Inhaltswissen modellieren

1. Ein volles Wasserglas steht stabil auf der Rückbank eines konstant geradeaus fahrenden Autos. Plötzlich tritt der Fahrer das Gaspedal durch und beschleunigt das Auto. Welche der folgenden Aussagen treffen zu?

- ☐ Weil sich das Glas bezüglich der Rückbank im Auto nicht bewegt, bleibt die Wasseroberfläche unverändert.
- ☐ Das Wasser wird mit dem Auto beschleunigt, so dass etwas Wasser in Fahrtrichtung über den Rand des Glases schwappt.
- ☐ Aufgrund der Trägheit des Wassers verändert sich die Wasseroberfläche nicht.
- ☐ Das Wasser behält zunächst seinen vorherigen Bewegungszustand bei, so dass etwas Wasser entgegen der Fahrtrichtung über den Rand des Glases schwappt.

3. Ein Bus fährt mit konstanter Geschwindigkeit auf horizontaler Strasse geradeaus. Welche der folgenden Aussagen treffen zu?

- ☐ Damit der Bus nicht langsamer wird, muss die Antriebskraft des Motors genau so gross sein wie der Luftwiderstand und die übrigen Reibungskräfte zusammen.
- ☐ Damit die Geschwindigkeit konstant bleibt, muss die Antriebskraft grösser sein als der Luftwiderstand und die übrigen Reibungskräfte zusammen.
- ☐ Damit die Geschwindigkeit nicht zunimmt, muss die Antriebskraft etwas geringer sein als der Luftwiderstand und die übrigen Reibungskräfte zusammen.
- ☐ Die Antriebskraft ist nur zum Beschleunigen erforderlich, bei konstanter Geschwindigkeit hingegen nicht.

basic Mechanics Conceptual Understanding Test (Hofer, 2015)

8. Somebody stands at the back of a stationary boat and throws a big stone horizontally with great momentum towards the back into the water. Which of the following statements are true?

- ☐ The boat moves in the direction of the stone that was thrown.
- ☐ The stone displaces water and therefore the boat only rocks slightly sideways.
- ☐ In principle, the same thing is happening when the nozzle of an inflated balloon is opened, and the balloon is whizzing through the air.
- ☐ The boat moves opposite to the direction of the throw.

11. The following three balls roll on a horizontal plane:

- Ball A rolls with velocity 1 m/s around a bend.
- Ball B starts with a velocity of 6 m/s, then its velocity continuously decreases.
- Ball C moves with an ever increasing velocity.

Which of the following statements are true?

- ☐ A horizontal force acts on ball A.
- ☐ A horizontal force acts on ball B.
- ☐ A horizontal force acts on ball C.

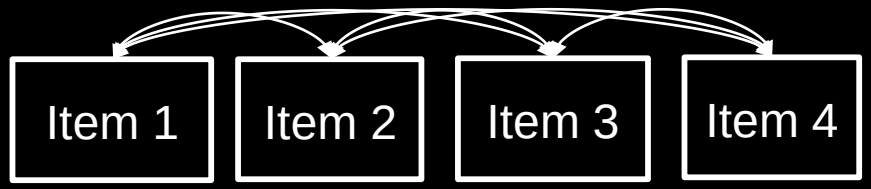
Option 1:

Summenscore/
Mittelwert

Intervention/Prä-
Post Veränderung

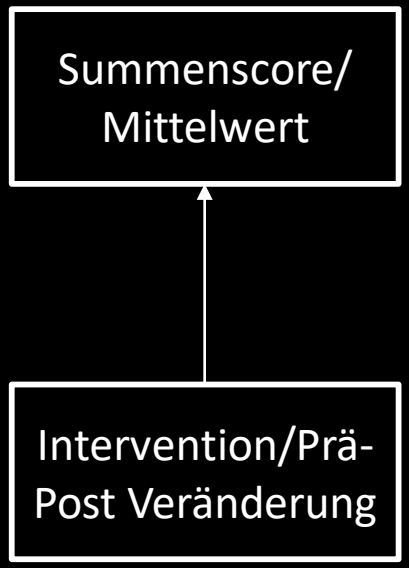
Option 2:

Faktoren
-analyse

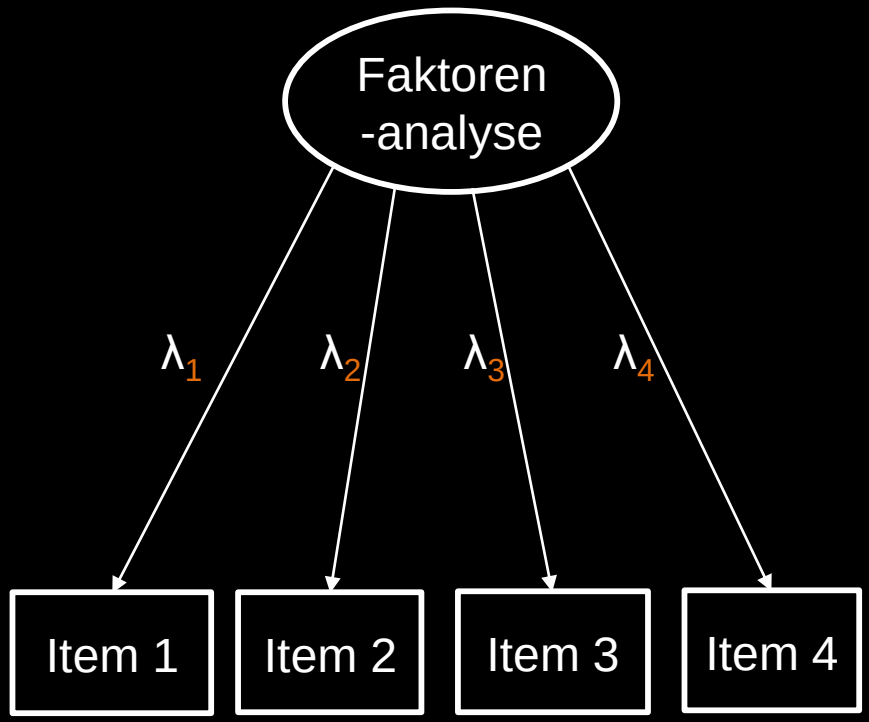


Ermittlung der Anzahl benötigter *Quellen gemeinsamer Varianz* (Faktoren),
um Interkorrelationen zwischen Items zu erklären/modellieren

Option 1:



Option 2:



Ermittlung der Anzahl benötigter *Quellen gemeinsamer Varianz* (Faktoren),
um Interkorrelationen zwischen Items zu erklären/modellieren

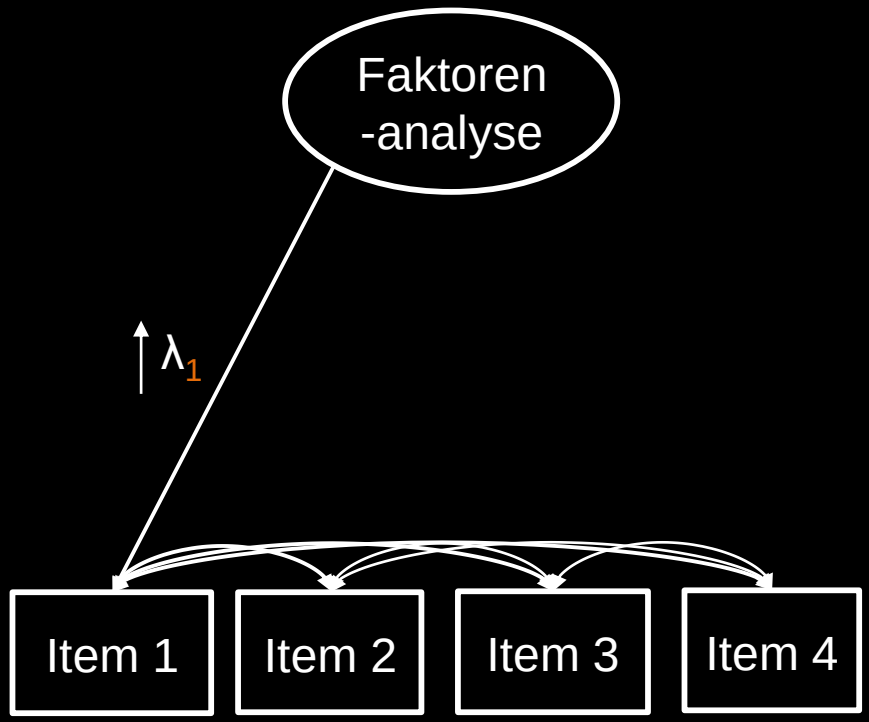
Schätzung der Stärke, mit welcher die gemeinsame Varianz
In jedes Item eingeht (Faktorladung λ_1-λ_4)

Option 1:

Summenscore/
Mittelwert

Intervention/Prä-
Post Veränderung

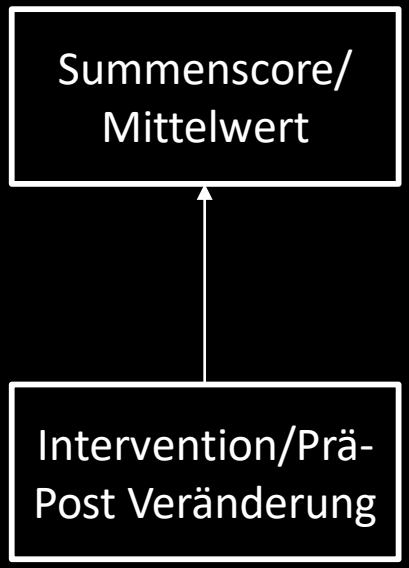
Option 2:



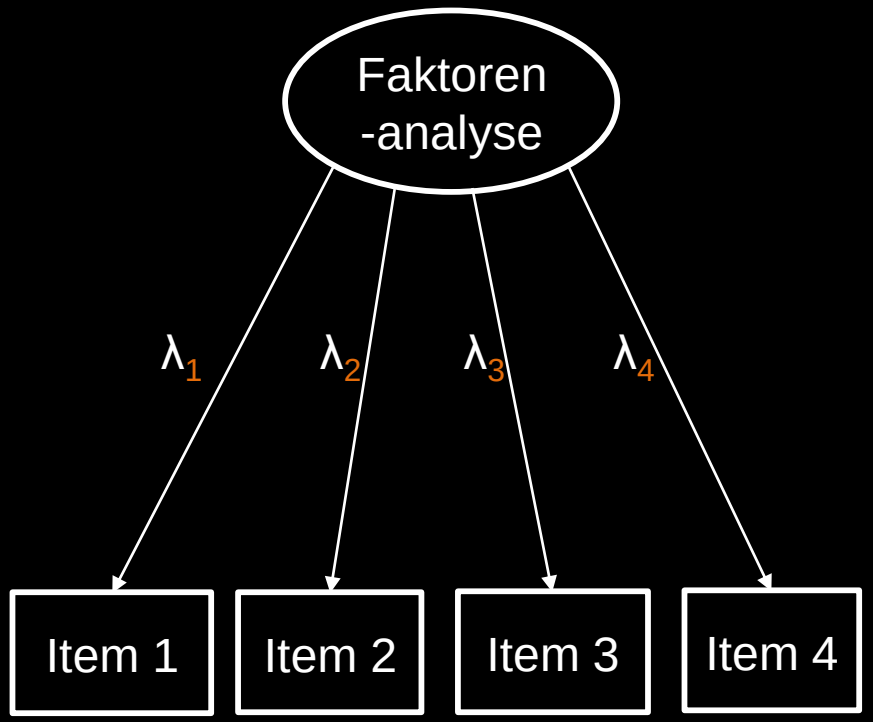
Items, die mit den anderen Items *hohe Interkorrelationen* aufweisen,
Erhalten *hohe Faktorladungen* (starke Indikatoren des g. Konstruktes)

Schätzung der Stärke, mit welcher die gemeinsame Varianz
In jedes Item eingeht (Faktorladung λ_1-λ_4)

Option 1:



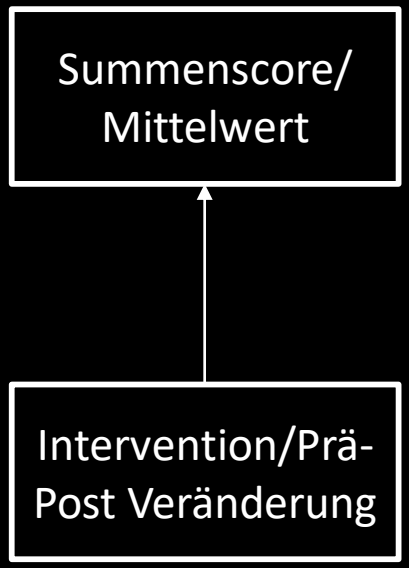
Option 2:



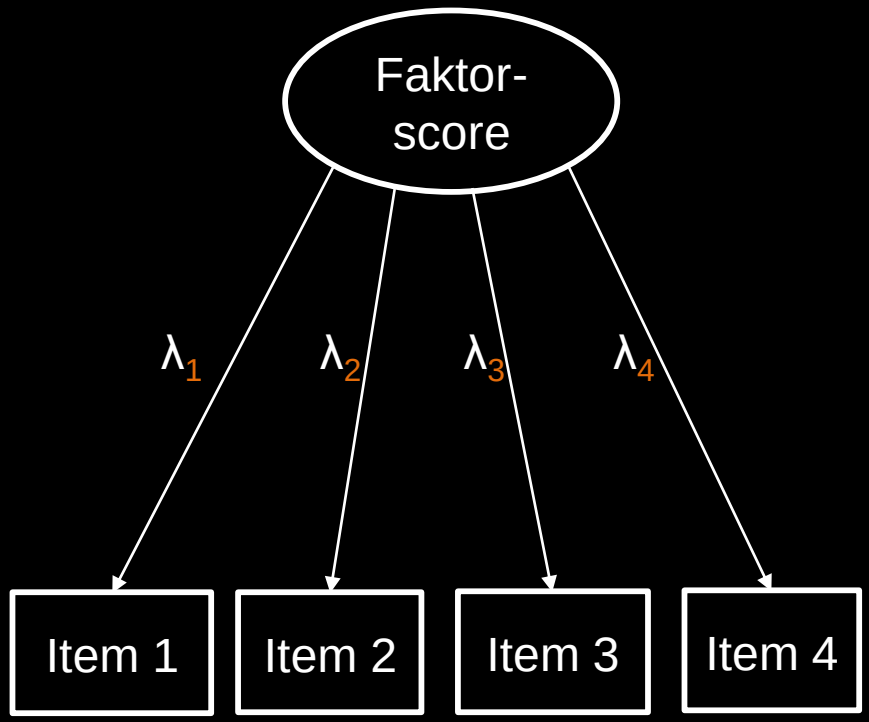
Als Messmodell:
Untersuchen, ob *die theoretisch erwartete Faktorenstruktur* (Anzahl & Ladungen) vorliegt

Als Skalierungsmodell:
Aus dem Messmodell werden geschätzte Messwerte gebildet

Option 1:



Option 2:



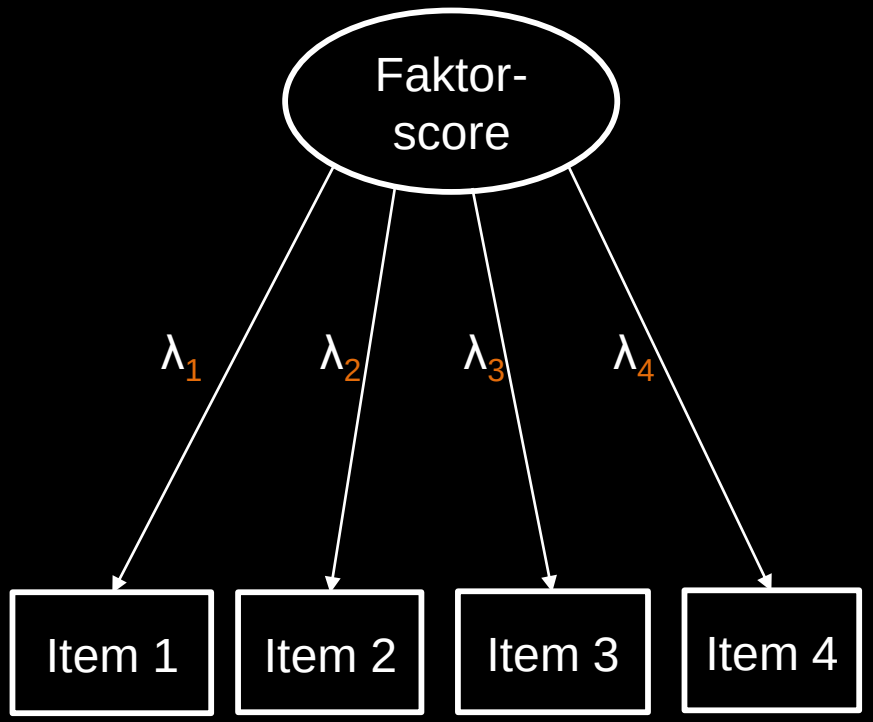
Faktorscore:
Nach Faktorladung gewichteter Wert aus allen Items

Als Skalierungsmodell:
Aus dem Messmodell werden geschätzte Messwerte gebildet

Option 1:

Summenscore/
Mittelwert

Option 2:



› Psychol Methods. 2022 Apr;27(2):234-260. doi: 10.1037/met0000367. Epub 2020 Oct 22.

Avoiding bias from sum scores in growth estimates: An examination of IRT-based approaches to scoring longitudinal survey responses

Megan Kuhfeld [1], James Soland [2]

Affiliations + expand

PMID: 33090818 DOI: 10.1037/met0000367

Abstract

A huge portion of what we know about how humans develop, learn, behave, and interact is based on survey data. Researchers use longitudinal growth modeling to understand the development of students on psychological and social-emotional learning constructs across elementary and middle school. In these designs, students are typically administered a consistent set of self-report survey items across multiple school years, and growth is measured either based on sum scores or scale scores produced based on item response theory (IRT) methods. Although there is great deal of guidance on scaling and linking IRT-based large-scale educational assessment to facilitate the estimation of examinee growth, little of this expertise is brought to bear in the scaling of

Behavior Research Methods (2020) 52:2287–2305
https://doi.org/10.3758/s13428-020-01398-0

Thinking twice about sum scores

Daniel McNeish [1] · Melissa Gordon Wolf [2]

Published online: 22 April 2020

Abstract

A common way to form scores from multiple-item scales is to sum responses of all items. Though sum scoring is often contrasted with factor analysis as a competing method, we review how factor analysis and sum scoring both fall under the larger umbrella of latent variable models, with sum scoring being a constrained version of a factor analysis. Despite similarities, reporting of psychometric properties for sum scored or factor analyzed scales are quite different. Further, if researchers use factor analysis to validate a scale but subsequently sum score the scale, this employs a model that differs from validation model. By framing sum scoring within a latent variable framework, our goal is to raise awareness that (a) sum scoring requires rather strict constraints, (b) imposing these constraints requires the same type of justification as any other latent variable model, and (c) sum scoring corresponds to a statistical model and is not a model-free arithmetic calculation. We discuss how unjustified sum scoring can have adverse effects on validity, reliability, and qualitative classification from sum score cut-offs. We also discuss considerations for how to use scale scores in subsequent analyses and how these choices can alter conclusions. The general goal is to encourage researchers to more critically evaluate how they obtain, justify, and use multiple-item scale scores.

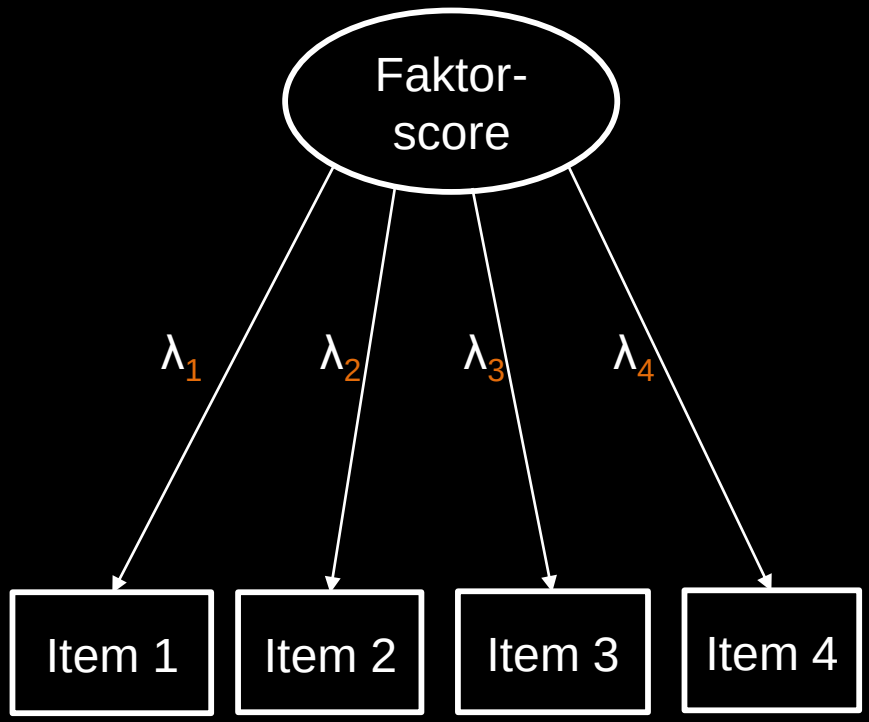
Kritik des *Summenscore + Alpha*
(+ Faktorenanalyse) - Ansatzes

Faktorscore:
Nach Faktorladung gewichteter Wert aus allen Items

Option 1:

Summenscore/
Mittelwert

Option 2:



Behavior Research Methods (2020) 52:2287–2305
https://doi.org/10.3758/s13428-020-01398-0

Check for updates

Thinking twice about sum scores

Daniel McNeish[1] · Melissa Gordon Wolf[2]

Published online: 22 April 2020

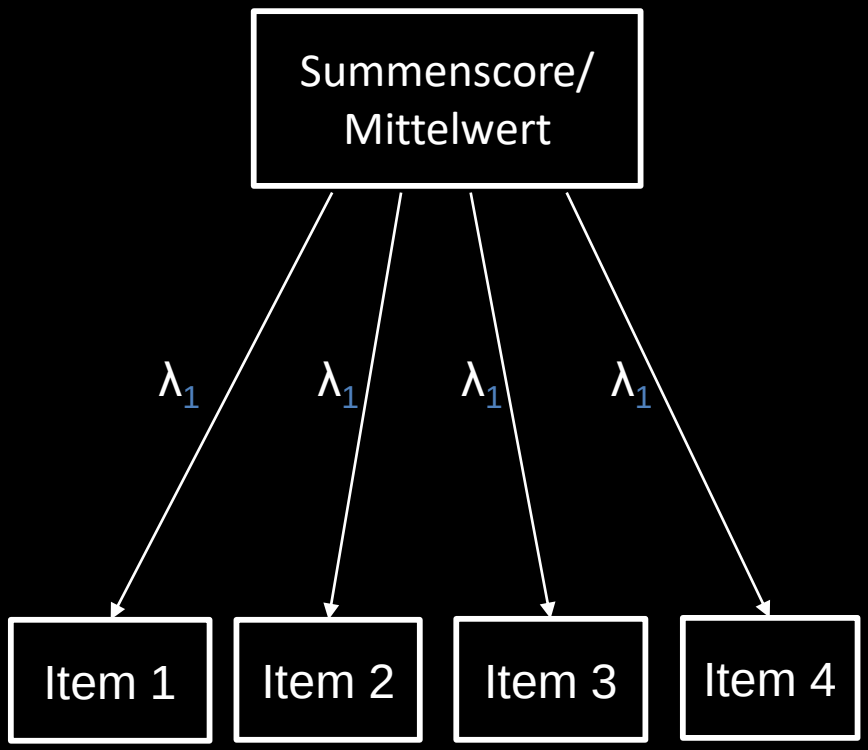
Abstract

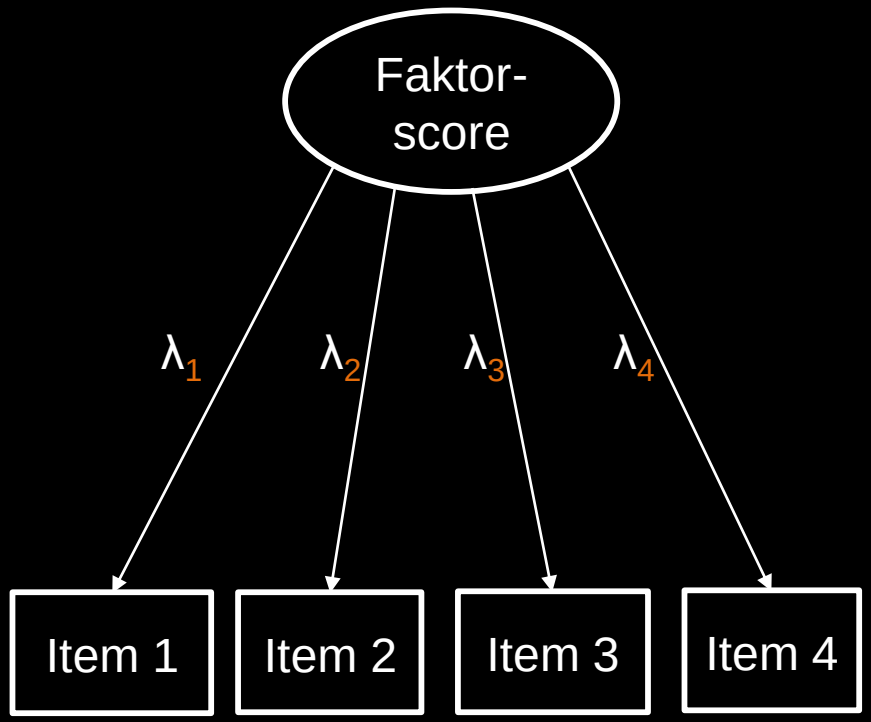
A common way to form scores from multiple-item scales is to sum responses of all items. Though sum scoring is often contrasted with factor analysis as a competing method, we review how factor analysis and sum scoring both fall under the larger umbrella of latent variable models, with sum scoring being a constrained version of a factor analysis. Despite similarities, reporting of psychometric properties for sum scored or factor analyzed scales are quite different. Further, if researchers use factor analysis to validate a scale but subsequently sum score the scale, this employs a model that differs from validation model. By framing sum scoring within a latent variable framework, our goal is to raise awareness that (a) sum scoring requires rather strict constraints, (b) imposing these constraints requires the same type of justification as any other latent variable model, and (c) sum scoring corresponds to a statistical model and is not a model-free arithmetic calculation. We discuss how unjustified sum scoring can have adverse effects on validity, reliability, and qualitative classification from sum score cut-offs. We also discuss considerations for how to use scale scores in subsequent analyses and how these choices can alter conclusions. The general goal is to encourage researchers to more critically evaluate how they obtain, justify, and use multiple-item scale scores.

The sum score is a *constrained version* of factor analysis
(McNeish & Wolf, 2020)

Faktorscore:
Nach Faktorladung gewichteter Wert aus allen Items

Option 1 = Option 2:





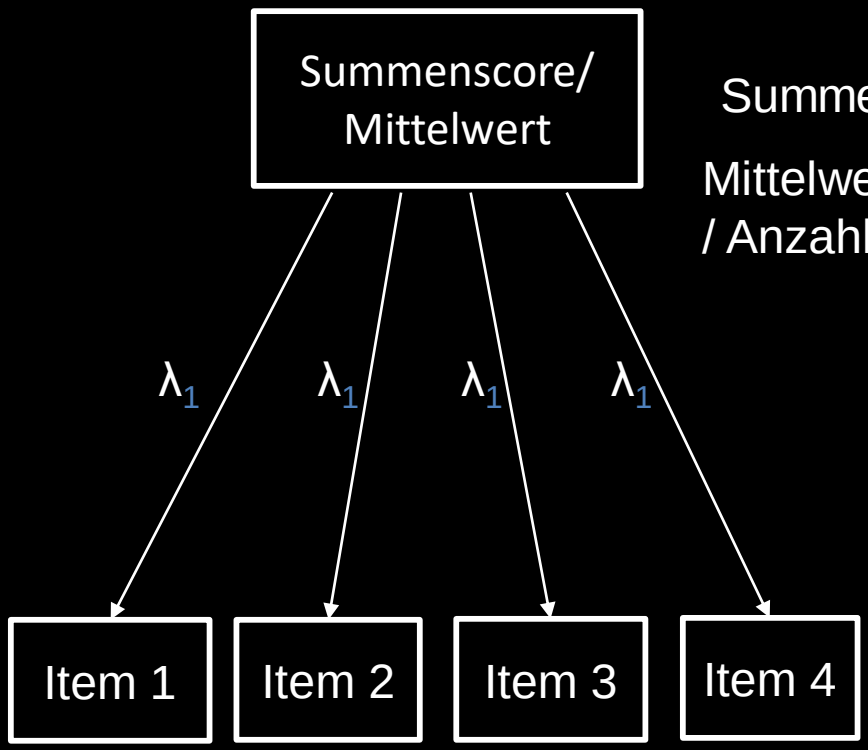
The sum score is a *constrained version* of factor analysis
(McNeish & Wolf, 2020)

Faktorscore:
Nach Faktorladung gewichteter Wert aus allen Items

Alle Items werden *gleich gewichtet*

Entspricht implizit der Annahme gleicher (std.) Faktorladungen

Option 1



Summe = 1*Item 1 + 1*Item 2 + 1*Item 3 + 1*Item 4

Mittelwert = (1*Item 1 + 1*Item 2 + 1*Item 3 + 1*Item 4) / Anzahl Items

0. Summenscore entspricht der Annahme gleicher Faktorladungen

1. Diese Annahme ist überprüfbar mittels Faktorenanalyse

2. Empirisch hält die Annahme (fast) nie

3. Schlussfolgerung:
Anstatt Summenscores/Mittelwerten sollten Faktorscores oder SEM verwendet werden (IRT/Personenschätzer)

The sum score is a *constrained version* of factor analysis
(McNeish & Wolf, 2020)

Alle Items werden *gleich gewichtet*

Entspricht implizit der Annahme gleicher (std.) Faktorladungen

Was denkt *ihr*?

Sollten wir anstatt Summenscores/Mittelwerte lieber generell Faktorscores/Struckturgleichungsmodellierung/Item Response Theorie-Schätzer verwenden?

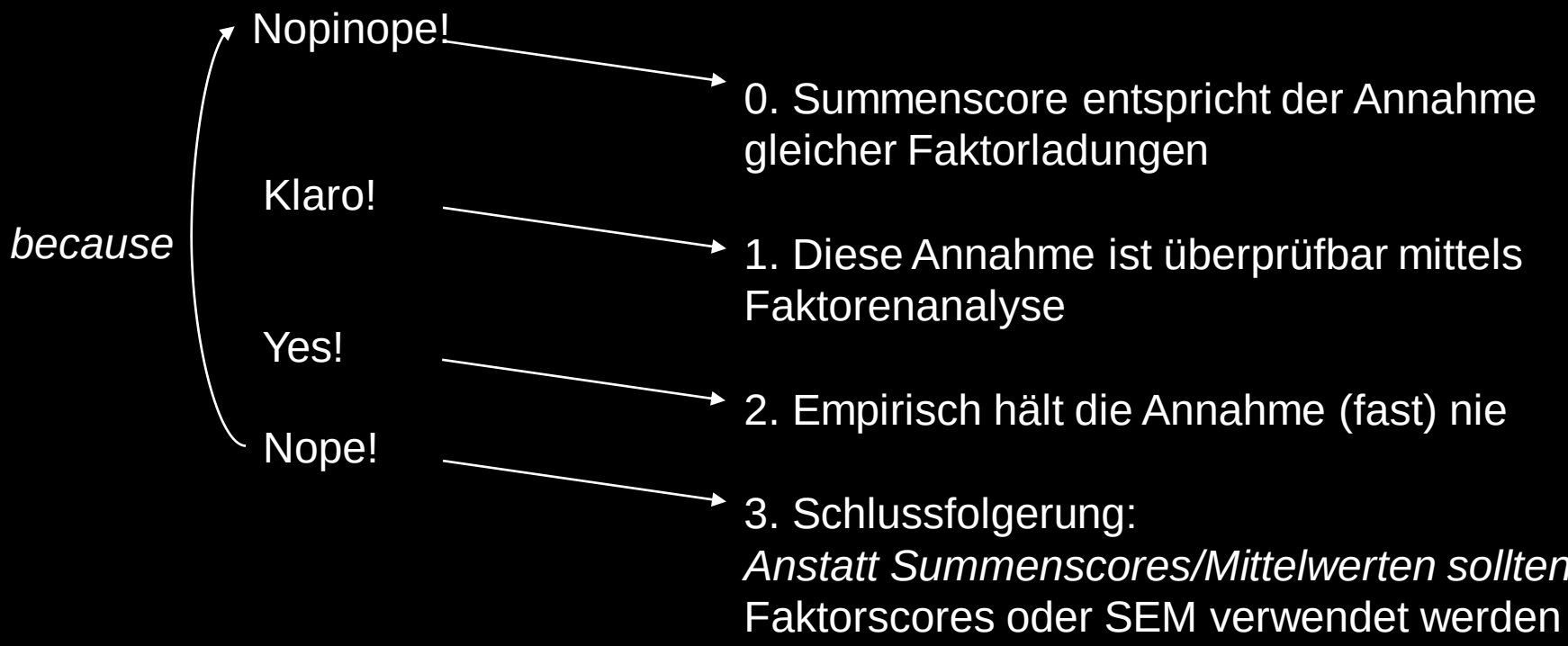
0. Summenscore entspricht der Annahme gleicher Faktorladungen

1. Diese Annahme ist überprüfbar mittels Faktorenanalyse

2. Empirisch hält die Annahme (fast) nie

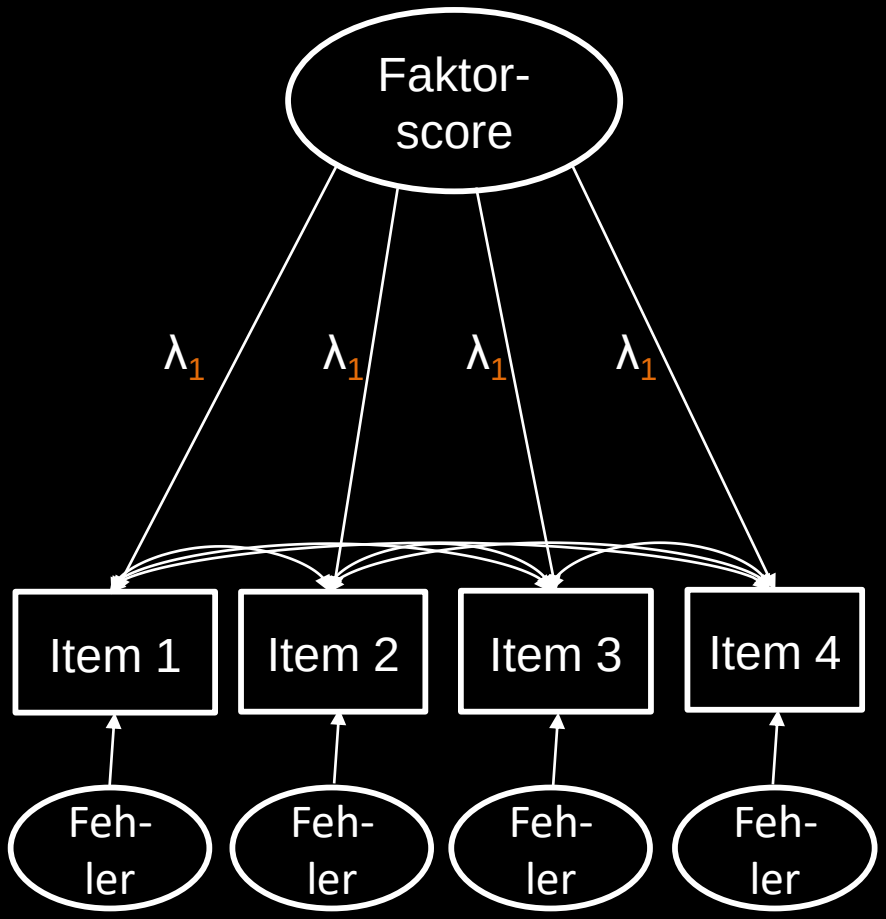
3. Schlussfolgerung:
Anstatt Summenscores/Mittelwerten sollten Faktorscores oder SEM verwendet werden

Das denk ich


Nopinope!
Klaro!
Yes!
Nope!
because
0. Summenscore entspricht der Annahme gleicher Faktorladungen
1. Diese Annahme ist überprüfbar mittels Faktorenanalyse
2. Empirisch hält die Annahme (fast) nie
3. Schlussfolgerung:
Anstatt Summenscores/Mittelwerten sollten
Faktorscores oder SEM verwendet werden

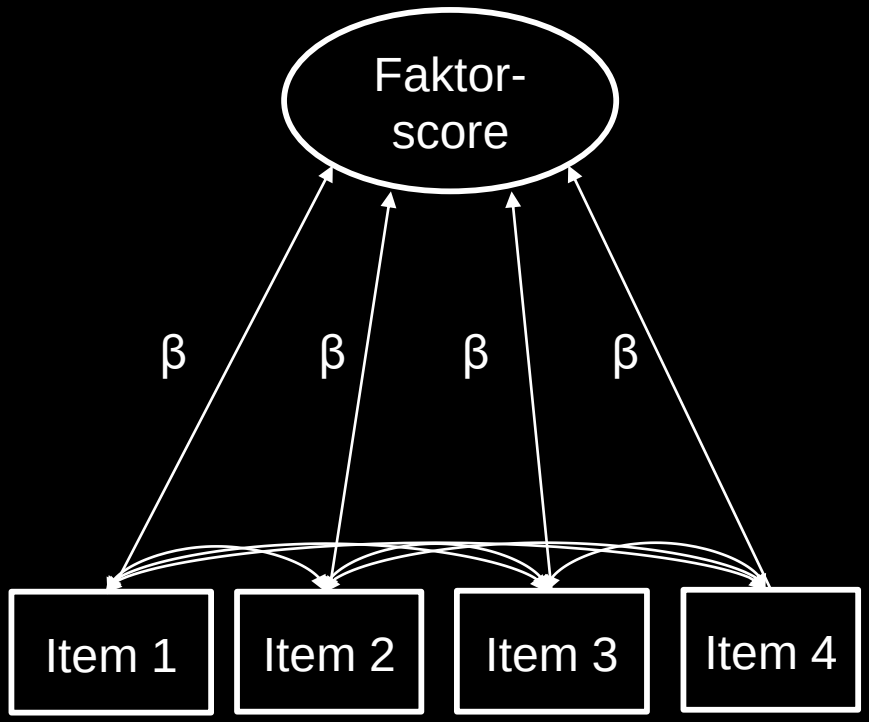
Das sag ich

0. Summenscore entspricht der Annahme **eines latenten Faktormodells** mit gleichen Faktorladungen



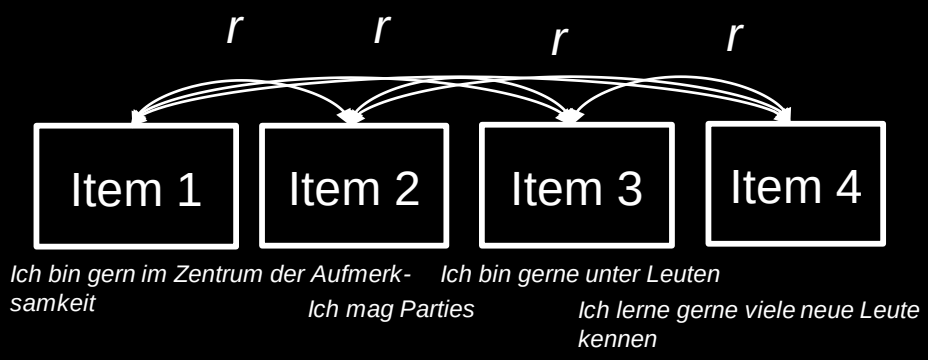
Das sag ich

0. Summenscore entspricht der Annahme **eines Kompositmodells** mit gleichen Faktorladungen

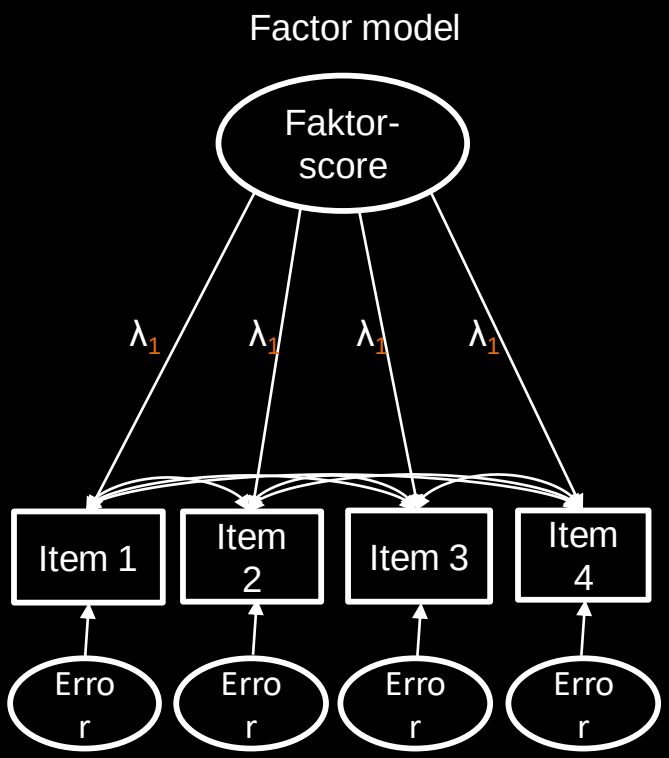


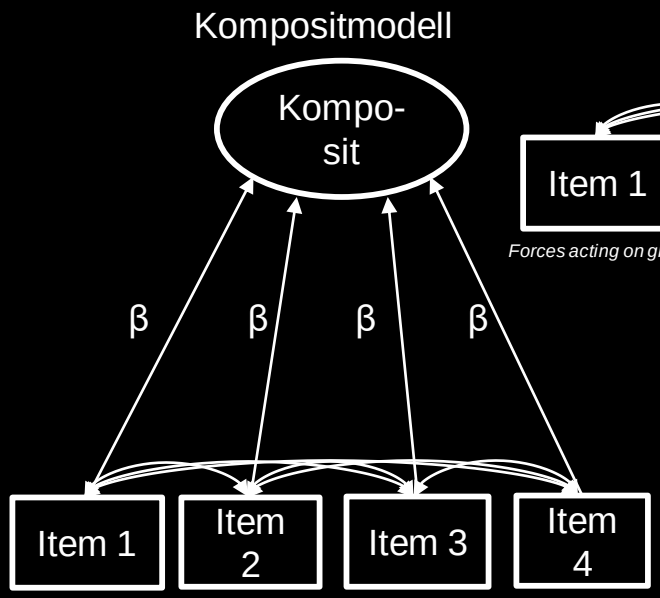
Das sag ich

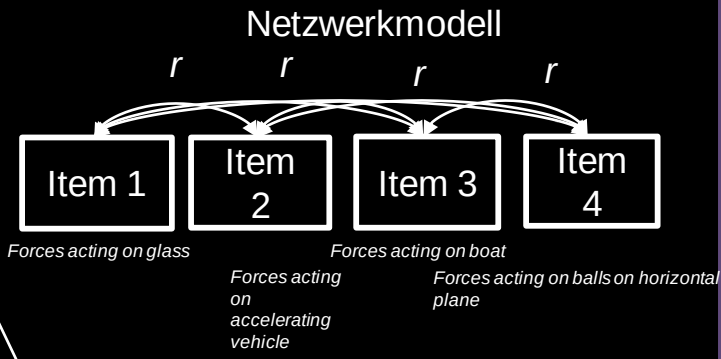
0. Summenscore entspricht der Annahme **eines Netzwerkmodells** mit gleichen Partialkorrelationen



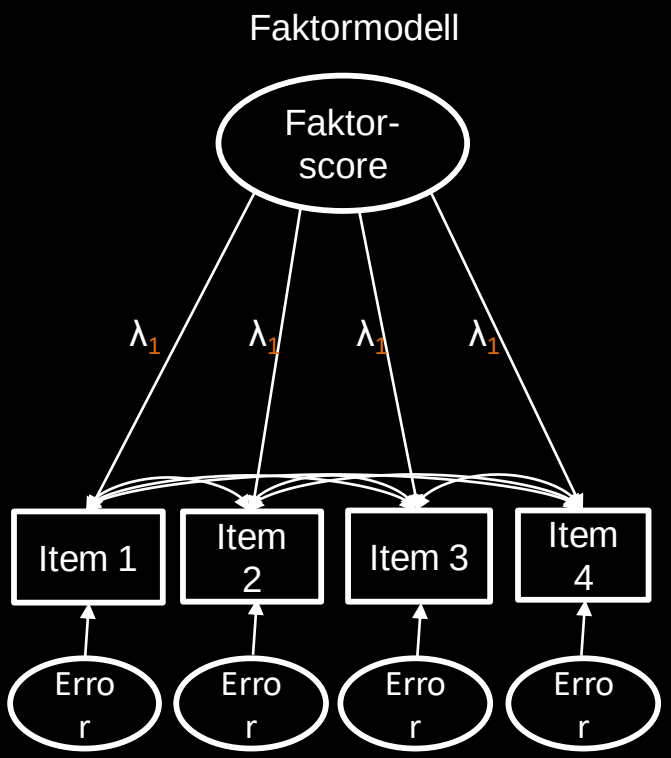
Empirisch (fast) nicht unterscheidbar (Edelsbrunner, 2022 für Referenzen)

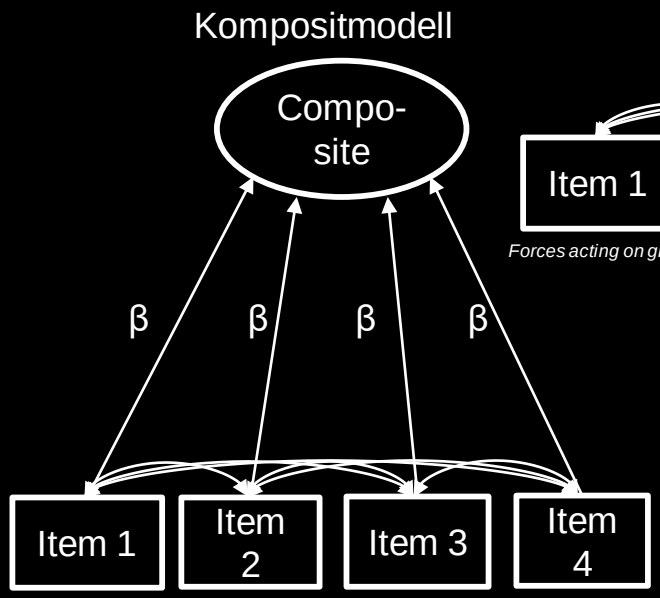


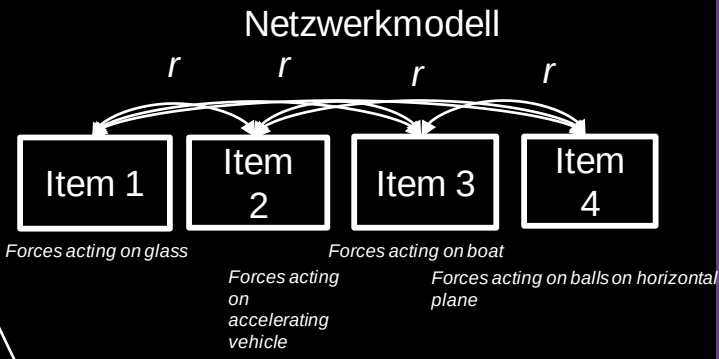




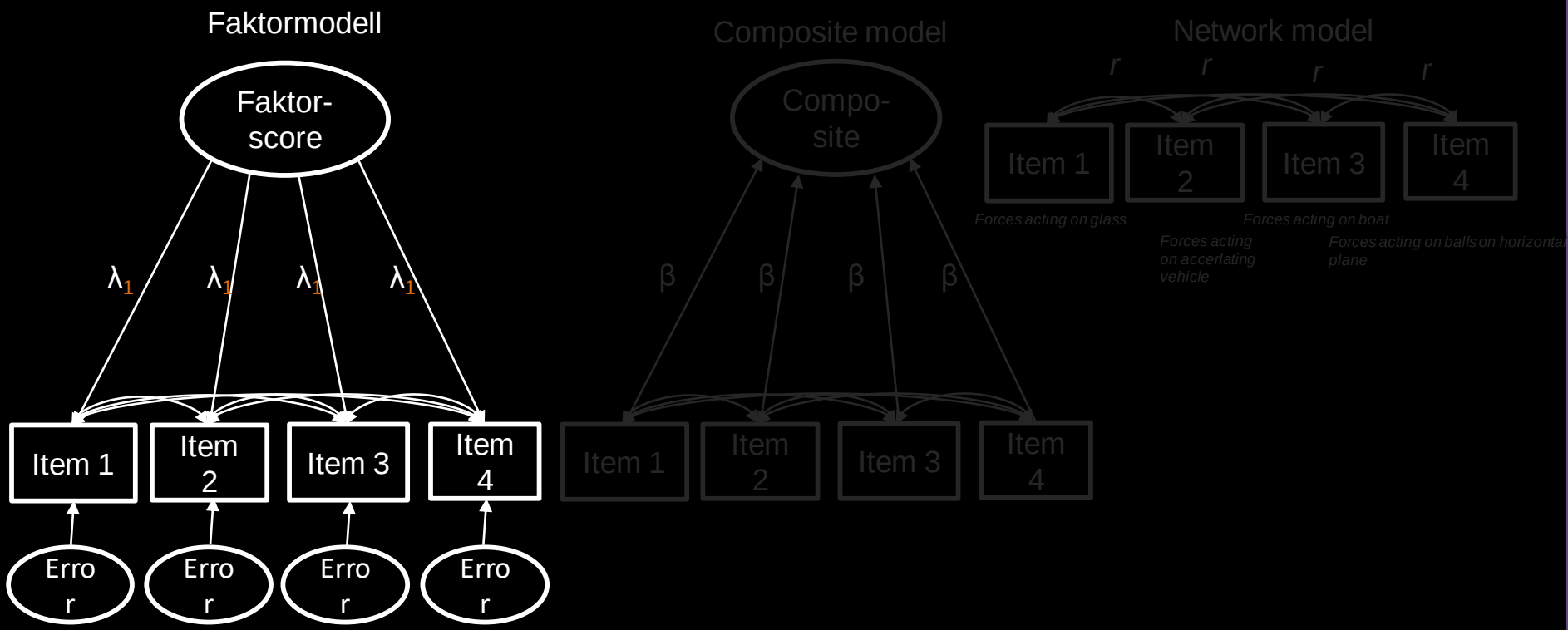
Theoretische Unterscheidung möglich (Edelsbrunner, 2022)



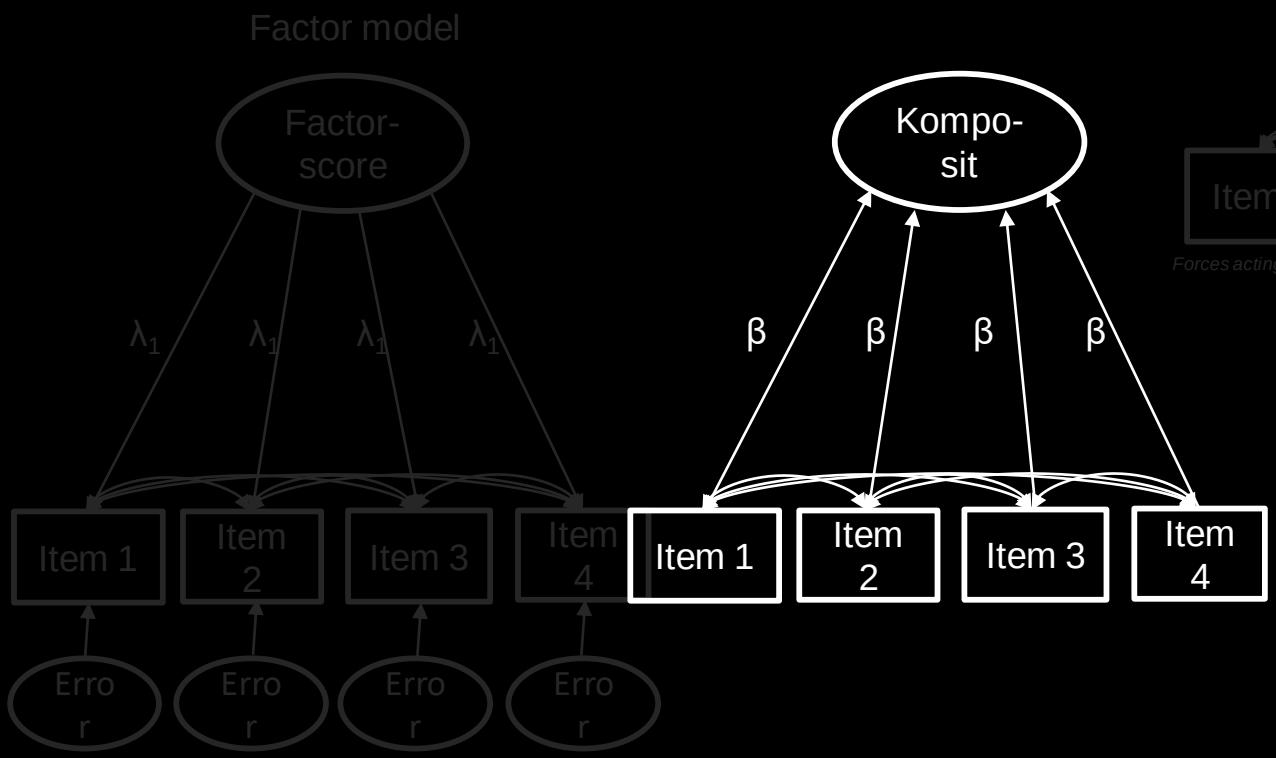


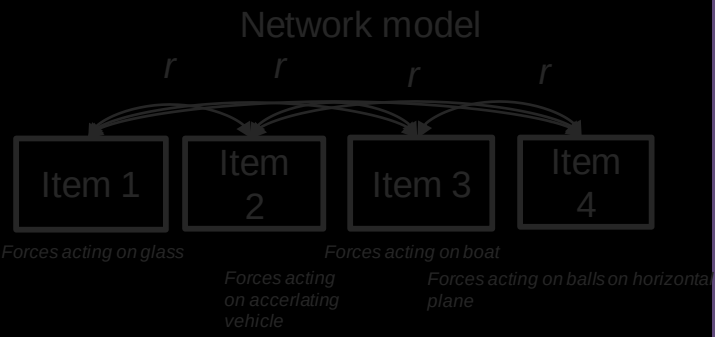


Korrekt wenn die Items *austauschbare Indikatoren* desselben Konstruktes sind

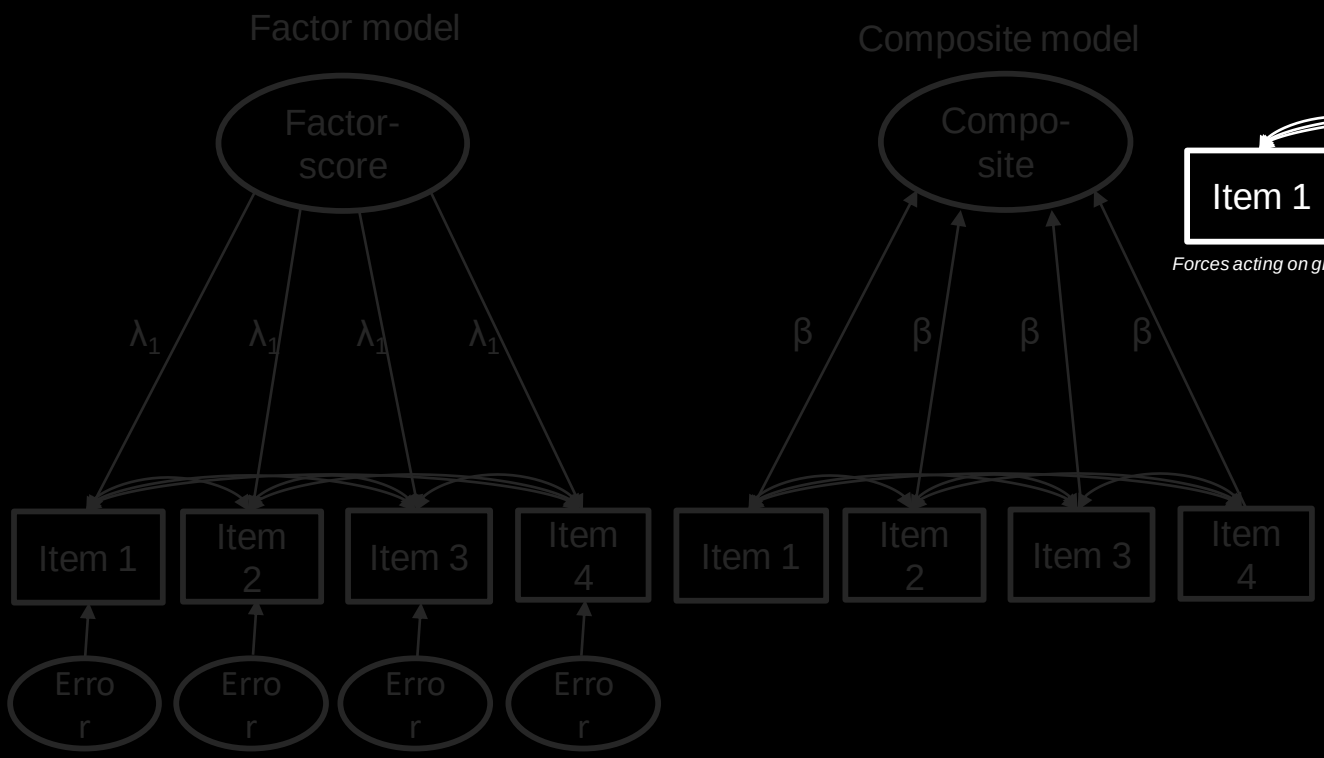


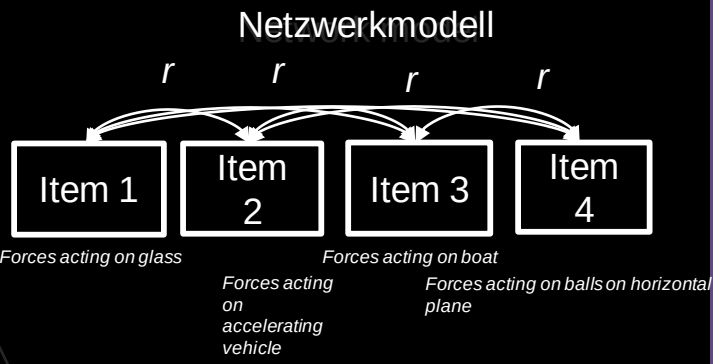
Korrekt wenn *jedes Item einen wichtigen Teil* des Konstruktes beisteuert



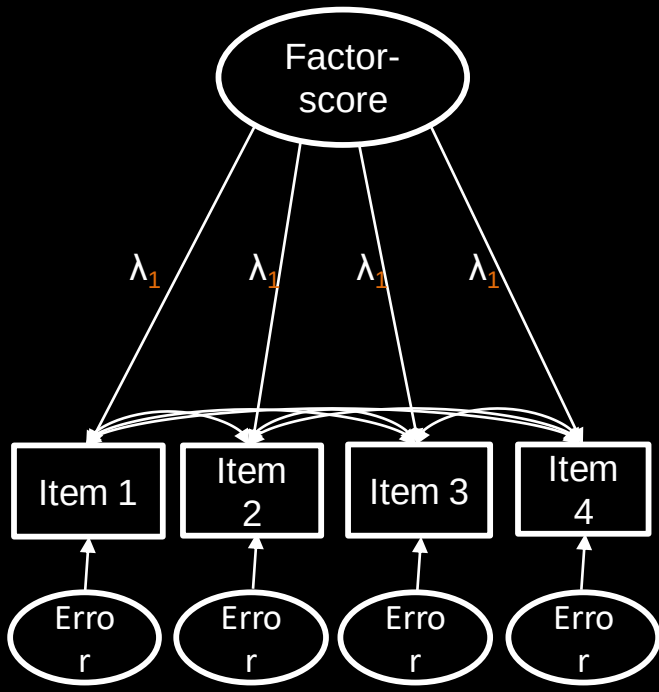


Korrekt wenn die *unterschiedlichen Teile* des Konstruktes *interagieren*

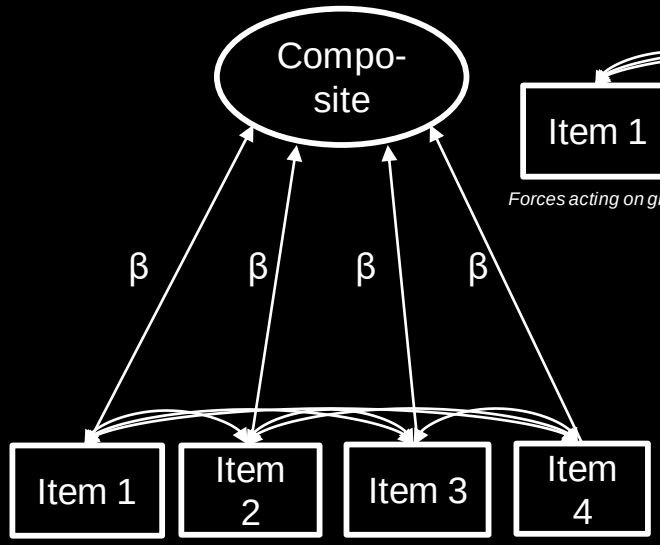


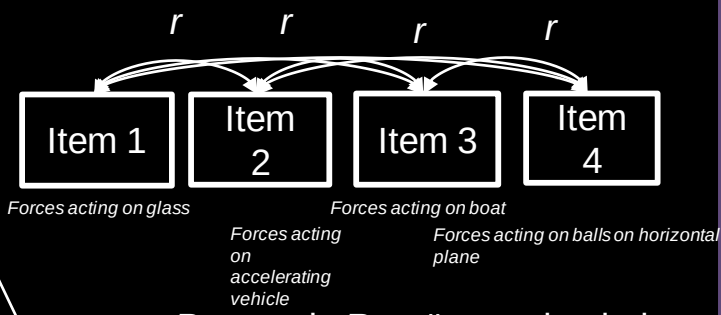


Passende Repräsentation *spezifischer Konzepte* (z.B., Newton's drittes Gesetz)



Valide Repräsentation sozial konstruierter Konstrukte (z.B. Mechanikverständnis, Mathelseitung, Englischverstädnis, Vokabular)





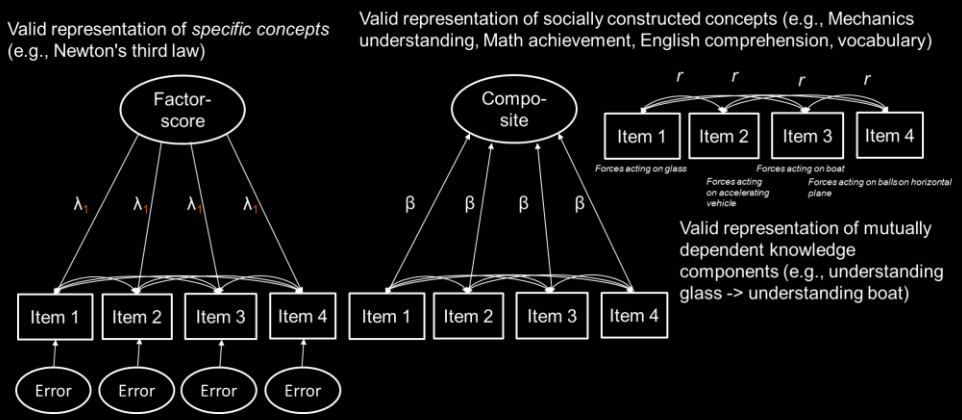
Passende Repräsentation bei gegenseitiger Abhängigkeiten der Aspekte (z.B. Verständnis Glas -> Verständnis Boot)

Implikation 1:
Die **Theorie** gibt das Modell vor, **nicht** die **Daten**

Implikation 2:
Sollten Items interkorrelieren?

Hohe *interne Konsistenz* wird häufig verlang
z.B. Cronbach's Alpha (/Omega) > .70
Als Indikatoren der *Reliabilität*
Taber 2018 (& Stadler et al., 2021)

Für Wissenskonstrukte die
Kompositen/Netzwerke abbilden:
Unpassend



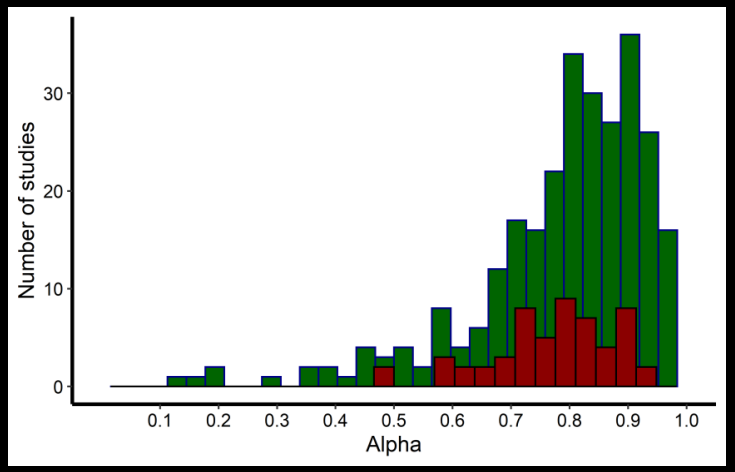
Meta-Analyse:

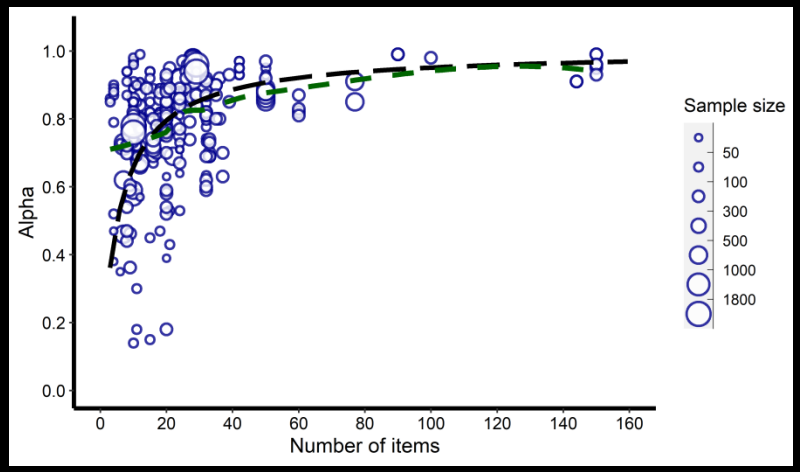
285 Alphas faus 52 Studien, die Lernen über mehrere Messzeitpunkte untersuchten

Ergebnisse: Mittleres Alpha = .85. Extreme Heterogenität (98%): Breite Vorhersageintervalle. Publikationsbias

Höher bei Jüngeren, nach Unterricht, über Entwicklung, in Mathe und Sprachen vs. Naturwissenschaften.

Number of items	90% Lower bound	Predicted Alpha	90% Upper bound
10	.18	.77	.94
20	.44	.84	.96
30	.55	.87	.96
40	.61	.89	.97
50	.65	.90	.97
60	.68	.91	.98
70	.70	.92	.98
80	.72	.92	.98
90	.73	.93	.98
100	.75	.93	.98

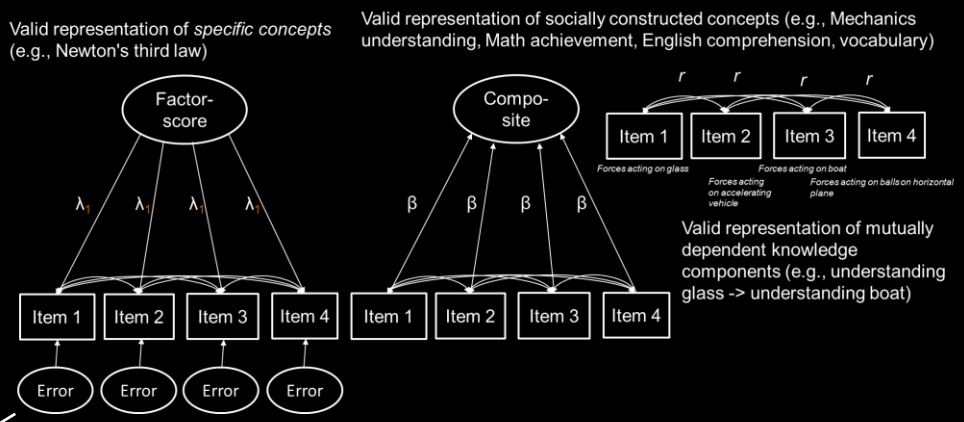




Implikation 2:
Sollten Items interkorrelieren?

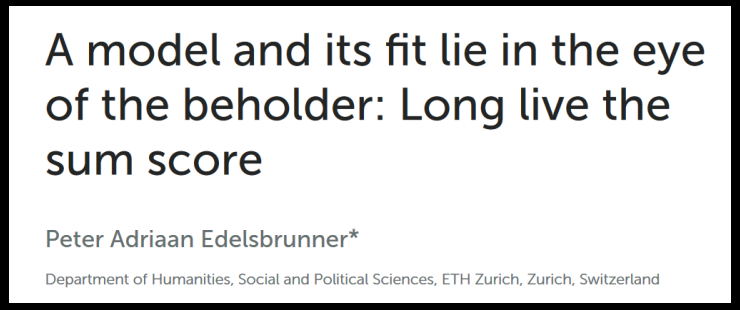
Implikation 1:
Die **Theorie** gibt das Modell vor, **nicht** die **Daten**

Implikation 4:
Wenn du ein **starkes Modell** findest, muss es **starke Annahmen** machen.



Implikation 3:
Wir benötigen **andere Indikatoren** für Validität und Reliabilität (z.B., Retest/Paralleltest Reliabilität, kognitive Interviews, Expertenratings).

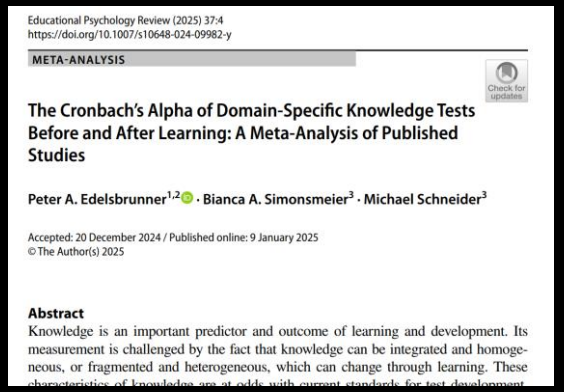
Dankeschön


A model and its fit lie in the eye of the beholder: Long live the sum score

Peter Adriaan Edelsbrunner*

Department of Humanities, Social and Political Sciences, ETH Zurich, Zurich, Switzerland

https://www.frontiersin.org/articles/10.3389/fpsyg.2022.986767/pdf


Educational Psychology Review (2025) 37:4
https://doi.org/10.1007/s10648-024-09982-y

META-ANALYSIS

The Cronbach's Alpha of Domain-Specific Knowledge Tests Before and After Learning: A Meta-Analysis of Published Studies

Peter A. Edelsbrunner[1,2] · Bianca A. Simonsmeier[3] · Michael Schneider[3]

Accepted: 20 December 2024 / Published online: 9 January 2025

Abstract
Knowledge is an important predictor and outcome of learning and development. Its measurement is challenged by the fact that knowledge can be integrated and homogeneous, or fragmented and heterogeneous, which can change through learning. These

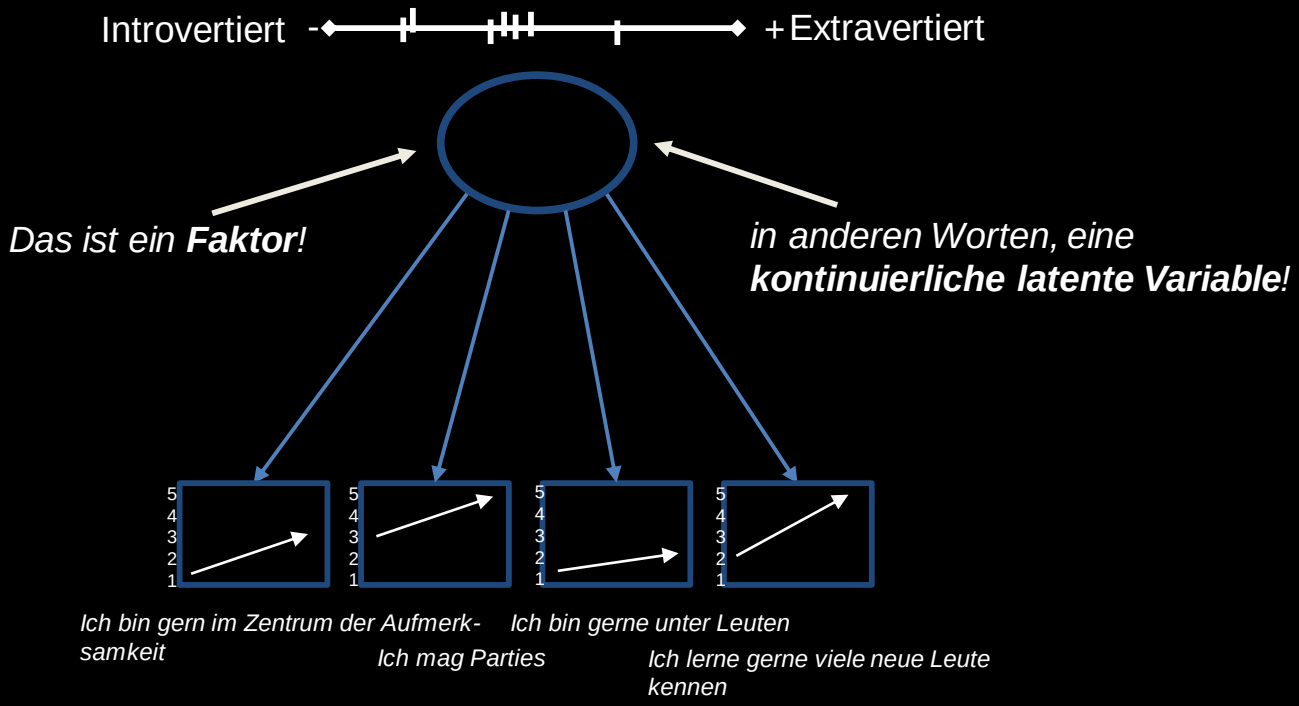
https://osf.io/m8d7t/download

Hol dir diese Präse:

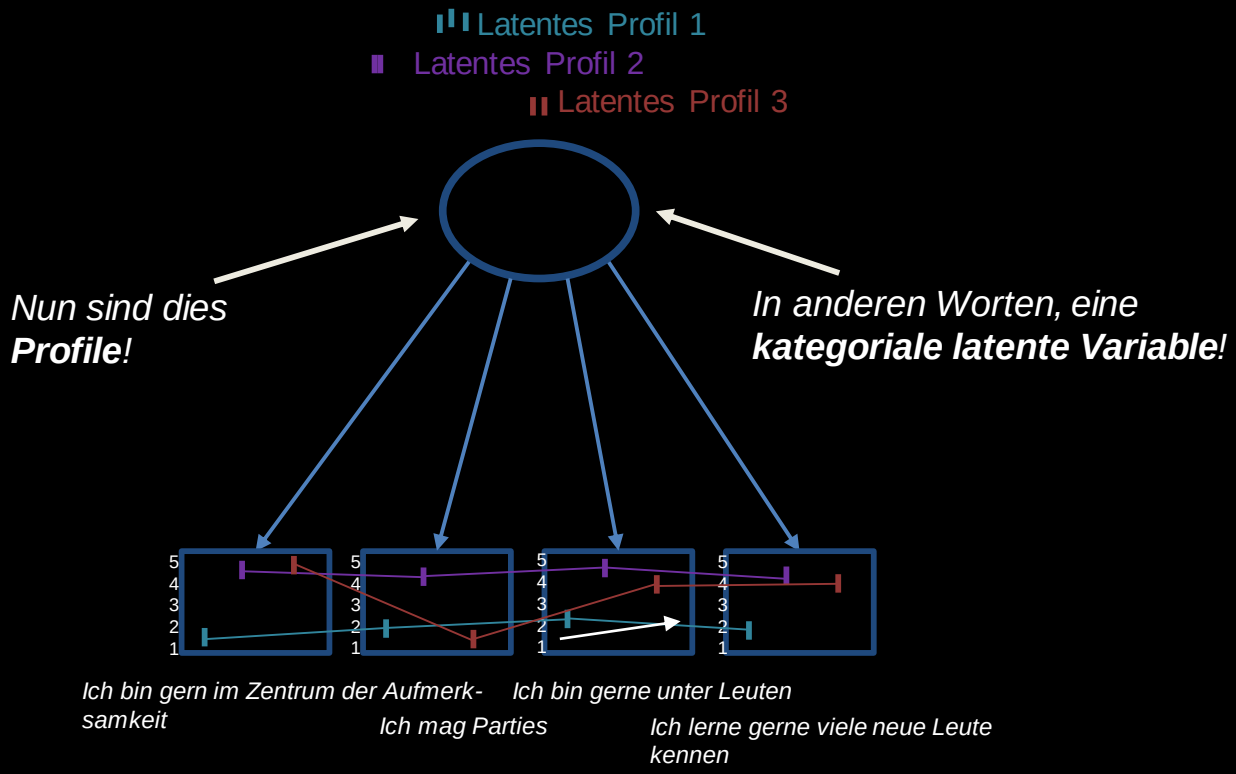
bit.ly/PeterE_presentations

	CTT	Rasch	IRT	CFA	EFA	G-Theory	Network	Mokken scaling	LCA/ LPA
Reliability estimation	+	~	~	~	-	+	-	~	~
Dimensiona-lity testing	-	~	~	+	+	-	~	+	~
Global fit	-	~	~	+	-	-	-	-	-
Item/person fit		+	-	~		-	-		-
Bivariate dependencies			~	~	~		+		
Non-linearity/subgroups								~	+
Deviations from assumptions								+	

+ MDS, Thurstonian Scaling, Fechnerian Scaling, Knowledge Space Theory, Cognitive diagnosis modeling



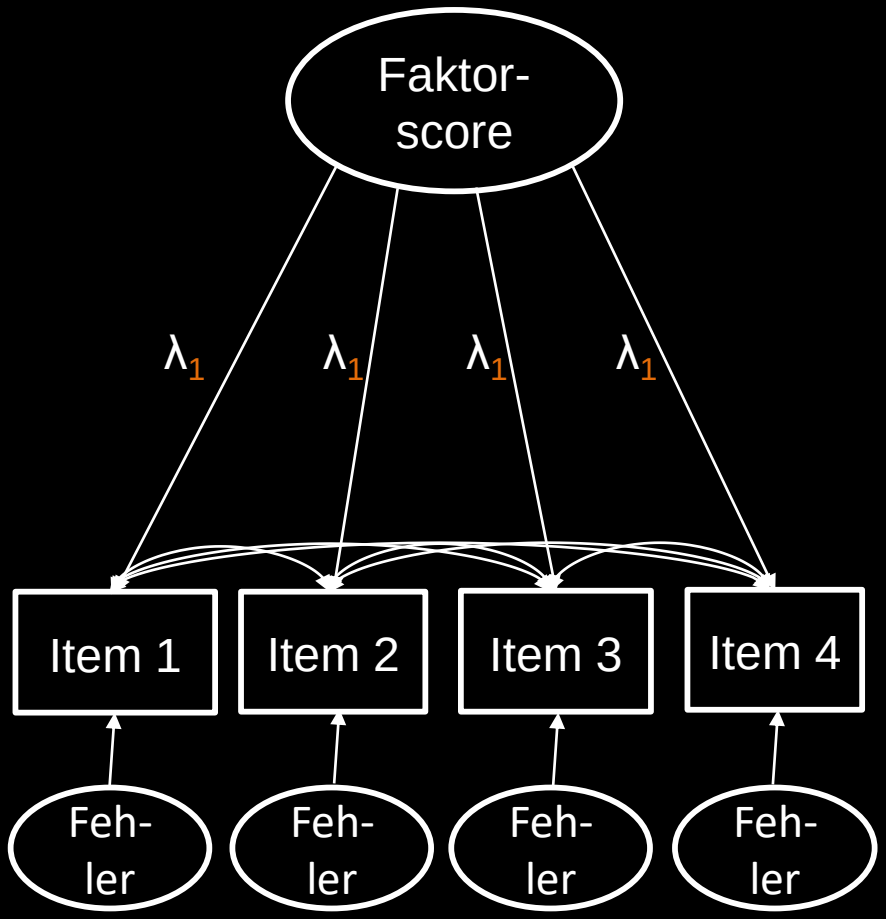
Höhere Extraversion: Erwartete Mittelwerte in Richtung zustimmender Antworten erhöhren sich linear (& proportional zu Faktorladung)



Unterschiedliche Profile von Extraversion: Erwartete Mittelwerte in Richtung zustimmender Antworten **unterscheiden sich zwischen Profilen**

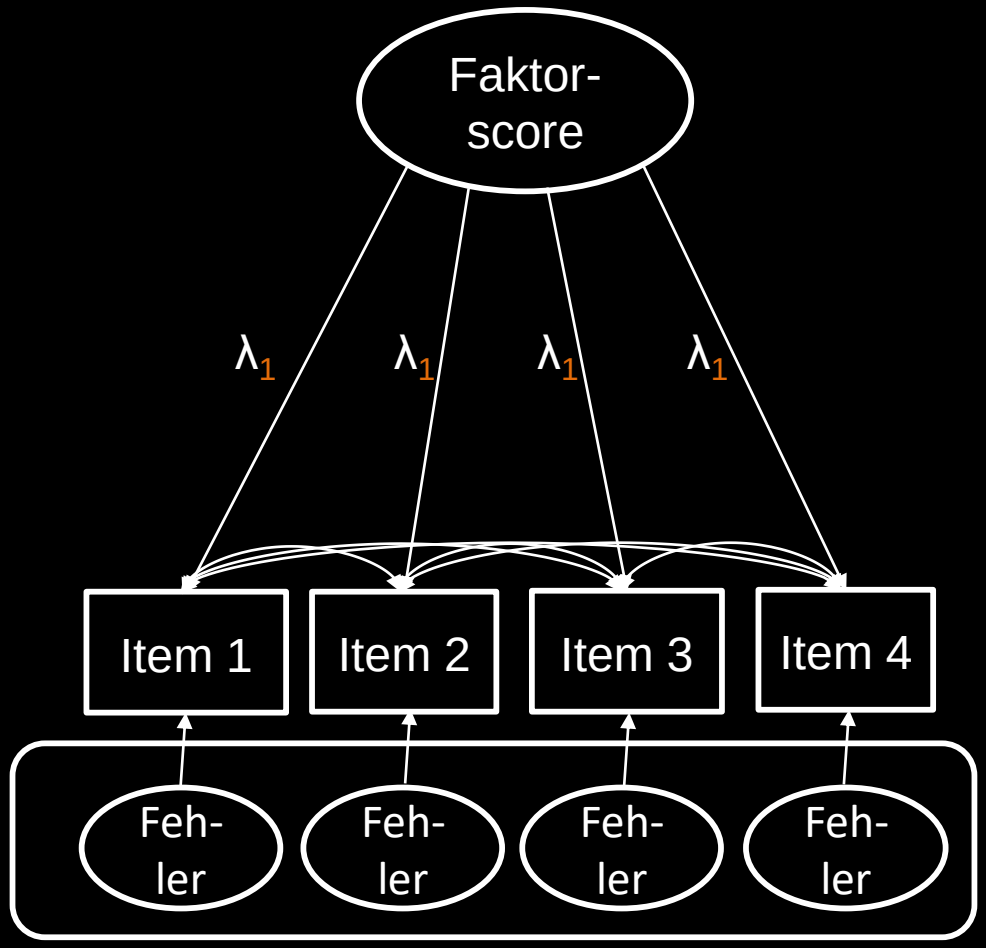
Das sag ich

0. Summenscore entspricht der Annahme **eines latenten Faktormodells** mit gleichen Faktorladungen



Das sag ich

0. Summenscore entspricht der Annahme **eines latenten Faktormodells** mit gleichen Faktorladungen



Kontext/Inhalte:
Austauschbar und
konstruktirrelevant?

Intrinsic cognitive load (Krieglstein et al., 2022)

Die Lerninhalte waren schwer zu verstehen
Die Erklärungen des Lerninhalts waren schwer nachvollziehbar
Die Lerninhalte waren komplex
Die Lerninhalte enthielten viele komplexe Informationen

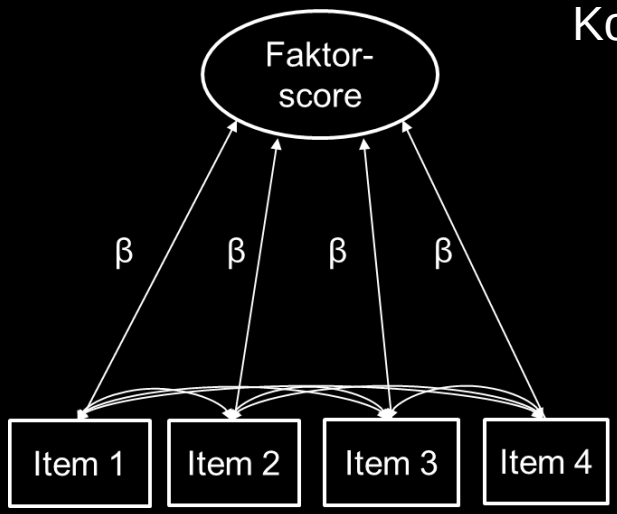
1 – trifft gar nicht zu
9 – trifft vollständig zu

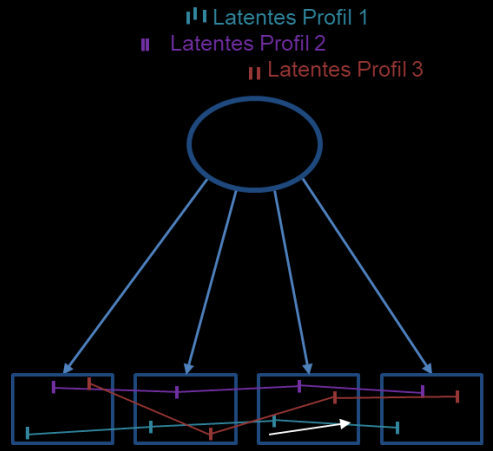
Das sag ich

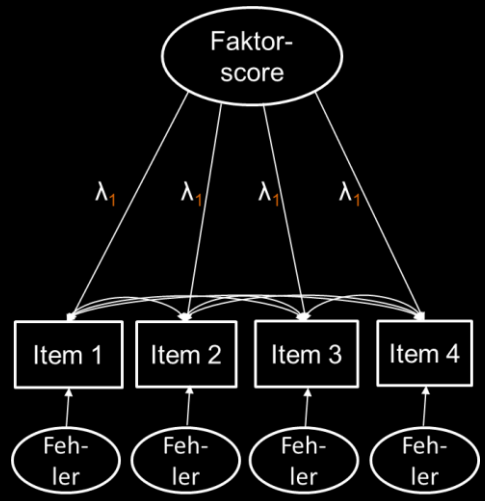
0. Summenscore entspricht der Annahme **eines latenten Faktormodells** mit gleichen Faktorladungen

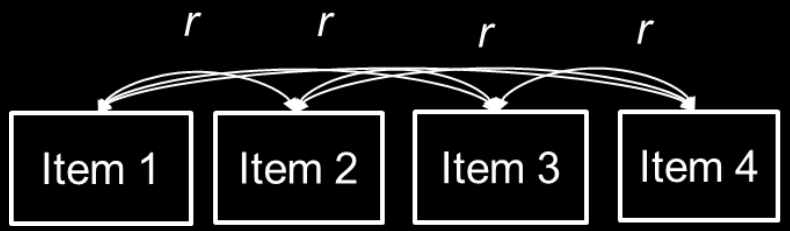
und die **Empirie**?

0. Es gibt **unterschiedliche Meta-Theorien**. Man sollte immer diejenige wählen, die **aus theoretischer Sicht** (*Austauschbarkeit*, *latente Eigenschaft* vs. *direkte Dynamiken*) am besten die Eigenschaften des modellierten Konstruktes beschreibt.









Das sag ich

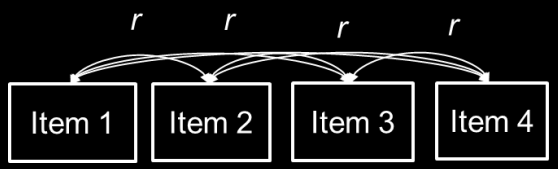
0. Latentes Faktormodell mit homogenen Faktorladungen

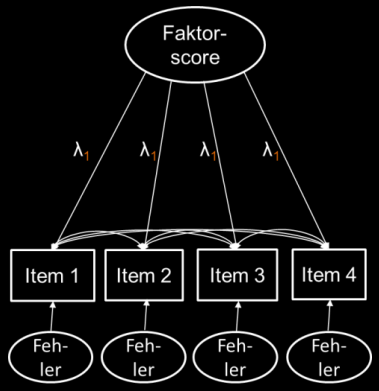
Summenscore

?

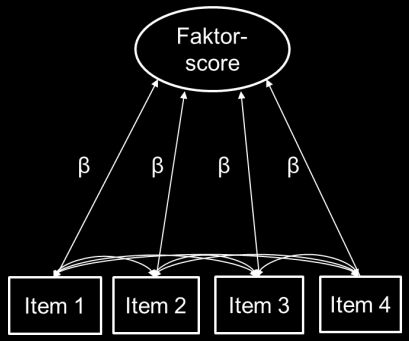
Affirmation der Konsequenz

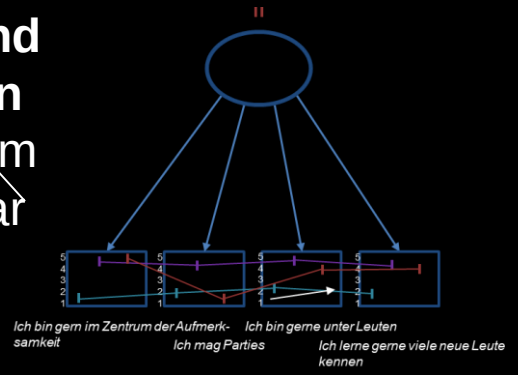
Umkehrfehler

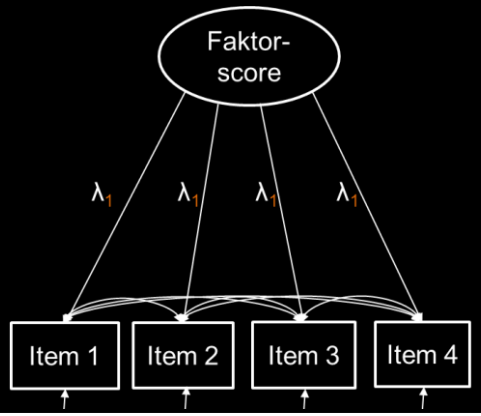




Modelle implizieren **dieselben Mittelwerte und Korrelationen** Empirisch kaum unterscheidbar







Bedeutet das im Umkehrschluss, dass ich immer, wenn ich einen Summenscore verwende, die Annahme mache, dass mein Test Rasch-homogen ist?

Wenn das **Rasch Modell** gilt, dann sind Summenscores **suffiziente Statistiken**

Äh… well…
Ja, das ist korrekt! :D

Faktorscores anstatt Summenscores zu verwenden entspricht nach naturwissenschaftlicher Messung der **Einmodellierung von Messfehler**
(Abweichung von Rasch Modell)

Das sag ich

0. Der Summenscore entspricht der Annahme des *theoretisch plausiblen Modells*

1. Dieses Modell kann man (deskriptiv) überprüfen, oder einfach (normativ) vorgeben

2. Empirisch hält die Annahme (fast) nie

3. Schlussfolgerung:
Scorebildung ist sinnvoll, wenn sie der Theorie oder dem Forschungsziel entspricht

Beispiel:

Inhaltswissen

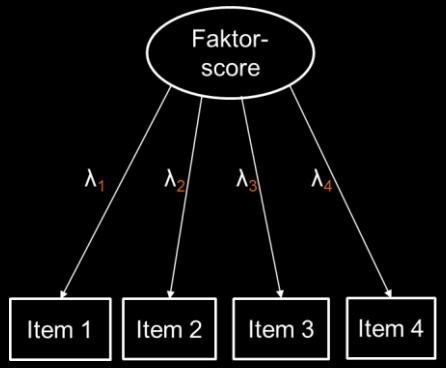
Hohe interne Konsistenz von Tests wird häufig gewünscht

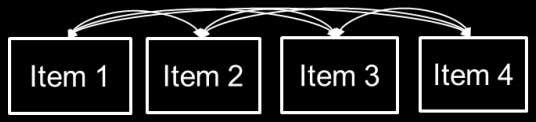
z.B. Cronbach’s Alpha (oder Omega)

Taber 2018 (& Stadler et al., 2021)

Für Wissenstests inadäquat

Theoretische Annahme:
Heterogen und **Mehrdimensional**



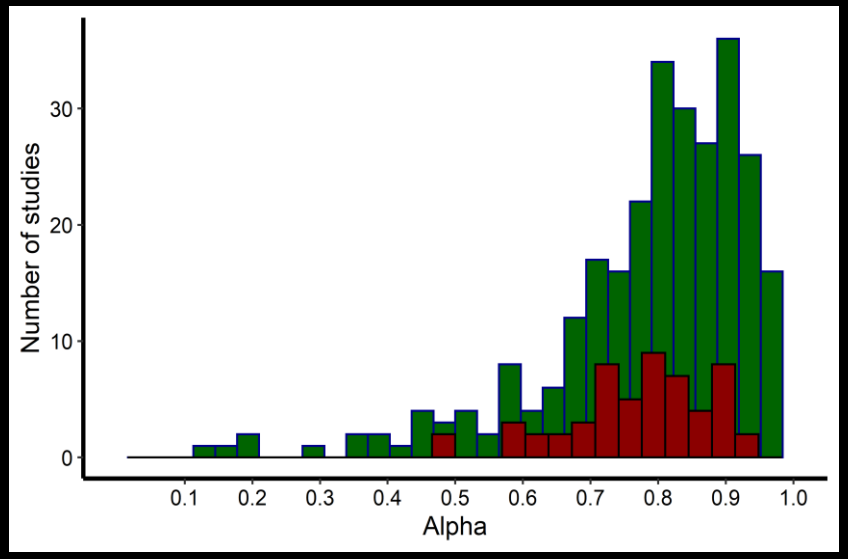


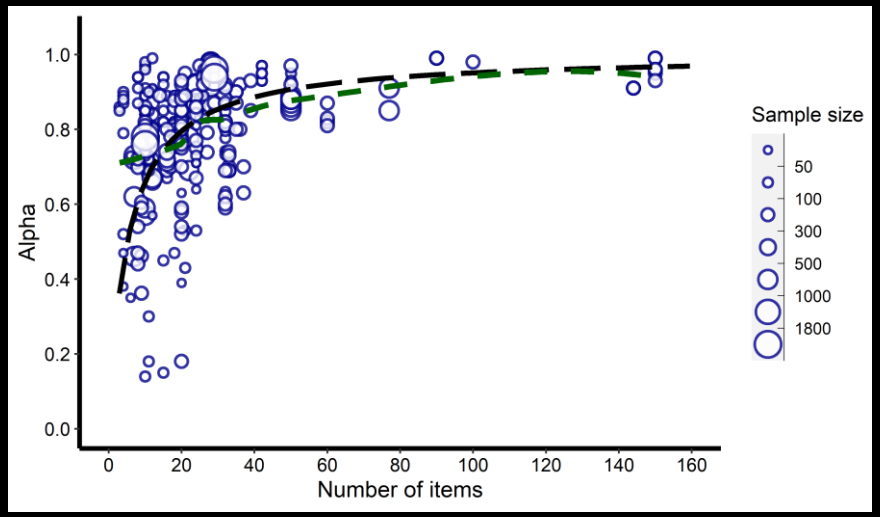
Beispiel:

Inhaltswissen

Meta-Analyse:

Mittlere Interkorrelation Wissensitems
r = .22

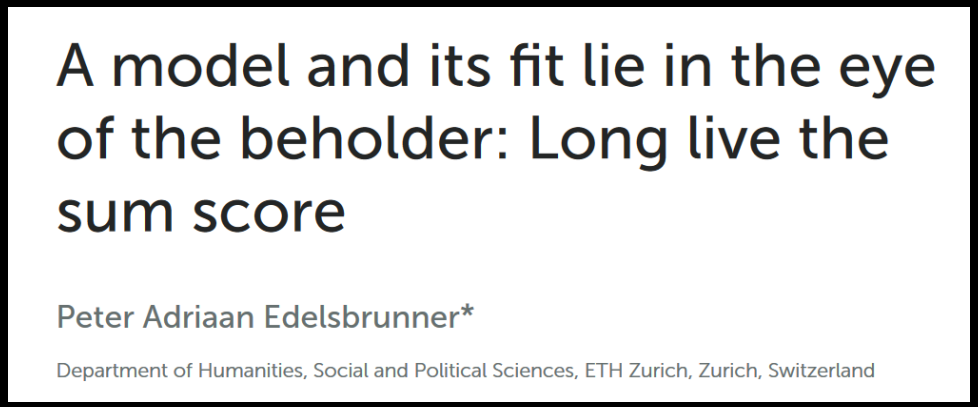




	CTT	Rasch	IRT	CFA	EFA	G-Theory	Netz-werk	Mokken	LCA/ LPA
Reliabilitätsschätzung	+	~	~	~	-	+	-	~	~
Dimensionalitätsprüfung	-	~	~	+	+	-	~	+	~
Globaler Fit	-	~	~	+	-	-	-	-	-
Item-/Personenfit		+	-	~		-	-		-
Bivariate Abhängigkeiten			~	~	~		+		
Nicht-Linearitäten/Subgruppen								~	+
Annahmenverletzung								+	

+ MDS, Thurstonian Scaling, Fechnerian Scaling, Knowledge Space Theory, Cognitive diagnosis modeling

Danke

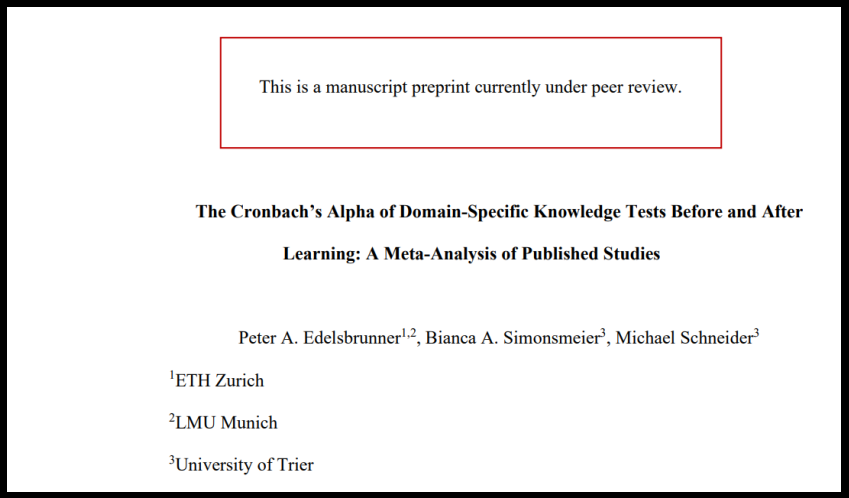


A model and its fit lie in the eye of the beholder: Long live the sum score

Peter Adriaan Edelsbrunner*

Department of Humanities, Social and Political Sciences, ETH Zurich, Zurich, Switzerland

https://www.frontiersin.org/articles/10.3389/fpsyg.2022.986767/pdf



This is a manuscript preprint currently under peer review.

The Cronbach's Alpha of Domain-Specific Knowledge Tests Before and After Learning: A Meta-Analysis of Published Studies

Peter A. Edelsbrunner[1,2], Bianca A. Simonsmeier[3], Michael Schneider[3]

[1]ETH Zurich

[2]LMU Munich

[3]University of Trier

https://osf.io/m8d7t/download

Zieht euch diese Präse:

bit.ly/PeterE_presentations

Behavior Research Methods (2020) 52:2287–2305
https://doi.org/10.3758/s13428-020-01398-0

Thinking twice about sum scores

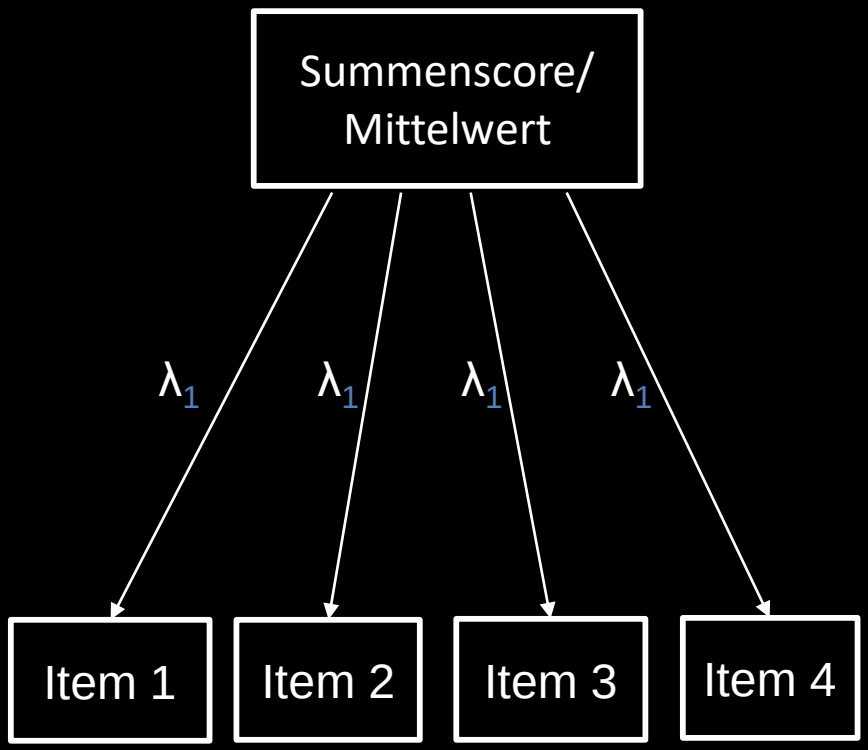
Daniel McNeish[1] · Melissa Gordon Wolf[2]

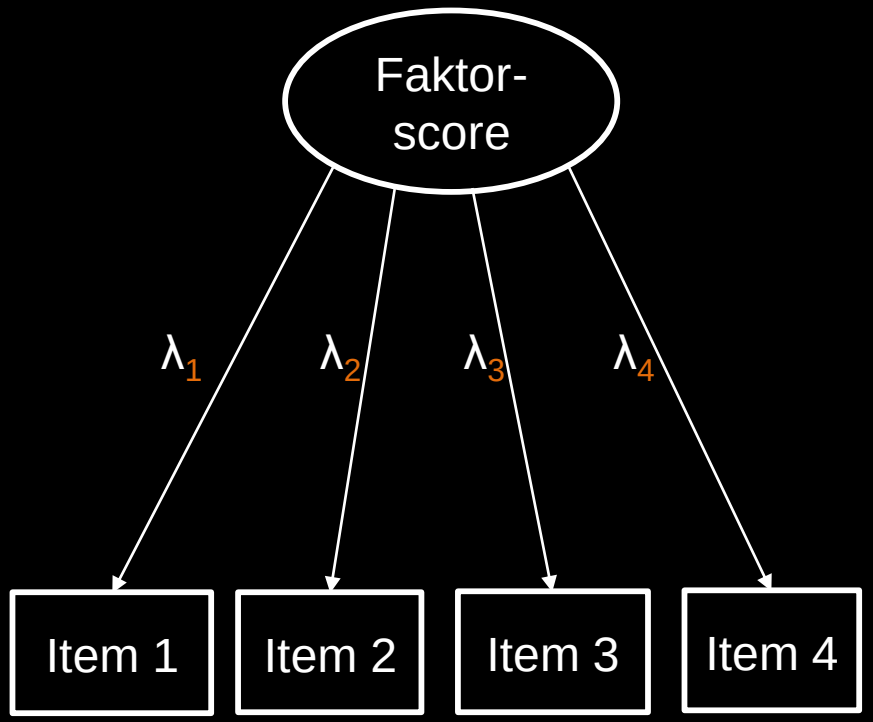
Published online: 22 April 2020

Abstract

A common way to form scores from multiple-item scales is to sum responses of all items. Though sum scoring is often contrasted with factor analysis as a competing method, we review how factor analysis and sum scoring both fall under the larger umbrella of latent variable models, with sum scoring being a constrained version of a factor analysis. Despite similarities, reporting of psychometric properties for sum scored or factor analyzed scales are quite different. Further, if researchers use factor analysis to validate a scale but subsequently sum score the scale, this employs a model that differs from validation model. By framing sum scoring within a latent variable framework, our goal is to raise awareness that (a) sum scoring requires rather strict constraints, (b) imposing these constraints requires the same type of justification as any other latent variable model, and (c) sum scoring corresponds to a statistical model and is not a model-free arithmetic calculation. We discuss how unjustified sum scoring can have adverse effects on validity, reliability, and qualitative classification from sum score cut-offs. We also discuss considerations for how to use scale scores in subsequent analyses and how these choices can alter conclusions. The general goal is to encourage researchers to more critically evaluate how they obtain, justify, and use multiple-item scale scores.

Option 1 = Option 2:





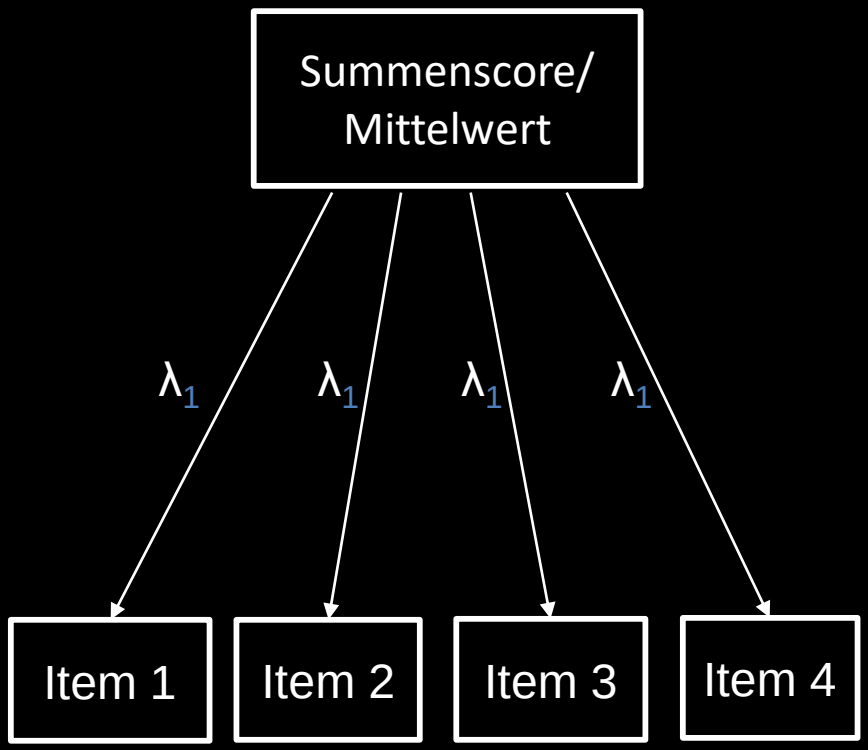
The sum score is a *constrained version* of factor analysis
(McNeish & Wolf, 2020)

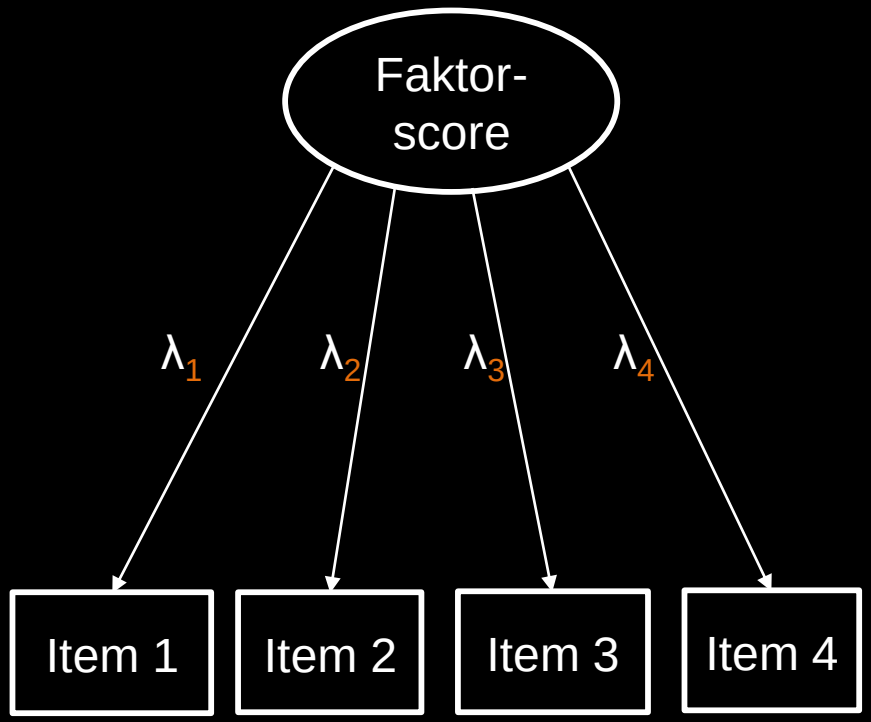
Faktorscore:
Nach Faktorladung gewichteter Wert aus allen Items

Alle Items werden *gleich gewichtet*

Entspricht implizit Faktorenanalyse mit gleichen Faktorladungen

Option 1 = Option 2:





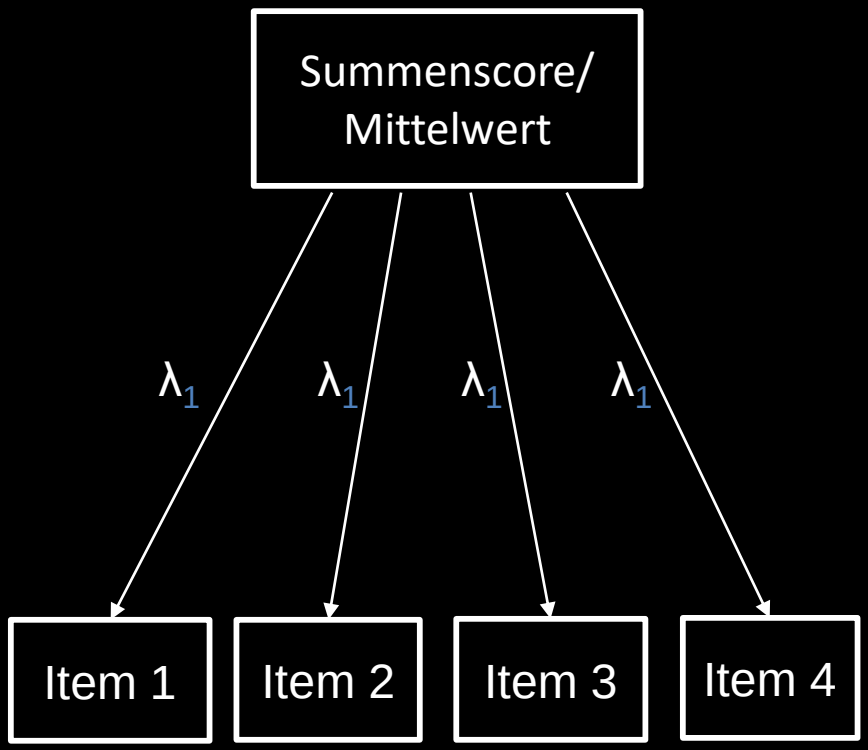
The sum score is a *constrained version* of factor analysis
(McNeish & Wolf, 2020)

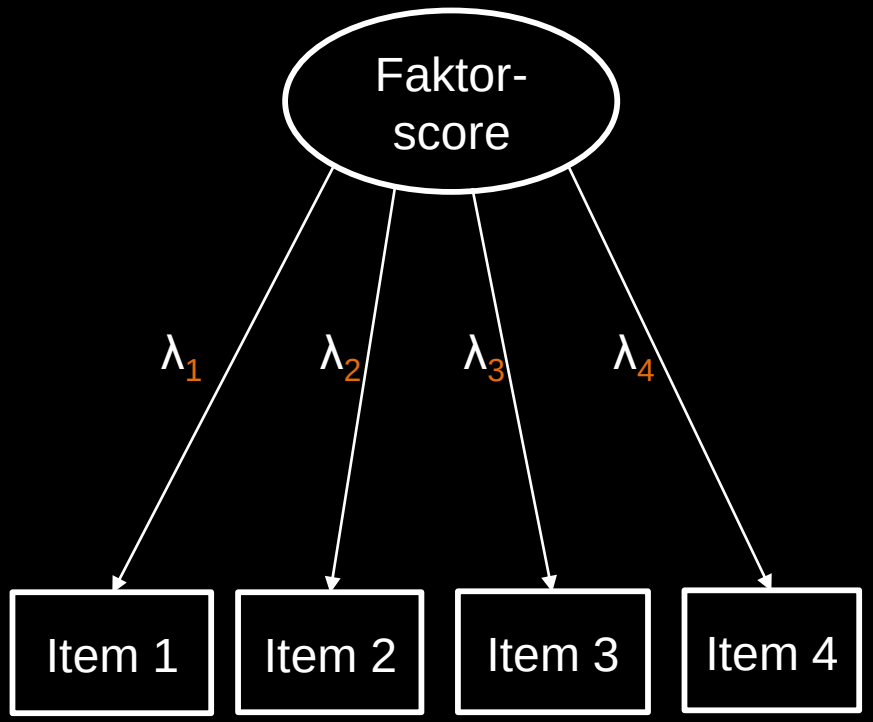
Alle Items werden *gleich gewichtet*

Entspricht implizit der Annahme gleicher Faktorladungen

Faktorscore:
Nach Faktorladung gewichteter Wert aus allen Items

Option 1 = Option 2:





Alle Items werden *gleich gewichtet*

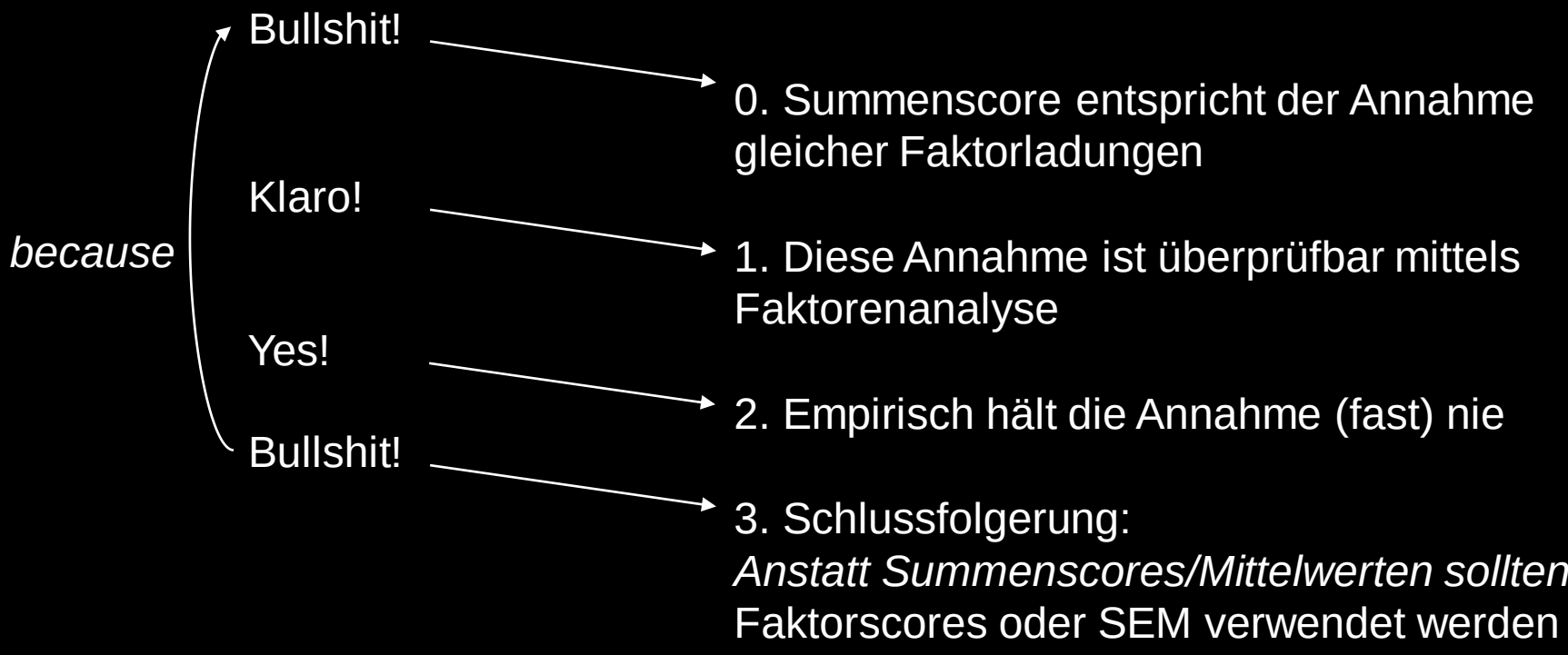
Entspricht implizit der *Annahme gleicher Faktorladungen*

Diese Annahme ist überprüfbar mittels Faktorenanalyse

Empirisch hält die Annahme nicht (fast nie)

Faktorscore:
Nach Faktorladung gewichteter Wert aus allen Items

Das sag ich


Bullshit!
Klaro!
Yes!
Bullshit!
because
0. Summenscore entspricht der Annahme gleicher Faktorladungen
1. Diese Annahme ist überprüfbar mittels Faktorenanalyse
2. Empirisch hält die Annahme (fast) nie
3. Schlussfolgerung:
Anstatt Summenscores/Mittelwerten sollten
Faktorscores oder SEM verwendet werden

Das sagen die Kommentare

0. Summenscore entspricht der Annahme gleicher Faktorladungen

1. Diese Annahme ist überprüfbar mittels Faktorenanalyse

2. Empirisch hält die Annahme (fast) nie

3. Schlussfolgerung:
Anstatt Summenscores/Mittelwerten sollten Faktorscores oder SEM verwendet werden

Rasch

Bedeutet das im Umkehrschluss, dass ich immer, wenn ich einen Summenscore verwende, die Annahme mache, dass mein Test Rasch-homogen ist?

well…
Ja, das ist korrekt!