Paper 200-31
Exploratory or Confirmatory Factor Analysis?
Diana D. Suhr, Ph.D.
University of Northern Colorado

Abstract
Exploratory factor analysis (EFA) could be described as orderly simplification of interrelated measures. EFA, traditionally, has been used to explore the possible underlying factor structure of a set of observed variables without imposing a preconceived structure on the outcome (Child, 1990). By performing EFA, the underlying factor structure is identified.

Confirmatory factor analysis (CFA) is a statistical technique used to verify the factor structure of a set of observed variables. CFA allows the researcher to test the hypothesis that a relationship between observed variables and their underlying latent constructs exists. The researcher uses knowledge of the theory, empirical research, or both, postulates the relationship pattern a priori and then tests the hypothesis statistically.

The process of data analysis with EFA and CFA will be explained. Examples with FACTOR and CALIS procedures will illustrate EFA and CFA statistical techniques.

Introduction
CFA and EFA are powerful statistical techniques. An example of CFA and EFA could occur with the development of measurement instruments, e.g. a satisfaction scale, attitudes toward health, customer service questionnaire. A blueprint is developed, questions written, a scale determined, the instrument pilot tested, data collected, and CFA completed. The blueprint identifies the factor structure or what we think it is. However, some questions may not measure what we thought they should. If the factor structure is not confirmed, EFA is the next step. EFA helps us determine what the factor structure looks like according to how participant responses. Exploratory factor analysis is essential to determine underlying constructs for a set of measured variables.

Confirmatory Factor Analysis
CFA allows the researcher to test the hypothesis that a relationship between the observed variables and their underlying latent construct(s) exists. The researcher uses knowledge of the theory, empirical research, or both, postulates the relationship pattern a priori and then tests the hypothesis statistically.
The use of CFA could be impacted by
- the research hypothesis being testing
- the requirement of sufficient sample size (e.g., 5-20 cases per parameter estimate)
- measurement instruments
- multivariate normality
- parameter identification
- outliers
- missing data
- interpretation of model fit indices (Schumacker & Lomax, 1996).

A suggested approach to CFA proceeds through the following process:
- review the relevant theory and research literature to support model specification
- specify a model (e.g., diagram, equations)
- determine model identification (e.g., if unique values can be found for parameter estimation; the number of degrees of freedom, df, for model testing is positive)
- collect data
- conduct preliminary descriptive statistical analysis (e.g., scaling, missing data, collinearity issues, outlier detection)
- estimate parameters in the model
- assess model fit
- present and interpret the results.

Statistics
Traditional statistical methods normally utilize one statistical test to determine the significance of the analysis. However, Structural Equation Modeling (SEM), CFA specifically, relies on several statistical tests to determine the adequacy of model fit to the data. The chi-square test indicates the amount of difference between expected and observed covariance matrices. A chi-square value close to zero indicates little difference between the expected and observed covariance matrices. In addition, the probability level must be greater than 0.05 when chi-square is close to zero.

The Comparative Fit Index (CFI) is equal to the discrepancy function adjusted for sample size. CFI ranges from 0 to 1 with a larger value indicating better model fit. Acceptable model fit is indicated by a CFI value of 0.90 or greater (Hu & Bentler, 1999).

Root Mean Square Error of Approximation (RMSEA) is related to residual in the model. RMSEA values range from 0 to 1 with a smaller RMSEA value indicating better model fit. Acceptable model fit is indicated by an RMSEA value of 0.06 or less (Hu & Bentler, 1999).

If model fit is acceptable, the parameter estimates are examined. The ratio of each parameter estimate to its standard error is distributed as a z statistic and is significant at the 0.05 level if its value exceeds 1.96 and at the 0.01 level it its value exceeds 2.56 (Hoyle, 1995). Unstandardized parameter estimates retain scaling information of variables and can only be interpreted with reference to the scales of the variables. Standardized parameter estimates are transformations of unstandardized estimates that remove scaling and can be used for informal comparisons of parameters throughout the model. Standardized estimates correspond to effect-size estimates.

In CFA, if unacceptable model fit is found, an EFA can be performed.

PROC CALIS
The PROC CALIS procedure (Covariance Analysis of Linear Structural Equations) estimates parameters and tests the appropriateness of structural equation models using covariance structural analysis. Although PROC CALIS was designed to specify linear relations, structural equation modeling (SEM) techniques have the flexibility to test nonlinear trends. CFA is a special case of SEM.

PROC CALIS and options for CFA
DATA =   specifies dataset to be analyzed
COV   covariance matrix
CORR   correlation matrix

Exploratory Factor Analysis
Psychologists searching for a neat and tidy description of human intellectual abilities lead to the development of factor analytic methods. Galton, a scientist during the 19th and 20th centuries, laid the foundations for factor analytic methods by developing quantitative methods to determine the interdependence between 2 variables. Karl Pearson was the first to explicitly define factor analysis. In 1902, Macdonnell was the first to publish an application of factor analysis, a comparison of physical characteristics between 3000 criminals and 1000 Cambridge undergraduates.

Factor analysis could be described as orderly simplification of interrelated measures. Traditionally factor analysis has been used to explore the possible underlying structure of a set of interrelated variables without imposing any preconceived structure on the outcome (Child, 1990). By performing exploratory factor analysis (EFA), the number of constructs and the underlying factor structure are identified.

EFA
- is a variable reduction technique which identifies the number of latent constructs and the underlying factor structure of a set of variables
- hypothesizes an underlying construct, a variable not measured directly
- estimates factors which influence responses on observed variables
- allows you to describe and identify the number of latent constructs (factors)
- includes unique factors, error due to unreliability in measurement
- traditionally has been used to explore the possible underlying factor structure of a set of measured variables without imposing any preconceived structure on the outcome (Child, 1990).

Goals of factor analysis are
1) to help an investigator determine the number of latent constructs underlying a set of items (variables)
2) to provide a means of explaining variation among variables (items) using a few newly created variables (factors), e.g., condensing information
3) to define the content or meaning of factors, e.g., latent constructs

Assumptions underlying EFA are
- Interval or ratio level of measurement
- Random sampling
- Relationship between observed variables is linear
- A normal distribution (each observed variable)
- A bivariate normal distribution (each pair of observed variables)
- Multivariate normality

Limitations of EFA are
- The correlations, the basis of factor analysis, describe relationships. No causal inferences can be made from correlations alone.
- the reliability of the measurement instrument (avoid an instrument with low reliability)
- sample size (larger sample → larger correlation)
    ◊ minimal number of cases for reliable results is more then 100 observations and 5 times the number of items
    ◊ since some subjects may not answer every item, a larger sample is desirable. For example, 30 items would require at least 150 cases (5*30), a sample of 200 subjects would allow for missing data
- sample selection
    ◊ Representative of population
    ◊ Do not pool populations
- variables could be sample specific, e.g., a unique quality possessed by a group does not generalize to the population
- nonnormal distribution of data

Factor Extraction
Factor analysis seeks to discover common factors. The technique for extracting factors attempts to take out as much common variance as possible in the first factor. Subsequent factors are, in turn, intended to account for the maximum amount of the remaining common variance until, hopefully, no common variance remains.

Direct extraction methods obtain the factor matrix directly from the correlation matrix by application of specified mathematical models. Most factor analysts agree that direct solutions are not sufficient. Adjustment to the frames of reference by rotation methods improves the interpretation of factor loadings by reducing some of the ambiguities which accompany the preliminary analysis (Child, 1990). The process of manipulating the reference axes is known as rotation.

Rotation applied to the reference axes means the axes are turned about the origin until some alternative position has been reached. The simplest case is when the axes are held at 90° to each other, orthogonal rotation. Rotating the axes through different angles gives an oblique rotation (not at 90° to each other).

Criteria for Extracting Factors
Determining the number of factors to extract in a factor analytic procedure means keeping the factors that account for the most variance in the data. Criteria for determining the number of factors are:
1) Kaiser's criterion, suggested by Guttman and adapted by Kaiser, considers factors with an eigenvalue greater than one as common factors (Nunnally, 1978)
2) Cattell's (1966) scree test. The name is based on an analogy between the debris, called scree, that collects at the bottom of a hill after a landslide, and the relatively meaningless factors that result from overextraction. On a scree plot, because each factor explains less variance than the preceding factors, an imaginary line connecting the markers for successive factors generally runs from top left of the graph to the bottom right. If there is a point below which factors explain relatively little variance and above which they explain substantially more, this usually appears as an "elbow" in the plot. This plot bears some physical resemblance to the profile of a hillside. The portion beyond the elbow corresponds to the rubble, or scree, that gathers. Cattell's guidelines call for retaining factors above the elbow and rejecting those below it.
3) Proportion of variance accounted for keeps a factor if it accounts for a predetermined amount of the variance (e.g., 5%, 10%).
4) Interpretability criteria
    a. Are there at least 3 items with significant loadings (>0.30)?
    b. Do the variables that load on a factor share some conceptual meaning?
    c. Do the variables that load on different factors seem to measure different constructs?
    d. Does the rotated factor pattern demonstrate simple structure? Are there relatively
        i. high loadings on one factor?
        ii. low loadings on other factors?

EFA decomposes an adjusted correlation matrix. Variables are standardized in EFA, e.g., mean=0, standard deviation=1, diagonals are adjusted for unique factors, 1-u. The amount of variance explained is equal to the trace of the matrix, the sum of the adjusted diagonals or communalities. Squared multiple correlations (SMC) are used as communality estimates on the diagonals. Observed variables are a linear combination of the underlying and unique factors. Factors are estimated, (X1 = b1F1 + b2F2 + . . . e1 where e1 is a unique factor).

Factors account for common variance in a data set. The amount of variance explained is the trace (sum of the diagonals) of the decomposed adjusted correlation matrix. Eigenvalues indicate the amount of variance explained by each factor. Eigenvectors are the weights that could be used to calculate factor scores. In common practice, factor scores are calculated with a mean or sum of measured variables that "load" on a factor.

The EFA Model is $Y = X\beta + E$
where   Y is a matrix of measured variables
          X is a matrix of common factors
          β is a matrix of weights (factor loadings)
          E is a matrix of unique factors, error variation

Communality is the variance of observed variables accounted for by a common factor. A large communality value indicates a strong influence by an underlying construct. Community is computed by summing squares of factor loadings

$$d_1^2 = 1 - \text{communality} = \% \text{ variance accounted for by the unique factor}$$
$$d_1 = \text{square root (1-community)} = \text{unique factor weight (parameter estimate)}$$

EFA Steps
1) initial extraction
   - each factor accounts for a maximum amount of variance that has not previously been accounted for by the other factors
   - factors are uncorrelated
   - eigenvalues represent amount of variance accounted for by each factor
2) determine number of factors to retain
   - scree test, look for elbow
   - proportion of variance
   - prior communality estimates are not perfectly accurate, cumulative proportion must equal 100% so some eigenvalues will be negative after factors are extracted, e.g., if 2 factors are extracted, cumulative proportion equals 100% with 6 items, then 4 items have negative eigenvalues
   - interpretability
     - at least 3 observed variables per factor with significant factors
     - common conceptual meaning
     - measure different constructs
     - rotated factor pattern has simple structure (no cross loadings)
3) rotation – a transformation
4) interpret solution
5) calculate factor scores
6) results in a table
7) prepare results, paper

PROC FACTOR and options for EFA
DATA =   specifies dataset to be analyzed
PRIORS =SMC   squared multiple correlations used as adjusted diagonals of the correlation matrix
METHOD =ML,ULS   specifies maximum likelihood and unweighted least squares methods
ROTATE =   PROMAX (ORTHOGONAL), VARIMAX(OBLIQUE)
SCREE   requests a scree plot of the eigenvalues
N =   specifies number of factors
MINEIGEN=1   specifies select factors with eigenvalues greater than 1
OUT =   data and estimated factor scores, use raw data and N=
FLAG =   include a flag (*) for factor loadings above a specified value
REORDER =

An example of SAS code to run EFA. priors specify the prior communality estimate

```
  proc factor method=ml priors=smc      maximum likelihood factor analysis
              method=uls priors=smc      unweighted least squares factor analysis
              method=prin priors=smc     principal factor analysis.
```

Similarities between CFA and EFA
- Both techniques are based on linear statistical models.
- Statistical tests associated with both methods are valid if certain assumptions are met.
- Both techniques assume a normal distribution.
- Both incorporate measured variables and latent constructs.

Differences between CFA and EFA
CFA requires specification of
- a model a priori
- the number of factors
- which items load on each factor
- a model supported by theory or previous research
- error explicitly

EFA
- determines the factor structure (model)
- explains a maximum amount of variance

Statistical Analysis
With background knowledge of confirmatory and exploratory factor analysis, we're ready to proceed to the statistical analysis!

Example 1 - Health Data and Participants
Example 1 hypothesizes two latent constructs related to wellness, physical (fitness, exercise, illness) and mental (stress, hardiness). CFA analyzes data from a study where researchers investigated the effects of hardiness, stress, fitness, and exercise on health problems (Roth, et al., 1989). College students (n=373) reported physical illness, stressful life events, exercise participation levels, perceived fitness levels, and hardiness components. Previously, multiple regression and SEM analyses examined the effects related to illness.

Subjects
Subjects were 163 men and 210 women enrolled in an introductory psychology course at a southern United States university. The mean age of the subjects was 21.7 (sd = 5.5).

Assessments
Fitness. Fitness Questionnaire (Roth & Fillingim, 1988) is a measure of self-perceived physical fitness. Respondents rate themselves on 12 items related to fitness and exercise capacity. The items are on an 11-point scale of 0 = very poor fitness to 5 = average fitness to 10 = excellent fitness. A total fitness score is calculated by summing the 12 ratings. Items include questions about strength, endurance, and general perceived fitness.

Exercise. Exercise Participation Questionnaire (Roth & Fillingim, 1988) assessed current exercise activities, frequency, duration, and intensity. An aerobic exercise participation score was calculated using responses to 15 common exercise activities and providing blank spaces to write in additional activities.

Illness. Seriousness of Illness Rating Scale (Wyler, Masuda, & Holmes, 1968) is a self-report checklist of commonly recognized physical symptoms and diseases and provides a measure of current and recent physical health problems. Each item is associated with a severity level. A total illness score is obtained
by adding the severity ratings of endorsed items (symptoms experienced within the last month).

Stress. Life Experience Survey (Sarason, Johnson, & Segal, 1978) is a measure used to access the occurrence and impact of stressful life experiences. Subjects indicate which events have occurred within the last month and rate the degree of impact on a 7-point scale (-3 = extremely negative impact, 0 = no impact, 3 = extremely positive impact). In the study, the total negative event score was used as an index of negative life stress (the absolute value of the sum of negative items).

Hardiness. In the study, hardiness included components of commitment, challenge, and control. A composite hardiness score was obtained by summing Z scores from scales on each component. The challenge component included one scale whereas the other components included 2 scales. Therefore, the challenge Z score was doubled when calculating the hardiness composite score. Commitment was assessed with the Alienation From Self and Alienation From Work scales of the Alienation Test (Maddi, Kobasa, & Hoover, 1979). Challenge was measured with the Security Scale of the California Life Goals Evaluation Schedule (Hahn, 1966). Control was assessed with the External Locus of Control Scale (Rotter, Seaman, & Liverant, 1962) and the Powerlessness Scale of the Alienation Test (Maddi, Kobasa, & Hoover, 1979).

CFA Diagram
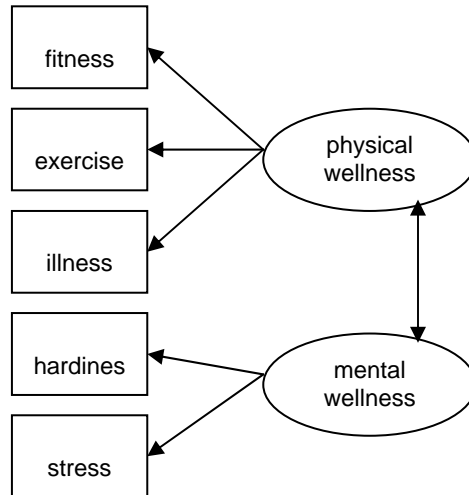
(error terms not
 shown on diagram)

SAS Code - CFA

```
data illfl (type=corr);
   input  _type_ $1-4     _name_ $6-13
          exercise 15-20  hardy 22-27
          fitness 29-34   stress 36-41
          illness 43-48;
cards;
n               373    373     373     373     373
mean            40.90   0.00  67.10   4.80  716.7
std             66.50   3.80  18.40   6.70 624.8
corr    exercise  1.00
corr    hardy    -0.03   1.00
corr    fitness   0.39   0.07   1.00
corr    stess    -0.05  -0.23  -0.13  1.00
corr    illness  -0.08  -0.16  -0.29  0.34  1.00
;;;;

proc calis data=illfl corr;
  lineqs
    hardy   = p1 F1 + e1,
    stress  = p2 F1 + e2,
    illness = p3 F1 + e3,
    fitness = p4 F2 + e4,
    exercise= p5 F2 + e5;
  std
    e1-e5 = vare1-vare5,
    F1 = 1,
    F2 = 1;
  cov
    F1 F2 = covf1f2;
VAR exercise fitness hardy stress illness;
```

Results

PROC CALIS procedure provides the number of observations, variables, estimated parameters, and informations (related to model specification). Notice measured variables have different scales.

```
The CALIS Procedure
Covariance Structure Analysis: Maximum Likelihood Estimation


Observations         373    Model Terms            1
Variables              5    Model Matrices         4
Informations          15    Parameters            11
```

```
Variable            Mean          Std Dev
exercise       40.90000         66.50000
fitness        67.10000         18.40000
hardy                 0          3.80000
stress          4.80000          6.70000
illness       716.70000       624.80000
```

Fit Statistics
Determine criteria a priori to access model fit and confirm the factor structure. Some of the criteria indicate acceptable model fit while other are close to meeting values for acceptable fit.
- Chi-square describes similarity of the observed and expected matrices. Acceptable model fit. Is indicated by a chi-square probability greater than or equal to 0.05. For this CFA model, the chi-square value is close to zero and p = 0.0478, almost the 0.05 value.
- RMSEA indicates the amount of unexplained variance or residual. The 0.0613 RMSEA value is larger than the 0.06 or less criteria.
- CFI (0.9640), NNI (0.9101), and NFI (0.9420) values meet the criteria (0.90 or larger) for acceptable model fit.

For purposes of this example, 3 fit statistics indicate acceptable fit and 2 fit statistics are close to indicating acceptable fit. The CFA analysis has confirmed the factor structure. If the analysis indicates unacceptable model fit, the factor structure cannot be confirmed, an exploratory factor analysis is the next step.

```
The CALIS Procedure
Covariance Structure Analysis: Maximum Likelihood Estimation


Fit Function . . .
Chi-Square                                            9.5941
Chi-Square DF                                              4
Pr > Chi-Square                                       0.0478
. . .
RMSEA Estimate                                        0.0613
. . .
Bentler's Comparative Fit Index                       0.9640
. . .
Bentler & Bonett's (1980) Non-normed Index            0.9101
Bentler & Bonett's (1980) NFI                         0.9420
. . .
```

Parameter Estimates
When acceptable model fit is found, the next step is to determine significant parameter estimates.
- A t value is calculated by dividing the parameter estimate by the standard error, 0.3213 / 0.1123 = 2.8587.
- Parameter estimates are significant at the 0.05 level if the t value exceeds 1.96 and at the 0.01 level if the t value exceeds 2.56.
- Parameter estimates for the confirmatory factor model are significant at the 0.01 level.

```
Manifest Variable Equations with Estimates
exercise =   0.3212*F2      +  1.0000 e5
Std Err      0.1123 p5
t Value      2.8587

fitness  =   1.2143*F2      +  1.0000 e4
Std Err      0.3804 p4
t Value      3.1923

hardy    =   0.2781*F1      +  1.0000 e1
Std Err      0.0673 p1
t Value      4.1293

stress   =  -0.4891*F1      +  1.0000 e2
Std Err      0.0748 p2
t Value     -6.5379

illness  =  -0.7028*F1      +  1.0000 e3
Std Err      0.0911 p3
t Value     -7.7157
```

Variances are significant at the 0.01 level for each error variance except vare4 (error variance for fitness).

```
             Variances of Exogenous Variables
                                    Standard
Variable Parameter      Estimate        Error    t Value
F1                       1.00000
F2                       1.00000
e1      vare1            0.92266        0.07198     12.82
e2      vare2            0.76074        0.07876      9.66
e3      vare3            0.50604        0.11735      4.31
e4      vare4           -0.47457        0.92224     -0.51
e5      vare5            0.89685        0.09209      9.74
```

Covariance between latent constructs is significant at the 0.01 level.

```
             Covariances Among Exogenous Variables
                                         Standard
Var1     Var2     Parameter      Estimate       Error    t Value
F1       F2       covf1f2         0.30404      0.11565      2.63
```

Correlation between latent constructs is 0.30. Latent constructs are uncorrelated.

```
Correlations Among Exogenous Variables
Var1     Var2     Parameter       Estimate
F1       F2       covf1f2          0.30404
```

Standardized Estimates
Report equations with standardized estimates when measured variables have different scales

```
Manifest Variable Equations with Standardized Estimates
exercise =   0.3212*F2      +  0.9470 e5
                 p5
fitness  =   1.2143*F2      +  1.0000 e4
                 p4
hardy    =   0.2781*F1      +  0.9606 e1
                 p1
stress   =  -0.4891*F1      +  0.8722 e2
                 p2
illness  =  -0.7028*F1      +  0.7114 e3
                 p3
```

Examples 2 and 3 - Achievement Data and Participants
Data is from the National Longitudinal Survey of Youth , a longitudinal study of achievement, behavior, and home environment. The original NLSY79 sample design enabled researchers to study longitudinal experiences of different age groups as well as analyze experiences of women, Hispanics, Blacks, and economically disadvantaged. The NLSY79 is a nationally representative sample of 12,686 young men and women who were 14- to 22-years old when first surveyed in 1979 (Baker, Keck, Mott, & Quinlan, 1993).

As part of the NLSY79, mothers and their children have been surveyed biennially since 1986. Although the NLSY79 initially analyzed labor market behavior and experiences, the child assessments were designed to measure academic achievement as well as psychological behaviors. The child sample consisted of all children born to NLSY79 women respondents participating in the study biennially since 1986. The number of children interviewed for the NLSY79 from 1988 to 1994 ranged from 4,971 to 7,089. The total number of cases available for the analysis is 2212.

Measurement Instrument
The PIAT (Peabody Individual Achievement Test) was developed following principles of item response theory (IRT). PIAT Reading Recognition, Reading Comprehension, and Mathematics Subtest scores are measured variables in the analysis Tests were admininstered in 1988, 1990, 1992, and 1994.

SAS Code – CFA

```
proc calis data=rawfl COV stderr;
  lineqs
    math88 = pm88f1 F1 + em88,
    math90 = pm90f1 F1 + em90,
    math92 = pm92f1 F1 + em92,
    math94 = pm94f1 F1 + em94,
    readr88 = prr88f2 F2 + err88,
    readr90 = prr90f2 F2 + err90,
    readr92 = prr92f2 F2 + err92,
    readr94 = prr94f2 F2 + err94,
    readc88 = prc88f2 F3 + erc88,
    readc90 = prc90f2 F3 + erc90,
    readc92 = prc92f2 F3 + erc92,
    readc94 = prc94f2 F3 + erc94;
  std
    em88 = varem88,
    em90 = varem90,
    em92 = varem92,
    em94 = varem94,
    err88 = varerr88,
    err90 = varerr90,
    err92 = varerr92,
    err94 = varerr94,
    erc88 = varerc88,
    erc90 = varerc90,
    erc92 = varerc92,
    erc94 = varerc94,
    F1 = 1, F2 = 1, F3 = 1;
  cov
    F1 F2 = covf1f2,
    F1 F3 = covf1f3,
    F2 F3 = covf2f3;
var math88  math90  math92  math94
    readr88 readr90 readr92 readr94
    readc88 readc90 readc92 readc94;
```

CFA Diagram
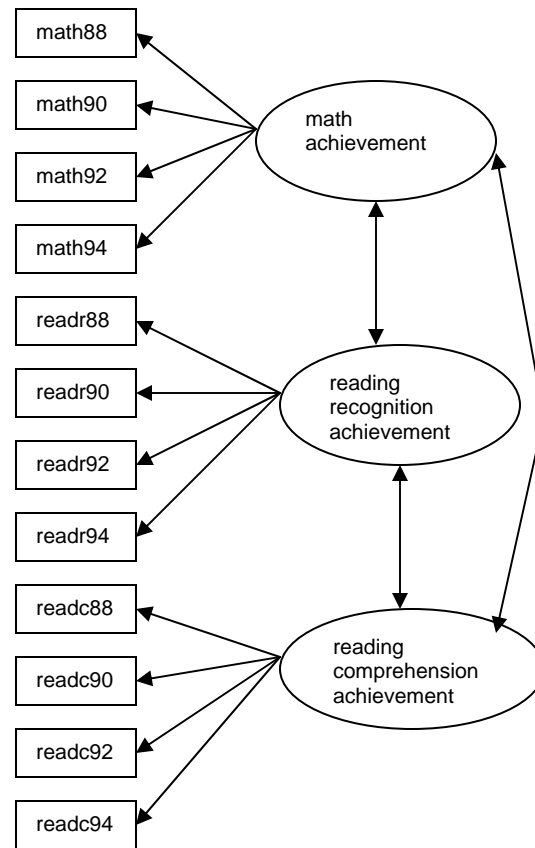
(error terms not shown on diagram)

Results

PROC CALIS procedure provides the number of observations, variables, estimated parameters, and informations (related to model specification).

```
The CALIS Procedure
Covariance Structure Analysis: Maximum Likelihood Estimation


Observations          995    Model Terms              1
Variables              12    Model Matrices           4
Informations           78    Parameters              25


Variable         Mean        Std Dev
math88       19.83920        9.91231
math90       33.93166       11.22261
math92       44.30050        9.84064
math94       50.70151       10.12559
readr88      22.35176       10.63057
readr90      36.95276       12.40289
readr92      48.04925       13.09248
readr94      56.00101       13.98530
readc88      21.24322        9.61194
readc90      34.78392       12.08132
readc92      44.42010       11.19870
readc94      48.54070       10.96891
```

Fit Statistics
Fit Statistics indicate unacceptable model fit.
- Chi-square is large. The model does not produce a small difference between observed and expected matrices.
- Chi-square probability (< 0.0001) is unacceptable. Criteria for acceptable model is pr > 0.05.
- Unexplained variance, residual, is un acceptable. Model RMSEA (0.2188) is greater than the 0.06 or less criteria.
- CFI ((0.8069), NNI (0.7595), and NFI (0.8038) are less than the acceptable criteria, 0.90.

```
Fit Function . . .
Chi-Square                                         2576.1560
Chi-Square DF                                            53
Pr > Chi-Square                                      <.0001
. . .
RMSEA Estimate                                       0.2188
. . .
Bentler's Comparative Fit Index                      0.8069
. . .
Bentler & Bonett's (1980) Non-normed Index           0.7595
Bentler & Bonett's (1980) NFI                        0.8038
. . .
```

The factor structure is not confirmed. No further investigation of the confirmatory model is necessary, parameter estimates, variances, covariances. Proceed with **exploratory factor analysis** to determine the factor structure**.**

SAS Code – EFA
Exploratory factor analysis with
- dataset rawfl
- method is maximum likelihood
- scree plot of eigenvalues
- diagonals of the correlation matrix are equal to squared multiple correclations
- measured variables are math, reading recognition, reading comprehension

```
proc factor data=rawfl method=ml scree priors=smc;
   var math88 math90 math92 math94
       readr88 readr90 readr92 readr94
       readc88 readc90 readc92 readc94;
```

Three factors are retained, cumulative variance is 1.0215.
- Preliminary eigenvalues are 43.7192325, 4.9699785, 1.4603799,
- Each factor explains 89%, 10% and 3% of the variance.

```
The FACTOR Procedure
Initial Factor Method: Maximum Likelihood
Preliminary Eigenvalues: Total = 49.0952995  Average = 4.09127496
        Eigenvalue    Difference     Proportion    Cumulative
    1    43.7192325    38.7492540        0.8905        0.8905
    2     4.9699785     3.5095987        0.1012        0.9917
    3     1.4603799     0.8226791        0.0297        1.0215
    4     0.6377007     0.5299327        0.0130        1.0345
    5     0.1077680     0.0819166        0.0022        1.0367
    6     0.0258514     0.0874121        0.0005        1.0372
    7    -0.0615606     0.1892583       -0.0013        1.0359
    8    -0.2508189     0.0309477       -0.0051        1.0308
    9    -0.2817666     0.0602684       -0.0057        1.0251
   10    -0.3420350     0.0922321       -0.0070        1.0181
   11    -0.4342670     0.0208965       -0.0088        1.0093
   12    -0.4551635                     -0.0093        1.0000
3 factors will be retained by the PROPORTION criterion.
```

Look for an "elbow" in the scree plot to determine the number of factors. The scree plot indicates 2 or 3 factors.
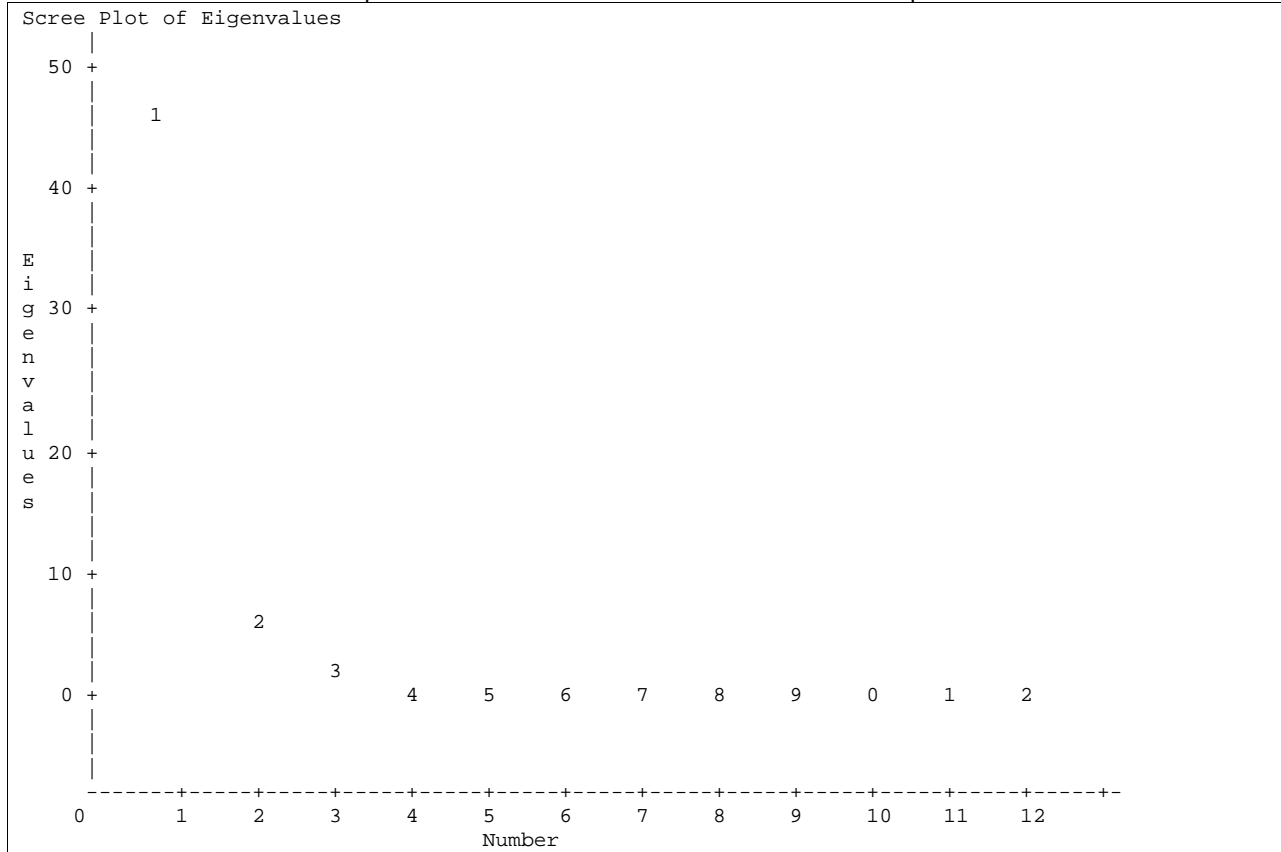
```
Scree Plot of Eigenvalues
          |
   50 +
          |
          |      1
          |
          |
          |
   40 +
          |
          |
E        |
i        |
g 30 +
e        |
n        |
v        |
a        |
l        |
u 20 +
e        |
s        |
          |
          |
   10 +
          |
          |            2
          |
          |          3
    0 +                    4     5     6     7     8     9     0     1     2
          |
          |
          |
      -------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-
       0     1     2     3     4     5     6     7     8     9    10    11    12
                                    Number
```

Hypothesis tests are both rejected, no common factors and 3 factors are sufficient.
In practice, we want to reject the first hypotheses and accept the second hypothesis.
Tucker and Lewis's Reliability Coefficient indicates good reliability. Reliability is a value between 0 and 1 with a
larger value indicating better reliability.

```
           Significance Tests Based on 995 Observations

                                                        Pr >
               Test                      DF    Chi-Square    ChiSq
H0: No common factors                    66    13065.6754    <.0001
HA: At least one common factor
H0: 3 Factors are sufficient             33      406.9236    <.0001
HA: More factors are needed


Chi-Square without Bartlett's Correction        409.74040
Akaike's Information Criterion                   343.74040
Schwarz's Bayesian Criterion                    181.94989
Tucker and Lewis's Reliability Coefficient        0.94247
```

Squared multiple correlations indicate amount of variance explained by each factor.

```
      Squared Canonical Correlations
   Factor1          Factor2         Factor3
0.98424906      0.88276725      0.65836583
```

Eigenvalues of the weighted reduced correlation matrix are 62.4882752, 7.5300396, and 1.9271076.
Proportion of variance explained is 0.8686, 0.1047, and 0.268.
Cumulative variance for 3 factors is 100%.

```
Eigenvalues of the Weighted Reduced Correlation Matrix:
Total = 71.9454163  Average = 5.99545136


       Eigenvalue    Difference    Proportion    Cumulative
    1   62.4882752    54.9582356        0.8686        0.8686
    2    7.5300396     5.6029319        0.1047        0.9732
    3    1.9271076     1.2700613        0.0268        1.0000
    4    0.6570464     0.4697879        0.0091        1.0091
    5    0.1872584     0.0202533        0.0026        1.0117
    6    0.1670051     0.0724117        0.0023        1.0141
    7    0.0945935     0.1415581        0.0013        1.0154
    8   -0.0469646     0.1141100       -0.0007        1.0147
    9   -0.1610746     0.0355226       -0.0022        1.0125
   10   -0.1965972     0.0469645       -0.0027        1.0097
   11   -0.2435617     0.2141498       -0.0034        1.0064
   12   -0.4577115                     -0.0064        1.0000
```

EFA with FACTOR and rotation
SAS Code – 3 factor model
```
proc factor data=rawfl method=ml rotate=v n=3 reorder plot out=facsub3 priors=smc;
   var math88 math90 math92 math94 readr88 readr90 readr92 readr94
       readc88 readc90 readc92 readc94;
```

Options added to the factor procedure are
- varimax rotation (orthogonal)
- reorder to arrange the pattern matrix from largest to smallest loading for each factor
- plot to plot factors
- n=3 to keep 3 factors
- out=facsub3 to save original data and factor scores

Results for the 3 factor model
Preliminary Eigenvalues – same as above
Significance Tests – same as above
Squared Canonical Correlations – same as above
Eigenvalues of the Weighted Reduced Correlation Matrix – same as above

Factor loadings illustrate correlations between items and factors. The REORDER option arranges factors loadings by factor from largest to smallest value.

```
The FACTOR Procedure
Rotation Method: Varimax


             Rotated Factor Pattern
             Factor1         Factor2         Factor3
readr92      0.79710         0.37288         0.31950
readr94      0.79624         0.26741         0.32138
readr90      0.67849         0.54244         0.31062
readc92      0.67528         0.32076         0.41293
readc90      0.61446         0.50233         0.32608
readc94      0.58909         0.21823         0.44571
- - - - - - - - - - - - - -
readr88      0.35420         0.88731         0.22191
readc88      0.33219         0.87255         0.22813
math88       0.24037         0.70172         0.43051
- - - - - -         - - - - - - - - - - - - - - -
math92       0.39578         0.31687         0.69707
math94       0.39013         0.22527         0.67765
math90       0.35250         0.51487         0.59444
```

Factor scores could be calculated by weighting each variable with the values from the rotated factor pattern matrix..
In common practice, factor scores are calculated without weights. A factor is calculated by using the mean or sum of
variables that load, are highly correlated with the factor. Factor scores could be calculated with a mean as illustrated
below.

Factor1 = mean(readr92, readr94, readr90, readc92, readc90, readc94);
Factor2 = mean(readr88, readc88, math88);
Factor3 = mean(math92, math94, math90);

Interpretability
Is there some conceptual meaning for each factor? Could the factors be given a name?
Factor1 could be called reading achievement.
Factor2 could be called basic skills achievement (math, reading recognition, reading comprehension).
Factor3 could be call math achievement.

Caution - The bolded loadings, larger than 0.50, are correlated with more than one factor

```
The FACTOR Procedure
Rotation Method: Varimax


             Rotated Factor Pattern
             Factor1         Factor2         Factor3
readr92       0.79710         0.37288         0.31950
readr94       0.79624         0.26741         0.32138
readr90       0.67849         0.54244         0.31062
readc92       0.67528         0.32076         0.41293
readc90       0.61446         0.50233         0.32608
readc94       0.58909         0.21823         0.44571
- - - - - - - - - - - - - - - - - - - - - -
readr88       0.35420         0.88731         0.22191
readc88       0.33219         0.87255         0.22813
math88        0.24037         0.70172         0.43051
- - - - - -                 - - - - - - - - - - - - - - - -
math92        0.39578         0.31687         0.69707
math94        0.39013         0.22527         0.67765
math90        0.35250         0.51487         0.59444
```

SAS Code – 2 factor model
```
proc factor data=rawfl method=ml priors=smc n=2 rotate=v reorder;
   var math88 math90 math92 math94
       readr88 readr90 readr92 readr94
       readc88 readc90 readc92 readc94;
```

(Option n=2 extracts 2 factors)

Results for the 2 factor model
Hypothesis tests are both rejected, no common factors and 3 factors are sufficient.
In practice, we want to reject the first hypotheses and accept the second hypothesis.
Tucker and Lewis's Reliability Coefficient indicates good reliability. Reliability is a value between 0 and 1 with a
larger value indicating better reliability.

```
The FACTOR Procedure
Initial Factor Method: Maximum Likelihood
Preliminary Eigenvalues – same as above



        Significance Tests Based on 995 Observations
                                                  Pr >
            Test                    DF    Chi-Square    ChiSq
H0: No common factors               66    13065.6754    <.0001
HA: At least one common factor
H0: 2 Factors are sufficient        43      947.6256    <.0001
HA: More factors are needed
```

```
Chi-Square without Bartlett's Correction        953.54123
Akaike's Information Criterion                  967.54123
Schwarz's Bayesian Criterion                    656.72329
Tucker and Lewis's Reliability Coefficient        0.89319
```

Squared multiple correlations indicate amount of variance explained by each factor.

```
Squared Canonical Correlations
    Factor1          Factor2
0.98290118      0.86924470
```

Eigenvalues of the weighted reduced correlation matrix are 57.4835709 and 6.6478737.
Proportion of variance explained is 0.8963 and 0.1037.
Cumulative variance for 2 factors is 100%.

```
Eigenvalues of the Weighted Reduced Correlation Matrix:
  Total = 64.1314404   Average = 5.3442867
       Eigenvalue    Difference    Proportion    Cumulative
    1   57.4835709   50.8356972        0.8963        0.8963
    2    6.6478737    5.5346312        0.1037        1.0000
    3    1.1132425    0.5742044        0.0174        1.0174
    4    0.5390381    0.4252208        0.0084        1.0258
    5    0.1138173    0.0915474        0.0018        1.0275
    6    0.0222699    0.0738907        0.0003        1.0279
    7   -0.0516208    0.1021049       -0.0008        1.0271
    8   -0.1537257    0.1222701       -0.0024        1.0247
    9   -0.2759958    0.1000048       -0.0043        1.0204
   10   -0.3760006    0.0716077       -0.0059        1.0145
   11   -0.4476083    0.0358123       -0.0070        1.0075
   12   -0.4834207                    -0.0075        1.0000
```

Factor loadings illustrate correlations between items and factors. The REORDER option arranges factors loadings by factor from largest to smallest value.

```
Rotated Factor Pattern
                Factor1          Factor2
readr94         0.82872          0.31507
readr92         0.81774          0.42138
readc92         0.77562          0.36634
readc94         0.72145          0.26511
readr90         0.71064          0.58111
readc90         0.67074          0.53838
math92          0.65248          0.38255
math94          0.64504          0.29031
math90          0.56547          0.56323
- - - - - - - - - - - - - - - - - - - - - - - -
readr88         0.36410          0.90935
readc88         0.34735          0.89771
math88          0.38175          0.73216
```

Factor scores could be calculated by weighting each variable with the values from the rotated factor pattern matrix..
In common practice, factor scores are calculated without weights. A factor is calculated by using the mean or sum of variables that load, are highly correlated with the factor. Factor scores could be calculated with a mean as illustrated below.

Factor1 = mean(readr92, readr94, readr90, readc92, readc90, readc94 math92, math94, math90);
Factor2 = mean(readr88, readc88, math88);

Interpretability
Is there some conceptual meaning for each factor? Could the factors be given a name?
Factor1 could be called reading and math achievement.
Factor2 could be called basic skills achievement (math, reading recognition, reading comprehension).

Caution - The bolded loadings, larger than 0.50, are correlated with more than one factor

```
Rotated Factor Pattern
               Factor1         Factor2
readr94        0.82872          0.31507
readr92        0.81774          0.42138
readc92        0.77562          0.36634
readc94        0.72145          0.26511
readr90        0.71064          0.58111
readc90        0.67074          0.53838
math92         0.65248          0.38255
math94         0.64504          0.29031
math90         0.56547          0.56323
- - - - - - - - - - - - - - - - - - - - - - - -
readr88        0.36410          0.90935
readc88        0.34735          0.89771
math88         0.38175          0.73216
```

2 or 3 Factor Model?

Reliability and interpretability plays a role in your decision of the factor structure. Reliability was determined for each factor using PROC CORR with options ALPHA NOCORR.

3 factor model reliabilities
- Basic Skills Factor is 0.943
- Math Factor is 0.867
- Reading Factor is 0.945

2 factor model reliabilities
- Basic Skills Factor is 0.943
- Math/Reading Factor is 0.949

Both models exhibit good reliability. Now we rely on interpretability. Which model would you select to represent the factor structure, a 2 or 3 factor model?

Conclusion

Confirmatory and Exploratory Factor Analysis are powerful statistical techniques. The techniques have similarities and differences. Determine the type of analysis a priori to answer research questions and maximize your knowledge.

WAM (Walk away message)

Select CFA to verify the factor structure and EFA to determine the factor structure.

References

Baker, P. C., Keck, C. K., Mott, F. L. & Quinlan, S. V. (1993). NLSY79 child handbook. A guide to the 1986-1990 National Longitudinal Survey of Youth Child Data (rev. ed.) Columbus, OH: Center for Human Resource Research, Ohio State University.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, 1, 245-276.

Child, D. (1990). The essentials of factor analysis, second edition. London: Cassel Educational Limited.

DeVellis, R. F. (1991). Scale Development: Theory and Applications. Newbury Park, California: Sage Publications.

Dunn, L. M. & Markwardt, J. C. (1970). Peabody individual achievement test manual. Circle Pines, MN: American Guidance Service.

Hahn, M. E. (1966). *California Life Goals Evaluation Schedule.* Palo Alto, CA: Western Psychological Services.

Hatcher, L. (1994). A step-by-step approach to using the SAS[®] System for factor analysis and structural equation modeling. Cary, NC: SAS Institute Inc.

Hoyle, R. H. (1995). The structural equation modeling approach: Basic concepts and fundamental issues. In Structural equation modeling: Concepts, issues, and applications, R. H. Hoyle (editor). Thousand Oaks, CA: Sage Publications, Inc., pp. 1-15.

Hu, L. & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6(1), 1-55.

Joreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis, *Psychometrika,* 34, 183-202.

Kline, R. B. (1998). *Principles and Practice of Structural Equation Modeling.* New York: The Guilford Press.

Maddi, S. R., Kobasa, S. C., & Hoover, M. (1979) An alienation test. *Journal of Humanistic Psychology,* 19, 73-76.

Nunnally, J. C. (1978). Psychometric theory, 2nd edition. New York: McGraw-Hill.

Roth, D. L., & Fillingim, R. B. (1988). *Assessment of self-reported exercies participation and self-perceived fitness levels.* Unpublished manuscript.

Roth, D. L., Wiebe, D. J., Fillingim, R. B., & Shay, K. A. (1989). "Life Events, Fitness, Hardiness, and Health: A Simultaneous Analysis of Proposed Stress-Resistance Effects". *Journal of Personality and Social Psychology,* 57(1), 136-142.

Rotter, J. B., Seeman, M., & Liverant, S. (1962). Internal vs. external locus of control: A major variable in behavior theory. In N. F. Washburne (Ed.), *Decisions, values, and groups* (pp. 473-516). Oxford, England: Pergamon Press.

Sarason, I. G., Johnson, J. H., & Siegel, J. M. (1978) Assessing the impact of life changes: Development of the Life Experiences Survey. *Journal of Consulting and Clinical Psychology,* 46, 932-946.

SAS$^{®}$ Language and Procedures, Version 6, First Edition. Cary, N.C.: SAS Institute, 1989.

SAS$^{®}$ Online Doc 9. Cary, N.C.: SAS Institute.

SAS$^{®}$ Procedures, Version 6, Third Edition. Cary, N.C.: SAS Institute, 1990.

SAS/STAT$^{®}$ User's Guide, Version 6, Fourth Edition, Volume 1. Cary, N.C.: SAS Institute, 1990.

SAS/STAT$^{®}$ User's Guide, Version 6, Fourth Edition, Volume 2. Cary, N.C.: SAS Institute, 1990.

Schumacker, R. E. & Lomax, R. G. (1996). *A Beginner's Guide to Structural Equation Modeling.* Mahwah, New Jersey: Lawrence Erlbaum Associates, Publishers.

Thorndike, R. M., Cunningham, G. K., Thorndike, R. L., & Hagen E. P. (1991). Measurement and evaluation in psychology and education. New York: Macmillan Publishing Company.

Truxillo, Catherine. (2003). Multivariate Statistical Methods: Practical Research Applications Course Notes. Cary, N.C.: SAS Institute.

Wyler, A. R., Masuda, M., & Holmes, T. H. (1968). Seriousness of Illness Rating Scale. *Journal of Psychosomatic Research,* 11, 363-374.

Contact Information

Diana Suhr is a Statistical Analyst in Institutional Research at the University of Northern Colorado. In 1999 she earned a Ph.D. in Educational Psychology at UNC. She has been a SAS programmer since 1984.
email: diana.suhr@unco.edu

Appendix A - Definitions

An **observed variable** can be measured directly, is sometimes called a **measured variable** or an **indicator** or a **manifest variable**.

A **latent construct** can be measured indirectly by determining its influence to responses on measured variables. A latent construct is also referred to as a **factor**, **underlying construct**, or **unobserved variable**.

Unique factors refer to unreliability due to measurement error and variation in the data. CFA specifies unique factors explicitly.

Factor scores are estimates of underlying latent constructs.

Eigenvalues indicate the amount of variance explained by each principal component or each factor.

Orthogonal, axis held at a 90 degree angle, perpendicular.

Obilque, axis other than a 90 degree angle.
An observed variable "**loads**" on a factors if it is highly correlated with the factor, has an eigenvector of greater magnitude on that factor.

Communality is the variance in observed variables accounted for by a common factors. Communality is more relevant to EFA (Hatcher, 1994).