE 6. Bi-plot for Model H_{2F} and 4 Selected Response Patterns.

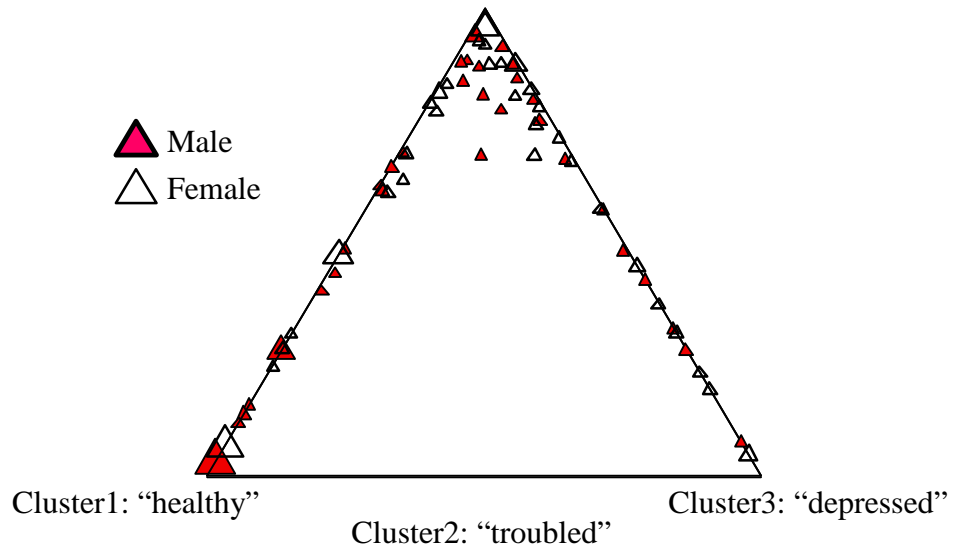


FIGURE 8: Tri-plot of the 64 Response Patterns for Males and Females based on the 3-class Model (H_{2c}).

Note: The area of each triangle is proportional to the estimated expected frequency associated with the corresponding response pattern (subject to a minimum size).

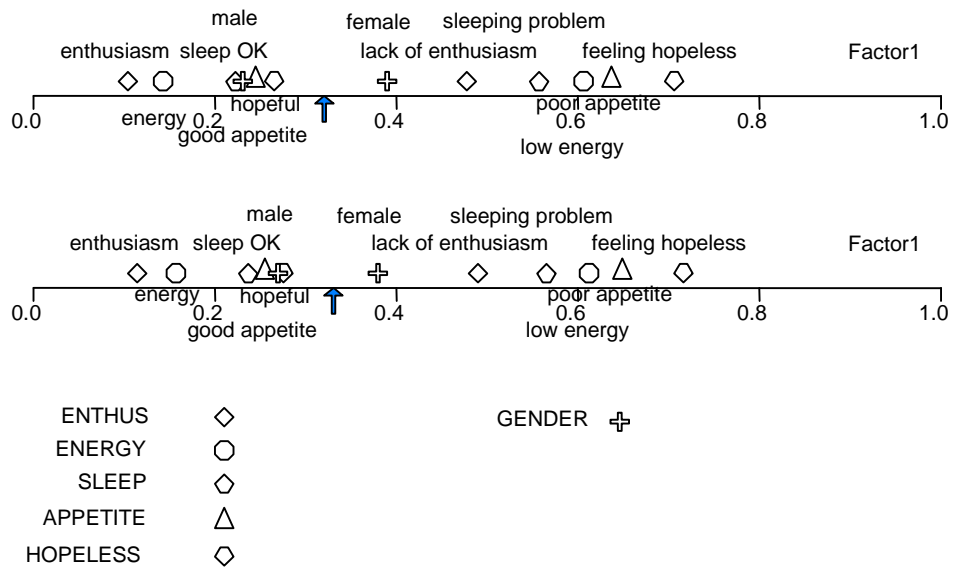


FIGURE 9: Uni-plots associated with the 3-level Factor Model:

Note: GENDER is treated as 'Active' in the top plot and 'Inactive' in the bottom plot.

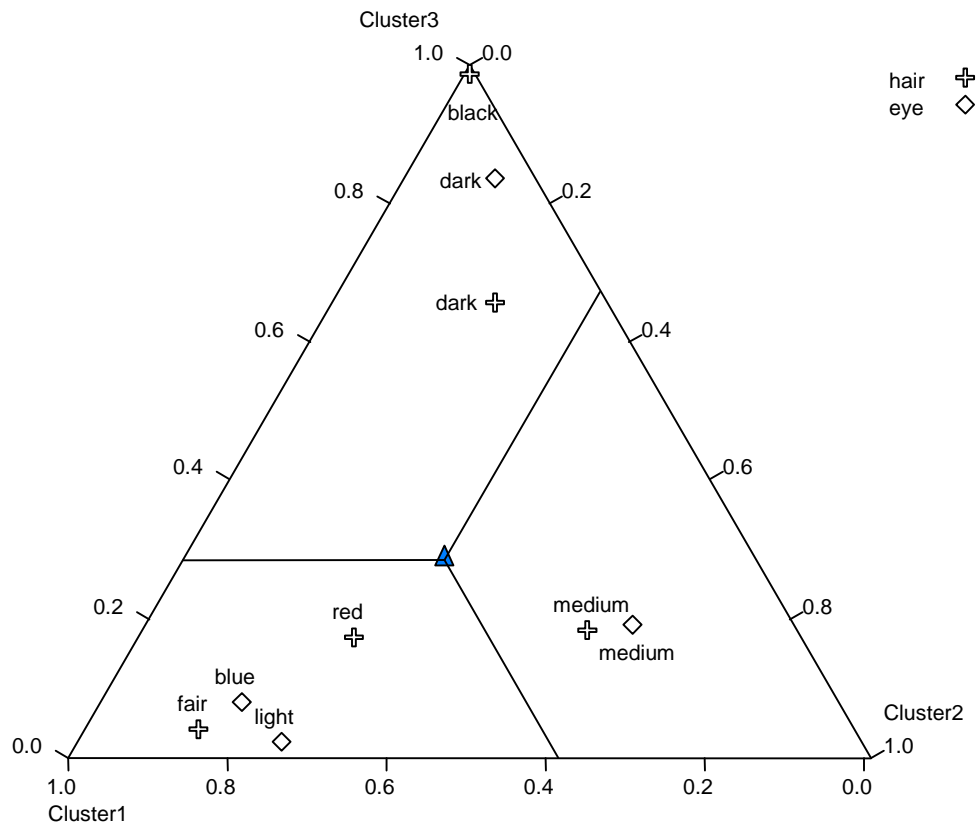


FIGURE 10. Tri-plot of Results from Model $H_{2C}(+)$ Fit to Fisher Data

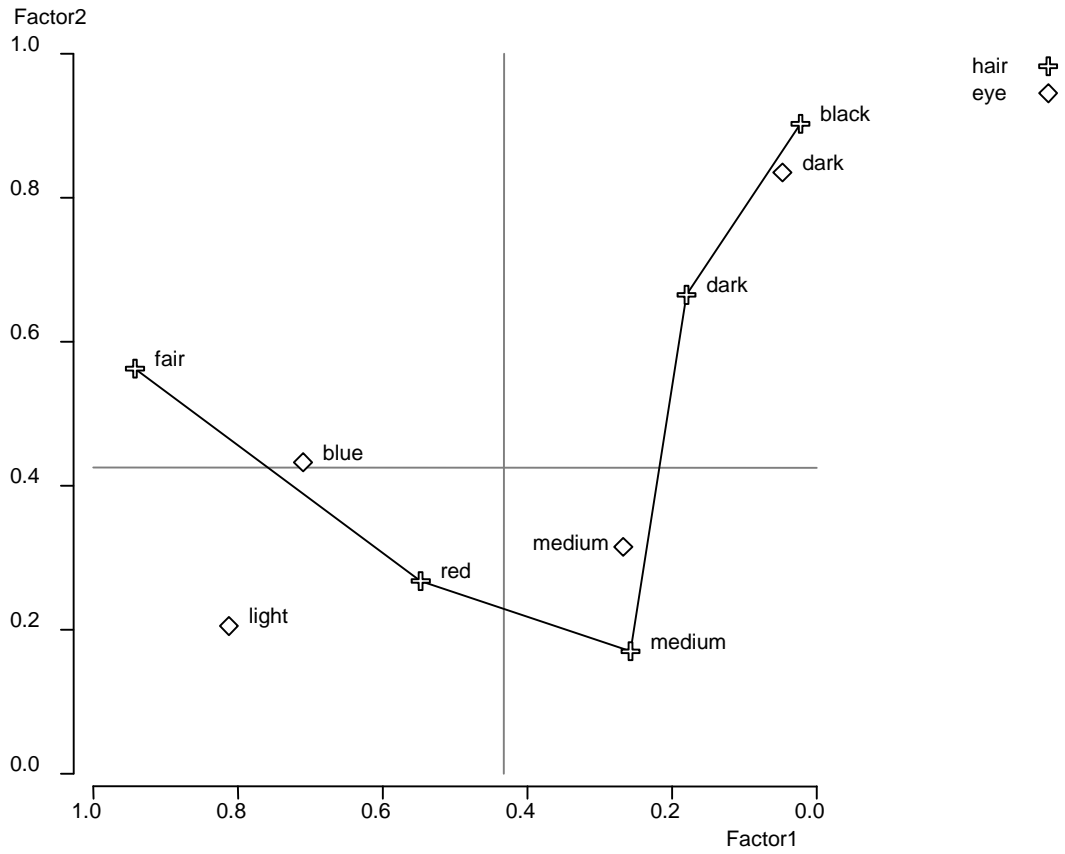


FIGURE 11. Bi-plot of Results from Model $H_{2F(+)}$ Fit to Fisher Data

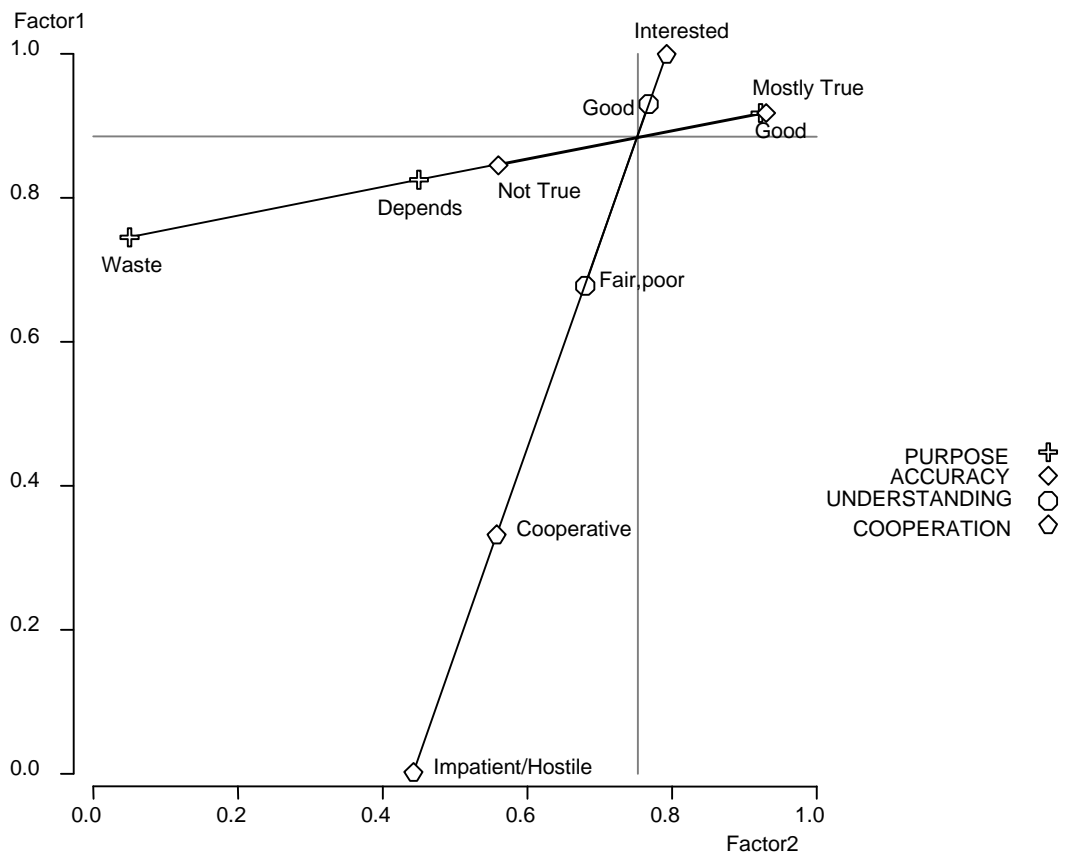


FIGURE 12. Bi-plot for Model H_{R2F} obtained using Aggregation Method 2.