

# Linear Factor Analysis - The Model

$$\mathbf{x} = \boldsymbol{\mu} + \boldsymbol{\Lambda}\mathbf{f} + \boldsymbol{\epsilon} \quad (1)$$

$\mathbf{x}$  :  $p$ -dimensional vector of continuous observed variables

$\mathbf{f}$  :  $q$ -dimensional vector of underlying latent factors. Often called “common factors”. Assume  $\mathbf{f}$  is random such that  $E(\mathbf{f}) = \mathbf{0}$  and  $Var(\mathbf{f}) = \boldsymbol{\Phi}$

$\boldsymbol{\epsilon}$ :  $p$ -dimensional vector of random error. Often called “unique factors” or “specific factors”. Assume  $E(\boldsymbol{\epsilon}) = \mathbf{0}$  and  $Var\boldsymbol{\epsilon} = \boldsymbol{\Psi}$ . Element along the diagonal of  $\boldsymbol{\Psi}$  often called “uniquenesses” or “specific variances”

$\boldsymbol{\Lambda}$  :  $p \times q$  matrix of scalars called “factor loadings”. This matrix describes how the observed variables  $\mathbf{x}$  are related to the latent factors  $\mathbf{f}$ .

$\boldsymbol{\mu}$  :  $p \times 1$  vector of scalars. Often ignored, most software assumes by default that  $\boldsymbol{\mu} = \mathbf{0}$  and analyze centered  $\mathbf{x}$  variables, i.e. analyze  $\mathbf{x} - \bar{\mathbf{x}}$

# Exploratory factor analysis vs. Confirmatory factor analysis (EFA vs. CFA)

EFA general purposes:

- To determine how many underlying factors are necessary to explain most of the correlations and variance in the data.
- To determine the relationship via **rotation** between each of these underlying factors with each of the observed variables in a meaningful way so that the factors can be interpreted and named.
- To weed out observed variables that do not tend to measure well the underlying factors shared by the other variables.
- To propose blocks of variables that may be subsequently be used to create a simple sum scale.
- To propose a CFA model

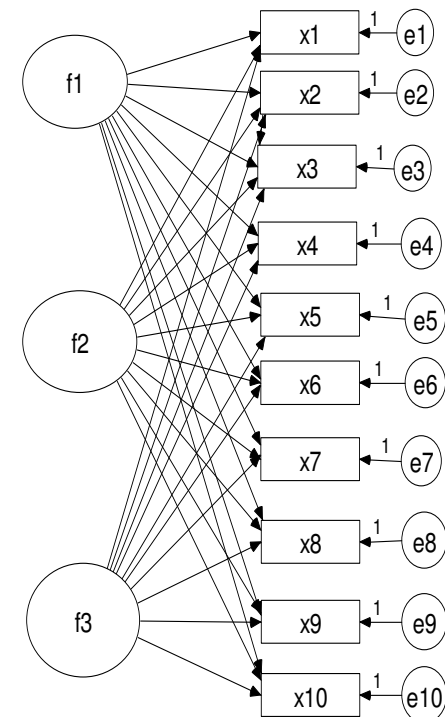
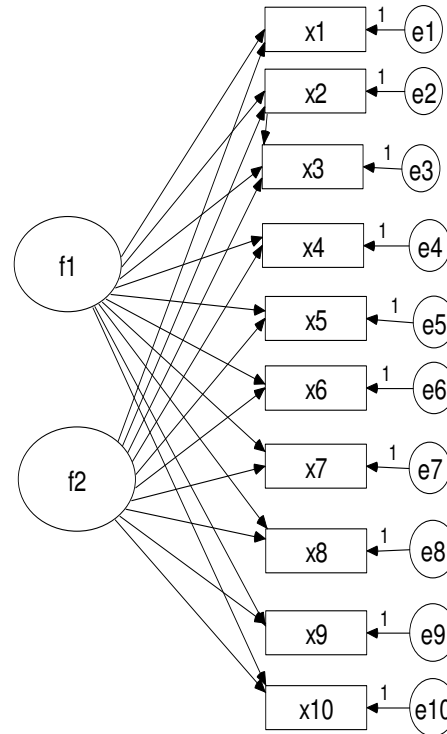
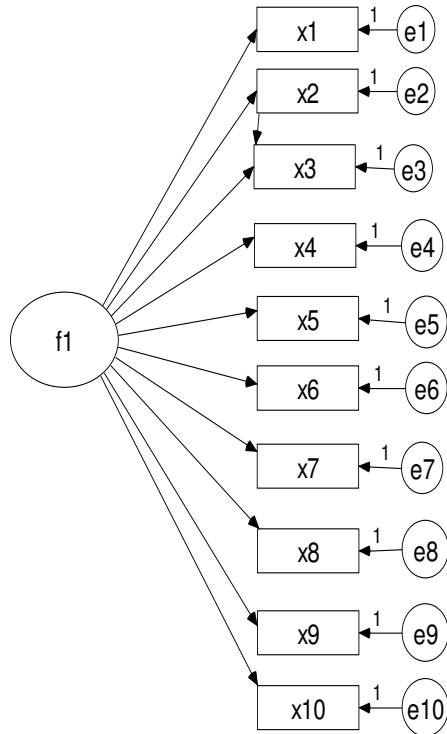
In EFA every element in  $\Lambda$  is estimated and it is assumed that  $\Psi$  is diagonal. Also, it is common to assume that  $Var(\mathbf{f}) = \Phi = \mathbf{I}$ , i.e. the factors are uncorrelated with variance 1 (but this is not a necessary assumption, it is dropped when examining oblique rotations).

# Exploratory factor analysis vs. Confirmatory factor analysis (EFA vs. CFA)

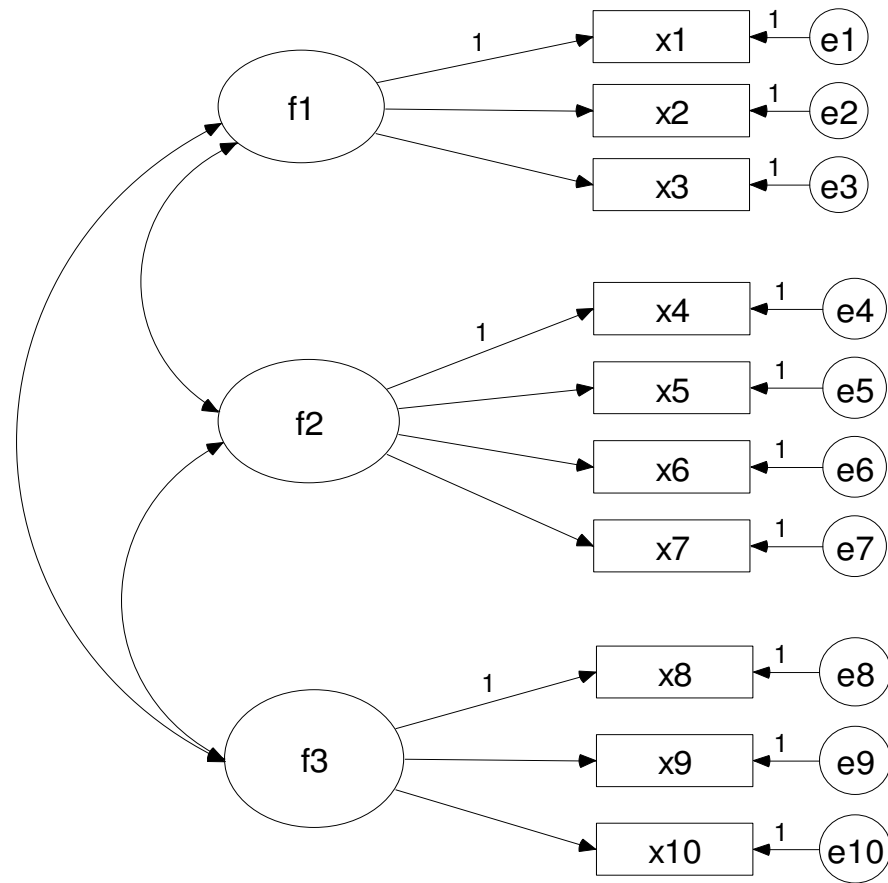
CFA general purposes:

- To define a measurement model for the relationship between multivariate observations and underlying factors
- To test the statistical significance of factor loadings and correlations. Note this testing cannot currently be done in the EFA model. Thus one may be interested in testing whether rotated factor loadings from an EFA that look "close to zero" are, in fact, significantly different from zero or not.
- To test whether the measurement model for one group is the same as the measurement model for some other group
- As a precursor to a Structural equation model

In CFA usually several elements in  $\Lambda$  are fixed to zero and it is possible to consider correlated  $\epsilon$  which means that  $\Psi$  is not necessarily diagonal. Furthermore, it is usually assumed that the factors are correlated so that no restriction is placed on  $\Phi$ .



In EFA every element in  $\Lambda$  is estimated and it is assumed that  $\Psi$  is diagonal. Also, it is common to assume that  $Var(\mathbf{f}) = \Phi = \mathbf{I}$ , i.e. the factors are uncorrelated with variance 1 (but this is not a necessary assumption, it is dropped when examining oblique rotations).



In CFA usually several elements in  $\Lambda$  are fixed to zero and it is possible to consider correlated  $\epsilon$  which means that  $\Psi$  is not necessarily diagonal. Furthermore, it is usually assumed that the factors are correlated so that no restriction is placed on  $\Phi$ .

# Linear Factor Analysis - The Model

$$\mathbf{x} = \boldsymbol{\mu} + \boldsymbol{\Lambda}\mathbf{f} + \boldsymbol{\epsilon}$$

An assumption is made for model (1) that  $Cov(\mathbf{f}, \boldsymbol{\epsilon}) = \mathbf{0}$ . This is a very critical assumption because what it means is that the variability in  $\mathbf{x}$  can be separated into two additive parts, one coming from the **common factors** and one coming from **specific factors** or the errors. That is there is no covariance between the two. Specifically,

$$Var(\mathbf{x}) = \boldsymbol{\Lambda}Var(\mathbf{f})\boldsymbol{\Lambda}' + Var(\boldsymbol{\epsilon}) + 2Cov(\mathbf{f}, \boldsymbol{\epsilon})$$

$$Var(\mathbf{x}) = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}.$$

It is assumed that equation (1) holds for each individual in the population and thus for  $i = 1 \dots n$  independently sampled individuals we have

$$\mathbf{x}_i = \boldsymbol{\mu} + \boldsymbol{\Lambda}\mathbf{f}_i + \boldsymbol{\epsilon}_i.$$

As an example of how to write model (1) , here are the equations when  $p = 5$  and  $q = 2$  and it is assumed that  $\Psi$  is a diagonal matrix, for  $i = 1 \dots n$  we have:

$$\begin{aligned} x_{1i} &= \mu_1 + \lambda_{11}f_{1i} + \lambda_{12}f_{2i} + \epsilon_{1i} \\ x_{2i} &= \mu_2 + \lambda_{21}f_{1i} + \lambda_{22}f_{2i} + \epsilon_{2i} \\ x_{3i} &= \mu_3 + \lambda_{31}f_{1i} + \lambda_{32}f_{2i} + \epsilon_{3i} \\ x_{4i} &= \mu_4 + \lambda_{41}f_{1i} + \lambda_{42}f_{2i} + \epsilon_{4i} \\ x_{5i} &= \mu_5 + \lambda_{51}f_{1i} + \lambda_{52}f_{2i} + \epsilon_{5i} \end{aligned}$$

or in matrix/vector notation

$$\begin{pmatrix} x_{1i} \\ x_{2i} \\ x_{3i} \\ x_{4i} \\ x_{5i} \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \\ \mu_5 \end{pmatrix} + \begin{pmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \\ \lambda_{31} & \lambda_{32} \\ \lambda_{41} & \lambda_{42} \\ \lambda_{51} & \lambda_{52} \end{pmatrix} \begin{pmatrix} f_{1i} \\ f_{2i} \end{pmatrix} + \begin{pmatrix} \epsilon_{1i} \\ \epsilon_{2i} \\ \epsilon_{3i} \\ \epsilon_{4i} \\ \epsilon_{5i} \end{pmatrix}$$

and we also must specify the variance structure of  $\mathbf{f}$  and  $\epsilon$

$$\text{Var}(\mathbf{f}_i) = \text{Var} \begin{pmatrix} f_{1i} \\ f_{2i} \end{pmatrix} = \Phi = \begin{pmatrix} \phi_{11} & \phi_{12} \\ \phi_{12} & \phi_{22} \end{pmatrix}$$

In EFA, the measurement errors are assumed uncorrelated, in CFA this assumption can be dropped

$$\text{Var}(\epsilon_i) = \text{Var} \begin{pmatrix} \epsilon_{1i} \\ \epsilon_{2i} \\ \epsilon_{3i} \\ \epsilon_{4i} \\ \epsilon_{5i} \end{pmatrix} = \Psi = \begin{pmatrix} \psi_1 & 0 & 0 & 0 & 0 \\ 0 & \psi_2 & 0 & 0 & 0 \\ 0 & 0 & \psi_3 & 0 & 0 \\ 0 & 0 & 0 & \psi_4 & 0 \\ 0 & 0 & 0 & 0 & \psi_5 \end{pmatrix}$$

# Communalities

- Communalities are the diagonal elements of  $\Lambda\Lambda'$  when the factors are assumed to be uncorrelated
- Communalities are the part of the variance of each observed variable which is due to the  $q$  underlying factors

$$\begin{aligned}x_1 &= \mu_1 + \lambda_{11}f_1 + \lambda_{12}f_2 + \epsilon_1 \\x_2 &= \mu_2 + \lambda_{21}f_1 + \lambda_{22}f_2 + \epsilon_2 \\x_3 &= \mu_3 + \lambda_{31}f_1 + \lambda_{32}f_2 + \epsilon_3 \\x_4 &= \mu_4 + \lambda_{41}f_1 + \lambda_{42}f_2 + \epsilon_4 \\x_5 &= \mu_5 + \lambda_{51}f_1 + \lambda_{52}f_2 + \epsilon_5\end{aligned}$$

$$\begin{aligned}Var(x_1) &= \lambda_{11}^2 + \lambda_{12}^2 + \psi_1 \\Var(x_2) &= \lambda_{21}^2 + \lambda_{22}^2 + \psi_2 \\Var(x_3) &= \lambda_{31}^2 + \lambda_{32}^2 + \psi_3 \\Var(x_4) &= \lambda_{41}^2 + \lambda_{42}^2 + \psi_4 \\Var(x_5) &= \lambda_{51}^2 + \lambda_{52}^2 + \psi_5\end{aligned}$$

The communality for  $x_2$  is  $\lambda_{21}^2 + \lambda_{22}^2$ . On the other hand,  $\psi_2$  is the part of the variability in  $x_2$  that is unique to  $x_2$ , i.e. is not shared with other observed variables, “is not common”.

# Modeling the covariance matrix - the natural choice for the linear factor analysis model

- Consider  $p$  continuous random variables  $\mathbf{x}$  from some population, let  $\Sigma$  represent the true covariance matrix of  $\mathbf{x}$  in the population. That is,  $Var(\mathbf{x}) = \Sigma$ .
- If  $\mathbf{x}$  were Normally distributed then everything about it is described simply by its mean and its covariance matrix. Thus if  $\mathbf{x}$  can be assumed Normally distributed, the natural thing to model is its mean and variance(covariance) matrix.
- Our goal is to find a parametric model  $\Sigma(\Theta)$  that describes  $\Sigma$  as closely as possible.
- $\mathbf{x} = \boldsymbol{\mu} + \mathbf{\Lambda}\mathbf{f} + \boldsymbol{\epsilon}$  is a model such that

$$Var(\mathbf{x}) = \mathbf{\Lambda}\Phi\mathbf{\Lambda}' + \Psi = \Sigma(\Theta) \quad (2)$$

# Model estimation

$$\text{Var}(\mathbf{x}) = \mathbf{\Lambda}\mathbf{\Phi}\mathbf{\Lambda}' + \mathbf{\Psi} = \mathbf{\Sigma}(\mathbf{\Theta}) \quad (3)$$

Here  $\mathbf{\Theta} = (\mathbf{\Lambda}, \mathbf{\Phi}, \mathbf{\Psi})$  are the parameters

- $\mathbf{\Sigma}(\mathbf{\Theta})$  is called the “model covariance matrix”
- Given  $n$  independent samples of  $\mathbf{x}$  and given that  $\mathbf{x}$  is Normally distributed then the sufficient statistic for  $\mathbf{\Sigma}$  is the sample covariance matrix  $\mathbf{S}$ . The  $p \times p$  sample covariance matrix is

$$\mathbf{S} = \frac{1}{n} \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'$$

- Given a model (i.e.  $\mathbf{\Sigma}(\mathbf{\Theta})$ ) and the data (i.e.  $\mathbf{S}$ ), the goal is to estimate  $\mathbf{\Theta}$  so that  $\mathbf{\Sigma}(\hat{\mathbf{\Theta}})$  is as close to  $\mathbf{S}$  as possible while still being parsimonious.
- Whether model (1) is an EFA or a CFA we will fit the model covariance matrix  $\mathbf{\Sigma}(\mathbf{\Theta})$  to the sample covariance matrix

# Model estimation - Fitting a correlation matrix

When the  $\mathbf{x}$  has been standardized

- Standardize  $\mathbf{x}$  to get  $\mathbf{Z}$ , i.e.  $\mathbf{Z} = \begin{pmatrix} \frac{x_1 - \bar{x}_1}{s_1} \\ \frac{x_2 - \bar{x}_2}{s_2} \\ \cdot \\ \cdot \\ \frac{x_p - \bar{x}_p}{s_p} \end{pmatrix}$

$$\text{Var}(\mathbf{Z}) = \boldsymbol{\rho} = \boldsymbol{\Lambda}^s \boldsymbol{\Lambda}^{s'} + \boldsymbol{\Psi}^s$$

- $\boldsymbol{\Lambda}^s$  are the “standardized factor loadings”
- We will estimate  $\boldsymbol{\rho}$  using  $\mathbf{R}$ , the sample correlation matrix
- For some estimation methods, e.g. maximum likelihood method,  $\boldsymbol{\Lambda}^s$  obtained by analyzing the correlation matrix is the same as rescaling the  $\boldsymbol{\Lambda}$  obtained by analyzing the covariance matrix by the observed standard deviations, i.e.  $\boldsymbol{\Lambda}^s = (\text{diag}(\mathbf{S}))^{-\frac{1}{2}} \boldsymbol{\Lambda}$ . Note THIS IS NOT TRUE in general.

## Estimating $\Theta$

There are many ways to estimate  $\Theta = (\Lambda, \Phi, \Psi)$  in the model covariance matrix  $\Sigma(\Theta)$ . All use some sort of discrepancy function

$$F = F(\mathbf{S}, \Sigma(\Theta))$$

A discrepancy function is just a rule that determines how you will find the “best” estimate for the parameters. As an analogy, in linear regression you have data  $y_i$ ,  $i = 1 \dots n$  and you have some model for each observed  $y_i$ , for example  $\beta_0 + \beta_1 x_i$ . The goal there is to find the values for parameters  $\beta_0$  and  $\beta_1$  that make the model as close to the data as possible. The discrepancy function is what we use to define “close”. A discrepancy function which could be used for this regression is  $F(\text{data}, \text{model}) = \sum_{i=1}^n (y_i - \beta_0 + \beta_1 x_i)^2$  which is the ordinary least squares discrepancy function, but there are other discrepancy functions that could be used for example  $F(\text{data}, \text{model}) = \sum_{i=1}^n w_i (y_i - \beta_0 + \beta_1 x_i)^2$  where  $w_i$  is some predetermined weight.

## Estimating $\Theta$

The common discrepancy functions used for the factor analysis model are

- Normal theory maximum likelihood
- Generalized least squares
- (Un)Weighted least squares
- Asymptotically distribution free (ADF)
- when  $\Phi = \mathbf{I}$  and  $\Psi$  is diagonal the principal factor method can be used which is very quick because it uses eigenvalue eigenvectors

We will call the estimates:  $\hat{\Theta} = (\hat{\Lambda}, \hat{\Psi}, \hat{\Phi})$ , and  $\hat{\Sigma} = \Sigma(\hat{\Theta})$  will be referred to as the “fitted model covariance matrix” or “Implied covariance” (in AMOS) or “estimated covariance matrix” (in MPLUS).

# Normal distribution and the likelihood function

Recall that maximum likelihood estimation asks us to find the  $\boldsymbol{\mu}$  and  $\boldsymbol{\Sigma}$  which maximize the likelihood  $L$  (or the log likelihood).

Given  $n$  i.i.d. random vectors  $\mathbf{x}_i \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  the log likelihood is

$$\log L = \frac{pn}{2} \log 2\pi - \frac{n}{2} \log |\boldsymbol{\Sigma}| - \frac{1}{2} \sum_{i=1}^n (\mathbf{x}_i - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu})$$

After a bit of matrix algebra, this can be transformed into

$$\log L = \text{constant} + \frac{n}{2} [\log |\boldsymbol{\Sigma}^{-1}| - \text{trace}(\mathbf{S}\boldsymbol{\Sigma}^{-1})] - \frac{n}{2} (\bar{\mathbf{x}} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\bar{\mathbf{x}} - \boldsymbol{\mu})$$

where  $\mathbf{S}$  is the sample covariance matrix. Since  $\boldsymbol{\mu}$  only appears in the last term, we can maximize the likelihood with respect to  $\boldsymbol{\mu}$  by minimizing the last term with respect to  $\boldsymbol{\mu}$ . This is clearly done when  $\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}}$

## Normal distribution and the likelihood function

So what we are left with doing is maximizing the remaining part with respect to the parameters in  $\Sigma$ , i.e. with respect to  $\Theta = (\Lambda, \Phi, \Psi)$ . So we want to maximize the following:

$$\log L = \frac{n}{2} [\log |\Sigma^{-1}| - \text{trace}(\mathbf{S}\Sigma^{-1})] \quad (4)$$

The maximization is done by iteration. Given starting values (best first guesses) for the parameters  $P_{sii}$ , it estimates  $\Lambda$  and  $\Phi$ , then given those estimates, it updates its estimate for  $P_{sii}$  and continues back and forth making the  $\log L$  as large as possible.

We will call the maximum likelihood estimators:  $\hat{\Theta} = (\hat{\Lambda}, \hat{\Psi}, \hat{\Phi})$ , and  $\hat{\Sigma} = \Sigma(\hat{\Theta})$  will be referred to as the “fitted model covariance matrix” or “Implied covariance” (in AMOS) or “estimated covariance matrix” (in MPLUS).

# Testing the model fit - the Chi-square test

Besides giving estimates for  $\Lambda$ ,  $\Phi$  and  $\Psi$ , Maximum Likelihood Provides a Goodness of Fit test.

- Use Likelihood Ratio Test, i.e.,

$$\begin{aligned} -2 \log \frac{L(\text{MLE of restricted model})}{L(\text{MLE of unrestricted model})} \\ &= 2\{\log(L(\mathbf{S})) - \log(L(\hat{\underline{\Sigma}}))\} \\ &= n\{\text{trace} \hat{\underline{\Sigma}}^{-1} \mathbf{S} - \log |\hat{\underline{\Sigma}}^{-1} \mathbf{S}| - p\} \end{aligned}$$

- IF the model fits the data well, this statistic should be small!
- Its distribution is asymptotically distributed  $\chi^2$  with degrees of freedom =  $(\frac{p(p+1)}{2} - \text{number of unique parameters in model})$ .
- We can determine if the the statistic is “small” enough by comparing to the  $\chi^2$  distribution and obtaining a p-value.

# The Chi-square test continued

- In general, the Hypothesis being tested is

$$H_0: \Sigma = \Sigma(\Theta) \text{ (your model)}$$

$H_A: \Sigma = \mathbf{S}$  the saturated model (i.e. the ‘model’ that estimates a separate parameter for every unique element of the covariance matrix)

- Since the degrees of freedom for the saturated model are 0, this means it fits the data perfectly
- So we are comparing  $\Sigma(\Theta)$  to a model that fits the data perfectly. Thus if it is not significantly different than the model that fits perfectly, it means it is pretty good
- Thus we are looking for the model where we DO NOT REJECT the  $H_0$  (i.e. find a big p-value)
- Note, the chi-square test has been proven to be **asymptotically** valid even when the data is not normally distributed (Amemiya and Anderson, 1985). Note: this is an asymptotic result.
- Some correction methods exist, e.g. Satorra Bentler correction to the Chi-square statistic when the data is non-normal.

## Other Goodness of Fit statistics

- The strict null hypothesis  $\Sigma = \Sigma(\theta)$  will almost always not be exactly true and with a large enough sample size will be rejected
- “Moreover a  $\chi^2$  test offers only a dichotomous decision strategy implied by a statistical decision rule and cannot be used to quantify the degree of fit along a continuum” (from Hu and Bentler (1995))
- Thus **many** fit indices have been developed to assess the degree of congruence between model and data
- “Like  $R^2$  in multiple regression, these indexes are meant to quantify something akin to variance accounted for, rather than to test null hypotheses. In particular these indexes generally quantify the extent to which the variation and covariation in the data are accounted for by the model”
- For details of each fit statistic and rules of thumb see the handout of the Appendix of the AMOS manual

# Using Fit statistics

Example text from one of my papers...

“Factor loadings were tested for statistical significance and measures were obtained to assess overall model fit. Chi-square was used to test the hypothesis that the relationships proposed in the model provide a plausible explanation of those that exist in the data. Since most models are either slightly misspecified or do not account for all measurement error, when sample sizes are large (as in the present study), a non-significant chi-square is rarely obtained (1,2). Therefore, other measures have been developed to assess the fit of the model; in the present study the Normed Fit Index (NFI) and the root mean squared error of approximation (RMSEA) were used. The NFI specifies the practical fit of the model to the data with acceptable fit equaling 0.90 or greater and a good fit reflected by values of 0.95 or above (3). RMSEA expresses fit per degree of freedom of the model; values of RMSEA of less than 0.08 imply an acceptable model fit and values of less than 0.05 imply a good fit (4). In the present study, 90% confidence intervals for the RMSEA measures are presented.”

1) Bentler PM, Bonett DG. Significance tests and goodness of fit in the analysis of covariance structures. *Psych Bulletin* 1980; 88:588-606.

2) Joreskog SG. Analysis of covariance structures. *Scand J Stat* 1981; 8: 65-92.

3) Joreskog SG, Sorbom D. LISREL 7: a guide to the program and applications. Chicago: SPSS, 1988.

4) Browne MW, Cudeck R. Single sample cross-validation indices for covariance structures. *Multivariate Behav Res* 1990; 24:445-55.

# Outline of Exploratory Factor Analysis

- Factor extraction method
  - principal component analysis  
Starts with correlation matrix and extracts factors (by calculating eigenvectors and eigenvalues) from it directly.
  - common factor analysis (Maximum Likelihood, Least squares, or Principal factor method)  
Starts with correlation matrix but then makes a “reduced correlation matrix” meant to represent the correlation of the “common part”, then extracts factors (by calculating eigenvectors and eigenvalues or using a discrepancy function) from the “reduced correlation matrix”.
- Choose the number of factors
  - examine eigenvalues: % variance explained, eigenvalue  $> 1$ , scree plot
  - Model fit statistics: Chi-square value, RMSR root mean square residuals for correlation matrix, RMSEA (Cudeck) function of chi-square taking into account sample size,  $< .05$  good fit,  $< .08$  ok fit

# PCA vs Factor model

Dimension of  $\mathbf{x}$  is  $p$

Factor model

$$\mathbf{x} = \boldsymbol{\mu} + \boldsymbol{\Lambda}\mathbf{f} + \boldsymbol{\epsilon} = \boldsymbol{\mu} + \lambda_1 f_1 + \lambda_2 f_2 + \dots + \lambda_p f_q + \boldsymbol{\epsilon}$$

$$Var(\mathbf{x}) = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}$$

Factor analysis is a stochastic model. There's a common part ( $q < p$ ) and a unique part. It models the measurement error within each variable.

PCA analysis (it is not really a model)

$$\mathbf{x} = \boldsymbol{\mu} + \boldsymbol{\Lambda}\mathbf{f} = \boldsymbol{\mu} + \lambda_1 f_1 + \lambda_2 f_2 + \dots + \lambda_p f_p$$

$$Var(\mathbf{x}) = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}'$$

PCA is a mathematical manipulation (eigenvalue-eigenvector decomposition) of the covariance matrix (i.e.  $Var(\mathbf{x})$ ). There is no unique part, the  $p$  factors completely explain the covariance matrix. Decide to keep the factors that explain a large proportion, but remaining part is not “unique”, it is not just random measurement error.

# PCA vs Factor model

- Lots of literature trying to claim that one of the two is better than the other
- Some people argue that PCA is better because it can always be performed (no worry about convergence in model estimation)
- Some people argue that FA is better because it is a model and reflects hypothesis about measurement and as a model it is possible to develop test statistics and do inference for parameters, also argue that it is better because its results do not depend on scale of observed variables
- Some people have tried to show analytically cases where they give identical results, show with simulation studies cases where they give similar results, find cases where they give very different results

# PCA vs Factor model

- Fact is, they are similar, and very often lead to qualitatively the same results
- “It is unclear why so many comments attempted to stress the differences between the methods. We do not believe that finding such differences would be desirable. We view the relative lack of difference under most circumstances to be reassuring and to provide support for the continued use of both methods.” from Velicer and Jackson (1990) Component analysis versus common factor analysis: Some issues in selecting an appropriate procedure”
- My feeling is that fitting the factor model makes the most theoretical sense. Which estimation method may vary, maximum likelihood can run into problems, principal factor method is a stable alternative. Fitting the factor model (rather than doing PCA) also leads into the more general SEM (which cannot be done with PCA). Note: PCA is limited to exploratory factor analysis.

# Outline of Exploratory Factor Analysis - Continued

- Interpreting the factors - Rotation
  - rotate to simple structure - orthogonal (varimax) or oblique (promax) correlated factors
  - examining size of loadings (stat sig?), cross loadings
- Determining quality of factors and items
  - subject matter interpretation
  - number of items loading on the factor
  - variance explained by the factor
  - eliminate poor items and factors and repeat EFA
- Possible next steps
  - Verify structure using confirmatory factor analysis
  - Calculate factor score estimates  $E(f|X)$ , Mplus produces a “factor determinacy” value which represents correlation between factor and the estimated factor score
  - Calculate simple sum scores of items associated with each  $f$  with simple structure, use Cronbach alpha as estimate of reliability

# Maximum number of factors possible to extract in EFA

- When the PCA method is used, it is possible to extract as many factors as there are variables.
- For the common factor methods which assume that the factors explain only the systematic part of the data (not the specific errors), the number of factors is constrained. The following must be true  $\frac{p(p+1)}{2} - pq - p + \frac{q(q-1)}{2} \geq 0$   
So for example when  $p = 3$  or  $p = 4$ ,  $q$  must be no greater than 1. When  $p = 20$ ,  $q$  can be up to 14.

This restriction is related to the degrees of freedom associated with the common factor model. The number of parameters being estimated must be less than or equal to the amount of unconstrained data. For the EFA model the d.f. are  $\frac{p(p+1)}{2} - pq - p + \frac{q(q-1)}{2}$ . The first term is the number of unique elements in the sample covariance matrix  $\mathbf{S}$ , then we are going to estimate  $pq$  factor loadings in  $\mathbf{\Lambda}$  and  $p$  variances in  $\mathbf{\Psi}$  and since we have to put restrictions in order to fix rotation we get those  $\frac{q(q-1)}{2}$  d.f. back. These  $\frac{q(q-1)}{2}$  d.f. correspond to fixing the  $\mathbf{\Lambda}'\mathbf{\Psi}^{-1}\mathbf{\Lambda}$  off diagonal elements being fixed to zero.

## Rotation in EFA - this is an issue related to model identification

No matter what estimation procedure, for exploratory factor analysis we get estimates that look like:

$$\hat{\Sigma} = \hat{\Lambda}\hat{\Lambda}' + \hat{\Psi}$$

We can get the exact same  $\hat{\Sigma}$  by taking

$$\hat{\Sigma} = \hat{\Lambda}TT'\hat{\Lambda}' + \hat{\Psi}$$

where  $T$  is a  $q \times q$  matrix such that  $TT' = \text{Identity matrix}$ .

Thus

$$\hat{\Sigma} = \hat{\Lambda}^*\hat{\Lambda}^{*'} + \hat{\Psi}$$

where  $\hat{\Lambda}^*$  is the **orthogonally rotated** factor loading matrix, i.e.  $\hat{\Lambda}^* = \hat{\Lambda}T$ . BUT, There are an infinite number of matrices  $T$  that satisfy  $TT' = I$ .

# Rotation

- Why do we bother to rotate? – We want to find out which variables “stick together”
- How to choose a  $T$ ? – Find one that gives close to “simple” structure to the factor loadings
- “Simple structure” means that each observed variable only loads on one factor. In an AMOS notation, each observed variable only has one arrow going to it from a factor. Simple structure allows you to say, for example, “observed variable 1 is a direct measure of only factor 1”, instead of “variable 1 is direct measure of factor 1, 2, and 3”
- What about if simple structure is not found by orthogonal rotation?
- NOTE: implicitly orthogonal rotation keeps the assumption that  $Var(\mathbf{f}) = I$ , i.e. the factors are uncorrelated.
- Maybe the assumption that  $Var(\mathbf{f}) = I$  should be dropped. This leads to **oblique rotations** of the factor loadings.
- We can search for

$$\hat{\Sigma} = \Lambda^{**} \Phi \hat{\Lambda}^{**'} + \hat{\Psi}$$

so that  $\hat{\Lambda}^{**}$  has “simple structure” and  $\Phi \neq I$

# Matching items to factors - Examine the rotated loadings

Throughout it is necessary to keep in mind whether the items loading on a given factor share some conceptual meaning.

- What is a BIG loading?
  - Common rule of thumb, if the absolute value of the standardized loading is  $> .3$ , the variable is relevant for the particular factor.
  - In 1994 Cudeck, R. and O'Dell, L. “Applications of standard error estimates in unrestricted factor analysis: Significance Tests for factor loadings and correlations” *Psychological Bulletin* argue that the  $> .3$  cutoff is too simplistic and encourage sig testing.
  - Now possible to get standard errors from SAS Proc factor when using method = ml and the se option.
  - Additional work on calculated standard errors is Ogasawara (1998) “Standard errors for rotation matrices with an application to the promax solution”. All of these results are based on asymptotic results, thus it is still not clear how useful they are for problems with small sample sizes.

## Matching items to factors - Continued

- Examine the communality. Rule of thumb, if the communality is less than .10 then the observed variable may be deleted. A communality  $< .10$  means that less than 10% of the variability in the observed variable is explained by all of the common factors.
- Items with strong loadings on multiple factors (4 options)
  - allow the item to load on more than one factor
  - choose the item to load on the factor it has the largest loading with
  - choose the item to load on the factor it conceptually goes best with
  - drop the item
- Cross Validation. If the sample size is large enough, an intuitive way to gain an idea about the stability of the factor structure is to split the sample randomly into two equal parts and then fit the model to each part. If the factor structure is similar, this tends to increase our confidence in the genuineness of the factors.

## Variance explained by factors

In EFA the sum of the total variance explained by  $k$  factors is same for any orthogonal rotation of the  $k$  factors. The difference is how the variance is spread out across the factors.

On the other hand, when an oblique rotation is used, summarizing the variance explained by the factors is tricky because we need to talk about the variance explained after adjusting for the other factors (since they are correlated).

# School Subjects example from BSMG

Possible to read just the correlation matrix into SAS for analysis.

```
data a (type = corr);
input _type_ $ _name_ $ gaelic english history math algebra geometry;
cards;
mean . 0 0 0 0 0 0
std . 1 1 1 1 1 1
n . 220 220 220 220 220 220
corr gaelic      1 .44 .41 .29 .33 .25
corr english    .44 1 .35 .35 .32 .33
corr history    .41 .35 1 .16 .19 .18
corr math       .29 .35 .16 1 .59 .47
corr algebra    .33 .32 .19 .59 1 .46
corr geometry   .25 .33 .18 .47 .46 1 ;
run;

proc contents data = a; run;

proc factor data = a;
var gaelic english history math algebra geometry;
run;
```

options include method = prinit, method = uls, method = ml. By default with no method stated, SAS performs principal component analysis. Other options available are regarding the rotation method which can be used with all of the estimation methods. They are rotate=varimax and rotate=promax

# School Subjects example from BSMG - PCA output

```
proc factor data = a rotate = promax;  
var gaelic english history math algebra geometry; run;
```

The FACTOR Procedure

```
Initial Factor Method: Principal Components  
Prior Commnality Estimates: ONE  
Eigenvalues of the Correlation Matrix: Total = 6 Average = 1
```

	Eigenvalue	Difference	Proportion	Cumulative
1	2.72868347	1.59989129	0.4548	0.4548
2	1.12879218	0.51350073	0.1881	0.6429
3	0.61529145	0.01248259	0.1025	0.7455
4	0.60280886	0.08029449	0.1005	0.8459
5	0.52251437	0.12060470	0.0871	0.9330
6	0.40190967		0.0670	1.0000

2 factors will be retained by the MINEIGEN criterion.

Factor Pattern

	Factor1	Factor2
gaelic	0.66080	0.44447
english	0.68846	0.28977
history	0.51636	0.63955
math	0.73562	-0.41702
algebra	0.74187	-0.37276
geometry	0.67817	-0.35410

Variance Explained by Each Factor

Factor1	Factor2
2.7286835	1.1287922

Final Commnality Estimates: Total = 3.857476

gaelic	english	history	math	algebra	geometry
0.63421803	0.55795080	0.67565080	0.71504052	0.68931666	0.58529884

# School Subjects example from BSMG - PCA OUTPUT

Prerotation Method: Varimax

Orthogonal Transformation Matrix

	1	2
1	0.77371	0.63354
2	-0.63354	0.77371

Rotated Factor Pattern

	Factor1	Factor2
gaelic	0.22967	0.76254
english	0.34909	0.66037
history	-0.00567	0.82196
math	0.83335	0.14340
algebra	0.81015	0.18160
geometry	0.74904	0.15568

Variance Explained by Each Factor

Factor1	Factor2
2.0865260	1.7709497

Final Communality Estimates: Total = 3.857476

gaelic	english	history	math	algebra	geometry
0.63421803	0.55795080	0.67565080	0.71504052	0.68931666	0.58529884

# School Subjects example from BSMG - PCA OUTPUT

Rotation Method: Promax (power = 3)

## Inter-Factor Correlations

	Factor1	Factor2
Factor1	1.00000	0.36730
Factor2	0.36730	1.00000

## Rotated Factor Pattern (Standardized Regression Coefficients)

	Factor1	Factor2
gaelic	0.09286	0.75757
english	0.23917	0.62522
history	-0.16759	0.86862
math	0.85268	-0.01983
algebra	0.82064	0.02525
geometry	0.76115	0.01045

## Variance Explained by Each Factor Eliminating Other Factors

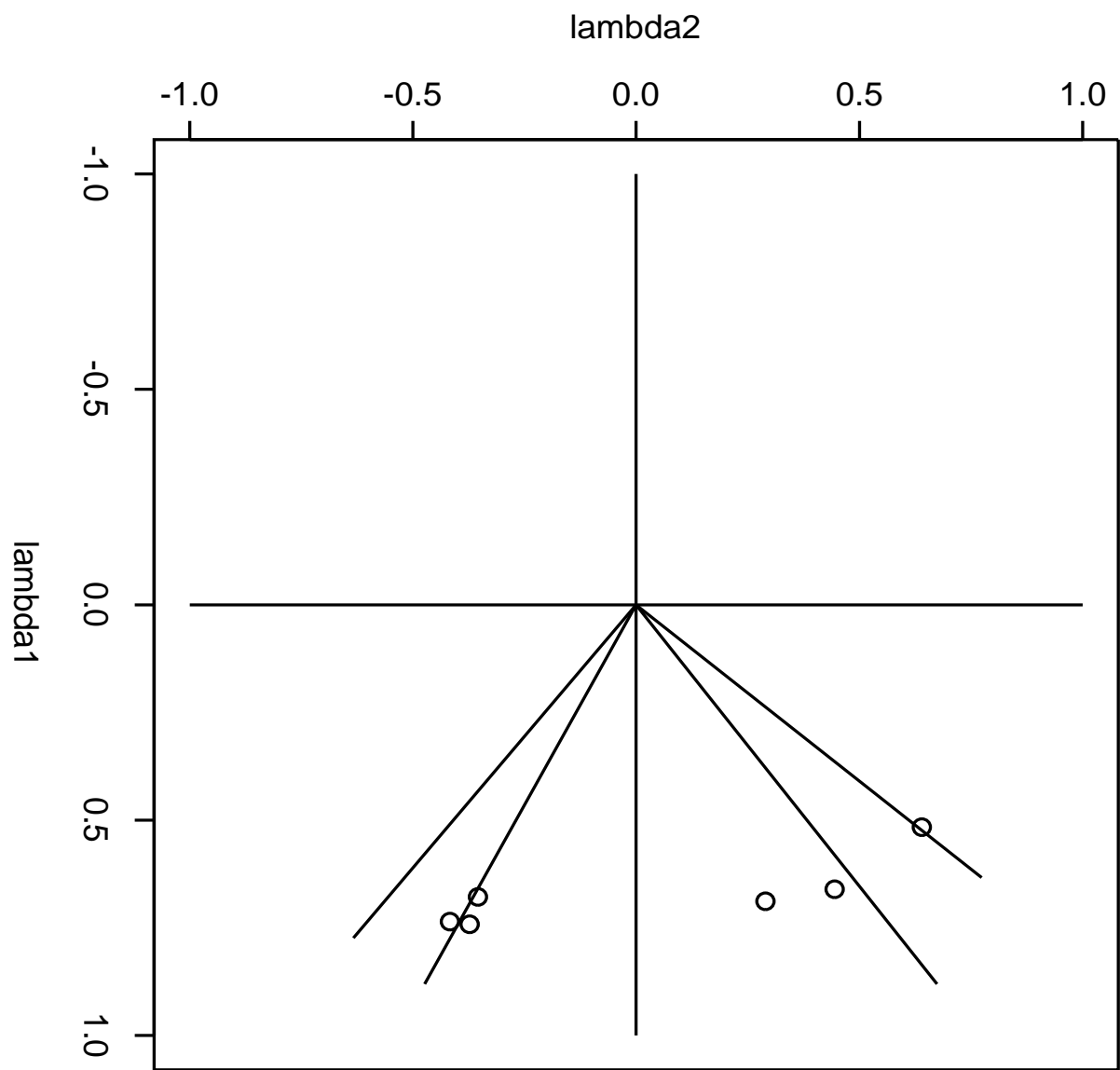
Factor1	Factor2
1.7940097	1.4883579

## Variance Explained by Each Factor Ignoring Other Factors

Factor1	Factor2
2.3691177	2.0634659

Final Communalities Estimates: Total = 3.857476

gaelic	english	history	math	algebra	geometry
0.63421803	0.55795080	0.67565080	0.71504052	0.68931666	0.58529884



# School Subjects example from BSMG - ML OUTPUT

```
proc factor data = a method = ml rotate = promax;  
var gaelic english history math algebra geometry; run;
```

The FACTOR Procedure

Initial Factor Method: Maximum Likelihood

Prior Communality Estimates: SMC

gaelic	english	history	math	algebra	geometry
0.30168715	0.29583987	0.20564352	0.41522406	0.41174029	0.29450749

Preliminary Eigenvalues: Total = 2.93847068 Average = 0.48974511

	Eigenvalue	Difference	Proportion	Cumulative
1	3.17047301	2.54848714	1.0790	1.0790
2	0.62198587	0.72927492	0.2117	1.2906
3	-.10728905	0.06189654	-0.0365	1.2541
4	-.16918559	0.08194532	-0.0576	1.1965
5	-.25113091	0.07525173	-0.0855	1.1111
6	-.32638264		-0.1111	1.0000

Significance Tests Based on 220 Observations

Test	DF	Chi-Square	Pr > ChiSq
H0: No common factors	15	308.5731	<.0001
HA: At least one common factor			
H0: 2 Factors are sufficient	4	2.1801	0.7027
HA: More factors are needed			

# School Subjects example from BSMG - ML OUTPUT

Eigenvalues of the Weighted Reduced Correlation Matrix: Total = 5.71554705 Average = 0.95259117

	Eigenvalue	Difference	Proportion	Cumulative
1	4.57559833	3.43564960	0.8006	0.8006
2	1.13994873	1.05025464	0.1994	1.0000
3	0.08969409	0.06384279	0.0157	1.0157
4	0.02585130	0.03607174	0.0045	1.0202
5	-.01022044	0.09510453	-0.0018	1.0184
6	-.10532497		-0.0184	1.0000

Factor Pattern  
Estimate/StdErr

	Factor1	Factor2
gaelic	0.55798 0.07342	0.42461 0.08368
english	0.56901 0.06148	0.28574 0.08579
history	0.39239 0.07933	0.44952 0.09364
math	0.73802 0.05948	-0.27943 0.06620
algebra	0.71847 0.04950	-0.20868 0.08511
geometry	0.59488 0.05347	-0.13331 0.08464

# School Subjects example from BSMG - ML OUTPUT

Rotation Method: Promax (power = 3)

Rotated Factor Pattern (Standardized Regression Coefficients)

Estimate/StdErr

	Factor1	Factor2
gaelic	0.06268 0.07879	0.66737 0.09821
english	0.19141 0.08863	0.51810 0.09333
history	-0.08658 0.05532	0.63580 0.08912
math	0.81101 0.07027	-0.04511 0.05589
algebra	0.73469 0.07277	0.02597 0.06756
geometry	0.57413 0.07394	0.06505 0.07831

# School Subjects example from BSMG - OUTPUT From MPLUS - ML

## DATA:

```
FILE IS "C:\Mplus\ph5482\schoolsubjects.txt";  
TYPE IS FULLCORR;  
NOOBSERVATIONS = 220;
```

## VARIABLE:

```
NAMES ARE gaelic english history math algebra geometry;  
USEVARIABLES ARE gaelic english history math algebra geometry;
```

## ANALYSIS:

```
TYPE IS EFA 1 3;  
ESTIMATOR IS ML;  
ITERATIONS = 1000;  
CONVERGENCE = 0.00005;
```

-----  
INPUT READING TERMINATED NORMALLY

## SUMMARY OF ANALYSIS Number of groups

1 Number of observations

220

Number of dependent variables 6

Number of independent variables 0

Number of continuous latent variables 0

Observed dependent variables

Continuous

GAELIC ENGLISH HISTORY MATH ALGEBRA GEOMETRY

## EIGENVALUES FOR SAMPLE CORRELATION MATRIX

	1	2	3	4	5	6
1	2.729	1.129	0.615	0.603	0.523	0.402

# MPLUS - ML results for fitting 1 factor

## RESULTS FOR EXPLORATORY FACTOR ANALYSIS

EXPLORATORY ANALYSIS WITH 1 FACTOR(S) :

CHI-SQUARE VALUE                    52.682  
DEGREES OF FREEDOM                    9  
PROBABILITY VALUE                    0.0000

RMSEA (ROOT MEAN SQUARE ERROR OF APPROXIMATION) :  
ESTIMATE (90 PERCENT C.I.) IS 0.149 ( 0.111 0.189)  
PROBABILITY RMSEA LE 0.05 IS 0.000

ROOT MEAN SQUARE RESIDUAL IS                    0.0987

ESTIMATED FACTOR LOADINGS

1

-----  
GAELIC            0.500  
ENGLISH           0.539  
HISTORY           0.350  
MATH              0.726  
ALGEBRA           0.729  
GEOMETRY         0.615

ESTIMATED RESIDUAL VARIANCES

	GAELIC	ENGLISH	HISTORY	MATH	ALGEBRA	Geometry
	-----	-----	-----	-----	-----	-----
1	0.750	0.710	0.878	0.473	0.468	0.621

# MPLUS - ML results for fitting 2 factors

EXPLORATORY ANALYSIS WITH 2 FACTOR(S) :

CHI-SQUARE VALUE                    2.233  
DEGREES OF FREEDOM                    4  
PROBABILITY VALUE                    0.6928

RMSEA (ROOT MEAN SQUARE ERROR OF APPROXIMATION) :  
ESTIMATE (90 PERCENT C.I.) IS   0.000 ( 0.000 0.077)  
PROBABILITY RMSEA LE 0.05 IS    0.854  
ROOT MEAN SQUARE RESIDUAL IS     0.0140

VARIMAX ROTATED LOADINGS

1                    2

-----

GAELIC	0.660	0.236
ENGLISH	0.550	0.321
HISTORY	0.591	0.083
MATH	0.169	0.771
ALGEBRA	0.217	0.716
GEOMETRY	0.213	0.571

PROMAX ROTATED LOADINGS

1                    2

-----

GAELIC	0.683	0.033
ENGLISH	0.533	0.167
HISTORY	0.648	-0.113
MATH	-0.031	0.805
ALGEBRA	0.040	0.726
GEOMETRY	0.077	0.566

PROMAX FACTOR CORRELATIONS

1                    2

-----

1	1.000	
2	0.524	1.000

ESTIMATED RESIDUAL VARIANCES

GAELIC                    ENGLISH                    HISTORY                    MATH                    ALGEBRA                    Geometry

-----

1	0.508	0.595	0.644	0.377	0.440	0.628
---	-------	-------	-------	-------	-------	-------

MUTHEN & MUTHEN 3463 Stoner Ave. Los Angeles, CA 90066 Tel: (310)

391-9971 Fax: (310) 391-8971 Web: [www.StatModel.com](http://www.StatModel.com) Support:

Support@StatModel.com Copyright (c) 1998-2004 Muthen & Muthen

# Observed versus Model Estimated correlation matrix

Add the Output: sampstat residual; command to the Mplus code

## SAMPLE STATISTICS

### Covariances/Correlations/Residual Correlations

	GAELIC	ENGLISH	HISTORY	MATH	ALGEBRA
GAELIC	1.000				
ENGLISH	0.440	1.000			
HISTORY	0.410	0.350	1.000		
MATH	0.290	0.350	0.160	1.000	
ALGEBRA	0.330	0.320	0.190	0.590	1.000
GEOMETRY	0.250	0.330	0.180	0.470	0.460

## ESTIMATED MODEL AND RESIDUALS (OBSERVED - ESTIMATED)

1-factor model

### Model Estimated Correlations

	GAELIC	ENGLISH	HISTORY	MATH	ALGEBRA
GAELIC	1.000				
ENGLISH	0.269	1.000			
HISTORY	0.175	0.188	1.000		
MATH	0.363	0.391	0.254	1.000	
ALGEBRA	0.365	0.393	0.255	0.529	1.000
GEOMETRY	0.308	0.331	0.215	0.447	0.449

### Residuals for Correlations

	GAELIC	ENGLISH	HISTORY	MATH	ALGEBRA
GAELIC	0.000				
ENGLISH	0.171	0.000			
HISTORY	0.235	0.162	0.000		
MATH	-0.073	-0.041	-0.094	0.000	
ALGEBRA	-0.035	-0.073	-0.065	0.061	0.000
GEOMETRY	-0.058	-0.001	-0.035	0.023	0.011

# Observed versus Model Estimated correlation matrix

Add the Output: sampstat residual; command to the Mplus code

## SAMPLE STATISTICS

### Covariances/Correlations/Residual Correlations

	GAELIC	ENGLISH	HISTORY	MATH	ALGEBRA
GAELIC	1.000				
ENGLISH	0.440	1.000			
HISTORY	0.410	0.350	1.000		
MATH	0.290	0.350	0.160	1.000	
ALGEBRA	0.330	0.320	0.190	0.590	1.000
GEOMETRY	0.250	0.330	0.180	0.470	0.460

## ESTIMATED MODEL AND RESIDUALS (OBSERVED - ESTIMATED)

2-factor model

### Model Estimated Correlations

	GAELIC	ENGLISH	HISTORY	MATH	ALGEBRA
GAELIC	1.000				
ENGLISH	0.439	1.000			
HISTORY	0.410	0.352	1.000		
MATH	0.293	0.340	0.164	1.000	
ALGEBRA	0.312	0.349	0.188	0.589	1.000
GEOMETRY	0.275	0.300	0.173	0.476	0.455

### Residuals for Correlations

	GAELIC	ENGLISH	HISTORY	MATH	ALGEBRA
GAELIC	0.000				
ENGLISH	0.001	0.000			
HISTORY	0.000	-0.002	0.000		
MATH	-0.003	0.010	-0.004	0.000	
ALGEBRA	0.018	-0.029	0.002	0.001	0.000
GEOMETRY	-0.025	0.030	0.007	-0.006	0.005

THESE ARE SOME ISSUES THAT PEOPLE HAVE INDICATED ARE CONCERNS WHEN MAKING THEIR DECISION ABOUT GENETIC TESTING	IS THIS A CONCERN FOR YOU?				
	1 Not at all	2 Slightly	3 Moderately	4 Quite a bit	5 Extremely
1. I might increase my sense of personal control over the condition after I receive the testing results	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
2. I worry about being faced with an uncertain diagnosis	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
3. I want to know what I should do to manage my risk for cancer.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
4. I hope that knowing the results will help reduce my uncertainty about the future.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
5. I fear being faced with the ambiguity of the results	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
6. I worry about being faced with a diagnosis I don't know what to do about	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
7. I hope to be able to make better health and lifestyle choices as a result.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
8. I am worried about being able to maintain health and life insurance coverage.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
9. I am worried about how my family will react to the testing information.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
10. I need information about the types of cancers I am at risk for.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
11. I am concerned about marital/family problems that might occur from obtaining the test results.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
12. I might be helped to make important future life decisions by knowing I carry the gene, e.g., getting married, having children.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
13. I need information about participation in future screening activities.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
14. I want to know how a positive test will affect my children and other family members.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
15. I am worried about my future life.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
16. I want information about the difference between the diagnosis of having the gene and getting cancer.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
17. I have financial concerns related to the screening.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
18. I want to know the financial and social implications	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
19. I worry about being targeted as a carrier of being identified as a carrier.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
20. I want to know my survival prospects.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

# Genetic Concerns Questionnaire - EFA - ML

```
proc factor data = geneticstest scree  
method = ml rotate = promax nfact=4 flag = .3; var c1-c20; run;
```

The FACTOR Procedure

Initial Factor Method: Maximum Likelihood

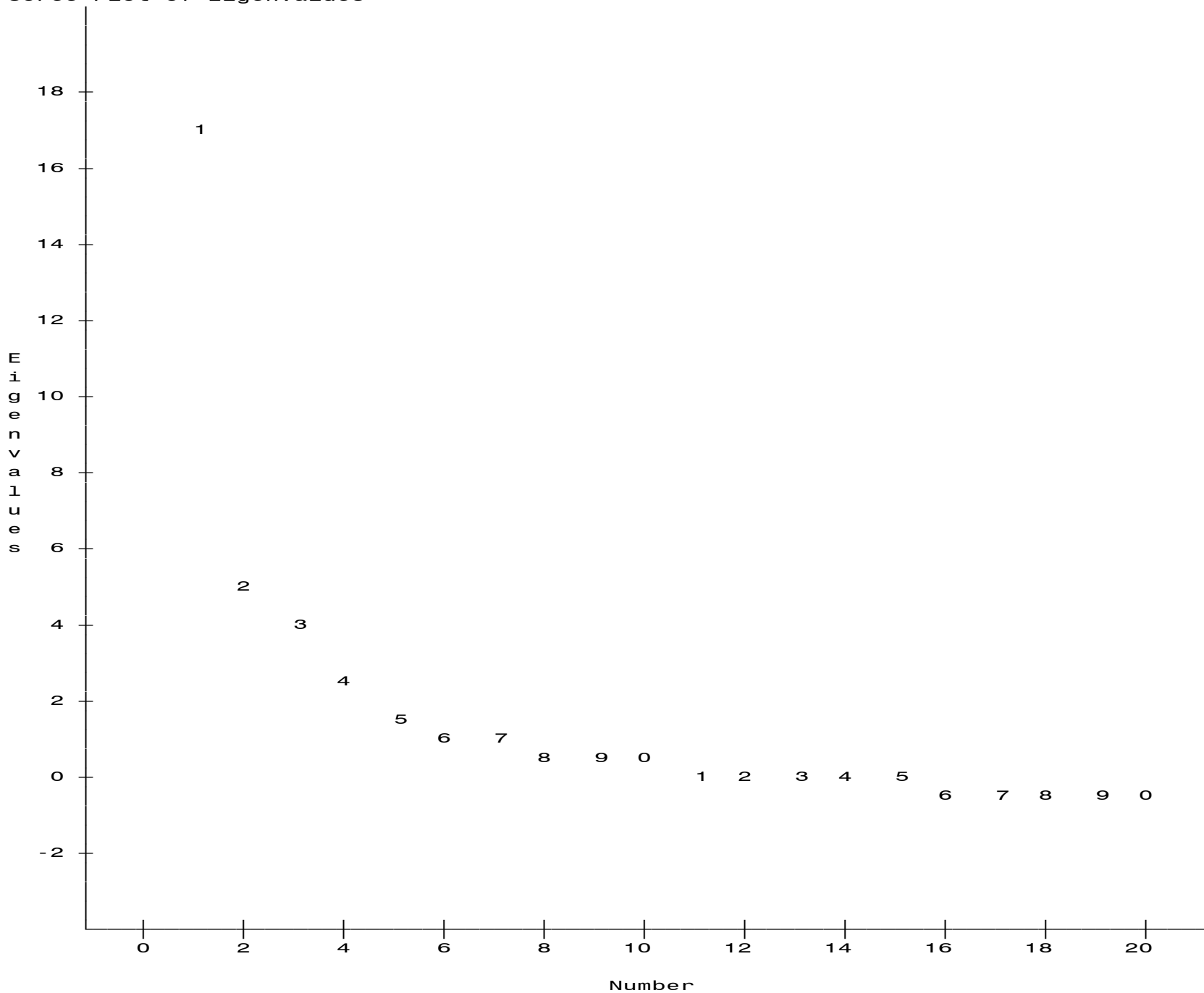
Prior Communalities Estimates: SMC

C1	C2	C3	C4	C5	C6	C7
0.41029924	0.69555063	0.69345231	0.54651245	0.51414191	0.58527718	0.62437419
C8	C9	C10	C11	C12	C13	C14
0.47598962	0.60194636	0.63636979	0.39761689	0.52851826	0.67491498	0.59745581
C15	C16	C17	C18	C19	C20	
0.67083904	0.58465071	0.52457147	0.64761079	0.69479515	0.61797719	

Preliminary Eigenvalues: Total = 30.372251 Average = 1.51861255

	Eigenvalue	Difference	Proportion	Cumulative
1	16.7672687	11.9963216	0.5521	0.5521
2	4.7709471	1.0027379	0.1571	0.7091
3	3.7682092	1.0293909	0.1241	0.8332
4	2.7388183	1.3083265	0.0902	0.9234
5	1.4304918	0.3590838	0.0471	0.9705
6	1.0714080	0.2230647	0.0353	1.0058
7	0.8483433	0.1408710	0.0279	1.0337
8	0.7074723	0.1696340	0.0233	1.0570
9	0.5378383	0.2754239	0.0177	1.0747
10	0.2624143	0.0221818	0.0086	1.0833
11	0.2402325	0.1761596	0.0079	1.0912
12	0.0640730	0.0939153	0.0021	1.0934
13	-0.0298423	0.1060550	-0.0010	1.0924
14	-0.1358972	0.1120360	-0.0045	1.0879
15	-0.2479332	0.0857999	-0.0082	1.0797
16	-0.3337331	0.0373563	-0.0110	1.0687
17	-0.3710894	0.0893819	-0.0122	1.0565
18	-0.4604714	0.0903286	-0.0152	1.0414
19	-0.5507999	0.1546994	-0.0181	1.0232
20	-0.7054993		-0.0232	1.0000

Scree Plot of Eigenvalues



# Genetic Concerns Questionnaire - EFA - ML

Rotated Factor Pattern (Standardized  
Regression Coefficients)

		Factor1	Factor2
C1	C1	57 *	-13
C2	C2	-4	77 *
C3	C3	77 *	0
C4	C4	15	53 *
C5	C5	-18	71 *
C6	C6	23	56 *
C7	C7	73 *	-9
C8	C8	14	26
C9	C9	-1	38 *
C10	C10	70 *	3
C11	C11	-4	36 *
C12	C12	63 *	-4
C13	C13	72 *	11
C14	C14	35 *	27
C15	C15	7	76 *
C16	C16	59 *	11
C17	C17	55 *	4
C18	C18	47 *	7
C19	C19	43 *	16
C20	C20	78 *	-4

Test	DF	Chi-Square	Pr > ChiSq
H0: No common factors	190	2126.5455	<.0001
HA: At least one common factor			
H0: 2 Factors are sufficient	151	818.6337	<.0001
HA: More factors are needed			

# Genetic Concerns Questionnaire - EFA - ML

Rotated Factor Pattern (Standardized Regression Coefficients)

		Factor1	Factor2	Factor3
C1	C1	51 *	-13	12
C2	C2	1	77 *	-5
C3	C3	86 *	1	-11
C4	C4	16	52 *	2
C5	C5	-19	68 *	6
C6	C6	23	54 *	5
C7	C7	83 *	-8	-15
C8	C8	-12	19	50 *
C9	C9	-4	36 *	10
C10	C10	66 *	3	7
C11	C11	-6	35 *	4
C12	C12	57 *	-4	11
C13	C13	70 *	11	5
C14	C14	31 *	25	11
C15	C15	10	77 *	-2
C16	C16	65 *	12	-8
C17	C17	30	-2	50 *
C18	C18	9	-2	73 *
C19	C19	-3	5	91 *
C20	C20	61 *	-8	33 *

Test	DF	Chi-Square	Pr > ChiSq
H0: No common factors	190	2126.5455	<.0001
HA: At least one common factor			
H0: 3 Factors are sufficient	133	593.5235	<.0001
HA: More factors are needed			

# Genetic Concerns Questionnaire - EFA - ML

Rotated Factor Pattern (Standardized  
Regression Coefficients)

		Factor1	Factor2	Factor3	Factor4
C1	C1	51 *	-4	13	-16
C2	C2	0	77 *	-4	7
C3	C3	85 *	-1	-11	4
C4	C4	17	68 *	6	-24
C5	C5	-19	63 *	8	10
C6	C6	23	45 *	3	21
C7	C7	82 *	-2	-15	-4
C8	C8	-11	28	55 *	-19
C9	C9	-11	5	1	73 *
C10	C10	66 *	6	7	-3
C11	C11	-12	8	-3	63 *
C12	C12	57 *	-7	9	7
C13	C13	69 *	7	4	12
C14	C14	26	-3	4	62 *
C15	C15	10	68 *	-2	17
C16	C16	67 *	13	-7	-4
C17	C17	28	-12	48 *	20
C18	C18	7	-3	74 *	4
C19	C19	-4	2	90 *	7
C20	C20	62 *	-2	35 *	-12

	Test	DF	Chi-Square	Pr > ChiSq
H0:	No common factors	190	2126.5455	<.0001
HA:	At least one common factor			
H0:	4 Factors are sufficient	116	454.4120	<.0001
HA:	More factors are needed			

# Genetic Concerns Questionnaire - EFA - ML

Rotated Factor Pattern (Standardized  
Regression Coefficients)

		Factor1	Factor2	Factor3	Factor4	Factor5
C1	C1	32 *	1	12	-16	32 *
C2	C2	-6	80 *	-7	8	8
C3	C3	91 *	-5	-8	3	-3
C4	C4	8	71 *	4	-23	10
C5	C5	-2	60 *	11	10	-40 *
C6	C6	24	44 *	4	21	-6
C7	C7	76 *	-2	-12	-5	11
C8	C8	1	23	59 *	-20	-28
C9	C9	-7	4	3	73 *	-8
C10	C10	43 *	13	5	-3	45 *
C11	C11	-16	10	-4	64 *	9
C12	C12	29	1	5	8	57 *
C13	C13	61 *	9	6	11	15
C14	C14	31 *	-6	6	62 *	-3
C15	C15	5	71 *	-5	18	9
C16	C16	67 *	12	-4	-5	0
C17	C17	24	-14	50 *	20	7
C18	C18	1	-4	75 *	4	7
C19	C19	-15	2	92 *	7	12
C20	C20	58 *	-4	39 *	-12	5

Pr >

Test

DF

Chi-Square

ChiSq

H0: No common factors	190	2126.5455	<.0001
HA: At least one common factor			
H0: 5 Factors are sufficient	100	374.4601	<.0001
HA: More factors are needed			

# Genetic Concerns Questionnaire - EFA - ML

1 factor	Akaike's Information Criterion	784.3265
	Schwarz's Bayesian Criterion	219.4148
2 factor	Akaike's Information Criterion	553.68549
	Schwarz's Bayesian Criterion	51.91098
3 factor	Akaike's Information Criterion	356.51308
	Schwarz's Bayesian Criterion	-85.44725
4 factor	Akaike's Information Criterion	246.24617
	Schwarz's Bayesian Criterion	-139.22299
5 factor	Akaike's Information Criterion	195.46089
	Schwarz's Bayesian Criterion	-136.84011

# Genetic Concerns Questionnaire - EFA - ML

Use the options “reorder” to get SAS to arrange output

Rotated Factor Pattern (Standardized Regression Coefficients)

		Factor1	Factor2	Factor3	Factor4
C3	C3	85 *	-1	-11	4
C7	C7	82 *	-2	-15	-4
C13	C13	69 *	7	4	12
C16	C16	67 *	13	-7	-4
C10	C10	66 *	6	7	-3
C20	C20	62 *	-2	35 *	-12
C12	C12	57 *	-7	9	7
C1	C1	51 *	-4	13	-16
C2	C2	0	77 *	-4	7
C15	C15	10	68 *	-2	17
C4	C4	17	68 *	6	-24
C5	C5	-19	63 *	8	10
C6	C6	23	45 *	3	21
C19	C19	-4	2	90 *	7
C18	C18	7	-3	74 *	4
C8	C8	-11	28	55 *	-19
C17	C17	28	-12	48 *	20
C9	C9	-11	5	1	73 *
C11	C11	-12	8	-3	63 *
C14	C14	26	-3	4	62 *

AND here are the associated Cronbach alphas using only the items with loading over .4. NOTE these were created separately.

.87                      .82                      .79                      .73

**Table 6.4** Content Areas and Factors on Which the 20 CGTS Items Loaded

Item	Content Area	Factor			
		1	2	3	4
C1	Personal control	.	.	.	.
C3	How to manage risk	.	.	.	.
C7	Make better lifestyle choices	.	.	.	.
C10	Need cancer information	.	.	.	.
C12	Helped to make future life decisions	.	.	.	.
C16	Information re diagnosis	.	.	.	.
C13	Need screening information	.	.	.	.
C20	Information re survival prospects	.	.	.	.
C6	Worry about the diagnosis	.	.	.	.
C2	Uncertain diagnosis	.	.	.	.
C4	Reduce uncertainty	.	.	.	.
C5	Fear ambiguity	.	.	.	.
C15	Worried about future life	.	.	.	.
C8	Health, life insurance	.	.	.	.
C18	Financial and social implications	.	.	.	.
C19	Being targeted as a carrier	.	.	.	.
C17	Financial concerns	.	.	.	.
C9	Family reactions	.	.	.	.
C11	Marital, family problems	.	.	.	.
C14	Effect of positive test on family members	.	.	.	.

④ C20 fits best with these items.

① C6 fits best with these items.

③ C17 fits best with these items.

② C14 fits best with these items.