Multidimensional Model Difficulties

Multidimensional Measurement Models (Fall 2023): Lecture 5

## Today’s Lecture

* Difficulties in Estimating and Using Multidimensional Measurement Models
  + Model statistical identification (Q-matrix Definitions)
  + Empirical identification
  + Differing definitions of dimensionality
  + Quantifying the number of dimensions
  + Latent variable indeterminacy
  + Estimation of models

## Statistical Identification

Statistical identification refers to the ability to estimate the parameters of a model uniquely

* Often, this becomes a system of equations where the number of parameters is less than the number of equations
* Much of the literature on statistical identification is based on the idea comes from confirmatory factor analysis (CFA) models
  + Although more difficult to show for IRT models, the CFA standard is often used
* Statistical identification is a necessary condition for model estimation
  + But, it is not a sufficient condition (see empirical identification)

## Two Types of Statistical Identification

1. Location and scale identification
2. Minimum numbers of items needed per latent variable

## Location and Scale Identification

* How the mean and standard deviation for each latent variable are determined
  + Generally, one constraint is needed for the mean and one for the standard deviation for each latent variable
  + The constraints can be to specify the mean and standard deviation for one latent variable (i.e., standardized latent variables)
  + Other constraints include setting marker items for the mean and standard deviation
    - Marker items for the mean are specified by placing a constraint on item intercepts
    - Marker items for the standard deviation are specified by placing a constraint on item discrimination parameters

## Minimum Number of Items Needed per Latent Variable

* For a single latent variable model (i.e., unidimensional model), the minimum number of items needed is three
* For a model with two latent variables and an estimated factor correlation, the minimum number of items needed is five
  + The Q-matrix must be specified so that one latent variable has at least three items and the other latent variable has at least two items
  + Note that this only applies if items measure only one latent variable

## Where do these numbers come from?

Consider the model-implied distribution in CFA model with three items measuring one latent variable:

Where:

* is the vector of item intercepts (size )
* is a column vector of factor loadings (size )
* is the vector of latent variable means (size )
* is the covariance matrix of the latent variable (i.e., the variance; size )
* is the (diagonal) covariance matrix of the residuals (size )

## Model-implied Mean Vector

For a unidimensional model with three items, the model-implied mean vector is:

From the data, we know there are three means (one per item):

So, without any constraints, we have seven unknown parameters but only three equations

* An underidentifed model

## Adding Constraints

If we set the factor mean to zero, we have the following:

Now, we have three unknown parameters and three equations

* A just identified model with
* But, what about if we wanted to estimate ?
  + We will get to that after looking at the model-implied covariance matrix

## Model-implied Covariance Matrix

For a unidimensional model with three items, the model-implied covariance matrix is:

From the data, we know there are three variances (one per item):

So, without any constraints, we have seven unknown parameters but only six equations

* An underidentified model

## Adding Constraints

If we set the factor variance to one, we have the following:

## Adding Constraints

Now, we have six unknown parameters and six equations

## Determining Parameters

With some algebra, we can determine the equations for each of the parameters

* The factor loadings come from the off-diagonal elements of the covariance matrix

The unique variances come from the diagonal elements of the covariance matrix along with the factor loadings

## Non-Standardized Factors (Mean)

With the result of the factor loadings, we can now estimate the factor mean with one additional constraint:

Let’s set the item intercept of the first item to zero:

## Non-Standardized Factors (Mean)

We now see that:

## Non-Standardized Factors (Variance)

With the result of the factor loadings, we can now estimate the factor variance with one additional constraint:

Let’s set the item discrimination of the first item to one:

* A similar derivation can be done for the other parameters

## My Derivation (90% Confidence)

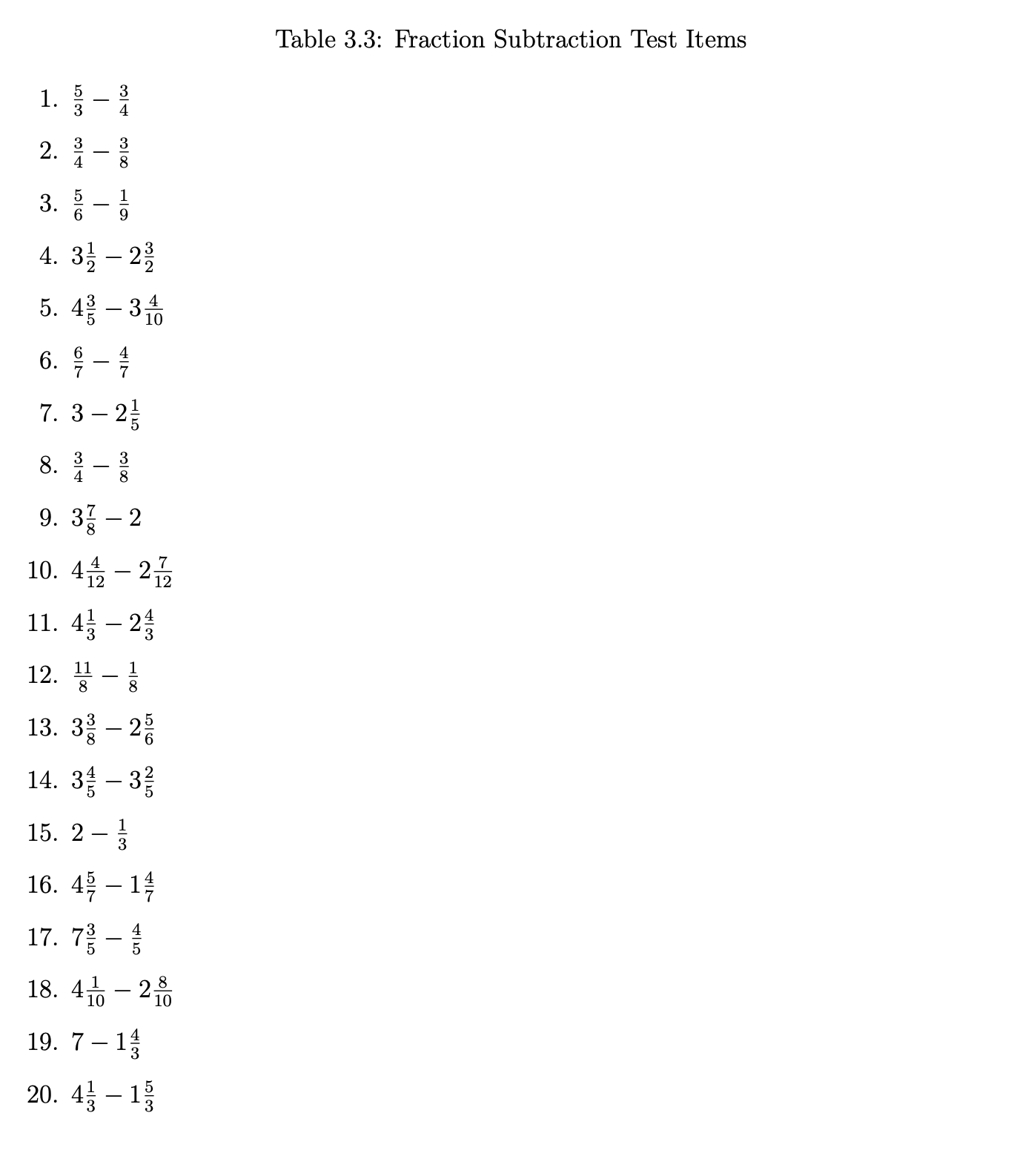
From my algebra, I get the following:

## Statistical Identification and the Q-matrix

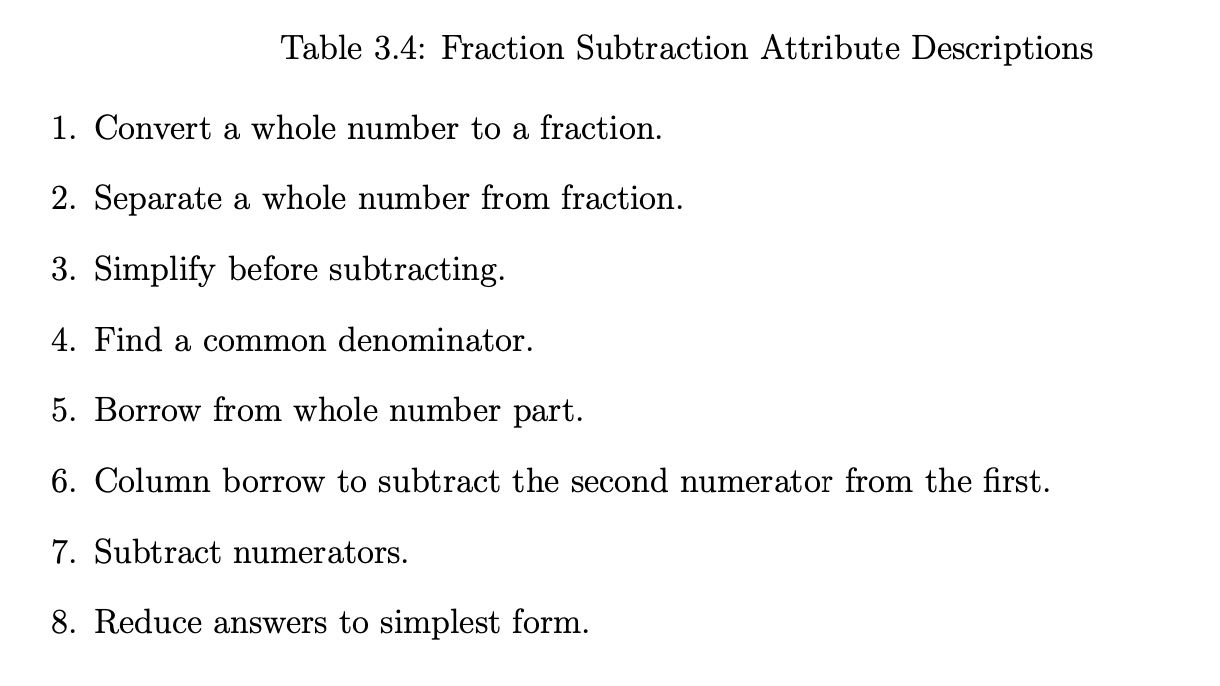
Q-matrices specify the alignment of items to latent variables

* But, not all Q-matrices are statistically identified
  + Our previous example used the most basic Q-matrix, a column vector of ones
* In general, Q-matrices need:
  + Three items per latent variable
  + At least some items that are unique to each latent variable

## Example Multidimensional Assessment

Tatsuoka’s fraction subtraction data: 

## Example Latent Variables

Tatsuoaka’s fraction subtraction data latent variables: 

## Example Q-matrices

* From the Tatsuoka Fraction Subtraction Q-matrix:

alpha1 alpha2 alpha3 alpha4 alpha5 alpha6 alpha7 alpha8  
Item1 0 0 0 1 0 1 1 0  
Item2 0 0 0 1 0 0 1 0  
Item3 0 0 0 1 0 0 1 0  
Item4 0 1 1 0 1 0 1 0  
Item5 0 1 0 1 0 0 1 1  
Item6 0 0 0 0 0 0 1 0  
Item7 1 1 0 0 0 0 1 0  
Item8 0 0 0 0 0 0 1 0  
Item9 0 1 0 0 0 0 0 0  
Item10 0 1 0 0 1 0 1 1  
Item11 0 1 0 0 1 0 1 0  
Item12 0 0 0 0 0 0 1 1  
Item13 0 1 0 1 1 0 1 0  
Item14 0 1 0 0 0 0 1 0  
Item15 1 0 0 0 0 0 1 0  
Item16 0 1 0 0 0 0 1 0  
Item17 0 1 0 0 1 0 1 0  
Item18 0 1 0 0 1 1 1 0  
Item19 1 1 1 0 1 0 1 0  
Item20 0 1 1 0 1 0 1 0

[1] "matrix" "array"

## More Examples

From our coding activity two weeks ago, the Q-matrix was:

theta1 theta2  
item1 1 0  
item2 1 0  
item3 1 0  
item4 1 0  
item5 1 0  
item6 0 1  
item7 0 1  
item8 0 1  
item9 0 1  
item10 0 1

This is often called a “simple structure” Q-matrix, although that term has been used in other places

* Within-item unidimensionality is another term that has been used

## Q-Matrix Math

Apart from showing the alignment, matrix operations on the Q-matrix can be helpful to determine if a Q-matrix is statistically identified

* To see the number of items measuring each latent variable and each pair of latent variables, use:

t(FSQmatrix) %\*% FSQmatrix

alpha1 alpha2 alpha3 alpha4 alpha5 alpha6 alpha7 alpha8  
alpha1 3 2 1 0 1 0 3 0  
alpha2 2 13 3 2 8 1 12 2  
alpha3 1 3 3 0 3 0 3 0  
alpha4 0 2 0 5 1 1 5 1  
alpha5 1 8 3 1 8 1 8 1  
alpha6 0 1 0 1 1 2 2 0  
alpha7 3 12 3 5 8 2 19 3  
alpha8 0 2 0 1 1 0 3 3

* Diagonal elements: Number of items measuring each latent variable
* Off-diagonal elements: Number of items measuring each pair of latent variables

## More Q-Matrix Math

Additionally, we can use the Q-matrix to build a quick sum-score for each latent variable:

Y = as.matrix(FSdata)  
sumScores = Y %\*% FSQmatrix  
head(sumScores)

alpha1 alpha2 alpha3 alpha4 alpha5 alpha6 alpha7 alpha8  
[1,] 3 9 3 0 6 1 12 2  
[2,] 3 12 3 3 8 1 17 2  
[3,] 2 4 1 2 1 0 9 0  
[4,] 0 8 2 4 5 2 14 3  
[5,] 0 1 1 0 1 0 4 1  
[6,] 0 2 0 0 0 0 3 0

## Sum Scores

Although sum scores aren’t what we want to use (way too many assumptions), they can help show how correlated traits may be:

cor(sumScores)

alpha1 alpha2 alpha3 alpha4 alpha5 alpha6 alpha7  
alpha1 1.0000000 0.8345775 0.8037195 0.6839375 0.7911124 0.6484896 0.8434539  
alpha2 0.8345775 1.0000000 0.8782167 0.7970944 0.9577298 0.7896634 0.9691487  
alpha3 0.8037195 0.8782167 1.0000000 0.6200224 0.9325677 0.6353779 0.8448615  
alpha4 0.6839375 0.7970944 0.6200224 1.0000000 0.7178911 0.8237720 0.8751911  
alpha5 0.7911124 0.9577298 0.9325677 0.7178911 1.0000000 0.7593760 0.9204327  
alpha6 0.6484896 0.7896634 0.6353779 0.8237720 0.7593760 1.0000000 0.8470215  
alpha7 0.8434539 0.9691487 0.8448615 0.8751911 0.9204327 0.8470215 1.0000000  
alpha8 0.6448976 0.8325724 0.6341646 0.7807335 0.7461302 0.6669306 0.8444153  
 alpha8  
alpha1 0.6448976  
alpha2 0.8325724  
alpha3 0.6341646  
alpha4 0.7807335  
alpha5 0.7461302  
alpha6 0.6669306  
alpha7 0.8444153  
alpha8 1.0000000

## Q-Matrix Limits

If exploratory analyses are to be used (and most times, they shouldn’t be used), then a very specific form of the Q-matrix indicates that the model is statistically identified

* Lower echelon form: The largest number of latent variables that can be statistically identified (with orthogonal/uncorrleated latent variables)
  + Zeroes at the top-right of the Q-matrix
    - First item measures only first latent variable
    - Second item measures both first and second latent variable

## Lower Echelon Form

The fraction subtraction Q-matrix in lower echelon form:

FSQmatrixLE = matrix(data = 1, nrow = nrow(FSQmatrix), ncol = ncol(FSQmatrix))  
for (i in 1:(ncol(FSQmatrixLE)-1)){  
 FSQmatrixLE[i,(i+1):ncol(FSQmatrixLE)] = 0  
}  
  
FSQmatrixLE

[,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8]  
 [1,] 1 0 0 0 0 0 0 0  
 [2,] 1 1 0 0 0 0 0 0  
 [3,] 1 1 1 0 0 0 0 0  
 [4,] 1 1 1 1 0 0 0 0  
 [5,] 1 1 1 1 1 0 0 0  
 [6,] 1 1 1 1 1 1 0 0  
 [7,] 1 1 1 1 1 1 1 0  
 [8,] 1 1 1 1 1 1 1 1  
 [9,] 1 1 1 1 1 1 1 1  
[10,] 1 1 1 1 1 1 1 1  
[11,] 1 1 1 1 1 1 1 1  
[12,] 1 1 1 1 1 1 1 1  
[13,] 1 1 1 1 1 1 1 1  
[14,] 1 1 1 1 1 1 1 1  
[15,] 1 1 1 1 1 1 1 1  
[16,] 1 1 1 1 1 1 1 1  
[17,] 1 1 1 1 1 1 1 1  
[18,] 1 1 1 1 1 1 1 1  
[19,] 1 1 1 1 1 1 1 1  
[20,] 1 1 1 1 1 1 1 1

## More Exploratory Q-Matrix Topics

* If factors need correlations, more zeros need to be added to the lower-echelon Q-matrix
* Moreover, the saturation of the Q-matrix with ones (up to lower-echelon form) is one way to use confirmatory methods for exploratory analyses
  + For instance, if some items are well known to measure some latent variables, the remainder can load onto all other latent variables

## Lower Echelon Form vs. Exploratory Analyses

* The lower echelon form is just-identified
  + The zeros show the number of constraints needed for statistical identification
* Exploratory analyses often use different constraints where it appears all loadings are estimated
  + These loadings are constrained–and the number of constraints is equal to the number of zeros in a lower echelon form Q-matrix
* One common constraint is to estimate the model such that:

Where is a diagonal matrix (see Lawley and Maxwell, 1971)

Lawley, D. N., & Maxwell, A. E. (1971). Factor analysis as a statistical method.

## Empirical Identification

Empirical identification is the ability for the data to support an estimable multidimensional model

* Essentially, the latent variables are estimable
* In practice this tends to mean:
  + The latent variable correlations are less than one (and have a positive semi-definite covariance matrix)
  + Any marker items (for loadings) have non-zero loading estimates

## Example Empirical Underidentification

nItems = 10  
nExaminees = 1000  
nFactors = 1  
  
Qmatrix = matrix(data = 0, nrow = nItems, ncol = 2)  
  
Qmatrix[1:10,1] = 1  
Qmatrix

[,1] [,2]  
 [1,] 1 0  
 [2,] 1 0  
 [3,] 1 0  
 [4,] 1 0  
 [5,] 1 0  
 [6,] 1 0  
 [7,] 1 0  
 [8,] 1 0  
 [9,] 1 0  
[10,] 1 0

# generate item intercepts  
mu = rnorm(n = nItems, mean = 0, sd = 1)  
  
# generate item slopes  
lambda = rlnorm(n = nItems, mean = 0, sd = 1)  
  
# uncorrelated univariate normal  
thetaZ = matrix(data = 0, nrow = nExaminees, ncol = nFactors)  
  
theta = thetaZ  
  
# generate data   
dataMat = matrix(data = NA, nrow = nExaminees, ncol = nItems)  
colnames(dataMat) = paste0("item", 1:nItems)  
for (item in 1:nItems){  
 logit = mu[item] + Qmatrix[item,1]\*lambda[item]\*theta[,1]  
 prob = exp(logit)/(1+exp(logit))  
 dataMat[,item] = rbinom(n = nExaminees, size = 1, prob = prob)  
}  
  
if (!require(mirt)) install.packages("mirt")  
library(mirt)  
  
Qmatrix2D = matrix(data = 0, nrow = nItems, ncol = 2)  
colnames(Qmatrix2D) = paste0("theta", 1:2)  
rownames(Qmatrix2D) = paste0("item", 1:nItems)  
  
Qmatrix2D[1:5,1] = 1  
Qmatrix2D[6:10,2] = 1  
  
COV = matrix(c(FALSE, TRUE, TRUE, FALSE), nrow = 2, ncol = 2)  
  
underIdentifiedModel = mirt(data = dataMat, model = mirt.model(input=Qmatrix2D, COV=COV), itemtype = "2PL", SE = TRUE, verbose=FALSE)  
coef(underIdentifiedModel)$GroupPars

MEAN\_1 MEAN\_2 COV\_11 COV\_21 COV\_22  
par 0 0 1 -0.48212415 1  
CI\_2.5 NA NA NA -1.03327203 NA  
CI\_97.5 NA NA NA 0.06902374 NA

underIdentifiedModel

Call:  
mirt(data = dataMat, model = mirt.model(input = Qmatrix2D, COV = COV),   
 itemtype = "2PL", SE = TRUE, verbose = FALSE)  
  
Full-information item factor analysis with 2 factor(s).  
Converged within 1e-04 tolerance after 246 EM iterations.  
mirt version: 1.38.1   
M-step optimizer: BFGS   
EM acceleration: Ramsay   
Number of rectangular quadrature: 31  
Latent density type: Gaussian   
  
Information matrix estimated with method: Oakes  
Second-order test: model is a possible local maximum  
Condition number of information matrix = 1787.844  
  
Log-likelihood = -6256.835  
Estimated parameters: 21   
AIC = 12555.67  
BIC = 12658.73; SABIC = 12592.04  
G2 (1002) = 890.19, p = 0.9951  
RMSEA = 0, CFI = NaN, TLI = NaN

## Empirical Underidentification Indicators

* Often, empirical underidentification presents itself in model results
  + (Maximum likelihood) algorithms that do not converge within default specifications
  + (Bayesian) posterior distributions lack convergence
  + Standard errors that are very large
  + Correlations that are very large
  + NaNs for model fit

## Differing Definitions of Dimensionality

Depending on the class of models used, the number of dimensions can be defined differently

* For instance, in CFA models, the number of dimensions is the number of latent variables
* In latent class models, the number of dimensions can be:
  + The number of classes -or-
  + The number of categorical variables
  + Also, Steinley and McDonald (2007) show how the number of latent classes plus one can be equivalent to the number of latent variables in a CFA model
* DCMs are latent class models, so the number of dimensions can be defined in multiple ways, also
  + The number of attributes (i.e., the number of latent variables)
  + The number of classes
* All of this is independent of models with method factors or other random effects
  + Each of which is a dimension

Steinley, D., & McDonald, R. P. (2007). Examining factor score distributions to determine the nature of latent spaces. Multivariate Behavioral Research, 42(1), 133-156.

## Quantifying the Number of Dimensions

In exploratory analyses, it can be difficult to easily determine the number of dimensions

* Matrix algebra-based methods (Principal components or other older methods) have highly unstable estimates of the number of dimensions
  + The number of dimensions can be estimated from the eigenvalues of the correlation matrix
    - But, this is not always reliable
* Likelihood-based exploratory methods put a set of constraints on the model for statistical identification (later class)
  + Easier to specify confirmatory models
* No uniform agreement on what level of correlation between latent variables indicates too high (so can be combined)

## Latent Variable Indeterminacy

Latent variable indeterminacy is the ability to rotate the latent variables without changing the model

* Often, this is a problem in exploratory analyses
  + Without knowledge of what your latent variables are, rotations are one way to try to derive meaning
* In confirmatory analyses, latent variables are constructed so they have meaning
  + Rotations are not needed–but still could be applied:

For instance, a model where the latent variables are correlated:

## Example Rotations (Confirmatory Model)

We can show that a model where:

And

Is equivalent to:

When:

Where is the lower triangle of the Cholesky decomposition of

## Practical Implications of Rotational Indeterinancy

Rotational indeterminancy is a problem if your latent variables lack a validity argument

* Exploratory analyses are the main problem

Although other solutions exist for confirmatory models, rotations aren’t needed as meaning is derived by construction of the items and the model

* So no need to rotate, philosophically

## Estimation of Models

Finally, estimation of models with some distributions can be difficult if not impossible

* CFA models (all items multivariate normal): No problems
  + Marginal distribution of data can be derived without numeric integration
* Any model that contains one or more items that are not multivariate normal:
  + Marginal ML methods must numerically integrate across the latent variables
  + Numeric integration requires a set of quadrature points (e.g., summation locations) per dimension
  + The total number of quadrature points is the product of the number of quadrature points per dimension
    - Exponential increases in the number of calculations needed per linear increase in latent variables

## Ramifications of Estimation Difficulties

Many estimation methods exist

* MML via Quasi-Newton or EM
  + EM via Gibbs sampling or Metropolis-Hastings (hybrid ML/MCMC; approximate estimates)
  + Numerical integration via sampling (approximate estimates)
  + Implications for model comparisons and replication
* Limited information methods
  + Use only tetra or polychoric correlations
  + Must specify constraints (Mplus’ delta vs. theta)
  + Does not fit full data
  + Implications for missing data
* Bayesian methods
  + Linear increase in calculations for linear increases in latent variables
  + Implications for model comparisons and replication

## Summary

* Statistical identification is a necessary condition for model estimation
  + But, it is not a sufficient condition (see empirical identification)
* Statistical identification is often the easiest difficulty to overcome
* Estimation difficulties are more often the problem
  + Especially for models with non-normal data
* Many long-running issues with rotational indeterminacy and with differing definitions of dimensionality
  + But, these are more philosophical than practical

## Discussion

Now, we will discuss Torres Irribarra & Arneson (2023)

* What is the argument about what should be considered when quantifying dimensions?
* Is this argument reasonable? Can you think of flaws in the argument?