On Test Scores (Part 1)

Structural Equation Modeling
Lecture #11

April 22, 2015

KU THE UNIVERSITY OF KANSAS

Today's Class

- Scores
  - Types of scores
    - Sum scores / test scores
    - Factor scores
  - Score contents
  - Relating sum scores to factor scores
  - Score reliability
- Why using scores alone in separate analysis, while done almost always, is not good practice

KU THE UNIVERSITY OF KANSAS

The Big Picture

- Overall, the purpose of this class and the main message of structural equation models is that multivariate analyses with (and without) measurement error should be conducted simultaneously
  - Error propagates

- There are many instances when one cannot do a simultaneous analysis
  - This lecture is an attempt to get you *as close to* results from a simultaneous analysis by getting you to understand the psychometric and statistical properties of using scores

KU THE UNIVERSITY OF KANSAS

WHAT'S IN A SUM SCORE?

The Purpose of this Lecture: Some Clarity on Score

- As I've been a student and a teacher I have found the topic of scores to be incomplete and often contradictory

- Some things I've heard:
  - "Sum scores are almost always okay"
  - "Factor scores (think GRE) are okay if they are from some strange sounding model…"
  - "…otherwise factor scores are the work of the devil"

- A question that I hearing: Why use Structural Equation Modeling (or CFA/IRT) when I can just use a sum of the items?
  - Sum of the items == sum score == total score == Add s**t up (ASU) model

- Sum score are used as:
  - Observed variables in secondary analyses
  - Results given to participants, patients, students, etc…

- Current practice in psychological/educational research seems to be:
  - Use a sum score until some reviewer (#3?) says you cannot use one
  - At that point, use a confirmatory factor model to verify that you have a one-factor scale
  - …then use a sum score

KU THE UNIVERSITY OF KANSAS

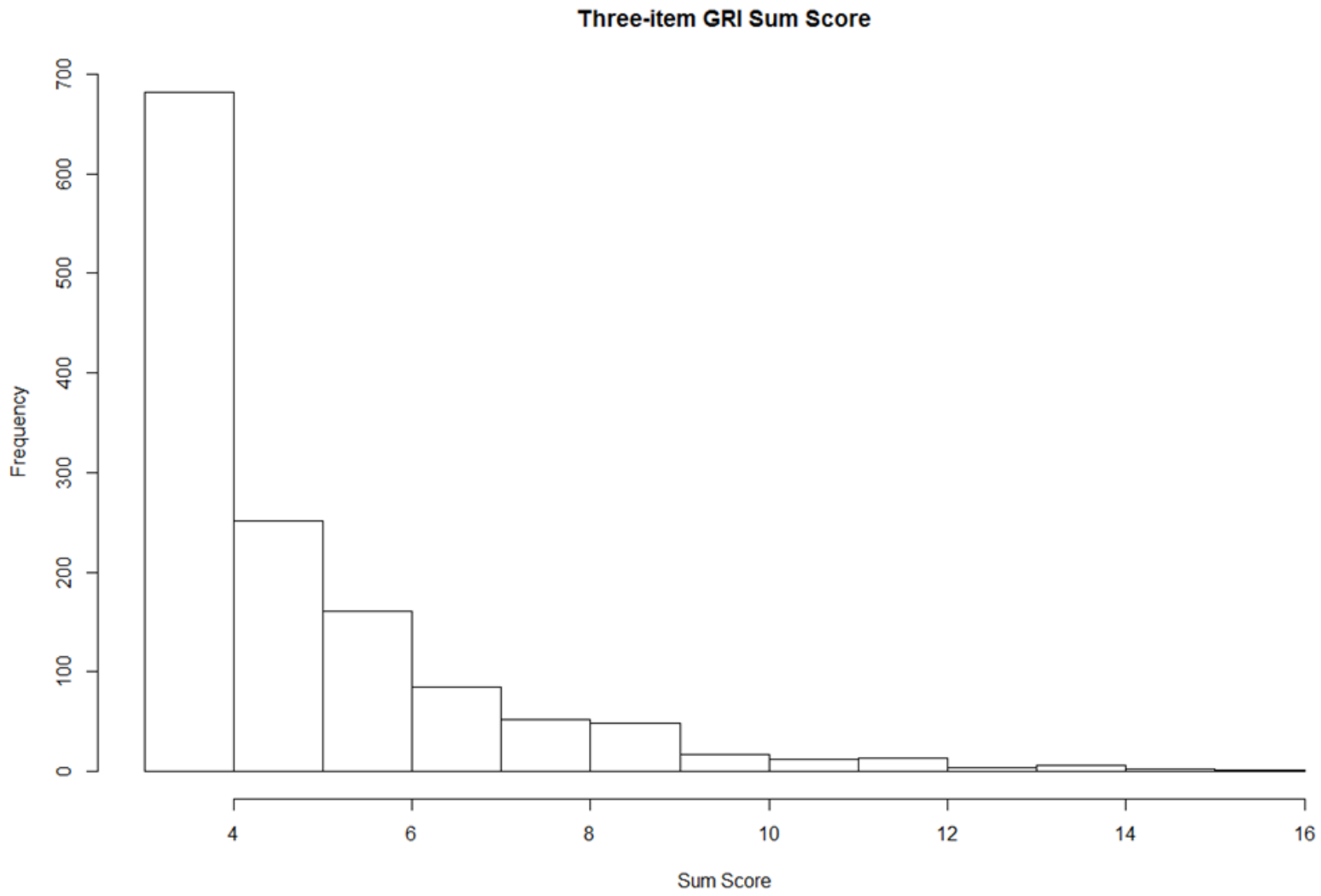
Demonstration Data

- To demonstrate the concepts appearing throughout this section, we will revisit the three-item GRI scale used in the lecture on Structural Equation Models
  - Items: GRI1, GRI3, and GRI 5

- As scores on each item ranged from 1 to 6 in integer units, this means sum scores must fall within a range of 3 to 18

KU THE UNIVERSITY OF KANSAS

Distribution of GRI Sum Scores



Psychometric Properties of as Sum Score

- The use of sum scores brings about a discussion about the psychometrics that underlie sum scores

- What you have learned about measurement so far likely falls under the category of CTT:
  - Writing items and building scales
  - Item analysis
  - Score interpretation
  - Evaluating reliability and construct validity

- Big picture: We will view CTT as model with a restrictive set of assumptions within a more general family of latent trait measurement models
  - Confirmatory Factor Analysis is a measurement model

Differences Among Measurement Models

- What is the **name of the latent trait** measured by a test?
  - Classical Test Theory (CTT) = "True Score" (T)
  - Confirmatory Factor Analysis (CFA) = "Factor Score" (F)
  - Item Response Theory (IRT) = "Theta" (θ)

- Fundamental difference in approach:
  - **CTT → unit of analysis is the WHOLE TEST** (item sum or mean)
    - **Sum = latent trait**, and the sum doesn't care how it was created
    - Only using the sum requires restrictive assumptions about the items
  - **CFA, IRT, and beyond → unit of analysis is the ITEM**
    - Model of how item response relates to an **estimated latent trait**
    - Different models for differing item response formats
    - Provides a framework for testing adequacy of measurement models

KU THE UNIVERSITY OF KANSAS

Classical Test Theory: Assumed Model

- In CTT, the TEST is the unit of analysis:

$$Y_{\text{Total}} = T + e$$

  - **True score T:**
    - Best estimate of 'latent trait': Mean over infinite replications
    - Scale of T is the same as the scale of Y_{Total}
  - **Error e:**
    - Expected value (mean) of 0, expected to be uncorrelated with T
    - Supposed to wash out over repeated observations

- **So the expected value of Y_{total} is T**
  - Put another way: should the model fit, Y_{total} is an unbiased estimate of T
  - The true score is why you created the sum in the first place→ your test purports to measure one thing, bringing about one sum score per person

- No distributional assumptions made…yet

- Even if your data fit a one-factor model, when using a sum score, the error portion is part of Y_{Total}
  - But, it is only one part of the error that is in a sum score

- Because the CTT model does not include individual items, items must be assumed exchangeable
  - If the model fits, then more items means better reliability

More CTT Basics

- A goal of CTT is to quantify reliability
  - Reliability is the proportion of variance in the sum score that is due to variation in the latent trait

- Reliability decomposition comes from Var(Y)
  - Var() function comes from the expected value in mathematical statistics
  - $E\big(g(x)\big) = \int g(x)f(x)dx$
    - Over the sample space/support of x with probability density function f(x)
    - Replace integral with a sum for discrete x (and pdf for probability mass function)
  - Mean: $\mu = E(x) = \int x\, f(x)dx$
  - Variance: $Var(x) = E\big((x-\mu)^2\big) = E\left[\big(x - E(x)\big)^2\right] = \int (x-\mu)^2\, f(x)dx$

- For CTT:

$$\mathrm{Var}(Y_{Total}) = \mathrm{Var}(T+e) = \mathrm{Var}(T) + Var(e) + 2Cov(T,e)$$

- But, $Cov(T,e) = 0$ as T and e are assumed independent, so

$$\mathrm{Var}(Y_{Total}) = \mathrm{Var}(T) + Var(e)$$

Moving from Variance to Reliability

- Reliability, as a proportion of variance in sum score due to the trait:

$$\rho = \frac{\text{Var}(T)}{\text{Var}(Y)} = \frac{\text{Var}(T)}{\text{Var}(T) + \text{Var}(e)}$$

  - $\text{Var}(Y)$ == variance of observed sum score
  - $\text{Var}(T)$ == variance of true score == variability in the unobserved latent trait == **individual differences**
  - $\text{Var}(e)$ == variance of error == **measurement error**

- Key question: how does one quantify reliability?
  - We will see that depends….

KU THE UNIVERSITY OF KANSAS

Draw Templin, Draw!

- Picture of distributions of Y, T, and e…

Parceling: Creating Another Type of Sum Score

- Another type of sum score is a parcel (sometimes called an item parcel or an item bundle)
  - A parcel then takes the places of the summed variables in a larger structural equation model

- There is some debate about what parceling assumes
  - There are some who believe a parcel assumes a CTT model:

$$Y_{\text{Total}} = T + e$$

  - There are others who parceling makes no assumptions, which is mathematically equivalent to:

$$Y_{\text{Total}} = e$$

- Either way:
  - What we are saying about CTT scores applies to parcels and parceling
  - Parceling is frequently done to hide model misfit, so it is like cheating

Potential Sources of Error in a Sum Score

- Measurement error
  - e.g., the e in $Y = T + e$

- Model misspecification error of various types:
  - Dimensionality misspecification error
    - e.g., Assuming one dimension when there is more than one present
  - Parameter constraint misspecification error
    - e.g., Assuming overly restrictive constraints (see next section and all of CTT)
  - Linear model functional misspecification error
    - e.g., Assuming a linear relationship between the factor and the items when a non-linear one is present
  - Outcome distribution misspecification error
    - e.g., Assuming Likert-type data to be continuous and using a normal distribution
  - Factor distribution misspecification error
    - e.g., Assuming your trait is normally distributed when it is categorical or a mixture distribution

- Missing data error
  - How you treat missing responses to items makes even more untenable assumptions

- ~~Sampling error~~
  - (meaning error in parameters due to small n) is **not a source of error** in a sum score
  - Note: measurement error is sampling error with respect to items instead of people

KU THE UNIVERSITY OF KANSAS

Why Error Matters

- Ignoring error will lead to inaccurate and potentially misleading results
  - Biased estimates (Type II error)
  - Biased standard errors of estimates (Type I error)

- Some sources of error matter more than others

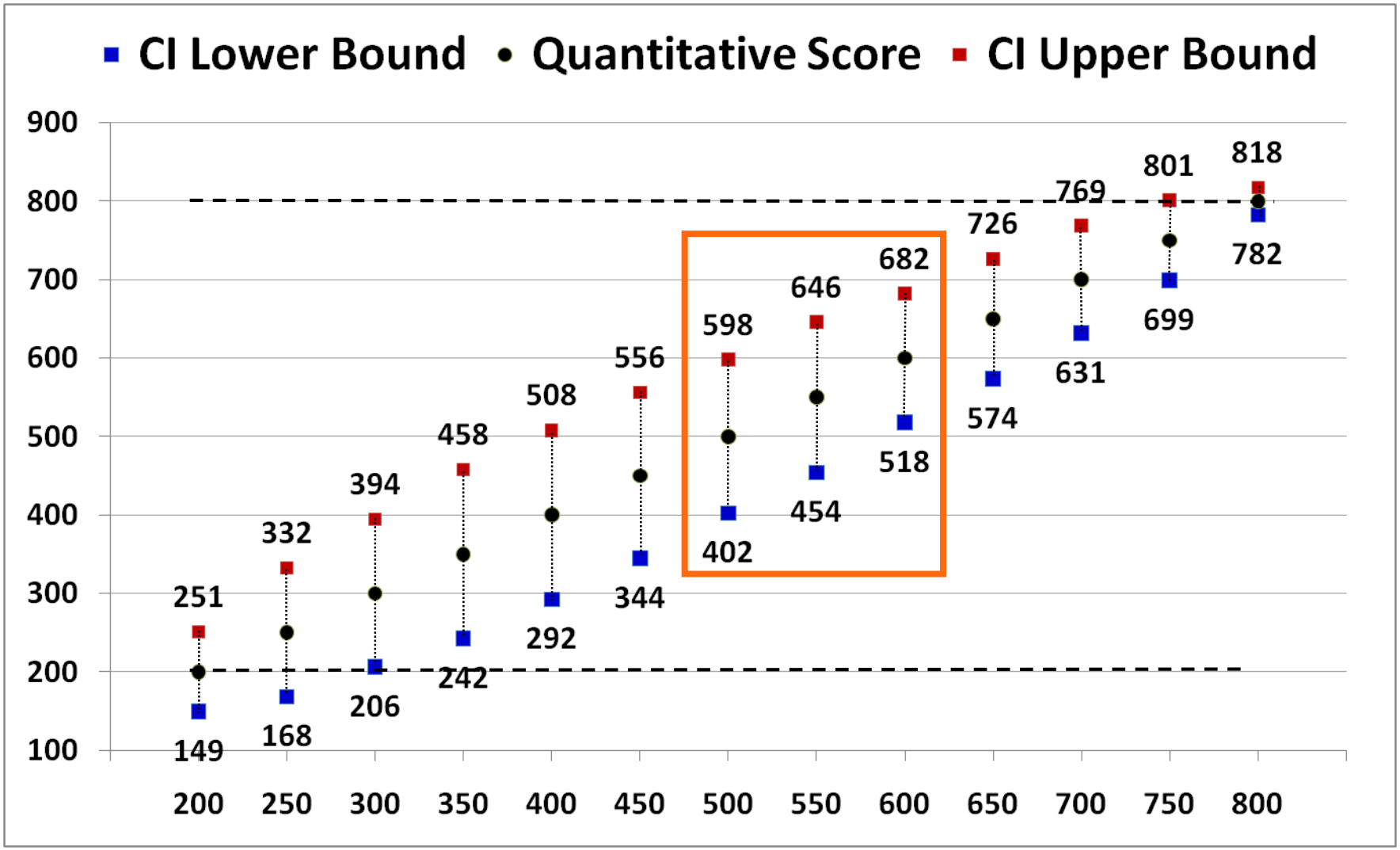
- Measurement error is often thought of as the worst, but I believe model misspecification error (of all five types from last slide) to be even worse than measurement error

KU THE UNIVERSITY OF KANSAS

95% Confidence Intervals: Quantitative (GRE 2011 Guide)

SEM ranges from 9 to 55

http://www.ets.org/s/gre/pdf/gre_guide.pdf



THE UNIVERSITY OF KANSAS

FACTOR SCORES

Factor Scores

- To describe a factor score, first remember the CFA model:

$$Y_{pi} = \mu_i + \lambda_i F_p + e_{pi}$$

- Simply put: A factor score is an estimated value for F_p, or $\hat{F}_p$

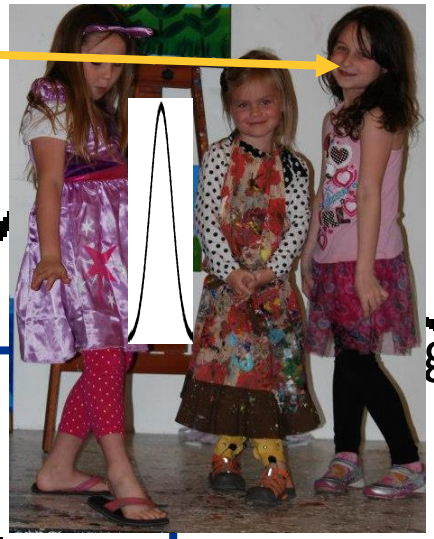
- There has long been a resistance to using factor scores in psychological research with the most common objection cited being the **indeterminacy of factor scores**
  - **Indeterminacy of factor scores == factor scores are not unique**

- Why are factor scores not unique? Because factor models must fix some parameters for identification
  - The values may be indeterminate—but in CFA and in ML versions of EFA **the rank order of the factor scores is unique**

A Depiction of Traditional Psychometric Estimates

Ordering provided by Daphne Templin

Latent Trait Score



Low

gh

Bieber

Cyrus

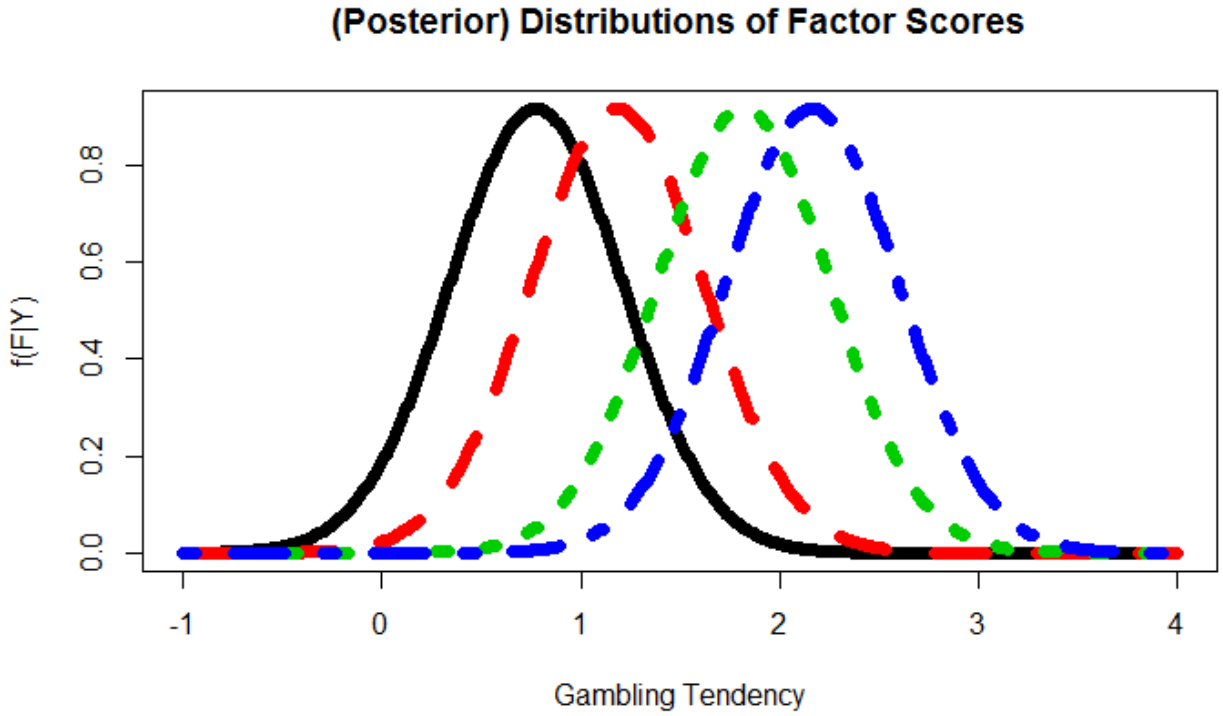
Swift

Factor scores provide a **weak ordering** of people (weak because of error): like ordinal-level measurement

THE UNIVERSITY OF KANSAS

Draw Templin, Draw!

Example factor scores and their distributions (discussed next)



A different version of factor model identification would change the numbers on the X-axis, but the shapes and order of the distributions would not change

Factor scores provide a **weak ordering** of people (weak because of error)

KU THE UNIVERSITY OF KANSAS

Factor Scores and Testing

- These factor scores are found using the same methods as are used in practice for finding test scores (like the GRE)
  - The only difference between such test scores and factor scores in this class is the distributional assumptions of the measurement model (IRT is CFA with assumed Bernoulli/Multinomial distributed items)
  - They behave the same

- That said, some in the testing industry don't quite realize how these work

See: http://images.pearsonassessments.com/images/tmrs/Responses_Walter_Stroup.pdf (p. 2)

> • **IRT does not rank order students or select test questions. IRT simply measures students' academic knowledge and skills on a scale (like a ruler) and, just as a child gets taller, when students increase their knowledge and skills, their test scores will increase. IRT provides a thorough and fair measurement of growth and mastery.**

More on Factor Scores

- Factor scores (by other names) are used in many domains
  - Item response theory (CFA with categorical items): GRE scores are factor scores
- Because the historical relationship between CFA and exploratory factor analysis, factor scores are widely avoided
  - In EFA factor meaning is unknown so rotations were used
- Further making the issue even more difficult, many **crazy** methods for determining factor scores have been developed
  - See http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3773873/
- We will only focus on one method for estimating factor scores that is used in nearly all fields based on the posterior distribution of the factor score given the data
  - Identical to methods described by Lawley and Maxwell (1971) of Bartlett (1936)
  - Also used in generalized linear mixed effects models where factor scores are called Best Linear Unbiased Predictors (or BLUPs)

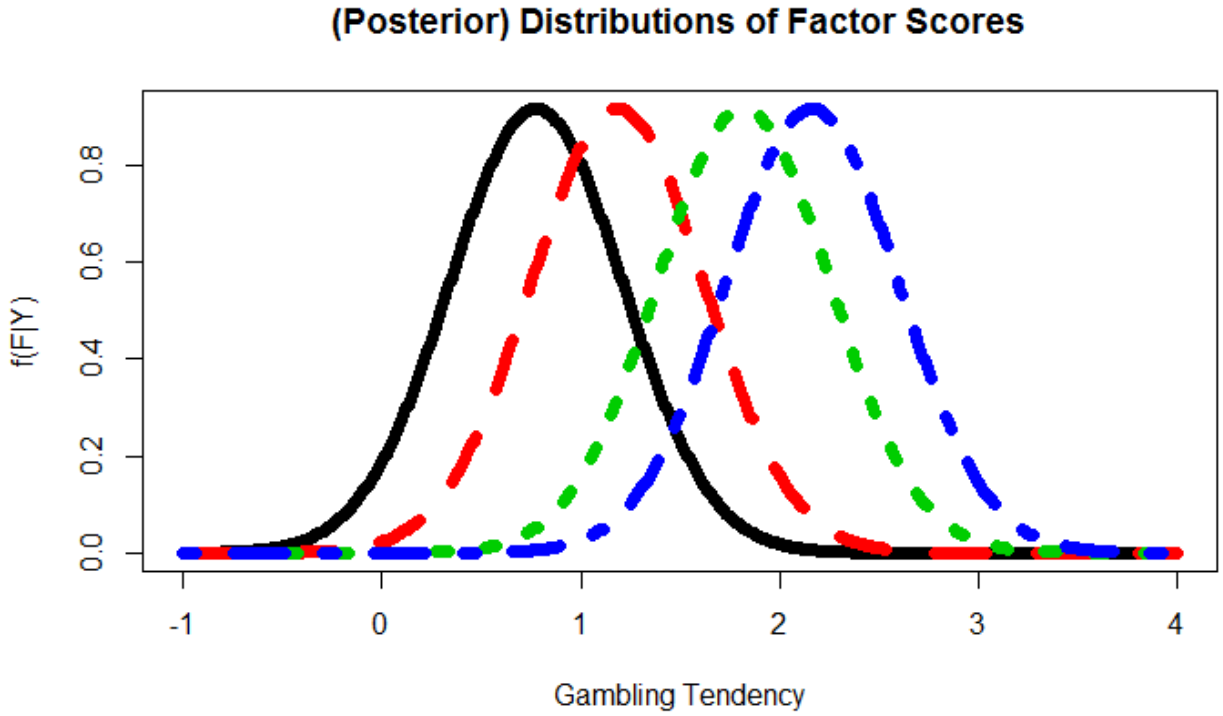
KU THE UNIVERSITY OF KANSAS

Factor Scores: The Big Picture

- A factor score is the estimate of a subject's unobserved latent trait
- Because this latent variable is not measured directly, it acts like it is missing data: you really cannot know with certainty its true value
- It is difficult to pin down what the missing data value (factor score value) should be precisely
  - Each factor score has a posterior distribution of possible values
  - Often, the mean of the posterior distribution is the "factor score"
    - In CFA, the mean is the most likely value
  - Depending on the test, there may be a lot of error (variability) in the distribution
- Therefore, the use of factor scores must reflect that the score is not known and is represented by a distribution

Draw Templin, Draw!

Example factor scores and their distributions (discussed next)



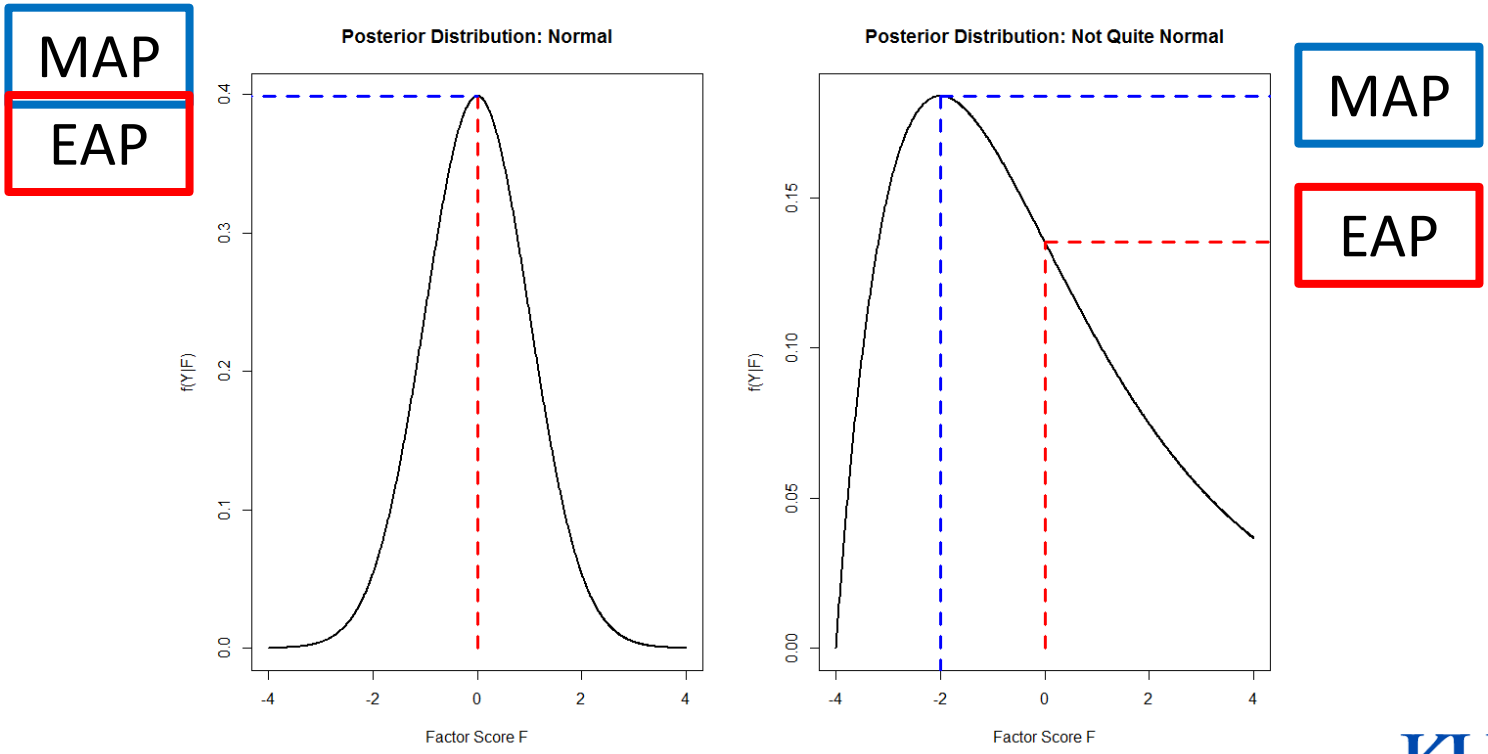
A different version of factor model identification would change the numbers on the X-axis, but the shapes and order of the distributions would not change

Factor scores provide a **weak ordering** of people (weak because of error)

KU THE UNIVERSITY OF KANSAS

How Distributions get Summarized into Scores

- There are two ways of providing a score from the factor score posterior distribution:
  - Expected a posteriori (EAP): the mean of the distribution
  - Maximum a posteriori (MAP): the most likely score from the distribution
- In CFA factor score distributions are normal (so EAP=MAP)



KU THE UNIVERSITY OF KANSAS

Additional Information on Factor Scores

- For EAP factor scores:
  - $\hat{F}_p = E\left(f\left(F_p|\mathbf{Y}\right)\right)$
  - $SE\left(\hat{F}_p\right) = \sqrt{Var\left(f\left(F_p|\mathbf{Y}\right)\right)}$
- For MAP factor scores:
  - $\hat{F}_p = \arg\max_{F_p} f\left(F_p|\mathbf{Y}\right)$
  - $SE\left(\hat{F}_p\right) = \left[\frac{\partial^2}{\partial F_p^2} f\left(F_p|\mathbf{Y}\right)\Big|_{\hat{F}_p}\right]^{-\frac{1}{2}}$ (square root of Fisher's information)

- For CFA (Normal Data/Normal Factor) measurement models:
  - MAP = EAP
  - Variance is identical across all people, regardless of score
- For non-CFA measurement models:
  - MAP ≠ EAP (but does with infinite items)
  - Standard error is a function of the factor score

KU THE UNIVERSITY OF KANSAS

Tying Factor Scores to Classical Test Theory

- Recall Classical Test Theory's model:

$$Y = T + E$$

- With reliability: $\rho = \frac{Var(T)}{Var(T)+Var(E)}$

- For factor scores:
  - $Var(T) = \sigma_F^2$: the (possibly estimated) variance of the factor
  - $Var(E) = SE\left(\hat{F}_p\right)^2$: From the posterior distribution of the factor score

- Therefore, reliability of factor scores can be computed using model estimated parameters
  - Caution: The factor model must fit to use these parameters!
  - Caveat: We'll soon see reliability for sum scores can be estimated by CFA model parameters

Factor Scores: Empirical Bayes Estimates

- **For most (if not all) latent variable techniques**, the factor scores come from Empirical Bayes estimation—meaning there is a prior distribution present
  - Empirical = some or all of the parameters of the distribution of the latent variable are estimated (i.e., factor mean and variance)
  - Bayes = comes from the use of Bayes' Theorem

- Prior == Assumed factor distribution with mean/variance

- This is true for all CFA, IRT, mixed/multilevel/hierarchical models
  - And is true for models that don't have a label (e.g., Poisson Factor Analysis?)

Bayes' Theorem

- Bayes' Theorem states the conditional distribution of a variable A (soon to be our factor score) given values of a variable B (soon to be our data) is:

For Categorical A, replace integral with sum

$$f(A|B) = \frac{f(B|A)f(A)}{f(B)} = \frac{f(B|A)f(A)}{\int_{a \in A} f(B|A=a)f(A=a)da}$$

- $f(A|B)$ is the **distribution** of A, conditional on B
  - We will come to know this as the posterior distribution of the factor score, conditional on the data observed or $f(\mathbf{F}|\mathbf{Y})$
- $f(B|A)$ is the **distribution** of B, conditional on A
  - We will come to know this as our measurement model or $f(\mathbf{Y}|\mathbf{F})$
- $f(A)$ is the marginal distribution of A
  - We will come to know this as the prior distribution of the factor or $f(\mathbf{F})$

KU THE UNIVERSITY OF KANSAS

Putting Together the Pieces of Empirical Bayes Factor Scores

$$f(A|B) = \frac{f(B|A)f(A)}{f(B)} = f(\mathbf{F}|\mathbf{Y}) = \frac{f(\mathbf{Y}|\mathbf{F})f(\mathbf{F})}{f(\mathbf{Y})}$$

- For $f(\mathbf{Y}|\mathbf{F})$, consider the measurement model (here CFA) for one item:

$$Y_{pi} = \mu_i + \lambda_i F_p + e_{pi}$$

Where: $e_{pi} \sim N(0, \psi_i^2)$

- Using expected values, we can show the distribution for this one item is:

$$f(Y_{pi}|F_p) \sim N(\mu_i + \lambda_i F_p, \psi_i^2)$$

- Therefore, for all I items, our conditional distribution is:

$$f(\mathbf{Y}|F_p) \sim N_I(\boldsymbol{\mu} + \boldsymbol{\Lambda} F_p, \boldsymbol{\Psi})$$

- With multiple factors, this becomes:

$$f(\mathbf{Y}|\mathbf{F}) \sim N_I(\boldsymbol{\mu} + \boldsymbol{\Lambda}\mathbf{F}, \boldsymbol{\Psi})$$

Putting Together the Pieces of Empirical Bayes Factor Scores

$$f(A|B) = \frac{f(B|A)f(A)}{f(B)} = f(\mathbf{F}|\mathbf{Y}) = \frac{f(\mathbf{Y}|\mathbf{F})f(\mathbf{F})}{f(\mathbf{Y})}$$

- For $f(\mathbf{F})$, consider the distribution assumed by the factor:
  - For one factor

$$f(F_p) \sim N(\mu_F, \sigma_F^2)$$

  - For multiple factors K

$$f(\mathbf{F}) \sim N_K(\boldsymbol{\mu}_F, \boldsymbol{\Phi})$$

- We must pick an identification method which determines if certain parameters of $\boldsymbol{\mu}_F$ and $\boldsymbol{\Phi}$ are fixed or are estimated
  - Any method identification works, so we keep $\boldsymbol{\mu}_F$ and $\boldsymbol{\Phi}$ throughout

Putting Together the Pieces of Empirical Bayes Factor Scores

$$f(A|B) = \frac{f(B|A)f(A)}{f(B)} = f(\mathbf{F}|\mathbf{Y}) = \frac{f(\mathbf{Y}|\mathbf{F})f(\mathbf{F})}{f(\mathbf{Y})}$$

- For $f(\mathbf{Y})$, we return to the model-implied mean vector and covariance matrix:

$$f(\mathbf{Y}) \sim N_I(\boldsymbol{\mu} + \boldsymbol{\Lambda}^T\boldsymbol{\mu}_F, \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T + \boldsymbol{\Psi})$$

KU THE UNIVERSITY OF KANSAS

A Quick Reminder About Types of Distributions

- For two random variables x and z, a conditional distribution is written as: $f(z|x)$

- The conditional distribution is also equal to the joint distribution divided by the marginal distribution of the conditioning random variable

$$f(z|x) = \frac{f(z,x)}{f(x)}$$

- Therefore, the joint distribution can be found by the product of the conditional and marginal distributions:

$$f(z,x) = f(z|x)f(x)$$

- We can use this result in our analysis:

$$f(\mathbf{Y}|\mathbf{F})f(\mathbf{F}) = f(\mathbf{Y},\mathbf{F})$$

A Quick Reminder about Multivariate Normal Distributions

- If $\mathbf{X}$ is distributed multivariate normally:

Conditional distributions of $\mathbf{X}$ are multivariate normal

- We can show that $f(\mathbf{Y}, \mathbf{F})$, the joint distribution of the data and the factors, is multivariate normal

- We can then use the result above (shown on the next slides) to show that our posterior distribution of the factor scores is also multivariate normal
  - This result **only** applies for measurement models assuming normally distributed data and normally distributed factors: CFA
  - For IRT (and other measurement models), this result will not hold—but this distribution is asymptotically normal as the number of items gets large

KU THE UNIVERSITY OF KANSAS

Conditional Distributions of MVN Variables are Multivariate Normal

- The conditional distribution of sets of variables from a MVN is also MVN

- If we were interested in the distribution of the first q variables, we partition three matrices:
  - The data: $[\mathbf{X}_{1:(N\ x\ q)} \mid \mathbf{X}_{2:(N\ x\ p-q)}]$
  - The mean vector: $\begin{bmatrix} \boldsymbol{\mu}_{1:(q\ x\ 1)} \\ \hline \boldsymbol{\mu}_{2:(p-q\ x\ 1)} \end{bmatrix}$
  - The covariance matrix: $\left[\begin{array}{c|c} \boldsymbol{\Sigma}_{11:(q\ x\ q)} & \boldsymbol{\Sigma}_{12:(q\ x\ p-q)} \\ \hline \boldsymbol{\Sigma}_{21:(p-q\ x\ q)} & \boldsymbol{\Sigma}_{22:(p-q\ x\ p-q)} \end{array}\right]$

KU THE UNIVERSITY OF KANSAS

Conditional Distributions of MVN Variables

- The, $f(\mathbf{X}_1|\mathbf{X}_2)$, conditional distribution of $\mathbf{X}_1$ given the values of $\mathbf{X}_2 = \mathbf{x}_2$ is then:

$$\mathbf{X}_1|\mathbf{X}_2 \sim N_q(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$$

Where (using our partitioned matrices):

$$\boldsymbol{\mu}^* = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2^T - \boldsymbol{\mu}_2)$$

And:

$$\boldsymbol{\Sigma}^* = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$$

Derive, Templin, Derive!

- The joint distribution of all I items and K factor scores is

$$f(\mathbf{Y}, \mathbf{F}) = f\left(\begin{bmatrix}\mathbf{Y}\\ \mathbf{F}\end{bmatrix}\right)$$

$$= N_{I+K}\left(\begin{bmatrix}\boldsymbol{\mu} + \boldsymbol{\Lambda}^T\boldsymbol{\mu}_F\\ \boldsymbol{\mu}_F\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T + \boldsymbol{\Psi} & \boldsymbol{\Lambda}\boldsymbol{\Phi}\\ \boldsymbol{\Phi}\boldsymbol{\Lambda}^T & \boldsymbol{\Phi}\end{bmatrix}\right)$$

- Using the conditional distributions of MVNs result:

$f\left(\mathbf{F}_p \middle| \mathbf{Y}_p\right)$ is MVN:

With mean: $\boldsymbol{\mu}_F + \boldsymbol{\Phi}\boldsymbol{\Lambda}^T(\boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T + \boldsymbol{\Psi})^{-1}\left(\mathbf{Y}_p^T - \boldsymbol{\mu}\right)$

And Covariance: $\boldsymbol{\Phi} - \boldsymbol{\Phi}\boldsymbol{\Lambda}^T(\boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T + \boldsymbol{\Psi})^{-1}\,\boldsymbol{\Lambda}\boldsymbol{\Phi}$

#WTFTemplin

What All That Math Means for Factor Scores

- **When using measurement models assuming normally distributed data and normally distributed factors (CFA)**:
  - The posterior distribution of the factor scores is MVN
  - Therefore, the **most likely** factor score (MAP) and the **expected** factor score (EAP) is given by the mean from the previous slides
  - The factor score is a function of the model parameter estimates and the data

LINKING SUM SCORES AND CTT TO MEASUREMENT MODELS VIA FACTOR SCORES

Connecting Sum Scores and Factor Scores

- Sum scores have a correlation of 1.0 with factor scores from a **parallel items** CFA model
  - Parallel items model: all factor loadings equal + all unique variances equal

- For example, here are the parallel items model equations for our three-item GRI example data:

$$GRI1_p = \mu_1 + \lambda F_p + e_{p1}; \qquad e_{p1} \sim N(0, \psi^2)$$
$$GRI3_p = \mu_1 + \lambda F_p + e_{p3}; \qquad e_{p3} \sim N(0, \psi^2)$$
$$GRI5_p = \mu_1 + \lambda F_p + e_{p5}; \qquad e_{p5} \sim N(0, \psi^2)$$

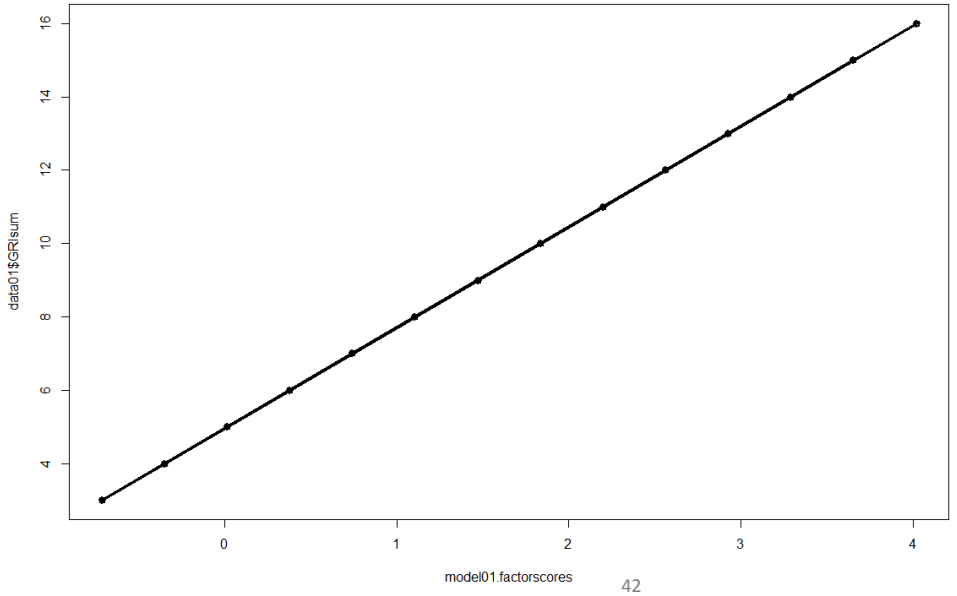
- With a common loading estimated, we will use a standardized factor identification (but we don't have to)

$$F_p \sim N(0, 1)$$

Comparing a PI Model Factor Score to a Sum Score

```
model01.lavaan = "
  GAMBLING =~ (LOADING)*GRI1+(LOADING)*GRI3+(LOADING)*GRI5

  GRI1 ~~ (UVAR)*GRI1
  GRI3 ~~ (UVAR)*GRI3
  GRI5 ~~ (UVAR)*GRI5

  GAMBLING ~~ GAMBLING
"
model01.fit = sem(model01.lavaan, data=data01, estimator = "MLR", mimic="Mplus", fixed.x=FALSE, std.lv=TRUE)
summary(model01.fit, fit.measures=TRUE, rsquare=TRUE, standardized=TRUE)

#get factor score estimates from the predict function:
model01.factorscores = predict(model01.fit)

#compare both on plot:
par(mfrow = c(1,1))
plot(model01.factorscores, data01$GRIsum, type ="o", lwd=3)

#compare with correlation
cor(model01.factorscores, data01$GRIsum)
```

```
> cor(model01.factorscores, data01$GRIsum)
          [,1]
GAMBLING    1
```



Comparing for Specific Scores

- To look more closely at factor scores versus sum scores, consider the following five people in the data set

```
> #Factor score of a person with GRI1==1, GRI3==1, & GRI5==1 ---- Sum Score = 3
> person111_id = data01[data01$GRI1==1 &data01$GRI3==1 & data01$GRI5==1,]$ID[1]
> model01.factorscores[person111_id]
[1] -0.7151343
>
> #Factor score of a person with GRI1==1, GRI3==1, & GRI5==2 ---- Sum Score = 4
> person112_id = data01[data01$GRI1==1 &data01$GRI3==1 & data01$GRI5==2,]$ID[1]
> model01.factorscores[person112_id]
[1] -0.3508875
>
> #Factor score of a person with GRI1==1, GRI3==2, & GRI5==1 ---- Sum Score = 4
> person121_id = data01[data01$GRI1==1 &data01$GRI3==2 & data01$GRI5==1,]$ID[1]
> model01.factorscores[person121_id]
[1] -0.3508875
>
> #Factor score of a person with GRI1==2, GRI3==1, & GRI5==1 ---- Sum Score = 4
> person211_id = data01[data01$GRI1==2 &data01$GRI3==1 & data01$GRI5==1,]$ID[1]
> model01.factorscores[person211_id]
[1] -0.3508875
>
> #Difference between factor scores for sum score 3 vs. sum score 4:
> model01.factorscores[person112_id] - model01.factorscores[person111_id]
[1] 0.3642468
>
> #Factor score of a person with GRI1==1, GRI3==1, & GRI5==3 ---- Sum Score = 5
> person113_id = data01[data01$GRI1==1 &data01$GRI3==1 & data01$GRI5==3,]$ID[1]
> model01.factorscores[person113_id]
[1] 0.01335935
>
> #Difference between factor scores for sum score 3 vs. sum score 4:
> model01.factorscores[person113_id] - model01.factorscores[person112_id]
[1] 0.3642468
```

Before We Get Too Far…Did The Model Fit?

- Good model fit…
- We could use the model
- So, we could use the factor scores or the sum scores
- But we won't!

Estimator		ML	Robust	
Minimum Function Test Statistic		49.386	19.176	
Degrees of freedom		4	4	
P-value (Chi-square)		0.000	0.001	
Scaling correction factor			2.575	
for the Yuan-Bentler correction (Mplus variant)				
Model test baseline model:				
Minimum Function Test Statistic		480.988	199.641	
Degrees of freedom		3	3	
P-value		0.000	0.000	
User model versus baseline model:				
Comparative Fit Index (CFI)		0.905	0.923	
Tucker-Lewis Index (TLI)		0.929	0.942	
Loglikelihood and Information Criteria:				
Loglikelihood user model (H0)		-5279.302	-5279.302	
Scaling correction factor			1.091	
for the MLR correction				
Loglikelihood unrestricted model (H1)		-5254.609	-5254.609	
Scaling correction factor			2.236	
for the MLR correction				
Number of free parameters		5	5	
Akaike (AIC)		10568.605	10568.605	
Bayesian (BIC)		10594.592	10594.592	
Sample-size adjusted Bayesian (BIC)		10578.709	10578.709	
Root Mean Square Error of Approximation:				
RMSEA		0.092	0.053	
90 Percent Confidence Interval	0.070	0.116	0.039	0.069
P-value RMSEA <= 0.05		0.001	0.331	
Standardized Root Mean Square Residual:				
SRMR		0.115	0.115	

And…About Reliability

- Factor score reliability is:

$$\rho = \frac{\sigma_F^2}{\sigma_F^2 + SE\left(F_p\right)^2}$$

- lavaan does not compute the factor score standard errors (Mplus does)…but that's okay, because we can grab them from the matrix algebra on p. 35

R Syntax for Computing SE of Factor Scores

```
> #getting more decimal places from model estimates estimates
> parameterEstimates(model01.fit)$est
 [1] 0.5759246 0.5759246 0.5759246 0.5860709 0.5860709 0.5860709 1.0000000 1.8226048 1.5479042 1.5928144 0.0000000 0.3614116
>
> #saving into matrices:
> lambda = matrix(.5759246, nrow=3, ncol=1)
> psi = diag(rep(.5860709,times=3))
> mu = matrix(c(1.8226048, 1.5479042, 1.5928144),nrow=3, ncol = 1)
> phi = matrix(1, nrow=1, ncol=1)
> mu_f = matrix(0, nrow=1, ncol=1)
>
> sigma = lambda%*%t(lambda) + psi
>
> x = matrix(cbind(data01$GRI1, data01$GRI3, data01$GRI5), ncol=3)
>
> #getting mean and variance of factor scores from slide 35:
> scores = t(phi %*% t(lambda) %*% solve(sigma)%*%(t(x) - mu%*%matrix(1,nrow=1, ncol=dim(x)[1])))
> varscores = phi - phi %*% t(lambda) %*% solve(sigma) %*% lambda %*% phi
>
> #the standard error of the factor score:
> sqrt(varscores)
          [,1]
[1,] 0.6088217
>
> #showing they match with lavaan's estimated scores
> plot(scores, model01.factorscores)
>
> #factor score reliability
> 1/(1+varscores)
          [,1]
[1,] 0.7295735
```

Reliability of Factor Score = .73

What about the reliability of our sum scores?

Classical Test Theory from a CFA Perspective

- In CTT the unit of analysis is the test score:

$$Y_{p,Total} = T_p + E_p$$

- In CFA the unit of analysis is the item:

$$Y_{pi} = \mu_{I_i} + \lambda_i F_p + e_{pi}$$

- To map CFA onto CTT, we must put these together:

$$Y_{p,Total} = \sum_{i=1}^{I} Y_{pi}$$

Further Unpacking of the Total Score Formula

- Because CFA is an item-based model, we can then substitute each item's model into the sum:

$$Y_{p,Total} = \sum_{i=1}^{I} Y_{pi} = \sum_{i=1}^{I} \left(\mu_{I_i} + \lambda_i F_p + e_{pi}\right)$$

$$= \sum_{i=1}^{I} \mu_{I_i} + \left(\sum_{i=1}^{I} \lambda_i\right) F_p + \sum_{i=1}^{I} e_{pi}$$

- Mapping this onto true score and error from CTT:

$$T = \sum_{i=1}^{I} \mu_{I_i} + \left(\sum_{i=1}^{I} \lambda_i\right) F_p \text{ and } E = \sum_{i=1}^{I} e_{pi}$$

CFA-Model Estimated Reliability of Sum Scores

- From:

$$T = \sum_{i=1}^{I} \mu_{I_i} + \left(\sum_{i=1}^{I} \lambda_i \right) F_p \text{ and } E = \sum_{i=1}^{I} e_{pi}$$

- $Var(T) = Var\left(\sum_{i=1}^{I} \mu_{I_i}\right) + \left(\sum_{i=1}^{I} \lambda_i^2\right) Var\left(F_p\right) =$

$$\left(\sum_{i=1}^{I} \lambda_i \right)^2 \sigma_F^2$$

- $Var(E) = Var\left(\sum_{i=1}^{I} e_{pi}\right) =$

$$\sum_{i=1}^{I} \psi_i^2$$

For models with correlated residuals, those add to Var(E)

KU THE UNIVERSITY OF KANSAS

CFA-Model Estimated Reliability of Sum Scores

- From the previous slide:

$$\rho = \frac{Var(T)}{Var(T) + Var(E)} = \frac{\left(\sum_{i=1}^{I} \lambda_i\right)^2 \sigma_F^2}{\left(\sum_{i=1}^{I} \lambda_i\right)^2 \sigma_F^2 + \sum_{i=1}^{I} \psi_i^2}$$

- And…we can do this in lavaan syntax:

```
model01.lavaan = "
  GAMBLING =~ (LOADING)*GRI1+(LOADING)*GRI3+(LOADING)*GRI5

  GRI1 ~~ (UVAR)*GRI1
  GRI3 ~~ (UVAR)*GRI3
  GRI5 ~~ (UVAR)*GRI5

  GAMBLING ~~ GAMBLING

  rho := ( (3*LOADING)^2 )/(((3*LOADING)^2)+3*UVAR)
"
model01.fit = sem(model01.lavaan, data=data01, estimator = "MLR", mimic="Mplus", fixed.x=FALSE, std.lv=TRUE)
summary(model01.fit, fit.measures=TRUE, rsquare=TRUE, standardized=TRUE)
```

- The estimated reliability is….

```
Defined parameters:
    rho               0.629    0.025   24.968    0.000    0.629    0.629
```


KU THE UNIVERSITY OF KANSAS

Notes on CFA-Estimated Reliabilities

- The CFA-Estimated reliability is for the sum score, not the factor score

- The sum score's reliability is .629 (SE = .025); the factor score's reliability is .73
  - The difference comes from additional sources of error in the factor score:
    - Sampling error
    - Error from the prior distribution (squishing the variance of the factor/error)

- The sum score's reliability is equal to the **Spearman Brown** reliability estimate
  - Therefore, CTT reliability estimates can come from CFA….

Comparing Other CFA Models with Sum Scores

- Another model to consider is the Tau-equivalent items model, which, for CFA, means equal loadings but different unique variances:

- For example, here are the parallel items model equations for our three-item GRI example data:

$$GRI1_p = \mu_1 + \lambda F_p + e_{p1}; \qquad e_{p1} \sim N(0, \psi_1^2)$$
$$GRI3_p = \mu_1 + \lambda F_p + e_{p3}; \qquad e_{p3} \sim N(0, \psi_3^2)$$
$$GRI5_p = \mu_1 + \lambda F_p + e_{p5}; \qquad e_{p5} \sim N(0, \psi_5^2)$$

- With a common loading estimated, we will use a standardized factor identification (but we don't have to)

$$F_p \sim N(0, 1)$$

The Tau Equivalent Model in lavaan

- Note: shown for didactic purposes (don't use this model)

```
#Tau Equivalent Model-------------------------------------------------------------------------
model02.lavaan = "
  GAMBLING =~ (LOADING)*GRI1+(LOADING)*GRI3+(LOADING)*GRI5

  GRI1 ~~ (UVAR1)*GRI1
  GRI3 ~~ (UVAR3)*GRI3
  GRI5 ~~ (UVAR5)*GRI5

  GAMBLING ~~ GAMBLING

  rho := ( (3*LOADING)^2 )/(((3*LOADING)^2)+UVAR1+UVAR3+UVAR5)
"
```

- Yielding model fit indices of:

```
Estimator                                         ML      Robust
Minimum Function Test Statistic               18.897       8.482
Degrees of freedom                                2           2
P-value (Chi-square)                          0.000       0.014
Scaling correction factor                                 2.228
  for the Yuan-Bentler correction (Mplus variant)
```

User model versus baseline model:

Comparative Fit Index (CFI)	0.965	0.967
Tucker-Lewis Index (TLI)	0.947	0.951

Root Mean Square Error of Approximation:

RMSEA		0.080	0.049	
90 Percent Confidence Interval	0.049	0.114	0.028	0.073
P-value RMSEA <= 0.05		0.053	0.475	

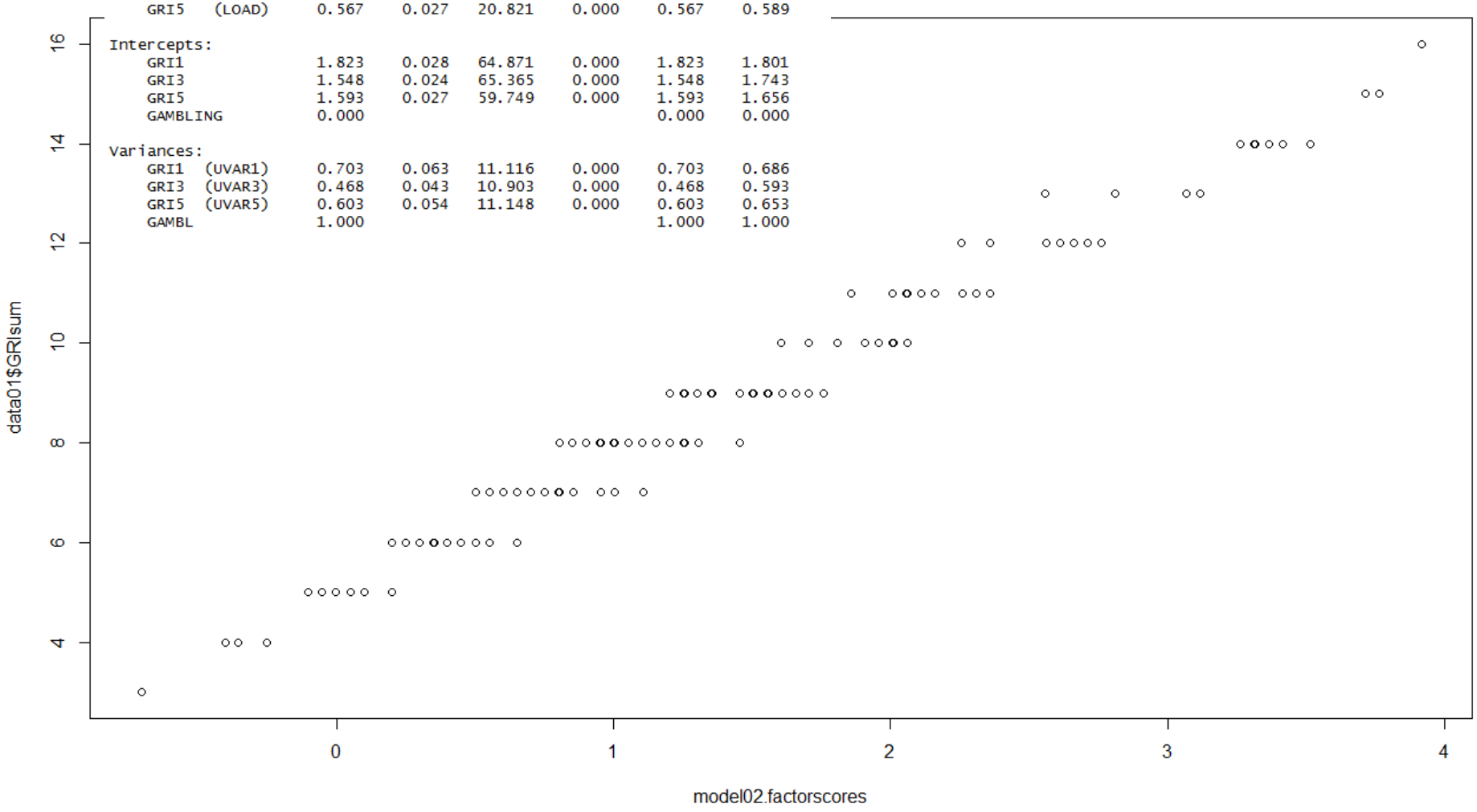
Standardized Root Mean Square Residual:

SRMR	0.040	0.040

Parameter Estimates vs. Factor Score vs. Sum Score

		Estimate	Std.err	Z-value	P(>|z|)	Std.lv	Std.all
Latent variables:							
GAMBLING =~							
GRI1	(LOAD)	0.567	0.027	20.821	0.000	0.567	0.560
GRI3	(LOAD)	0.567	0.027	20.821	0.000	0.567	0.638
GRI5	(LOAD)	0.567	0.027	20.821	0.000	0.567	0.589
Intercepts:							
GRI1		1.823	0.028	64.871	0.000	1.823	1.801
GRI3		1.548	0.024	65.365	0.000	1.548	1.743
GRI5		1.593	0.027	59.749	0.000	1.593	1.656
GAMBLING		0.000				0.000	0.000
Variances:							
GRI1	(UVAR1)	0.703	0.063	11.116	0.000	0.703	0.686
GRI3	(UVAR3)	0.468	0.043	10.903	0.000	0.468	0.593
GRI5	(UVAR5)	0.603	0.054	11.148	0.000	0.603	0.653
GAMBL		1.000				1.000	1.000

```
> cor(model02.factorscores, data01$GRIsum)
                [,1]
GAMBLING 0.9945267
```



Factor vs. Sum Score…by item

- Now what matters is which item had a higher score…
  - Items with higher information (loading^2/unique variance) result in bigger jumps in factor score relative to items with lower information

```
> scoremat
 GRI1 GRI3 GRI5 SUMSC       FS-PI        FS-TE
    1    1    1     3 -0.71513425 -0.704060532
    1    1    2     4 -0.35088745 -0.353082562
    1    2    1     4 -0.35088745 -0.251461804
    2    1    1     4 -0.35088745 -0.402560941
    1    1    3     5  0.01335935 -0.002104592
>
```

```
Variances:
    GRI1   (UVAR1)      0.703
    GRI3   (UVAR3)      0.468
    GRI5   (UVAR5)      0.603
```

Tau Equivalent Reliability for Factor and Sum Scores

- Factor score reliability estimate: .73

```
> 1/(1+varscores)
             [,1]
[1,] 0.7279126
```

- Sum score reliability estimate: .62

```
Defined parameters:
    rho               0.620    0.027   23.220    0.000    0.620    0.594
```

- The sum score reliability is actually coefficient alpha
  - Cronbach's alpha (1951) /Guttman's Lambda 6 (1945)

- HUGE NOTE: THIS IS WHY RELIABILTY IS NOT AN INDEX OF MODEL FIT
  - IT CAN BE SHOWN TO DEPEND ON PARAMETERS THAT WILL BE BIASED UNDER MISFITTING MODELS

Finally…the Unrestricted CFA Model

- All of the previous slides were to get us to see the relationship between sum scores and CFA models
  - We would never estimate either…we would use an unrestricted CFA model
  - Here is what happens with an that unrestricted CFA model

```
model03.lavaan = "
GAMBLING =~ (LOADING1)*GRI1+(LOADING3)*GRI3+(LOADING5)*GRI5

GRI1 ~~ (UVAR1)*GRI1
GRI3 ~~ (UVAR3)*GRI3
GRI5 ~~ (UVAR5)*GRI5

GAMBLING ~~ GAMBLING

rho := ( (LOADING1+LOADING3+LOADING5)^2 )/(((LOADING1+LOADING3+LOADING5)^2)+UVAR1+UVAR3+UVAR5)
"
```

- This model fits perfectly—so no need to check model fit
- Compared to the other two models (we reject CTT)

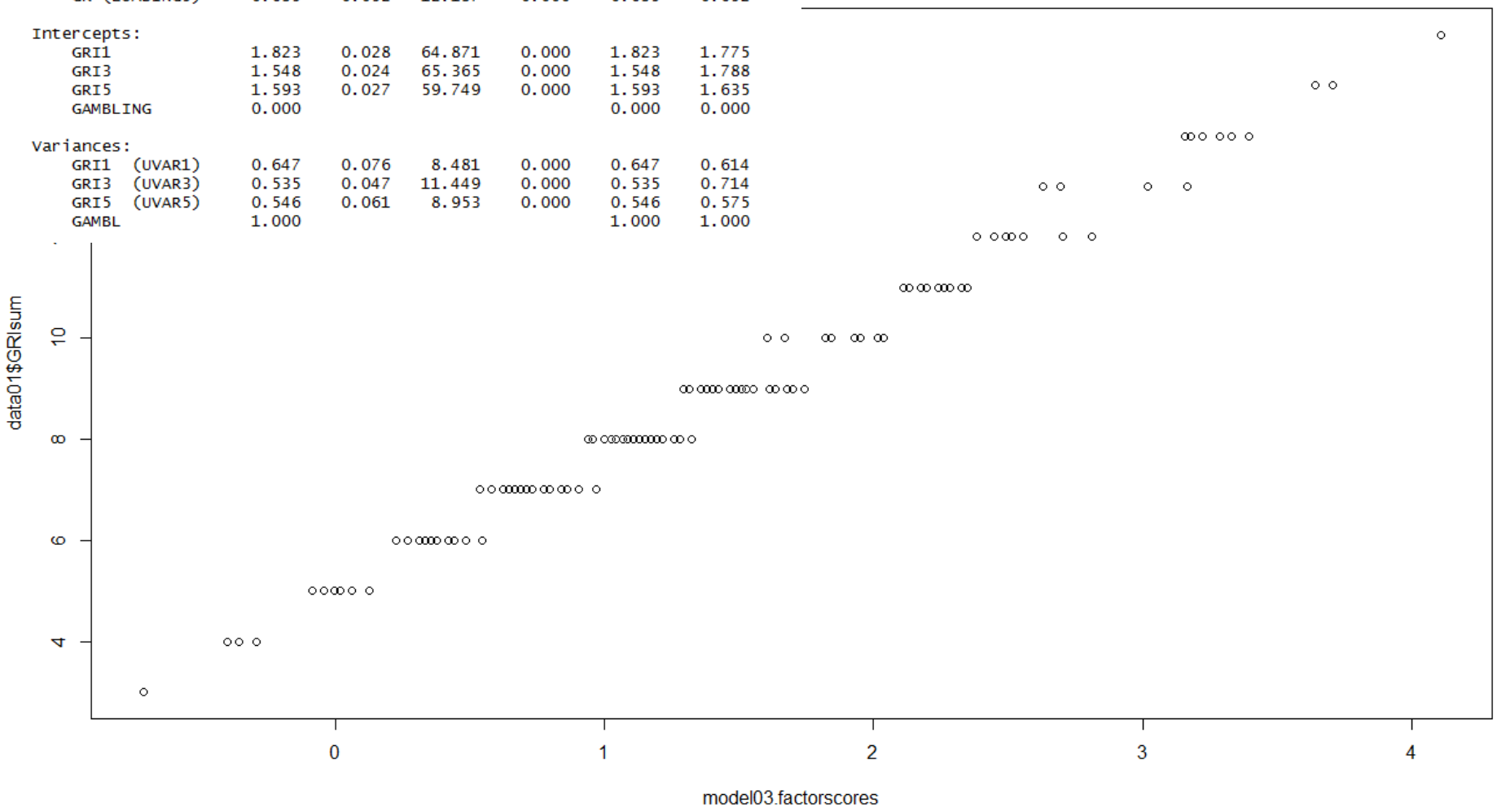
```
> anova(model01.fit, model02.fit, model03.fit)
Scaled Chi Square Difference Test (method = "satorra.bentler.2001")

            Df   AIC   BIC  Chisq Chisq diff Df diff Pr(>Chisq)
model03.fit  0 10527 10574  0.000
model02.fit  2 10542 10578 18.897     8.4825       2   0.014390 *
model01.fit  4 10569 10595 49.386    10.4306       2   0.005433 **
```


KU THE UNIVERSITY OF KANSAS

Parameter Estimates vs. Factor Score vs. Sum Score

	Estimate	Std.err	Z-value	P(>|z|)	Std.lv	Std.all
Latent variables:						
GAMBLING =~						
GR (LOADING1)	0.638	0.052	12.249	0.000	0.638	0.621
GR (LOADING3)	0.463	0.046	10.124	0.000	0.463	0.535
GR (LOADING5)	0.635	0.052	12.137	0.000	0.635	0.652
Intercepts:						
GRI1	1.823	0.028	64.871	0.000	1.823	1.775
GRI3	1.548	0.024	65.365	0.000	1.548	1.788
GRI5	1.593	0.027	59.749	0.000	1.593	1.635
GAMBLING	0.000				0.000	0.000
Variances:						
GRI1 (UVAR1)	0.647	0.076	8.481	0.000	0.647	0.614
GRI3 (UVAR3)	0.535	0.047	11.449	0.000	0.535	0.714
GRI5 (UVAR5)	0.546	0.061	8.953	0.000	0.546	0.575
GAMBL	1.000				1.000	1.000

```
> cor(model03.factorscores, data01$GRIsum)
             [,1]
GAMBLING 0.9975023
```



KU THE UNIVERSITY OF KANSAS

Factor Scores by Sum Score…by item

- Now what matters is which item had a higher score…
  - Items with higher information (loading^2/unique variance) result in bigger jumps in factor score relative to items with lower information

```
> scoremat
 GRI1 GRI3 GRI5 SUMSC        FS-PI         FS-TE       FS-CFA
    1    1    1     3 -0.71513425 -0.704060532 -0.7132409
    1    1    2     4 -0.35088745 -0.353082562 -0.2931139
    1    2    1     4 -0.35088745 -0.251461804 -0.4003987
    2    1    1     4 -0.35088745 -0.402560941 -0.3573275
    1    1    3     5  0.01335935 -0.002104592  0.1270130
```

```
                   Estimate  Std.err  Z-value  P(>|z|)   Std.lv  Std.all
Latent variables:
  GAMBLING =~
    GR (LOADING1)     0.638    0.052   12.249    0.000    0.638    0.621
    GR (LOADING3)     0.463    0.046   10.124    0.000    0.463    0.535
    GR (LOADING5)     0.635    0.052   12.137    0.000    0.635    0.652

Intercepts:
    GRI1              1.823    0.028   64.871    0.000    1.823    1.775
    GRI3              1.548    0.024   65.365    0.000    1.548    1.788
    GRI5              1.593    0.027   59.749    0.000    1.593    1.635
    GAMBLING          0.000                       0.000    0.000

Variances:
    GRI1   (UVAR1)    0.647    0.076    8.481    0.000    0.647    0.614
    GRI3   (UVAR3)    0.535    0.047   11.449    0.000    0.535    0.714
    GRI5   (UVAR5)    0.546    0.061    8.953    0.000    0.546    0.575
    GAMBL             1.000                       1.000    1.000

    info1             0.629    0.157    4.015    0.000    0.629    0.629
    info3             0.401    0.094    4.265    0.000    0.401    0.401
    info5             0.739    0.173    4.268    0.000    0.739    0.739
```


KU THE UNIVERSITY OF KANSAS

CFA Equivalent Reliability for Factor and Sum Scores

- Factor score reliability estimate: .734

```
> 1/(1+varscores)
          [,1]
[1,] 0.7346829
```

- Sum score reliability estimate: .636

```
Defined parameters:
    rho               0.636    0.025   25.248    0.000    0.636    0.632
```

- The sum score reliability is sometimes called coefficient omega (see McDonald, 1999)

- If all three models fit the data then

Omega > Alpha > Spearman Brown

But…the differences are very small

Potential Sources of Error in a Factor Score

- Measurement error
  - e.g., the $SE(\hat{F})$
- Model misspecification error of various types:
  - ~~Dimensionality misspecification error~~
    - ~~e.g., Assuming one dimension when there is more than one present~~
  - ~~Parameter constraint misspecification error~~
    - ~~e.g., Assuming overly restrictive constraints (see next section and all of CTT)~~
  - ~~Linear model functional misspecification error~~
    - ~~e.g., Assuming a linear relationship between the factor and the items when a non-linear one is present~~
  - Outcome distribution misspecification error
    - e.g., Assuming Likert-type data to be continuous and using a normal distribution
  - Factor distribution misspecification error
    - e.g., Assuming your trait is normally distributed when it is categorical or a mixture distribution
- ~~Missing data error~~
  - ~~How you treat missing responses to items makes even more untenable assumptions~~
- Sampling error
- Prior Distribution Error
  - e.g., factor scores are "shrunken estimates"

KU THE UNIVERSITY OF KANSAS

So….?

- Up to this point we have seen
  - Assumptions underlying sum scores
  - Definitions of factor scores
  - How sum scores imply a very specific CFA model

- We have also seen a history of reliability:
  - Spearman Brown (1910): Parallel items model
    - Equal loadings/unique variances
  - Guttman/Cronbch Alpha (1945,1953): Tau equivalent items model
    - Equal loadings
  - Coefficient omega (source unknown): Unrestricted CFA model
  - Reliability for factor scores
  - Also note: the next step is conditional reliability (IRT models)

- **The point is that if you are ever reporting scores but not using them in subsequent analyses, then use a factor score**

- But what we haven't seen is what to do when we cannot use a simultaneous analysis/SEM
  - And that answer will have to come during the next lecture…

KU THE UNIVERSITY OF KANSAS

WRAPPING UP

Wrapping Up

- Today was our first pass at trying to show how SEM and CFA relate to what is frequently done in most analyses
- Next time we will start again with factor scores and work to make the sources of error minimized
- To do so, we'll have to learn a little about missing data, multiple imputation, and Bayesian analyses
  - Hence needing two lectures on the topic!
- These are the most important lectures this semester
  - It provides a WHY and HOW for doing science with errors of measurement
  - Take this information and compare it to how people use scores…like: http://www.ets.org/s/gre/pdf/gre_guide.pdf

KU THE UNIVERSITY OF KANSAS