

SEM 4: Modellfit & Modellvergleiche



We are happy to share our materials openly:

The content of these [Open Educational Resources](#) by [Lehrstuhl für Psychologische Methodenlehre und Diagnostik, Ludwig-Maximilians-Universität München](#) is licensed under [CC BY-SA 4.0](#). The CC Attribution-ShareAlike 4.0 International license means that you can reuse or transform the content of our materials for any purpose as long as you cite our original materials and share your derivatives under the same license.

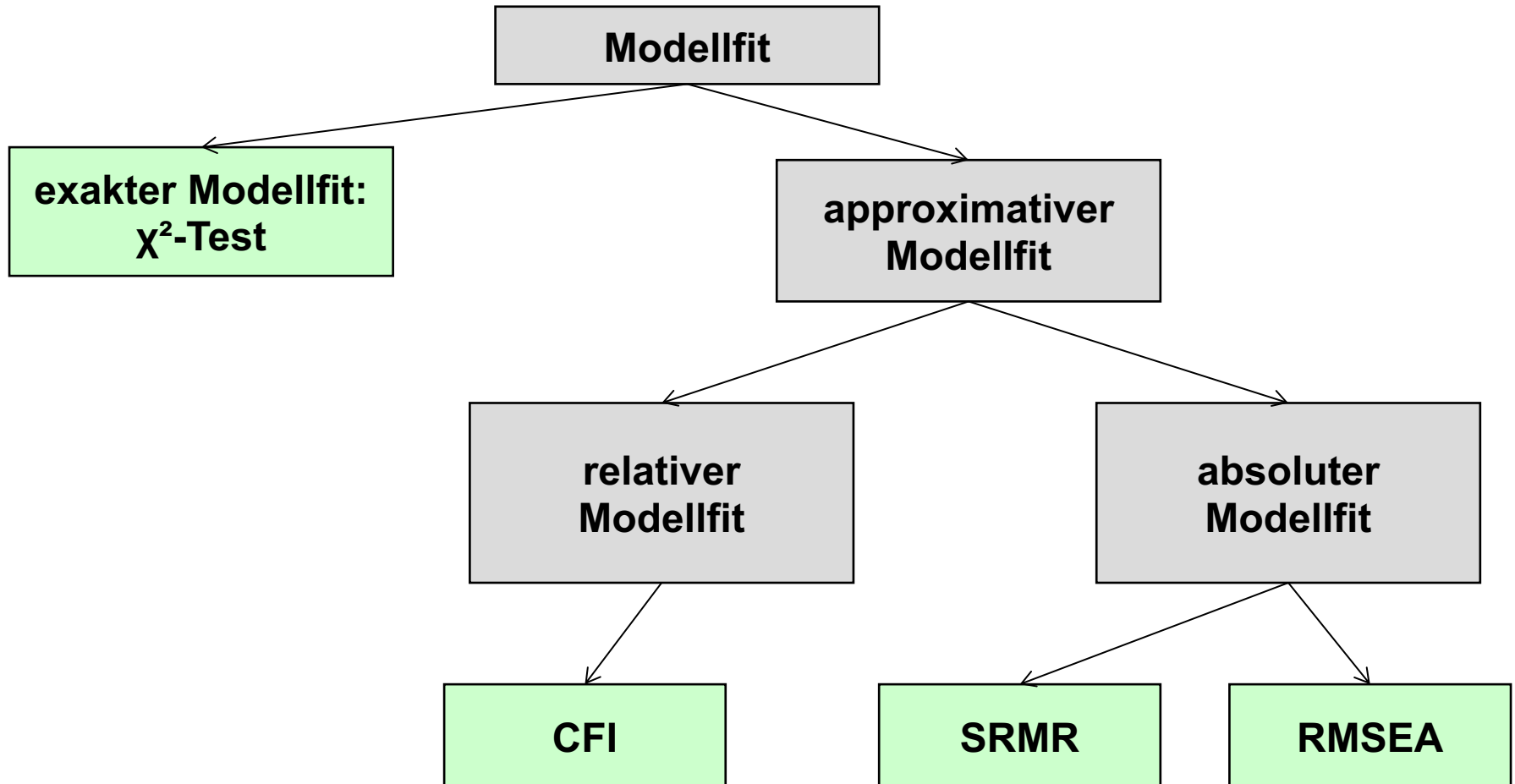
- Die Modellparameter wurden mithilfe von z.B. ML-Schätzung optimiert, so dass sie so gut wie möglich die empirischen Daten abbilden. Zwei Fragen:
- Ist „so gut wie möglich“ auch „gut genug“? → Modellfit-Indizes
- Wenn ich zwei Modelle miteinander vergleiche – welches kann die Daten besser abbilden? → Modellvergleiche

Empirische Kovarianzmatrix			
S	x1	x2	x3
x1	1.36	0.41	0.58
x2	-	1.39	0.45
x3	-	-	1.28

Implizierte Kovarianzmatrix von Modellparameter-Set 2:			
$\hat{\Sigma}(\theta_2)$	x1	x2	x3
x1	1.3	0.48	0.48
x2	-	1.4	0.48
x3	-	-	1.3

Näher kommt man mit einer Parameteroptimierung innerhalb dieses spezifischen Strukturmodells nicht an die empirischen Daten heran ...

Modelltest & Fit-Indizes



Exakter Modell-Fit (exact fit): χ^2 -Test. Leitet sich aus Minimum der Diskrepanzfunktion ab:

$$(N - 1)F_{ML}[S, \Sigma(\hat{\Theta})] \sim X^2(df) \text{ mit } df = p - q$$

p : Anzahl bekannter Parameter

q : Anzahl zu schätzender Parameter

N : Anzahl der Versuchspersonen

➤ Die Hypothesen lauten hierbei:

- $H_0: \Sigma = S$
- $H_1: \Sigma \neq S$

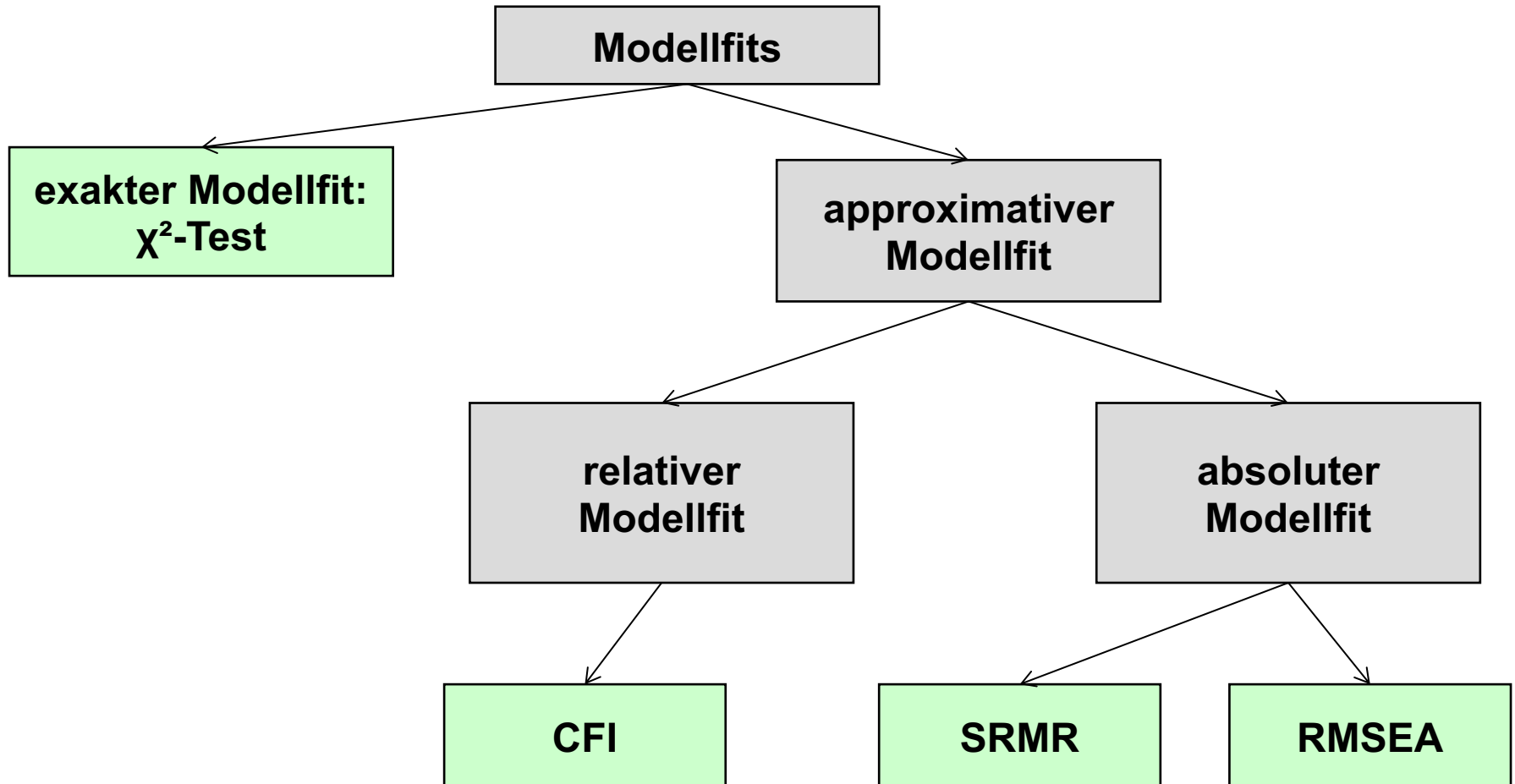
- D.h.: Der finale Wert der ML-Diskrepanzfunktion wird mit $(N-1)$ multipliziert; unter der H_0 ist dieser Wert χ^2 -verteilt mit $df = p - q$
- Ergibt der Signifikanztest für diesen χ^2 -Wert $p < .05$ wird die H_0 abgelehnt → die modell-implizierte Kovarianzmatrix weicht signifikant von der empirischen ab. Bzw.: Wenn der Test *nicht* signifikant wird, spricht das für einen guten Fit.
- Exkurs: die χ^2 -Verteilung (sprich: chi oder ki) ist die Verteilung, wenn man mehrere, unabhängige standardnormalverteilte Zufallsvariablen quadriert und aufsummiert. Sie sieht so ähnlich aus wie und ist verwandt mit der F-Verteilung (z.B. Omnibustest in ANOVA und Regression)

Die Teststärke (Power) des χ^2 -Tests steigt mit der Stichprobengröße an:

- Große Stichproben: Schon geringe Abweichungen (Residuen zwischen implizierter und empirischer Kovarianz-Matrix) von einem perfekten Modell-Fit können zur Ablehnung des Modells führen.
- Kleine Stichproben: Auch große Abweichungen von einem perfekten Modell führen nicht zwangsläufig zur Ablehnung des Modells.
- Test ist also sehr abhängig von der Stichprobengröße; Interpretation muss im Licht der Stichprobengröße geschehen.

Außerdem: „Verdrehte“ Testlogik

- Die „Wunschhypothese“ ist die H_0 („das Modell passt“).
- Wir wollen also, dass der Test nicht signifikant wird.
- Allerdings bedeutet ein nicht-signifikant gewordener Hypothesentest NICHT, dass die H_0 wahr ist. Es bedeutet lediglich, dass wir nicht genug Evidenz hatten, die H_1 abzulehnen.



1. **Absolute Fit-Indizes:** Quantifizieren, wie gut ein Modell die Daten reproduziert; es wird kein Vergleichsmodell benötigt
2. **Relative Fit-Indizes:** Vergleich mit einem Nullmodell (auch Independence-Modell genannt), in dem alle Variablen unkorreliert sind (d.h., es gibt außer den Fehlertermen keine latenten Variablen und keine Zusammenhänge zwischen den Variablen).
 - Das Nullmodell hat üblicherweise einen sehr schlechten Fit
 - Relative Fit-Indizes geben die proportionale Verbesserung der Anpassung gegenüber einem restriktiveren Nullmodell an.

Welche Indizes?

Beauducel und Wittmann (2005) empfehlen folgende Angaben:

- χ^2 -Wert + Überschreitungswahrscheinlichkeit
- RMSEA
- SRMR
- CFI

$$RMSEA = \sqrt{\frac{\chi^2 - df}{N \cdot df}}$$

- Der RMSEA ist ein absoluter Fit-Index
- Der Wertebereich liegt zwischen 0 und 1; der optimale Wert beträgt Null
- Typische Empfehlung (nach Hu & Bentler, 1998):
RMSEA \leq .06 bei $N > 250$; \leq .08 bei $N < 250$
- Der RMSEA ist sensitiv gegenüber fehlspezifizierten Ladungen und Pfaden
- Berücksichtigt sowohl Stichprobengröße als auch Modellkomplexität
- Je komplexer ein Modell, desto weniger Freiheitsgrade $\rightarrow N \cdot df$ im Nenner wird kleiner \rightarrow RMSEA wird größer. D.h.: sparsame Modelle werden vom RMSEA „belohnt“

- Der SRMR ist ein absoluter Fit-Index
- Wertebereich liegt zwischen 0 und 1; der optimale Wert beträgt Null
- Der SRMR gibt durchschnittliche Abweichungen der beobachteten von der implizierten Korrelationsmatrix an (standardisiertes Maß)
- r_{jk} = Differenz zwischen dem beobachteten und dem geschätzten Korrelationskoeffizient in allen Zellen der Matrix; p = Anzahl der manifesten Variablen; S = modellimplizierte Standardabweichung/Kovarianz; σ = aus Stichprobe geschätzte Standardabweichung/Kovarianz
- Typische Empfehlung (nach Hu & Bentler, 1998): SRMR < 0.11
- Im Gegensatz zum RMSEA berücksichtigt der SRMR weder Modellkomplexität noch die Stichprobengröße
- SRMR hoch → Prüfen, ob Kovarianz zwischen latenten Variablen nötig sind oder ein oder mehrere weitere Faktoren/latente Variablen nötig sind.

$$SRMR = \sqrt{\sum_j \sum_{k < j} \frac{r_{jk}^2}{e}}$$
$$r_{jk} = \frac{S_{jk}}{S_j \cdot S_k} - \frac{\hat{\sigma}_{jk}}{\hat{\sigma}_j \cdot \hat{\sigma}_k}$$
$$e = \frac{p \cdot (p + 1)}{2}$$

$$CFI = 1 - \frac{\chi^2_M - df_M}{\chi^2_N - df_N}$$

- Der CFI ist ein relativer Fit-Index
- Der CFI nimmt einen Vergleich des getesteten Modells mit einem restriktiveren Nullmodell vor
 - Nullmodell: Alle manifesten Variablen sind unkorreliert, es gibt keine latenten Variablen
- χ^2_M : Teststatistik des eigentlichen Modells; χ^2_N : Teststatistik des Nullmodells
- Der Wertebereich liegt zwischen 0 und 1; der optimale Wert beträgt 1
- Typische Empfehlung (nach Hu & Bentler, 1998): $CFI > 0.95$
- Vorteil: insensitiv gegenüber der Stichprobengröße

Beauducel und Wittmann (2005), nach Hu und Bentler (1998, 1999), empfehlen folgende Cut-Off Werte bei einer Schätzung nach der ML-Methode:

- $RMSEA \leq .06$ bei $N > 250$; $\leq .08$ bei $N < 250$
- $SRMR \leq .11$
- $CFI > .95$
- χ^2 -Wert mit df und zugehörigem p-Wert

Achtung:

- Fit-Indizes können sich aufgrund der angewandten Methode (z.B. zwischen GLS, ML oder ADF) stark unterscheiden!
- Ein Modell mit ausgezeichnetem statistischen Fit kann inhaltlich völlig sinnfrei sein!
- Die Hu und Bentler-Cutoffs sind mehr als umstritten (vgl. Heene et al., 2011)! Sie sind in der aktuellen SEM-Literatur jedoch Quasistandard. Am zuverlässigsten entdeckt immer noch der χ^2 -Test Fehlspezifikationen im Modell.

- Führt die Bewertung des Gesamtmodells zu einem positiven Ergebnis (das Modell wurde nicht abgelehnt), dann stellt sich die Frage, ob sich jeder einzelne geschätzte Parameter signifikant von null unterscheidet.
- **Achtung, häufiger Irrtum:** Ein global passendes Modell impliziert nicht, dass alle Pfade signifikant sind!
- Jeder geschätzte Modellparameter (Varianz einer exogenen latenten Variable, Pfadkoeffizient, ...) lässt sich testen:
 - H_0 : der Parameter unterscheidet sich nicht von Null.
 - H_1 : der Parameter unterscheidet sich von Null.

- Signifikanztest für einzelne Modellparameter:
 - H_0 : Parameter = 0
 - Zur Überprüfung der H_0 wird eine Teststatistik berechnet: Die *Critical Ratio* C.R. ist der geschätzte Modellparameter geteilt durch seinen geschätzten Standardfehler:

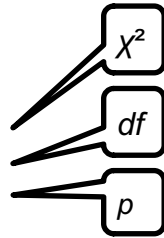
$$C.R. = \frac{\hat{\theta}}{se_{\theta}}$$

- Falls die H_0 gilt (und die manifesten Variablen eine multivariate NV in der Population aufweisen), dann ist C.R. eine standardnormalverteilte Variable → z-Wert
- Im lavaan-Output in R erhalten wir p-Werte für den Test mit H_0 : Parameter = 0

Beispiel

```
> summary(fit, standardized = T, fit.measures = T)
lavaan (0.6-1) converged normally after 35 iterations
```

```
Number of observations      301
Estimator                   ML
Model Fit Test Statistic   85.306
Degrees of freedom         24
P-value (Chi-square)       0.000
```



Model test baseline model:

```
Minimum Function Test Statistic  918.852
Degrees of freedom                36
P-value                           0.000
```

User model versus baseline model:

```
Comparative Fit Index (CFI)      0.931
Tucker-Lewis Index (TLI)         0.896
```



Loglikelihood and Information Criteria:

```
Loglikelihood user model (H0)    -3737.745
Loglikelihood unrestricted model (H1) -3695.092
```

```
Number of free parameters        21
Akaike (AIC)                     7517.490
Bayesian (BIC)                   7595.339
Sample-size adjusted Bayesian (BIC) 7528.739
```

Root Mean Square Error of Approximation:

```
RMSEA                            0.092
90 Percent Confidence Interval    0.071 0.114
P-value RMSEA <= 0.05           0.001
```

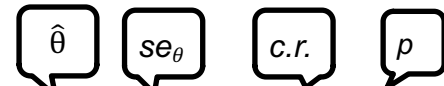


Standardized Root Mean Square Residual:

```
SRMR                             0.065
```



Latent Variables:



	Estimate	Std.Err	z-value	P(> z)	Std.lv	Std.all
visual =~						
x1	1.000				0.900	0.772
x2	0.554	0.100	5.554	0.000	0.498	0.424
x3	0.729	0.109	6.685	0.000	0.656	0.581
textual =~						
x4	1.000				0.990	0.852
x5	1.113	0.065	17.014	0.000	1.102	0.855
x6	0.926	0.055	16.703	0.000	0.917	0.838
speed =~						
x7	1.000				0.619	0.570
x8	1.180	0.165	7.152	0.000	0.731	0.723
x9	1.082	0.151	7.155	0.000	0.670	0.665

Covariances:

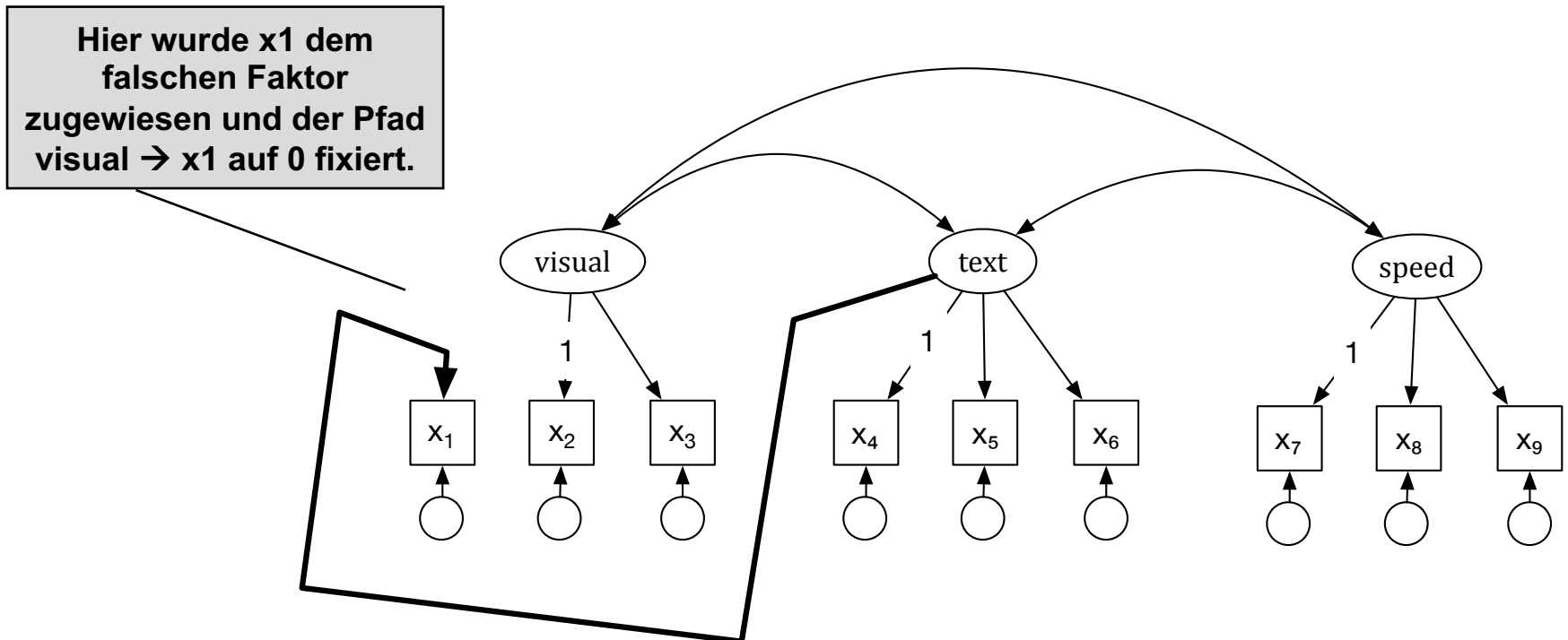
	Estimate	Std.Err	z-value	P(> z)	Std.lv	Std.all
visual ~						
textual	0.408	0.074	5.552	0.000	0.459	0.459
speed	0.262	0.056	4.660	0.000	0.471	0.471
textual ~						
speed	0.173	0.049	3.518	0.000	0.283	0.283

Variances:

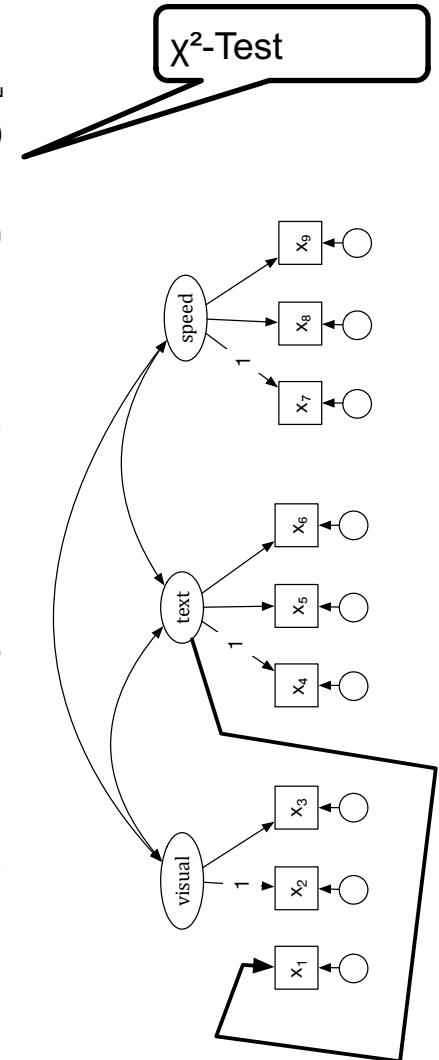
	Estimate	Std.Err	z-value	P(> z)	Std.lv	Std.all
.x1	0.549	0.114	4.833	0.000	0.549	0.404
.x2	1.134	0.102	11.146	0.000	1.134	0.821
.x3	0.844	0.091	9.317	0.000	0.844	0.662
.x4	0.371	0.048	7.779	0.000	0.371	0.275
.x5	0.446	0.058	7.642	0.000	0.446	0.269
.x6	0.356	0.043	8.277	0.000	0.356	0.298
.x7	0.799	0.081	9.823	0.000	0.799	0.676
.x8	0.488	0.074	6.573	0.000	0.488	0.477
.x9	0.566	0.071	8.003	0.000	0.566	0.558
visual	0.809	0.145	5.564	0.000	1.000	1.000
textual	0.979	0.112	8.737	0.000	1.000	1.000
speed	0.384	0.086	4.451	0.000	1.000	1.000

Modifikationsindizes

- Jeder fixierte oder eingeschränkte Parameter (constraints) verringert die Flexibilität in der Struktur der implizierten Kovarianzmatrix → sobald man einen bisher fixierten Parameter freigibt, verbessert sich der Fit
- Modifikationsindizes geben einen Überblick, wie sich der Modellfit verbessert, wenn man einen fixierten Parameter freigibt



Number of observations	301
Estimator	ML
Minimum Function Test Statistic	141.549
Degrees of freedom	24
P-value (Chi-square)	0.000
User model versus baseline model:	
Comparative Fit Index (CFI)	0.867
Root Mean Square Error of Approximation:	
RMSEA	0.128
Standardized Root Mean Square Residual:	
SRMR	0.097



Parameter,
den man
„befreien“
könnte

mi = modification index: Ungefähre
Verbesserung (d.h.: Verkleinerung) der χ^2 -
Statistik, wenn man den Parameter frei lässt →
die faktische Verbesserung kann von diesem
Wert abweichen!

epc = expected parameter
change: Welchen Wert
hätte der Modellparameter
(ungefähr), wenn er frei
geschätzt würde?

```
> modindices(fit3, sort.=TRUE, standardized=FALSE)
```

	lhs	op	rhs	mi	epc
28	visual	=~	x1	59.521	1.202
54	x3	~~	x1	31.842	0.361
76	x7	~~	x8	24.610	0.569
31	visual	=~	x9	24.364	0.730
75	x1	~~	x9	20.874	0.244
42	speed	=~	x1	19.267	0.530
29	visual	=~	x7	12.962	-0.570
52	x3	~~	x5	12.281	-0.173
26	visual	=~	x5	11.692	-0.416

Number of observations		301
Estimator		ML
Minimum Function Test Statistic	141.549	75.439
Degrees of freedom	24	23
P-value (Chi-square)		0.000

$$\Delta\chi^2 = 66$$

User model versus baseline model:

Comparative Fit Index (CFI) ~~0.867~~ **0.941**

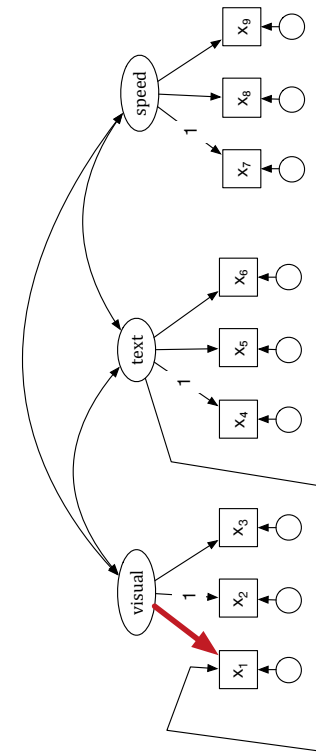
Root Mean Square Error of Approximation:

RMSEA ~~0.128~~ **0.087**

Standardized Root Mean Square Residual:

SRMR ~~0.097~~ **0.061**

visual ~ x1: estimate = 1.239



- Nur MIs > 5 (oder, je nach Autor, > 10) sollten überhaupt in Betracht gezogen werden. Alles darunter ist vermutlich nur Rauschen
- Die Modifikationen sollten theoriegeleitet Sinn machen (→ allerdings: warum hat man es dann nicht gleich so in das Modell eingebaut?). **Jede Modifikation am Strukturmodell ändert die zugrundeliegende Theorie!**
- Jede datengeleitete Modifikation ist als explorativ anzusehen, und sollte in einer frischen Stichprobe kreuzvalidiert werden.
- „It is shown that even under favorable conditions, models arising from **specification searches** must be viewed with caution“ (McCallum, 1986)
 - specification search = solange mit Hilfe der Modifikationsindizes rumspielen, bis das Modell einen akzeptablen Fit hat.

Voraussetzungen:

1. Linearität
2. Zutreffen der angenommenen Verteilungen
3. Keine (Multi-) Kollinearität
4. Ausreichende Stichprobengröße
5. Angemessene Anzahl von Indikatoren pro latenter Variable

Linearität: SEM nur dann sinnvoll, wenn

- die Zusammenhänge linear sind.
- keine Ausreißer vorliegen.
Streudiagramme, Boxplots, Histogramme checken!

Verteilungsannahmen: Siehe Folien zu Schätzmethoden (Vorlesung SEM 3)

Keine (Multi-) Kollinearität

- Von (Multi-) Kollinearität wird gesprochen, wenn zwei oder mehrere Items sehr hoch miteinander korrelieren.
- Sehr hoch korrelierende Indikatoren eines Faktors können insbesondere im Rahmen von ML-Schätzungen zu Schätzproblemen führen: keine Konvergenz, d.h. keine Parameterschätzungen
- Daher sollte als ungefähre Richtlinie darauf geachtet werden, dass keine sehr hoch korrelierenden ($r > .85$) Items als Indikatoren verwendet werden.
- Streng genommen stellt (Multi-) Kollinearität keine Verletzung von Modellannahmen dar, führt aber praktisch häufig zu Konvergenzproblemen.

Stichprobengröße

- Bei SEM können bei kleinen Stichproben Schätzprobleme auftreten.
- Stichprobengrößen:
 - Marsh, Hau, Balla und Grayson (1998) empfehlen Stichprobengrößen von $N > 100$ für ML Schätzer.
 - Kline (2005) and Weston & Gore (2006): $N > 200$
 - Andere Empfehlung: mindestens $n=10$ pro geschätztem Modellparameter
 - Benötigtes N hängt ab von Modellkomplexität, durchschnittlicher Ladungshöhe, Verteilungsverletzungen, ...
 - Am empfehlenswertesten sind Poweranalysen für ein konkretes Modell mit hypothetischen Ladungshöhen (z.B. mit dem `simsem` package; Pornprasertmanit, 2014)

Anzahl von Indikatoren pro latenter Variable

- Anzahl von Items pro latenter Variable wirkt sich auf die Ergebnisse einer CFA aus.
 - Empfehlung: Es sollten mindestens drei oder vier Items pro latenter Variable vorhanden sein.
 - Identifizierbarkeit des Modells wird damit erhöht.
- Mehr Items pro Faktor führen (Marsh, Hau, Balla & Grayson, 1998)
 - häufiger zu angemessenen Lösungen.
 - zu exakteren und stabileren Parameterschätzungen.
 - zu weniger nicht-konvergierenden Lösungen.
 - zu reliableren Faktoren.

Exkurs: Kritische Anmerkung zur Modelltestung

Hu und Bentler (1999) schlagen folgende Cutoff-Werte für ML-Schätzalgorithmus vor (vgl. Beauducel & Wittmann, 2005):

- RMSEA < 0.06 bei $N > 250$; < 0.08 bei $N \leq 250$
- SRMR < 0.11
- CFI > 0.95

Sind diese Cutoff-Werte nun generalisierbar?

Marsh (2004): „...there is some evidence to suggest that even the old cutoff values are overly demanding in relation to a normative criterion of appropriateness based on the best existing psychological instruments. (...) [the] cutoff values proposed by Hu and Bentler (1998, 1999) appear to be largely unobtainable in appropriate practice”.

Hu und Bentler (1999):

Cutoff-Grenzen stammen aus Simulationen mit Modellen, deren Faktorladungen .70 bis .80 betragen haben.

Empirische Realität:

- Peterson (2000): Metaanalytisch ermittelte durchschnittliche Ladung bei Fragebögen: .32
- French (1951; vgl. Carroll, 1993): Durchschnittliche Kommunalität bei Leistungstests: .56
- NEO-FFI (Borkenau & Ostendorff, 1993): Ladungsrange .30 bis .60

Eine interessante Fragestellung lautet nun:

Sind die Ergebnisse von Hu und Bentler (1999) invariant gegenüber verschiedenen Ladungshöhen (und weiteren Faktoren aus der ursprünglichen Simulation von Hu und Bentler)?

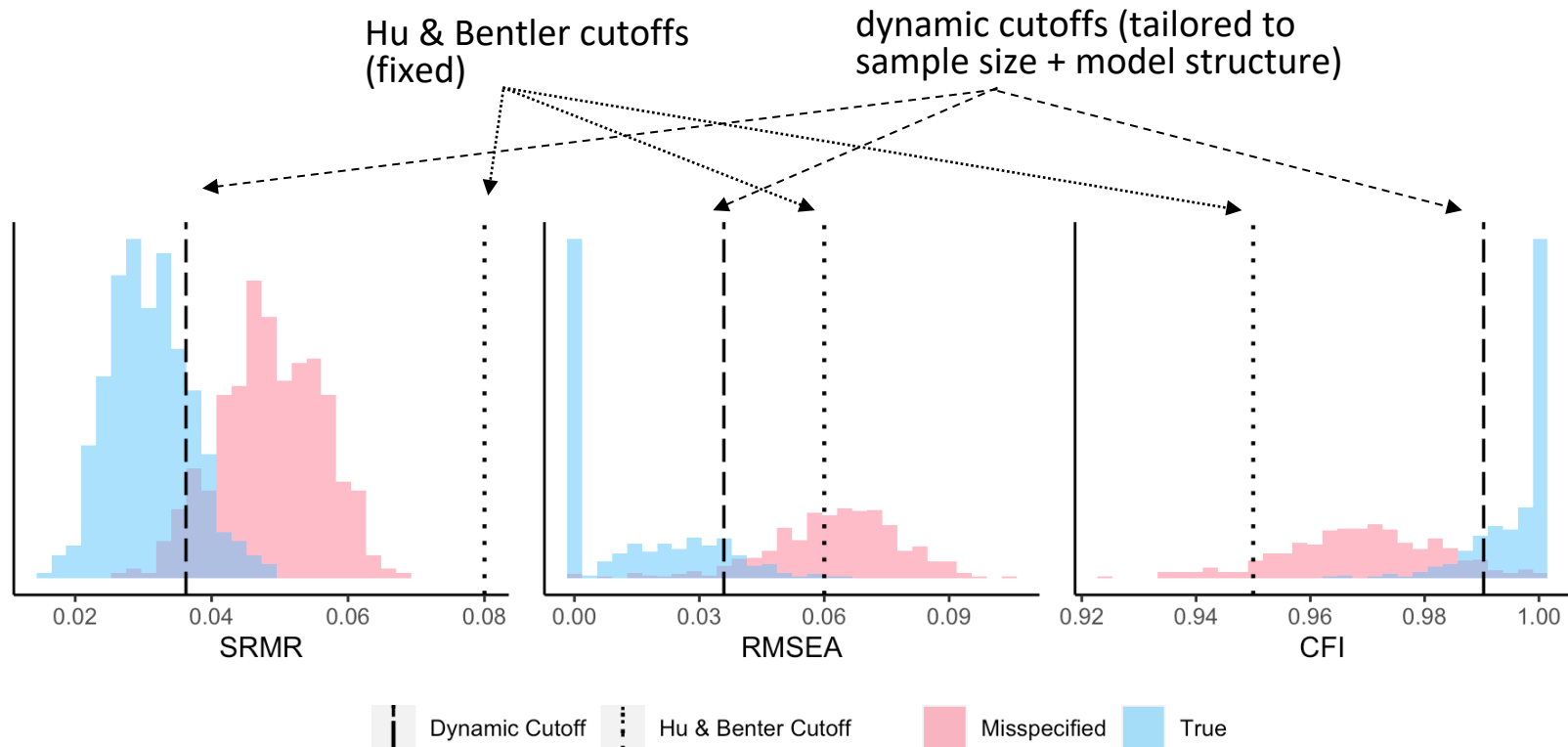
Simulationsstudie (Heene et al., 2011): Masking misfit in confirmatory factor analysis by increasing unique variances: A cautionary note on the usefulness of cutoff values of fit indices. *Psychological Methods*, 16(3), 319.

- Vorgehen: Spezifikation zweier Populationsmodelle (einfach vs. komplex) und zweier fehlspezifizierter Modelle analog zu Hu und Bentler (1999)
- Variation der Faktorladungen λ : Zufallsziehung aus Gleichverteilungen:
 - hoch: $\lambda \in [.70, .90]$
 - mittel: $\lambda \in [.50, .70]$
 - niedrig: $\lambda \in [.30, .50]$
- Stichprobengrößen: 150, 250, 500, 1000, 2500
- 1000 Replikationen je Simulationsbedingung
- Arithmetisches Mittel, Standardabweichung und β -Fehler der Indizes berechnet

Ergebnisse bei realistischen Ladungsmatrizen:

- CFI bei nicht-passenden Modellen auch von Faktorladungshöhe stark beeinflusst.
- RMSEA und SRMR lehnen bei Cutoff-Werten nach Hu und Bentler (1999) zu selten misspezifizierte Modelle ab:
 - β -Fehler bis zu 100%
- χ^2 -Test verliert an Teststärke, je niedriger Faktorladungen (bzw. Reliabilitäten) sind, ist jedoch vergleichsweise zuverlässigste Teststatistik
- Darüber hinaus gibt es neuere Studien, die zeigen, dass mit bestimmten Schätzmethoden (z.B. WLSMV statt Maximum Likelihood) die Fit Indizes anders berechnet werden müssen (Savalei, 2018).
- In lavaan finden sich diese „korrekteren“ Berechnungen als „Robust CFI“ und „Robust RMSEA“, wenn in der *cfa()* Funktion *estimator* = „WLSMV“ gesetzt wird

- „The new generation of fit index cutoffs are dynamic in that they change given the user’s model statement, model type, and sample size.“



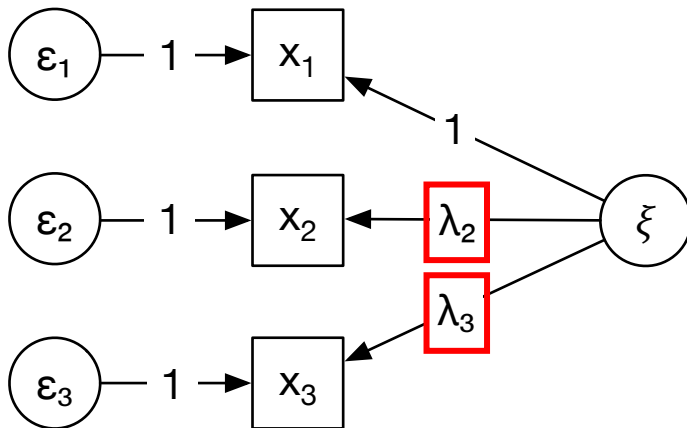
- „This model would not have fit well by Hu & Bentler’s traditional cutoffs [...], but the fit is even poorer when compared to the DFI cutoffs tailored to the user’s model.“

Modellvergleiche

- Ein einfacheres Modell ist unter einem komplexeren Modell geschachtelt, wenn man durch Parameterfixierungen das komplexe zu dem einfachen Modell transformieren kann

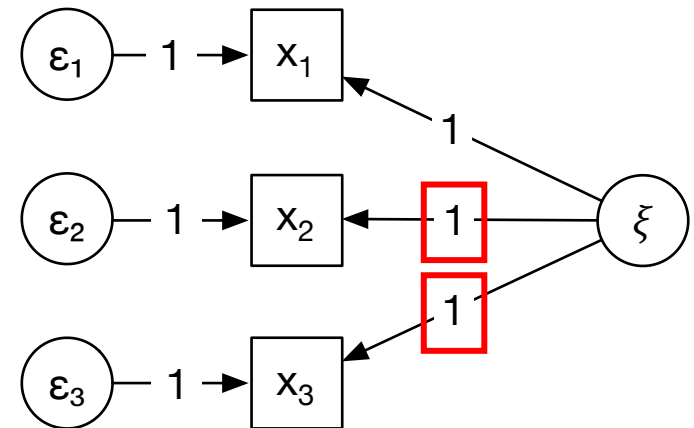
Komplex:

- 6 Modellparameter
- $df = 0$



Einfach(er):

- 4 Modellparameter
- $df = 2$



- Beide zu vergleichenden Modelle müssen auf derselben Kovarianzmatrix basieren. D.h., beide müssen dieselben manifesten Variablen enthalten *und* auf demselben Datensatz beruhen.
 - D.h., auch manifeste Variablen, die nach der Fixierung eines Pfades auf 0 „frei flottierend“ im graphischen Modell sind (ohne Verbindungen zum Rest des Modells), müssen drin bleiben. Inhaltlich bildet das ja eine testbare Hypothese ab: es wird explizit angenommen, dass die frei flottierende Variable keine signifikanten Zusammenhänge zu allen anderen Variablen im Modell hat.
- Wenn zwei Modelle dieselbe Anzahl von Freiheitsgraden haben, können sie nicht geschachtelt sein - das geht nur, wenn ein Modell komplexer ist als das andere (d.h., mehr Modellparameter hat; d.h., weniger Freiheitsgrade hat).
- Um vom komplexen Modell zum (geschachteltem) einfachen Modell zu kommen, dürfen nur Parameter fixiert werden (typischerweise auf 0 oder 1). Es ist nicht erlaubt, Parameter freizugeben.

- Geschachtelte Modelle können per χ^2 -Differenzentest gegeneinander getestet werden
- Teststatistik: Differenz der χ^2 -Werte beider Modelle
$$\Delta\chi^2 = \chi^2_{\text{einfach}} - \chi^2_{\text{komplex}}$$
- Freiheitsgrade des Modellvergleichs:
$$\Delta df = df_{\text{einfach}} - df_{\text{komplex}}$$
- Ein einfacheres Modell hat immer einen schlechteren Modellfit (Grenzfall: gleich gut) als ein komplexeres Modell. Die Frage ist: Ist es nur unwesentlich schlechter?
 - Ein *signifikanter* Modellvergleich zeigt an, dass das einfachere Modell signifikant schlechter ist als das komplexe \rightarrow beim komplexen Modell bleiben.
 - Wenn das einfachere Modell *nicht signifikant* schlechter ist als das komplexe \rightarrow das einfache Modell beibehalten (Occams Razor)

```
HS.model <- '  
  visual    =~ 1*x1 + x2 + x3  
  textual   =~ 1*x4 + x5 + x6  
  speed     =~ 1*x7 + x8 + x9  
'  
  
HS.model.fixed <- '  
  visual    =~ 1*x1 + 1*x2 + 1*x3  
  textual   =~ 1*x4 + 1*x5 + 1*x6  
  speed     =~ 1*x7 + 1*x8 + 1*x9  
'  
  
# Modellvergleich per anova()-Funktion  
anova(fit, fit.fixed)  
  
Chi Square Difference Test
```

	Df	AIC	Chisq	Chisq diff	Df diff	Pr(>Chisq)
fit	24	7517.5	85.305			
fit.fixed	30	7527.6	107.411	22.105	6	0.001159 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Interpretation des
Beispiels: Man
sollte beim
komplexeren
bleiben; das
vereinfachte Modell
ist *zu* einfach (weil
signifikant
schlechter).

- Nicht-geschachtelte Modelle beziehen sich wie geschachtelte auf die gleichen Daten
- Nicht-geschachtelte Modelle können *nicht* per χ^2 -Differenzentest verglichen werden
- Alternative für nicht-geschachtelte Modelle (ohne ins Detail zu gehen): *AIC* (Akaike Information Criterion) und *BIC* (Bayesian Information Criterion)
- AIC/BIC errechnen sich aus dem Modell-Fit und der Anzahl an zu schätzenden Parametern im Modell → sie sind eine Gewichtung aus Fit und Komplexität
- Je kleiner AIC/BIC, desto besser passt das Modell zu den Daten.
Achtung: Im Gegensatz zum χ^2 -Wert kann der AIC/BIC des komplexeren Modells sowohl größer als auch kleiner als der AIC/BIC des einfacheren Modells sein!
- Absolute Höhe bei AIC/BIC irrelevant; entscheidend ist nur die Differenz von AIC beider Modelle, also Δ AIC/BIC
- AIC/BIC kann auch bei geschachtelten Modellen angewendet werden; bietet jedoch keinen Signifikanztest

```
HS.model <- '  
  visual   =~ x1 + x2 + x3  
  textual  =~ x4 + x5 + x6  
  speed    =~ x7 + x8 + x9  
,  
  
HS.model.fixed <- '  
  visual   =~ 1*x1 + 1*x2 + 1*x3  
  textual  =~ 1*x4 + 1*x5 + 1*x6  
  speed    =~ 1*x7 + 1*x8 + 1*x9  
,
```

```
# Modellvergleich per anova()-Funktion  
anova(fit, fit.fixed)
```

Chi Square Difference Test

	Df	AIC	Chisq	Chisq diff	Df diff	Pr(>Chisq)
fit	24	7517.5	85.305			
fit.fixed	30	7527.6	107.411	22.105	6	0.001159 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Δ AIC = 10.1: Das komplexere
Modell „fit“ ist zu bevorzugen

Nicht-geschachteltes Modell - AIC

```
HS.model <- '  
  visual   =~ 1*x1 + x2  
  textual  =~ 1*x4 + x6  
  speed    =~ 1*x7 + x8 + x9'
```

```
HS.model2 <- '  
  visual   =~ 1*x2 + 1*x3  
  textual  =~ 1*x4 + x5  
  speed    =~ 1*x7 + x8 + x9'
```

```
anova(fit, fit2)  
AIC(fit)  
AIC(fit2)
```

Bei nicht-geschachtelten
Modellen (siehe lavaan-
Warning): χ^2 Test *nicht*
interpretieren!

```
> anova(fit, fit2)  
Chi Square Difference Test  


|      | Df | AIC    | BIC    | Chisq  | Chisq diff | Df diff | Pr(>Chisq) |
|------|----|--------|--------|--------|------------|---------|------------|
| fit  | 11 | 5937.6 | 6000.6 | 50.728 |            |         |            |
| fit2 | 12 | 6029.3 | 6088.6 | 52.597 | 1.8685     | 1       | 0.1716     |


```

Warnmeldung:

In lavTestLRT(object = new("lavaan", version = "0.6.3", call = lavaan::lavaan(model = HS.model,
lavaan WARNING: some models are based on a different set of observed variables

```
>  
> AIC(fit)  
[1] 5937.589  
> AIC(fit2)  
[1] 6029.264
```

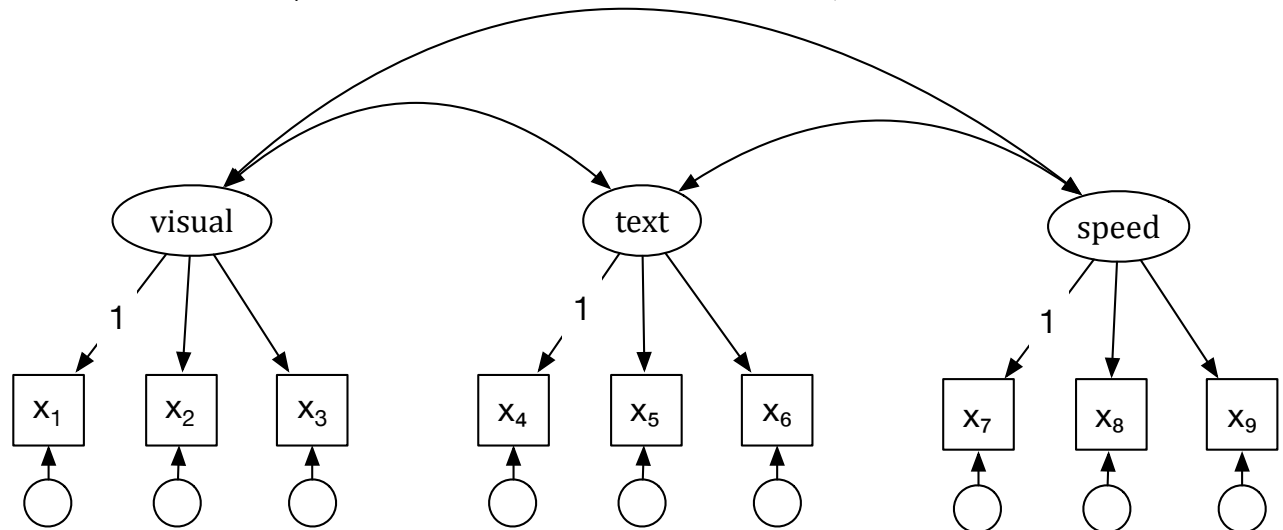

- Sowohl bei Modellvergleichen mit dem χ^2 -Differenzentest als auch dem AIC/BIC hat man keine Garantie, dass das „bessere“ Modell näher dran ist am tatsächlichen kausalen Modell in der Population.
- Dies gilt vor allem, wenn das bessere Modell beim exakten Modelltest und den Fit Indices immer noch schlecht abschneidet.
- Sowohl χ^2 -Differenzentest als auch das AIC/BIC sollten eher als prädiktive Kriterien angesehen werden. Sie können helfen, das Modell zu finden, das die manifesten Variablen am besten vorhersagt.

Ein Beispiel, komplett durcheinexerziert

Beispiel CFA: HolzingerSwineford1939

- Bekannter Datensatz, basierend auf Holzinger & Swineford (1939)
- 9 beobachtete Indikatoren messen 3 latente Faktoren:
 - Faktor „visual“ durch x1 (Visual perception), x2 (Cubes) und x3 (Lozenges)
 - Faktor „text“ durch x4 (Paragraph comprehension), x5 (Sentence completion) und x6 (Word meaning)
 - Faktor „speed“ durch x7 (Speeded addition), x8 (Speeded counting of dots) und x9 (Speeded discrimination straight and curved capitals)
- N=301
- Annahme: 3 korrelierte Faktoren

```
HS.model <- '  
  visual =~ x1 + x2 + x3  
  textual =~ x4 + x5 + x6  
  speed  =~ x7 + x8 + x9  
,  
  
fit <- cfa(model = HS.model,  
  data = HolzingerSwineford1939)  
  
summary(fit, standardized = TRUE, fit.measures = TRUE)
```



lavaan (0.6-1) converged normally after 35 iterations

Number of observations	301
Estimator	ML
Model Fit Test Statistic	85.306
Degrees of freedom	24
P-value (Chi-square)	0.000

User model versus baseline model:

Comparative Fit Index (CFI)	0.931
-----------------------------	-------

Root Mean Square Error of Approximation:

RMSEA	0.092
90 Percent Confidence Interval	0.071 0.114
P-value RMSEA ≤ 0.05	0.001

Standardized Root Mean Square Residual:

SRMR	0.065
------	-------

- Aufgrund der Fit-Indizes würden wir das Modell im Ganzen ablehnen (nach Hu & Bentler gesprochen).
- Generell gilt: wird das Modell im Ganzen abgelehnt, ist es auch sinnfrei die Schätzungen der Parameter zu interpretieren.
- Zur Veranschaulichung wollen wir im Folgenden die Schätzungen der Parameter trotzdem interpretieren.

Latent Variables:

	Estimate	Std.Err	z-value	P(> z)	Std.lv	Std.all
visual =~						
x1	1.000				0.900	0.772
x2	0.554	0.100	5.554	0.000	0.498	0.424
x3	0.729	0.109	6.685	0.000	0.656	0.581
textual =~						
x4	1.000				0.990	0.852
x5	1.113	0.065	17.014	0.000	1.102	0.855
x6	0.926	0.055	16.703	0.000	0.917	0.838
speed =~						
x7	1.000				0.619	0.570
x8	1.180	0.165	7.152	0.000	0.731	0.723
x9	1.082	0.151	7.155	0.000	0.670	0.665

- Pfadkoeffizienten zwischen manifesten Indikatoren und latenten Variablen (mit ULI).
- Bei einem Unterschied von 1 Einheit in der latenten Variablen wird ein Unterschied von 0.554 Einheiten in der manifesten Variablen x2 erwartet.
- Da sie unstandardisiert sind, können wir sie in ihrer relativen Höhe zueinander nicht interpretieren.
- Keine Aussage darüber, ob es ein starker oder schwacher Zusammenhang ist.

Latent Variables:

	Estimate	Std.Err	z-value	P(> z)	Std.lv	Std.all
visual =~						
x1	1.000				0.900	0.772
x2	0.554	0.100	5.554	0.000	0.498	0.424
x3	0.729	0.109	6.685	0.000	0.656	0.581
textual =~						
x4	1.000				0.990	0.852
x5	1.113	0.065	17.014	0.000	1.102	0.855
x6	0.926	0.055	16.703	0.000	0.917	0.838
speed =~						
x7	1.000				0.619	0.570
x8	1.180	0.165	7.152	0.000	0.731	0.723
x9	1.082	0.151	7.155	0.000	0.670	0.665

- **Std.all** steht für den Fall, dass die Varianzen **aller** manifesten und aller latenten Variablen (mit Ausnahme der Fehler) standardisiert und damit auf 1 festgelegt wurden.
- Jetzt können wir die jeweiligen **standardisierten** Ladungen in ihrer relativen Ausprägung zueinander interpretieren.
- Auch eine Aussage darüber, wie stark der Zusammenhang zwischen latenter und manifester Variable ist, ist nun möglich.

Latent Variables:

	Estimate	Std.Err	z-value	P(> z)	Std.lv	Std.all
visual =~						
x1	1.000				0.900	0.772
x2	0.554	0.100	5.554	0.000	0.498	0.424
x3	0.729	0.109	6.685	0.000	0.656	0.581
textual =~						
x4	1.000				0.990	0.852
x5	1.113	0.065	17.014	0.000	1.102	0.855
x6	0.926	0.055	16.703	0.000	0.917	0.838
speed =~						
x7	1.000				0.619	0.570
x8	1.180	0.165	7.152	0.000	0.731	0.723
x9	1.082	0.151	7.155	0.000	0.670	0.665

- **Std.lv** steht für den Fall, dass die Metrik der latenten Variable mit UVI festgelegt wurde (im Vergleich zu ULI in der Spalte „Estimate“).
- Deshalb erhält man hier auch eine geschätzte Ladung für jedes Item: es muss keine Ladung auf 1 fixiert werden, um der latenten Variable eine Metrik zu geben.

Covariances:

	Estimate	Std.Err	z-value	P(> z)	Std.lv	Std.all
visual ~~						
textual	0.408	0.074	5.552	0.000	0.459	0.459
speed	0.262	0.056	4.660	0.000	0.471	0.471
textual ~~						
speed	0.173	0.049	3.518	0.000	0.283	0.283

Wie oben können wir die Ausprägungen der Kovarianzen nicht wirklich gut interpretieren (vgl. VL Statistik 1)

- In der standardisierten Lösung std.all entsprechen die Kovarianzen den **Korrelationen**.
- Diese erlauben uns nun Aussagen über die Stärke des Zusammenhangs.

Variances:

	Estimate	Std.Err	z-value	P(> z)	Std.lv	Std.all
.x1	0.549	0.114	4.833	0.000	0.549	0.404
.x2	1.134	0.102	11.146	0.000	1.134	0.821
.x3	0.844	0.091	9.317	0.000	0.844	0.662
.x4	0.371	0.048	7.779	0.000	0.371	0.275
.x5	0.446	0.058	7.642	0.000	0.446	0.269
.x6	0.356	0.043	8.277	0.000	0.356	0.298
.x7	0.799	0.081	9.823	0.000	0.799	0.676
.x8	0.488	0.074	6.573	0.000	0.488	0.477
.x9	0.566	0.071	8.003	0.000	0.566	0.558
visual	0.809	0.145	5.564	0.000	1.000	1.000
textual	0.979	0.112	8.737	0.000	1.000	1.000
speed	0.384	0.086	4.451	0.000	1.000	1.000

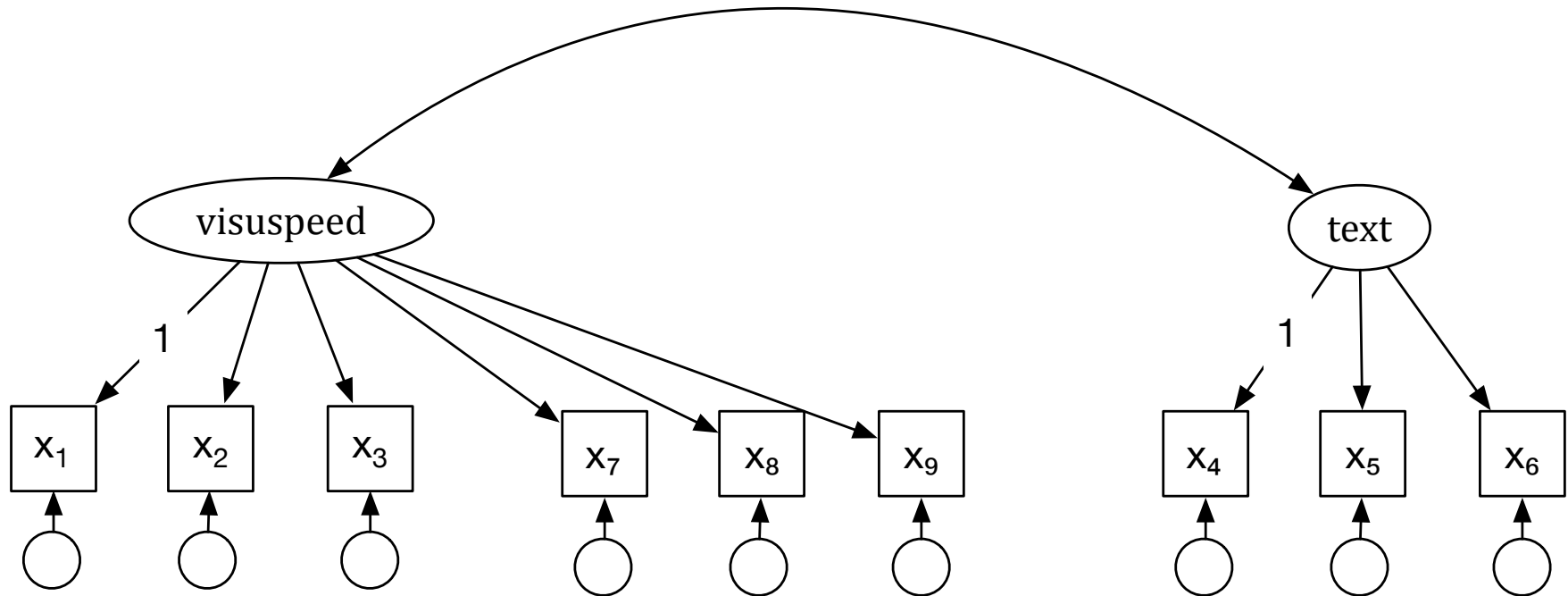
- Erinnerung: von allen Fehlern und allen exogenen latenten Variablen müssen Varianzen geschätzt werden (sofern sie nicht fixiert sind).
- Für alle **endogenen** Variablen in der Tabelle entspricht der Estimate der **Fehlervarianz** dieser Variablen.
- Für alle **exogenen** Variablen in der Tabelle entspricht der Estimate der **Varianz**.
- In der absoluten Ausprägung hat die Varianz wie oben keine wirkliche Interpretation. Man müsste bspw. schon die Gesamtvarianz einer endogenen Variablen wissen, um abzuschätzen, ob die Fehlervarianz vergleichsweise hoch ist.

Variances:

	Estimate	Std.Err	z-value	P(> z)	Std.lv	Std.all
.x1	0.549	0.114	4.833	0.000	0.549	0.404
.x2	1.134	0.102	11.146	0.000	1.134	0.821
.x3	0.844	0.091	9.317	0.000	0.844	0.662
.x4	0.371	0.048	7.779	0.000	0.371	0.275
.x5	0.446	0.058	7.642	0.000	0.446	0.269
.x6	0.356	0.043	8.277	0.000	0.356	0.298
.x7	0.799	0.081	9.823	0.000	0.799	0.676
.x8	0.488	0.074	6.573	0.000	0.488	0.477
.x9	0.566	0.071	8.003	0.000	0.566	0.558
visual	0.809	0.145	5.564	0.000	1.000	1.000
textual	0.979	0.112	8.737	0.000	1.000	1.000
speed	0.384	0.086	4.451	0.000	1.000	1.000

- Die std.all Lösung erhält man, wenn die Varianzen aller Variablen (mit Ausnahme der Fehler) auf 1 gesetzt werden.
- Wichtig: das entspricht nicht einer Fixierung, sondern einer Standardisierung.
- Wenn wir jetzt die Fehlervarianz für den Fall kennen, dass die Gesamtvarianz 1 ist, können wir deren Höhe nun beurteilen.
- Bsp. x1:
 - Fehlervarianz = 0.404, wenn die Gesamtvarianz 1 ist.
 - “Systematische Varianz“, also alles, was z.B. durch latente Variablen (hier *visual*) erklärt wird = Gesamtvarianz – Fehlervarianz = $1 - 0.404 = 0.596$
 - Ca. 60% der Gesamtvarianz von x1 wird durch die latente Variable *visual* erklärt.
 - Dies entspricht wiederum der quadrierten Ladung (s. Slide 47: $0.772^2 = 0.596$)

```
HS.model2 <- '  
  visuspeed =~ x1 + x2 + x3 + x7 + x8 + x9  
  textual =~ x4 + x5 + x6  
,
```



lavaan (0.6-1) converged normally after 30 iterations

Number of observations	301
Estimator	ML
Model Fit Test Statistic	151.470
Degrees of freedom	26
P-value (Chi-square)	0.000

Aufgrund der Fit-Indizes
würden wir auch dieses
Modell ablehnen.

User model versus baseline model:

Comparative Fit Index (CFI)	0.858
-----------------------------	-------

Root Mean Square Error of Approximation:

RMSEA	0.127
90 Percent Confidence Interval	0.108 0.147
P-value RMSEA ≤ 0.05	0.000

Standardized Root Mean Square Residual:

SRMR	0.088
------	-------

```
> lavTestLRT(fit, fit2, model.names=c(fit = "3-Faktor", fit2 = "2-Faktor"))
```

Chi Square Difference Test

	Df	AIC	BIC	Chisq	Chisq diff	Df diff	Pr(>Chisq)
3-Faktor	24	7517.5	7595.3	85.305			
2-Faktor	26	7579.7	7650.1	151.470	66.165	2	4.29e-15 ***

- Der Vergleich des AIC und BIC legt nahe, dass das 3-Faktor Modell besser als das 2-Faktor Modell zu den Daten passt.
- Aber wir haben ja schon beim Modellfit gesehen, dass das Modell eh unsere Daten nicht ausreichend gut beschreibt.

Die χ^2 -Differenz wird nicht interpretiert, da die Modelle nicht geschachtelt sind

- Fit-Indizes quantifizieren, wie nahe die (gefittete) modell-implizierte Kovarianzmatrix an der empirischen Kovarianzmatrix ist (d.h.: Wie gut kann das Modell die Daten abbilden?)
- Unterscheide den exakten Modellfit (basierend auf dem χ^2 -Test), relative und absolute Modellfits
- Die am häufigsten herangezogenen Cutoff-Werte (d.h., ab wann ist ein Fitindex gut / schlecht?) sind die von Hu & Bentler (1999).
 - RMSEA < 0.06 bei $N > 250$; < 0.08 bei $N \leq 250$; SRMR < 0.11; CFI > 0.95
- Aber: Kritik an diesen starren Cutoffs, passen oft nicht zu den tatsächlichen Modellen
- Modellvergleiche geben an, welches von zwei Modellen *relativ gesehen* besser zu den Daten passt
 - Geschachtelte Modelle: χ^2 -Test oder AIC
 - Nicht-geschachtelte Modelle: AIC