Dynamic Fit Index Cutoffs

Dan McNeish &
Melissa Wolf

Supported by:
Institute of Educational Sciences
US Dept of Education
R305D220003

30,000 Foot Overview

1. History of SEM/CFA Model Fit

2. Hu & Bentler (1999)

3. Potential Issues With Hu & Bentler Cutoffs

4. Dynamic Fit Index Cutoffs

5. Simulations

Brief History of SEM Model Fit

Fit in CFA/SEM

- Unlike regression, SEM fit not about variance explained
- In SEM, fit is about whether the fitted model reproduces the observed means and covariances

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \boldsymbol{\Sigma}$$

OR

$$\hat{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}$$

Fit in CFA/SEM

- Initial idea for how to test SEM was to use NHST
  - H0: $\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \boldsymbol{\Sigma}$
- Inferentially test whether model-implied equals observed
  - Test of exact fit
  - χ^2 most common test statistic
- Clear hypotheses, clear interpretation
  - Arguments against whether exact fit is meaningful

Fit in CFA/SEM

- Initi
- Infe... rved
- Clea

"In applications of the analysis of covariance structures in the social sciences it is implausible that any model that we use is anything more than an approximation to reality.

Since a null hypothesis that a model fits exactly in some population is known a priori to be false, it seems pointless even to try and test whether it is true"

Browne & Cudeck (1993), p. 137

Fit in CFA/SEM

- Initi
- Infe… rved
- Clea
  - Arguments against whether exact fit is meaningful

> "In most empirical work the model is tentative and is regarded as only an approximation of reality.
>
> Hence the statistical problem is not one of testing a given hypothesis (which a priori may be considered false) but rather one of fitting the model to the data and deciding whether the fit is adequate"
>
> Jöreskog & Sörbom (1982), p. 408

Fit in CFA/SEM

- Approximate fit indices (RMSEA, CFI) were developed
  - Exact fit tests *presence* of misspecification
  - Approximate fit captures *magnitude* of misspecification

- Kind of like an effect size for fit
  - In regression, significant treatment effect doesn't imply practical difference
  - In SEM, lack of exact fit doesn't imply complete lack of utility

- Issue with effect sizes is that definition of "small", "large", "good", or "bad" is subjective

Fit in CFA/SEM

- Heuristic approaches to define what value of fit indices indicates good fit in SEM:

> "Practical experience has made us feel that a value of the RMSEA of about 0.05 or less would indicate a close fit of the model … We are also of the opinion that a value of about 0.08 or less for the RMSEA would indicate a reasonable error of approximation and would not want to employ a model with an RMSEA greater than 0.10"
>
> Browne & Cudeck (1993), p. 141

Fit in CFA/SEM

- Heuristic approaches to define what value of fit indices indicates good fit in SEM:

> "Experience will be required to establish values of the indices [CFI] that are associated with various degrees of meaningfulness of results. In our experience, models with over fit indices of less than .9 can usually be improved substantially"
>
> Bentler & Bonett (1980), p. 600

Hu & Bentler (1999)

Hu and Bentler (1999)

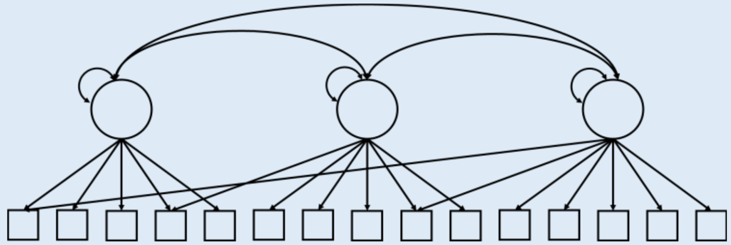
- Personal experience and unsubstantiated heuristics were guiding fit index use for years after they were introduced

- HB tried to determine objective benchmarks for fit indices
  - Also want to verify if heuristic suggestions are reasonable (p. 4)

- They conducted a large simulation, to see which values of fit indices were actually sensitive to misspecification

Hu & Bentler Conditions

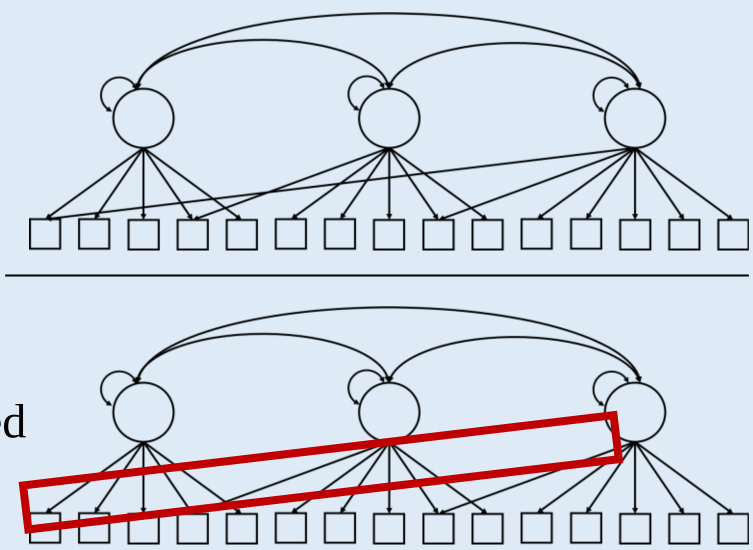
Data were generated from this model

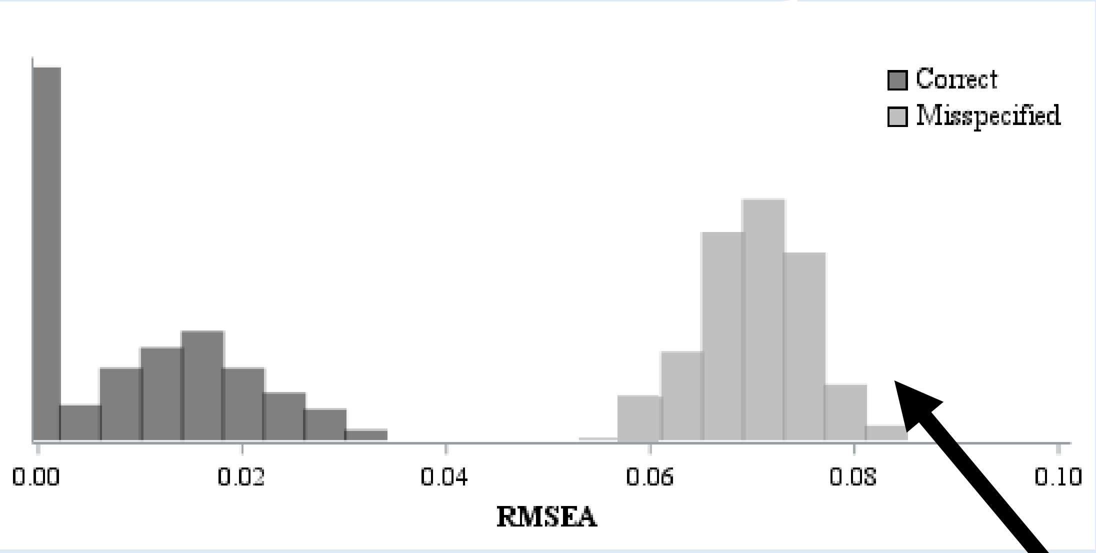
	Complex Model
True	

Hu & Bentler Conditions

Then correct and misspecified models were fit to each dataset

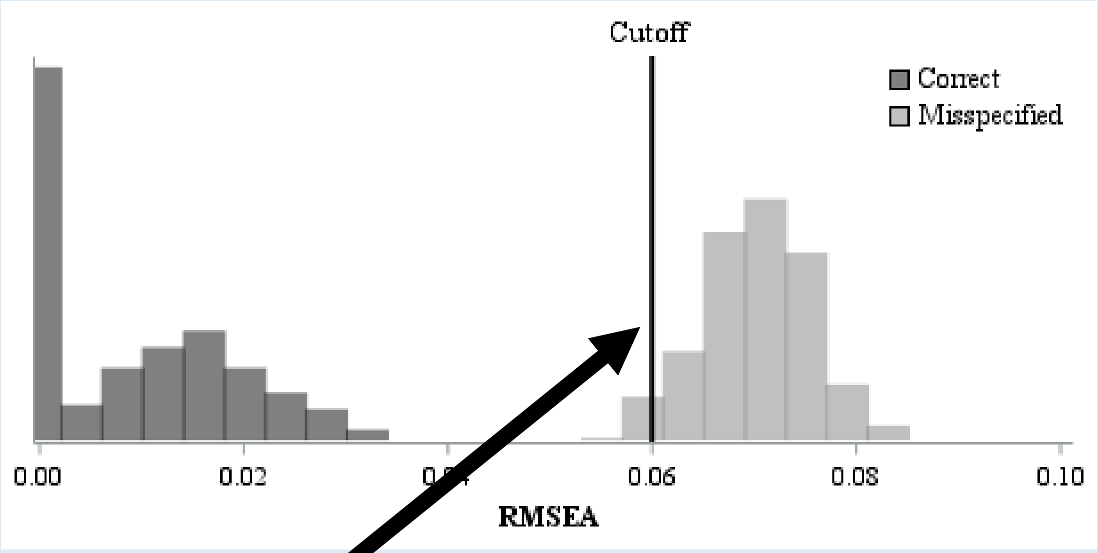
	Complex Model
True	
Misspecified	





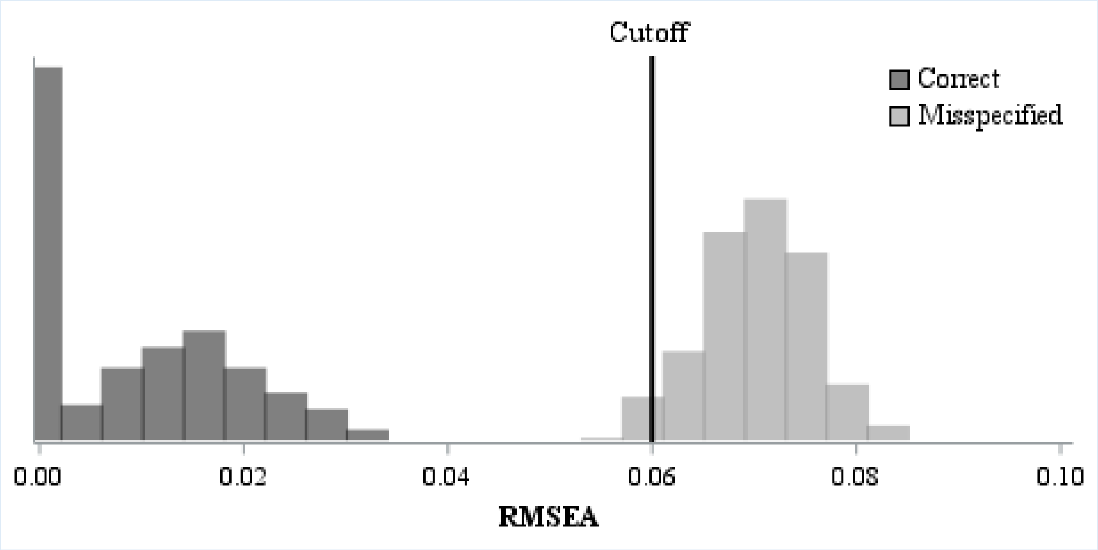
RMSEA distribution when model is correct

RMSEA distribution when model is misspecified



Cutoff is value that is ~95% sensitive to misspecification

(without rejecting more than ~5% of true models)



Basis for

RMSEA < .06
CFI > .95

Hu and Bentler Impact

- Hard to overstate the impact of HB on behavioral sciences
  - Over 110,000 citations
  - Over 13,000 in 2022 alone

- Used as primary source of validity evidence for measurement scales
  - Affects how we conceive latent constructs and how we obtain scores

- Even if researchers are not citing HB directly, they are likely using scales/instruments that cite or are implicitly guided by HB

H&B Citations
In Context

- Hu & Bentler (1999) – 111,296 Google Scholar citations

H&B Citations In Context

- Hu & Bentler (1999) – 111,296 Google Scholar citations

- Marx (1875) *Das Kapital* 68,807
- Smith (1776) *Wealth of Nations* 36,331

- Combined 105,138

H&B Citations In Context

- Hu & Bentler (1999) – 111,296 Google Scholar citations

- Darwin (1869) *Origin of Species* 62,600
- God (5000 BCE) *Holy Bible* ~30,000

- Combined ~92,600

Main Point

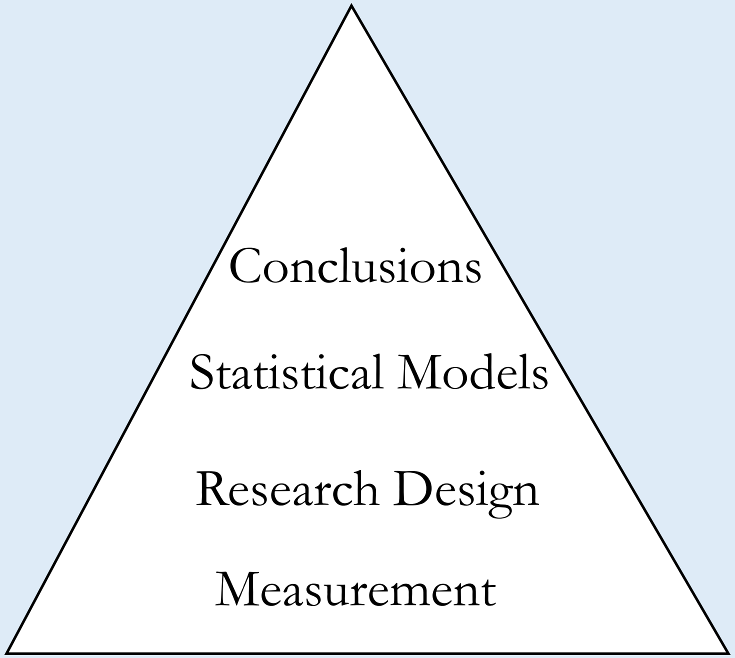
- Measurement is foundational to behavioral sciences because many variables of interest are latent constructs
  - Quality of conclusions limited by quality of the measurement

- Guidelines from Hu and Bentler (1999) essentially have come to determine what is considered "good" or "bad" measurement

- The quality of our body of knowledge is somewhat dependent on the cutoffs from Hu and Bentler (1999) functioning well

Analytic Pyramid

- Conclusions can't be trusted if statistical analysis is done incorrectly
- Statistical Models cannot correct for design flaws
  - Who cares if treatment is significant if group comparisons are confounded?
- Research Design is irrelevant if our measures do not capture what they intend to capture
  - Who cares if groups are comparable if you're comparing noise?



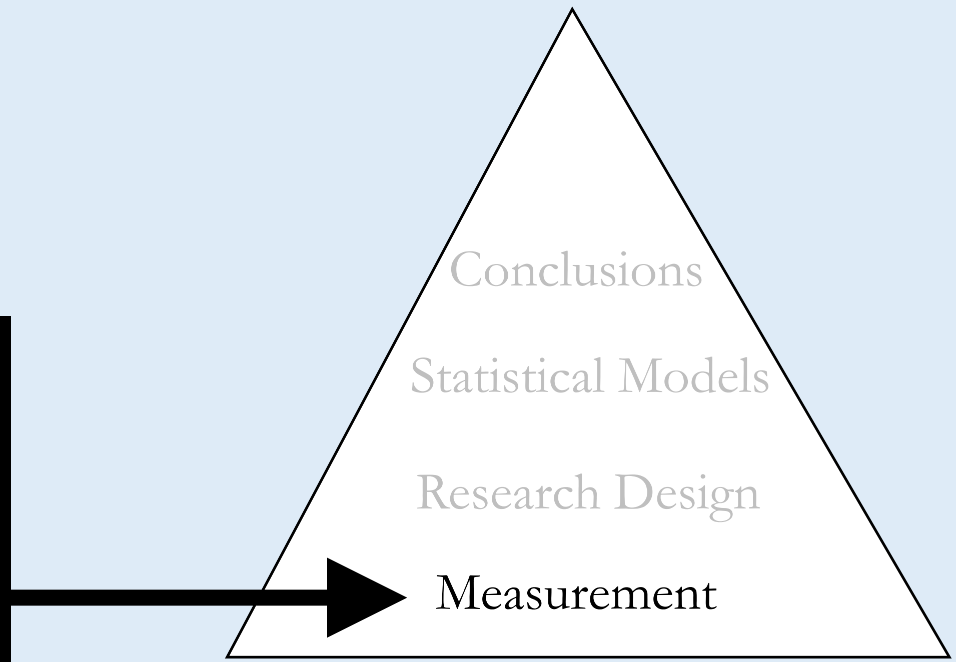
(Adapted from Flake, 2021)

Analytic Pyramid

If we get this wrong, everything else comes crashing down

HB cutoffs have been entrusted as a primary way to evaluate this step

What happens if the HB cutoffs don't work as intended?



(Adapted from Flake, 2021)

It ain’t what you don’t know that gets you into trouble. It’s what you know for sure that just ain’t so – Mark Twain

Potential Issues With Hu & Bentler Cutoffs

Issue

- Despite popularity, HB cutoffs not infallible
- Studies note that if you run the similar simulation with different conditions, <u>the cutoffs change</u>

1. Model Characteristics
   - Degrees of freedom
   - Number of indicators or factors
   - Estimator
   - Loading strength
2. Data Characteristics
   - Missing data
   - Response Scale
   - Categorical vs. Continuous

- Beauducel & Wittman, 2005
- Chen et al. (2008)
- Cole & Preacher, 2014
- Davey, Salva, & Luo, 2005
- Fan & Sivo (2007)
- Grieff & Heene, 2017
- Hancock & Mueller (2011)
- Heene et al. (2011)
- Jackson (2007)
- Kenny & McCoach (2003)
- Kenny et al. (2015)
- Kang et al., (2016)
- Lance et al., 2016
- Marsh et al. (2004)
- McNeish et al. (2018)
- McNeish & Hancock, 2018
- Miles & Shevlin (2007)
- Monroe & Cai 2015
- Nieman & Mai (2018)
- O'Boyle & Williams, 2011
- Saris et al. (2009)
- Savalei (2012)
- Savalei (2020)
- Shi et al. (2019)
- Shi, Lee, & Terry, 2018
- Shi et al., 2018
- Sivo et al. (2006)
- Steiger, 2000
- Williams, et al. 2020
- Williams & O'Boyle, 2011
- Xia & Yang, 2018
- Xia & Yang, 2019
- Zhang & Savalei 2020

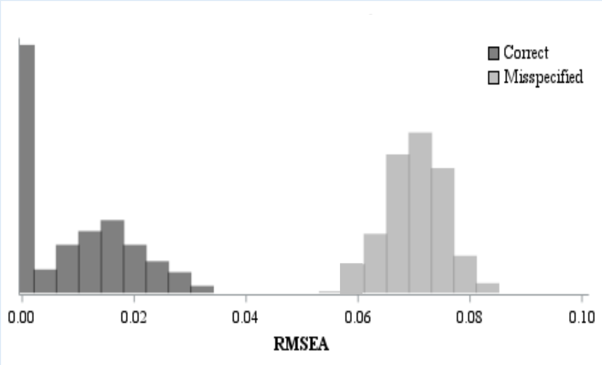
Redo HB with Different Conditions

- Follow HB's protocol exactly

- Change the loadings or the number of items per factor
  - 3, 5, or 7 items per factor
  - .60, .75, or .90 loadings

- Track RMSEA that optimally distinguishes correct from misspecified models

		Items Per Factor		
	Loadings	3	5	7
RMSEA	0.60			
	0.75			
	0.90			



Redo HB with Different Conditions

- Loadings of 0.75 with 5 items per factor were the conditions in the original study
- If you replicate these conditions, you get HB's cutoff of .06 for RMSEA

		Items Per Factor		
	Loadings	3	5	7
RMSEA	0.60			
	0.75		.061	
	0.90			

Redo HB with Different Conditions

- Fewer items lead to more lenient cutoffs to detect the same misspecification

- More items lead to more strict cutoffs to detect the same misspecification

		Items Per Factor		
	Loadings	3	5	7
RMSEA	0.60			
	0.75	.080	.061	.044
	0.90			

Redo HB with Different Conditions

- Stronger loadings lead to more lenient cutoffs to detect the exact same misspecification

- Weaker loadings lead to more strict cutoffs to detect the same misspecification

		Items Per Factor		
	Loadings	3	5	7
RMSEA	0.60		.044	
	0.75	.080	.061	.044
	0.90		.090	

Redo HB with Different Conditions

- Changing multiple conditions simultaneously produces interactions effects

- Multiway interactions make changes in optimal cutoffs difficult to predict

		Items Per Factor		
	Loadings	3	5	7
RMSEA	0.60	.041	.044	.028
	0.75	.080	.061	.044
	0.90	.161	.090	.062

Redo HB with Different Conditions

- Changing multiple conditions simultaneously produces interactions
- Multiway int changes in o difficult to p

Items Per Factor	
5	7
.044	.028
.061	.044
.090	.062

If the goal is for cutoffs to be sensitive to an omitted 0.50 cross-loading, the RMSEA value corresponding to that misspecification changes markedly as a function of model characteristics

An RMSEA of 0.06 has <u>very</u> different sensitivity to misfit depending on characteristics

Hu & Bentler (1999)
As a Power Analysis

Power Analysis

- HB simulation was essentially a power analysis
- In power analysis for sample size planning, goal is to determine *N* where a test is 80% sensitive to a non-null effect of predefined size
- In HB, goal was to determine fit index value that is ~95% sensitive to predefined misspecification

Power Analysis

<u>Underlying idea is the same</u>:

Determine value of a target quantity that optimizes sensitivity to the presence of some phenomena

<u>Power Analysis</u>
Target : N
Phenomena : Non-Null Group Difference

<u>Hu & Bentler</u>
Target: RMSEA, CFI
Phenomena: Meaningful Misspecification

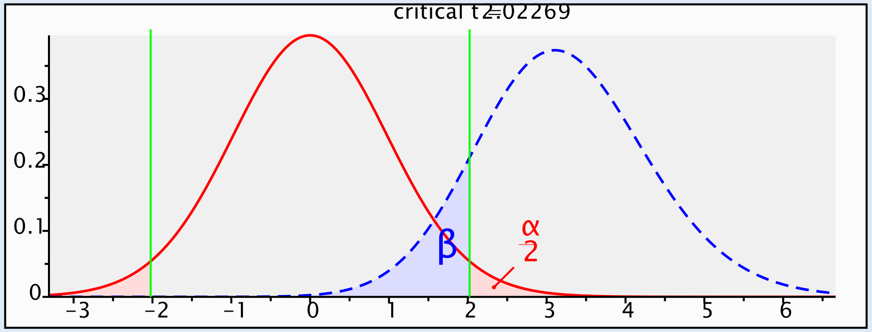
- HB s
- In po … determine N wh … f prede
- In H … 95% sensi

Sample Size Planning

- Required sample size (target quantity) changes based on model/design characteristics
  - Within-subjects designs need smaller N than between-subjects

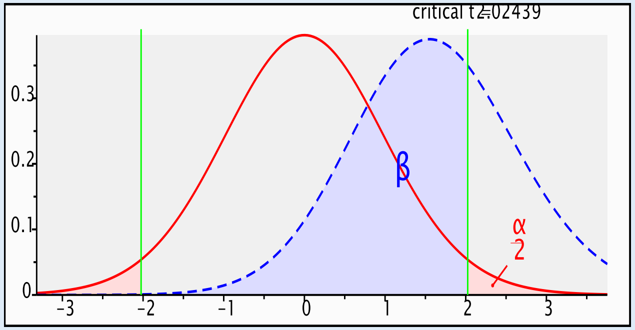
- No single universal N that satisfies all scenarios

- Custom power analysis reflect how different conditions affect N
  - Produces optimal and efficient N for specific scenario



Within-Subjects
N =40
d=0.50

Power = 87%

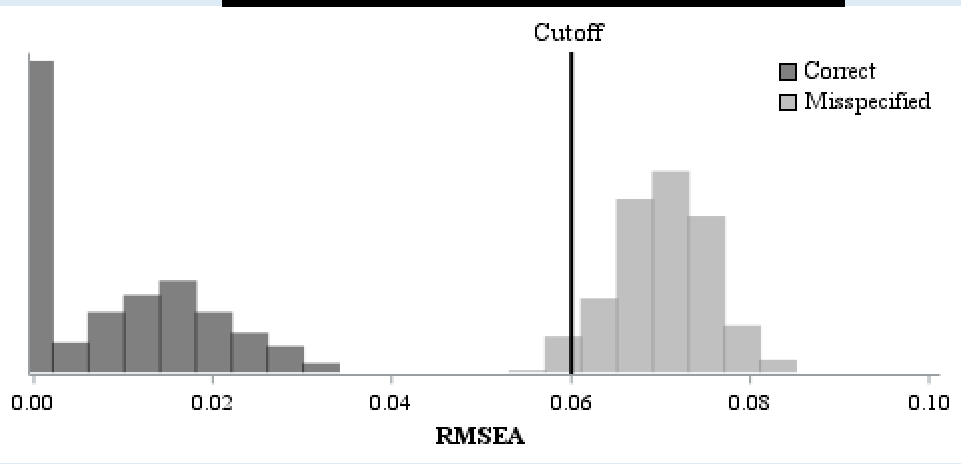


Between-Subjects
N =40
d=0.50

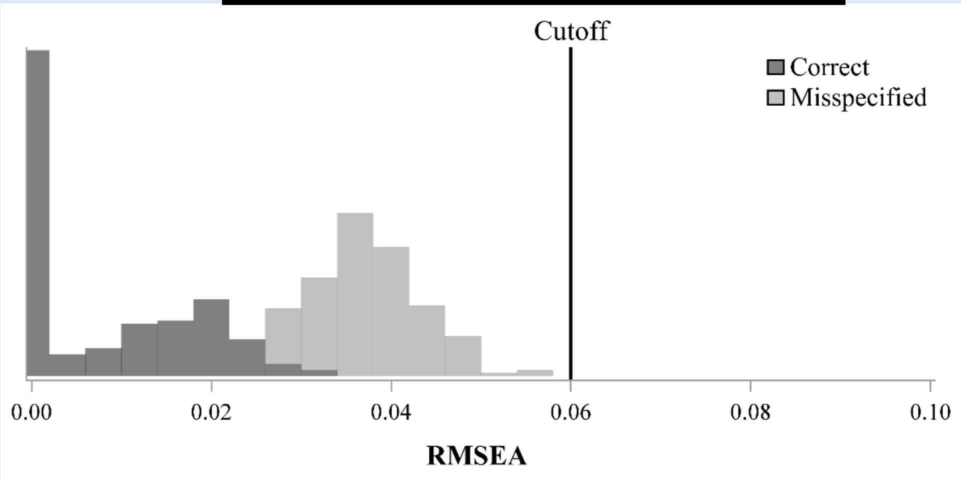
Power = 34%

Same Idea for Fit Indices

5 Items, 0.75 loadings



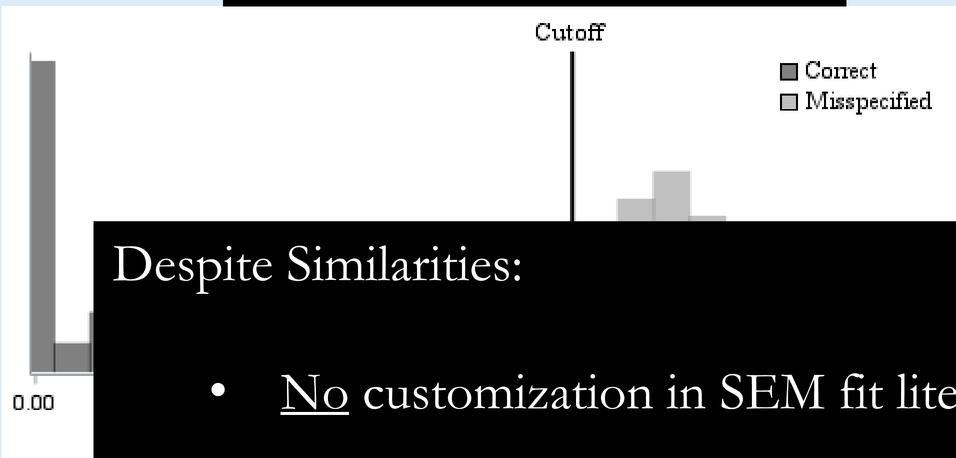
5 Items, 0.55 loadings



- Location and Dispersion of these distributions change
- Cutoffs that optimally distinguishes between distributions will also change
- E.g., HB cutoffs has ~100% sensitivity in top plot but 0% sensitivity in bottom plot

Same Idea for Fit Indices

5 Items, 0.75 loadings



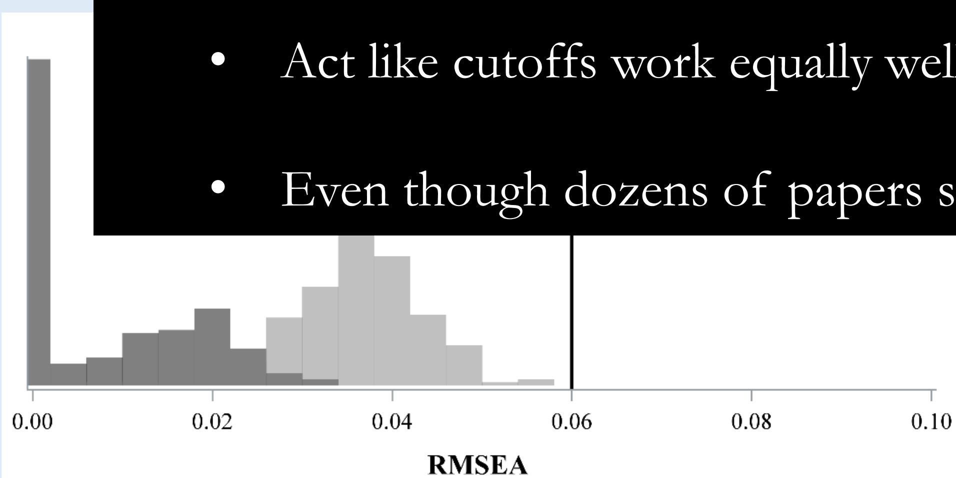
Despite Similarities:

- No customization in SEM fit literature
- Generalize cutoffs from single simulation 25 years ago
- Act like cutoffs work equally well in all scenarios
- Even though dozens of papers show that they don't

on of these

distinguishes
will also

100%
but 0%
sensitivity in bottom plot



If Sample Size Planning Were Like Fit Indices …

- Imagine a simulation in 1997
- Finds that 80% power in independent samples t-test with $d = 0.50$ occurs at N=120
- Every study now uses N=120 thinking that makes the study adequately powered, for any design
- Non-significant tests would be ambiguous
  - Is the effect null?
  - Or was there insufficient power to detect non-null effect?

Ambiguity in Scale Validation

- However, this is essentially how scale validation currently operates

- Current use of HB cutoffs make it hard to interpret scale validations

- HB cutoffs confounds model characteristics with misfit
  - Does the model actually fit?
  - Or does the model just have characteristics where HB cutoffs are not sensitive to misfit?

Ambiguity in Scale Validation

- However, this is essentially how scale validation currently operates
- Current u

“Our primary conclusion is simple. If you wish your model to fit … ensure that your measures are unreliable” (Miles & Shevlin, 2007, p. 874)

cale validations

- HB cutoffs confounds model characteristics with misfit
    - Does the model actually fit?
    - Or does the model just have characteristics where HB cutoffs are not sensitive to misfit?

- Many studies have pointed out these issues with little change to practice
- Few proposed & accessible alternatives to use instead
- Major challenge is therefore to create new methods that bridge methodological and empirical research
    - Otherwise, we'll just continue complaining to each other while being ignored by empirical researchers

Dynamic Fit
Index Cutoffs

Making SEM More Like Power Analysis

- Millsap (2007) pointed out similarities of cutoffs and power analysis
  - Proposed deriving custom cutoffs for every model
  - Published multiple papers/chapters on the idea
  - However, the idea did not catch on

- Presumably, issues are that custom simulation is hard to do
  - Empirical researchers don't know simulation
  - A lot of work for a methodologist to do from scratch/low incentives
  - Also requires defining a misspecification to which indices should be sensitive

- Custom simulation is inaccessible and HB cutoffs are too easy/accepted

Dynamic Fit Index Cutoffs

- DFI is a framework/software to try to make custom simulation more accessible by

  1. Trying to reproduce HB's simulation for your model
  2. Using an algorithm to internally determine a misspecification to test
  3. Automating writing fit index Monte Carlo code
  4. Automating execution of Monte Carlo code
  5. Collate all results

Dynamic Fit Index Cutoffs

- DFI is a framework/software to try to make custom simulation more accessible by

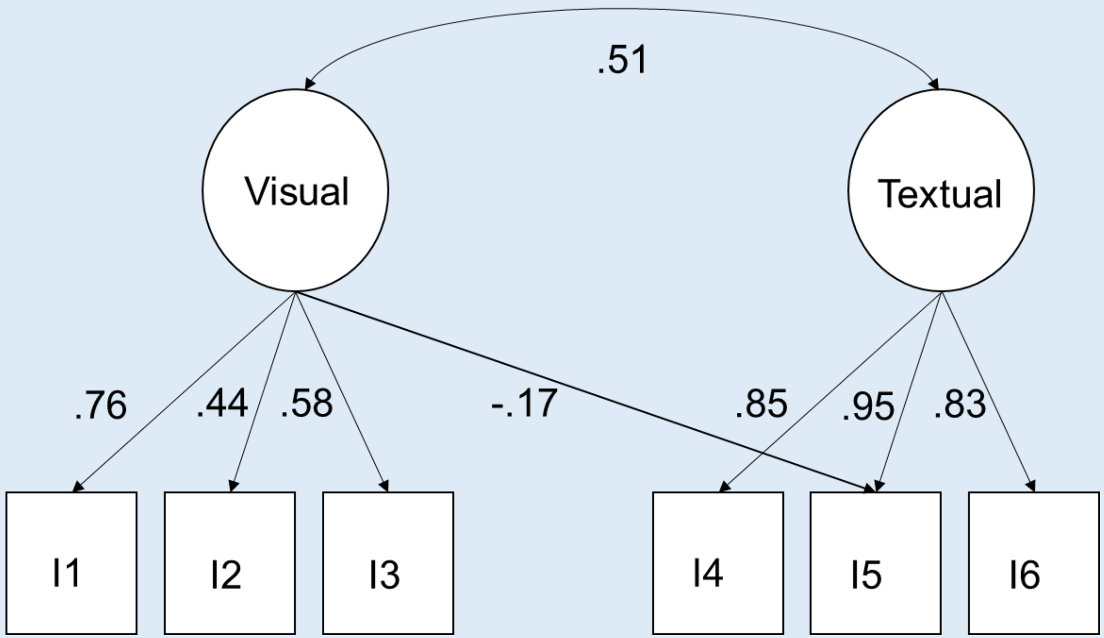
DFI is essentially a power analysis package for fit indices

G*Power for fit indices

Makes otherwise complicated process accessible to empirical researchers by automating the difficult parts

1. T
2. U … to test
3. A
4. A
5. C

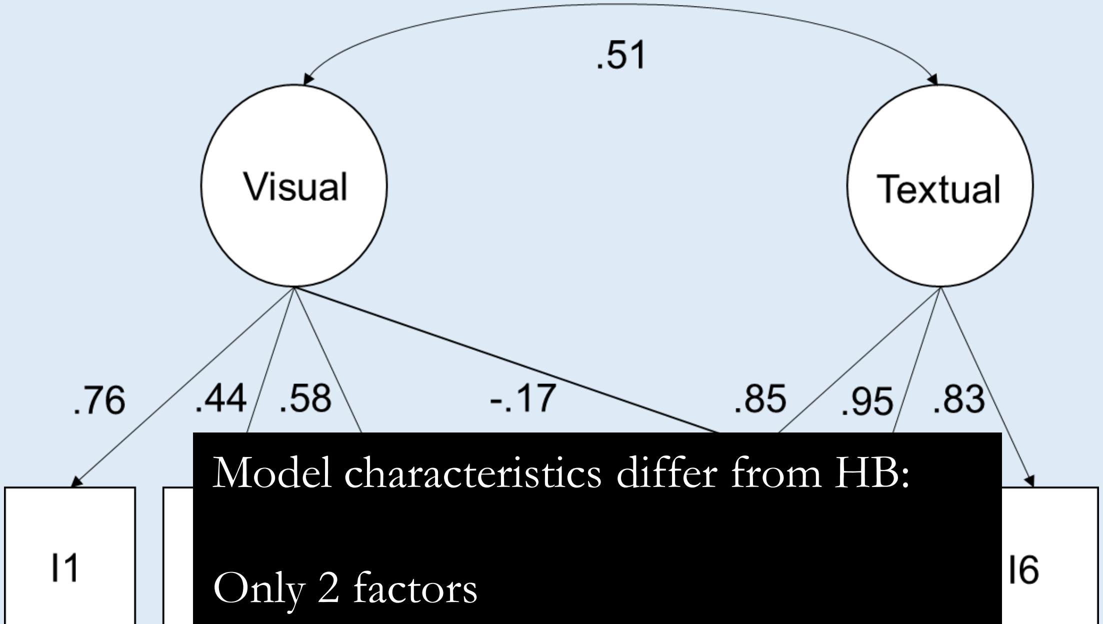
Holzinger & Swineford (1939)



$\chi^2(7)=16$	$p = .026$
SRMR	.043
RMSEA	.065
CFI	.986

- Model does not fit exactly (N = 301)
- SRMR and CFI satisfy HB cutoffs
- RMSEA is close

Holzinger & Swineford (1939)



$\chi^2(7)=16$	$p= .026$
SRMR	.043
RMSEA	.065
CFI	.986

- Mo 1)
- SR
- RM

Model characteristics differ from HB:

Only 2 factors
Only 6 items (vs 15 in HB)
Only 3 items per factor (vs. 5 in HB)
Only 7 df (vs 87 in HB)
Heterogeneous loadings (vs .70 -.80)

Unclear how well HB cutoffs might perform with these model characteristics

Three Options

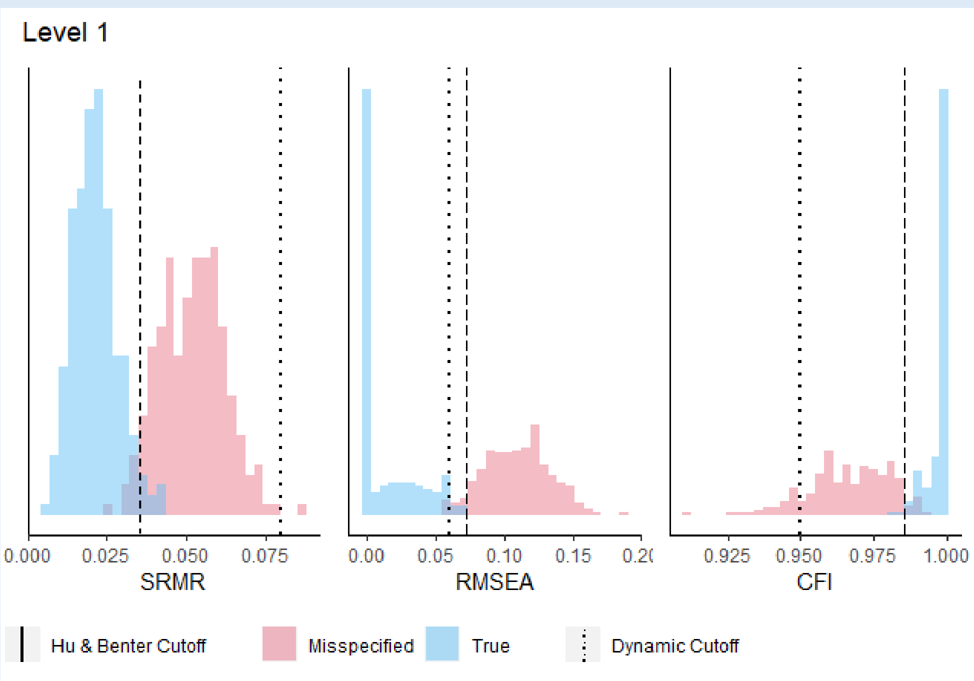
1. Directly Interface with `lavaan` Object

```
HS.model <- '
visual =~  x1 + x2 + x3 + x5
textual =~ x4 + x5 + x6 '

fit <- lavaan::cfa(HS.model, data = dat)

dynamic::cfaHB(fit)
```

- Store `lavaan` result as an object
- Use object as argument in function from `dynamic` R package
  - `cfaHB` is function for multifactor CFA models, mimicking HB
  - `cfaOne` is function for one-factor CFA

Output

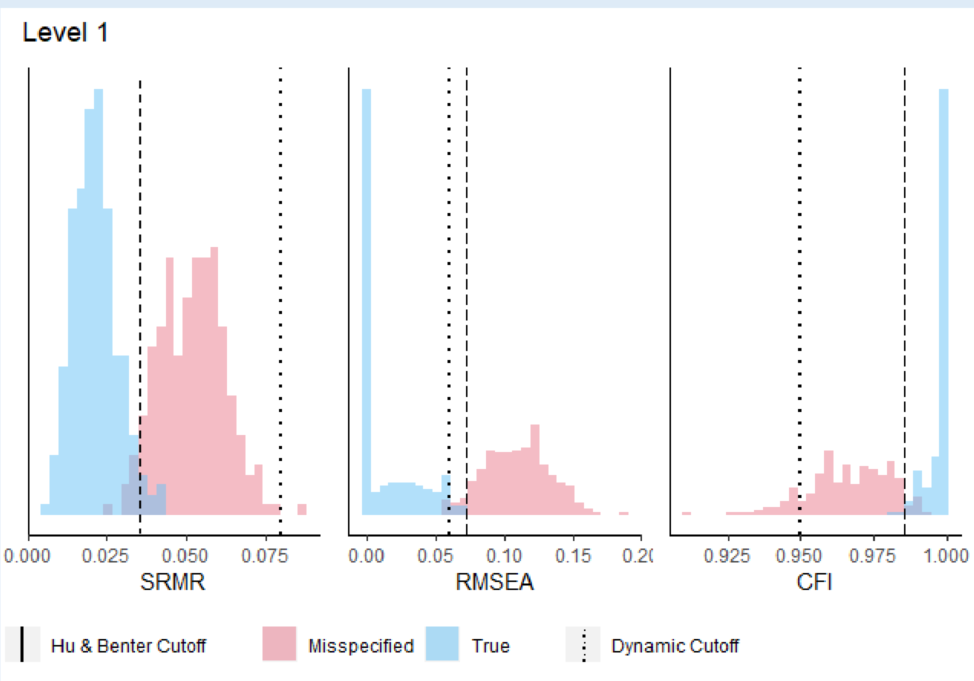
```
Your DFI cutoffs:
                SRMR RMSEA  CFI Magnitude
Level 1: 95/5  .036  .073 .986       .436
Level 1: 90/10   --    --   --

Empirical fit indices:
 Chi-Square  df p-value   SRMR   RMSEA    CFI
     15.917   7   0.026  0.043   0.065  0.986

> Sys.time()-start Time difference of 21.70764 secs
```



- Cutoffs sensitive to an omitted 0.436 cross-loading:
  - SRMR < .036
  - RMSEA < .073
  - CFI > .986
- Also shows plots of fit indices from simulation when fitted assumed was assumed correct (in blue) and assumed misspecified (in red)

Output

```
Your DFI cutoffs:
                 SRMR RMSEA  CFI Magnitude
Level 1: 95/5    .036  .073 .986      .436
Level 1: 90/10     --    --   --

Empirical fit indices:
 Chi-Square  df p-value   SRMR   RMSEA    CFI
     15.917   7   0.026  0.043   0.065  0.986
> Sys.time()-start Time difference of 21.70764 secs
```



- Cutoffs sensitive to an omitted 0.436 cross-loading:
  - SRMR < .036
  - RMSEA < .073
  - CFI > .986

Even though CFI looks great according to HB, CFI appears to be less receptive under these conditions, so the cutoff is stricter

Conversely, RMSEA is more receptive under these conditions, so the cutoff is larger than HB

Three Options

2. Manually write out model in R

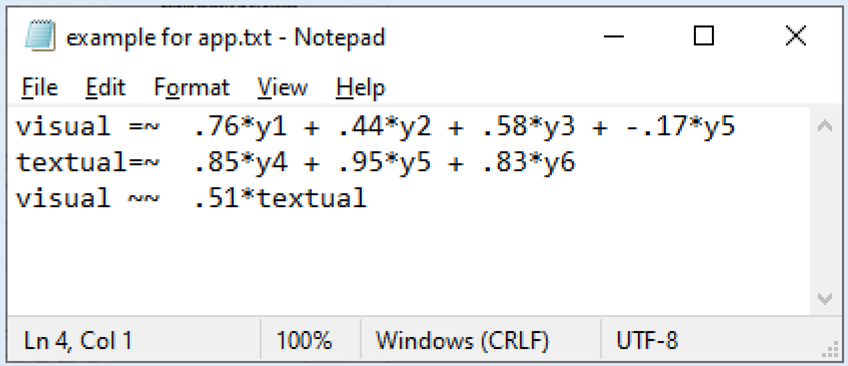
```
HS<- '
visual   =~  .76*x1 + .58*x2 + .44*x3 + -.17*x5
textual =~  .85*x4 + .95*x5 + .83*x6
visual   ~~  .51*textual '

dynamic::cfaHB(HS, n= 301, manual=TRUE)
```

- If not a `lavaan` user, the model can be written out with the standardized estimates from another software

- Use this as the object in the `dynamic` function, include the sample size and `manual=TRUE` to let software know the model was manually entered

Three Options

3. Manually write out model in Shiny app


example for app.txt - Notepad

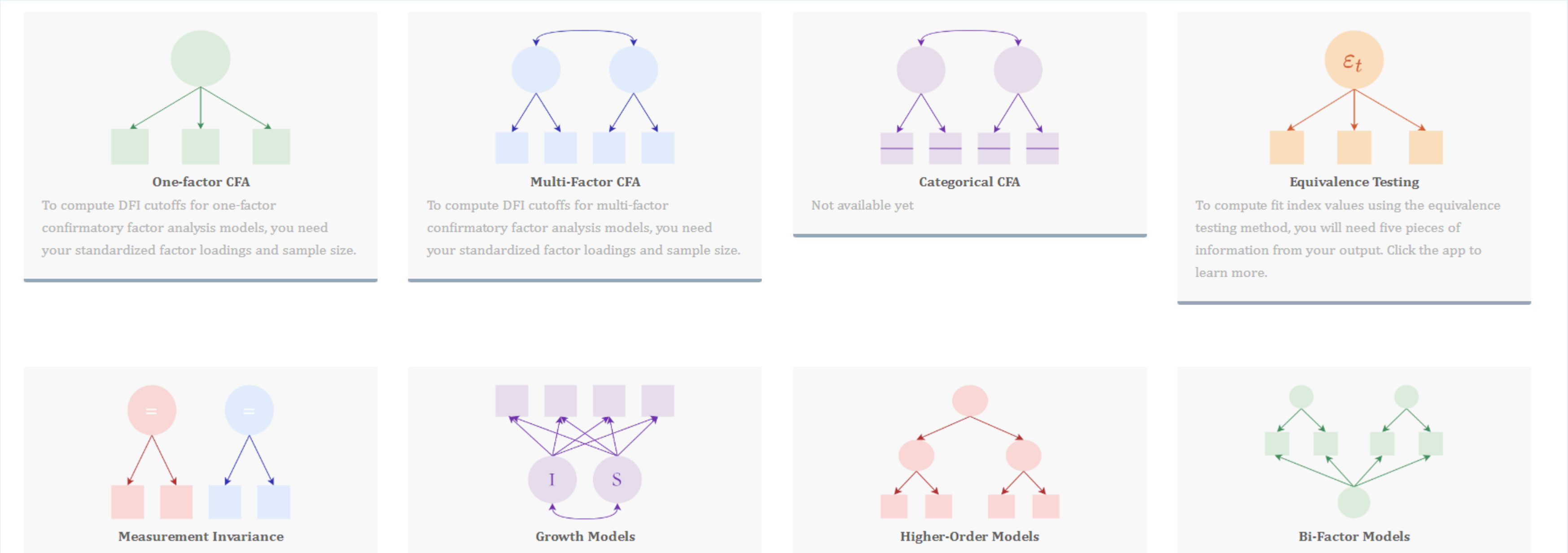
File Edit Format View Help

```
visual =~  .76*y1 + .44*y2 + .58*y3 + -.17*y5
textual=~  .85*y4 + .95*y5 + .83*y6
visual ~~  .51*textual
```

Ln 4, Col 1 | 100% | Windows (CRLF) | UTF-8

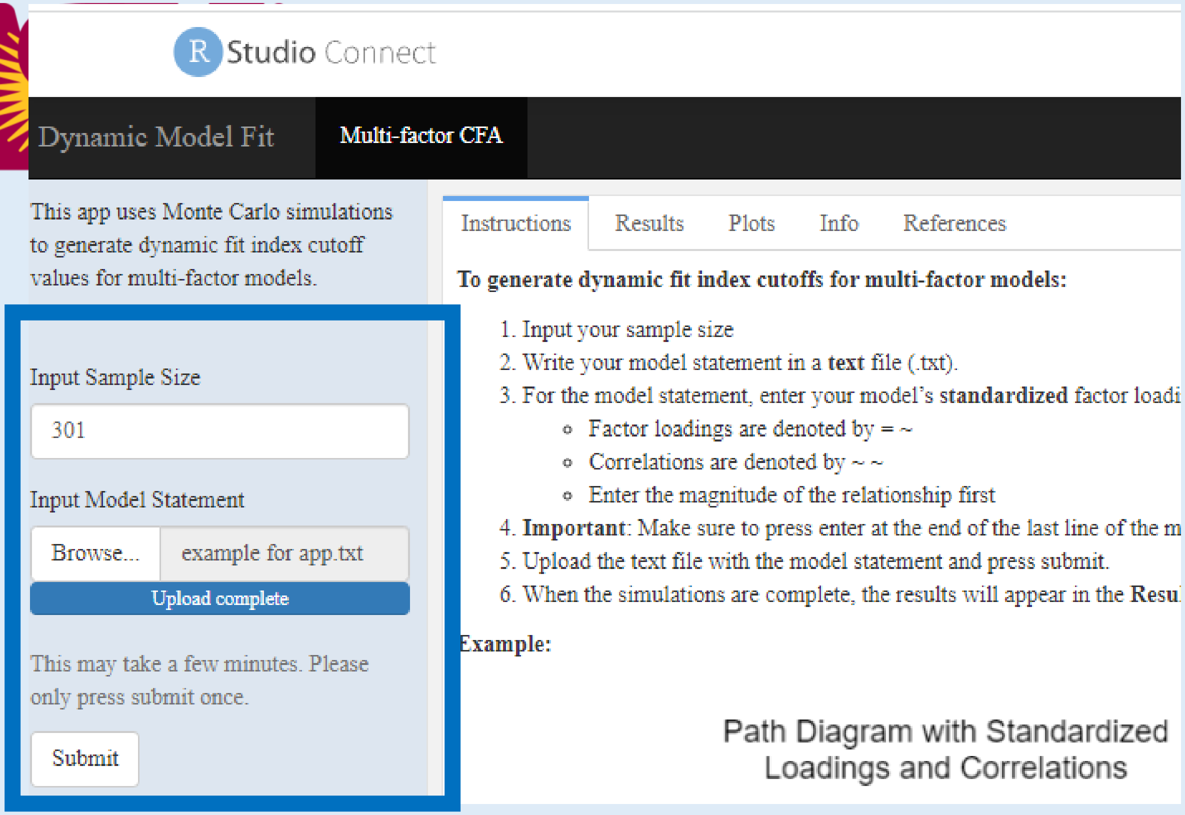
- Write model with standardized estimates in a .txt
- Go to www.dynamicfit.app

Three Options



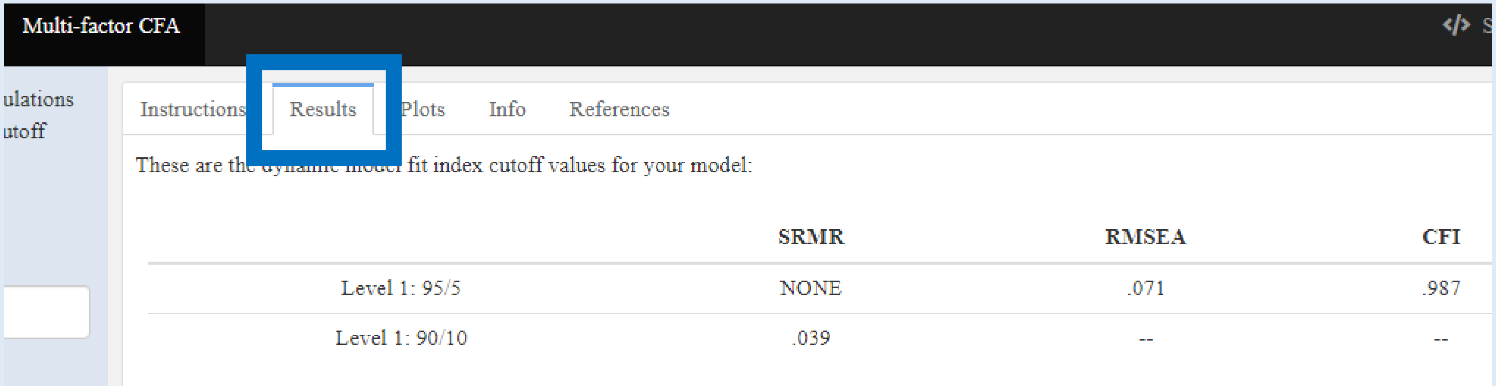
- Choose your model type
  - Not all model types available in R package are currently available on Shiny app

Three Options



- Enter your sample size
- Upload the .txt file in the “input model statement’ box
- Click submit
  - Shiny app runs virtually and is slower than running DFI locally in R
  - Shiny also has fewer options than R package (e.g., estimator, how many replications)

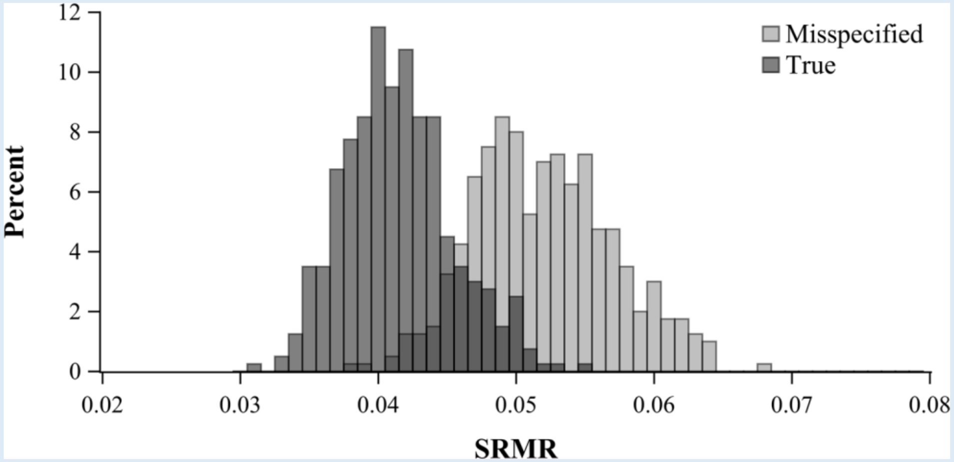
Results Tab



	SRMR	RMSEA	CFI
Level 1: 95/5	NONE	.071	.987
Level 1: 90/10	.039	--	--

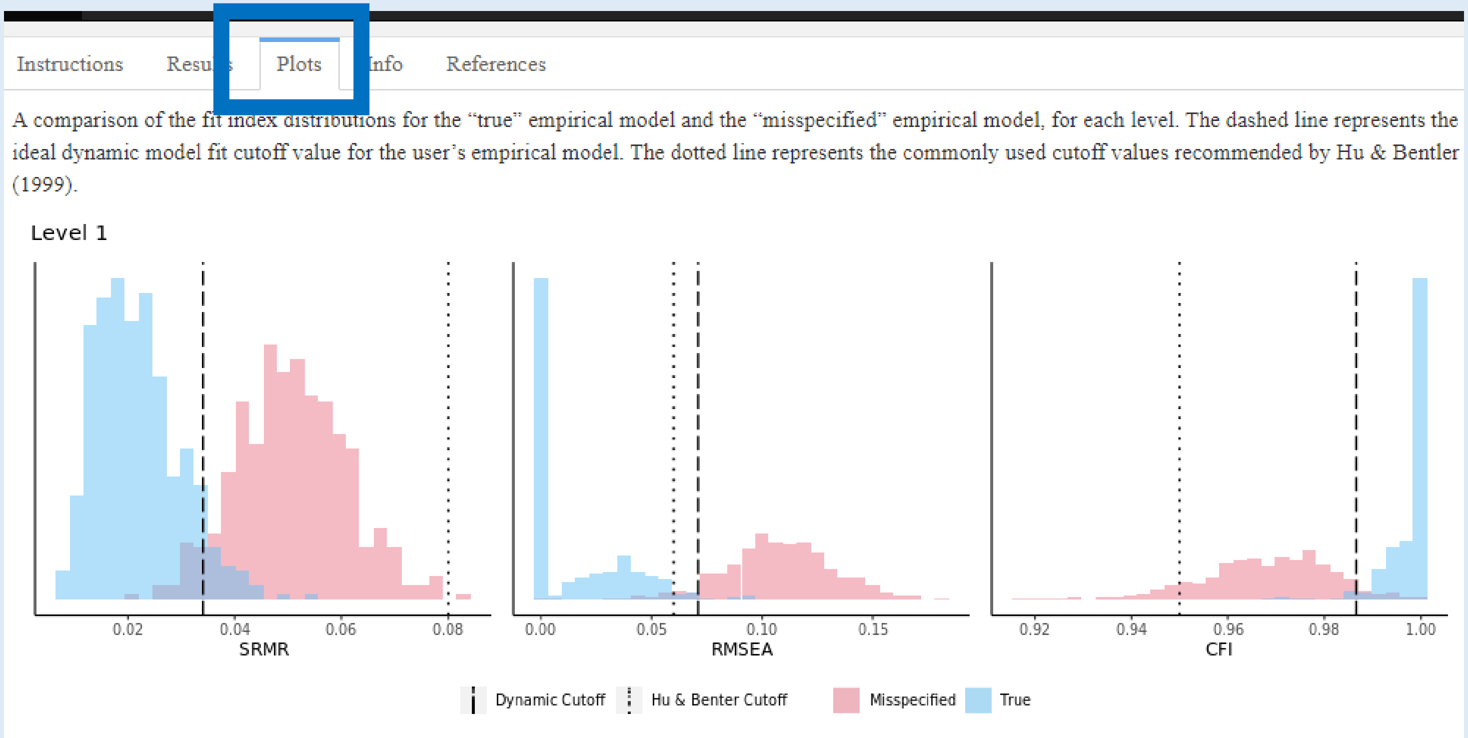
- Very similar cutoffs
- Small difference in cutoffs due to rounding loadings
- SRMR also only has 90% sensitivity rather than 95% sensitivity

Unavailable DFI cutoffs



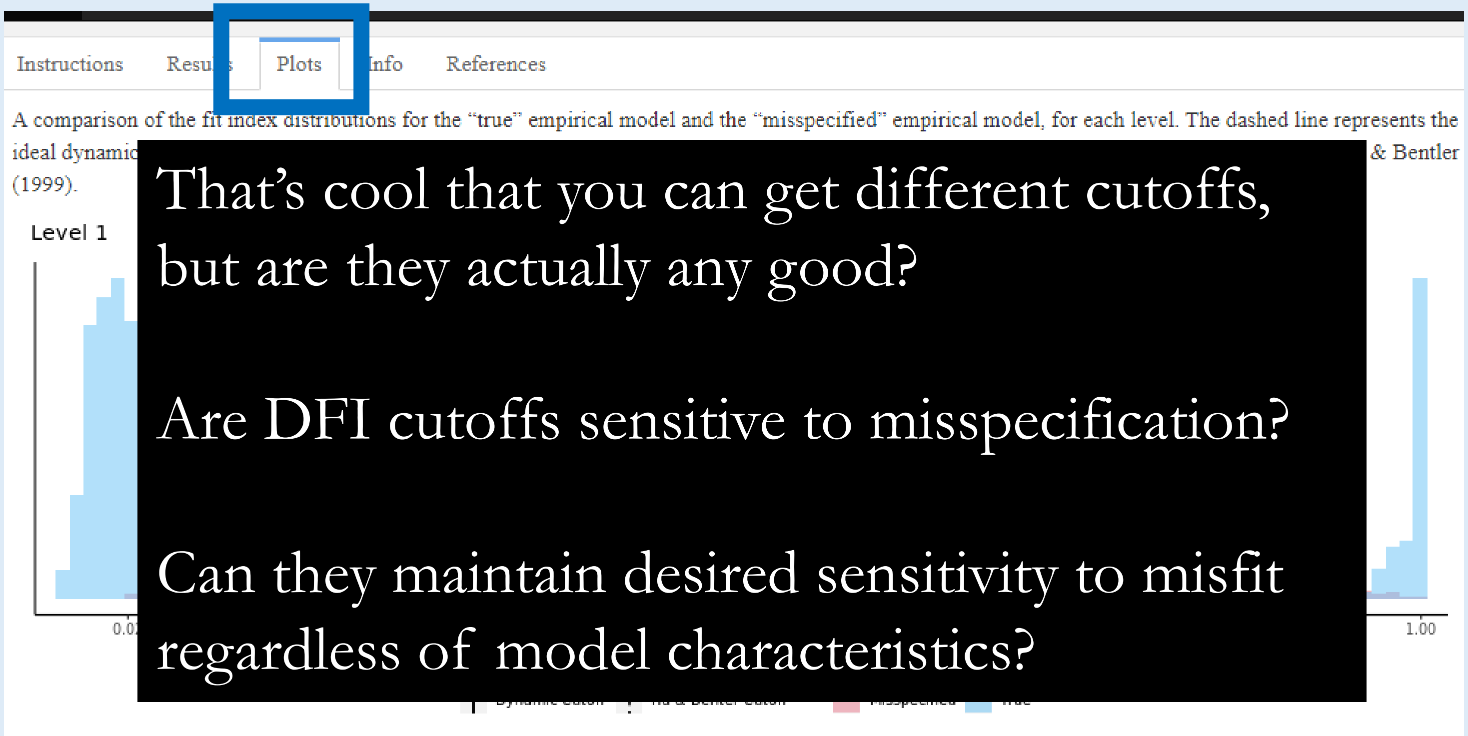
- When sampling variability is large, distributions overlap
- Not always a value that is consistently able to reject misspecified models and retain correct models
- If this occurs, DFI will try to find cutoffs with slightly lower sensitivity
- If sensitivity is too low, no cutoffs will be produced

Plots Tab



- Same plots as before

Plots Tab



That’s cool that you can get different cutoffs, but are they actually any good?

Are DFI cutoffs sensitive to misspecification?

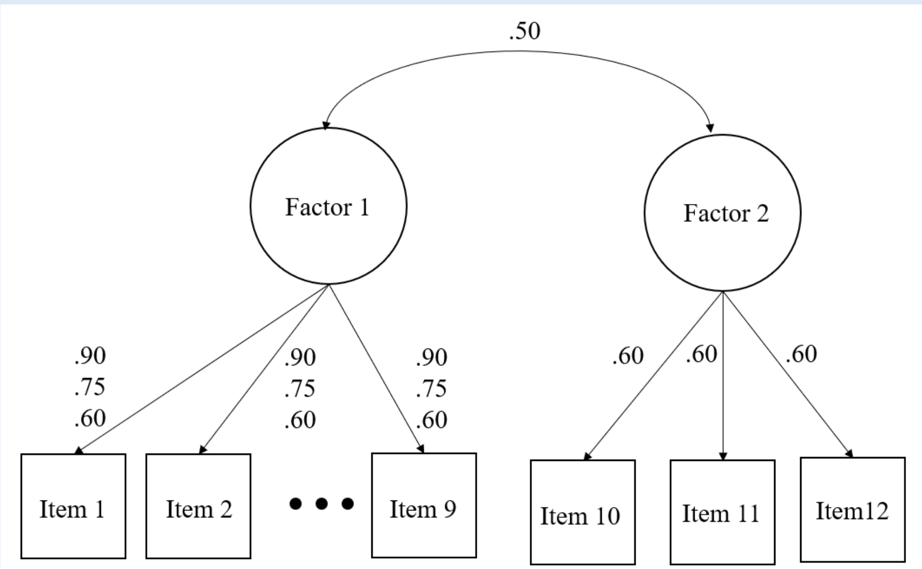
Can they maintain desired sensitivity to misfit regardless of model characteristics?

- Same plots as before

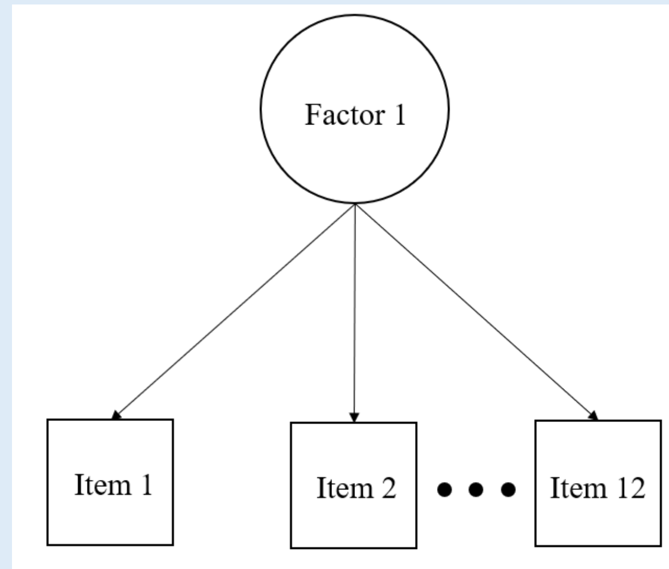
One
Factor CFA

Simulation

Multidimensional



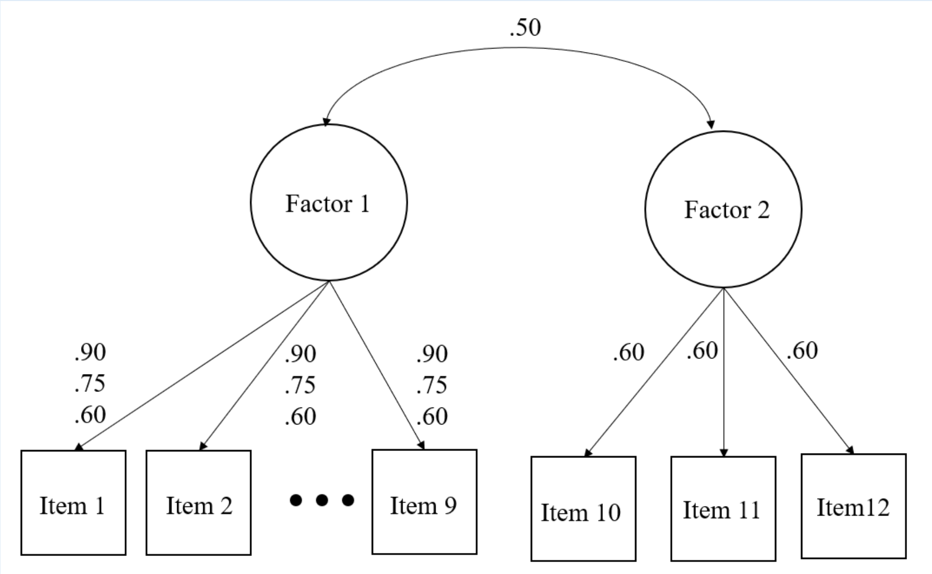
Unidimensional



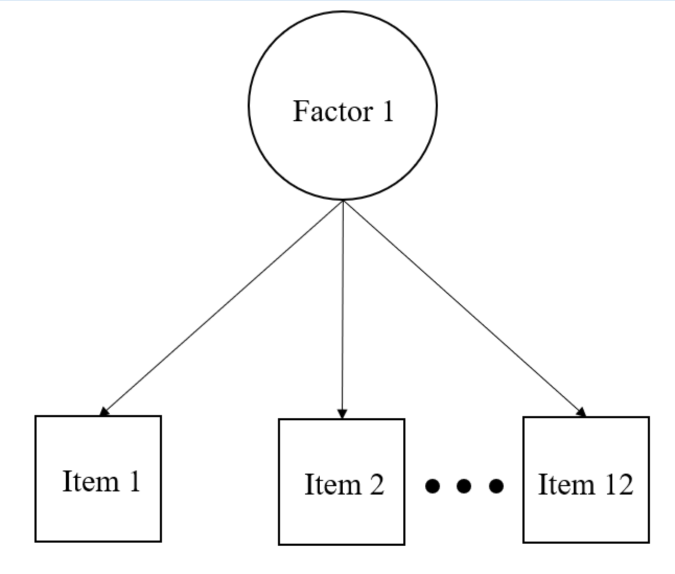
- Generate uni or multidimensional data
- Fit one factor model to all data
- Cutoffs should reject one-factor model for multidim data
- Cutoffs should not reject one-factor model for unidim data

Simulation

Multidimensional

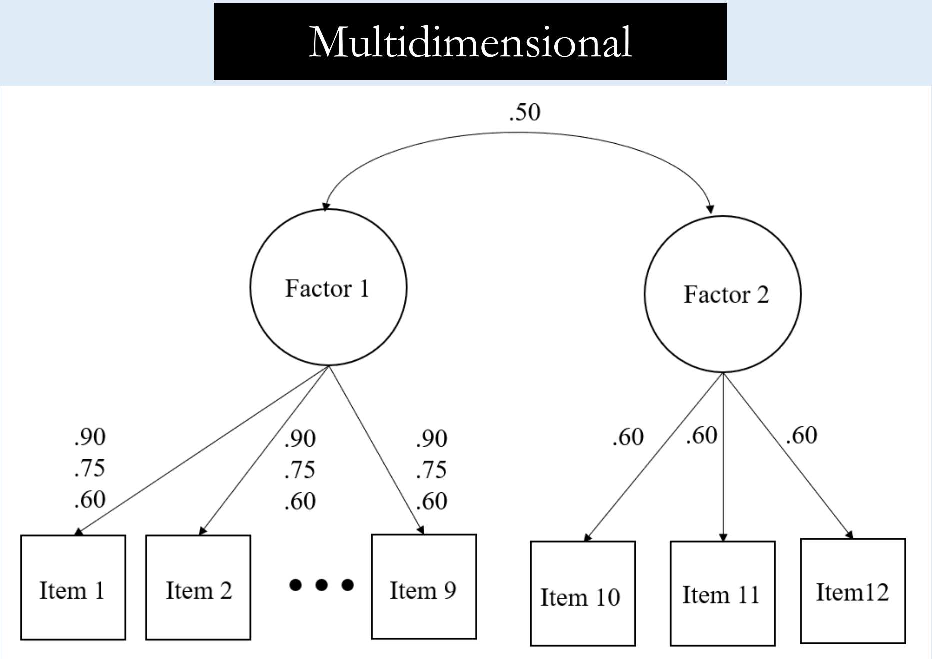


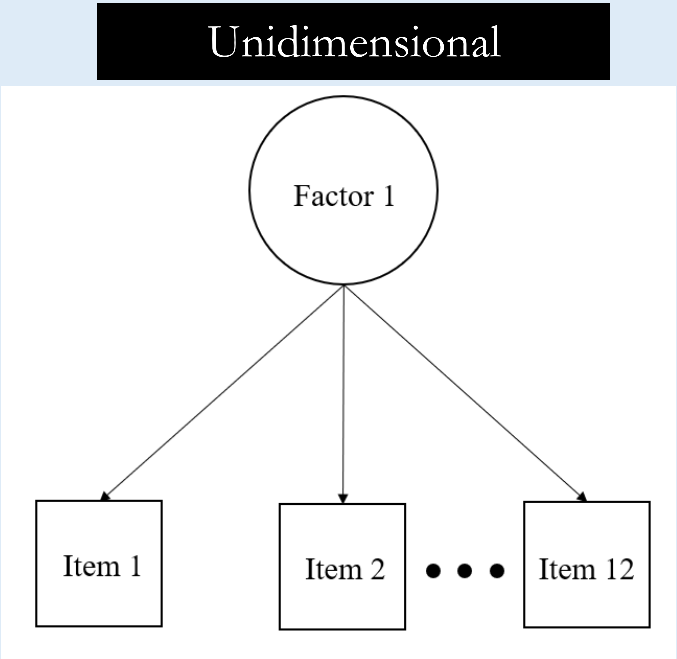
Unidimensional



- Items= 8 or 12
- N = 250 or 400
- Loadings = 0.60, 0.75, or 0.90
  - Only loadings for Factor 1 items are manipulated

Simulation





- Factor 2 items are constant at 0.60
  - Keeps the magnitude of misspecification constant across conditions
  - Makes the standardized loadings in fitted model around 0.30
  - I.e., a one-factor model with six .75 loadings and two .30 loadings is plausible rather than six .75 loadings and two .10 loadings

Rejection Rates: Unidimensional Data

		8 items			12 items		
Load	Cutoff	SRMR	RMSEA	CFI	SRMR	RMSEA	CFI
.90	DFI	0%	1%	2%	0%	0%	0%
	HB	0%	0%	0%	0%	0%	0%
.75	DFI	0%	2%	2%	0%	0%	0%
	HB	0%	0%	0%	0%	0%	0%
.60	DFI	0%	2%	3%	0%	0%	0%
	HB	0%	0%	0%	0%	0%	0%

When data are truly unidimensional, either cutoff rarely rejects the model

HB Rejection Rates: Multidimensional Data

		8 items				12 items	
N	Load	SRMR	RMSEA	CFI	SRMR	RMSEA	CFI
250	.90	0%	84%	1%	0%	75%	0%
	.75	1%	85%	51%	0%	73%	58%
	.60	3%	84%	93%	0%	70%	99%
400	.90	0%	92%	0%	0%	71%	0%
	.75	0%	90%	49%	0%	69%	58%
	.60	0%	91%	97%	0%	66%	95%

The HB SRMR cutoff has almost no ability to detect that one-factor model is inappropriate for multidimensional data

HB Rejection Rates: Multidimensional Data

		8 items			12 items		
N	Load	SRMR	RMSEA	CFI	SRMR	RMSEA	CFI
250	.90	0%	84%	1%	0%	75%	0%
	.75	1%	85%	51%	0%	73%	58%
	.60	3%	84%	93%	0%	70%	99%
400	.90	0%	92%	0%	0%	71%	0%
	.75	0%	90%	49%	0%	69%	58%
	.60	0%	91%	97%	0%	66%	95%

HB RMSEA cutoff is sensitive to misfit with 8 items, less so for models with 12 items

HB Rejection Rates: Multidimensional Data

		8 items				12 items	
N	Load	SRMR	RMSEA	CFI	SRMR	RMSEA	CFI
250	.90	0%	84%	1%	0%	75%	0%
	.75	1%	85%	51%	0%	73%	58%
	.60	3%	84%	93%	0%	70%	99%
400	.90	0%	92%	0%	0%	71%	0%
	.75	0%	90%	49%	0%	69%	58%
	.60	0%	91%	97%	0%	66%	95%

HB CFI cutoff depends heavily on loadings.
0% sensitivity for strong loadings,
100% sensitivity for weak loadings

DFI Rejection Rates: Multidimensional Data

		8 items			12 items		
N	Load	SRMR	RMSEA	CFI	SRMR	RMSEA	CFI
250	.90	99%	97%	97%	100%	99%	99%
	.75	100%	96%	96%	100%	99%	100%
	.60	100%	97%	98%	100%	98%	99%
400	.90	99%	99%	99%	100%	96%	96%
	.75	100%	99%	99%	100%	99%	97%
	.60	100%	100%	99%	100%	97%	97%

DFI cutoffs are consistently sensitive to misspecification across all conditions, for all indices

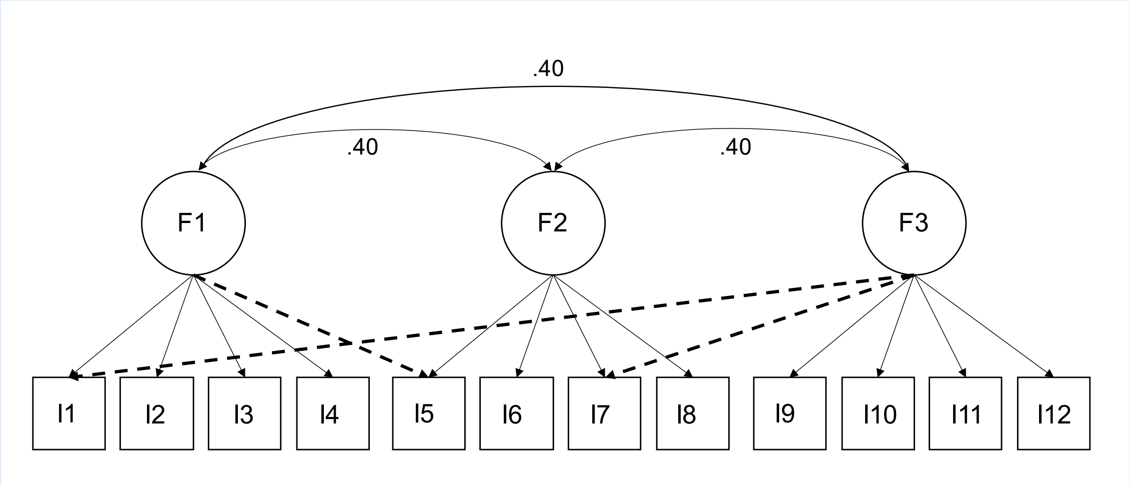
DFI Rejection Rates: Multidimensional Data

		8 items			12 items		
N	Load	SRMR	RMSEA	CFI	SRMR	RMSEA	CFI
250	.90					99%	99%
	.75					99%	100%
	.60					98%	99%
400	.90	99%	99%	99%	100%	96%	96%
	.75	100%	99%	99%	100%	99%	97%
	.60	100%	100%	99%	100%	97%	97%

DFI adapts to whatever the conditions are to provide cutoffs that are appropriately scaled to be sensitive to misspecification

Three
Factor CFA

Simulation



- 25% of items have 0.30 cross-loadings
- Fitted model omitted all cross loadings
- Items = 12 or 24
- Loadings = 0.60, 0.70, or 0.80
- N = 400 or 1000

HB Rejection Rates: Multidimensional Data

		12 Items		24 Items	
N	Load	RMSEA	CFI	RMSEA	CFI
400	.80	100%	100%	100%	100%
	.70	98%	96%	78%	100%
	.60	37%	84%	2%	100%
1000	.80	100%	100%	100%	100%
	.70	100%	100%	89%	100%
	.60	26%	97%	0%	100%

CFI not greatly affected in these conditions

RMSEA sensitivity varies from 0 to 100%

DFI Rejection Rates: Multidimensional Data

		12 Items		24 Items	
N	Load	RMSEA	CFI	RMSEA	CFI
400	.80	100%	100%	100%	100%
	.70	94%	99%	100%	100%
	.60	97%	96%	100%	100%
1000	.80	100%	100%	100%	100%
	.70	96%	100%	100%	100%
	.60	97%	100%	99%	100%

DFI consistently sensitive to omitted cross-loadings regardless of conditions or index

Categorical Responses

Categorical Factor Analysis

- HB only address ML estimation with continuous responses

- Categorical factor analysis uses limited information estimators like WLSMV or ULSMV

- Applying HB cutoffs to limited information estimators for categorical data leads to poor results

Categorical Factor Analysis

- HB only a … sponses
- Categorica … estimators like WLSM…
- Applying … ors for categorica…

"Applying the conventional cutoffs to ULS and DWLS can lead in the long run to the accumulation of models with severe misfit that are nonetheless considered acceptable.

[fit indices] all appear to be insensitive to model misspecification if Hu and Bentler's cutoff values are applied"

Xia & Yang (2019) p. 420-421

Categorical Responses

- To date, no alternative cutoffs for categorical models have been suggested

- Simulations show that there is not single cutoff because sensitivity to misfit is a function of data characteristics like number of categories and the distribution of the responses

- Also, which estimator is used (ULS vs DWLS)

Categorical Responses

- To date, no alternative cutoffs for categorical models have been suggested

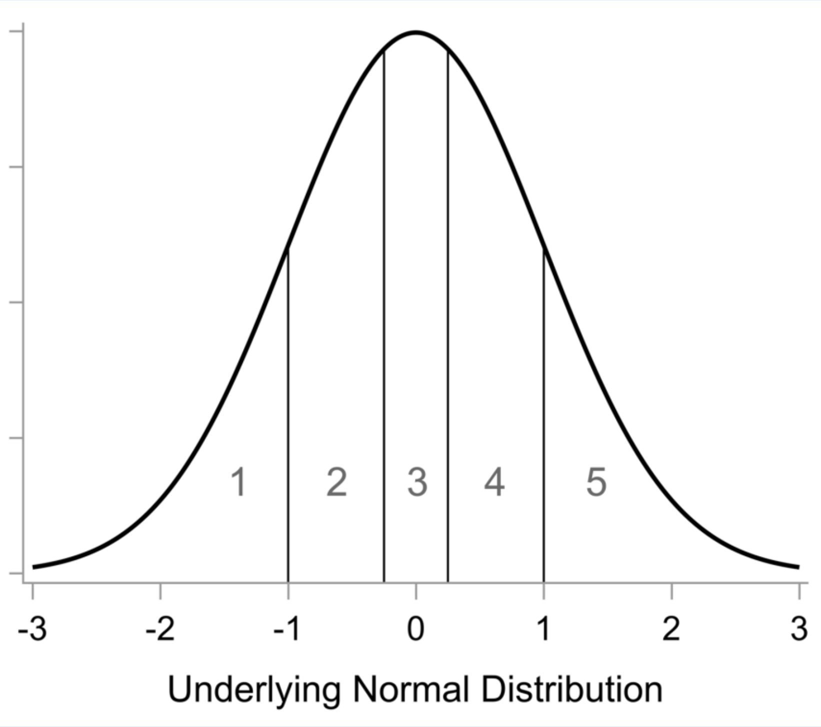
If no single cutoff exists because there is too much fluctuation, custom simulation and DFI might be helpful to produce cutoffs tailored to the specific conditions

- Simulat[illegible]ause sensitiv[illegible]tics like numbe[illegible]sponses

- Also, which estimator is used (ULS vs DWLS)

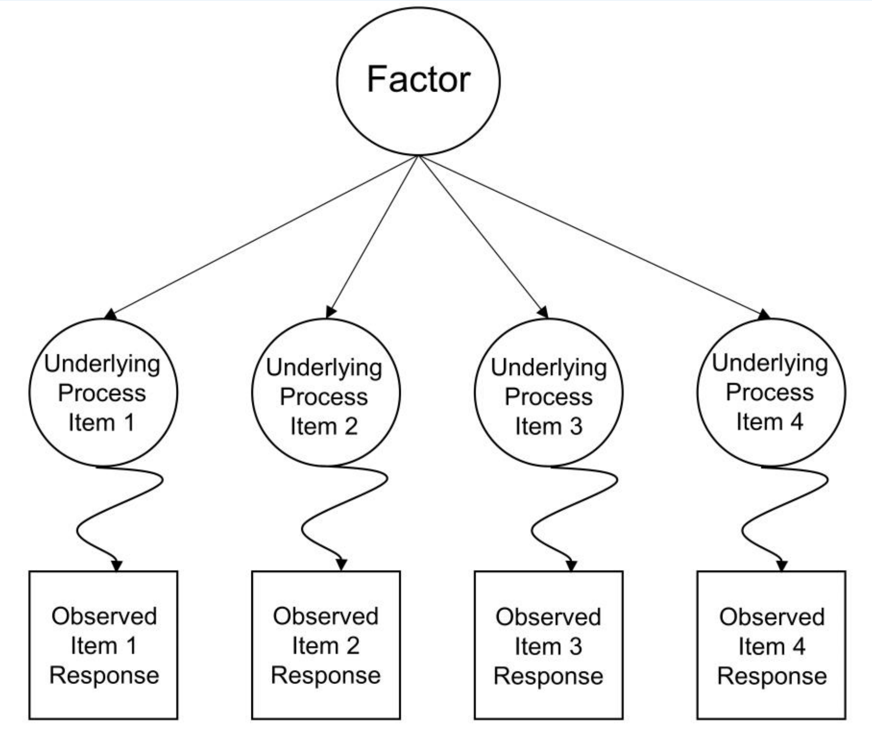
Limited Information Estimators

- Basic idea is that categorical responses are considered a coarse discretization of a truly normal process
- The true response is continuous, but the response scale forces responses into bins
- The cutoff between bins are called <u>thresholds</u>



Limited Information Estimators

- Categorical factor analysis decomposes the polychoric correlations between the latent underlying processes
- As opposed to covariance between observed categorical data
  - Observed covariance is attenuated



Categorical DFI

- DFI simulations can be altered to generate categorical data

1. Generate MVN data from model-implied covariance
2. Bin MVN data based on number and location of thresholds

- Will simulate categorical data with underlying normal process
- Will simulate same number of categories and same proportion of responses in each category as original data
  - Allows for mix of continuous and categorical responses
  - Continuous is just special case with 0 thresholds

Categorical DFI

- DFI simulations can be altered to generate categorical data

Method implemented in the `catOne` and `catHB` functions in the `dynamic` R package

1. Generate M ed covariance
2. Bin MVN d ocation of thresholds

- Will simulate categorical data with underlying normal process
- Will simulate same number of categories and same proportion of responses in each category as original data
  - Allows for mix of continuous and categorical responses
  - Continuous is just special case with 0 thresholds

One-Factor Simulation

- Same model and conditions as previous one-factor simulation

- Only difference is data characteristics:

- 3 or 5 category responses
- N = 400 or 1000
- Balanced or Skewed response distribution
  - Balanced – symmetric/bell-shaped
  - Skewed – majority of responses are in highest category

One-Factor: HB Cutoffs

12 Items		3 Categories						5 Categories					
		Balanced			Skewed			Balanced			Skewed		
N	L	SRMR	RMSEA	CFI	SRMR	RMSEA	CFI	SRMR	RMSEA	CFI	SRMR	RMSEA	CFI
400	.90	0	92	0	1	50	0	0	98	0	0	96	0
	.75	0	89	19	3	27	14	0	96	32	0	84	27
	.60	1	56	91	9	10	76	0	87	95	0	58	91
1000	.90	0	96	0	0	60	0	0	100	0	0	99	0
	.75	0	94	15	0	24	10	0	100	34	0	94	27
	.60	0	68	98	0	3	88	0	97	100	0	66	98

SRMR has essentially 0% sensitivity to misfit

One-Factor: HB Cutoffs

12 Items		3 Categories						5 Categories					
		Balanced			Skewed			Balanced			Skewed		
N	L	SRMR	RMSEA	CFI	SRMR	RMSEA	CFI	SRMR	RMSEA	CFI	SRMR	RMSEA	CFI
400	.90	0	92	0	1	50	0	0	98	0	0	96	0
	.75	0	89	19	3	27	14	0	96	32	0	84	27
	.60	1	56	91	9	10	76	0	87	95	0	58	91
1000	.90	0	96	0	0	60	0	0	100	0	0	99	0
	.75	0	94	15	0	24	10	0	100	34	0	94	27
	.60	0	68	98	0	3	88	0	97	100	0	66	98

RMSEA sensitivity 3%-100%,
heavily dependent on conditions

One-Factor: HB Cutoffs

12 Items		3 Categories						5 Categories					
		Balanced			Skewed			Balanced			Skewed		
N	L	SRMR	RMSEA	CFI	SRMR	RMSEA	CFI	SRMR	RMSEA	CFI	SRMR	RMSEA	CFI
400	.90	0	92	0	1	50	0	0	98	0	0	96	0
	.75	0	89	19	3	27	14	0	96	32	0	84	27
	.60	1	56	91	9	10	76	0	87	95	0	58	91
1000	.90	0	96	0	0	60	0	0	100	0	0	99	0
	.75	0	94	15	0	24	10	0	100	34	0	94	27
	.60	0	68	98	0	3	88	0	97	100	0	66	98

CFI sensitivity 0% to 100%,
strongly related to strength of factor loadings

One-Factor: DFI cutoffs

12 Items		3 Categories						5 Categories					
		Balanced			Skewed			Balanced			Skewed		
N	L	SRMR	RMSEA	CFI	SRMR	RMSEA	CFI	SRMR	RMSEA	CFI	SRMR	RMSEA	CFI
400	.90	100	100	100	100	100	99	100	100	100	100	100	100
	.75	100	100	100	100	100	100	100	100	100	100	100	100
	.60	100	99	100	---	96	96	100	100	100	99	99	99
1000	.90	99	98	99	98	97	97	99	98	97	97	97	97
	.75	100	100	100	100	100	100	100	100	100	100	100	100
	.60	100	100	100	100	100	100	100	100	100	100	100	100

DFI cutoffs have sensitivity consistently near 100% for all conditions

Multifactor Simulation

- Same model as previous multifactor simulation
  - Misspecification is omitted cross-loadings on 25% of items

- Similar conditions:
  - Loadings = 0.60 or 0.75
  - Items = 12 or 24 (4 or 8 per factor)
  - N = 500 or 1000
  - Categories = 3 or 5
  - Distribution = Balanced or Skewed

Multifactor: HB Cutoffs

		3 Categories						5 Categories					
		Balanced			Skewed			Balanced			Skewed		
N	L	SRMR	RMSEA	CFI	SRMR	RMSEA	CFI	SRMR	RMSEA	CFI	SRMR	RMSEA	CFI
500	.75	0	89	18	1	34	20	0	100	33	0	95	27
	.60	0	6	36	0	1	37	0	29	53	0	9	42
1000	.75	0	98	13	0	34	12	0	100	37	0	98	24
	.60	0	1	40	0	0	29	0	24	58	0	0	42

Sensitivity of traditional cutoffs varies widely between conditions.

Multifactor: DFI Cutoffs

		3 Categories						5 Categories					
		Balanced			Skewed			Balanced			Skewed		
N	L	SRMR	RMSEA	CFI	SRMR	RMSEA	CFI	SRMR	RMSEA	CFI	SRMR	RMSEA	CFI
500	.75	100	98	100	100	97	100	100	99	100	100	99	100
	.60	---	98	100	---	---	---	100	99	100	---	94	97
1000	.75	100	100	100	98	98	100	100	100	100	100	99	100
	.60	100	99	97	94	97	97	100	99	100	100	99	100

Sensitivity of DFI cutoffs consistently near 100%
(when DFI cutoffs are available)

Likert
Responses

Likert Responses

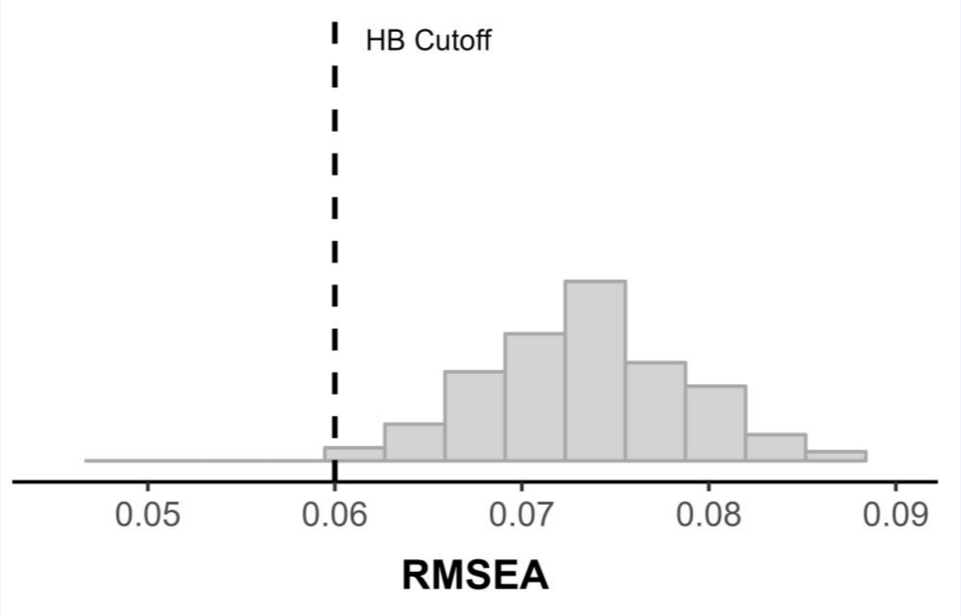
- Most behavioral science scales solicit Likert responses
  - E.g., Flake et al. (2017) report 81% of scales use Likert responses
- Great that DFI supports categorical models, but Likert responses are usually treated as continuous
- Kind of halfway between categorical and continuous
  - Unclear if this affects the sensitivity of HB cutoffs

Redo HB with Different Conditions

- Will follow HB's protocol again except that that responses are generated to be:

1. Multivariate Normal
2. 5-point Likert with symmetric distribution
3. 5-point Likert with skewed distribution

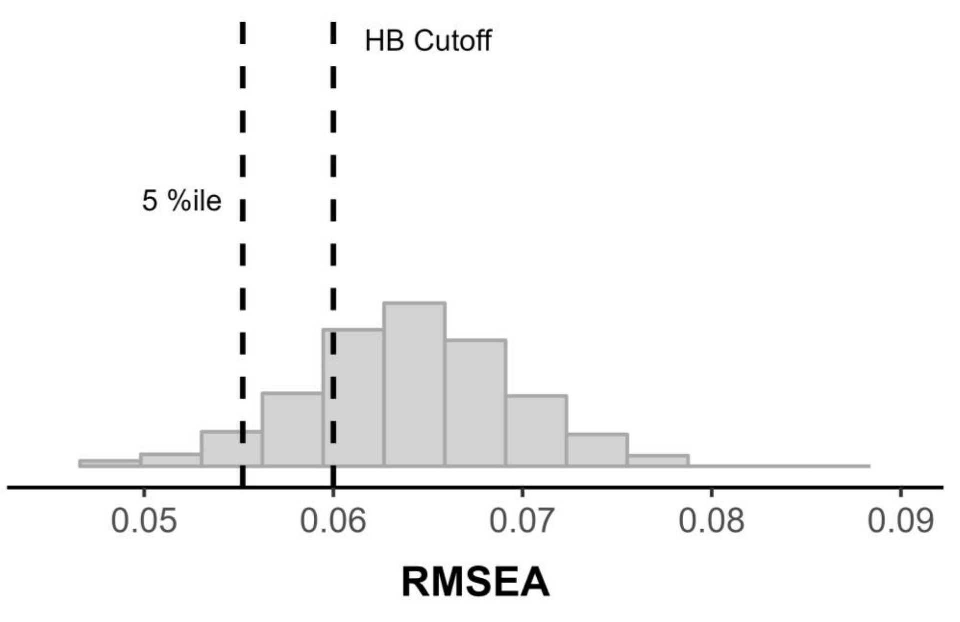
Redo HB with Different Conditions



Multivariate Normal

- With MVN data, 0.06 RMSEA cutoff works well
- 99% of misspecified models are rejected
- Matches HB's conclusions

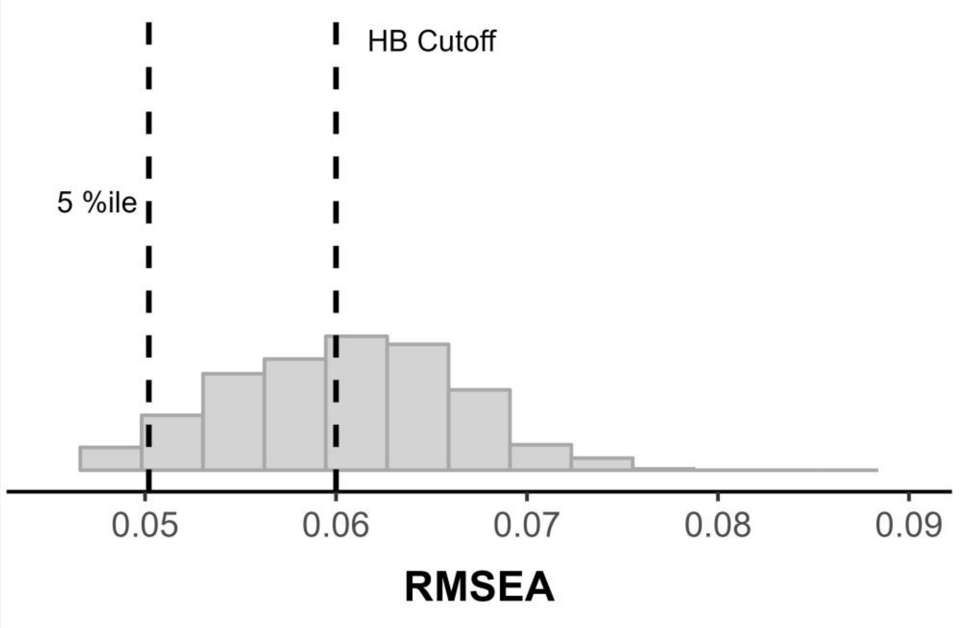
Redo HB with Different Conditions



5-Point Likert Symmetric

- With symmetric 5-point Likert responses, the RMSEA distribution shifts left
  - Likert responses contain less information, so they encode and pass on less misfit information than continuous responses
- Sensitivity to misfit for HB cutoffs is now 77%

Redo HB with Different Conditions



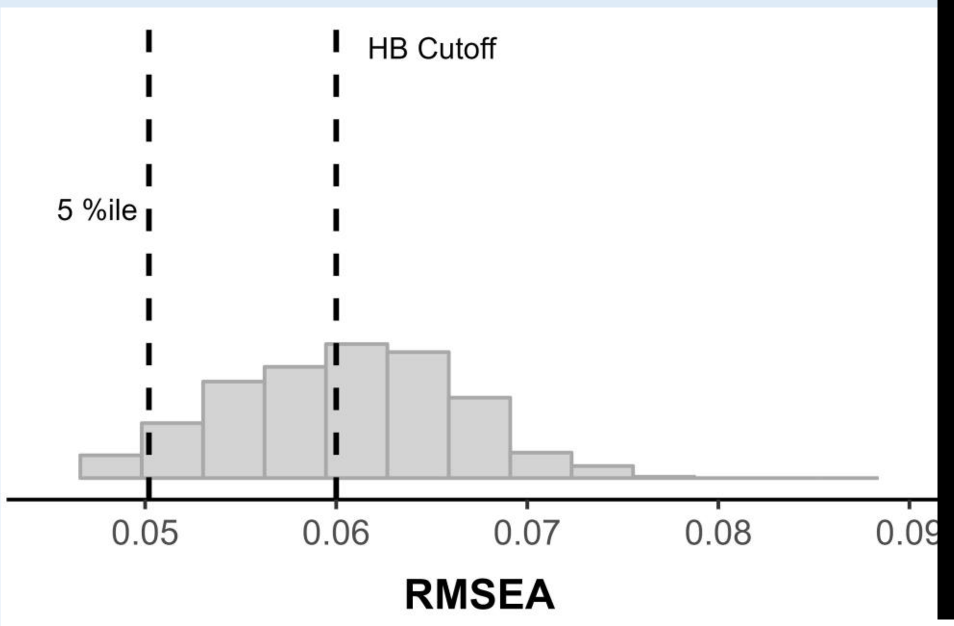
5-Point Likert Skewed

- With skewed 5-point Likert responses, the RMSEA distribution shifts further left
- Sensitivity of HB cutoff to misfit is now 53%
- HB cutoffs too lenient for Likert data
  - "meaningful" misspecification occurs at smaller RMSEA value, even when all conditions identical to HB's simulation

Main Point

Model and data characteristics are relevant

- Tailoring only to model characteristics is not enough
- Big disconnect: most data are Likert, but cutoffs assume continuous



5-Point Likert
Skewed

Likert data

- "meaningful" misspecification occurs at smaller RMSEA value, even when all conditions identical to HB's simulation

Altering DFI Simulations

- DFI for categorical data uses thresholds to discretize data
- But models that treat Likert responses as continuous don’t have thresholds
- Model output not informative for discretizing
  - Empirical data can be used instead

Altering DFI Simulations

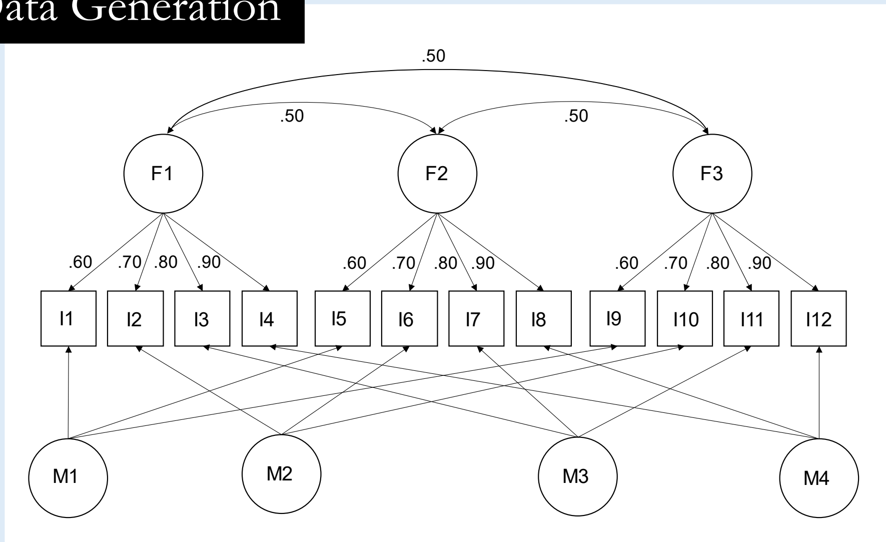
1. Generate MVN data from model-implied covariance
2. Take number of categories and category proportions for each item from the empirical data
3. Convert proportions to thresholds
4. Discretize simulated MVN based on "pseudo" thresholds

- Creates Likert data with similar properties as empirical data
- DFI simulations based on Likert data rather than continuous data
- Get cutoffs for treating Likert as continuous

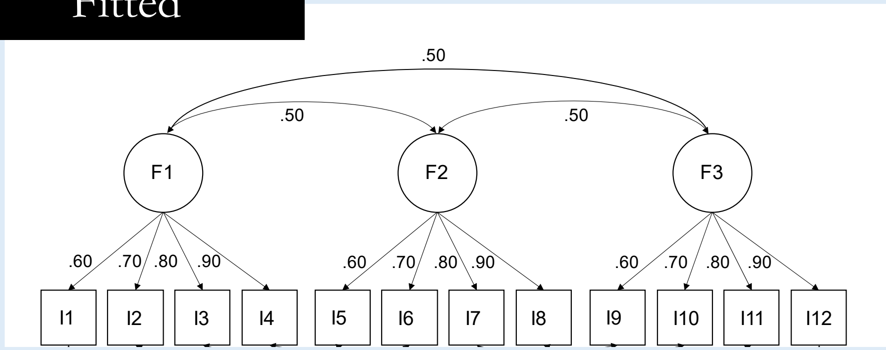
Simulation

- Generate 3-factor model with methods factors
  - N = 400 or 800
  - 5-point symmetric or 5-point skewed
- Intended to mimic multiple reporter models
- Fitted model omits all methods factors and just includes substantive factors

Data Generation



Fitted

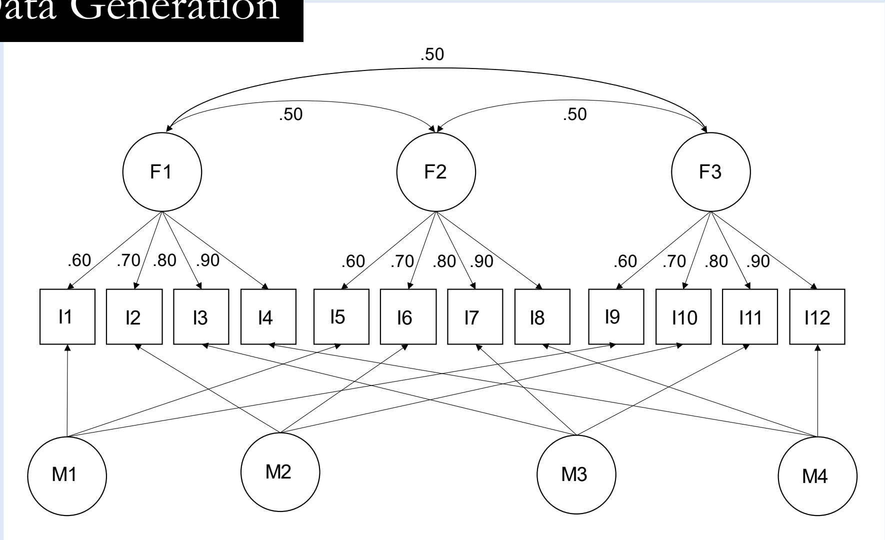


Simulation

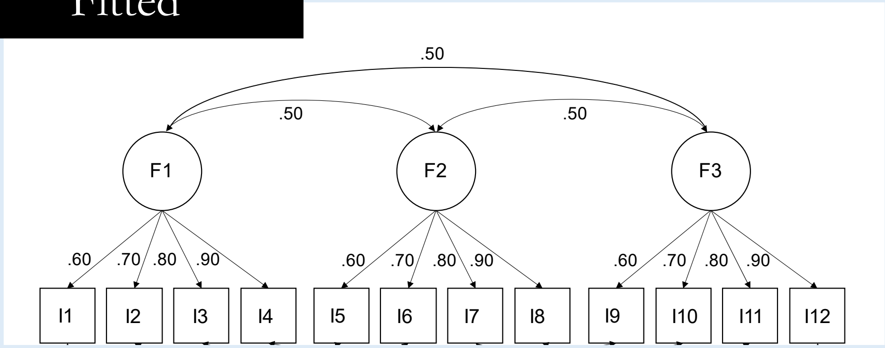
- DFI cutoffs for each fitted model calculated two ways
  1. Using MVN data in DFI simulations
  2. Using Likert data in DFI simulations

- Goal is to compare differences in cutoffs (and their sensitivity) based on how DFI simulates data

Data Generation



Fitted



Differences in Sensitivity

	Method	RMSEA		CFI	
		Balanced	Skewed	Balanced	Skewed
N=400	MVN	67%	75%	87%	89%
	Likert				
N=800	MVN	36%	54%	81%	83%
	Likert				

MVN DFI cutoffs have reduced sensitivity to misfit when the empirical data are 5-point Likert responses

Differences in Sensitivity

	Method	RMSEA		CFI	
		Balanced	Skewed	Balanced	Skewed
N=400	MVN	67%	75%	87%	89%
	Likert	97%	98%	98%	96%
N=800	MVN	36%	54%	81%	83%
	Likert	96%	98%	98%	94%

Likert DFI maintains consistent sensitivity to misspecification

Differences in Sensitivity

	Method	RMSEA		CFI	
		Balanced	Skewed	Balanced	Skewed
N=400	[illegible]	[illegible]	[illegible]	[illegible]	89%
	[illegible]	[illegible]	[illegible]	[illegible]	96%
N=800	MVN	36%	54%	81%	83%
	Likert	96%	98%	98%	94%

`likertOne` and `likertHB` functions in the `dynamic` R package are designed explicitly for models that treat Likert responses as continuous

Lots of Limitations

1. **Who says cutoffs are a good idea anyway?**

- AERA/APA/NCME standards lists 5 approaches to validity
- Fit indices are half of one approach (internal structure)
- DFI focuses on fit indices because that is the most common approach

Lots of Limitations

1. **Who says cutoffs are a good idea anyway?**

Fit indices are overwhelmingly the most common approach

Some argue that the fit indices set up silly rules

DFI is just constrained optimization – smarter way to play a game governed by silly rules

- AERA … lidity
- Fit ind …
- DFI f … mon approach

Lots of Limitations

2. Global fit is not the only type of fit to consider

- Measures like RMSEA and CFI are global indices that try to distill fit across the whole model into a single number

- There are also local fit measures that assess differences in observed and implied covariances for each covariance element

- Global fit is much more commonly reported, but both types are recommended in comprehensive fit analysis

Lots of Limitations

4. **Invariance testing is important step in validation**

- Invariance similarly uses cutoffs derived some a limited set of simulation conditions (Cheung & Rensvold, 2002)

- DFI not yet extended to testing measurement invariance, but it is a clear future direction to encompass all steps of the validation process

Thank you!

dmcneish@asu.edu

@dmcneish18

www.dynamicfit.app

https://github.com/melissagwolf/dynamic