GDCP Methodenworkshop 2014

Einführung in quantitative Forschungsmethoden

Christoph Kulgemeyer

&

Christoph Gut-Glanzmann

Klassische Testtheorie

Ausgangslage: ein gängiges Problem

- Wir wollen testen, ob eine neuartige Unterrichtsmethode zu besseren Ergebnissen führt.
- Z.B. Vergleich Frontalunterricht gegen Gruppenpuzzle bei der Einführung des Kraftbegriffs – Vergleich zweier Gruppen
- Wir könnten erheben:
  - Motivation
  - Fachwissen
  - Interesse
  - etc.

Was ist ein Test?

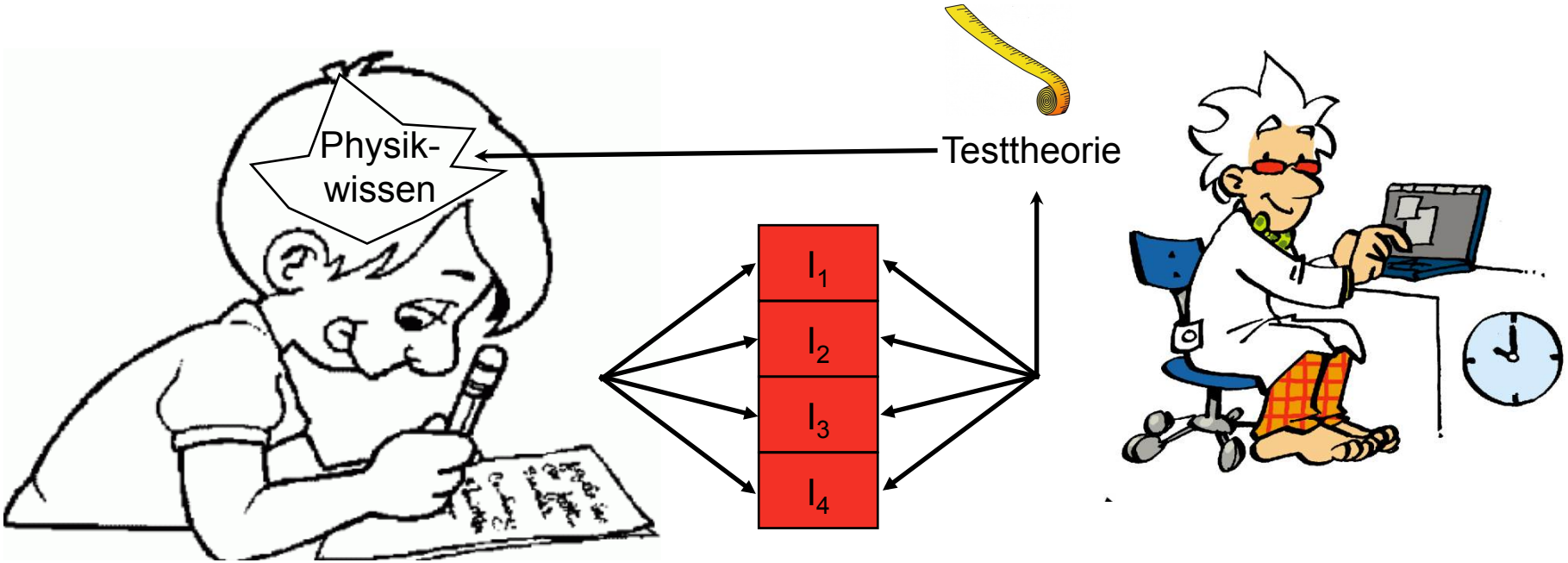
- Test: Entweder ein statistisches Verfahren, um Hypothesen auf Signifikanz prüfen zu können - oder ein Messinstrument

- „Ein wissenschaftliches Routineverfahren zur Untersuchung eines oder mehrerer empirisch abgrenzbarer Persönlichkeitsmerkmale mit dem Ziel einer möglichst quantitativen Aussage über den relativen Grad einer individueller Merkmalsausprägung“ (Lienert & Raatz, 1998)

- Verschiedene Arten von Tests:
  - Leistungstest: Leistung nach Kritierien als richtig oder falsch beurteilen - z.B. Fachwissen
  - Persönlichkeitstest: Ausprägung eines Persönlichkeitsmerkmals - z.B. Interesse

Was ist Testtheorie?

- Die Testtheorie beschäftigt sich mit der Frage, wie aus einer Anzahl Verhaltensbeobachtungen von Versuchsperson in bestimmten Situationen auf die „wahre“ Ausprägung eines Persönlichkeitsmerkmals geschlossen werden kann.
- Dabei werden die „Verhaltensbeobachtungen“ (z.B. Itemantworten) als manifeste Variable bezeichnet und die Merkmalsausprägung als latente Variable



KTT - Klassische Testtheorie

- „Klassisch“, weil die erste Testtheorie - naturwissenschaftliche Messungen als Vorbild
- Grundannahme: jede Messung ist fehlerbehaftet (z.B. Testmotivation, Müdigkeit, Ratewahrscheinlichkeit). Die KTT ist eine Messfehlertheorie.
- D.h. der wahre Wert ergibt sich als Mittelwert über unendlich viel Messungen - jedes Item kann man als einzelne Messung begreifen
- Unterscheidung zwischen
  - Messwerte X (Verhaltenbeobachtungen)
  - wahre Werte T
  - (zufällige) Messfehler F
- Jede Messung ist mit einem zufälligen Fehler behaftet.
  - X = T+F

Beispiel

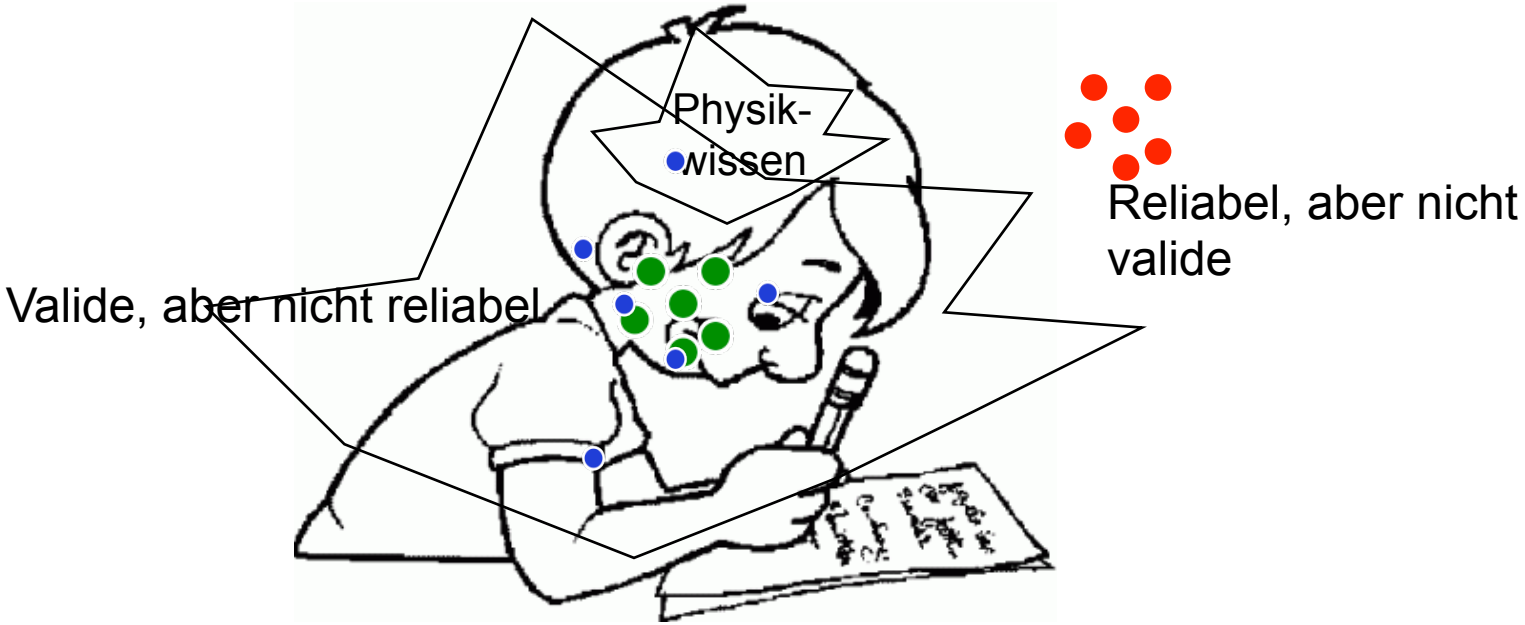
- Wie kommt man von den gesammelten Daten zu einer Aussage über die Fähigkeit einer Person?
  - Lienert: $X = N_R$
  - oder: $X = N_R - N_F$
- Wie kommt man von den gesammelten Daten zu einer Aussage über die Schwierigkeit eines Items?
  - Lienert: $P = 100\ (N_R / N)$
  - $P = 100\ (N_R / N_F)$

	I1	I2	I3	I4	I5
P1	0	1	1	1	0
P2	0	1	1	0	0
P3	1	0	0	0	0
...	...	...	...	...	...
P300	1	1	1	0	0

- Personenfähigkeit ist abhängig von der Itemstichprobe.
- Wie werden die einzelnen Items gewichtet?

Woher weiß ich klassisch, dass mein Test gut ist?

- Validität: Aus meinen Messwerten kann ich begründet Schlüsse über das Merkmal ziehen
- Reliabilität: Das Messinstrument soll genau messen – und zwar jedes Item (=Testaufgabe)
- Objektivität: Das Messergebnis soll unabhängig von äußeren Bedingungen sein (z.B. Testauswerter, Testdurchführerin)



Objektivität

- Durchführungsobjektivität
  - Die Durchführung eines Tests darf nicht von Mal zu Mal unterschiedlich sein.
  - Durchführungsvorschriften: Die Durchführung ist unabhängig vom Testleiter (z.B. vorgeschriebener Ablauf, vorgeschriebene Zeit,...)

- Auswertungsobjektivität:
  - Jeder Auswerter sollte die gleichen Punktwerte eines Probanden ermitteln
  - Auswertungsvorschriften und „Interraterreliabilität “: Die Auswertung ist unabhängig vom Auswerter

- Interpretationsobjektivität:
  - Jeder Auswerter soll zu den gleichen Interpretationen der Testergebnisse gelangen

Reliabilität

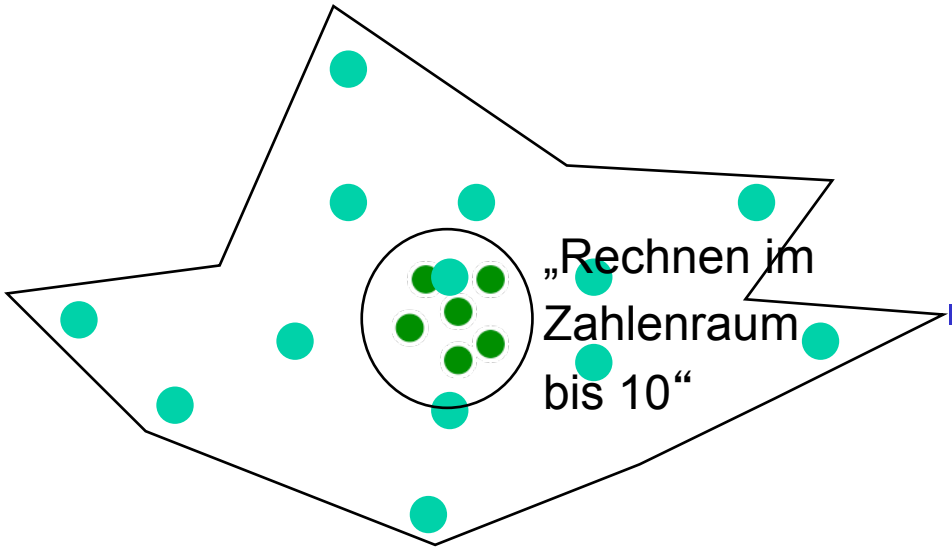
- Split-Half-Test Reliabilität
  - Idee: Der Test wird in zwei Hälften geteilt, jeder soll zu demselben Ergebnis kommen. Ergebnisse können korreliert werden.

- Interne Konsistenz: Cronbachs α
  - Jedes Item wird als eigener „Test“ aufgefasst! Wie sehr kommen alle Items zu derselben Einschätzung des Personenmerkmals?
  - In SPSS: ANALYSIEREN – SKALIEREN- RELIABILITÄTSANALYSE (Unter STATISTIK „SKALA WENN ITEM GELÖSCHT“ anklicken)
    - Cronbachs α ist eine Verrechnung von Itemanzahl und Iteminterkorrelation
    - Itemanzahl: Anzahl der Messungen erhöht Genauigkeit
    - Interkorrelation: Wie sehr messen die Items dasselbe?
    - Viele Konventionen... oftmals Cronbachs $\alpha > 0{,}7$

Validität

- Früher: Eigenschaft eines Tests, heute: wie angemessen ist die Interpretation der Messwerte, d.h. der Schlüsse, die ich aus den Messwerten ziehe?

- Inhaltsvalidität: inwieweit erfassen die Testitems das zu messende Merkmal?



- schlechte Inhaltsvalidität, hohes Cronbachs Alpha - „zu Tode homogenisiert“. Alphamaximierung ist keine gute Testentwicklung!
- niedriges Alpha, gute Inhaltsvalidität - übliches Problem fachdidaktischer Messung (Psychologie: Alpha größer 0,8! Nur durch viele Items erreichbar)
- Methoden: Modell, Expertenbefragung

Validität

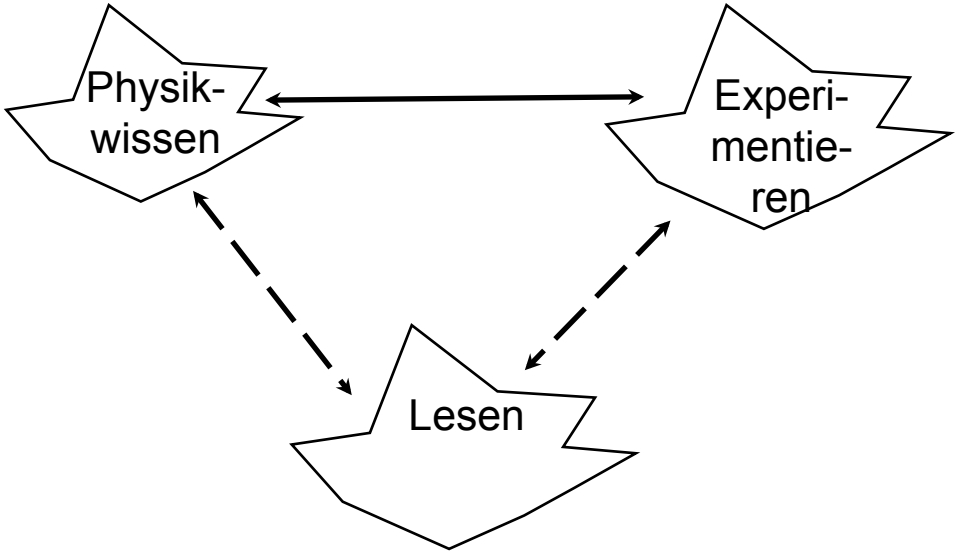
- Früher: Eigenschaft eines Tests, heute: wie angemessen ist die Interpretation der Messwerte, d.h. der Schlüsse, die ich aus den Messwerten ziehe?

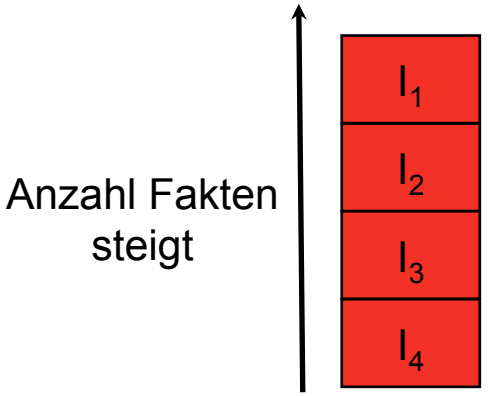
- Inhaltsvalidität: inwieweit repräsentieren die Testitems das zu messende Merkmal?
  - Methoden: Expertenbefragung, Modelle

- Kriteriumsvalidität: Vorhersage eines Kriteriums außerhalb des Tests
  - Konkurrente Validität: Wie sehr wird durch den Test ein verwandtes Merkmal vorhergesagt? Z.B. Fachwissen und Fachnote
  - Prognostische Validität: Wie sehr gelingt es spätere Leistungen vorherzusagen? Z.B. Fachwissen und Abschneiden in Prüfung

Validität

- Konstruktvalidität: Übergreifendes Konzept, inwieweit kann das Testergebnis erklärt werden?
  - z.B. Zusammenhang zu anderen Merkmalen (nomologisches Netzwerk)
  - z.B. Innerer Zusammenhang (ist die Schwierigkeit der Testitems theoretisch begründbar?)





Probleme der KTT

- Die KTT stellt eigentlich keine Verbindung zwischen Fähigkeit und Itembeantwortung her. Sie beschäftigt sich in den Gütekriterien nur damit, die Messfehler klein zu halten
- Alle Personen müssen dieselben Items gelöst haben, um die Personenfähigkeit zu bestimmen - es gibt keine allgemeine Skala für Personenfähigkeit
- Alle Items müssen von denselben Personen gelöst werden, um die Schwierigkeit zu kennen - es gibt keine allgemeine Skala für Itemschwierigkeit
- Die Annahme, dass jedes Item bis auf statistische Schwankungen richtig beantwortet wird, wenn die Fähigkeit vorliegt, ist schwer haltbar
- Aber: in der Praxis große Erfolge, 95 % aller Tests sind klassisch

Exkurs: Konstruktion von Likert-Skalen

Skalen

	Volle Zustimmung				Volle Ablehnung
■ Stufe Dich selbst ein!					
■ Der Unterricht hat mir heute Spaß gemacht.	☐	☐	☐	☐	☐
■ Ich konnte selbst mit über die Inhalte bestimmen.	☐	☐	☐	☐	☐

- Skalenniveaus:
- Nominal (Rot, Grün, Gelb – keine qualitativen Unterschiede)
- Ordinal (Sport ist beliebter als Physik – Abstufung, aber keine Abstandsmaße)
- Intervall (Hans ist drei Mal besser in Physik als Timo)
- Likert-Skala? Äquidistanz (Stimmt gar nicht, Stimmt wenig, Stimmt teils, Stimmt ziemlich, stimmt vollständig)
- Kodierung von 1 bis 5 – Missing Data?

Exkurs: Likert-Skalen

- Viele Fragebögen zu Personenmerkmalen verwenden Likert-Skalen.
- Die Probanden sollen einer Aussage zustimmen oder sie ablehnen, z.B. „Ich freue mich immer auf den Physikunterricht“
- Für die Antwort gibt es ein gestuftes Schema
  - Bsp. 4-stufig:
  - 1: „trifft voll zu“ | „2: trifft eher zu“ | 3: „trifft eher nicht zu“ | 4: „trifft gar nicht zu“
  - Bsp. 5-stufig:
  - 1: „trifft voll zu“ | „2: trifft eher zu“ | 3: „unentschieden“ | 4: „trifft eher nicht zu“ | 5: „trifft gar nicht zu“

Exkurs: Likert-Skalen

- Soll man 4- oder 5-stufige Antworten verwenden?
  - 5-stufig
    - Antworttendenz zur Mitte
    - „unentschieden “ schwer von „weder noch “ oder „weiß nicht “ abgrenzbar
  - 4-stufig
    - Mittelwert nicht klar interpretierbar
    - Probanden werden zu Tendenzentscheidungen gezwungen, die sie vielleicht eigentlich nicht treffen möchten.
  - Mögliche Erweiterung bei beiden Skalen
    - 6: „Ich möchte die Frage nicht beantworten “ | „Ich habe dazu keine Meinung “ | „weder/noch “
    - Problem: Diese Option erzeugt systematisch „fehlende Werte

Exkurs: Likert-Skalen

- Es handelt sich um ordinale Daten. Häufig werden die Daten jedoch als a) intervallskaliert oder zumindest b) als äquidistante Ränge angenommen. Zumindest werden sie so behandelt, obwohl a) gar nicht und b) eigentlich nicht zulässig ist!
- Was eigentlich alles aufgrund der Datenqualität „ordinal “ nicht geht:
  - Mittelwerte und Standardabweichungen berechnen
    --> Median angeben
  - Skalenreliabilitäten auf Grundlage von Produkt-Moment-Korrelationen oder Varianzen berechnen
    --> Rangkorrelationen zugrundelegen
- Deshalb wird häufig bewusst über die Probleme hinweggesehen:„Augen zu und durch “. (Selbst Bortz drückt sich um klare Aussagen.)
  Man sollte zumindest den Probanden nahelegen, dass die Stufen äquidistant sein sollen, z.B. durch eine grafische Skala von 1 bis 5.

Checkliste Formulierung von Aufgaben

- Habe ich Fremdwörter vermieden?
- Testet jedes Item nur ein Konstrukt?
- Habe ich mehrdeutige Begriffe vermieden?
- Lassen meine Antworten eindeutige Aussagen zu?
- Habe ich soziale Erwünschtheit berücksichtigt? („Wenn mich jemand ärgert, möchte ich manchmal am liebsten zuschlagen. “)

- Kann ich mir sicher sein, dass die Aufgaben so verstanden werden, wie ich sie gemeint habe? (Kognitive Validität, Lautes Denken)

Übungsphase 1

Explorative Faktorenanalyse

Differenzierte Ziele

Es gibt verschiedene «faktoranalystische» Methoden mit unterschiedlichen Zielen (Auswahl):

- **Hauptkomponenten-Analyse** (PCA: Principal Component Analysis)
  Ziel: Datenreduktion von vielen Items zu wenigen Supervariablen, sog. Komponenten (= beste Linearkombinationen der Items). Die Komponenten sollen als **Sammelbegriffe** der Item verstanden werden.

- **Hauptachsen-Analyse** (PAF: Principal Axis Factor Analysis)
  Ziel: möglichst umfassende Reproduktion der Datenstruktur durch möglichst wenige **Faktoren**, die kausal interpretiert werden können.
  Alternative: Maximum-Likelihood-Analyse (stabiler)

⇒ Beide Verfahren werden benutzt, um Skalen bzgl.
Interpretierbarkeit (e.g. Dimensionen) und
Homogenität (durch Ausschluss von Items) zu optimieren

Fragestellung: Itemauswahl

- **Gegeben:**
  Antworten von 650 SuS auf die Itemgruppe *Familie* (HarmoS-Fragebogen)

x_1	Fam_1	Wir sprechen zu Hause über Themen aus der Schule.
x_2	Fam_2	Wir gehen mit der Familie oft nach draussen.
x_3	Fam_3	Wir besuchen mit der Familie Museen, Lehrpfade usw.
x_4	Fam_4	Wenn wir in den Ferien sind, schauen wir uns Sachen an diesem Ort an und unternehmen dort Ausflüge.
x_5	Fam_5	Ich gehe in Bibliotheken, Mediotheken und sehe mir Bücher und Filme zu Themen an oder leihe sie aus.
x_6	Fam_6	Ich sehe im Fernsehen Nachrichtensendungen.
x_7	Fam_7	Ich lese Tageszeitungen.

Fragestellung: Korrelationsmatrix

	Fam_1	Fam_2	Fam_3	Fam_4	Fam_5	Fam_6	Fam_7
Fam_1	1	.242 ***	.191 ***	.146 ***	.242 ***	.078 *	.111 **
Fam_2		1	.308 ***	.279 ***	.247 ***	.024	.145 ***
Fam_3			1	.339 ***	.244 ***	.062	.079 *
Fam_4				1	.174 **	.051	.117 ***
Fam_5					1	.019	.150 ***
Fam_6						1	.368 ***
Fam_7							1

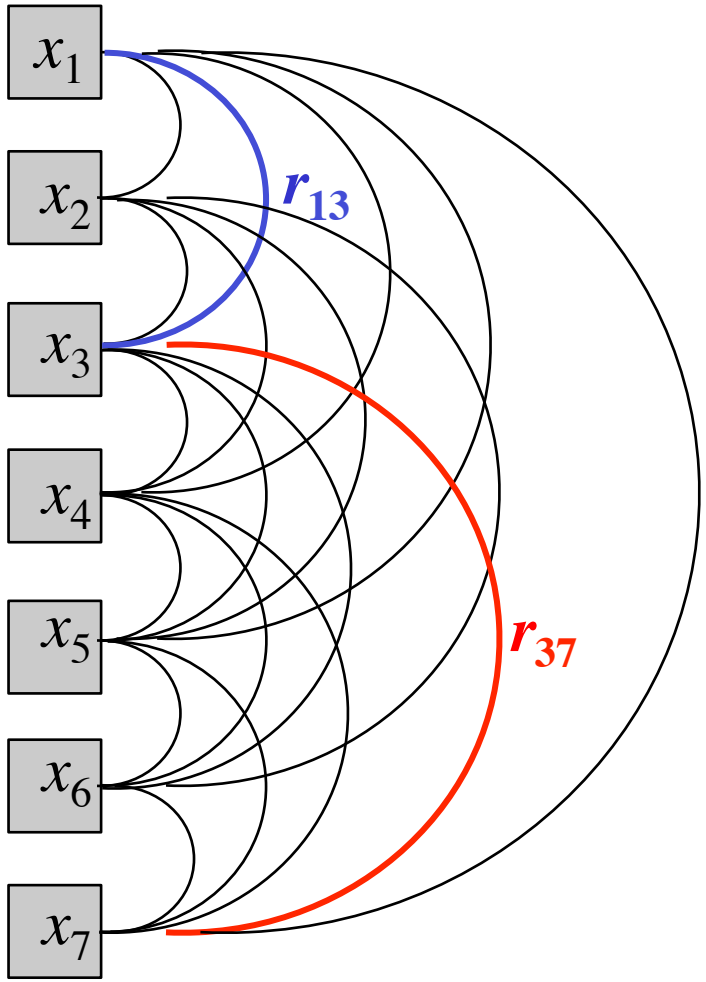
Korrelationen: Kendal-Tau-b * / ** / *** Signifikanzen (2-seitig)

- **Frage bei Hauptkomponenten:**
  Zeigen die signifikanten Korrelationen an, dass die Items durch eine Supervariable (Sammelbegriffs) ersetzt werden können?
- **Frage bei Hauptachsen:**
  Zeigen die signifikanten Korrelationen an, dass die Antworten auf die Items gemeinsame Ursachen haben (Ausprägungen auf latente Variablen)

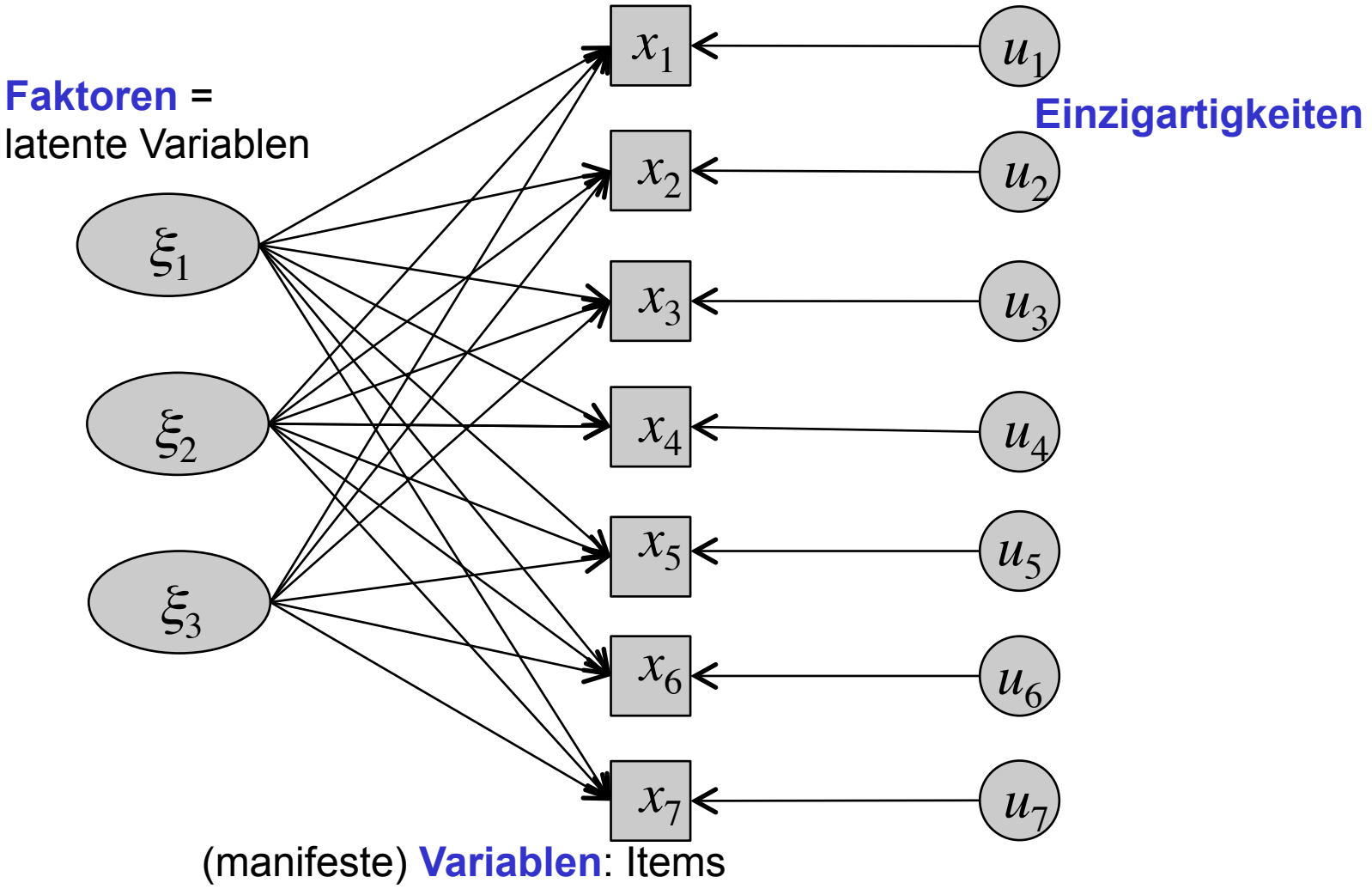
2 “fachdidaktische” Fragen = 1 mathematische Frage

Mathematische Frage

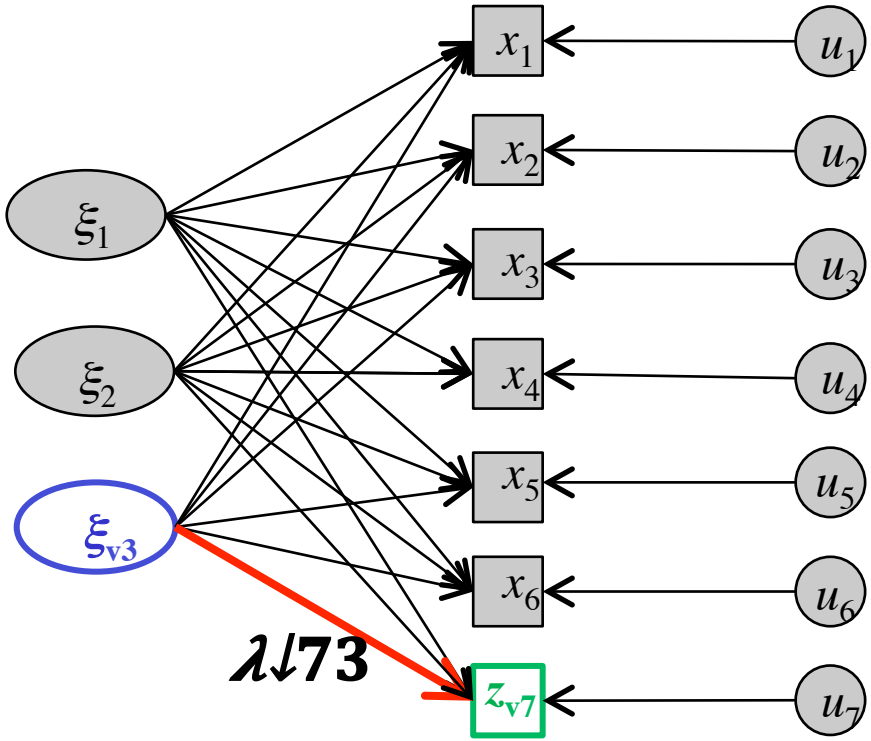
Lassen sich die Korrelationen r_{ij} $(i, j = 1, .. 7)$ zwischen den 7 Variablen x_1 , ... x_7 (Items) mithilfe von Linearkombinationen weniger Faktoren (latenten Variablen) reproduzieren?



Mathematisches Grundmodell



Basis: Lineare Zusammenhänge standardisierter Variablen

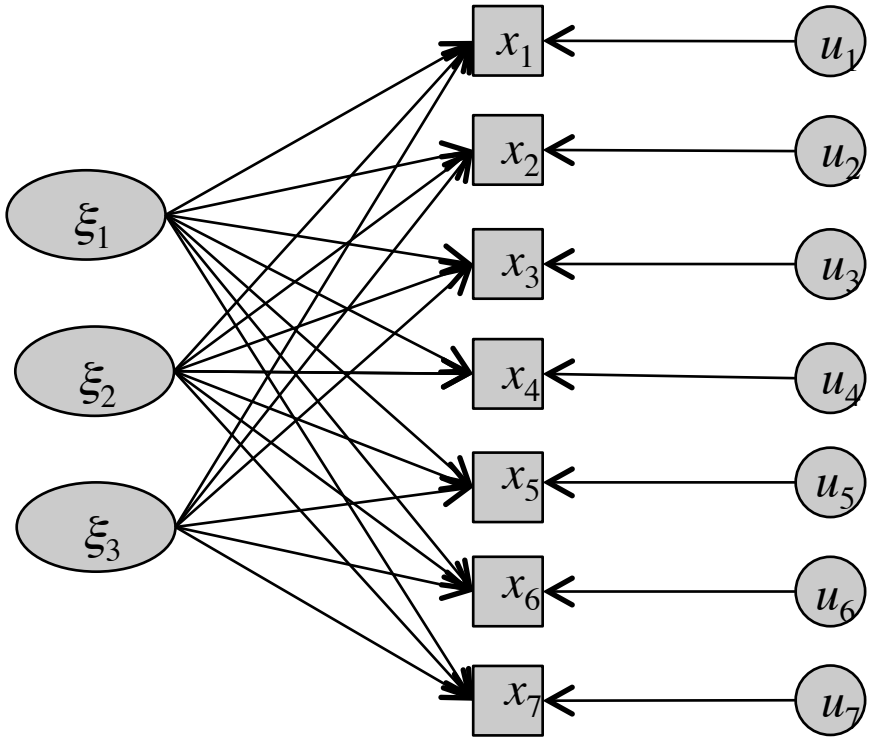


stand. Wert der Person v auf Item i — Ladung des Items i auf Faktor ξ_3

$$z_{vi} = \lambda_{i1}\,\xi_{v1} + \lambda_{i2}\,\xi_{v2} + \lambda_{i3}\,\xi_{v3} + u_{vi}$$

Ausprägung der Person v auf dem Faktor ξ_1

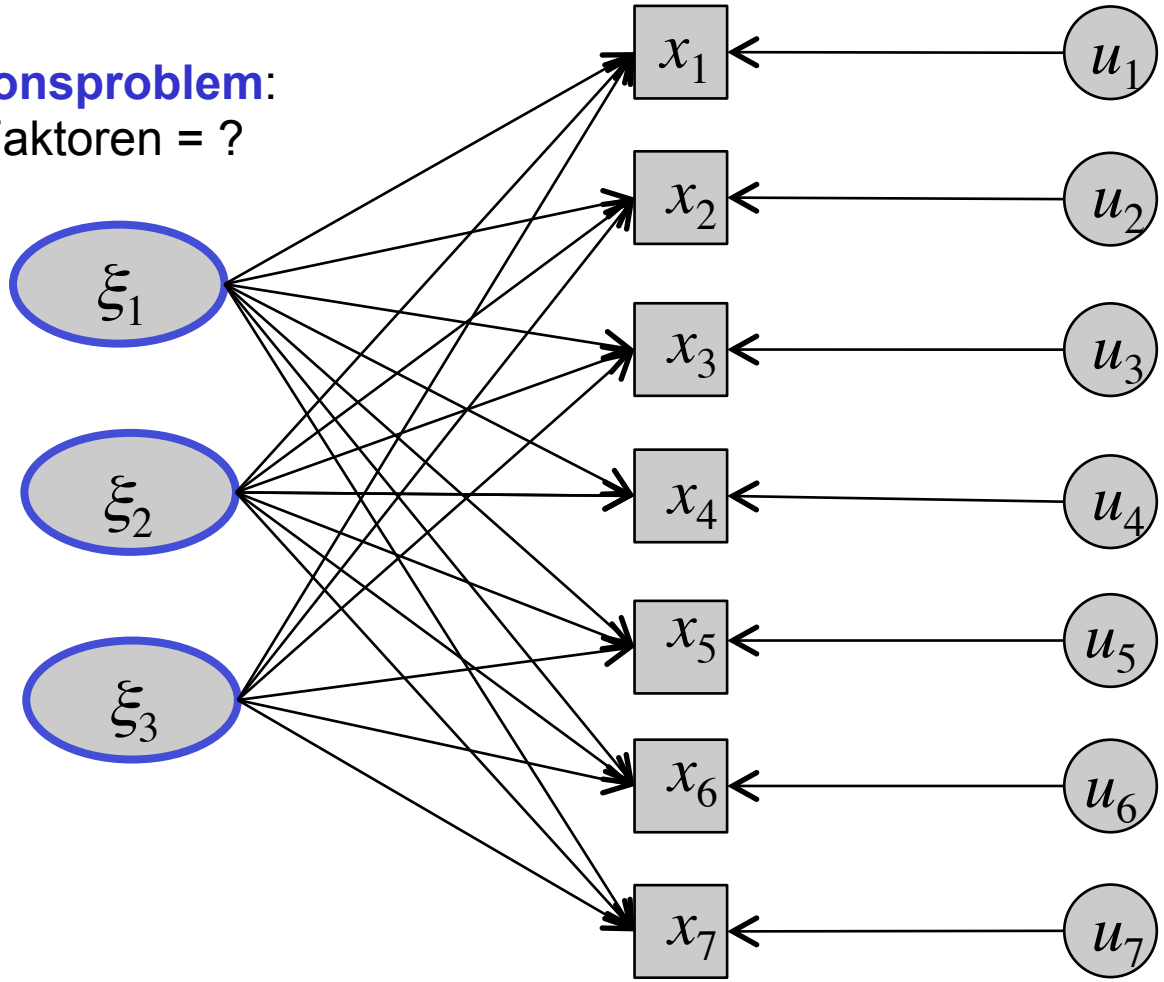
Problem: Vielzahl an Modellen (1)



- Eine Korrelationsmatrix kann stets durch mehrere Modelle auf äquivalente Weise reproduziert werden.
- Die verschiedenen Modelle müssen theoretisch unterschiedlich bewertet werden.
- Die Frage, welches Modell das „Richtige“ ist, kann nicht statistisch (empirisch) beantwortet werden, sondern nur fachdidaktisch bzw. psychologisch.

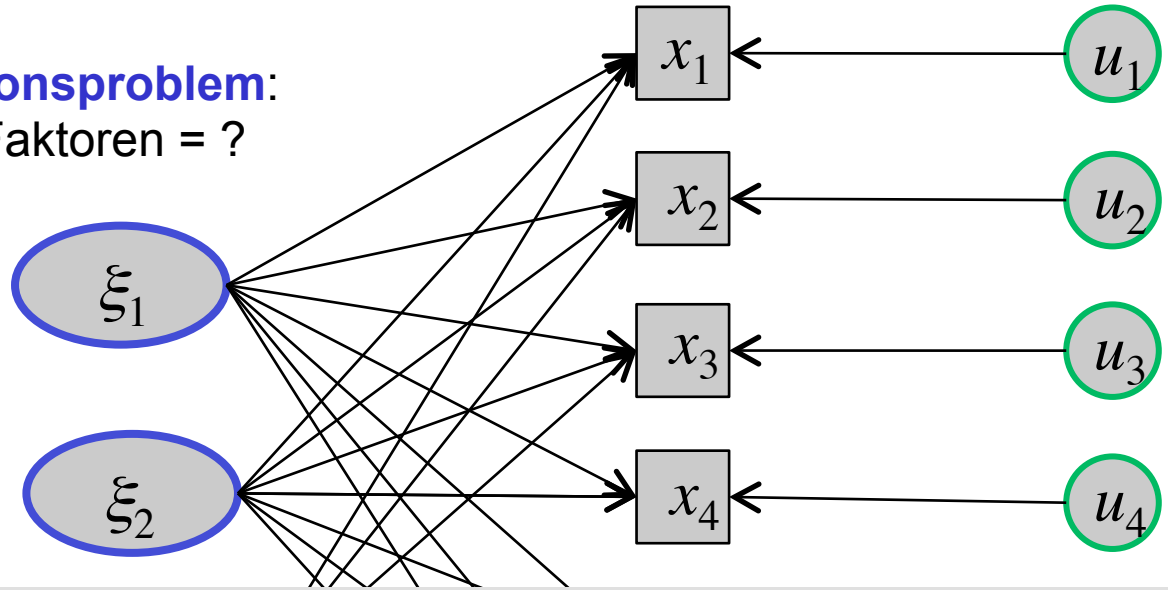
Problem: Vielzahl an Modellen (2)

Extraktionsproblem:
Anzahl Faktoren = ?



Problem: Vielzahl an Modellen (3)

Extraktionsproblem:
Anzahl Faktoren = ?



Kommunalitäten-problem:
Wieviel Varianz soll durch die Faktoren erklärt werden?

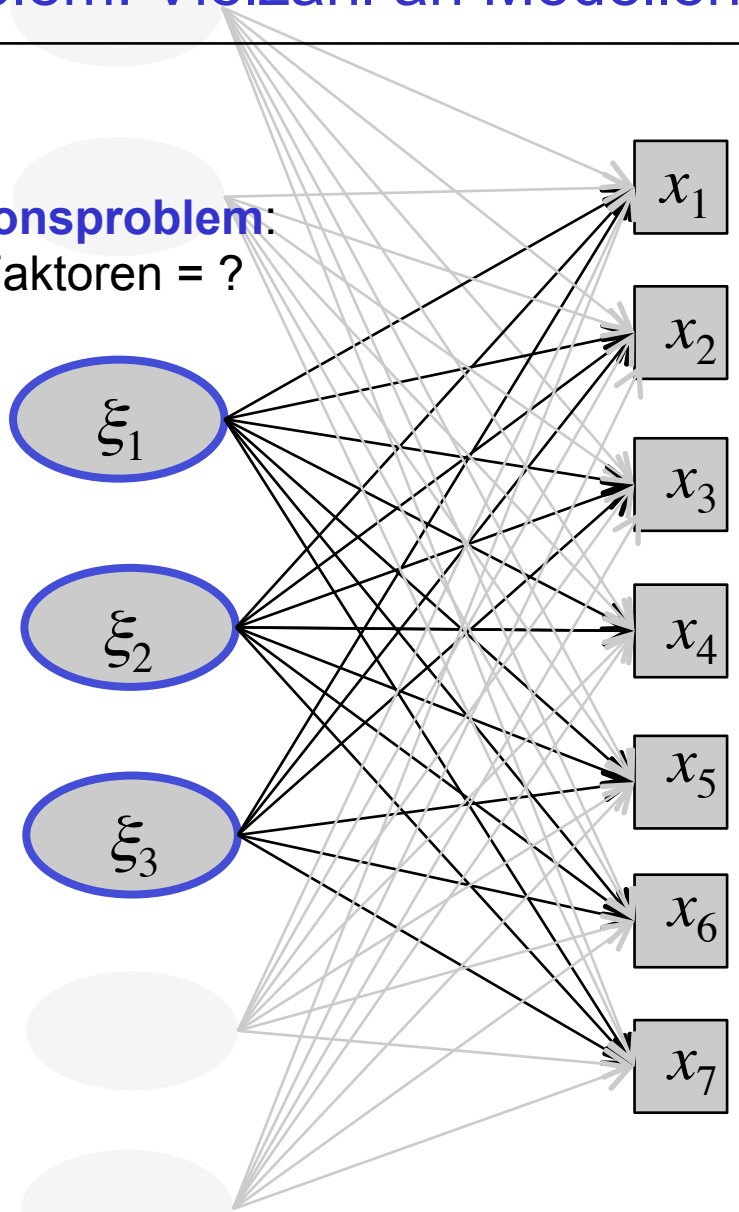
Hauptachsen-Analyse
Es werden für jedes Item eine «unbekannte» latente Variablen angenommen, die die spezifische Varianz und Fehlervarianz des Items erklärt bzw. enthält.

Beispiel einer möglichen «unbekannten Variablen» beim Item Fam_6:

„Ich lese keine Tageszeitungen,
nicht weil ich nicht interessiert wäre,
sondern weil wir zuhause keine Zeitung haben."

Problem: Vielzahl an Modellen (4)

Extraktionsproblem:
Anzahl Faktoren = ?

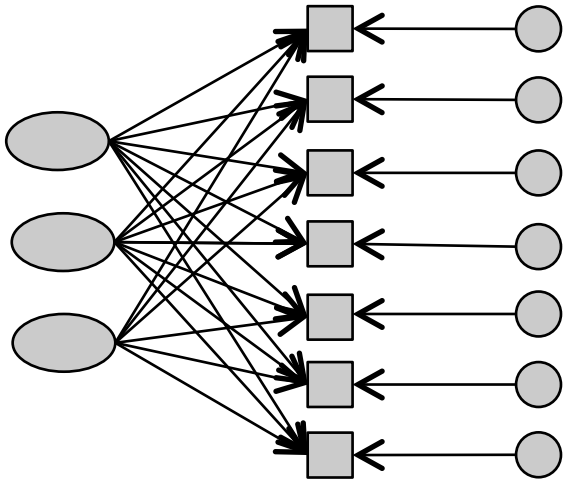


Kommunalitäten-problem:
Wieviel Varianz soll durch die Faktoren erklärt werden?

Hauptkomponenten-Analyse
keine zusätzlichen Variablen angenommen

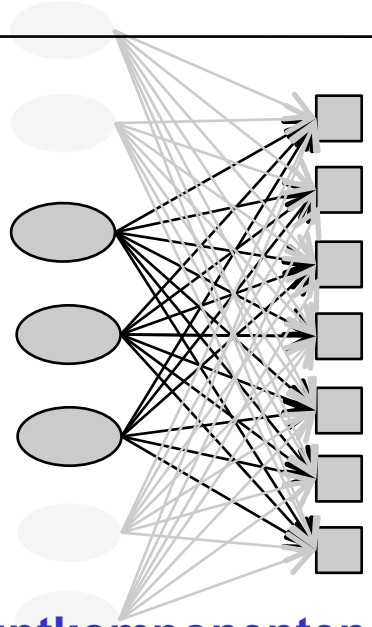
Das Modell extrahiert gleichviele Faktoren wie Items, die sämtliche Varianz in den Items erklären. Interpretiert werden jedoch nur die Faktoren mit den grössten Eigenwerten.

Problem: Vielzahl an Modellen (5)



Hauptachsen-Analyse

- Besser bei kleinen Stichproben, wenig Items pro Faktor , grossen spezifischen Varianzen einzelner Items
- Unterschiede vernachlässigbar bei grossen Stichproben und Itemgruppen



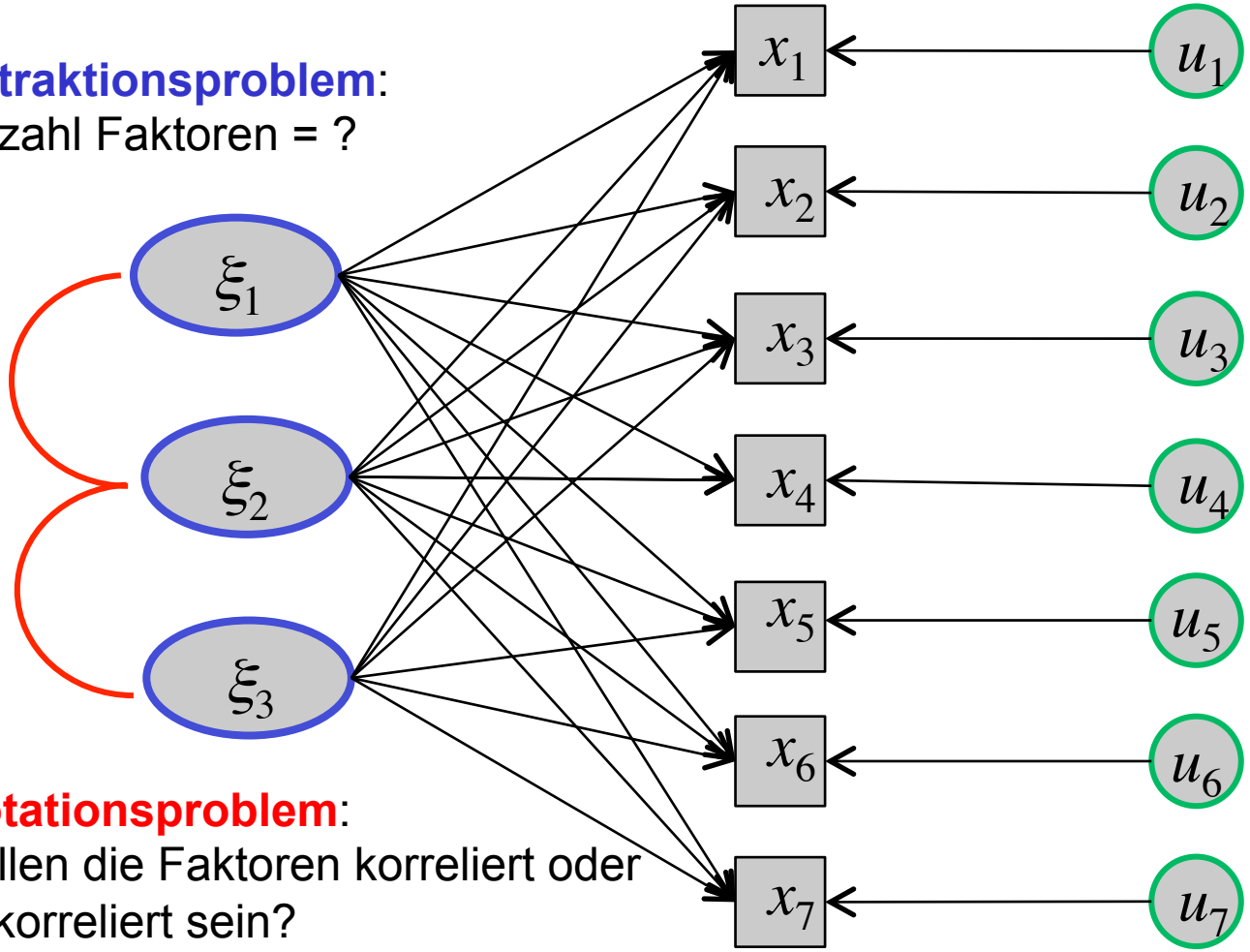
Hauptkomponenten-Analyse

- Psychometrisch fragwürdig, weil Konstrukte keine spezifische Varianzen oder Fehlervarianzen von Items erklären sollten.
- instabil (Stichprobenabhängigkeit gross) bei kleinen Stichproben

Problem: Vielzahl an Modellen (6)

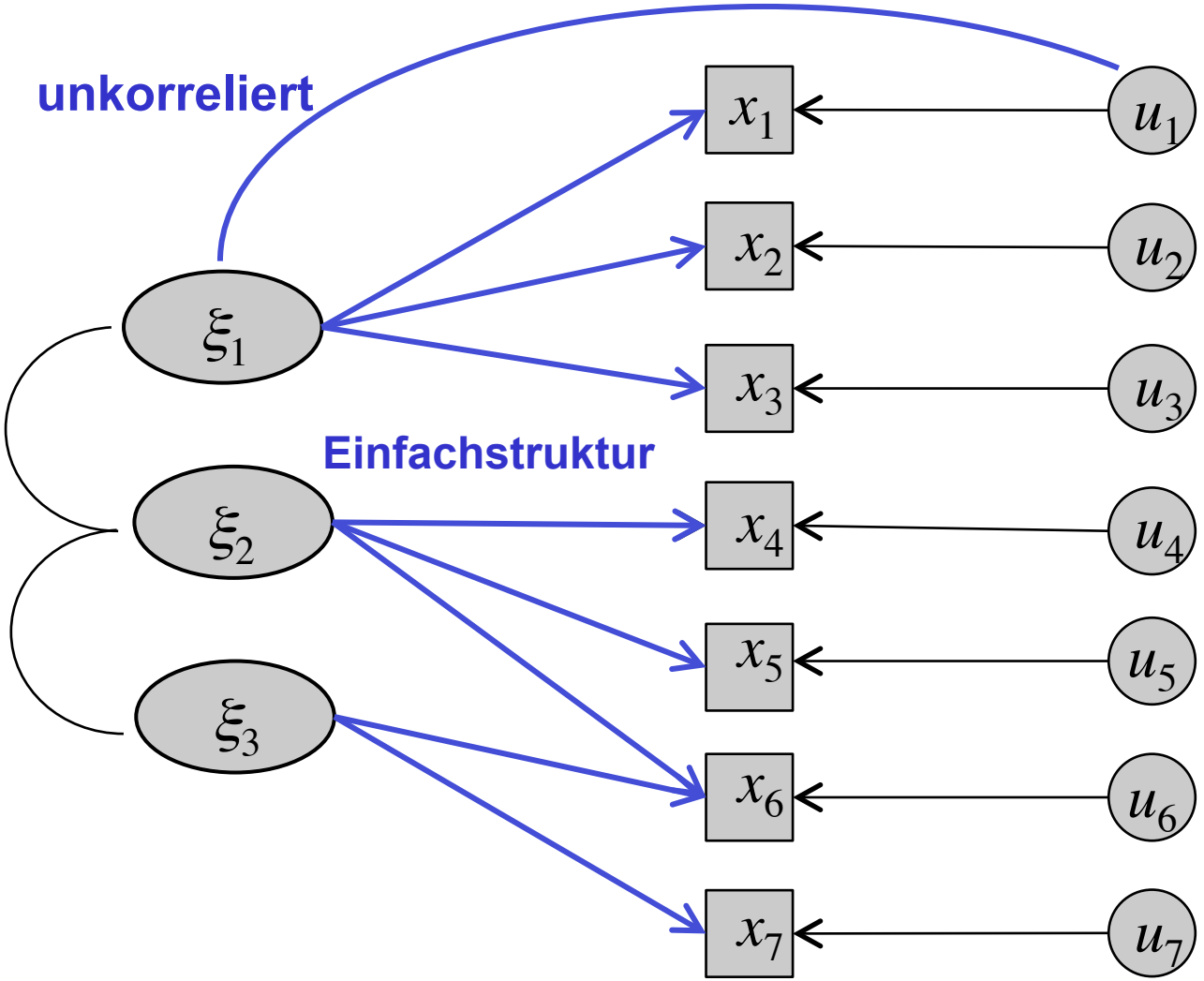
Extraktionsproblem:
Anzahl Faktoren = ?

Kommunalitäten-problem:
Wieviel Varianz soll durch die Faktoren erklärt werden?



Rotationsproblem:
Sollen die Faktoren korreliert oder unkorreliert sein?

Problem: Vielzahl an Lösungen



Welches ist die «richtige» Lösung: Mathematisch wird diejenige Lösung angestrebt, bei der unter bestimmten Randbedingungen (keine Korrelationen mit unbekannten Variablen) eine möglichst gute Einfachstruktur vorliegt.

Beispiel (SPSS / PSPP)

SPSS / PSPP: Wahl der Extraktionsmethode

- **Sollen die Korrelationen der Items durch latente Variablen erklärt werden?**
  Ja => Hauptachsenanalyse

- **Sollen die Daten zu Sammelbegriffen zusammengefasst werden können?**
  Ja => Hauptkomponentenanalyse

Träger von Interessen? Lebensweltbezug der Interessensinhalte?

Fam_1	Wir sprechen zu Hause über Themen aus der Schule.
Fam_2	Wir gehen mit der Familie oft nach draussen.
Fam_3	Wir besuchen mit der Familie Museen, Lehrpfade usw.
Fam_4	Wenn wir in den Ferien sind, schauen wir uns Sachen an diesem Ort an und unternehmen dort Ausflüge.
Fam_5	Ich gehe in Bibliotheken, Mediotheken und sehe mir Bücher und Filme zu Themen an oder leihe sie aus.
Fam_6	Ich sehe im Fernsehen Nachrichtensendungen.
Fam_7	Ich lese Tageszeitungen.

SPSS / PSPP: Eignung der Daten für Faktoranalyse (1)

- **Ist Korrelationsmatrix signifikant von Null verschieden?**
  Variablen normalverteilt ⇒ **Barlett-Test**
  Variablen nicht normalverteilt ⇒ kein Test vorhanden

Nichtparametrische Tests

Kolmogorov-Smirnov-Anpassungstest

		Wir sprechen zu Hause über Themen aus der Schule.	Wir gehen mit der Familie oft nach draussen.	Wir besuchen mit der Familie Museen, Lehrpfade usw.	Wenn wir in den Ferien sind, schauen wir uns Sachen an diesem Ort an und unternehmen dort Ausflüge.	Ich gehe in Bibliotheken, Mediotheken und sehe mir Bücher und Filme zu Themen an oder leihe sie aus.	Ich sehe im Fernsehen Nachrichtens endungen.	Ich lese Tageszeitung en.
N		669	673	666	671	673	673	673
Parameter der Normalverteilung[a,b]	Mittelwert	2.51	2.08	1.53	2.73	1.86	2.51	2.63
	Standardabweichung	.737	.868	.660	.943	.938	.897	.952
Extremste Differenzen	Absolut	.255	.270	.343	.228	.262	.224	.213
	Positiv	.247	.270	.343	.165	.262	.224	.213
	Negativ	-.255	-.201	-.211	-.228	-.178	-.199	-.184
Kolmogorov-Smirnov-Z		6.695	7.005	8.852	5.917	6.788	5.817	5.528
Asymptotische Signifikanz (2-seitig)		.000	.000	.000	.000	.000	.000	.000

a. Die zu testende Verteilung ist eine Normalverteilung.

b. Aus den Daten berechnet.

$p < 0.05$ ⇒ Variabel zu mehr als 95% W'keit nicht normalverteilt

SPSS / PSPP: Eignung der Daten für Faktoranalyse (2)

- **Eignet sich die Itemauswahl (Test) für eine Faktoranalyse?**

KMO-Koeffizient < .50 ⇒ keine Faktoranalyse durchführen
⇒ Stichprobenumfang vergrössern
KMO-Koeffizient > .80 ⇒ Stichprobe gut bis sehr gut geeignet

Faktorenanalyse

KMO- und Bartlett-Test

Maß der Stichprobeneignung nach Kaiser-Meyer-Olkin.		.681
Bartlett-Test auf Sphärizität	Ungefähres Chi-Quadrat	552.239
	df	21
	Signifikanz nach Bartlett	.000

KMO > 0.6 ⇒
SuS-Stichprobe mässig geeignet

iO, sofern Daten einer Normalverteilung folgten

SPSS / PSPP: Eignung der Daten für Faktoranalyse (2)

- **Eigenen sie die einzelnen Items für eine Faktoranalyse?**

Anti-Image-Matrizen

		Wir sprechen zu Hause über Themen aus der Schule.	Wir gehen mit der Familie oft nach draussen.	Wir besuchen mit der Familie Museen, Lehrpfade usw.	Wenn wir in den Ferien sind, schauen wir uns Sachen an diesem Ort an und unternehmen dort Ausflüge.	Ich gehe in Bibliotheken, Mediotheken und sehe mir Bücher und Filme zu Themen an oder leihe sie aus.	Ich sehe im Fernsehen Nachrichtens endungen.	Ich lese Tageszeitung en.
Anti-Image-Korrelation	Wir sprechen zu Hause über Themen aus der Schule.	.784[a]	-.152	-.077	-.055	-.178	-.047	-.026
	Wir gehen mit der Familie oft nach draussen.	-.152	.744[a]	-.207	-.170	-.134	.085	-.113
	Wir besuchen mit der Familie Museen, Lehrpfade usw.	-.077	-.207	.720[a]	-.273	-.140	-.072	.042
	Wenn wir in den Ferien sind, schauen wir uns Sachen an diesem Ort an und unternehmen dort Ausflüge.	-.055	-.170	-.273	.748[a]	-.072	-.018	-.050
	Ich gehe in Bibliotheken, Mediotheken und sehe mir Bücher und Filme zu Themen an oder leihe sie aus.	-.178	-.134	-.140	-.072	.742[a]	.085	-.161
	Ich sehe im Fernsehen Nachrichtensendungen.	-.0					.488[a]	-.430
	Ich lese Tageszeitungen.	-.0					-.430	.560[a]

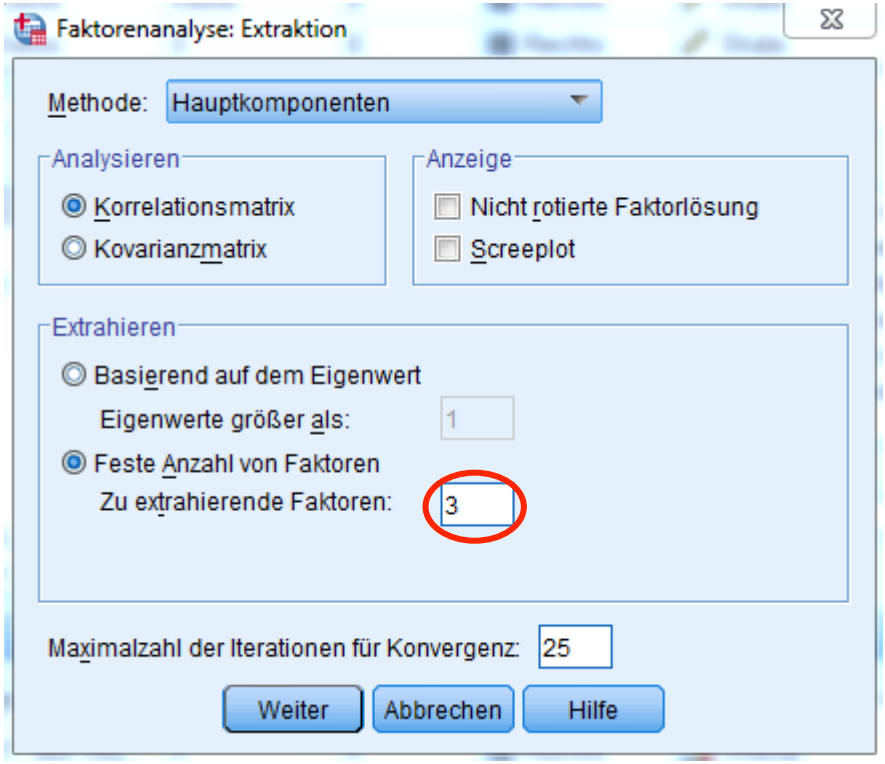
a. Maß der Stichprobeneignung

Item zeigt zu wenig gemeinsame Varianz mit anderen Items

PSPP unterstützt die Berechnung von Anti-Image-Matrizen nicht!

SPSS / PSPP: Anzahl Faktoren festlegen (1)

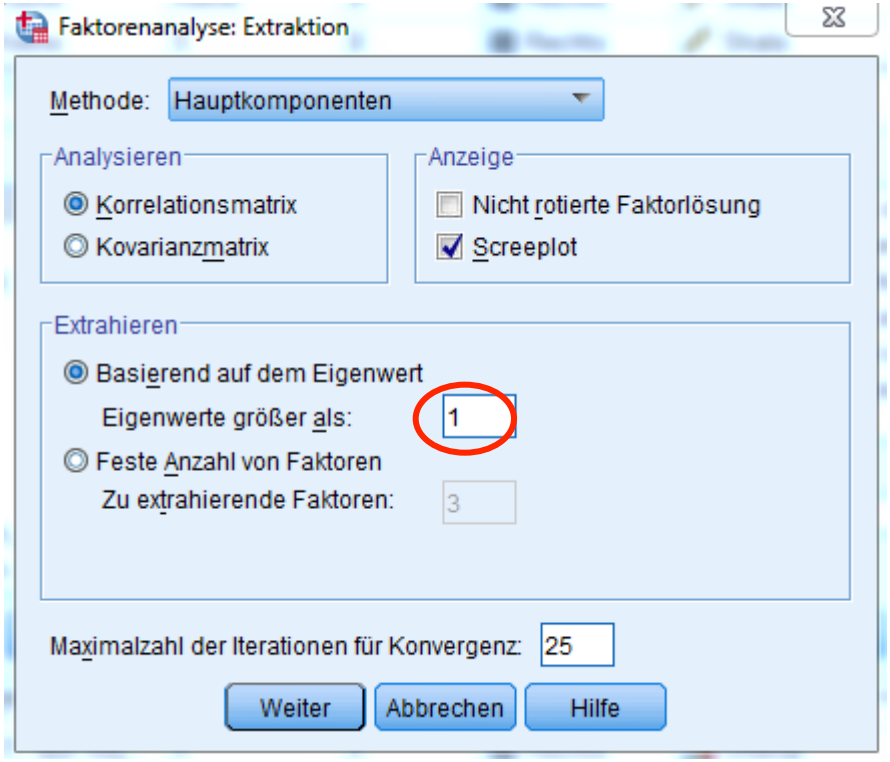
- **Interpretierbarkeit der Faktoren:**
  Wichtigster Massstab: Nur sinnvoll zu interpretierende Faktoren extrahieren.
- **Hypothetisches Modell**
  theoriebasierte Annahme ⇒ Voreinstellung fix wählen: mind. 4 Items pro Faktor



Voreinstellung: 3 Faktoren

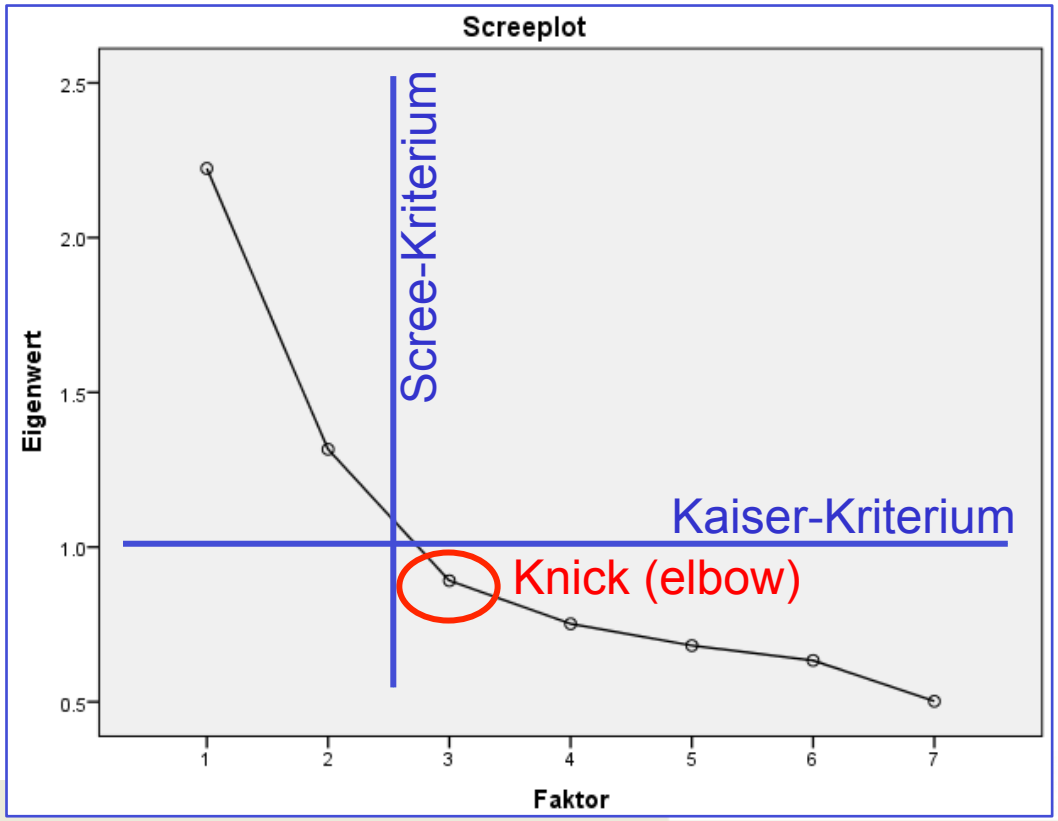
SPSS / PSPP: Anzahl Faktoren festlegen (2)

- **Kaiser-Kriterium: Eigenwert ≥ 1:**
  Eigenwert eines Faktors gibt an, von wie vielen Items die Varianz durch den Faktor erklärt wird.



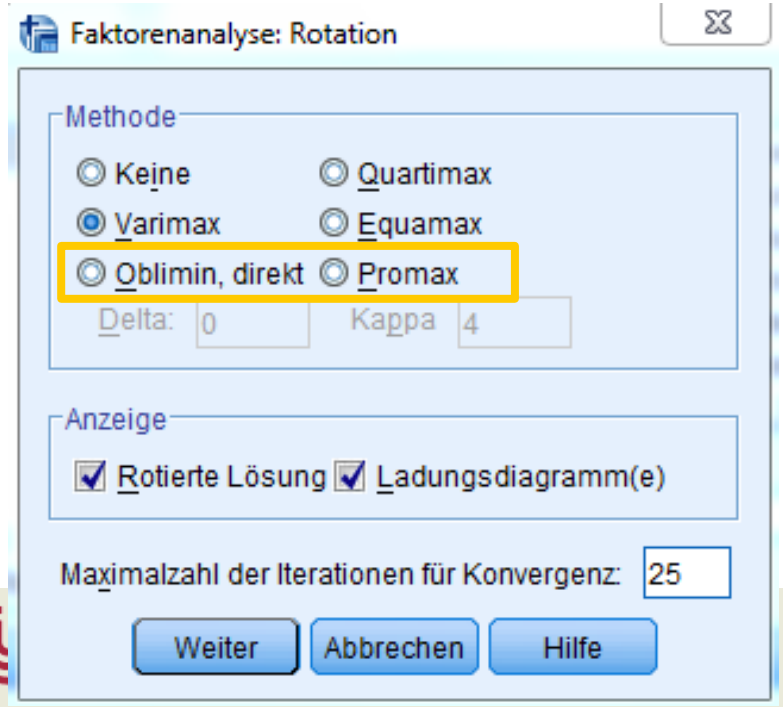
SPSS / PSPP: Anzahl Faktoren festlegen (3)

- **Scree-Plot:**
  Die Hauptkomponenten-Analyse rechnet mit gleichviel Faktoren wie Items.
  Der erste Faktor erklärt die grösstmögliche Varianz aller Items (Eigenwert).
  Der zweite Faktor erklärt den grösstmöglichen Teil der Restvarianz usw.
- **Scree-Kriterium:**
  Faktoren ab dem Punkt vernachlässigen, wo die Eigenwerte nicht mehr merklich abnehmen.



SPSS / PSPP: Rotationsmethode wählen

- **Wird erwartet, dass die Faktoren nicht korrelieren?**
  (Annahme für das Auffinden möglichst gut zu interpretierender Skalen bzw. Dimensionen)
  Ja ⇒ eine orthogonale Rotation wählen: am besten **Varimax-Methode**
- **Wird erwartet, dass die Faktoren korrelieren?**
  (Annahme für das Auffinden möglichst homogener Skalen)
  Ja ⇒ eine oblique Rotation wählen: am besten **Promax-Methode**



PSPP unterstützt die Berechnung von obliquen Rotationen nicht!

SPSS / PSPP: Faktoreninterpretation

- **Ladungen < 0.30 nicht anzeigen lassen**
  Items mit Ladung > 0.5 müssen für Interpretation des Faktors herangezogen werden
- **Keine Einfachstruktur: Mehrfachladungen in rotierter Faktorenmatrix!**

  korrelierte Faktoren ⇒ Items mit Mehrfachladungen (> .50) in Mustermatrix entfernen / Faktorzahl verändern

  unkorrelierte Faktoren ⇒ Items nur entfernen, wenn Inhaltsvalidität nicht leidet
- **Positive und negative Faktorladungen vorhanden!**
  ⇒ Itempolung checken

Mustermatrix[a]

	Komponente	
	1	2
Wir gehen mit der Familie oft nach draussen.	.711	
Wir besuchen mit der Familie Museen, Lehrpfade usw.	.701	
Wenn wir in den Ferien sind, schauen wir uns Sachen an diesem Ort an und unternehmen dort Ausflüge.	.651	
Ich gehe in Bibliotheken, Mediotheken und sehe mir Bücher und Filme zu Themen an oder leihe sie aus.	.611	
Wir sprechen zu Hause über Themen aus der Schule.	.544	
Ich sehe im Fernsehen Nachrichtensendungen.		.870
Ich lese Tageszeitungen.		.812

Extraktionsmethode: Hauptkomponentenanalyse.
Rotationsmethode: Promax mit Kaiser-Normalisierung.

a. Die Rotation ist in 3 Iterationen konvergiert.

SPSS / PSPP: Ergebnisse (1)

- **Faktorenanzahl:**
  Kriterium Eigenwert > 1
- **Extraktionsmethode:**
  Hauptachsen
- **Rotationsmethode:**
  Promax
  (geeignet für das Auffinden möglichst einfach zu interpretierender Skalen)

Faktorenmatrix[a]

	Faktor 1	Faktor 2
Wir sprechen zu Hause über Themen aus der Schule.	.411	-.087
Wir gehen mit der Familie oft nach draussen.	.557	-.212
Wir besuchen mit der Familie Museen, Lehrpfade usw.	.547	-.215
Wenn wir in den Ferien sind, schauen wir uns Sachen an diesem Ort an und unternehmen dort Ausflüge.	.504	-.161
Ich gehe in Bibliotheken, Mediotheken und sehe mir Bücher und Filme zu Themen an oder leihe sie aus.	.473	-.109
Ich sehe im Fernsehen Nachrichtensendungen.	.266	.529
Ich lese Tageszeitungen.	.466	.566

Extraktionsmethode: Hauptachsen-Faktorenanalyse.

a. Es wurde versucht, 2 Faktoren zu extrahieren. Es werden mehr als 25 Iterationen benötigt. (Konvergenz=.004). Die Extraktion wurde abgebrochen.

Mustermatrix[a]

	Faktor 1	Faktor 2
Wir sprechen zu Hause über Themen aus der Schule.	.406	.042
Wir gehen mit der Familie oft nach draussen.	.605	-.036
Wir besuchen mit der Familie Museen, Lehrpfade usw.	.598	-.042
Wenn wir in den Ferien sind, schauen wir uns Sachen an diesem Ort an und unternehmen dort Ausflüge.	.530	-.002
Ich gehe in Bibliotheken, Mediotheken und sehe mir Bücher und Filme zu Themen an oder leihe sie aus.	.473	.040
Ich sehe im Fernsehen Nachrichtensendungen.	-.074	.608
Ich lese Tageszeitungen.	.079	.707

Extraktionsmethode: Hauptachsen-Faktorenanalyse.
Rotationsmethode: Promax mit Kaiser-Normalisierung.

a. Die Rotation ist in 3 Iterationen konvergiert.

Maximum-Likelihood-Analyse verwenden ⇐

SPSS / PSPP: Ergebnisse (2)

- **Faktorenanzahl:**
  Kriterium Eigenwert > 1
- **Extraktionsmethode:**
  Maximum-Likelihood
- **Rotationsmethode:**
  Promax

Faktorenmatrix[a]

	Faktor 1	Faktor 2
Wir sprechen zu Hause über Themen aus der Schule.	.127	.391
Wir gehen mit der Familie oft nach draussen.	.166	.561
Wir besuchen mit der Familie Museen, Lehrpfade usw.	.095	.599
Wenn wir in den Ferien sind, schauen wir uns Sachen an diesem Ort an und unternehmen dort Ausflüge.	.135	.524
Ich gehe in Bibliotheken, Mediotheken und sehe mir Bücher und Filme zu Themen an oder leihe sie aus.	.199	.435
Ich sehe im Fernsehen Nachrichtensendungen.	.425	-.021
Ich lese Tageszeitungen.	.999	.000

Extraktionsmethode: Maximum-Likelihood.

a. 2 Faktoren extrahiert. Es werden 10 Iterationen benötigt.

Mustermatrix[a]

	Faktor 1	Faktor 2
Wir sprechen zu Hause über Themen aus der Schule.	.406	.017
Wir gehen mit der Familie oft nach draussen.	.582	.008
Wir besuchen mit der Familie Museen, Lehrpfade usw.	.621	-.073
Wenn wir in den Ferien sind, schauen wir uns Sachen an diesem Ort an und unternehmen dort Ausflüge.	.545	-.013
Ich gehe in Bibliotheken, Mediotheken und sehe mir Bücher und Filme zu Themen an oder leihe sie aus.	.453	.077
Ich sehe im Fernsehen Nachrichtensendungen.	-.015	.429
Ich lese Tageszeitungen.	.015	.996

Extraktionsmethode: Maximum-Likelihood.
Rotationsmethode: Promax mit Kaiser-Normalisierung.

a. Die Rotation ist in 3 Iterationen konvergiert.

Einfachstruktur

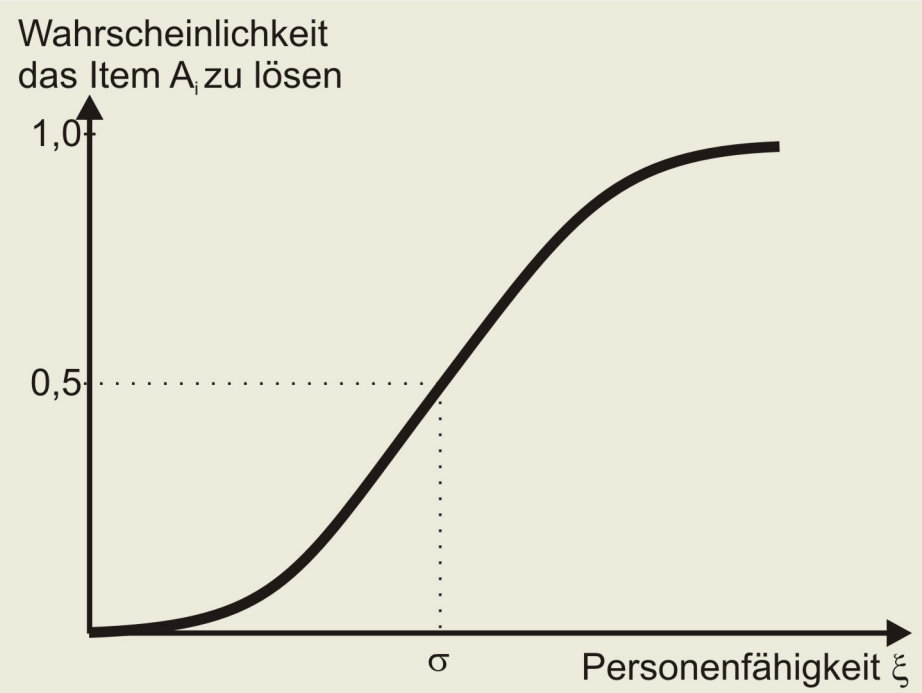
Übungsphase 2

Probabilistische Testtheorie

Andere Ansätze

- Idee: Die Wahrscheinlichkeit, dass eine Person ein Item löst, soll von zwei Faktoren abhängen:
  - Personenfähigkeit ξ
  - Itemschwierigkeit σ
- Guttmann
  - Wenn ξ > σ wird das Item sicher gelöst. (Deterministisches Modell)
- Linear
  - Die Wahrscheinlichkeit steigt linear mit der Personenfähigkeit

Itemcharakteristik



Eigenschaften Probabilistischer Modelle

- Abhängigkeit der Lösungswahrscheinlichkeit ausschließlich von Personenfähigkeit und Itemschwierigkeit

- Spezifische Objektivität: Es hängt nicht von den ausgewählten Items ab, welche Person als besser eingeschätzt wird (Missing Datas!), es hängt nicht von den Personen ab, welche Items schwieriger sind

- Lokale stochastische Unabhängigkeit: Die Lösung der Items ist unabhängig voneinander (Produkt LW der Items = LW des Tests), also nur vom zu erfassenden Konstrukt abhängig.
- Latente Variable wird konstant gehalten, manifeste Variablen dürfen nicht mehr korrelieren

- Die Leistung von Georg Rasch war, mathematisch bewiesen zu haben, dass die Gültigkeit seines Modells diese Eigenschaften nach sich zieht

Logistischer Zusammenhang

- Logistische Funktionen
- Rasch-Modell

$$p_{Lösung} = \frac{e^{F-S}}{1+e^{F-S}}$$

- Funktionaler Zusammenhang zwischen Itemschwierigkeit S und Personenfähigkeit F. P ist die Wahrscheinlichkeit, ein Item korrekt zu lösen.

Itemcharakteristik

- Wie kommt man aus den Daten zu den beiden Parametern?
  - Schätzung der Parameter mit Hilfe eines Computers. (ConceptMap, CONQUEST,...)
- Funktioniert mit allen Testdaten - es gibt immer ein Ergebnis!
- Wie überprüft man, ob das Modell mit seinen Parametern die beobachteten Daten gut beschreibt?

Vorgehen bei der Rasch-Skalierung

- In getrennten Verfahren werden Itemparameter (Maß für Schwierigkeit der Items) und Personenparameter (Maß für Fähigkeit der Personen) berechnet

- Dazu werden numerische Optimierungsverfahren verwendet: Maximierung der Wahrscheinlichkeit, dass die realen Daten auftreten unter der Annahme verschiedener Item-/Personenparameter

- Gültigkeit des Modells ergibt sich durch zwei Aspekte:
  - Itemhomogenität: Items messen dieselbe Fähigkeit und unterscheiden sich nur in Schwierigkeit, nicht in Trennschärfe
  - Personenhomogenität: Personen bearbeiten den Test aufgrund derselben Fähigkeit, d.h. Schwierigkeit der Items hängt nicht von Stichprobe ab

Überprüfung des Itemhomogenität

- Nach Optimierungsvorgang (verschiedene Schätzalgorithmen möglich) werden Itemparameter ausgegeben.
- Güte der Passung der Items zum Modell: Overfit/ Underfit
- Mittelwert auf 1 normiert
- Konvention: zwischen 0,8 und 1,2

All_1_3_shw.txt

TERM 1: item

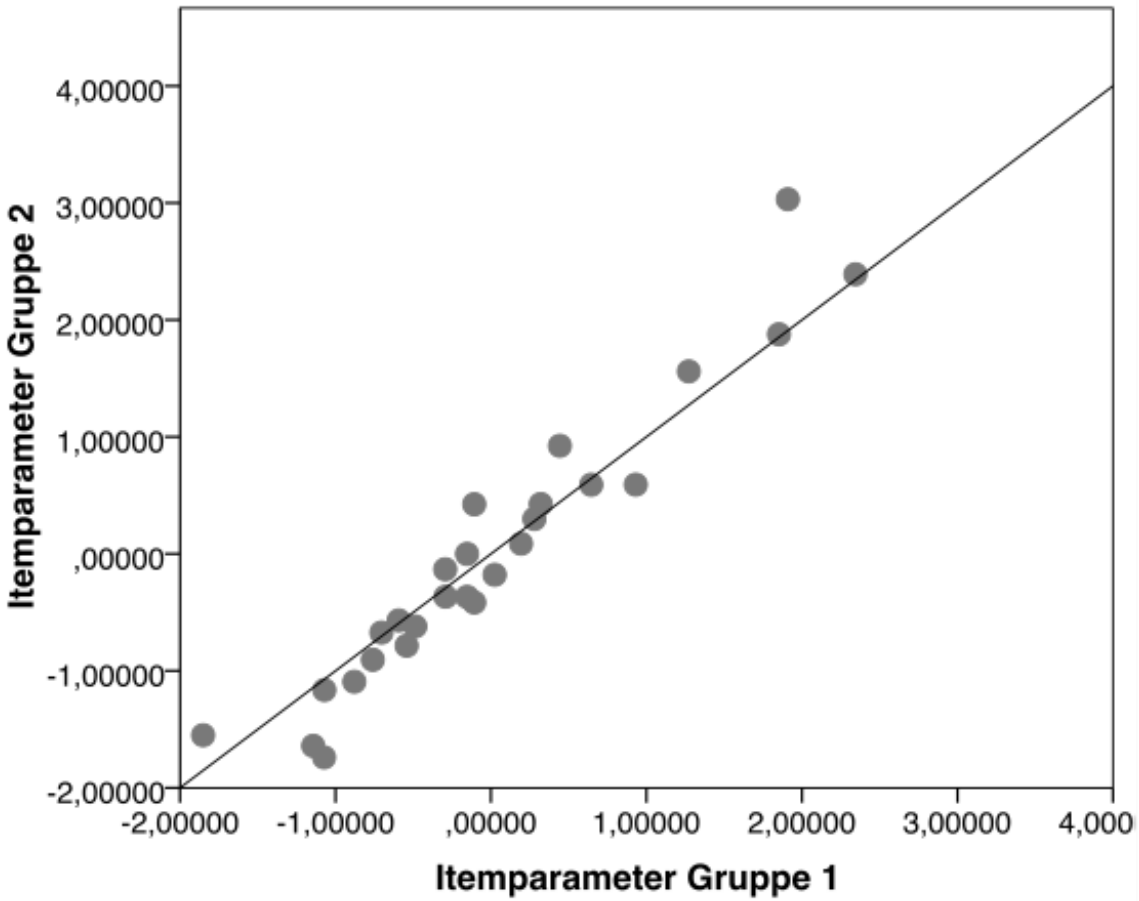
	VARIABLES			UNWEIGHTED FIT			WEIGHTED FIT		
	item	ESTIMATE	ERROR^	MNSQ	CI	T	MNSQ	CI	T
1	T1.1	0.558	0.122	1.16	(0.74, 1.26)	1.2	1.06	(0.81, 1.19)	0.6
2	T1.2	0.650	0.122	1.19	(0.74, 1.26)	1.4	1.15	(0.80, 1.20)	1.4
3	T1.4a	-0.619	0.119	0.94	(0.74, 1.26)	-0.4	0.96	(0.87, 1.13)	-0.6
4	T1.4b	-0.229	0.119	0.92	(0.74, 1.26)	-0.6	0.95	(0.87, 1.13)	-0.7
5	T1.8	-0.855	0.101	1.07	(0.82, 1.18)	0.8	1.05	(0.90, 1.10)	0.9
6	T3.2a	0.338	0.120	1.11	(0.74, 1.26)	0.9	1.13	(0.83, 1.17)	1.5
7	T3.2b	-0.385	0.119	1.05	(0.74, 1.26)	0.4	1.05	(0.87, 1.13)	0.7
8	T4.5	0.651	0.122	1.29	(0.74, 1.26)	2.1	1.14	(0.80, 1.20)	1.4
9	T5.3	0.558	0.122	1.02	(0.74, 1.26)	0.2	1.00	(0.81, 1.19)	0.1
10	T5.8	0.513	0.122	0.97	(0.74, 1.26)	-0.2	1.06	(0.81, 1.19)	0.6
11	T6.3	0.264	0.118	0.90	(0.76, 1.24)	-0.8	0.93	(0.86, 1.14)	-1.0
12	T6.4	1.261	0.112	1.36	(0.82, 1.18)	3.7	1.06	(0.81, 1.19)	0.7
13	T6.5a	1.521	0.128	0.75	(0.76, 1.24)	-2.2	0.96	(0.70, 1.30)	-0.2
14	T6.5b	-1.151	0.118	0.87	(0.76, 1.24)	-1.0	0.91	(0.84, 1.16)	-1.1
17	T6.8	-0.139	0.117	0.89	(0.76, 1.24)	-0.9	0.93	(0.88, 1.12)	-1.1
18	T6.9	1.429	0.129	1.21	(0.74, 1.26)	1.6	1.12	(0.68, 1.32)	0.7
19	W1.1	0.778	0.133	1.43	(0.69, 1.31)	2.5	1.02	(0.65, 1.35)	0.2
20	W1.10	-0.737	0.126	1.00	(0.69, 1.31)	0.1	1.00	(0.84, 1.16)	-0.0
21	W1.12	-0.031	0.119	0.94	(0.74, 1.26)	-0.4	0.95	(0.86, 1.14)	-0.7
22	W1.2	0.518	0.132	1.15	(0.69, 1.31)	1.0	1.03	(0.70, 1.30)	0.3
23	W1.3	0.360	0.131	1.01	(0.69, 1.31)	0.1	1.01	(0.73, 1.27)	0.1
24	W1.4	0.296	0.108	1.03	(0.81, 1.19)	0.4	1.00	(0.87, 1.13)	-0.0
25	W1.5	-0.561	0.127	0.93	(0.69, 1.31)	-0.4	0.98	(0.83, 1.17)	-0.3
26	W1.6	-0.142	0.075	1.06	(0.89, 1.11)	1.1	1.05	(0.94, 1.06)	1.4

Überprüfung der Personenhomogenität

- Nach Optimierungsvorgang (verschiedene Schätzalgorithmen möglich) werden Personenparameter ausgegeben.

- Im Prinzip könnte hier auch nach Over/ Underfit sortiert werden – dies verändert die Stichprobe

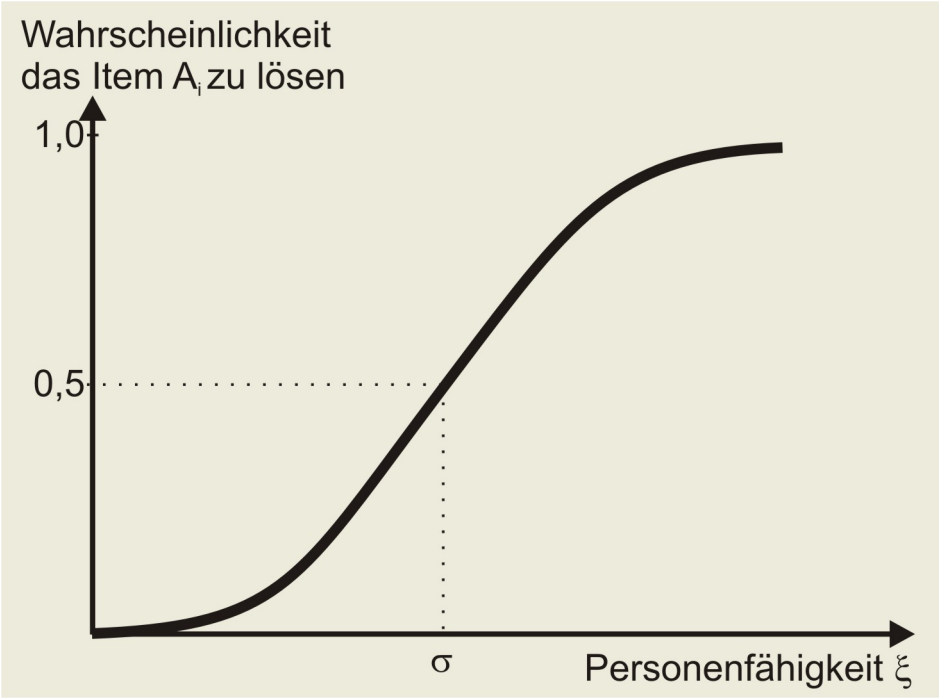
- Graphischer Modelltest: Stichprobe in zwei zufällige Hälften teilen und Itemparameter getrennt berechnen lassen



Skalenbildung

- Itemparameter und Personenparameter werden auf derselben Skala angegeben.
- Beides sind einheitenlose Werte (logarithmierte Verhältnisse von Wahrscheinlichkeiten)
- Gleichheit von Personen- und Itemparameter bedeutet, dass Lösungswahrscheinlichkeit 50 % entspricht.
- Sonderfall PISA-Skala: Mittelwert auf 50 Standardabweichung auf 100
- Metrisches Skalenniveau

$$p_{Lösung} = \frac{e^{F-S}}{1+e^{F-S}}$$



Weitere Vorteile Rasch-skalierter Tests

- Item- und Personenparameter sind vergleichbar auf derselben interpretierbaren Skala - Kompetenzstufen?
- Die Anzahl gelöster Items ist eine erschöpfende Statistik für die Fähigkeitsparameter der Person

- Es kann gezeigt werden, in welchen Items/ Personengruppen ein Instrument nicht Rasch-konform ist und somit das Konstrukt nicht erhebt

- Fähigkeitsanalyse durch Annahme mehrerer Rasch-Modelle (Dimensionen)
  - Z.B. Bayes Information Criterion: Verrechnung von Likelihood des Modells und Anzahl der Modellparameter

Literaturempfehlungen

- Krüger, Parchmann, Schecker (2014, Hrsg.). Methoden in der naturwissenschaftsdidaktischen Forschung. Berlin: Springer.
  - enthält u.a. zu KTT und PTT Kapitel mit praktischen Anwendungsbeispielen

- Bühner (2006): Einführung in die Test- und Fragebogenkonstruktion. München: Pearson.
  - eher praktisch orientiert, sehr anschaulich

- Rost (2004). Lehrbuch Testtheorie - Testkonstruktion. Bern: Huber.
  - sehr detailliert und auf PTT konzentriert

Übungsphase 3