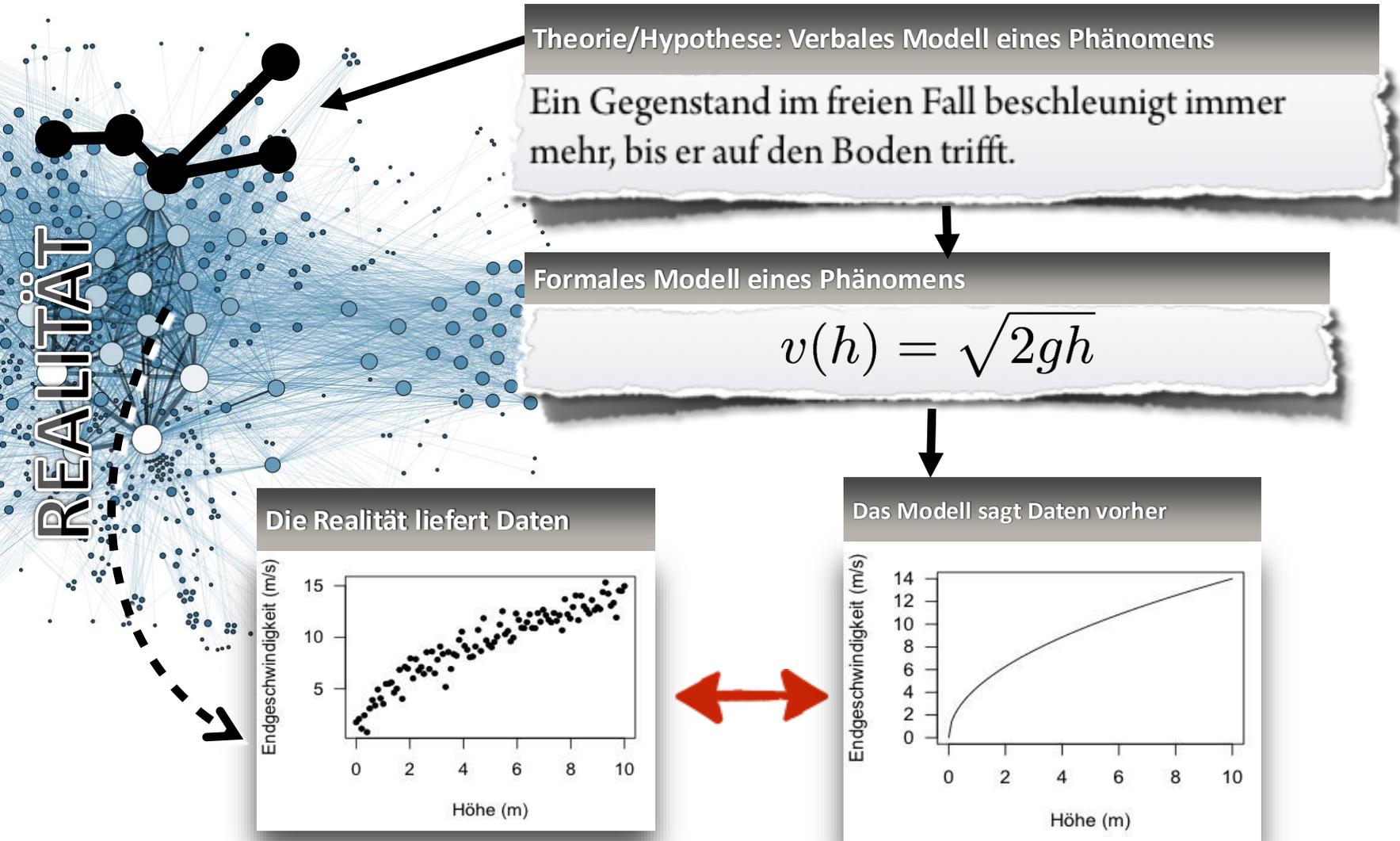


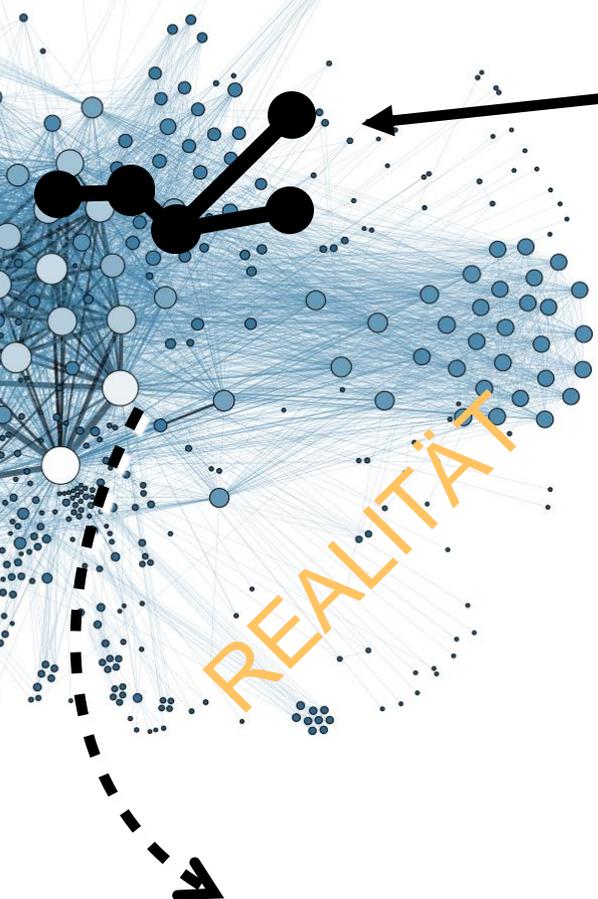
SEM 1: Einführung in Lineare Strukturgleichungsmodelle



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.



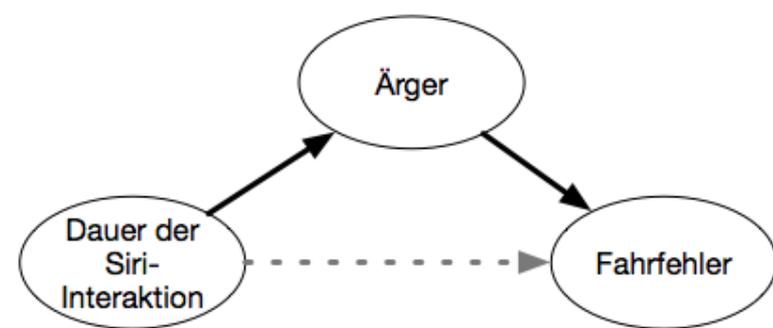


REALITÄT

Theorie/ Hypothese:
Verbales Modell eines Phänomens

Längere Interaktionen mit Siri führen nur deshalb zu mehr Unfällen, weil sie Ärger beim Fahrer auslösen.

Formales Kausalmodell eines Phänomens



Die Realität liefert Daten

	mimic	voice	hrtrt	att1	att2	att3	att4	errrs	siri
mimic	1.12								
voice	0.22	1.15							
heartrate	-0.05	0.05	1.19						
att1	0.01	0.01	0.15	0.94					
att2	0.04	0.03	0.02	0.13	1.09				
att3	0.05	-0.13	-0.03	-0.02	0.05	1.07			
att4	-0.06	0.11	-0.12	0.02	0.07	0.29	1.46		
errors	0.55	0.48	0.07	0.18	0.33	0.42	0.71	2.93	
siri	0.30	0.19	0.05	0.13	0.21	0.19	0.13	0.97	1.29

S empirische Kovarianzmatrix



Das Modell sagt Daten vorher

	mimic	voice	hertrt	att1	att2	att3	att4	errors	siri
mimic	1.49								
voice	0.24	1.12							
heartrate	0.16	0.08	1.05						
att1	0.01	0.01	0.00	1.04					
att2	0.02	0.01	0.01	0.08	1.16				
att3	0.02	0.01	0.01	0.06	0.12	1.09			
att4	0.03	0.02	0.01	0.12	0.25	0.18	1.37		
errors	0.79	0.40	0.26	0.16	0.32	0.24	0.48	2.78	
siri	0.36	0.18	0.12	0.03	0.06	0.04	0.09	0.70	1.00

modell-implizierte Kovarianzmatrix

$\hat{\Sigma}$



Ein **Lineares Strukturgleichungsmodell (SEM)** kann betrachtet werden als die folgende Kombination:

- Kausales Modell: Kausalbeziehungen zwischen Variablen sind formalisiert als Graph (DAG, s. letzte VL)
- Annahme linearer Zusammenhänge (meist ohne Interaktionen) zwischen allen Variablen sowie (ursprünglich) Normalverteilung aller Variablen: Pfadkoeffizienten analog zu Regressionskoeffizienten aus einer (multiplen) linearen Regression
- Eventuell zusätzliche Berücksichtigung nicht beobachteter Variablen: Messmodelle für latente Variablen

Wichtig:

- In der Psychologie sind nicht direkt beobachtbare Variablen eher die Regel als die Ausnahme (z.B. Intelligenz, Persönlichkeit, Werte, ...)

Lineare Strukturgleichungsmodelle sind ein sehr beliebtes Multifunktionsstool, das in verschiedenen Kontexten eingesetzt werden kann.

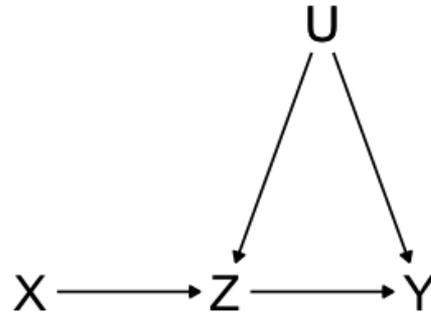
Strukturgleichungsmodell (SEM) als Oberbegriff / Kombination von ...

- *Pfadanalyse*: Mehrere direkte kausale Effekte von beobachteten („manifesten“) Variablen gemeinsam schätzen (DAG + Zusatzannahmen)
- *Strukturelle Regressionsanalyse*: Pfadanalyse, bei der (auch) nicht beobachtete („latente“) Variablen enthalten sind
- *Konfirmatorische Faktorenanalyse*: Fokus auf der Modellierung von latenten Variablen und ungerichtete Zusammenhänge zwischen diesen (vgl. VL Testtheorie/Diagnostik im Bachelor)

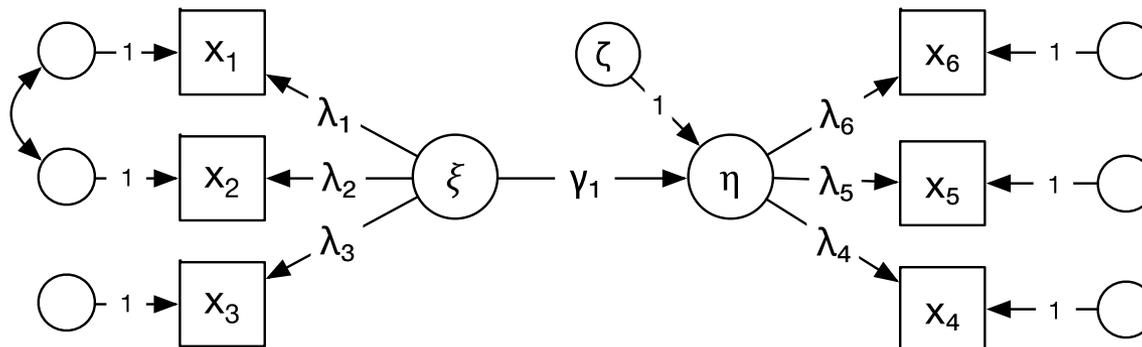
SEMs sind für die Psychologie besonders nützlich, weil sie ...

- komplexe sprachliche Theorien leicht formalisieren und prüfen
- und dabei Messfehler berücksichtigen können.

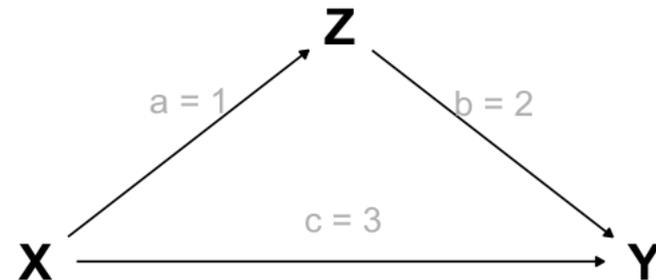
- Die grafische Darstellung von DAGs besteht klassischerweise nur aus Variablennamen und gerichteten Pfeilen.



- Bei Strukturgleichungsmodellen haben sich erweiterte Konventionen für die grafische Darstellung etabliert.
- Wir werden diese zusätzliche Notation in dieser Vorlesung Schritt für Schritt einführen und am Ende nochmal (vollständig) zusammenfassen.



```
> n <- 1000
> a <- 1
> b <- 2
> c <- 3
> set.seed(1)
> x <- rnorm(n)
> z <- rnorm(n, mean = a * x)
> y <- rnorm(n, mean = b * z + c * x)
> dat <- data.frame(x = x, y = y, z = z)
```

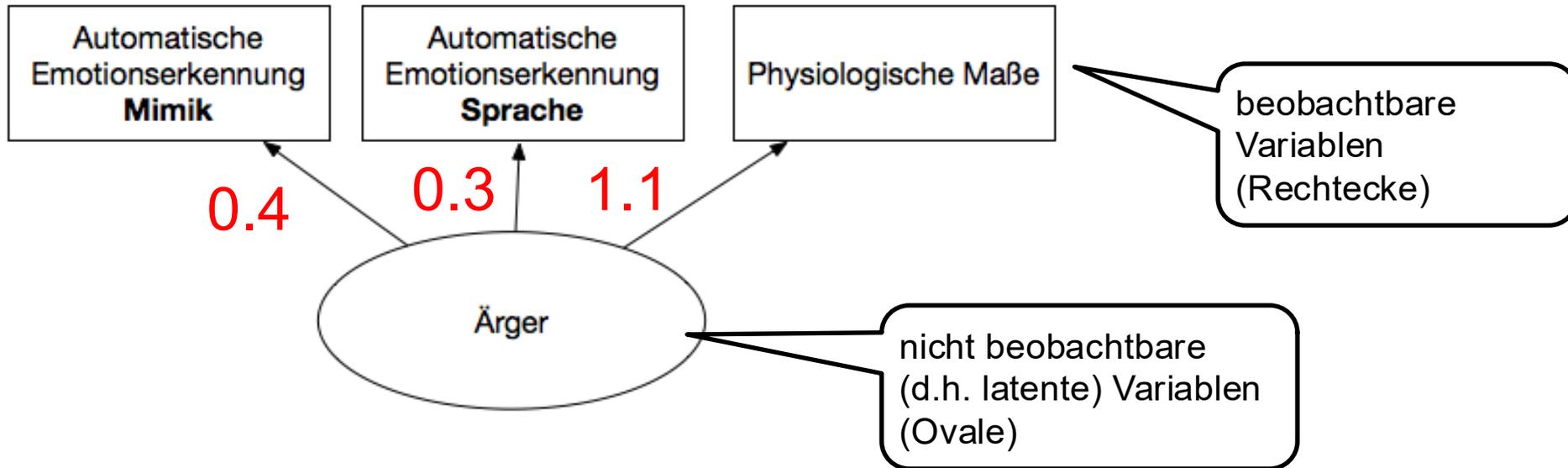


Regressionsanalyse

```
> coef(lm(z ~ x))
(Intercept)          x
-0.01618698  1.00643254
> coef(lm(y ~ z + x))
(Intercept)          z          x
 0.0162384  2.0220170  3.0270318
```

Pfadanalyse

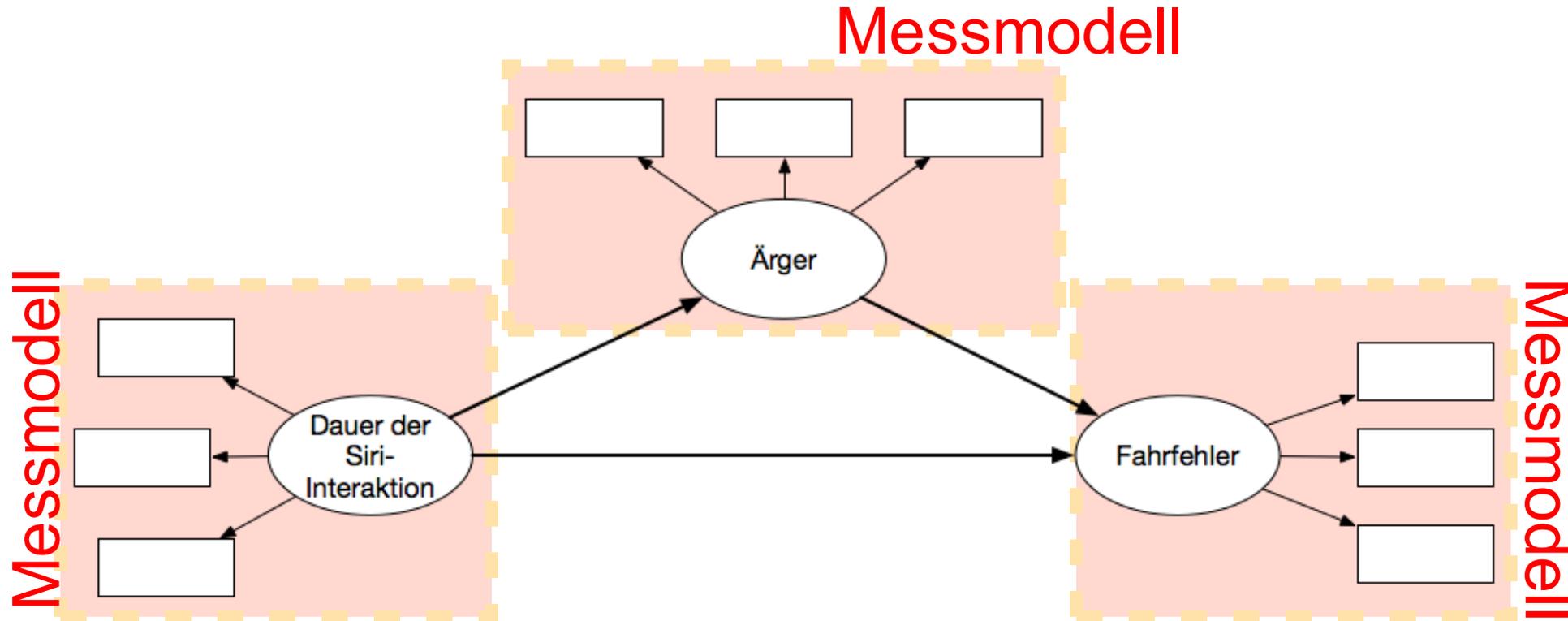
```
> library(lavaan)
>
> model <- "
+   z ~ a * x
+   y ~ b * z + c * x
+ "
> fit <- sem(model, data = dat)
> coef(fit)
      a      b      c  z~~z  y~~y
1.006 2.022 3.027 1.080 1.059
```

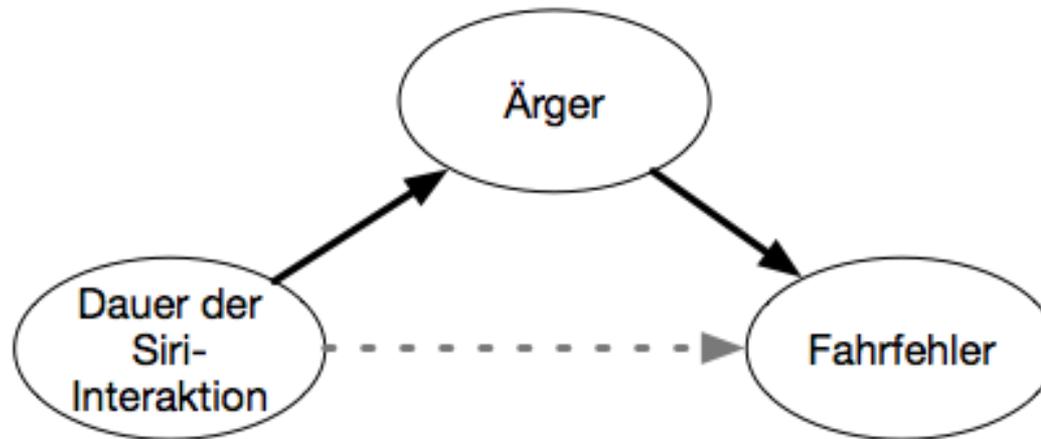


- Latente Konstrukte modellieren
- Optimale Gewichtung der Indikatoren

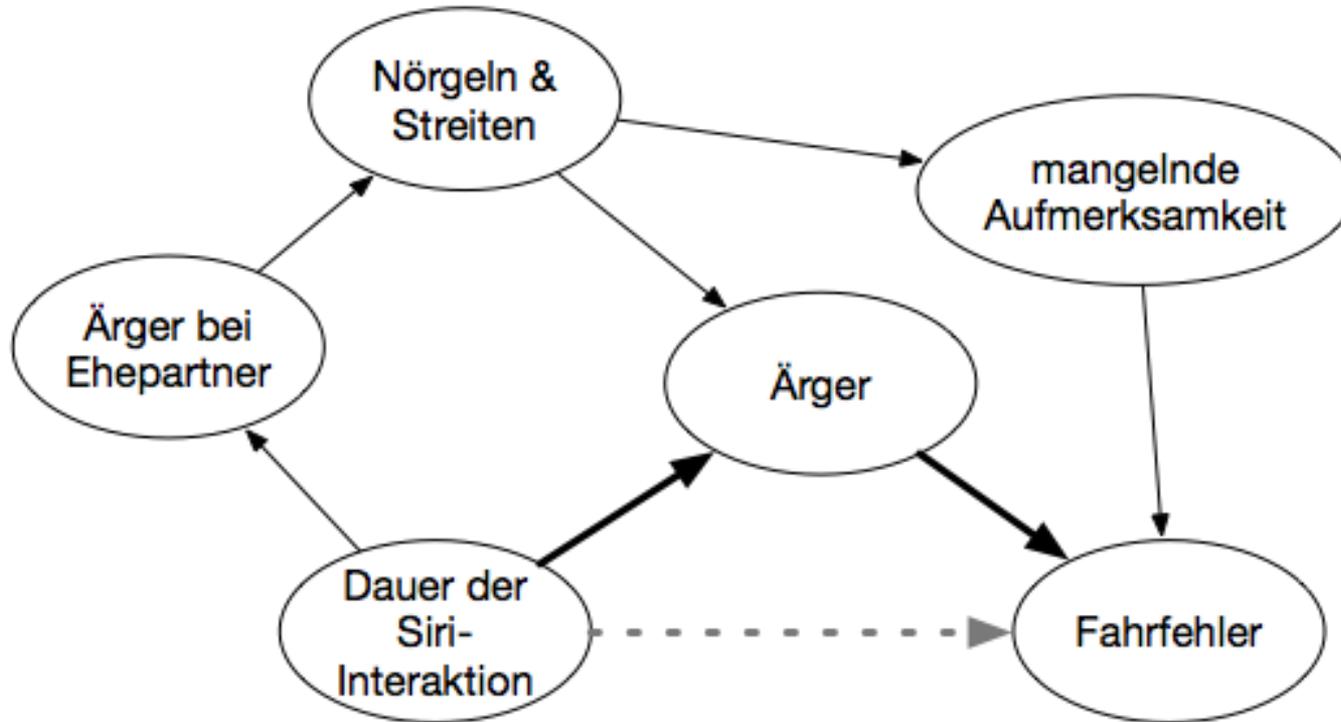
Wie werden die einzelnen latenten Variablen „gemessen“?

➤ Auf welche „Indikatoren“ wirken sie sich jeweils aus.

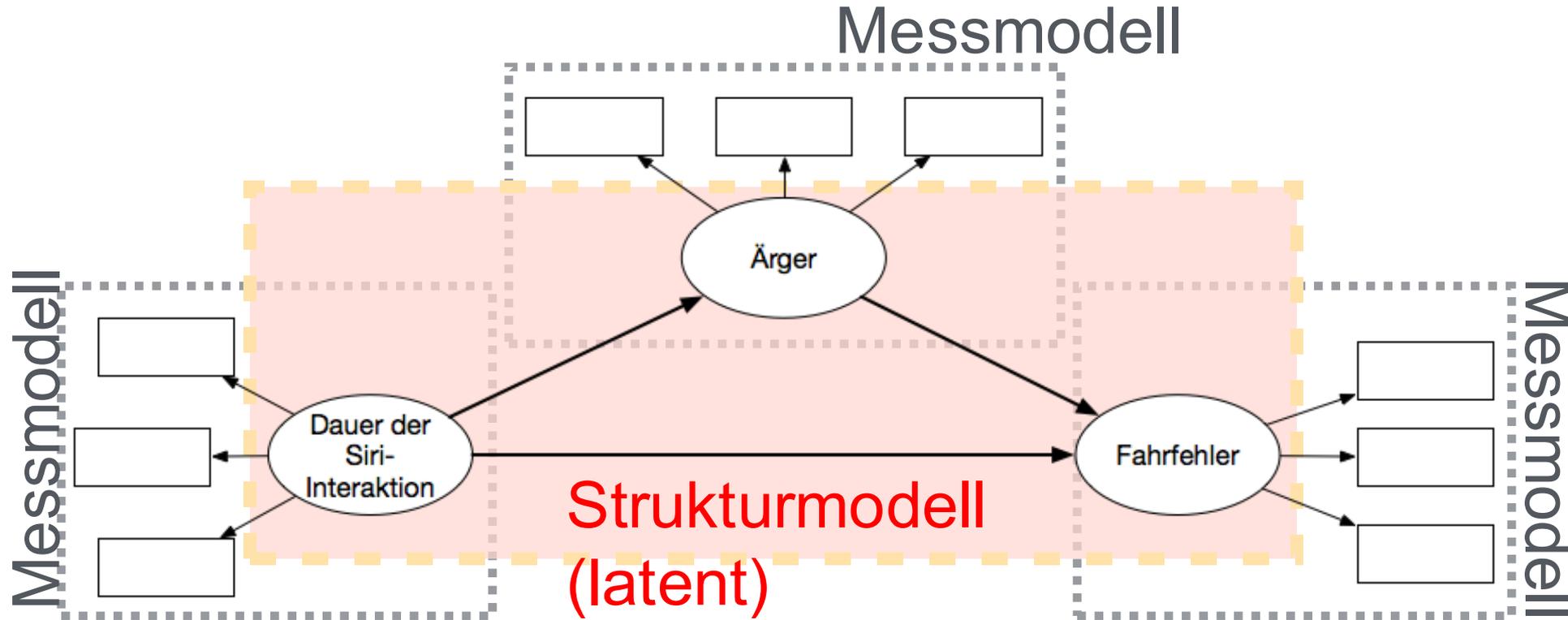




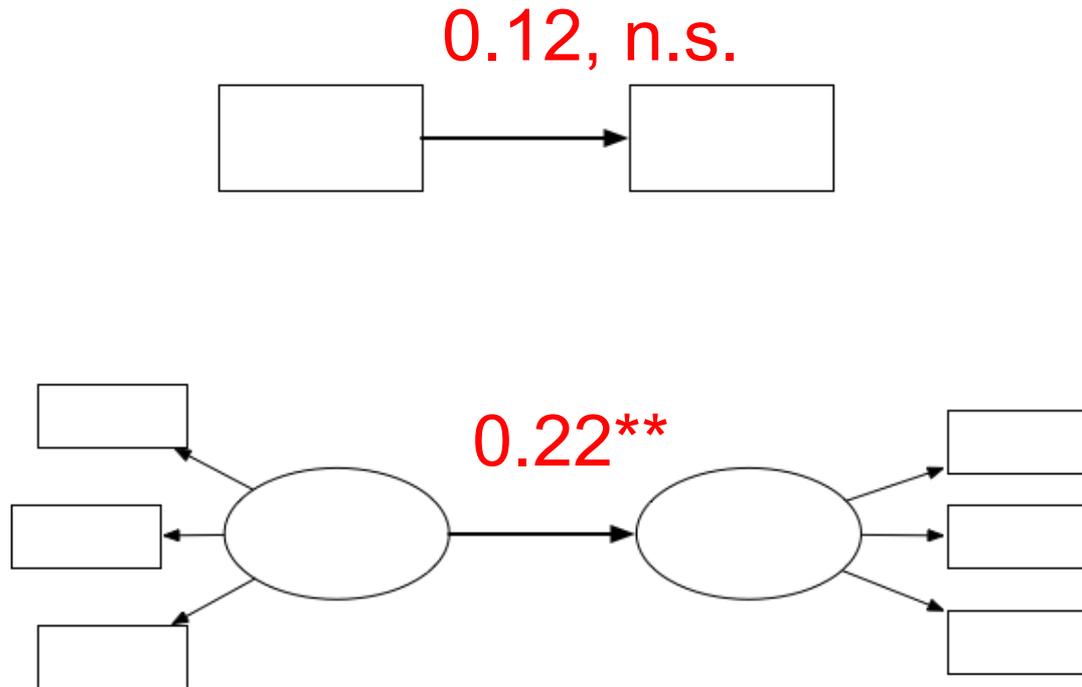
- “Multiequation models”:
Erweiterung von linearer Regression
- Indirekte Effekte



- “Multiequation models”:
Erweiterung von linearer Regression
- Indirekte Effekte



- **Messmodell:** beschreibt die Verknüpfung zwischen einer latenten Variable und ihren Indikatoren bzw. manifesten Variablen
→ spezifiziert die Operationalisierung des Konstrukts
- **Strukturmodell:** beschreibt die Verknüpfung zwischen latenten Variablen, oder zwischen latenten Variablen und manifesten Variablen, welche nicht als Indikatoren für latente Variablen verwendet werden

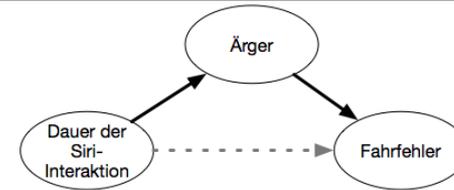


- Schätzung der messfehlerbereinigten Zusammenhänge auf latenter Ebene

Was leistet die SEM-Maschine?

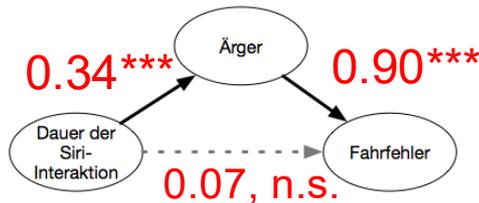
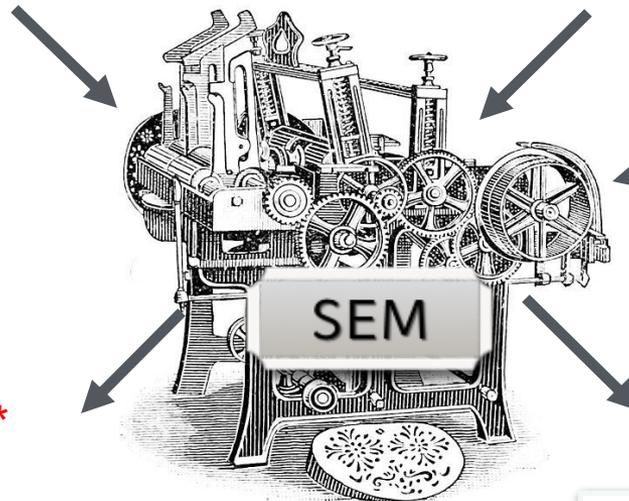
	mimic	voice	heartrate	att1
1	0.857	-0.167	-2.656	-1.032
2	-1.736	0.134	-0.725	0.636
3	-0.519	0.585	-1.422	0.037
4	2.195	0.820	1.869	0.743
5	2.221	0.278	0.099	-0.308
6	1.933	0.313	0.265	1.696

Empirische Daten



Kausale
Strukturannahmen

Weitere Annahmen:
z.B. Linearität, keine
Interaktionen, normal-
verteilte Variablen



Quantifizierung

der kausalen Pfade

Wenn die kausale Struktur

tatsächlich so ist wie angenommen!

	mimic	voice	hrtrt	att1	att2	att3	att4	errs	siri
mimic	1.12								
voice	0.22	1.15							
heartrate	-0.05	0.05	1.19						
att1	0.01	0.01	0.15	0.94					
att2	-0.04	0.03	0.02	0.13	1.09				
att3	0.05	-0.13	-0.03	-0.02	0.05	1.07			
att4	-0.06	0.11	-0.12	0.02	0.07	0.29	0.46		
errors	0.55	0.48	0.07	0.18	0.33	0.42	0.71	2.93	
siri	0.30	0.19	0.05	0.13	0.21	0.19	0.13	0.97	1.29

	mimic	voice	heartrate	att1	att2	att3	att4	errors	siri
mimic	1.49								
voice	0.24	1.12							
heartrate	0.16	0.08	1.05						
att1	0.01	0.01	0.00	1.04					
att2	0.02	0.01	0.01	0.06	1.16				
att3	0.02	0.01	0.01	0.05	0.12	1.09			
att4	0.03	0.02	0.01	0.12	0.25	0.18	1.37		
errors	0.79	0.40	0.26	0.16	0.32	0.24	0.46	2.78	
siri	0.36	0.16	0.12	0.03	0.06	0.04	0.06	0.96	1.00

Einschätzung des
Modell-Fit

Passt das Modell zu den Daten?

Die Realität liefert Daten

	mimic	voice	hrtrt	att1	att2	att3	att4	errrs	siri
mimic	1.12								
voice	0.22	1.15							
heartrate	-0.05	0.05	1.19						
att1	0.01	0.01	0.15	0.94					
att2	0.04	0.03	0.02	0.13	1.09				
att3	0.05	-0.13	-0.03	-0.02	0.05	1.07			
att4	-0.06	0.11	-0.12	0.02	0.07	0.29	1.46		
errors	0.55	0.48	0.07	0.18	0.33	0.42	0.71	2.93	
siri	0.30	0.19	0.05	0.13	0.21	0.19	0.13	0.97	1.29

Das Modell sagt Daten vorher

	mimic	voice	hertrt	att1	att2	att3	att4	errors	siri
mimic	1.49								
voice	0.24	1.12							
heartrate	0.16	0.08	1.05						
att1	0.01	0.01	0.00	1.04					
att2	0.02	0.01	0.01	0.08	1.16				
att3	0.02	0.01	0.01	0.06	0.12	1.09			
att4	0.03	0.02	0.01	0.12	0.25	0.18	1.37		
errors	0.79	0.40	0.26	0.16	0.32	0.24	0.48	2.78	
siri	0.36	0.18	0.12	0.03	0.06	0.04	0.09	0.70	1.00

S empirische
Kovarianzmatrix

modell-implizierte
Kovarianzmatrix $\hat{\Sigma}$

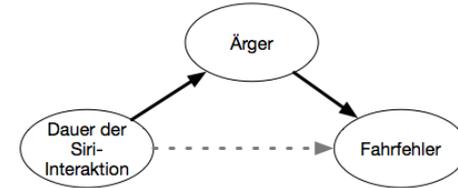
Welches Modell ist konsistent(er) zu den empirischen Daten?

→ “Fit-Indizes”

χ^2 RMSEA GFI NFI CFI CAIC BIC PNFI
 WRMR SRMR RFI AGFI AIC
 ... MECVI ...

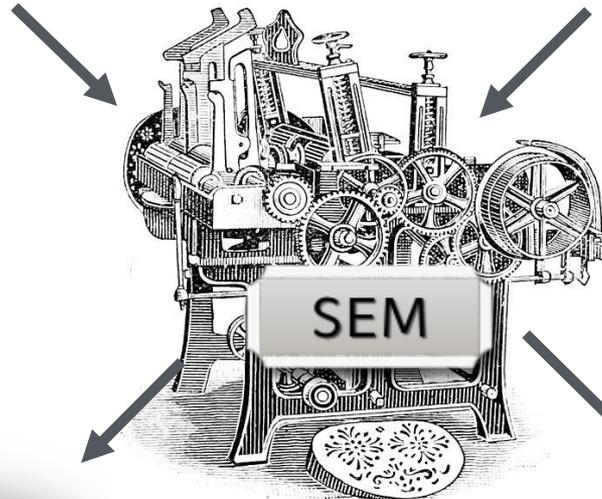
	mimic	voice	heartrate	att1
1	0.857	-0.167	-2.656	-1.032
2	-1.736	0.134	-0.725	0.636
3	-0.519	0.585	-1.422	0.037
4	2.195	0.820	1.869	0.743
5	2.221	0.278	0.099	-0.308
6	1.933	0.313	0.265	1.696

Empirische Daten

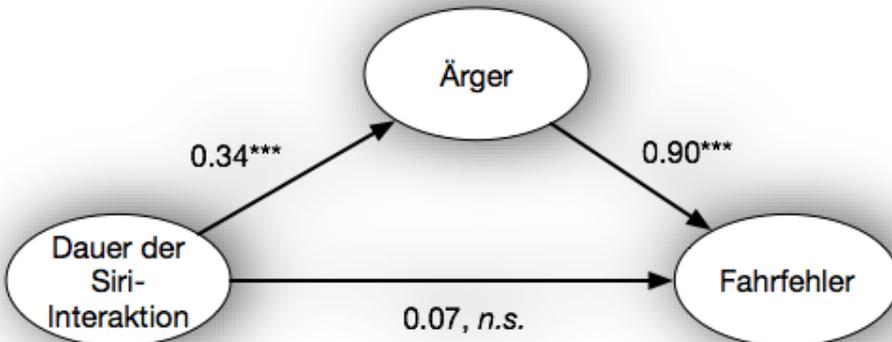


kausale
Strukturannahmen

Weitere Annahmen:
z.B. Linearität, keine
Interaktionen, normal-
verteilte Variablen



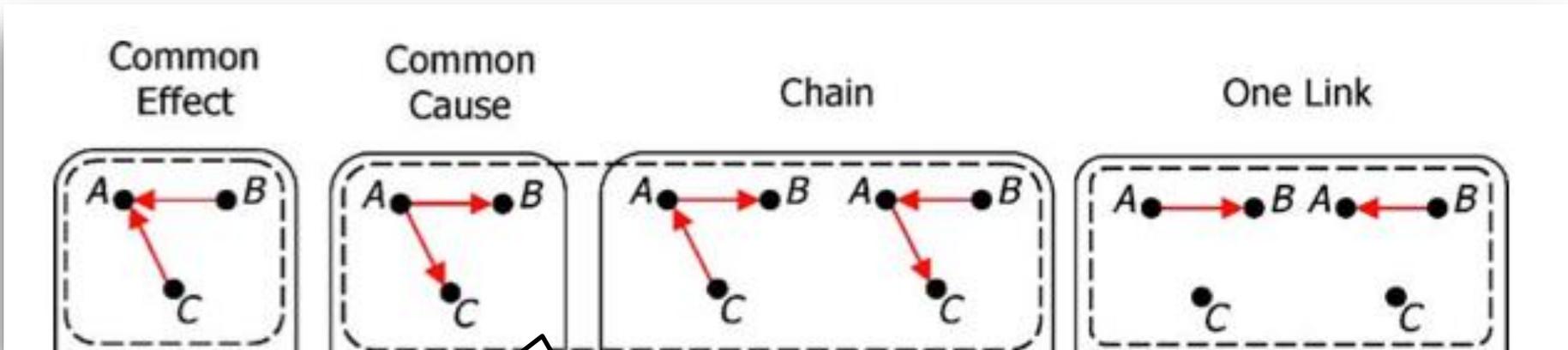
Quantifizierung
der kausalen Pfade



Einschätzung des
Modell-Fit

$\chi^2(4) = 4.338, p = .362$
CFI = 0.996
SRMR = 0.032
RMSEA = 0.024





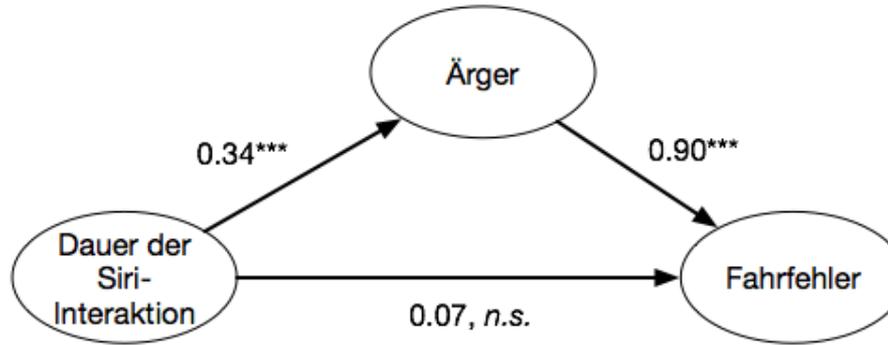
Steyvers, Tenenbaum, Wagenmakers, & Blum (2003)

Markov-äquivalente
Klasse

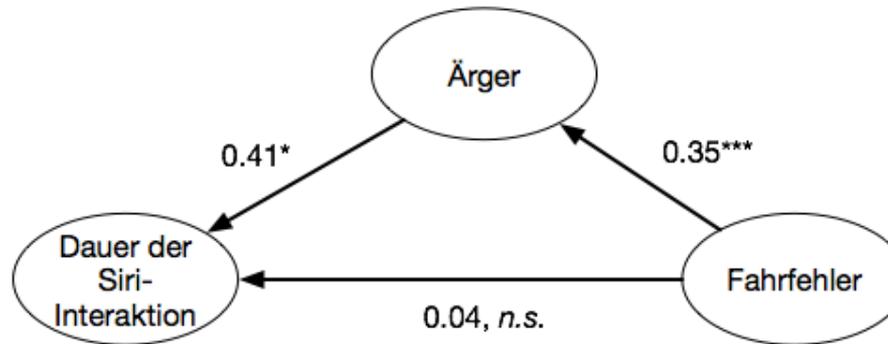
- Ohne weiteres Wissen kann aufgrund von Beobachtungsdaten nicht zwischen Markov-äquivalenten Strukturen entschieden werden
- Alle Modelle innerhalb einer gestrichelten Umrandung sind äquivalent → Wir können Modelle ausschließen, die offensichtlich nicht zu den Daten passen - wir können aber kein Modell verifizieren, bloß weil es gut zu den Daten passt
- Kausale Struktur wird a priori angenommen → Sind die Daten konsistent mit diesen Annahmen? → Daten sind gleich konsistent mit mehreren Modellen
- Der Nachweis von Kausalität ist keine Frage der Datenauswertung, sondern eine Frage des experimentellen Designs (z.B. durch Manipulation)

Ausblick: Äquivalente Modelle (aufgrund der Daten nicht unterscheidbar)

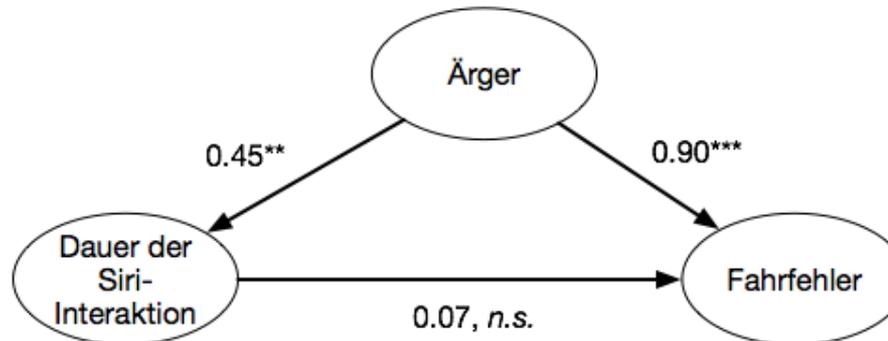
+ 3 weitere äquivalente Modelle!



$\chi^2(4) = 4.338, p = .362$
CFI = 0.996
SRMR = 0.032
RMSEA = 0.024



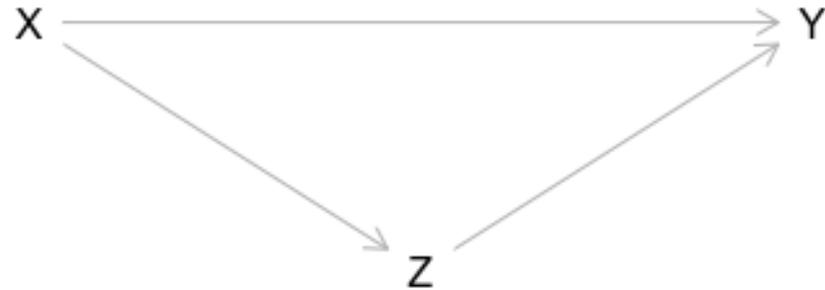
$\chi^2(4) = 4.338, p = .362$
CFI = 0.996
SRMR = 0.032
RMSEA = 0.024



$\chi^2(4) = 4.338, p = .362$
CFI = 0.996
SRMR = 0.032
RMSEA = 0.024

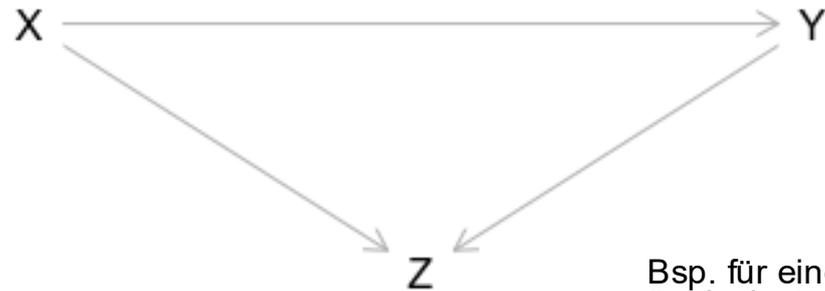
- „If a model is consistent with reality, then the data should be consistent with the model. But if the data are consistent with the model, this does not imply that the model corresponds to reality.“ (Bollen 1989)
- Verdeutlicht den Wert und die Wichtigkeit sauberer Theoriearbeit!
- Es ist unabdingbar, die Rolle einer jeden Variable theoretisch zu begründen.
- Warum wird eine Variable ins Modell aufgenommen? Ist sie ein Mediator oder ein Confounder? Was sagen Theorie und Empirie?

```
> library(dagitty)
>
> dag <- dagitty('dag{
+   X -> Y
+   X -> Z -> Y
+   X[pos="1, 1"]
+   Y[pos="2, 1"]
+   Z[pos="1.5, 2"]
+   }')
>
> plot(dag)
```



Ursprünglicher DAG

```
> length(equivalentDAGs(dag))
[1] 6
> plot(equivalentDAGs(dag)[[2]])
```



Bsp. für einen
äquivalenten DAG



- ✓ (Relativ) einfache Übersetzung:
komplexes verbales Modell → formales Modell
- ✓ Latente Modellierung
- ✓ Berücksichtigung von Messfehlern



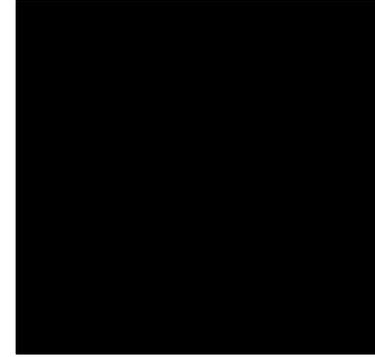
-
- Nur kausale Konsistenz prüfen - nicht Kausalstrukturen “beweisen”
 - Es gibt immer äquivalente Modelle
 - z.T. Kontroversen und kein Konsens um manche Fit-Indizes, Schätzmethoden, etc. (s. nächste VLs)

- Auf den folgenden Folien, werden wir die wichtigsten Konventionen zur grafischen Darstellung von SEMs, sowie die Definition von damit verbundenen neuen Begriffen nochmal möglichst vollständig auflisten.
- Wir versuchen uns in unseren Veranstaltungen soweit wie möglich daran zu halten.
- Ein vielen Stellen verwenden wir Vereinfachungen der ausführlichen Notation, die hoffentlich aus dem Kontext nachvollziehbar sind.
- Hinweis: In der Praxis existieren sehr viele verschiedene Variationen für einzelne Details der grafischen Darstellungen von SEMs.

Manifeste Variable:

Variable, die direkt beobachtbar ist.

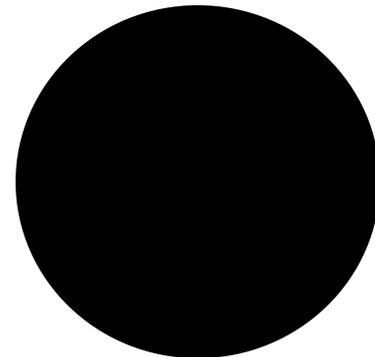
- Ein einzelnes Item in einem Fragebogen
- Ein „physisches“ Merkmal einer Person (Augenfarbe, Körpergröße, etc.)
- Sonstige beobachtbare Indikatoren wie Einkommen, Anzahl der Kinder, Studiengang)



Latente Variable:

Variable, die nicht direkt beobachtbar ist.

- bspw. psychologische Konstrukte wie Intelligenz oder Persönlichkeit (Neurotizismus, Need for Closure o.ä.)
- Nur indirekt beobachtbar, durch die Ausprägung auf einer manifesten Variable (die *ursächlich* auf die Ausprägung der latenten Variable zurückzuführen ist)



Mögliche Verbindungen zwischen zwei Variablen

- **Gerichteter Zusammenhang:**

- Grafisch: einfacher Pfeil 
- Numerisch: Stärke eines gerichteten (linearen) Zusammenhangs = Regressionsgewicht (synonym: Ladung, Gewicht, Pfad)
- Kausale Interpretation: direkter kausaler Effekt

- **Ungerichteter Zusammenhang:**

- Grafisch: Doppelpfeil 
- Numerisch: Stärke eines ungerichteten Zusammenhangs = Kovarianz/Korrelation
- Kausale Interpretation: Abgekürzte Notation für fehlenden Confounder (d.h. $X \leftrightarrow Y$ äquivalent zu $X \leftarrow U \rightarrow Y$ mit U unbekannt)

- Ein (gerichteter) Zusammenhang muss theoretisch begründet sein!

Endogene Variable = abhängige Variable:

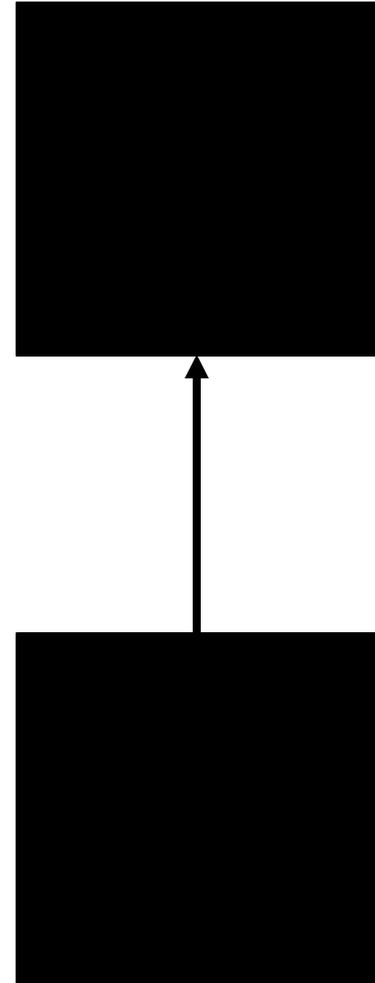
Variable, die durch andere Variablen erklärt/beeinflusst wird

- die meisten manifesten Variablen
- manche latenten Variablen
- grafisch: mind. 1 *gerichteter* Pfeil zeigt auf die endogene Variable (Doppelpfeile gelten nicht!)
- Benötigen zugehörige Fehlervariable (da wir nie Variablen *perfekt* aus anderen vorhersagen)

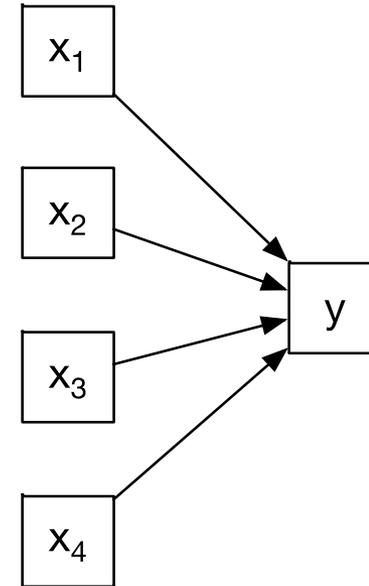
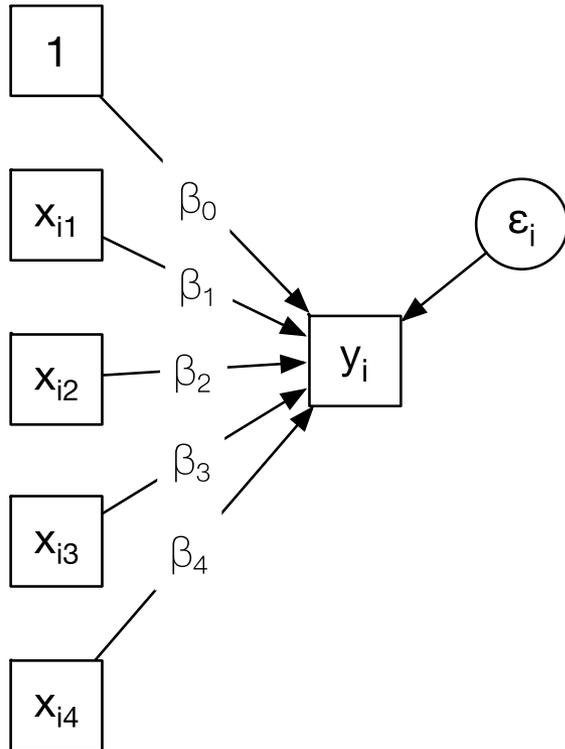
Exogene Variable = unabhängige Variable:

Variable, die nicht durch andere Variablen erklärt/beeinflusst wird

- alle Residuen/Fehlervariablen
- manche latente Variablen
- manche manifeste Variablen
- grafisch: kein gerichteter Pfeil zeigt auf die exogene Variable



Intuition zu Fehlervariablen: Multiple lineare Regression als SEM



Vereinfachte Darstellung:

- ohne Intercept
- ohne Personenindex i
- Manchmal ohne Fehlervariable ϵ

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \epsilon_i \quad (i = 1, 2, \dots, n)$$

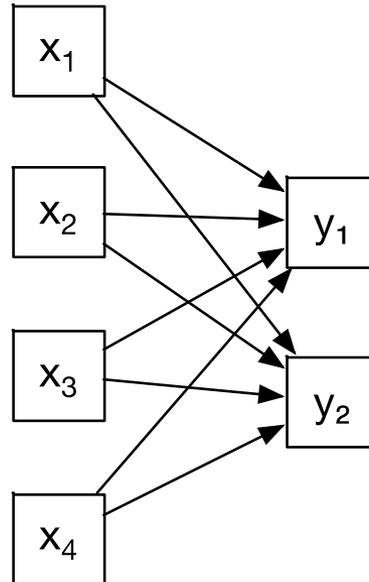
Analyse mit Linearer Regression

```
> n <- 1000
> b0 <- 0
> b1 <- 1
> b2 <- 2
> b3 <- 3
> b4 <- 4
> set.seed(1)
> x1 <- rnorm(n)
> x2 <- rnorm(n)
> x3 <- rnorm(n)
> x4 <- rnorm(n)
> y <- rnorm(n, mean = b0 + b1 * x1 +
+      b2 * x2 + b3 * x3 + b4 * x4)
> dat <- data.frame(x1 = x1, x2 = x2,
+      x3 = x3, x4 = x4, y = y)
```

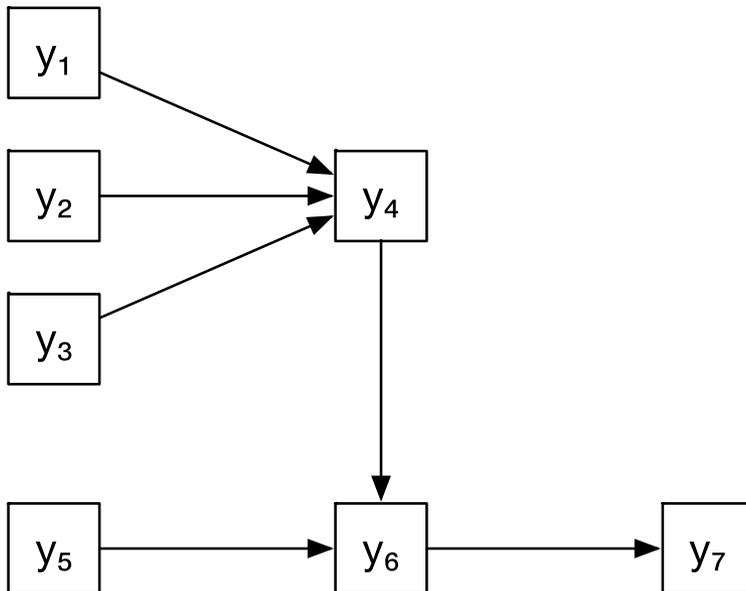
```
> coef(lm(y ~ x1 + x2 + x3 + x4))
(Intercept)          x1          x2          x3          x4
-0.01924457  0.96671025  2.01083692  2.92940219  4.00300811
```

Analyse als Pfadmodell

```
> library(lavaan)
>
> model <- "
+   y ~ x1 + x2 + x3 + x4
+ "
> fit <- sem(model, data = dat,
+   meanstructure = TRUE)
> coef(fit)
      y~x1  y~x2  y~x3  y~x4  y~~y  y~1
0.967  2.011  2.929  4.003  0.971 -0.019
```



- Pfadanalyse: Modelliert kausale Beziehungen zwischen beobachteten Variablen -> alle Variablen sind beobachtet (manifest), bis auf die Fehler
- Entspricht einem System von (linearen) Regressionsgleichungen



$$y_4 = \beta_{01} + \beta_1 y_1 + \beta_2 y_2 + \beta_3 y_3 + e_1$$
$$y_6 = \beta_{02} + \beta_4 y_4 + \beta_5 y_5 + e_2$$
$$y_7 = \beta_{03} + \beta_7 y_6 + e_3$$

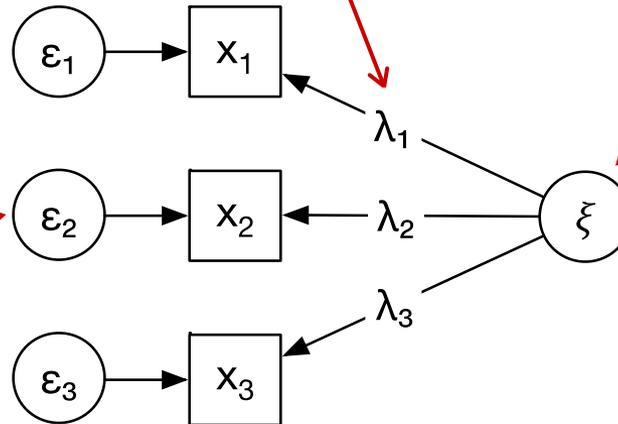
Beispiel I

Latente Variablen: Messmodell

Ladung „lambda“ (Pfadkoeffizient,
Regressionsgewicht) der latenten
Variable auf eine manifeste Variable,
gerichteter Pfeil

exogene latente Variable „xi“ (kein Pfeil
zeigt darauf), nicht beobachtbar, nicht
durch eine andere Variable erklärt

Latente Fehlervariable „epsilon“, bei
allen endogenen Variablen (min. ein Pfeil
zeigt darauf)



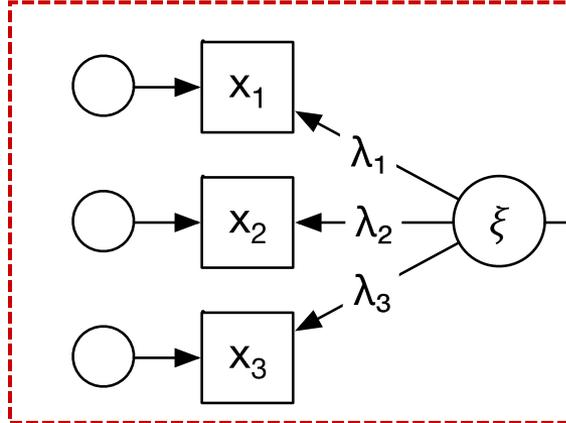
Manifeste (=beobachtete) Variable X

- Kausale Beziehungen zwischen latenten Variablen

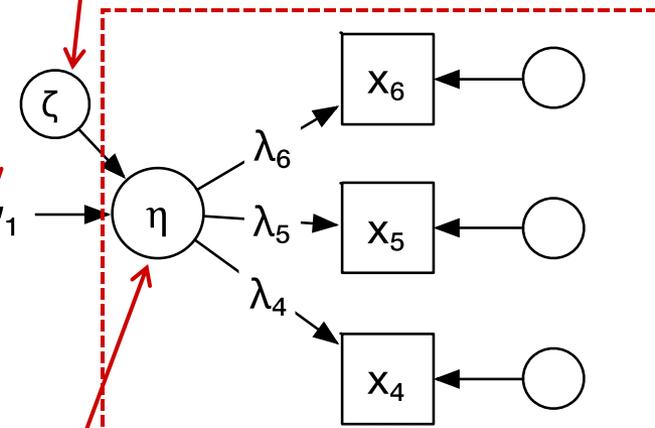
Gerichtete Ladung/Pfadkoeffizient
zwischen latenten Variablen („gamma“)

Latente Fehlervariable „zeta“,
einer endogenen latenten
Variable: Welcher Varianzanteil
in η kann nicht durch ξ erklärt
werden?

Messmodell der exogenen Prädiktorvariable ξ



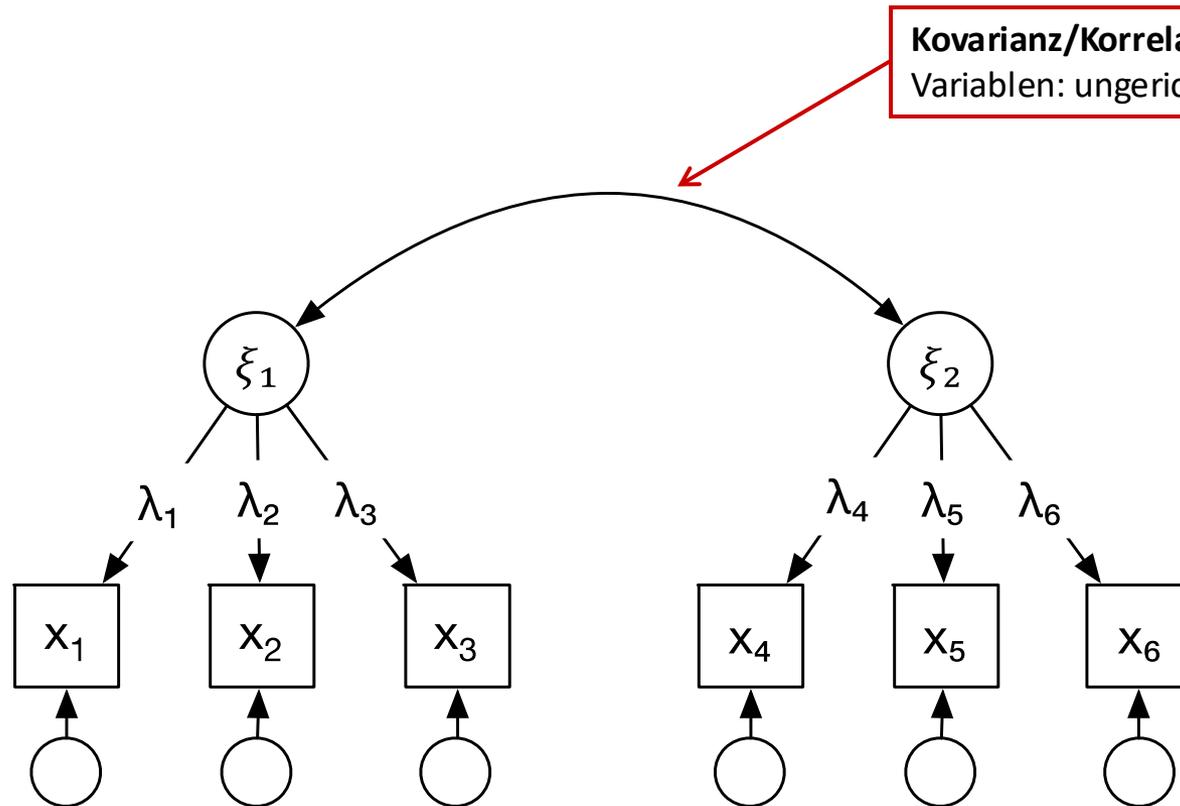
Messmodell der endogenen Kriteriumsvariable η



endogene latente Variable „eta“ (mind. ein Pfeil zeigt darauf), nicht direkt beobachtbar, aber durch Indikatoren gemessen und durch eine andere Variable (hier: „xi“) erklärt

Beispiel III

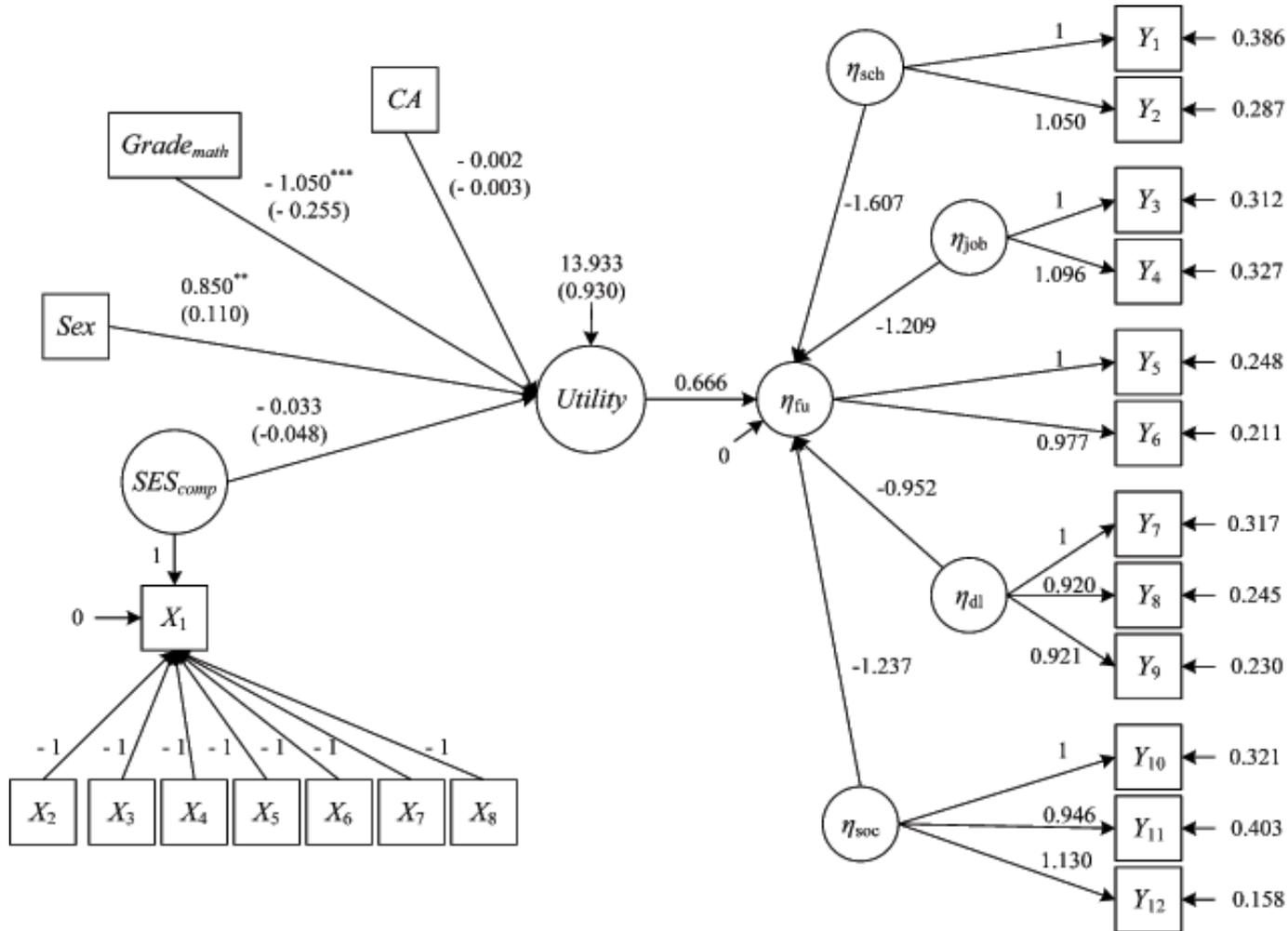
Confirmatory Factor Analysis (CFA)



Kovarianz/Korrelation zweier
Variablen: ungerichteter Pfeil

Vereinfachte Darstellung:
Fehlerterme ohne Label

Beispiel für komplexes SEM



- beobachtete, so genannte manifeste Variable
- latente Variable/Fehlervariable
- ↔ Kovarianz oder Korrelation
- ← semipartielles Regressionsgewicht oder Ladung (λ , Lambda) für Items und γ (Gamma) für latente Variablen
- ε Fehlervariable einer manifesten Variablen (Epsilon)
- ζ Fehlervariable einer latenten abhängigen Variablen (Zeta)
- ξ latente unabhängige Variable (Ksi), von der ein Pfeil ausgeht und auf die kein Pfeil zeigt
- η latente abhängige Variable (Eta), auf die ein Pfeil zeigt
- X Indikatoren der latenten Variablen