

Von stimme zu bis stimme überhaupt nicht zu: Grundlagen der Skalenentwicklung und Testtheorien

Dr. Philipp K. Masur | philipp.masur@uni-mainz.de

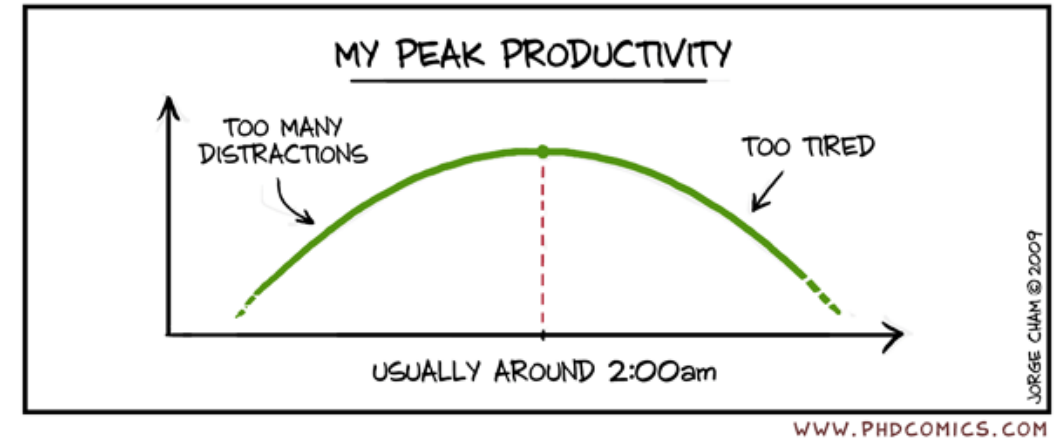
Würzburg | 23. Januar 2020

Willkommen!

Ich heie Philipp und beschftige mich gerne mit Messtheorie.

Ablauf

Uhrzeit	Thema
10:00 - 10:45:	Einführung
10:45 - 12:00:	Erstmal alles genau anschauen.
12:00 - 13:00:	Strukturen entdecken - leicht gemacht?
13:00 - 14:00:	Mittagessen
14:00 - 14:45:	Aber es gibt doch Theorie!
14:45 - 15:15:	Alles eine Frage der Güte, oder?
15:15 - 15:35:	Kaffeepause
15:35 - 17:00:	Probabili-was?



Hinweise zum Workshops

Theoretischer Teil

- Der Workshop besteht zu gleichen Teilen aus *theoretischer* und *praktischer* Beschäftigung mit Testtheorie
- Im Theorieteil
 - Gerne direkt Fragen stellen
 - Eigene Probleme und Erfahrungen gerne mitteilen
 - Diskussion erwünscht
- Es geht darum die Grundideen der unterschiedlichen Testtheorien gut zu verstehen

Praktischer Teil

- Umgang im Workshop
 - Wir lernen gemeinsam R-Code auszuführen
 - In Übungen wird das Gelernte selbst angewendet
 - Der Workshop soll interaktiv sein, Fragen also gerne *direkt* stellen
- Wichtige Hinweise
 - Es wird nicht alles auf Anhieb klappen (keine Panik!)
 - Syntax-basiert arbeiten bedeutet, dass kleinste "Abweichungen" im Code zu Fehlern führen

Materialien

Bitte hier runterladen:

<http://philippmasur.de/blog/workshop/>

Dann "entzippen" und den Ordner auf den Desktop ziehen.

Mini-Einführung in R

So definiert man eine Zahl oder einen Vektor:

```
# Definieren einer Zahl  
x <- 1  
  
# Ausgeben der Variable  
x
```

```
[1] 1  
|
```

```
# Definieren eines Vektors  
vektor1 <- c(1,2,3,4,5)  
vektor2 <- c(5,4,3,2,1)  
  
# Ausgeben des Vektors  
vektor1
```

```
[1] 1 2 3 4 5  
|
```

So kann man bestimmte Komponenten einer Variable ausgeben

```
# Teile eines Vektors ausgeben  
vektor2[1:2]
```

```
[1] 5 4
```

```
# Direkt abspeichern  
y <- vektor1[1]  
y
```

```
[1] 1
```

```
# Logical tests  
x != y
```

```
[1] FALSE
```

Mini-Einführung in R

In R arbeiten wir mit Objekten, nicht mit dem Datensatz selbst:

```
# Erstellen eines Datensatzes
data <- data.frame(var1 = vektor1,
                   var2 = vektor2)

# Ausgeben des Datensatzes
data
```

```
  var1 var2
1     1    5
2     2    4
3     3    3
4     4    2
5     5    1
:
```

```
# Rekodieren einer Variablen
data$var1 <- data$var1 + 5
data$var1
```

Wir nutzen **Funktionen** um statische Prozeduren auszuführen:

```
# Mittelwert berechnen
mean(data$var1)
```

```
[1] 8
```

```
summary(data)
```

```
      var1      var2
Min.   : 6  Min.   :1
1st Qu.: 7  1st Qu.:2
Median : 8  Median :3
Mean   : 8  Mean   :3
3rd Qu.: 9  3rd Qu.:4
Max.   :10  Max.   :5
```

Das 'tidyverse' als Werkzeugbox



- Data Wrangling kann sehr komplex und schwierig sein
- Das 'tidyverse' ist eine Sammlung von ineinandergreifenden Paketen
- Eigene Codierlogik und Philosophie
- Prominentes Beispiel: ggplot2 (welches wir auch kennenlernen)

Vergleich zwischen klassischer und 'tidyverse'-Codierlogik

```
# klassisch
summary(data[1:2,])

# tidyverse
data %>%
  select(var1, var2) %>%
  summary
```

Zentrales Werkzeug ist die sogenannte "pipe" (`%>%`). Hiermit lassen sich einzelnen Arbeitsschritte miteinander verbinden.

```
# klassisch
m <- mean(data$var1)
sd <- sd(data$var1)
cbind(m, sd)

# tidyverse
data %>%
  summarize(m = mean(var1),
            sd = sd(var1))
```

Der Shortcut für "`%>%`" ist: *ctrl+shift+m*

Einführung in die klassische Testtheorie

Wo steckt der wahre Wert?

Ausgangslage

- In den Sozialwissenschaften interessieren uns häufig abstrakte Konstrukte bzw. Merkmale (z.B. Emotionen, Einstellungen, Intelligenz, Medienkompetenz, Persönlichkeitseigenschaften...)
- Diese lassen sich häufig nicht *direkt* messen, sondern müssen *indirekt* über beobachtbare Indikatoren (z. B. Fragen in Form von Items) erfasst werden
- Es gilt möglichst unterschiedliche Aspekte des abstrakten Konstrukts über *mehrere* Fragen (Items) abzudecken

Attitudes and Emotions



Warum überhaupt Testtheorie?

- Wir wollen den Zusammenhang zwischen einem zu messenden Merkmal (z.B. eine Einstellung) und der Beantwortung einer bzw. mehrerer Items erklären
- Testtheorie bedeutet Messungen in einen statistischen Zusammenhang mit dem eigentlichen Merkmal zu bringen
- Testtheorie definiert, wie eine konkrete Messung zustande kommt und welche Einflussgrößen auf die Messung wirken
- Man unterscheidet zwischen:
 - **klassischer Testtheorie** (True-Score-Theorie)
 - **probabilistischer Testtheorie** (Item Response Theory)

Typische Fragestellungen

1. Was versteht man unter einem wahren Messwert?
2. Welche Auswirkungen haben Messfehler?
3. Wie lässt sich überprüfen, ob verschiedene Messinstrumente dasselbe Merkmal erfassen?
4. Wie kann man die Zuverlässigkeit einer Messung bestimmen?
5. Wie ist die Reliabilität eines Test definiert und wie lässt sie sich bestimmen?
6. Wie kann die Zuverlässigkeit eines Test gesteigert werden bzw. Items für einen Test optimal ausgewählt werden?

Wichtige Begriffe

- **Latente Variablen:** die nicht direkt beobachtbaren, aber uns interessierenden Merkmale (später auch Faktoren genannt), wie z.B. Intelligenz, Meinungen, Motive, Persönlichkeitseigenschaften, u.v.m.
- **Manifeste Variablen:** Messbare Indikatoren, wie z.B. die Antwort auf eine Frage (es sind aber auch andere Arten von Indikatoren denkbar!)
- **Wahrer Wert (True-Score):** Der Anteil der Messung, der auf das Merkmal zurückzuführen ist und den es deswegen möglichst gut zu ermitteln gilt.
- **Messfehler:** Der Anteil einer Messung, der nicht auf das Merkmal zurückzuführen ist (i.e., unsystematische Einflüsse, wie z.B. nicht Verstehen der Frageinstruktion, Verwechseln der Antwortkästchen, Fehler bei der Datenübertragung, etc...)

Annahmen der klassischen Testtheorie

- Die klassische Testtheorie beruht auf dem Prinzip der **Varianzzerlegung**: Jeder beobachtete Messwert Y besteht aus dem **wahren Wert** τ und dem **Messfehler** ϵ . D.h. wie können die Varianz des Messwertes additiv in die Varianz des True-Scores und die Varianz des Messfehlers zerlegen:

$$Y_i = \tau_i + \epsilon_i$$

- Der Messfehler ist unabhängig vom wahren Wert und schwankt zufällig über Personen und Messgelegenheiten. D.h. Messfehler- und der True-Score-Variablen sind unkorreliert:

$$Cor(\tau_i, \epsilon_i) = 0$$

- Der Erwartungswert (Mittelwert) des zufälligen Messfehlers ist 0, seine Varianz die sogenannte Messfehlervarianz:

$$\epsilon_i = 0$$

Reliabilität

- Die additive **Varianzzerlegung** ist damit die Grundlage für die Definition eines der wichtigsten Kennwerte der klassischen Testtheorie: der **Reliabilität** einer beobachteten Variablen (des manifesten Indikators!)
- Sie ist definiert als Anteil des wahren Wertes τ am gemessenen Wert:

$$Rel(Y_i) = \frac{Var(\tau_i)}{Var(Y_i)} = \frac{Var(\tau_i)}{Var(\tau_i) + Var(\epsilon_i)}$$

- Die Reliabilität ist damit ein Maß für die **Messfehlerfreiheit** einer Messung

Schätzung der Reliabilität

- Bei einer **Mehrfachmessung** gilt, dass die Kovarianz von zwei Items derselben Skala, der Varianz des wahren Wertes entspricht:

$$\text{` } Var(\tau) = Cov(Y_1, Y_2) \text{ `}$$

- Bei **Test-Retest** wird dieselbe Messung mehrfach durchgeführt und die Korrelation der beiden Messungen entspricht der Reliabilität der Variable:

$$\text{` } Cor(Y_{T1}, Y_{T2}) = Rel(Y) \text{ `}$$

- Beim **Parallel-Test** werden verschiedene Messinstrumente, die dasselbe Merkmal messen, verwendet (z. B. ***X*** und ***Y***). Deren Kovarianz entspricht wiederum der Varianz des True-Score:

$$\text{` } Var(\tau_i) = Cov(X_i, Y_i) \text{ `}$$

True-Score-Modelle vs. (komplexere) Faktormodelle

- Klassische True-Score-Modelle unterscheiden nur zwischen **wahrem Wert** τ und dem **Messfehler** ϵ
- Faktormodelle (mehr als zwei Items) gehen dagegen davon aus, dass der Anteil, der nicht durch die latente Variable erklärt wird, nicht nur Messfehlereinflüsse repräsentiert wird.
- Daneben gibt es einen weiteren wahren True-Score-Anteil, der nicht durch Faktoren gebunden ist, sondern durch ein (oder mehrere) weitere Merkmale erklärt wird:

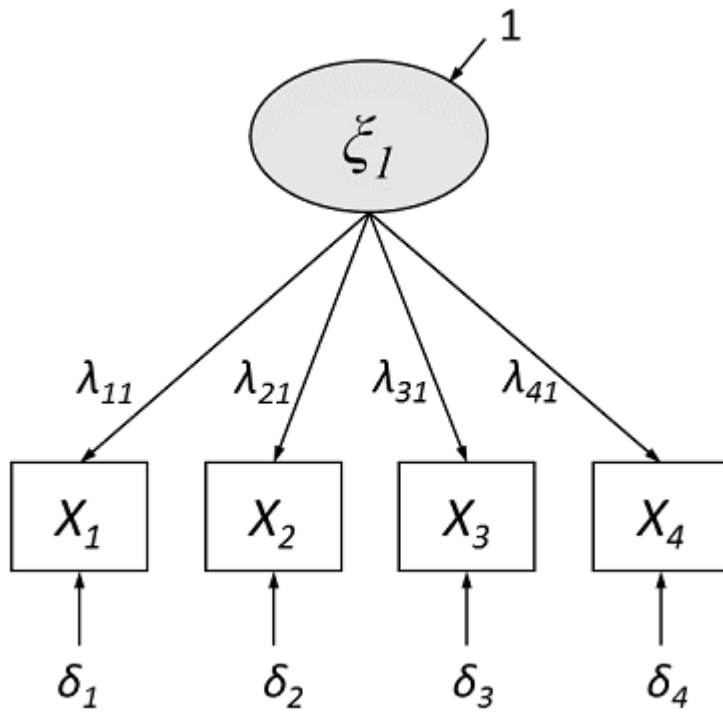
$$Y_i = \tau_{ig} + \tau_{is} + \epsilon_i$$

τ_g = **Gemeinsame Varianz** = die Varianz einer Variable, die mit allen anderen Indikatoren geteilt wird

τ_s = **Spezifische Varianz** = Itemspezifische Varianz, die nicht durch die Interkorrelation mit anderen Variablen erklärt werden kann (durch ein oder mehrere "andere" Merkmale erklärbar)

ϵ_{mi} = **Fehlervarianz** = Varianz, die auf unsystematische Messfehler zurückgeht

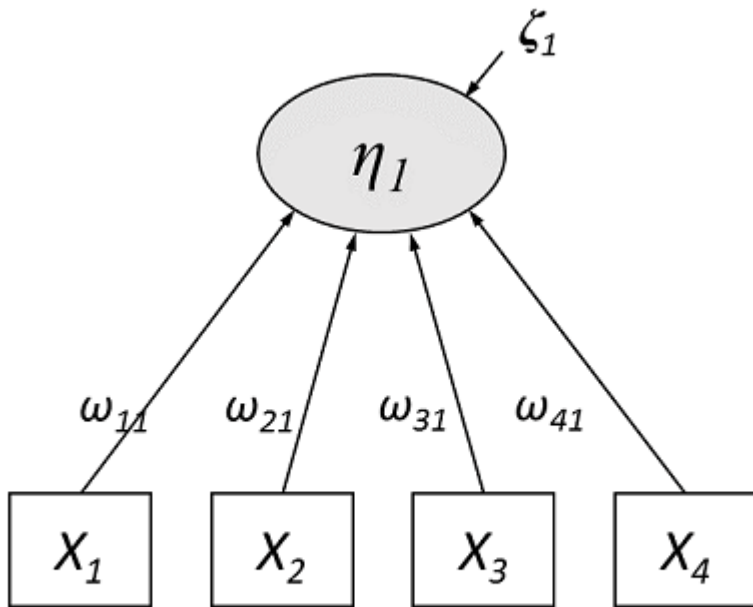
Reflektives Modell



- ξ = latente Variable (nicht beobachtet)
- λ = Faktorladung (= Regressionskoeffizient)
- X = manifeste Variable (Indikator)
- δ = Messfehler (= Itemspezifische + Fehlervarianz)

Achtung: Im reflektiven Messmodell erklärt die latente Variable die Indikatoren und nicht umgekehrt!

Kurzer Exkurs: Formatives Messmodell



- η = latente Variable (nicht beobachtet)
- ω = Regressionskoeffizient (= Gewichtung)
- X = manifeste Variable (Indikator)
- ζ = Messfehler des latenten Konstruktes

Achtung: Im formativen Messmodell erklären die Indikatoren die Variable!

Formativ oder reflektiv?

Frage 1: Würde eine Änderungen der Ausprägung der latenten Variable zu einer Veränderung der Indikatoren führen?

Antwort: ja = reflektives Modell

Frage 2: Würde die Elimination eines Indikators den konzeptionellen Inhalt der latenten Variablen verändern?

Antwort: ja = formatives Modell

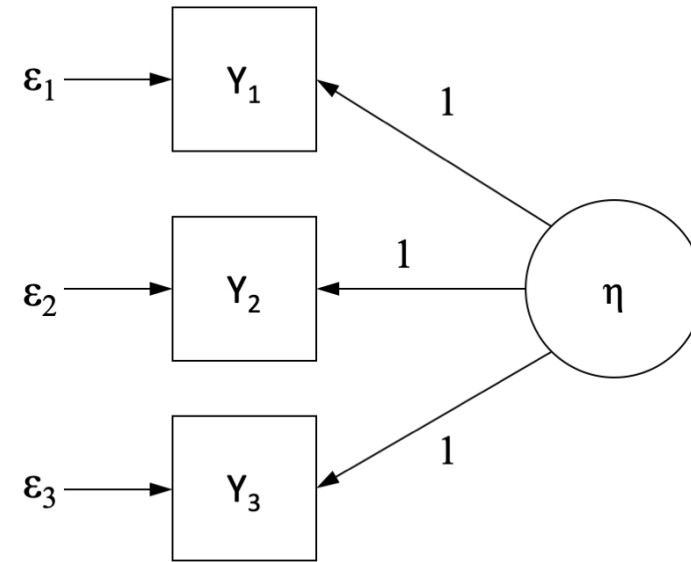
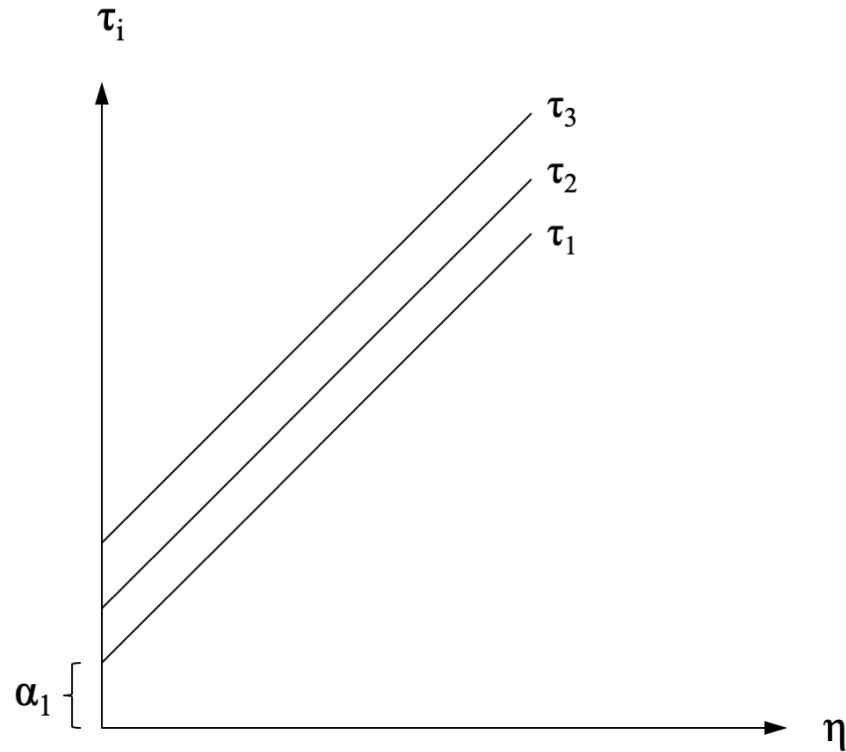
Frage 3: Sollte eine Änderung eines Indikators zu einer Veränderung der anderen Indikatoren führen?

Antwort: ja = reflektives Modell

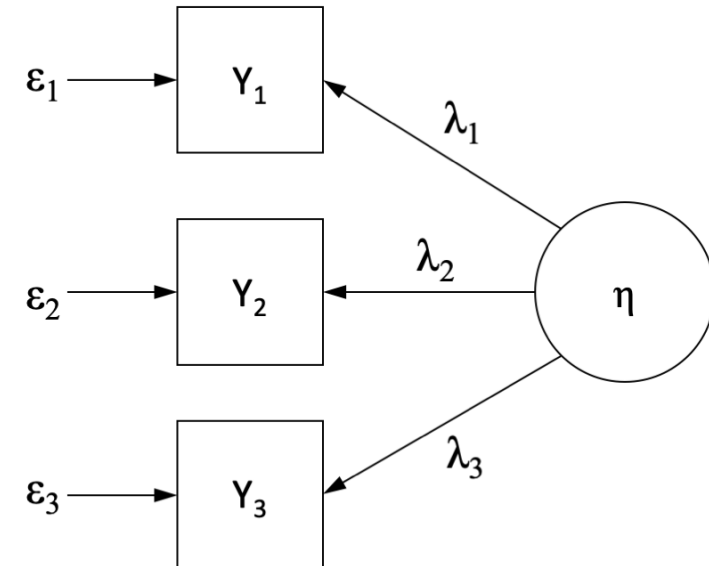
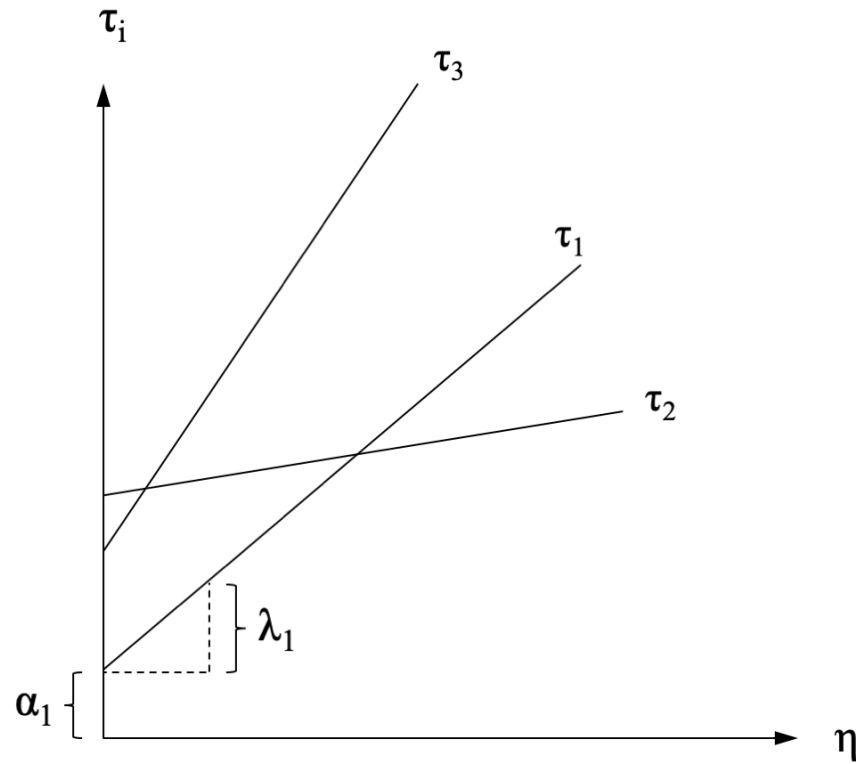
Ein letzter, aber wichtiger Aspekt: tau-Äquivalenz

- Bei der Mehrfachmessung können sich die einzelnen Items (i.e. Messungen) auf unterschiedlichen "Ebenen" unterscheiden:
 - Varianz des wahren Wertes: τ
 - Varianz der Messfehlereinflüsse: ϵ
 - Varianz in der Schwierigkeit (häufig als α bezeichnet)
- Je nachdem, welche dieser Eigenschaften variieren, unterscheiden wir in folgende Modelle
 - **essentiell τ -äquivalent**
 - **essentiell τ -parallel**
 - **τ -äquivalent**
 - **τ -parallel**
 - **τ -kongenerisch**
- In der Praxis verwenden wir wenn dann nahezu essentiell τ -äquivalente Modelle, primär aber eher τ -kongenerische Modelle!

Modell essentiell tau-äquivalenter Variablen



Modell tau-kongenerischer Variablen



Vergleich der Messmodelle

Modell	Wahrer_Wert	Messfehler	Schwierigkeit
essentiell tau-äquivalent	korreliert perfekt zwischen Items	dürfen variieren	darf variieren
essentiell tau-parallel	korreliert perfekt zwischen Items	bei allen Variablen gleich groß	darf variieren
tau-äquivalent	korreliert perfekt zwischen Items	dürfen variieren	kein unterschied zwischen den Items
tau-parallel	korreliert perfekt zwischen Items	bei allen Variablen gleich groß	kein unterschied zwischen den Items
tau-kongenerisch	unterschiedliche Steigungsparameter	dürfen variieren	darf variieren

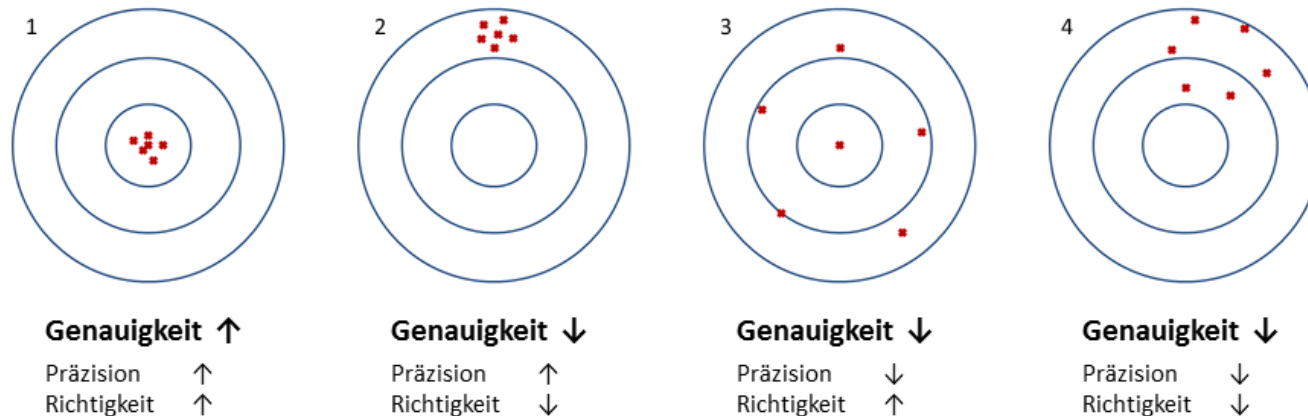
Warum ist gutes Messen wichtig?

Berücksichtigung von Messfehlern

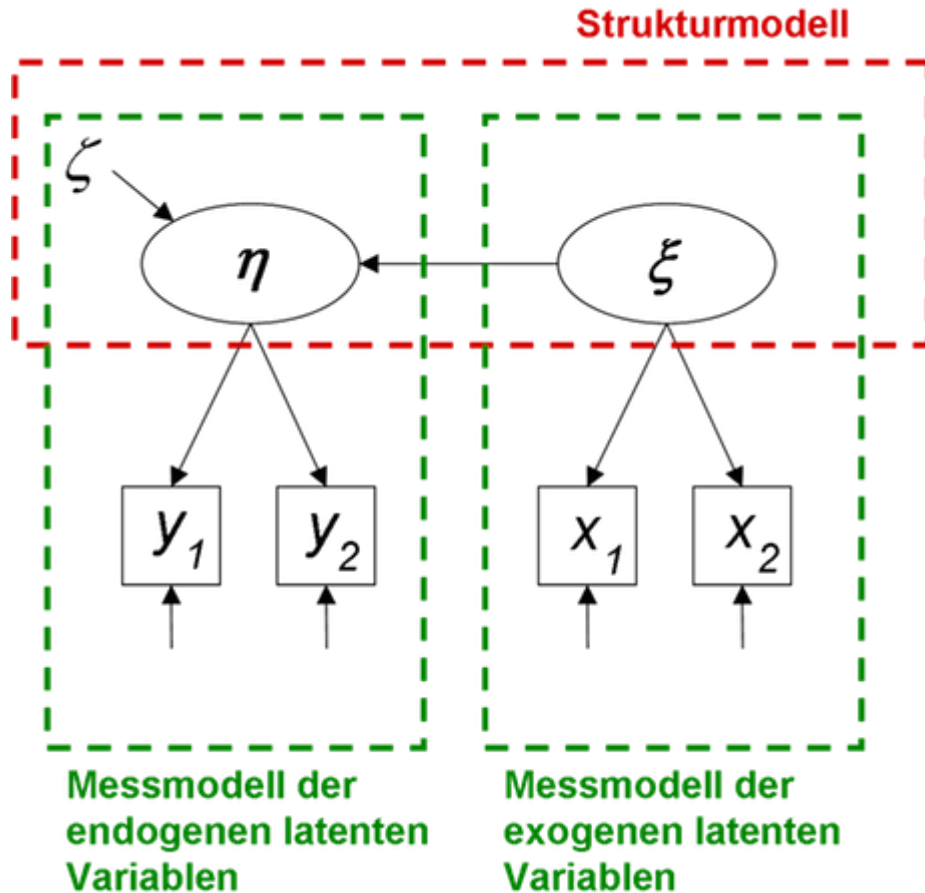
"Generally, ignoring measurement error leads to inconsistent estimators and to inaccurate assessments of the relation between the underlying latent variables." (Bollen, 1989, p. 179)

- Da es praktisch immer Messfehler gibt, sollten wir diese nicht ignorieren
- Die Zusammenhänge zwischen Variablen mit Messfehlern sind zumeist "nach unten" verzerrt
- Da wir es in der Kommunikationswissenschaft mit kleinen Effekten zu tun haben, wollen wir diese gern klar herausstellen

Merke: Der Zusammenhang zwischen den latenten Faktoren ist deutlich stärker als die Korrelation zwischen den einfachen Mittelwerten *derselben* Variablen, da letztere noch immer Messfehler (und itemspezifische Varianz) enthalten.



Berücksichtigung von Messfehler im Strukturgleichungsmodell



- Man kann zwar die sog. Factor-Scores abspeichern, aber auch dort ist der Messfehler enthalten (dazu später mehr!)
- Will man Regressions- oder Pfad-Modelle mit latenten Variablen durchführen, geht das in einem Strukturgleichungsmodell
- Technisch: Man korreliert hier messtheoretisch die **wahren Werte** der Variablen!

Wie geht eigentlich Skalenentwicklung?

Idealtypischer Ablauf I

Schritt 1: Beschäftigung mit Theorie und bestehender Literatur

- Definition des Merkmals
- Entwicklung des theoretischen **Messmodells**

Schritt 2: Formulierung eines (umfangreichen) Itempools

- Umfangreiche Sammlung und Entwicklung von möglichen **Items**, die den Merkmalsraum möglichs komplett abdecken
- Überprüfung/Antizipation der **Ausgewogenheit** bei Merkmalsaspekten und Itemschwierigkeiten
- **Inhaltsvalidität** anhand von Expertenbeurteilung (ggf. Anpassung und Ergänzung, mehrere Runden)

Schritt 3a: Empirische Untersuchung zur Faktorexploration (selten sinnvoll, da wir meisten theoriegeleitet entwickeln)

- **Itemanalyse** (ggf. Ausschluss unpassender Items)
- **Explorative Faktorenanalyse** zur Identifikation von Faktoren (ggf. Ausschluss unpassender Items)

Idealtypischer Ablauf II

Schritt 3b: Empirische Untersuchung zur Überprüfung des theoretischen Modells

- **Itemanalyse (ggf. Ausschluss unpassender Items)
- **Konfirmatorische Überprüfung** des Messmodells (CFA oder IRT, ggf. Ausschluss unpassender Items)
- Entwicklung des besten Messmodells und Beurteilung der **Reliabilität** und **faktorielle Validität**

(ggf. Schritt 3c: Wiederholung von Schritt 3b)

Schritt 4: Erneute Überprüfung des (neuen) Modells

- Erneute Überprüfung des entwickelten Messmodells an einer neuen (repräsentativen) Stichprobe
- Beurteilung der **(Re-Test-)Reliabilität** und der **faktoriellen Validität**
- Beurteilung der **diskriminanten, konvergenten** und **Kriteriumsvalidität**

Erstmal alles genau anschauen

Deskriptive Analysen, Itemschwierigkeiten und Abdeckung des Merkmalraumes

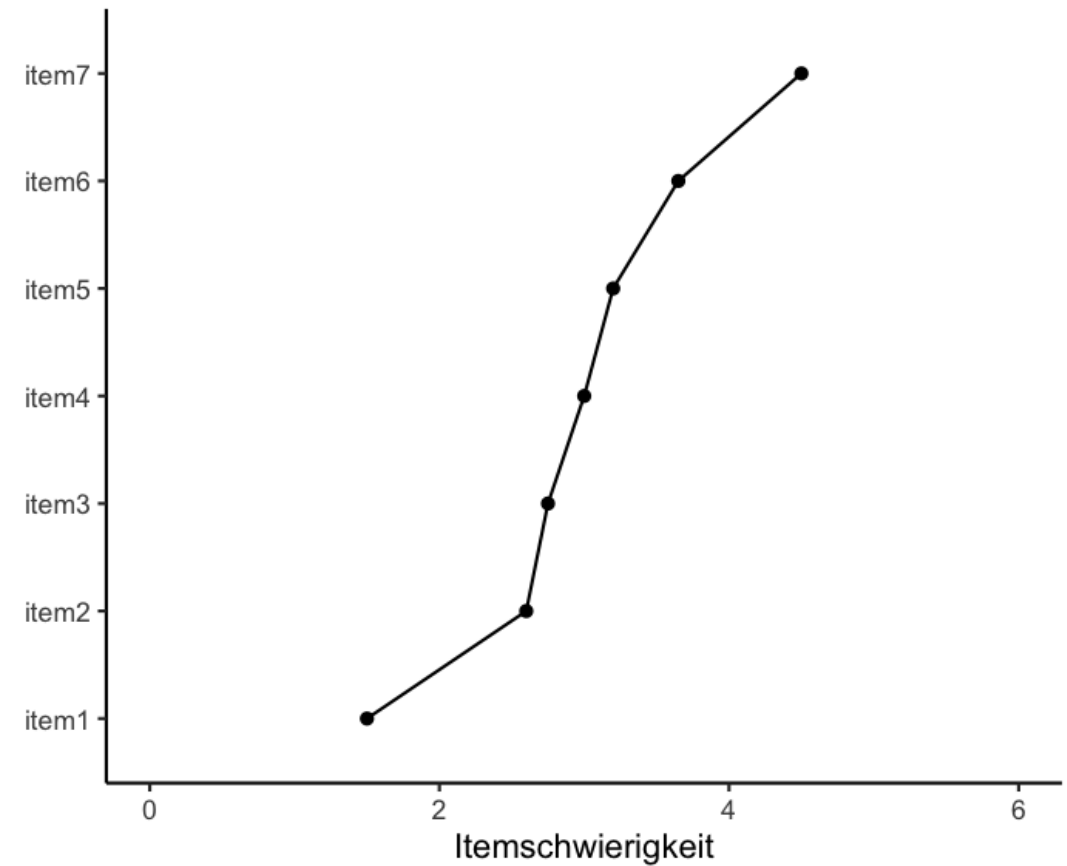
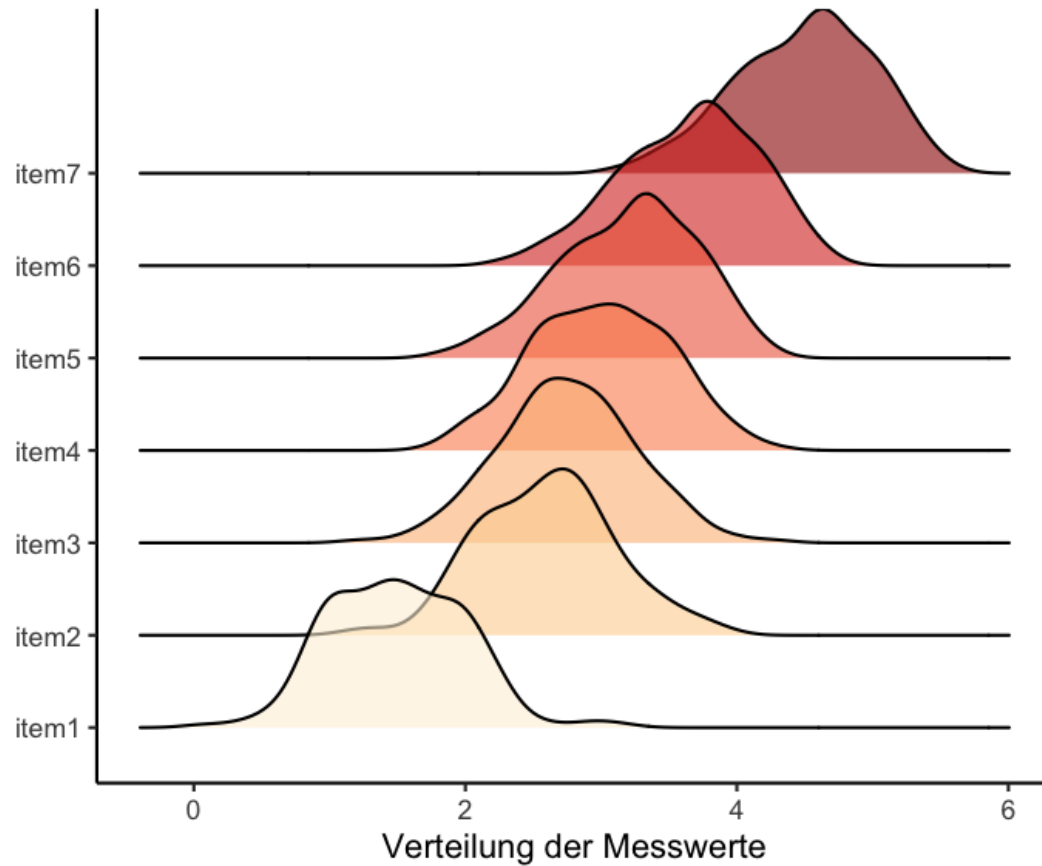
Itemanalyse

- Unabhängig von der Art des Messmodells bzw. der Art der Modellüberprüfung (EFA, CFA, oder IRT) sollte man in einem ersten Schritt ein gutes Verständnis für die Items entwickeln
- Dazu gehört im ersten Schritt, die Beantwortung folgender Fragen:
 - Gibt es Items mit auffallend vielen "Missing values"?
 - Gibt es Boden- oder Deckeneffekte?
 - Gibt es Items mit auffallender Schiefe oder Wölbung?
 - Wie sieht die Verteilung der Itemschwierigkeiten aus?
- In einem zweiten Schritt, fragen wir:
 - Wie stark korrelieren die Items untereinander?

Itemschwierigkeit

- Die **Schwierigkeit** eines Items beschreibt, ob die "meisten" Personen eher hohe (= leicht) oder niedrige Werte (= schwer) im Sinne der Merkmalsausprägung haben
- Bei Wissenstest, die logisch codiert sind (0 = falsche Antwort, 1 = richtige Antwort), ist dies leicht zu verstehen:
 - Wenn die meisten Personen eine Frage richtig beantworten (81%), dann ist die Itemschwierigkeit niedrig (.81)
 - Wenn nur wenige Personen die Antwort auf eine Frage wissen (nur 5%), dann ist die Schwierigkeit hoch (= .05)
- Der Mittelwert einer Variable entspricht der Schwierigkeit des Items
 - Bei nicht-dichotomen Items macht es dennoch Sinn, die Items zu reskalieren (Range: 0-1)
- Bei einer Mehrfachmessung sollte sich die **Schwierigkeit** der Items (bestenfalls jedoch nicht ihre **Diskriminanzfähigkeit**) ausgewogen unterscheiden

Ausgewogene Schwierigkeitsverteilung



Warum brauchen wir Ausgewogenheit?

- Beispiel: Ziel ist es zwischen Menschen zu "diskriminieren", die unterschiedliche Ausprägungen auf einer **Ausländerfeindlichkeitsskala** haben. Die Skala sieht bisher folgendermaßen aus:

Item	Formulierung
1	Wer als Ausländer in Deutschland bleiben will, muss die deutsche Kultur übernehmen.
2	Wenn Arbeitsplätze knapp werden, sollte man die Ausländer wieder in ihre Heimat zurückschicken.
3	Wer als Ausländer in Deutschland bleiben will, muss die deutsche Kultur übernehmen.

Frage: Kann eine solche Skala zwischen beispielsweise "AfD-Wählern" und "Rechtsradikalen" unterscheiden?

Lösung: Schwierigeres Item hinzufügen, wie z. B. "Unter Hitler war nicht alles schlecht."

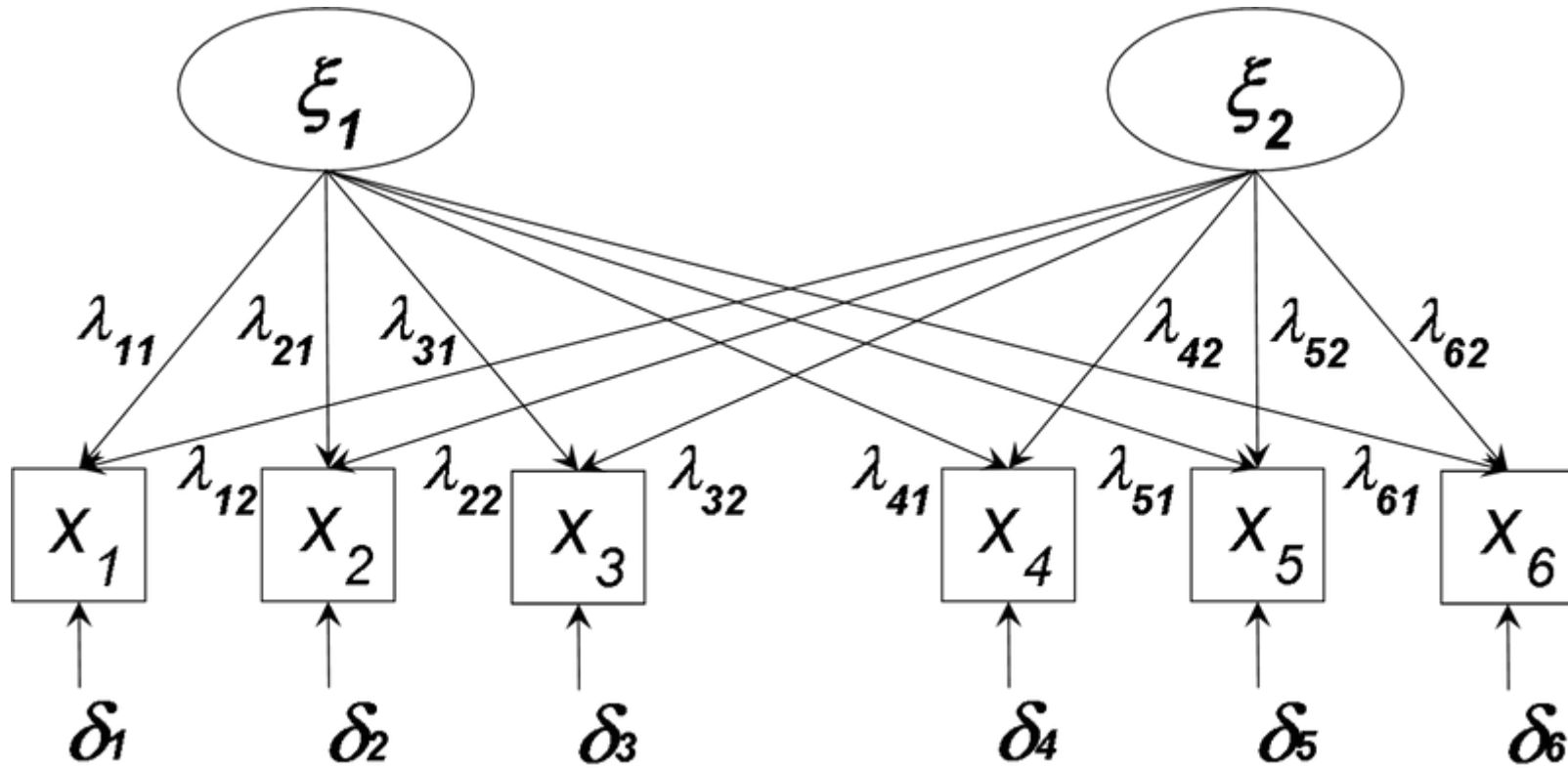
Strukturen entdecken - leicht gemacht?

Explorative Faktorenanalyse

Ziele der explorativen Faktorenanalyse

- **Strukturentdeckendes** Verfahren
- Hauptziel ist es, eine oder mehrere latente Variablen zu identifizieren, die die **gemeinsame Varianz** mehrerer Indikatorvariablen erklären können (sog. Common Factor Model)
- Das Ziel ist damit auch eine **Reduzierung** mehrerer Items auf übergeordnete Faktoren
- Dabei wird die Kovarianz- oder Korrelationsmatrix der Indikatoren so transformiert, dass sich eine Faktorstruktur ergibt
- Die Faktorenlösung soll möglichst eine **Einfachstruktur** haben, d.h. jeder Indikator gehört bestenfalls zu genau einem Faktor
- Praktisch sind jedoch immer **Doppelladungen** vorhanden, die im besten Fall jedoch sehr klein sind

Ladungsstruktur bei der EFA



Arbeitsschritte bei einer EFA

1 Eignungstest

2 Wahl der Extraktionsmethode

3 Bestimmung der Faktorenanzahl

4 Wahl der Rotationsmethode

5 Durchführung der Faktorenanalyse

6 Betrachtung der Ladungen, Reliabilität und Validität

7 Betrachtung der Modellgüte

8 ggf. Modellmodifikation

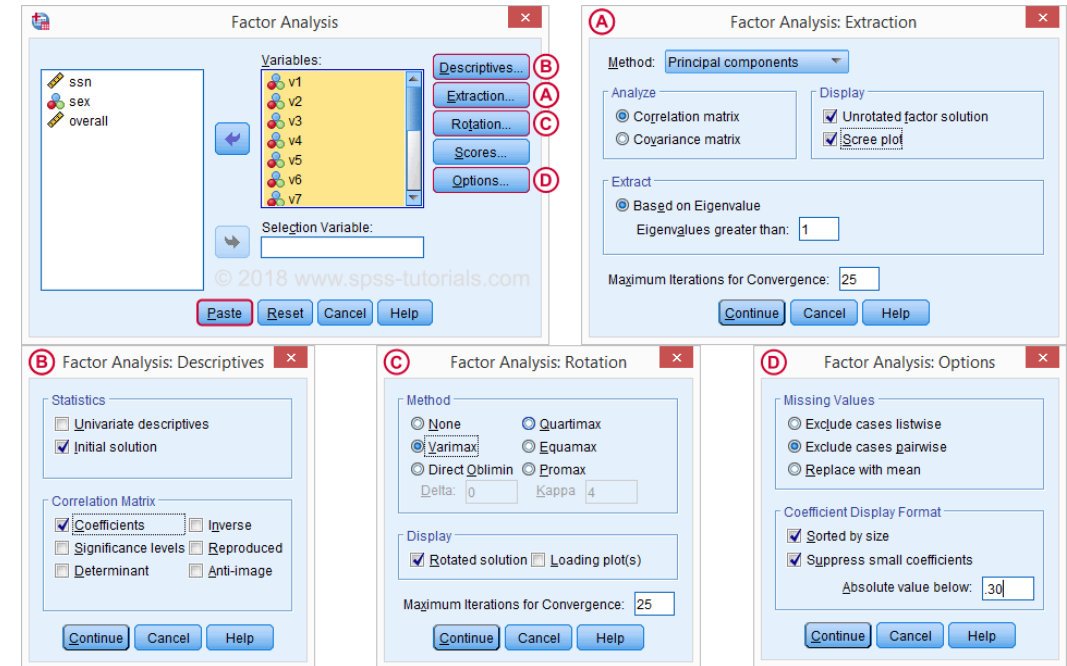
9 ggf. erneute Durchführung von Schritt 5 bis 7

Exkurs: Der Fluch der SPSS-Standard Einstellungen

Viele kennen oder nutzen vielleicht die SPSS-Standard Einstellungen zur "Faktorenanalyse":

- **Hauptkomponentenanalyse** als Extraktionsmethode
- **Kaiser-Kriterium** (d.h. Eigenwerte > 1) als Bestimmungsmerkmal für die Anzahl der zu extrahierenden Faktoren
- orthogonale **Varimax**-Rotation.

Spätestens seit dem bekannten Beitrag von Fabrigar et al. (1999), sollte klar sein, dass dies keine geeignete Vorgehensweise ist!



Wahl der Extraktionsmethode

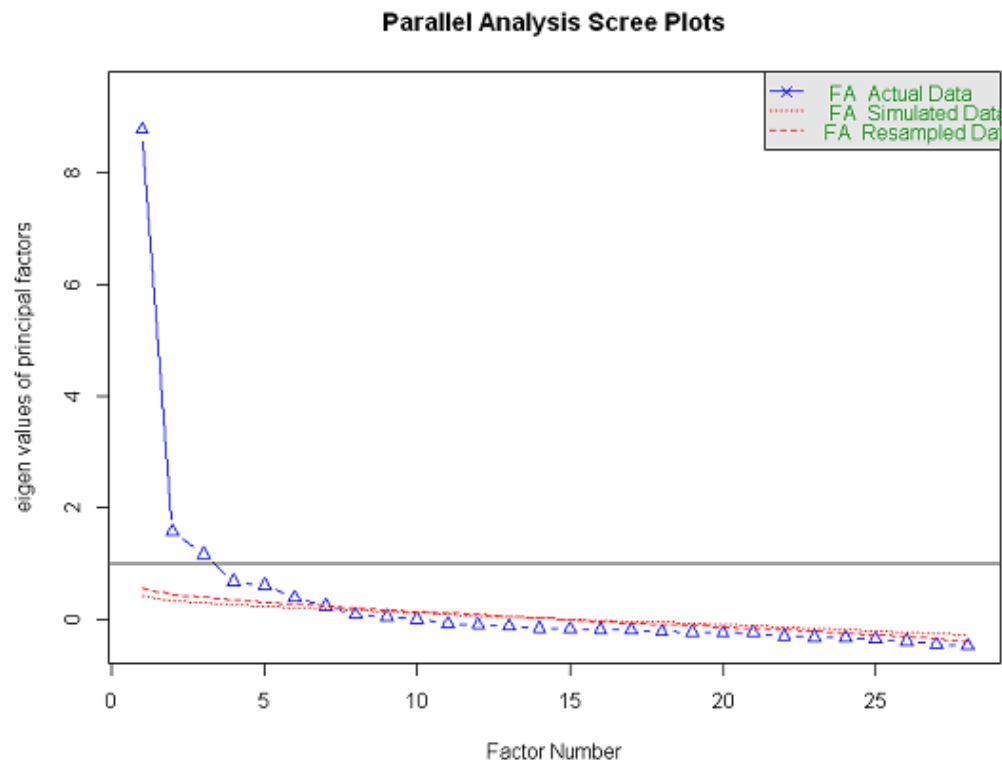
- Die **Hauptkomponentenanalyse** ist problematisch, da sie kein reflektives Messmodell unterstellt und davon ausgeht, dass es ausschließlich gemeinsame Varianz gibt
 - alle Varianz in den beobachteten Variablen durch gemeinsame Faktoren erklärt
 - Wir wissen jedoch, dass die Varianz einer Variable aus mehreren Bestandteilen besteht!
- Die häufigsten Verfahren für *echte* Faktorenanalysen sind:
 - **Hauptachsenanalyse** (Principal Axis Factoring, PAF)
 - **Maximum Likelihood Factor Analysis** (ML FA)
- Hier wird im Sinne der Testtheorie davon ausgegangen, dass nicht alle Varianz der beobachteten Variablen durch zugrundeliegende gemeinsame Faktoren erklärt werden kann

Bestimmung der Faktoranzahl

- Vor der Schätzung der Ladungen und Fehlervarianzen muss geklärt sein, wie viele Faktoren extrahiert werden müssen
- Es gibt kein allgemeingültiges bzw. anerkanntes Kriterium!
- Die Anzahl der Faktoren muss entsprechenden auf verschiedenen Kriterien beruhen
 - **Inhaltliche Plausibilität** (wichtig!)
 - Reproduzierbarkeit der **Korrelationsmatrix** durch die Ladungen der Faktoren
- Es gibt zahlreiche Möglichkeiten, die optimale Anzahl Faktoren zu bestimmen, darunter z.B:
 - Explizite **Theorie** zur Dimensionalität (dann aber eigentlich konfirmatorische Faktorenanalyse!)
 - Analyse der **Eigenwerte** der Faktoren (Kaiser-Kriterium; besser aber Point-of-Inflexion, Parallelanalyse)
 - Very Simple Structure-Kriterium, d.h. eine Einfachstruktur (ohne Doppelladungen) wird mit der empirischen Ladungsstruktur verglichen (häufig sehr andere Ergebnisse als die anderen Verfahren)

Beste Variante: Parallelanalyse

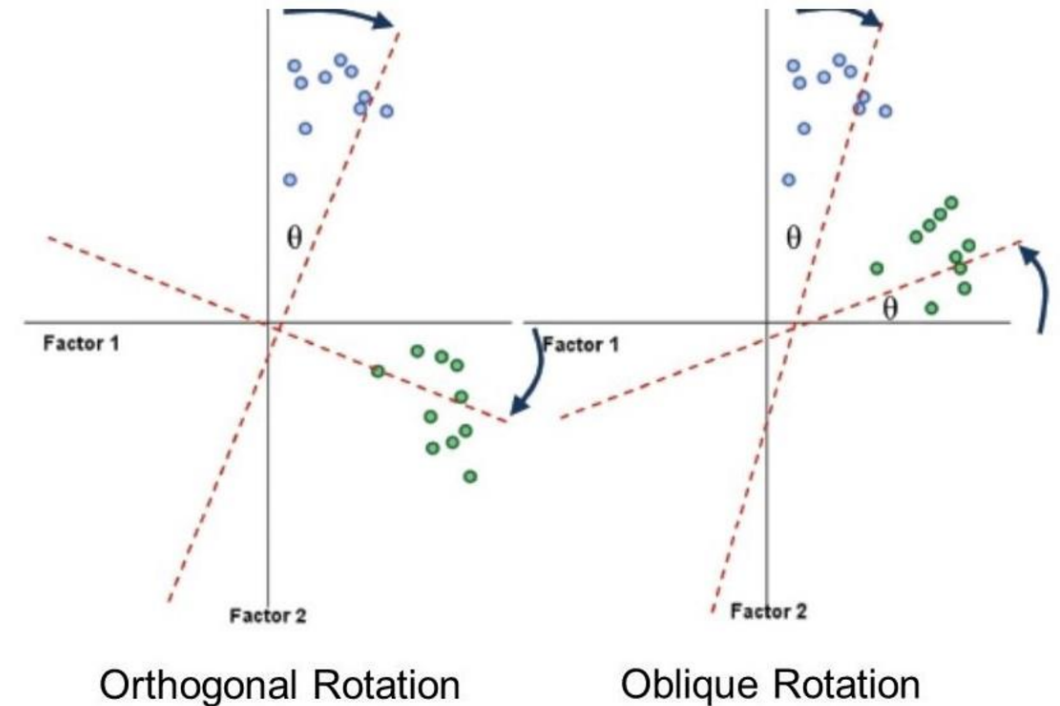
- Nach Horn wird der Eigenwertverlauf mit zufällig **simulierten Werten** verglichen
- Eigenwerte aus Stichprobendaten können auch bei in der Population unkorrelierten Variablen größer Eins werden



- Extrahiert werden Faktoren, deren empirisch beobachtbare Eigenwerte über einem Eigenwertverlauf von Zufallswerten liegen
- Auch das Ergebnis einer Parallelanalyse wird in einem **Screeplot** visualisiert
- In diesem Beispiel würde man auf Basis der Grafik 7 Faktoren differenzieren
- Nach Angaben verschiedener Autoren stellt dieses Verfahren die beste Extraktionsmethode dar

Rotationsmethode

- Um die Interpretierbarkeit der Faktoren zu gewährleisten, werden Faktorstrukturen rotiert (bis möglichst eine Einfachstruktur entsteht)
- Wir unterscheiden dabei zwei Arten der Rotation:
 - **Orthogonale Rotation** (Varimax), die immer zu unkorrelierten Faktoren führt (in den meisten Fällen unrealistisch!)
 - **Oblique Rotation** (Promax, Oblimin), die zu korrelierten Faktoren führt
- Tipp: Immer oblique rotieren!!



Beispiel: Vertrauen

```
fa(efa2, nfactors = 7, fm = "pa", rotate="Promax") %>% print( cut = .3, sort=TRUE)
```

Factor Analysis using method = pa

Call: fa(r = efa2, nfactors = 7, rotate = "Promax", fm = "pa")

Standardized loadings (pattern matrix) based upon correlation matrix

	item	PA2	PA5	PA1	PA7	PA6	PA3	PA4	h2	u2	com
M001_10	10	0.91							0.73	0.27	1.0
M001_09	9	0.88							0.75	0.25	1.0
M001_11	11	0.69							0.63	0.37	1.3
M001_12	12	0.66							0.50	0.50	1.1
M001_07	7	0.48							0.45	0.55	1.9
M002_07	21		0.77						0.64	0.36	1.0
M002_08	22		0.76						0.60	0.40	1.1
M002_06	20		0.73						0.53	0.47	1.2
M002_05	19		0.63						0.59	0.41	1.4
M002_13	27			0.68					0.56	0.44	1.1
M002_14	28			0.67					0.60	0.40	1.2
M001_03	3			0.59					0.35	0.65	1.4
M001_04	4			0.57		0.35			0.58	0.42	2.4
M002_10	24			0.49				0.33	0.62	0.38	1.8
M001_14	14				0.82				0.53	0.47	1.1
M001_01	1				0.67				0.48	0.52	1.1
M001_02	2				0.56				0.52	0.48	1.4
M001_13	13				0.51				0.59	0.41	2.0
M002_02	16								0.20	0.80	2.5
M001_08	8					0.72			0.69	0.31	1.1
M001_05	5					0.70			0.62	0.38	1.2

Kurzer Hinweis

- Auch bei der explorativen Faktorenanalyse wird ein Messmodell geschätzt, dessen Güte anhand der Daten evaluiert werden kann
- Auch wenn dies standardmäßig nicht gemacht wird: es ist sehr sinnvoll den **Model-Fit** zu berichten!
- Dazu später mehr...

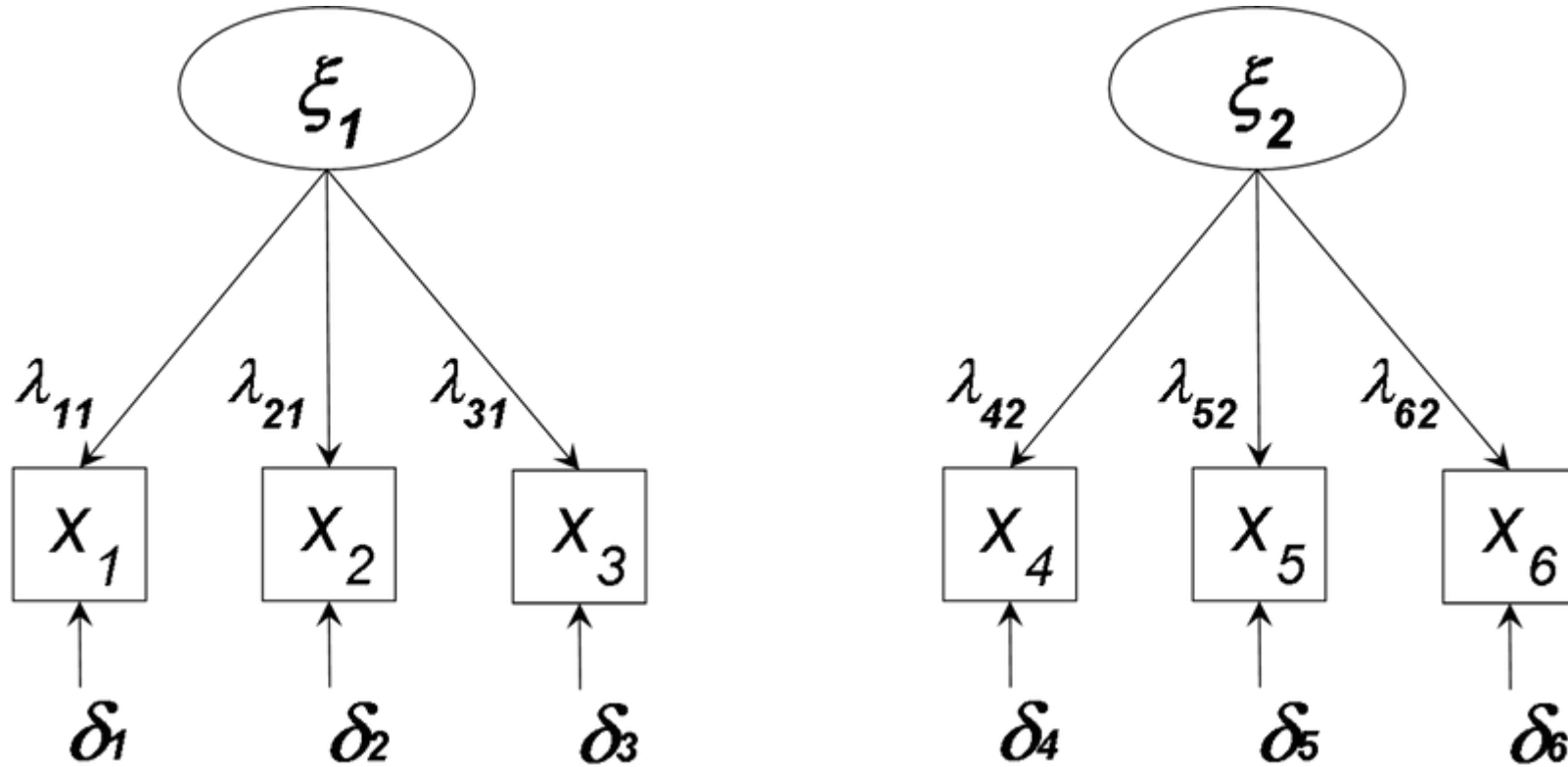
Aber es gibt doch Theorie!

Konfirmatorische Faktorenanalyse

Grundlagen der CFA

- Meistens haben wir konkrete **Vermutungen**, welche Items welchen Faktoren zugeordnet werden sollten
- Hypothesen basieren häufig auf **Theorien**, die ebenso Hinweise geben, wie die uns interessierenden Konstrukte zu operationalisieren sind
- So können wir die **Dimensionalität** unseres Konstruktes oftmals im Vorhinein bestimmen
- Anstatt explorativ vorzugehen und diese Annahmen zu ignorieren, können wir das hypothetische Messmodell direkt **überprüfen**

Ladungsstruktur der CFA (restringiert)



Arbeitsschritte bei einer CFA

1. Modellspezifikation auf Basis der Theorie
2. Modellidentifikation und -schätzung
3. Betrachtung der Ladungen
4. Betrachtung der Modellgüte
5. ggf. Modellmodifikation (z. B. über Modifikationsindices)
6. ggf. erneute Durchführung von Schritt 1 bis 5

Kurzes Beispiel

```
library(lavaan)

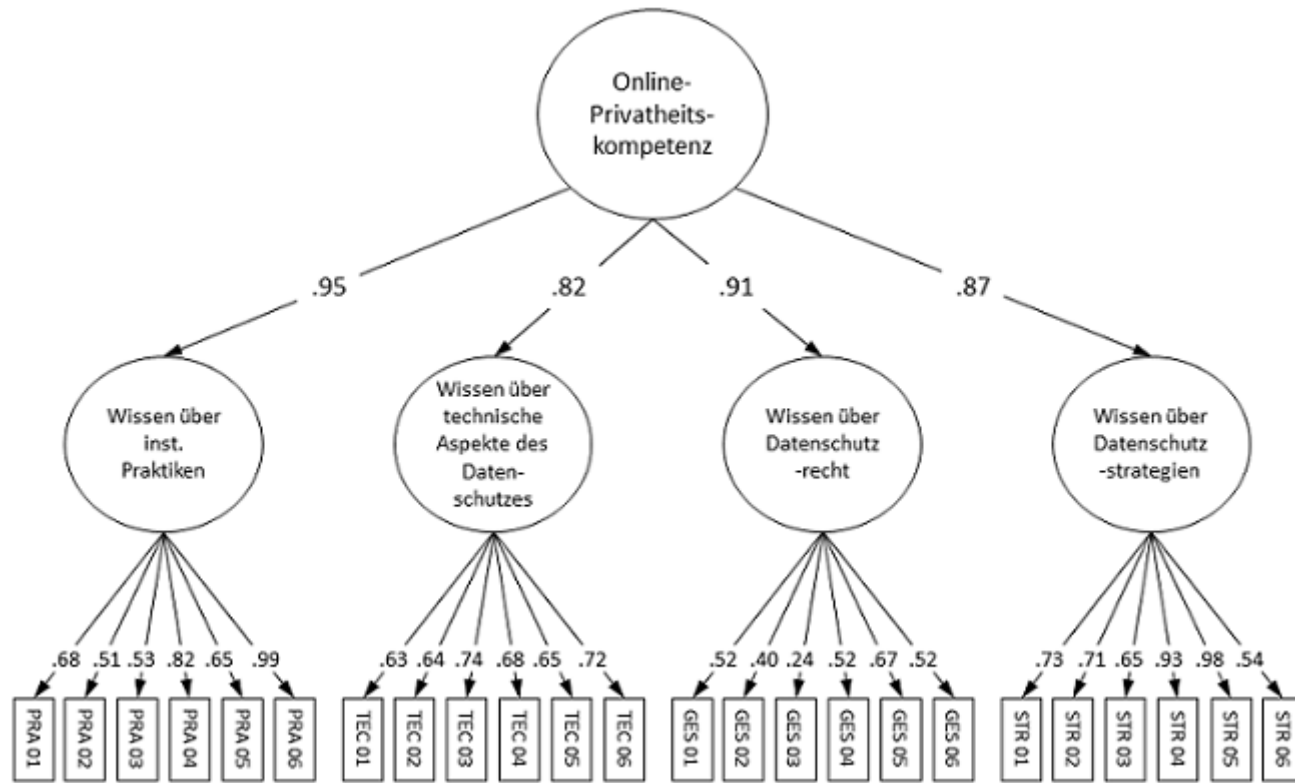
trust_model <- "
  trust =~ VT01_01 + VT01_02 + VT01_03 + VT01_04
"

trust_fit <- cfa(trust_model, data = efa)

parameterEstimates(trust_fit, standardized = T) %>%
  subset(op == "=~") %>%
  select(-ci.lower, -ci.upper, -std.lv, -std.noxx)
```

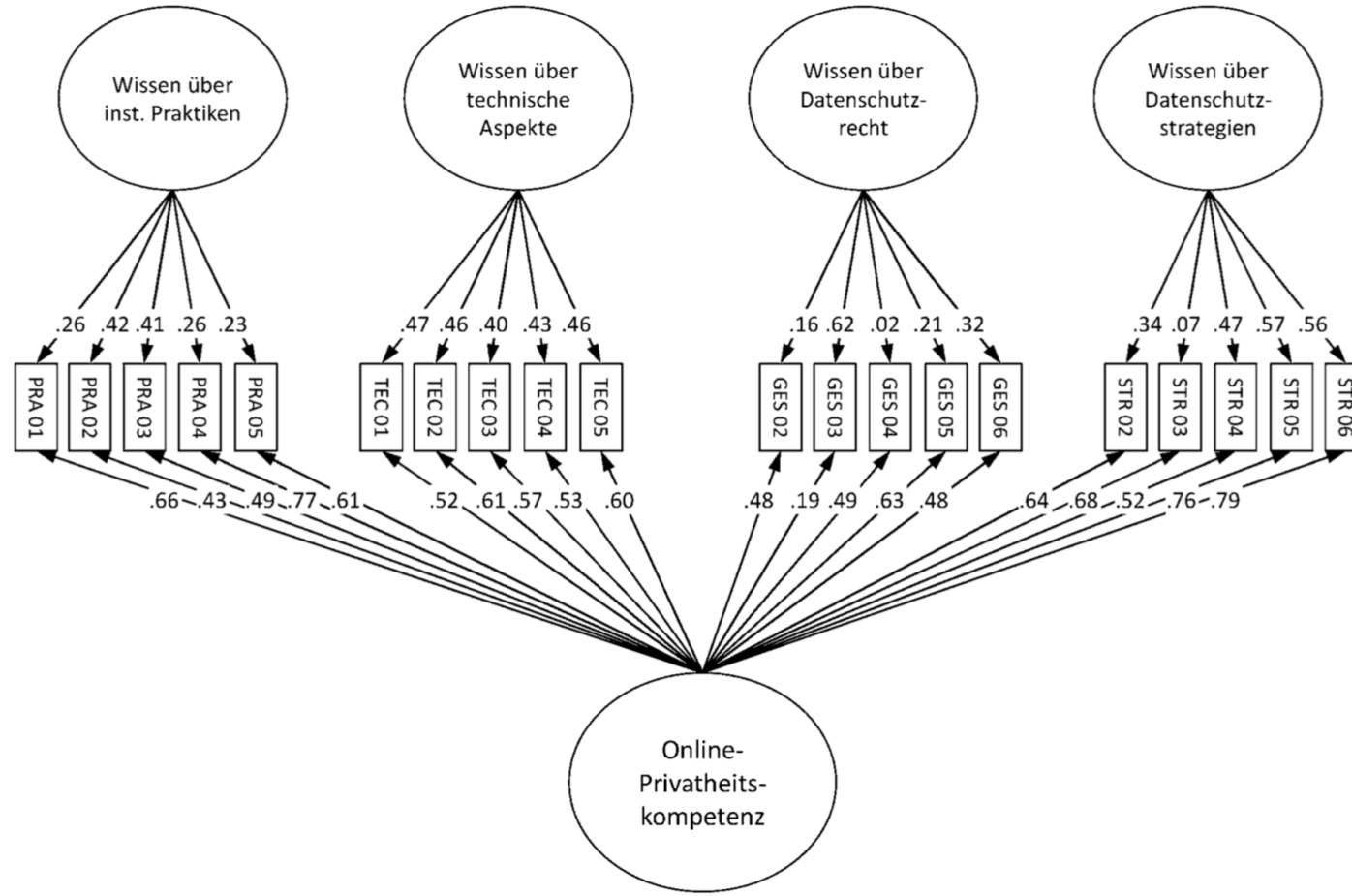
	lhs	op	rhs	est	se	z	pvalue	std.all
1	trust	=~	VT01_01	1.000	0.000	NA	NA	0.777
2	trust	=~	VT01_02	0.592	0.065	9.058	0	0.562
3	trust	=~	VT01_03	0.758	0.086	8.809	0	0.545
4	trust	=~	VT01_04	0.997	0.089	11.180	0	0.792

Beispiel: Second-Order-Modell



Quelle: Masur, Deutsch & Trepte, 2017

Beispiel: Bifaktor-Modell (Reise, 2012)



Alles eine Frage der Güte, oder?

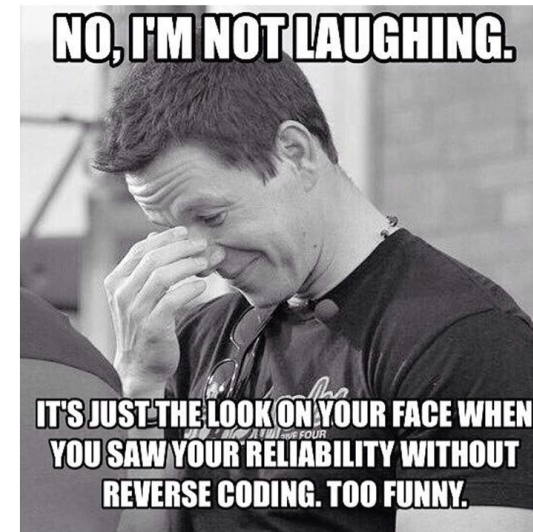
Reliabilität, Validität, und weitere Gütekriterien...

Unterschiedliche Arten von Reliabilität

- Wie bereits kennengelernt, kann man die **Reliabilität** folgendermaßen bestimmen:
 - Paralleltest-Reliabilität
 - Split-Half-Reliabilität
 - Retest-Reliabilität
- Im Falle von Faktorenmodellen und einmaliger Erfassung, wird aber häufiger die **Interne Konsistenz** berechnet:
 - Maß dafür, wie stark die Items einer Skala miteinander *zusammenhängen*.
 - Interne Konsistenz stellt gewissermaßen einen Umweg dar, die Messgenauigkeit eines Instruments zu erheben, wenn kein Retest oder Paralleltest zur Reliabilitätsbestimmung zur Verfügung steht
 - Es erfolgt die Reliabilitätsmessung also intern, wobei jedes Item gewissermaßen als *Paralleltest* behandelt und mit jedem anderen Item korreliert wird (Interkorrelationsmatrix)

Cronbach's Alpha, McDonald's Omega & AVE

- Standard: Cronbach's α
 - neuere Bezeichnung ist "tau-equivalent reliability"
 - Maßzahl für die interne Konsistenz einer Skala
 - denkbar als "durchschnittliche Korrelation" zwischen den Items
 - gilt streng genommen nur für τ -äquivalente Modelle (scheitert meistens an der Realität!)
- Alternative: McDonald's ω
 - auch "composite reliability" oder "congeneric reliability" genannt
 - gilt für τ -kongenerische Modelle
- Alternative: Average Variance Extracted (AVE)
 - Anteil der Varianz, der durch das latente Konstrukt in den Variablen (im Gegensatz zur Messfehlervarianz) erklärt wird
 - Kann prozentual interpretiert werden



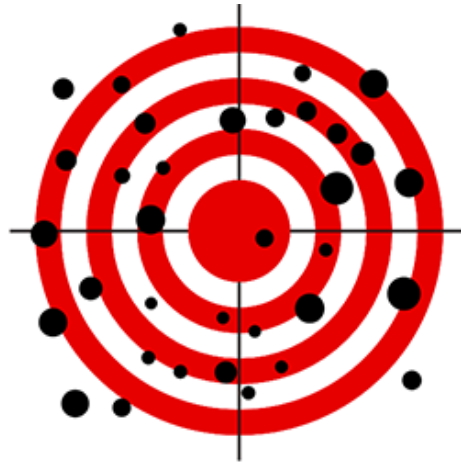
Unterschiedliche Arten von Validität

- **Inhaltsvalidität:** Betrifft die Frage, ob Items wirklich dazu geeignet sind ein bestimmtes Konstrukt zu erfassen (letztlich nur durch Expertenurteile überprüfbar; Interrater-Übereinstimmungsmaße mögliche Quantifizierung)
- **Konstruktvalidität:** Intersubjektiv (empirisch-quantitativ) nachprüfbare Hinweise darauf, dass tatsächlich das relevante Konstrukt gemessen wird und kein anderes
 - *Interne Struktur/Faktorielle Validität*
 - *Diskriminante und Konvergente Validität*
- **Kriteriumsvalidität:** Wie gut lassen sich Ergebnisse anderer Tests oder Verhaltensweisen durch das Testergebnis vorhersagen?

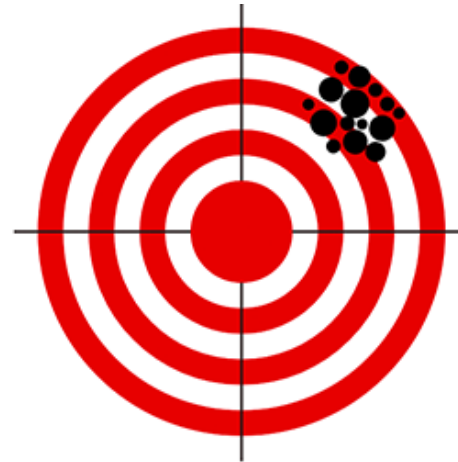
Der Einfluss von Reliabilität und Validität



Unreliable & Unvalid



Unreliable, But Valid



Reliable, Not Valid



Both Reliable & Valid

Beurteilung der Modell-Güte

Ein zentrales Ziel der Strukturgleichungsmodellierung (also auch EFA und CFA!) ist die Evaluation des postulierten Kausalmodells:

- Lässt sich das **a-priori formulierte Hypothesensystem** anhand der empirisch erhobenen Daten bestätigen?
- Auch einem Messmodell liegt ein Hypothesensystem zugrunde!

Nach Jöreskog & Sörbom (1993) können Modellevaluationen drei unterschiedliche Intentionen zu Grunde liegen:

1. Evaluation des Gesamtmodells
2. Vergleichende Evaluation alternativer Modelle
3. Modifikation der Modellstruktur

Arten von Prüfungen

- Plausibilitätsprüfung der **Parameterschätzungen**
 - Die Schätzung gilt allgemein als unplausibel, wenn negative Varianzen, Kommunalitäten > 1 oder Korrelationen > 1 auftreten
 - Die Parametermatrizen sind dann nicht positiv definiert, wodurch einzelnen Gütekriterien nicht berechnet werden können
- Prüfung der **Gesamtgüte**/des **Model-Fits**
 - Eine hohe Güte ist dann gegeben, wenn die mit Hilfe der Parameterschätzer berechneten Varianzen und Kovarianzen mit den empirischen übereinstimmen
 - In der Literatur gibt es dafür eine Vielzahl an Fit-Indizes, mit deren Hilfe man die Passung eines Modells (auch Model-Fit genannt) beurteilen kann

Fit-Indizes

- Es gibt unzählige **Fit-Indizes**, die im Laufe der Beschäftigung mit Pfadanalysen und Strukturgleichungsmodellen entwickelt wurden
- Für jeden Index gibt es Daumenregeln (z.B. Hu & Bentler, 1999; Hair et al., 2012), an denen man die Güte eines Modells evaluieren kann
- Ist gilt jedoch: Jedes Modell ist anders und für jedes Modell gelten damit andere Maßstäbe!
- Am häufigsten werden die folgenden Indizes berichtet:
 - Chi-Quadrat-Test
 - Root Mean Square Error of Approximation (RMSEA)
 - Tucker-Lewis-Index (TLI)
 - Comparative Fit Index (CFI)
 - und viele mehr...

Chi-Quadrat-Test

- Es wird getestet, inwiefern sich das postulierte Modell von den Daten unterscheidet
 - $H_0: S = \Sigma$ (d.h. empirische und modelltheoretische Kovarianzmatrizen sind gleich)
 - $H_1: S \neq \Sigma$ (d.h. empirische Varianz-Kovarianz-Matrix entspricht einer beliebig positiv definierten Matrix)
- Der p -Wert gibt hier die **Ablehnungswahrscheinlichkeit der Fehlhypothese** an und sollte deswegen nicht signifikant sein

Problem: Für Modelle mit einer Stichprobe größer als 400 wird der χ^2 -Test fast immer signifikant (Kenny, 2014), da die Power für den Test steigt (genauso wie man bei großen Stichproben leichter signifikante Effekte findet!)

Root Mean Square Error of Approximation (RMSEA)

- Ist ein absolutes Maß für die Güte des Modells
- $< .08$ sein (Hu & Bentler, 1999), bei größeren Stichproben $< .07$ (Hair et al., 2012)
- Wird standardmäßig mit 90%-Konfidenzintervallen angegeben (fragt mich nicht warum 90% anstatt 95%...)

Merke: Der χ^2 -Test und der RMSEA überprüfen wie gut die modelltheoretische Kovarianzmatrix Σ eine strenge Funktion allein der Modellparameter darstellt (**Badness-of-fit**).

Dies ist meistens unrealistisch, weswegen zusätzlich auch **deskriptive Gütekriterien** angeschaut werden, die versuchen, Hinweise darauf zu geben, ob eine Differenz zwischen S und Σ in der Anwendungspraxis vernachlässigt werden kann.

Goodness-of-Fit-Indizes

Tucker Lewis Index (TLI)

- Basiert auf den Verhältnissen zwischen Freiheitsgraden und χ^2 -Wert des Nullmodells und des gefitteten Modells
- Bei kleinen Stichproben: $> .97$, bei großen Stichproben: $> .95$ (Hair et al., 2012)

Comparative Fit Index (CFI)

- Basiert auf der Differenz zwischen χ^2 -Wert und den Freiheitsgraden des Nullmodells und des gefitteten Modells
- Kann nur Werte zwischen 0 und 1 annehmen, 1 entspricht einer perfekten Passung
- Bei kleinen Stichproben: $> .97$, bei großen Stichproben: $> .95$ (Hair et al., 2012)

Daumenregeln zur Modelüberprüfung (Hair et al., 2012)

TABLE 4 Characteristics of Different Fit Indices Demonstrating Goodness-of-Fit Across Different Model Situations

No. of Stat. vars. (<i>m</i>)	<i>N</i> < 250			<i>N</i> > 250		
	<i>m</i> ≤ 12	12 < <i>m</i> < 30	<i>m</i> ≥ 30	<i>m</i> < 12	12 < <i>m</i> < 30	<i>m</i> ≥ 30
χ^2	Insignificant <i>p</i> -values expected	Significant <i>p</i> -values even with good fit	Significant <i>p</i> -values expected	Insignificant <i>p</i> -values even with good fit	Significant <i>p</i> -values expected	Significant <i>p</i> -values expected
CFI or TLI	.97 or better	.95 or better	Above .92	.95 or better	Above .92	Above .90
RNI	May not diagnose misspecification well	.95 or better	Above .92	.95 or better, not used with <i>N</i> > 1,000	Above .92, not used with <i>N</i> > 1,000	Above .90, not used with <i>N</i> > 1,000
SRMR	Biased upward, use other indices	.08 or less (with CFI of .95 or higher)	Less than .09 (with CFI above .92)	Biased upward; use other indices	.08 or less (with CFI above .92)	.08 or less (with CFI above .92)
RMSEA	Values < .08 with CFI = .97 or higher	Values < .08 with CFI of .95 or higher	Values < .08 with CFI above .92	Values < .07 with CFI of .97 or higher	Values < .07 with CFI of .92 or higher	Values < .07 with CFI of .90 or higher

Note: *m* = number of observed variables; *N* applies to number of observations per group when applying CFA to multiple groups at the same time.

Beispiel: Vertrauen

```
pmstats::fit_table(trust_fit, rmsea_ci = TRUE, print = TRUE)
```

```
# A tibble: 1 x 9
  chisq    df pvalue cfi    tli  rmsea rmsea.ci.lower rmsea.ci.upper srmr
<chr> <dbl> <chr>  <chr> <chr> <chr> <chr>          <chr>      <chr>
1 29.28     2 < .001 .92    .77  .20   .14           .27        .06
```

```
semTools::reliability(trust_fit)
```

```
      trust    total
alpha 0.7649690 0.7649690
omega 0.7690536 0.7690536
omega2 0.7690536 0.7690536
omega3 0.7600089 0.7600089
avevar 0.4645915 0.4645915
```

- Weder der Modellpassung, noch die Reliabilität ist optimal

Bespiel Vertrauen

```
modificationindices(trust_fit) %>%  
  arrange(desc(mi)) %>% mutate_if(is.numeric, round, 2)
```

	lhs	op	rhs	mi	epc	sepc.lv	sepc.all	sepc.nox
1	VT01_02	~~	VT01_03	28.95	0.16	0.16	0.34	0.34
2	VT01_01	~~	VT01_04	28.95	0.36	0.36	1.24	1.24
3	VT01_03	~~	VT01_04	9.04	-0.13	-0.13	-0.30	-0.30
4	VT01_01	~~	VT01_02	9.04	-0.10	-0.10	-0.30	-0.30
5	VT01_01	~~	VT01_03	2.03	-0.06	-0.06	-0.14	-0.14
6	VT01_02	~~	VT01_04	2.03	-0.05	-0.05	-0.15	-0.15

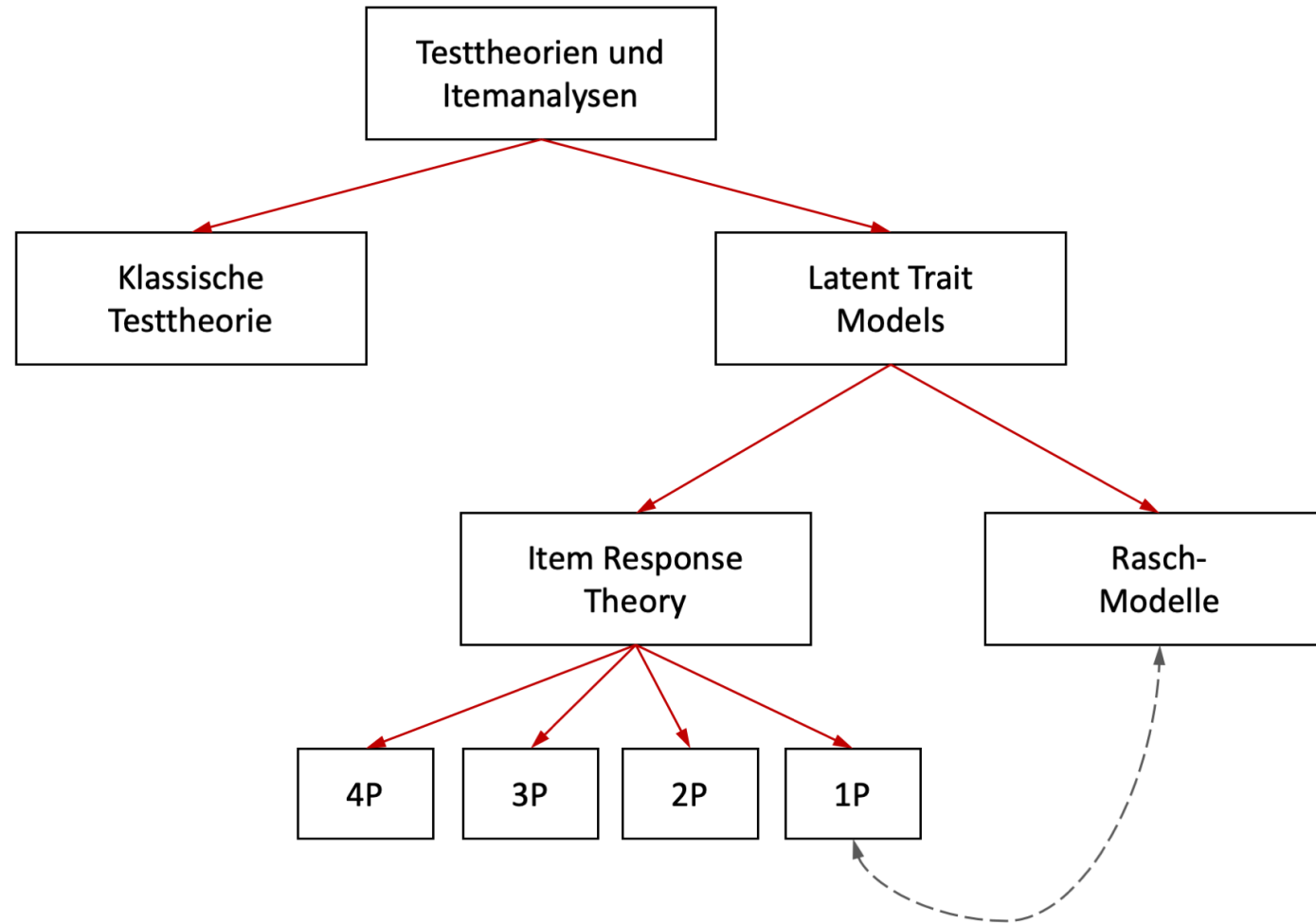
```
trust_model2 <- "  
  trust1 =~ VT01_01 + VT01_04  
  trust2 =~ VT01_02 + VT01_03  
"  
trust_fit2 <- cfa(trust_model2, data = efa)  
pmstats::fit_table(trust_fit2, print = TRUE)
```

```
# A tibble: 1 x 7  
  chisq    df pvalue cfi    tli    rmsea srmr  
  <chr> <dbl> <chr>  <chr> <chr> <chr> <chr>  
1 1.07      1 .300    > .99 > .99 .01    .01
```

Probabili-was?

Item Reponse Theory, Rasch-Modelle und probabilistische Testtheorie

Übersicht

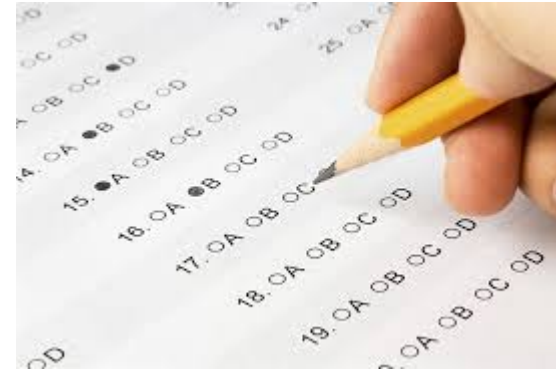


Nachteile klassischer Testtheorie

- Stichprobenabhängigkeit
 - Kennwerte sind immer abhängig von der gewählten Stichprobe
 - Erschwert die Generalisierbarkeit eines Tests (Retests notwendig!)
- Ausschließlich Evaluation der Item- und Gesamtgüte
 - Passung von Personen zum Modell wird nicht berücksichtigt
- Intervallskalen vorausgesetzt
- Eingeschränkter Fokus und Aufmerksamkeit bei der Skalenentwicklung
 - Schwierigkeit spielt untergeordnete Rolle
 - Dadurch häufig sehr homogene Schwierigkeitsgrade (mittlerer Schwierigkeitsgrad)
 - Kann sehr problematisch für eine gute Diagnostik sein

Was ist nochmal Testtheorie (allgemein?)

- Das Zuordnen von Beobachtungen zu internen Traits
 - Test-Scores im Fragenbogen -> Intelligenz
 - Richtigkeit der Antworten -> Wissen
 - Antworten auf Frageitems -> Einstellung
- Oftmals versuchen wir auch diskrete (binäre) Antworten, unbeobachteten Traits zuzuordnen, die eigentlich kontinuierlich sind
 - Populäre Methode: Aufsummieren zum Summenscore



Beispiel: Wahrgenommenes verfügbares Vermögen

Latenter Trait: Wir wollen wissen, wie "reich" sich Personen fühlen

Mögliche Fragemodalität: "Wenn ich möchte, könnte ich mir im nächsten Monat wahrscheinlich das Folgende leisten:"

- Eine Tasse Kaffee



Beispiel: Wahrgenommenes verfügbares Vermögen

Latenter Trait: Wir wollen wissen, wie "reich" sich Personen fühlen

Mögliche Fragemodalität: "Wenn ich möchte, könnte ich mir im nächsten Monat wahrscheinlich das Folgende leisten:"

- Eine Tasse Kaffee
- 10 Euro sparen



Beispiel: Wahrgenommenes verfügbares Vermögen

Latenter Trait: Wir wollen wissen, wie "reich" sich Personen fühlen

Mögliche Fragemodalität: "Wenn ich möchte, könnte ich mir im nächsten Monat wahrscheinlich das Folgende leisten:"

- Eine Tasse Kaffee
- 10 Euro sparen
- Eine Jacke kaufen



Beispiel: Wahrgenommenes verfügbares Vermögen

Latenter Trait: Wir wollen wissen, wie "reich" sich Personen fühlen

Mögliche Fragemodalität: "Wenn ich möchte, könnte ich mir im nächsten Monat wahrscheinlich das Folgende leisten:"

- Eine Tasse Kaffee
- 10 Euro sparen
- Eine Jacke kaufen
- Ein iPhone kaufen



Beispiel: Wahrgenommenes verfügbares Vermögen

Latenter Trait: Wir wollen wissen, wie "reich" sich Personen fühlen

Mögliche Fragemodalität: "Wenn ich möchte, könnte ich mir im nächsten Monat wahrscheinlich das Folgende leisten:"

- Eine Tasse Kaffee
- 10 Euro sparen
- Eine Jacke kaufen
- Ein iPhone kaufen
- Sich einen Privatjet zulegen



Personen und Items auf derselben Skala

Personen



Items



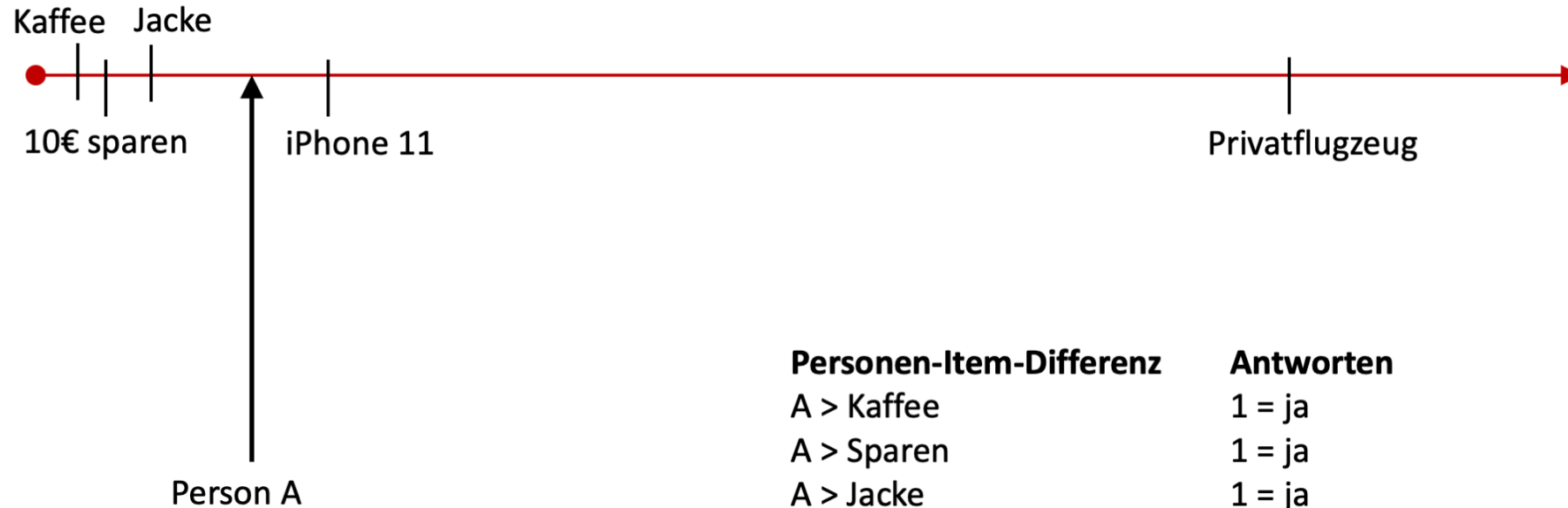
Latenter Trait



Zuordnen von binären Antworten auf die Skala

- Manche **Items** benötigen mehr Vermögen
 - Jacke vs. Privatflugzeug
- Manche **Personen** sind reicher als andere
 - Normaler Mensch vs. Kim Kardashian
 - Wenn Vermögen einer Person > Kosten eines Items = Positive Antwort
- Position der positiven Antwort sagt uns etwas darüber, wo die Person auf dem **latenten Trait** liegt
 - keine positive Antwort = Sehr wenig Vermögen
 - nur positive Antworten = Sehr großes Vermögen

Antwortmuster einer Person



Personen-Item-Differenz	Antworten
$A > \text{Kaffee}$	1 = ja
$A > \text{Sparen}$	1 = ja
$A > \text{Jacke}$	1 = ja
$A < \text{iphone}$	0 = nein
$A < \text{Privatflugzeug}$	0 = nein

Probabilistisches Zuordnen

- Das Zuordnen über und innerhalb von Individuen wird jedoch **nicht komplett konsistent** sein
 - Menschen schätzen die Kosten unterschiedlich ein
 - Nicht jeder weiß genau, wieviel er oder sie wirklich zur Verfügung hat
 - Das wahrgenommene Vermögen verändert sich über die Zeit
 - ...
- Das Zuordnen ist **probabilistisch**, d.h. es gibt nur eine Wahrscheinlichkeit und Messfehler
 - z. B: es ist wahrscheinlich, dass eine reiche Person sich ein Flugzeug leisten kann, aber sicher sind wir uns nicht

Probabilistisches Zuordnen



	Person B	Person A	Ingesamt
$\Pr(\text{Kaffee} = 1)$	0.65	0.95	0.80
$\Pr(\text{Sparen} = 1)$	0.45	0.75	0.60
$\Pr(\text{Jacke} = 1)$	0.40	0.70	0.55
$\Pr(\text{iPhone} = 1)$	0.15	0.45	0.30
$\Pr(\text{Flugzeug} = 1)$	0.00	0.00	0.00

Transformieren der Wahrscheinlichkeit

- Auch wenn wir Wahrscheinlichkeiten (halbwegs) interpretieren können, sind sie problematisch
- Sie sind grundsätzlich schwer zu modellieren
 - Problem: Range zwischen 0 und 1
 - Reminder: Logistische Regression!
- Besser wäre z. B. eine Range zwischen $-\infty$ und $+\infty$
- Eine Lösung ist die Transformation über den **natürlichen Logarithmus** der "Odds":
 - $\text{Logit} = \ln(\text{Pr} / (1-\text{Pr}))$
 - z. B.: $0 = \ln(0.5 / (1-0.5))$

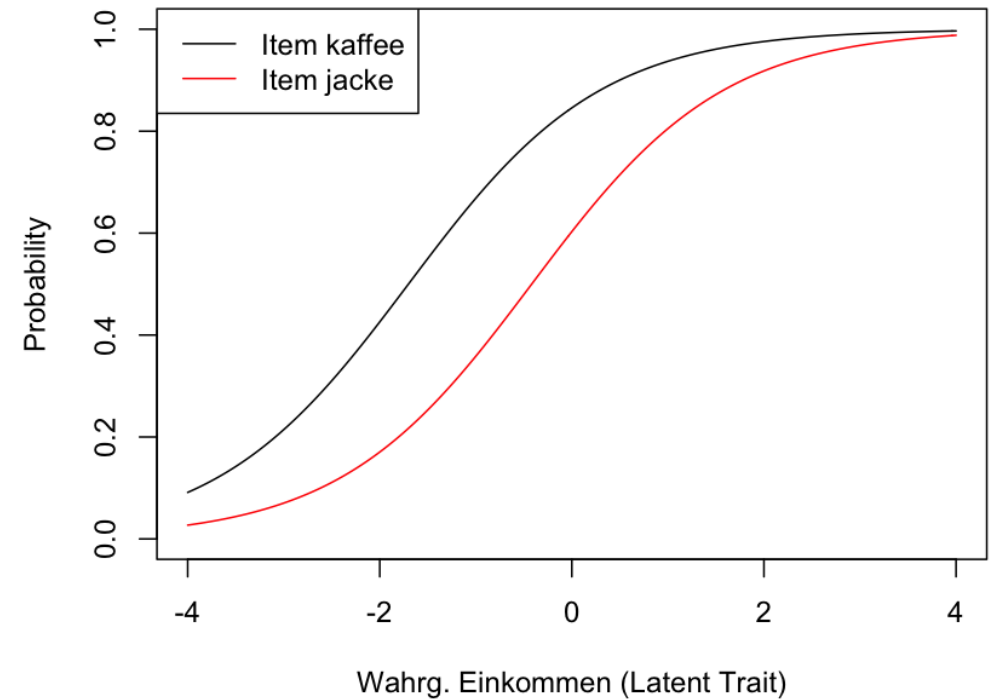
Item Characteristic Curves

```
library(eRm)

# Itemselektion
rdat <- raschdat1 %>%
  select(flugzeug = I14,
         kaffee = I1,
         sparen = I23,
         jacke = I12,
         phone = I25)

# Schätzen des Modells
rasch <- RM(rdat)

# Plott ICC curves
plotjointICC(rasch,
             item.subset = c(2,4),
             xlab = "Wahrg. Einkommen (Latent Trait)",
             ylab = "Probability",
             main = "")
```



Statistisches Modell

- Die Grundidee ist des dahinterliegenden Modells ist verhältnismäßig simpel:

$$\textit{Logit}_{item=1} = \textit{Vermögen}_{person} - \textit{Kosten}_{item}$$

- Dieses Modell heißt auch **Rasch-Modell** oder **1PL-Modell**. Es lässt sich folgendermaßen formalisieren:

$$Y_{ij} = \theta_j - b_i$$

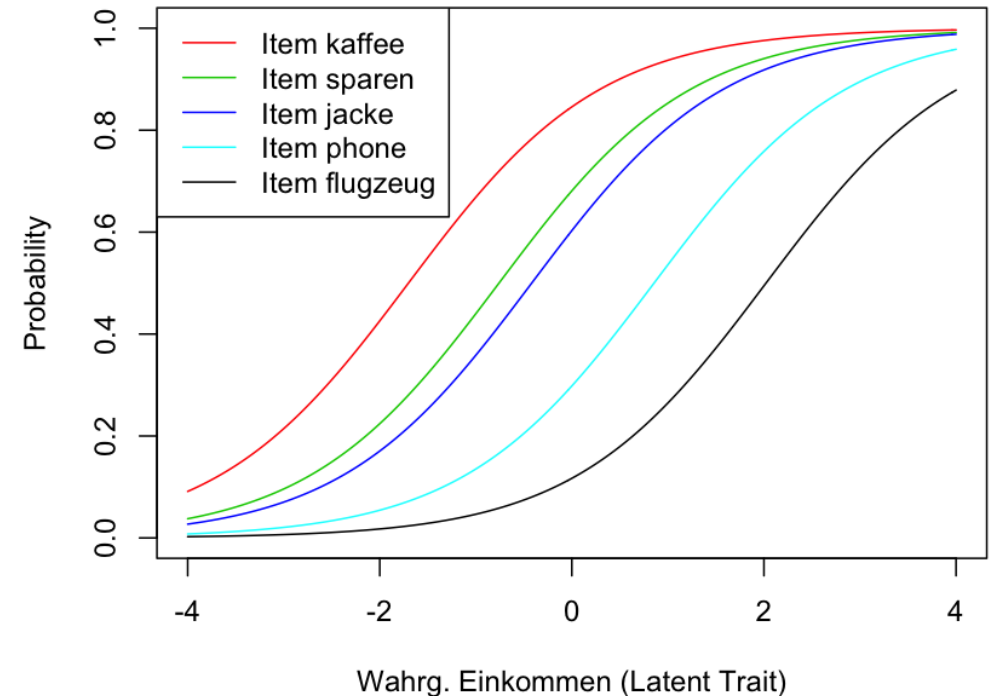
- Y_i = Logit, dass Item i von Person j positiv beantwortet wird
- θ_j = Latente Fähigkeit von Person j
- b_i = Schwierigkeit des Items i

Merke: Wir haben es im Grunde mit einem essentiell τ -äquivalenten Modell zu tun! Außer in ihrer Schwierigkeit unterscheiden sich die Items nicht!

Items "informieren" über unterschiedliche Trait-Level

```
plotjointICC(rasch,  
             item.subset = (1:5),  
             xlab = "Wahrg. Einkommen (Latent Trait)",  
             ylab = "Probability",  
             main = "")
```

	Person B	Person A	Ingesamt
Pr(Kaffee = 1)	0.65	0.95	0.80
Pr(Sparen = 1)	0.45	0.75	0.60
Pr(Jacke = 1)	0.40	0.70	0.55
Pr(iPhone = 1)	0.15	0.45	0.30
Pr(Flugzeug = 1)	0.00	0.00	0.00



Das Rasch-Modell

- In der Diagnostik haben IRT-Modelle längst die klassische Testtheorie abgelöst
- Gerade das **Rasch-Modell** hat einige vorteilhafte Eigenschaften
 - Spezifische Objektivität
 - Jede Person sollte zwei unterschiedliche Items gleichartig einordnen
 - Jedes Item sollte unterschiedliche Personen gleichartig einordnen
 - Einfachheit der Berechnung
 - bei einem passenden Modell ist der einfache Summenscore ausreichend (!)
 - die einfache Anzahl an positiver Items sagt uns etwas über den latenten Trait
 - Stichprobenunabhängigkeit

Gütekriterien

- Auch in der IRT-Welt gibt es natürlich Gütekriterien
- Wir unterscheiden in
 - **Itemfit:** Wie gut passen die Items zum Konstrukt?
 - **Personenfit:** Wie gut lassen sich einzelne Personen durch das Modell erklären?
- Wir bekommen also Gütemaße für alle Items und alle Einzelpersonen!
 - Damit können wir einen Test wesentlich umfangreicher beurteilen
- Weiterhin gibt es auch hier eine Art Reliabilitätsgesamtindex

Itemfit: Item-Infit und -Outfit

Itemfit Statistics:

	Chisq	df	p-value	Outfit	MSQ	Infit	MSQ	Outfit t	Infit t
flugzeug	55.616	90	0.998	0.611	0.709	-0.98	-1.96		
kaffee	66.267	90	0.971	0.728	0.944	-0.80	-0.36		
sparen	84.369	90	0.648	0.927	0.952	-0.31	-0.35		
jacke	71.384	90	0.926	0.784	0.829	-1.31	-1.45		
phone	75.316	90	0.866	0.828	0.912	-0.79	-0.73		

Mean-square Value Implication for Measurement

> 2.0	Distorts or degrades the measurement system. May be caused by only one or two observations.
1.5 - 2.0	Unproductive for construction of measurement, but not degrading.
0.5 - 1.5	Productive for measurement.
< 0.5	Less productive for measurement, but not degrading. May produce misleadingly high reliability coefficients.

Personfit: Person Infit und Outfit

Frage: Diskriminieren die Items Personen gleichsam? Passt das Antwortverhalten einer Person zu dem Modell?

- Problematisch in unserem Beispiel: Eine Person hat angegeben sich ein iPhone, aber keine Jacke leisten zu können
- Regeln: Weniger als 5% der Stichprobe sollten einen In- und Outfit > 1.96 haben.

```
pp_fit <- personfit(pp)
prop.table(table(pp_fit$p.infitZ > "1.96"))
```

FALSE	TRUE
0.98901099	0.01098901

```
prop.table(table(pp_fit$p.outfitZ > "1.96"))
```

FALSE	TRUE
0.97802198	0.02197802

Person Separation Reliability

```
SepRel(pp)
```

Separation Reliability: 0.4618

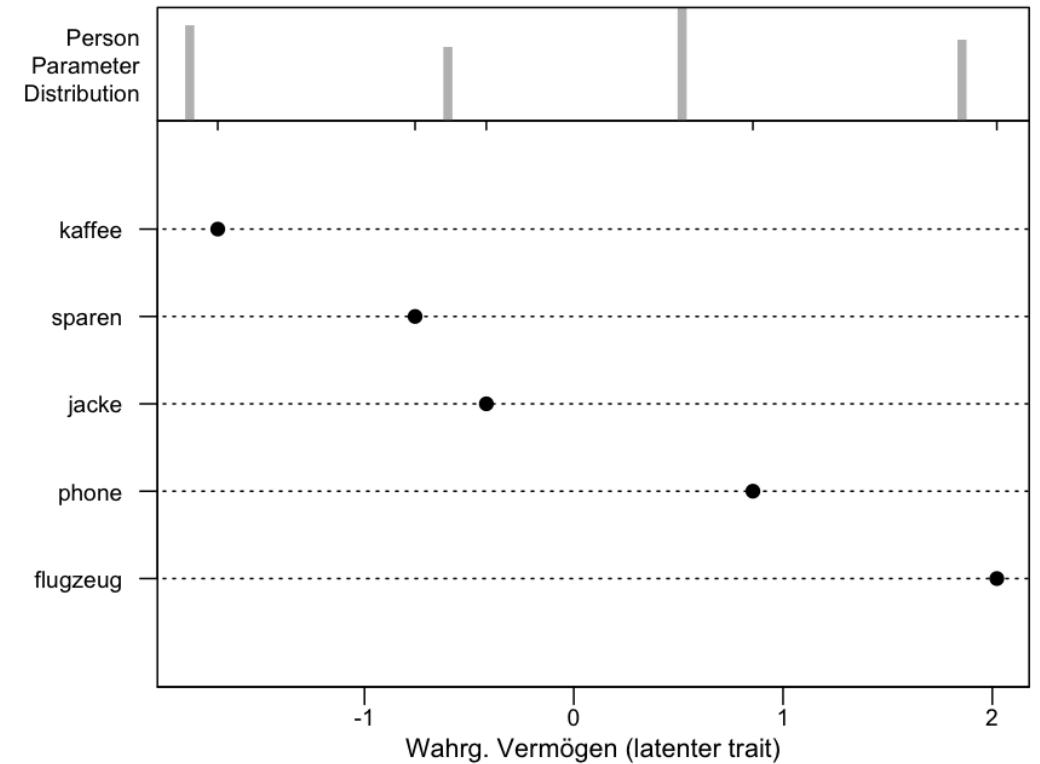
- Anteil der Personenvarianz, der nicht durch Messfehler erklärt wird
- Ähnlich wie Cronbach's Alpha (also hier nicht besonders gut!)

Aus der Dokumentation des **eRm**-Paketes:

"Please note that the concept of reliability and associated problems are fundamentally different between IRT and CTT (Classical Test Theory). Separation reliability is more like a workaround to make the "change" from CTT to IRT easier for users by providing something "familiar." Hence, we recommend not to put too much emphasis on this particular measure and use it with caution."

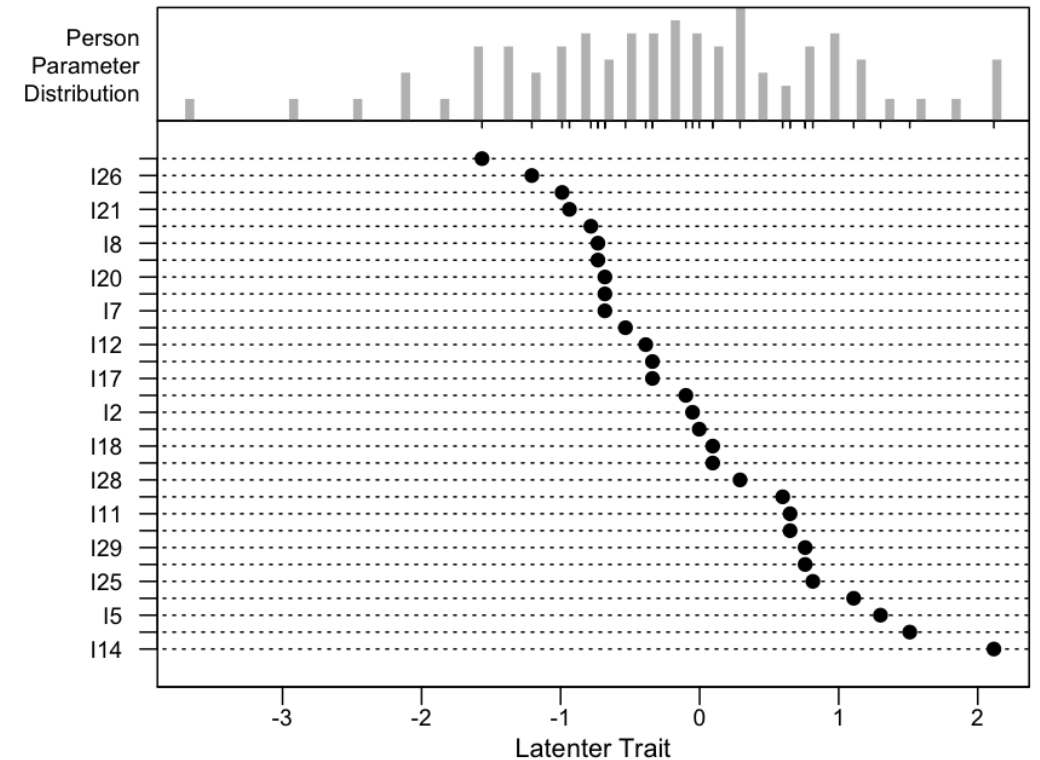
Zusammenfassung: Person-Item-Maps

```
plotPImap(rasch,  
  sorted = T,  
  cex.gen = .8,  
  main = "",  
  latdim = "Wahrg. Vermögen (latenter trait)")
```



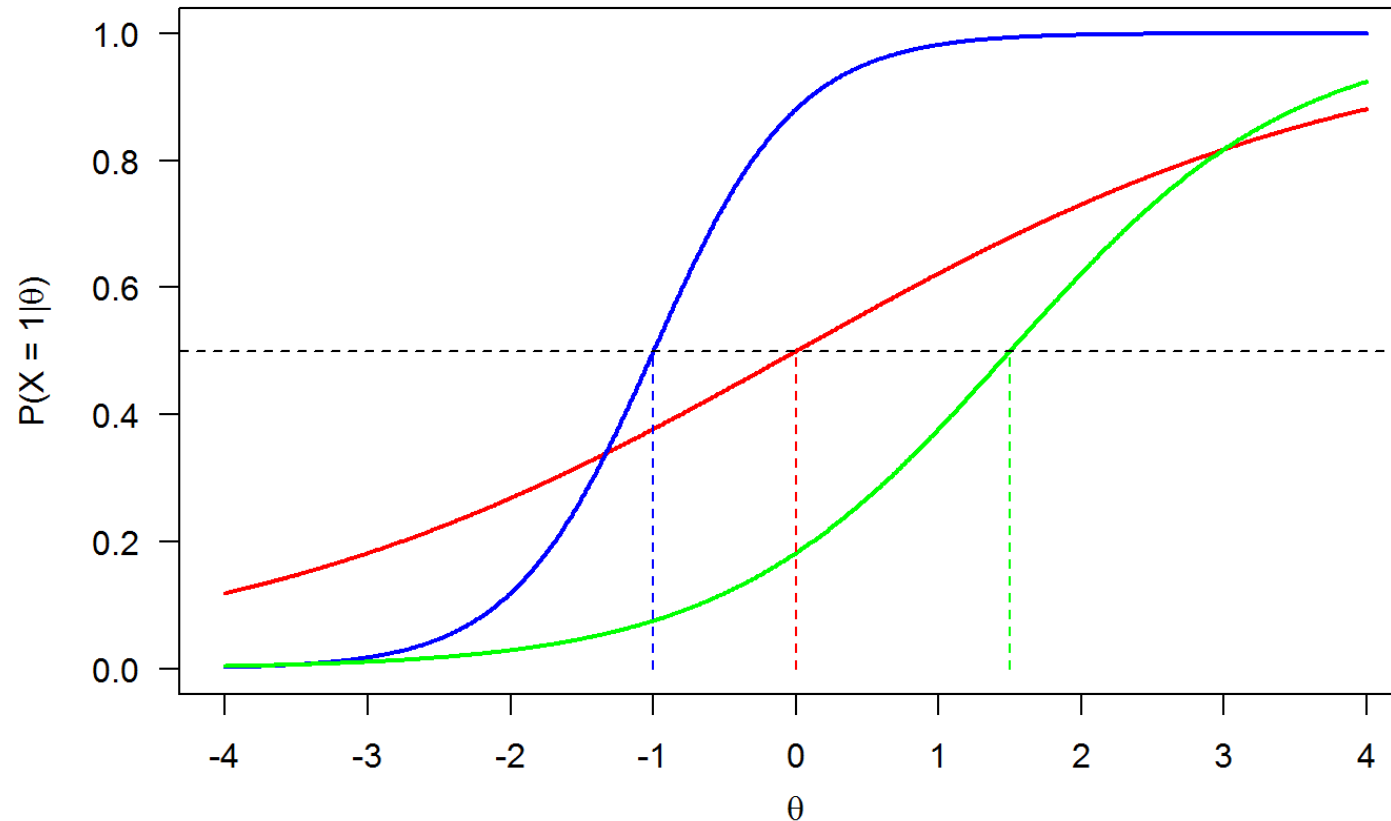
Beispiel für ein komplexeres Modell

```
rasch2 <- RM(raschdat1)
plotPImap(rasch2,
  sorted = T,
  cex.gen = .8,
  main = "",
  latdim = "Latenter Trait")
```



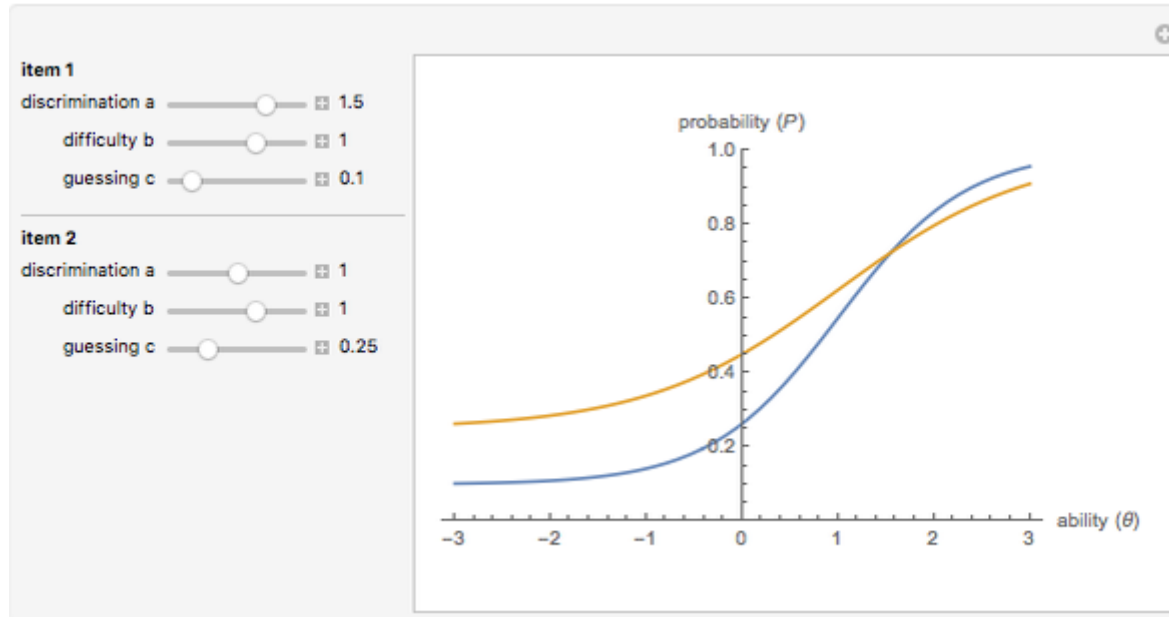
Ausblick 1: 2PL-Modell (variierender Diskriminanzparameter)

- Auch in der IRT-Welt gibt es τ -kongenerische Modelle
- Bei sogenannten 2PL-Modellen darf zusätzlich der Diskriminanzparameter variieren (im Grunde die Güte des Items!)



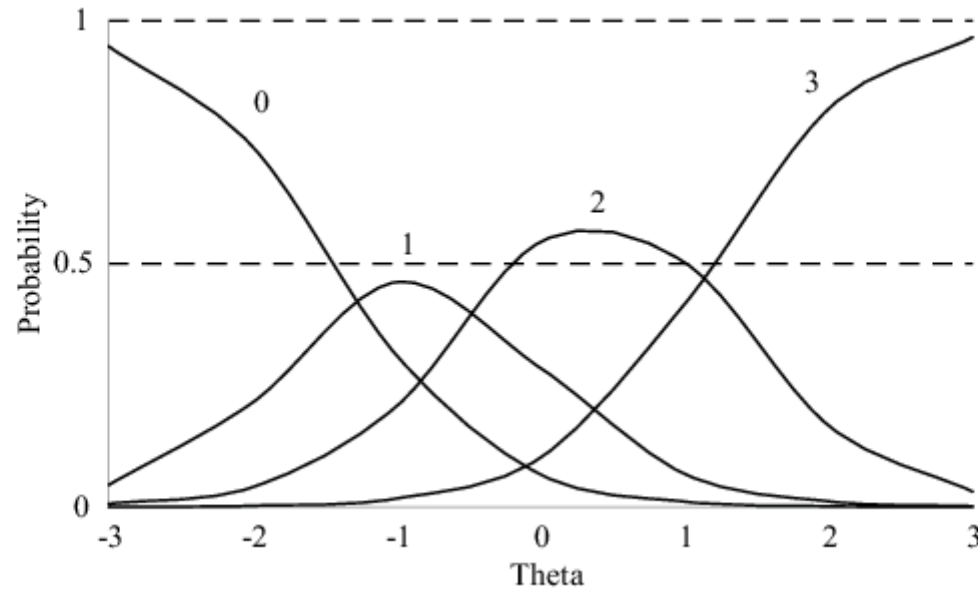
Ausblick 2: 3PL-Modell (Ratewahrscheinlichkeit)

- Man kann zusätzlich auch die Ratewahrscheinlichkeit modellieren
- Ein typisches Wahr/Falsch-Item hat schließlich eine Ratewahrscheinlichkeit von 50%!



Ausblick 3: Graded Response Model

- Und auch in der IRT-Welt, können Items mehrere Antwortoptionen haben...



Fazit

Stimmst du zu oder nicht zu?

Fazit

- Die **explorative Faktorenanalyse** wird noch immer sehr häufig in der Forschung eingesetzt
- Da man die Faktorstruktur (die Anzahl Faktoren) und inhaltliche Konzeption der Faktoren nicht beeinflussen kann, ist der Einsatz vielfach fraglich (warum haben wir sonst Theorie?)
- Die **bisherige Standardpraxis** mit Hauptkomponentenanalyse, Varimax-Rotation und Kaiser-Kriterium ist **schlechte Praxis**
- Wenn dann sollte man die **Hauptachsen-** oder **ML-Faktorenanalyse** mit obliquer Rotation und Parallelanalyse umsetzen
- Wenn man Hypothesen zur Faktorstruktur hat, kann und sollte man lieber gleich eine **konfirmatorische Faktorenanalyse** durchführen
- In vielen Fällen kann die **Item-Response-Theorie** helfen, validere und reliablere Messinstrumente zu entwickeln; insbesondere, wenn es um objektive Tests und Diagnostik geht!

Zum Abschluss...

- Testtheorien sind sehr weit entwickelt
 - In der **Praxis** (z. B. Pisa, Eignungstest) erfolgreich umgesetzt (IRT!)
 - Ermöglichen die **Herleitung** komplexer Messmodelle
 - Definieren deutlich, wie ein Konstrukt auch **validiert** werden sollte
- Forschungsalltag
 - Häufig werden unvalidierte, ungetestete Messinstrumente übernommen (Fragwürdiges Ideal: "Nimm was bereits benutzt wurde")
 - Bisher wenig Berücksichtigung des Messfehlers (weil primär Mittelwerte benutzt werden!)
 - Ad-hoc-Entwicklung von Kurzskalen mit fragwürdiger Validität
- Was bedeutet das für die Robustheit unserer Ergebnisse?

Empfehlungen

1. Auch (oder gerade) häufig benutzte Skalen explizit überprüfen (CFA!)
2. Im Zweifel zusätzliche Items entwickeln, damit mehr Spielraum bei der Überprüfung gegeben ist
3. Wenn eigene Skala entwickelt werden soll: Entwicklungsphasen beachten und (wenn möglich) mehrere Studien durchführen
4. Wenn möglich, Messfehler "herausrechnen": Strukturgleichungsmodellierung oder zumindest Factorscores!
5. Grundsätzlich eine Messung kritisch angehen: Messe ich hier vielleicht nur meine Vorstellung vom Konstrukt?

Noch offene Fragen?

Literatur

- Beaujean, A. A. (2014). *Latent Variable Modeling Using R: A Step-by-Step Guide*. Routledge.
- Bühner, M. (2011). *Einführung in die Test- und Fragebogenkonstruktion (3. Aufl.)*. Hallbergmoos: Pearson. (Kap. 2, 6)
- DiStefano C, Zhu, M. & Mîndrila, D. (2009). Understanding and Using Factor Scores: Considerations for the Applied Researcher. *Practical Assessment, Research & Evaluation*, 14(20), 1-11.
- Eid, M., Gollwitzer, M. & Schmitt, M. (2017). *Statistik und Forschungsmethoden (5. Auflage)*. Weinheim: Beltz Verlag.
- Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. London: Sage.
- Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, 4, 272–299.
- Hartmann, T. & Reinecke, L. (2013). Skalenkonstruktion in der Kommunikationswissenschaft. In W. Möhring & D. Schlütz (Hrsg.), *Handbuch standardisierte Erhebungsverfahren in der Kommunikationswissenschaft*. Berlin: Springer.

Literatur

- Masur, P. K., Teutsch, D. & Trepte, S. (2017). Entwicklung und Validierung der Online-Privatheitskompetenzskala (OPLIS) [engl. Development and validation of the online privacy literacy scale]. *Diagnostica*, 63, 256-268. <https://doi.org/10.1026/0012-1924/a000179>
- Reise, S. P. (2012). The Rediscovery of bifactor measurement models. *Multivariate Behavioral Research*, 47, 667 – 696.
- Weiber, R., & Mülhhaus, D. (2014). *Strukturgleichungsmodellierung: Eine anwendungsorientierte Einführung in die Kausalanalyse*. Heidelberg: Springer.