Long Live the Sum Score

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

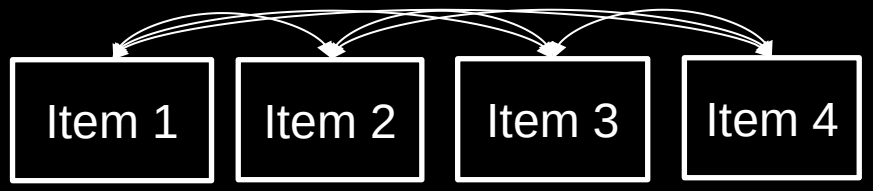
Option 1:

Summenscore/
Mittelwert

Intervention/Prä-
Post Veränderung

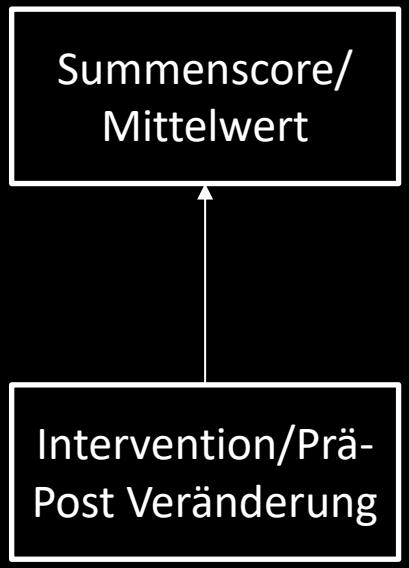
Option 2:

Faktoren
-analyse

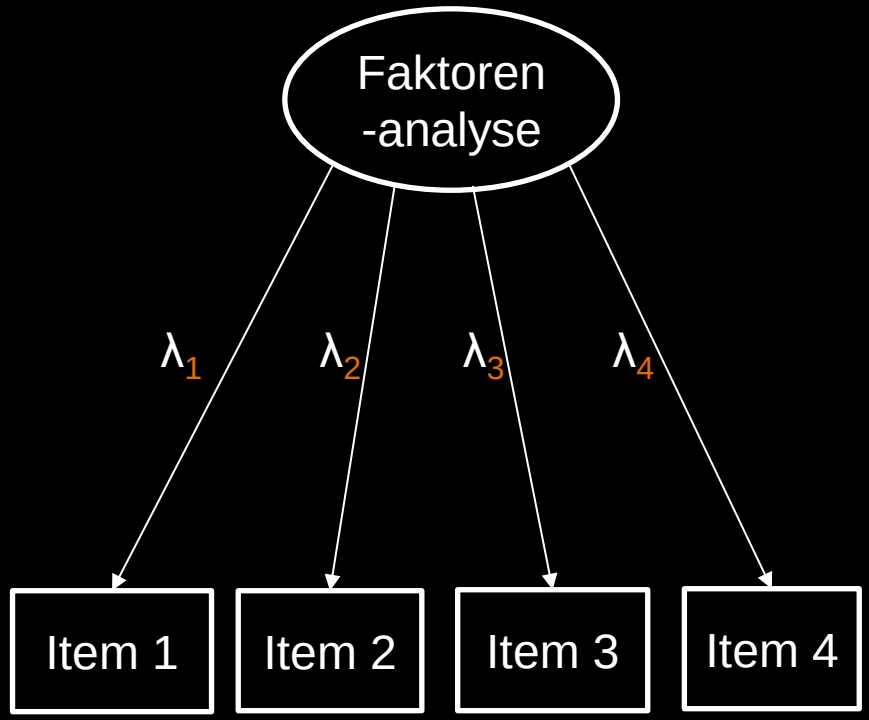


Ermittlung der Anzahl benötigter *Quellen gemeinsamer Varianz* (Faktoren),
um Interkorrelationen zwischen Items zu erklären/modellieren

Option 1:



Option 2:



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

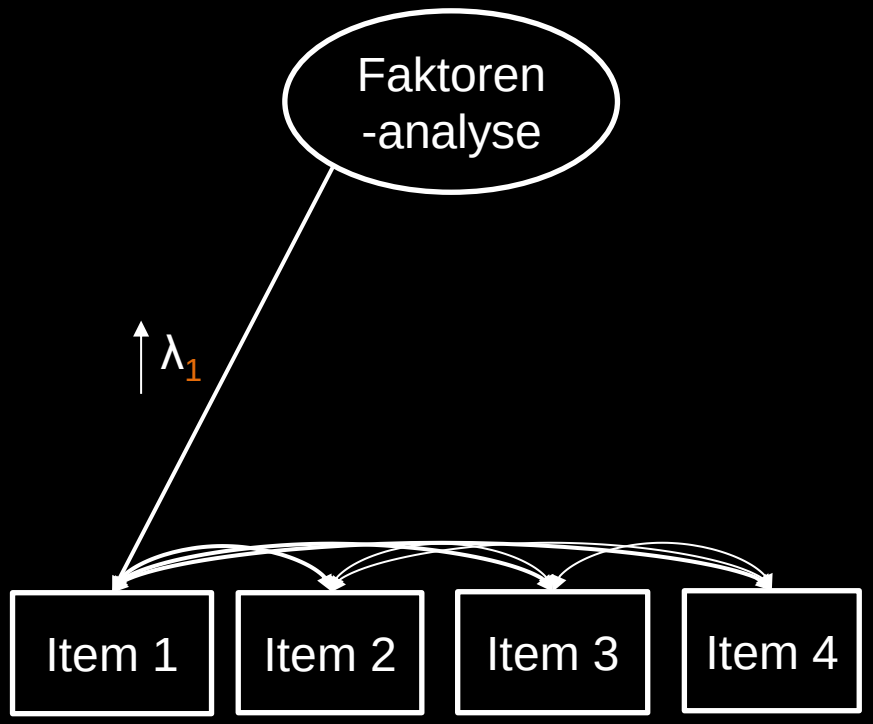
Schätzung der Stärke, mit welcher die gemeinsame Varianz
In jedes Item eingeht (Faktorladung λ_1-λ_4)

Option 1:

Summenscore/
Mittelwert

Intervention/Prä-
Post Veränderung

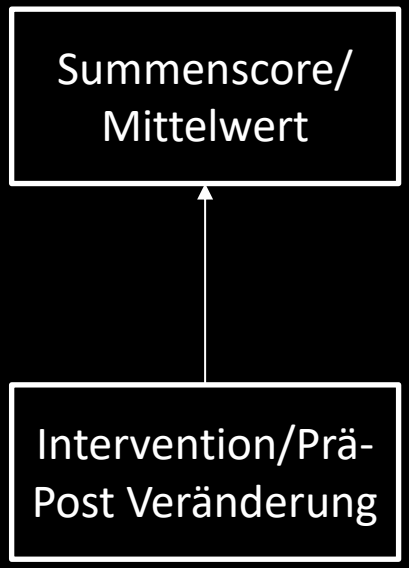
Option 2:



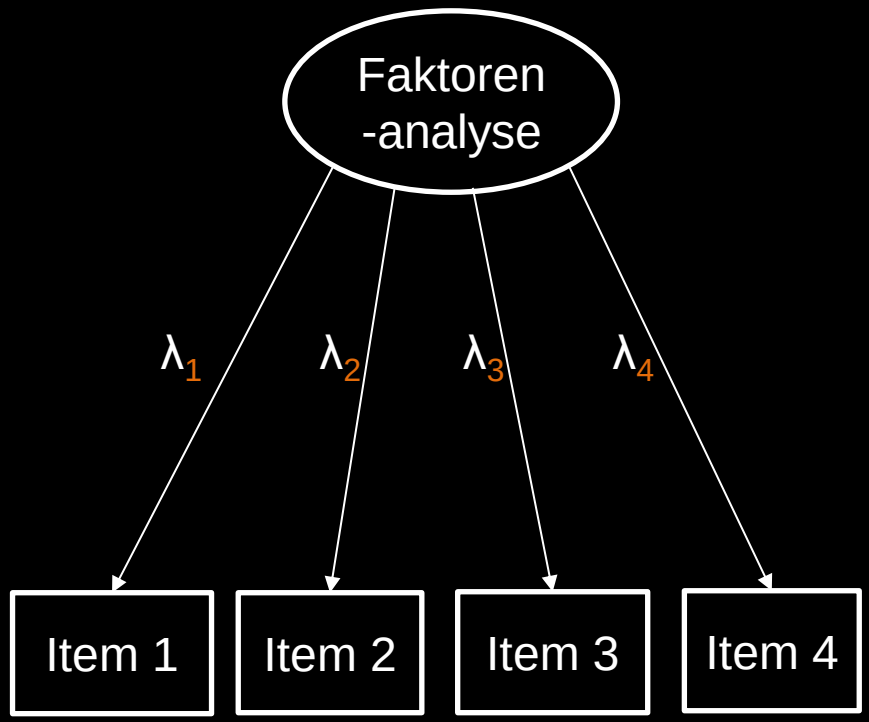
Items, die mit den anderen Items *hohe Interkorrelationen* aufweisen,
Erhalten *hohe Faktorladungen* (starke Indikatoren des g. Konstruktes)

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

Option 1:



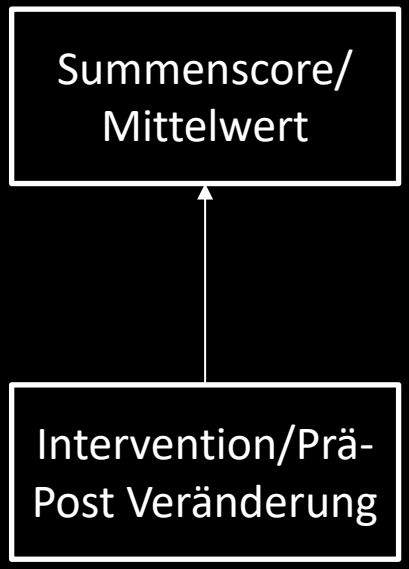
Option 2:



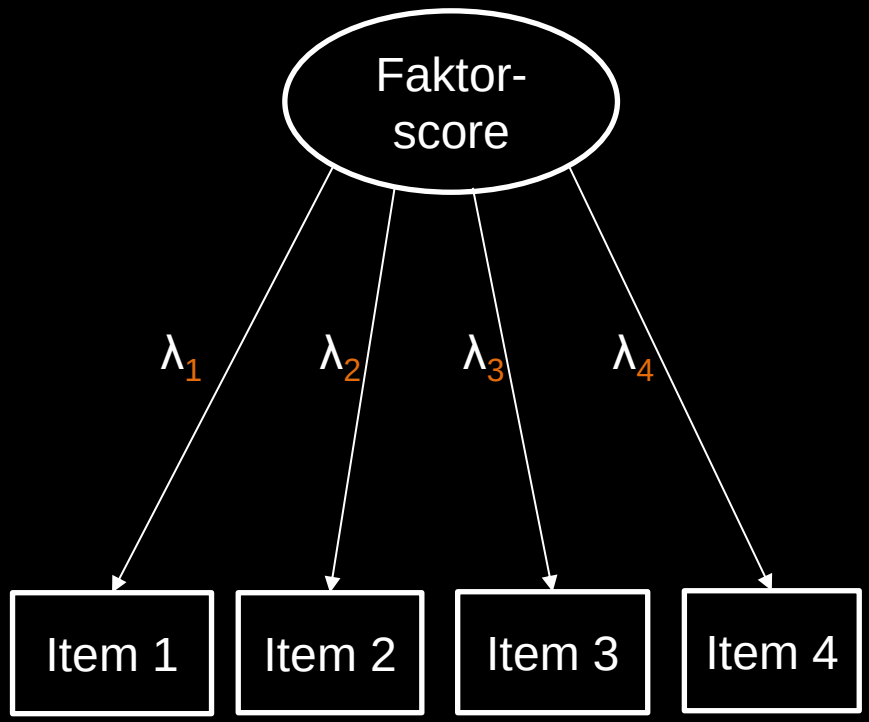
Als Messmodell:
Untersuchen, ob *die theoretisch erwartete Faktorenstruktur* (Anzahl & Ladungen) vorliegt

Als Skalierungsmodell:
Aus dem Messmodell werden geschätzte Messwerte gebildet

Option 1:



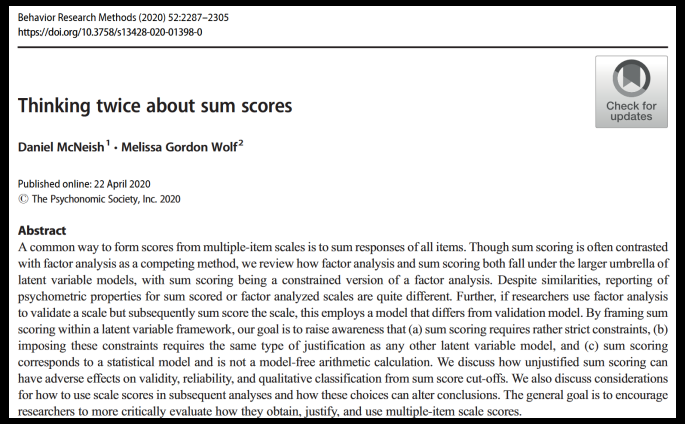
Option 2:



Faktorscore:
Nach Faktorladung gewichteter Wert aus allen Items

Option 1:

Summenscore/
Mittelwert



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

Thinking twice about sum scores

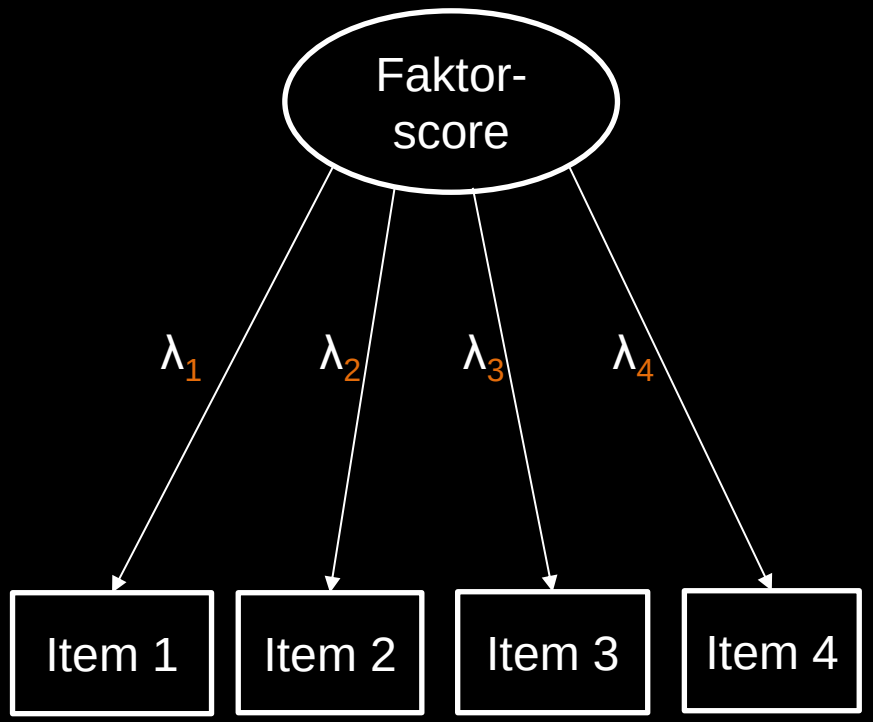
Daniel McNeish[1] · Melissa Gordon Wolf[2]

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 *Sum + Alpha*
(+ Faktorenanalyse) - Ansatzes

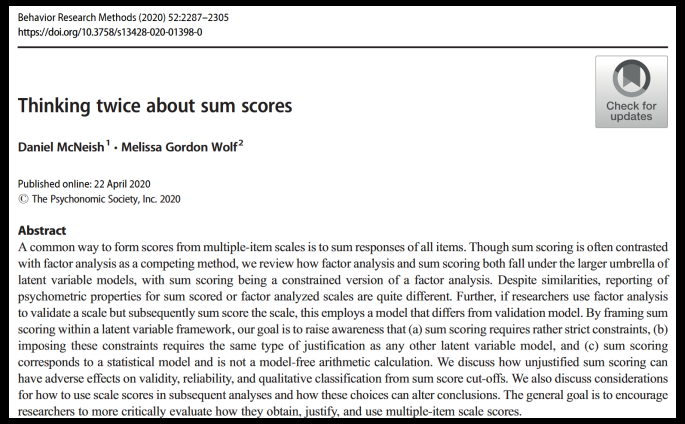
Option 2:



Faktorscore:
Nach Faktorladung gewichteter Wert aus allen Items

Option 1:

Summenscore/
Mittelwert



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

Thinking twice about sum scores

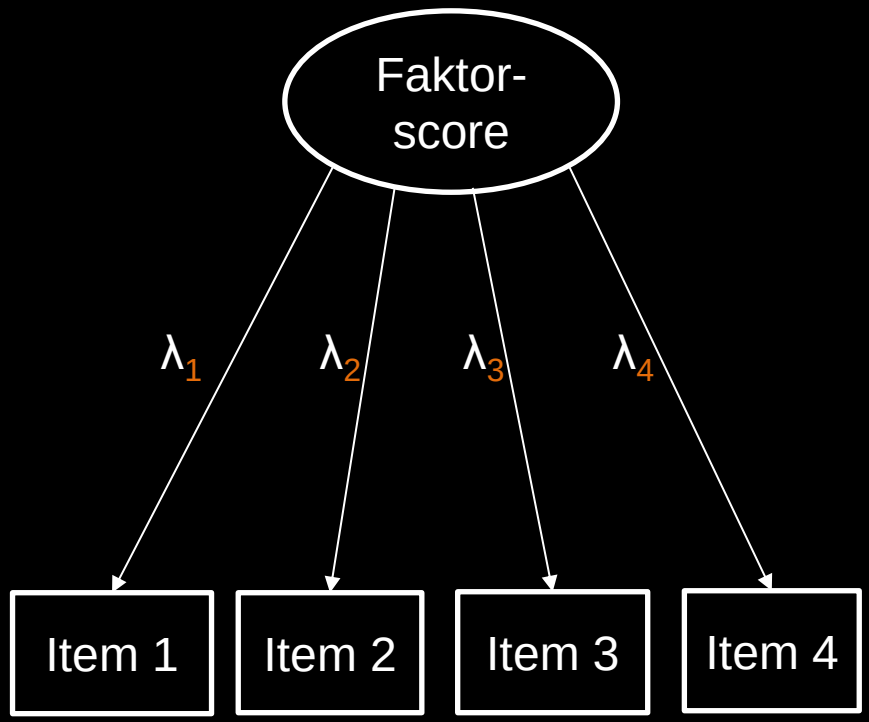
Daniel McNeish[1] · Melissa Gordon Wolf[2]

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.

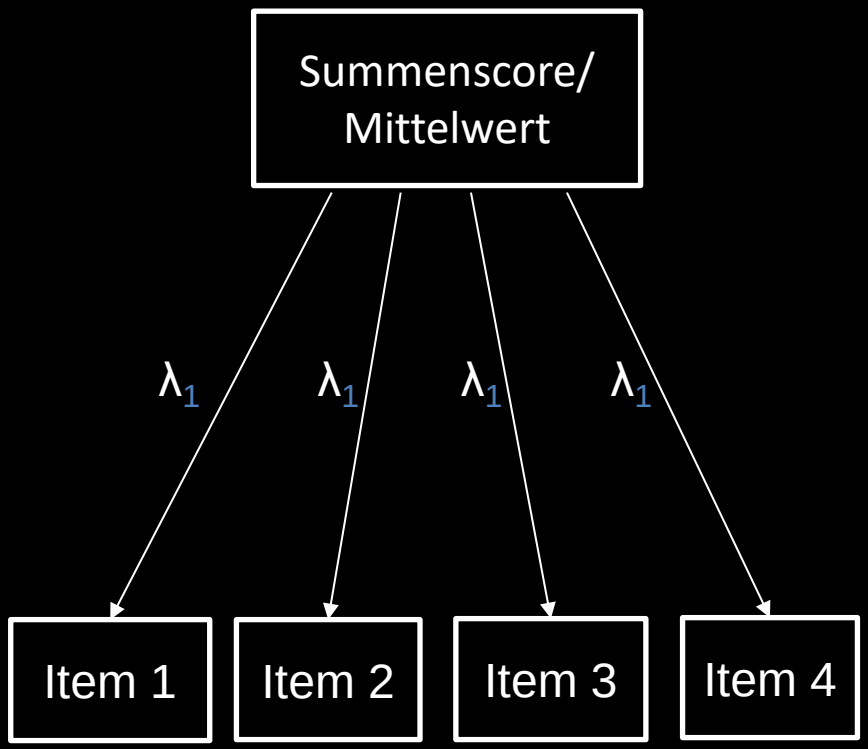
The sum score is a *constrained version* of factor analysis
(McNeish & Wolf, 2020)

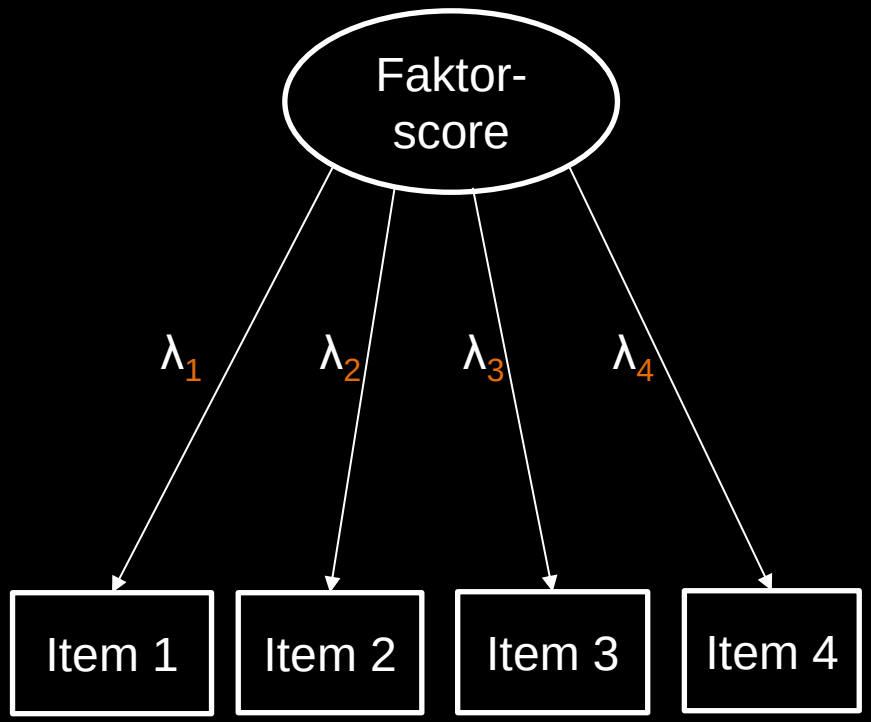
Option 2:



Faktorscore:
Nach Faktorladung gewichteter Wert aus allen Items

Option 1 = Option 2:





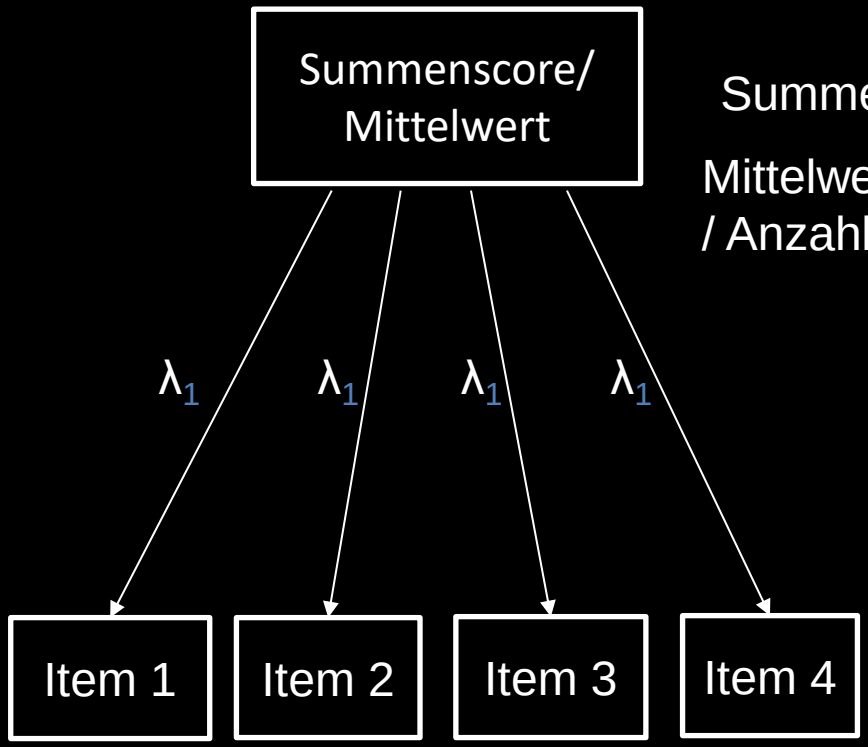
The sum score is a *constrained version* of factor analysis
(McNeish & Wolf, 2020)

Alle Items werden *gleich gewichtet*

Faktorscore:
Nach Faktorladung gewichteter Wert aus allen Items

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?

bit.ly/Faktorscores

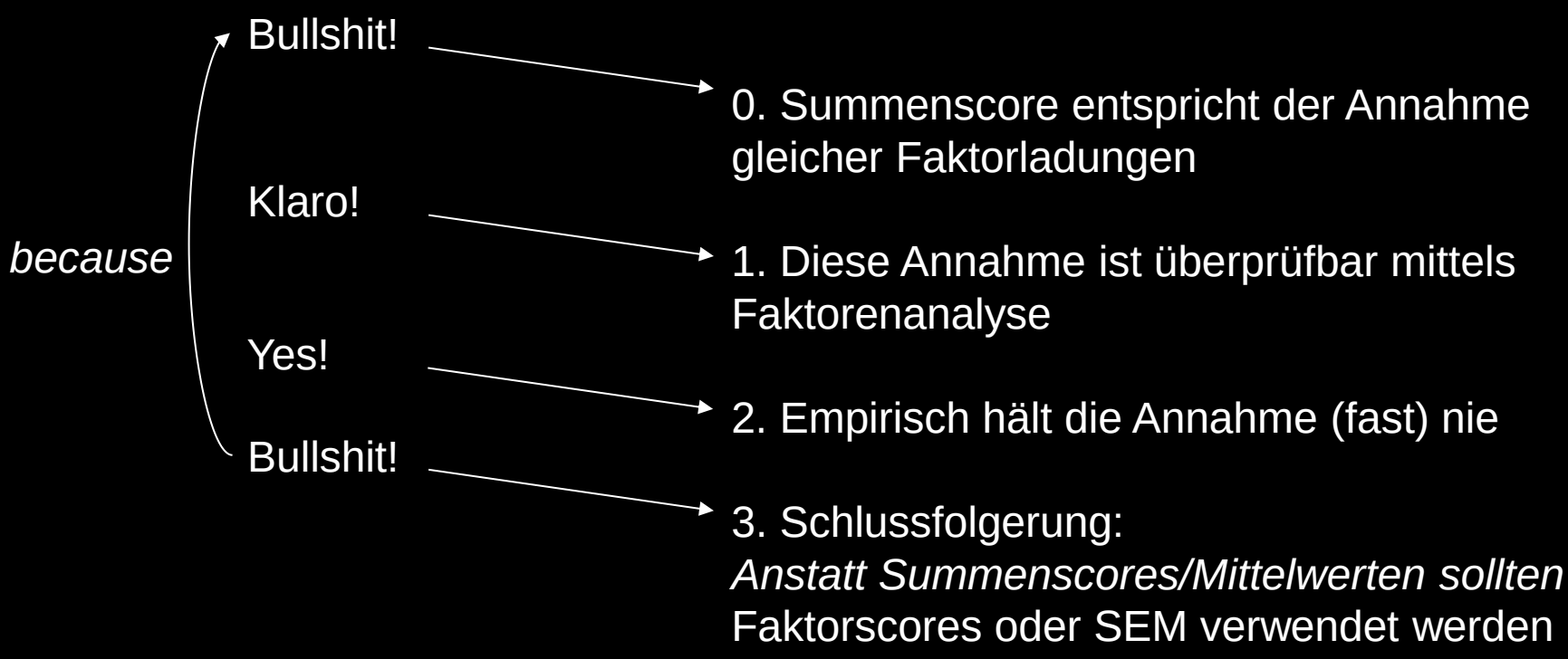
0. Summenscore entspricht der Annahme gleicher Faktorladungen

1. Diese Annahme ist überprüfbar mittels Faktorenanalyse

2. Empirisch hält die Annahme (fast) nie

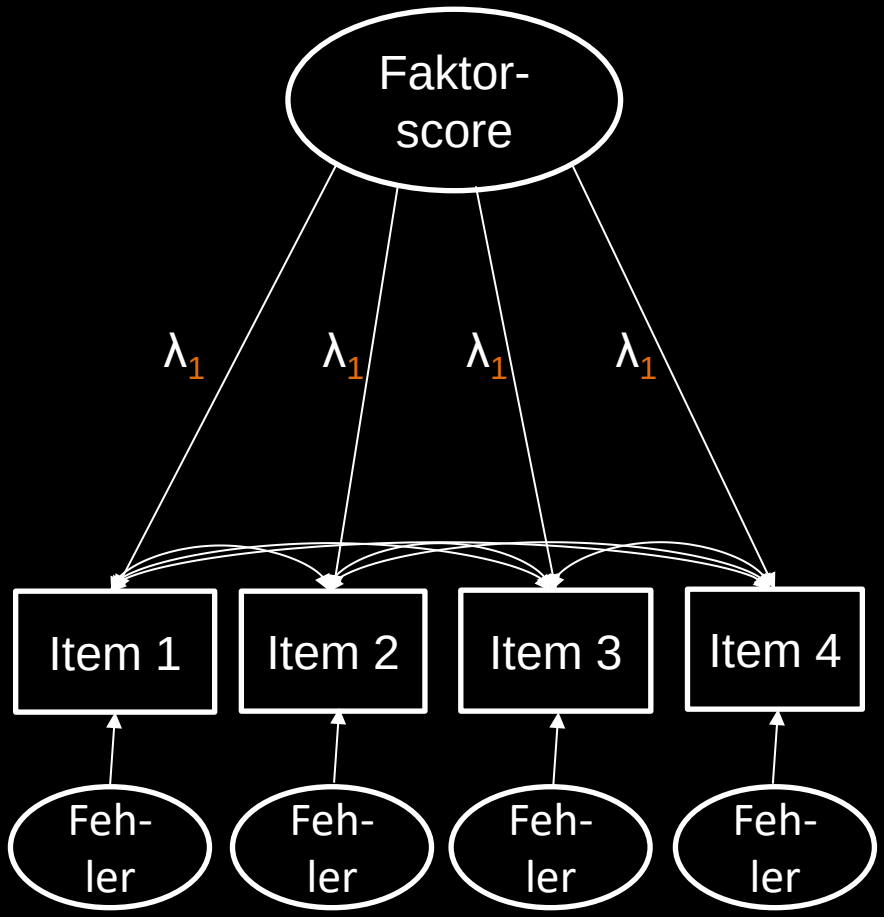
3. Schlussfolgerung:
Anstatt Summenscores/Mittelwerten sollten Faktorscores oder SEM verwendet werden

Das sag ich


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

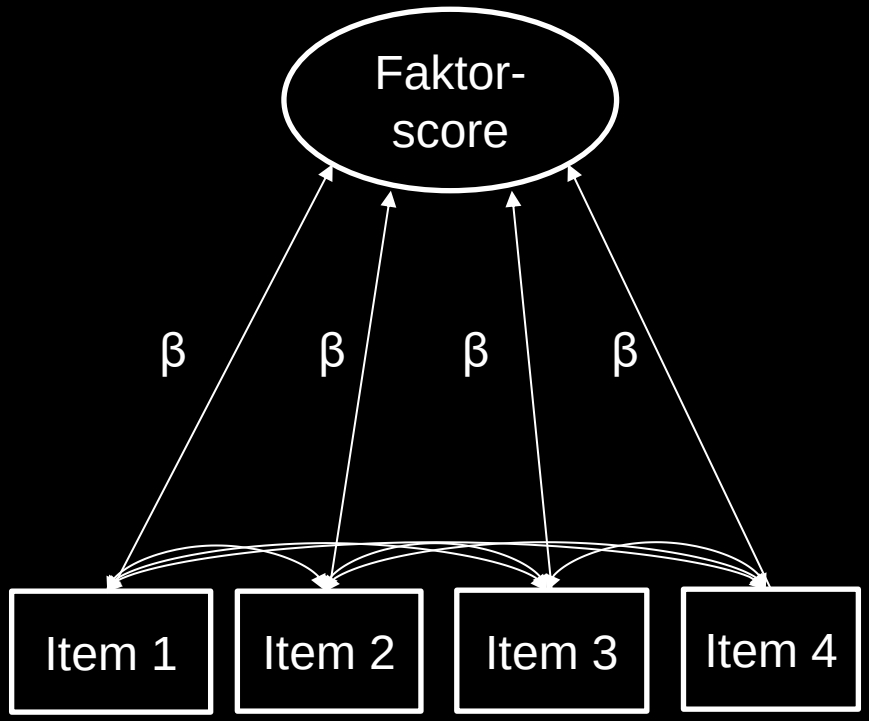
Das sag ich

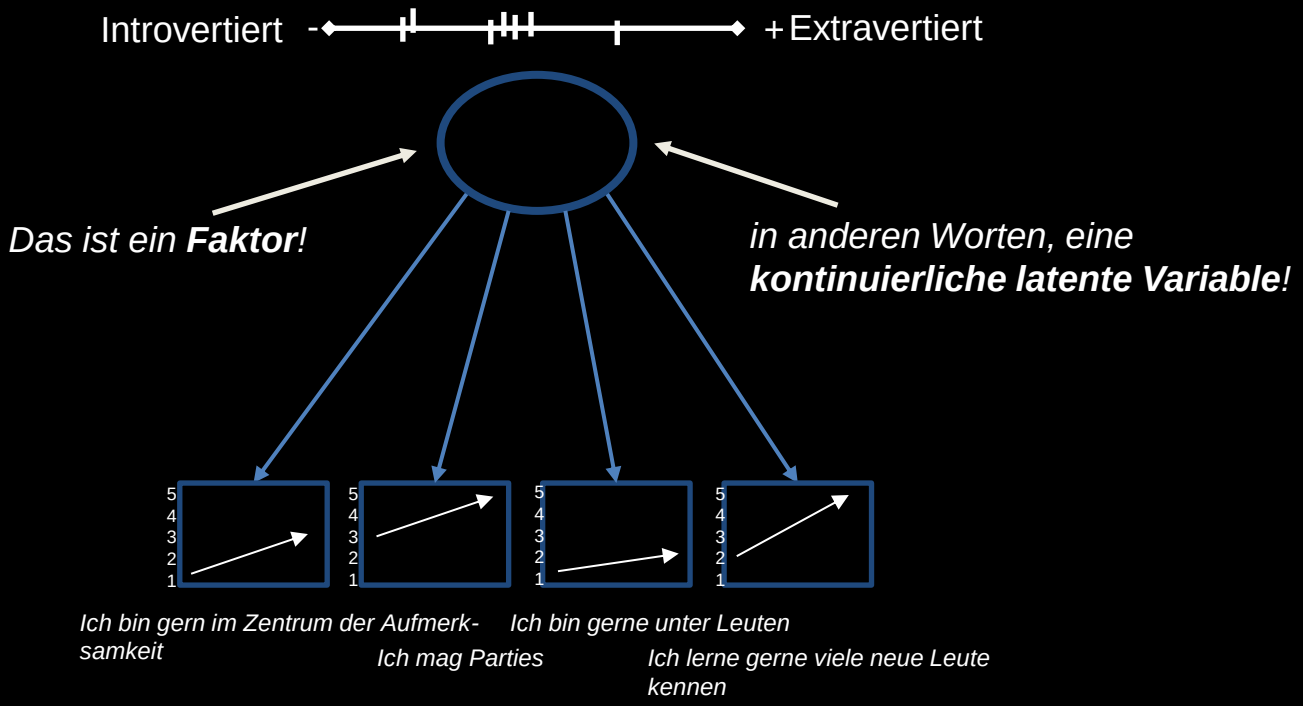
0. Summenscore entspricht der Annahme **eines latenten Faktormodells** mit gleichen Faktorladungen



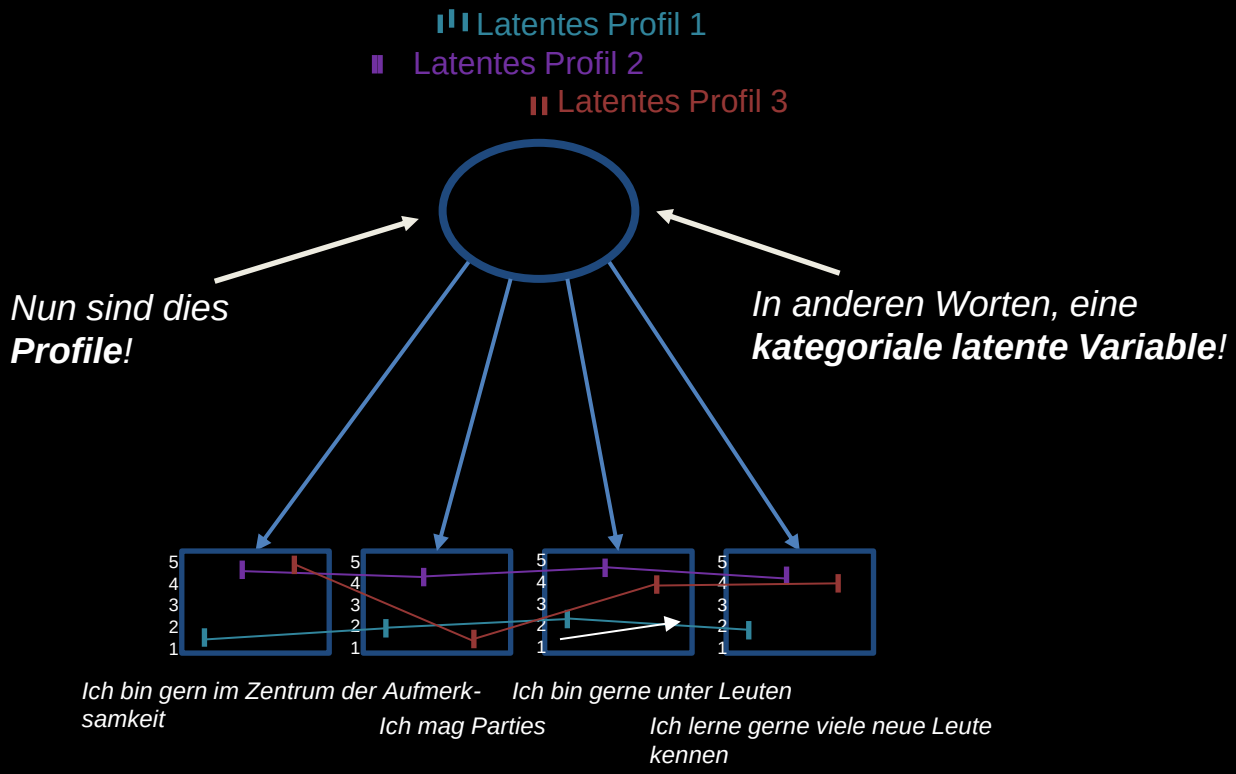
Das sag ich

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





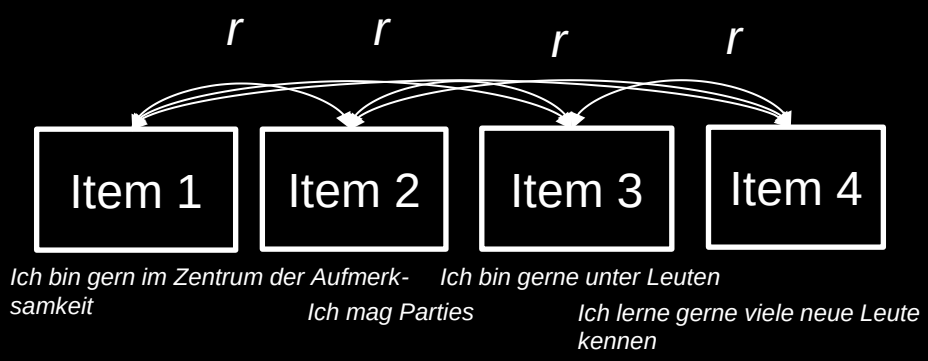
Höhere Extraversion: Erwartete Mittelwerte in Richtung zustimmender Antworten erhöhren sich linear (& proportional zu Faktorladung)



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

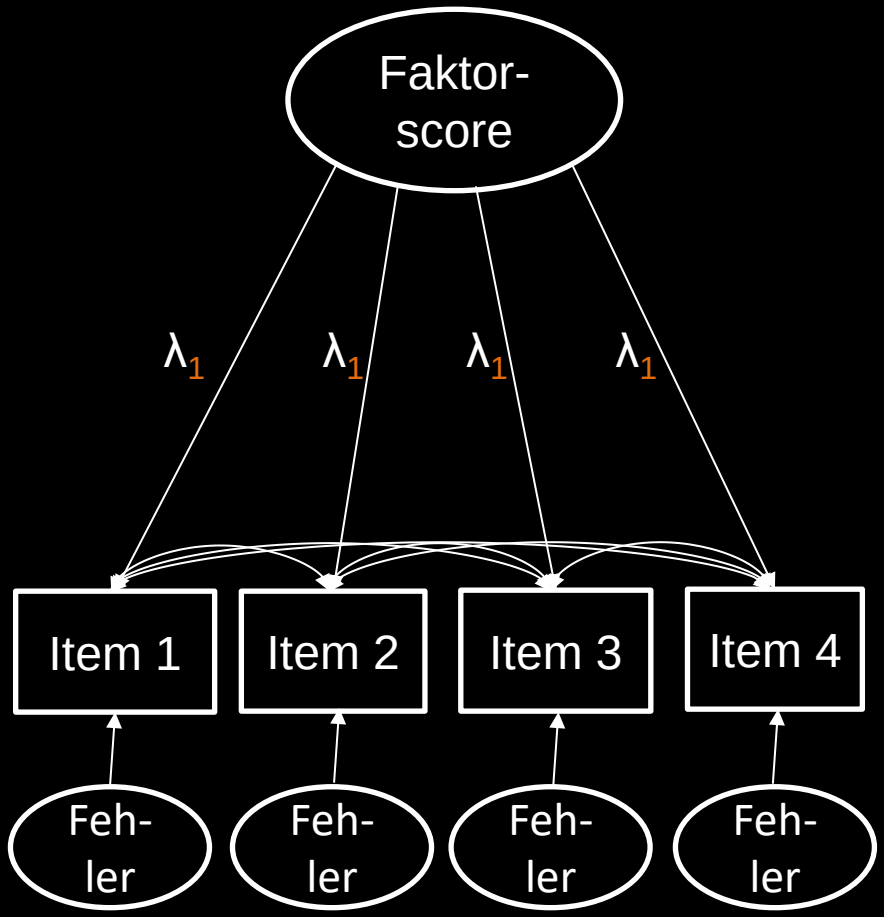
Das sag ich

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



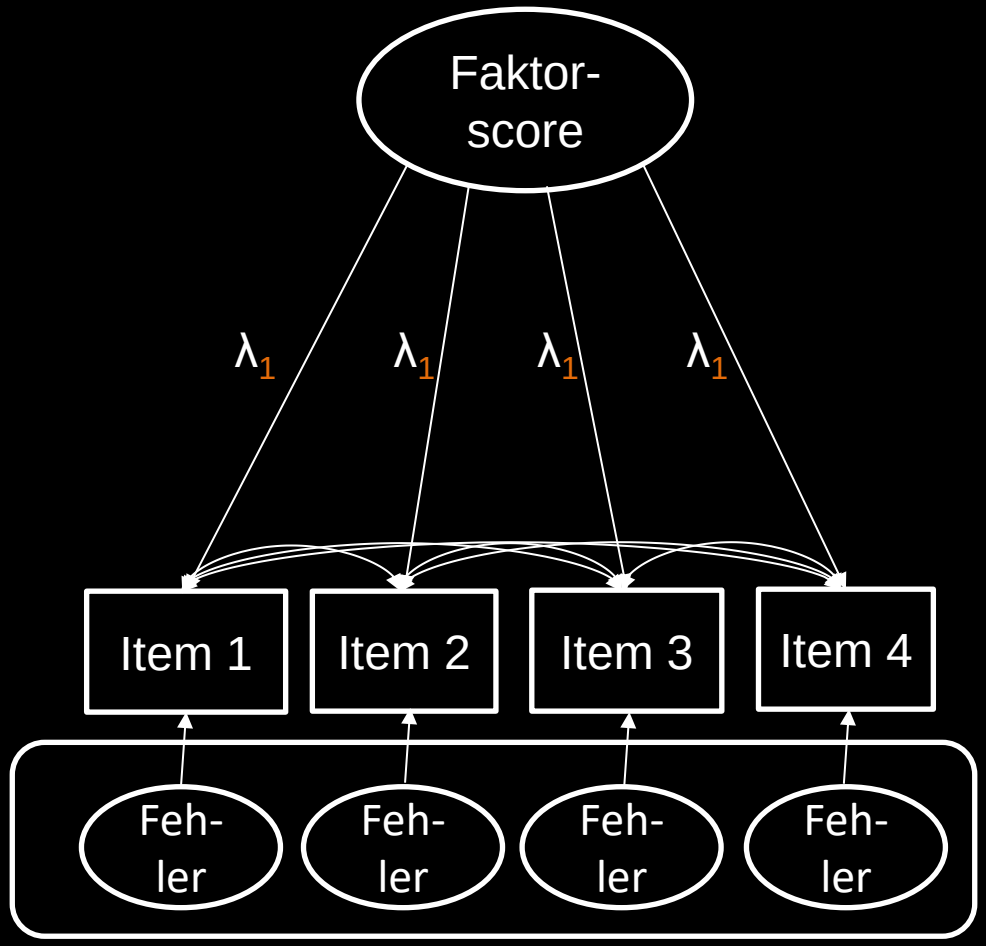
Das sag ich

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



Das sag ich

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



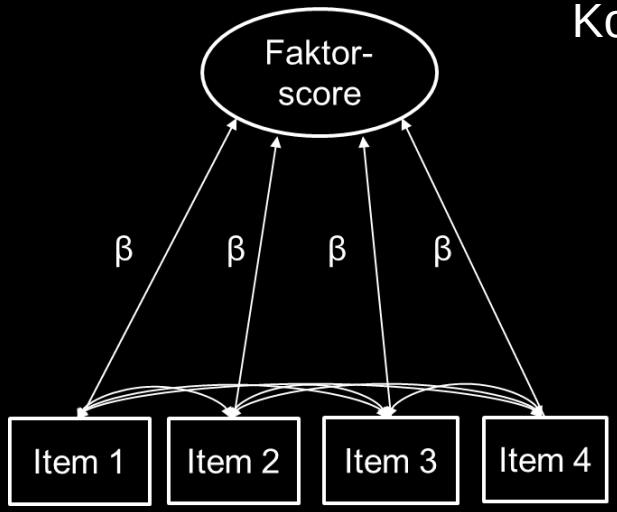
Kontext/Inhalte:
Austauschbar und
konstruktirrelevant?

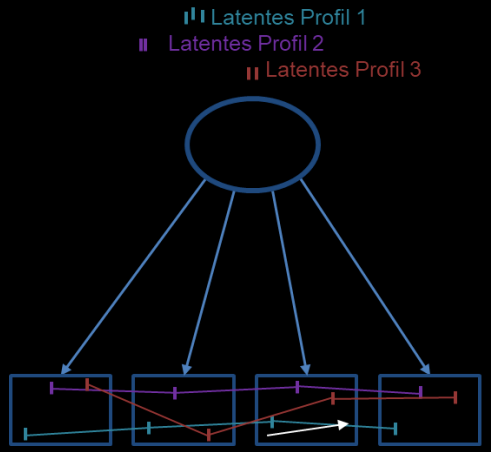
Das sag ich

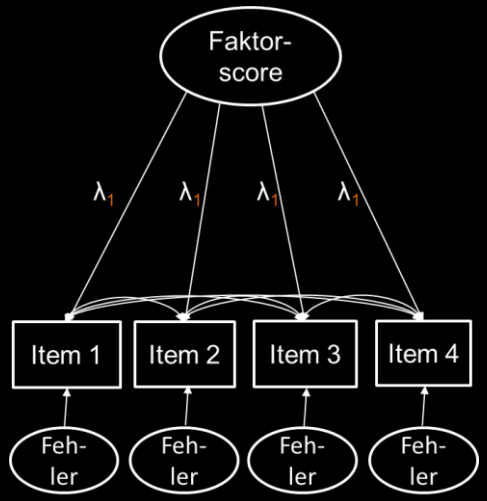
0. Summenscore entspricht der Annahme **eines latenten Faktormodells** mit gleichen Faktorladungen

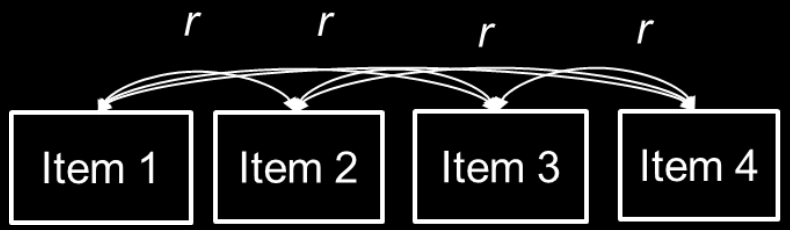
und die **Empirie**?

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

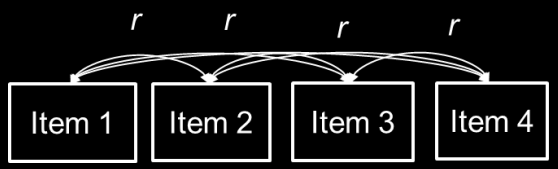
0. Latentes Faktormodell mit homogenen Faktorladungen

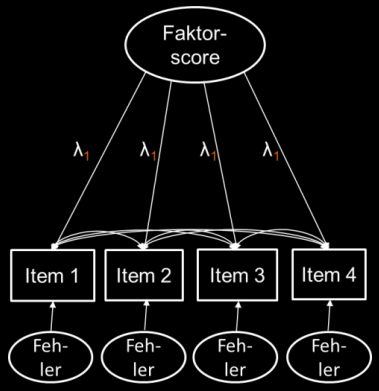
Summenscore

?

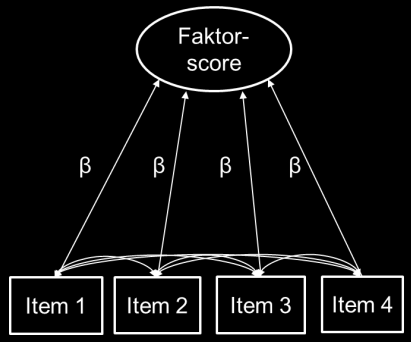
Affirmation der Konsequenz

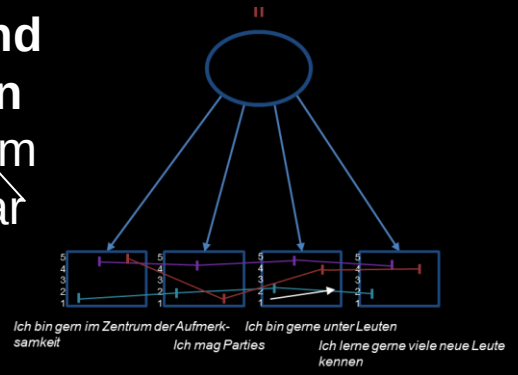
Umkehrfehler

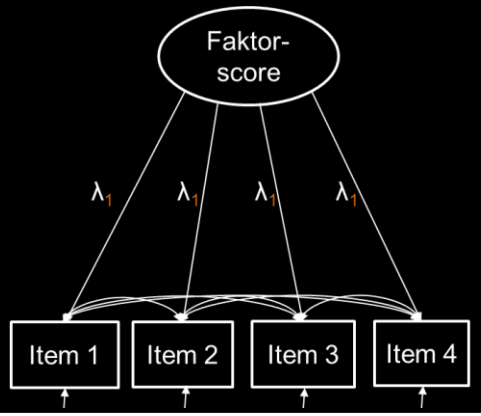




Modelle implizieren **dieselben Mittelwerte und Korrelationen** Empirisch kaum unterscheidbar







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

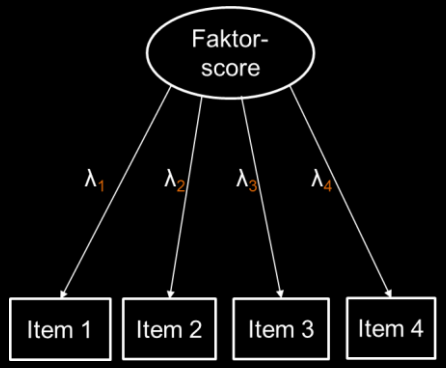
Hohe interne Konsistenz von Tests wird häufig gewünscht

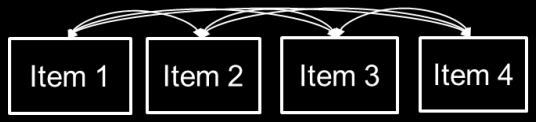
z.B. Cronbach’s Alpha (oder Omega)

Taber 2018 (& Stadler et al., 2021)

Für Wissenstests inadäquat

Theoretische Annahme:
Heterogen und **Mehrdimensional**



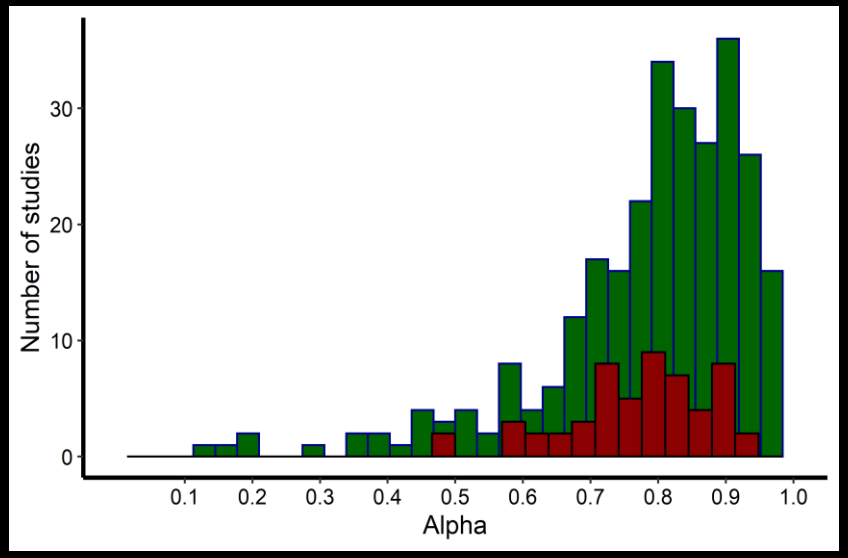


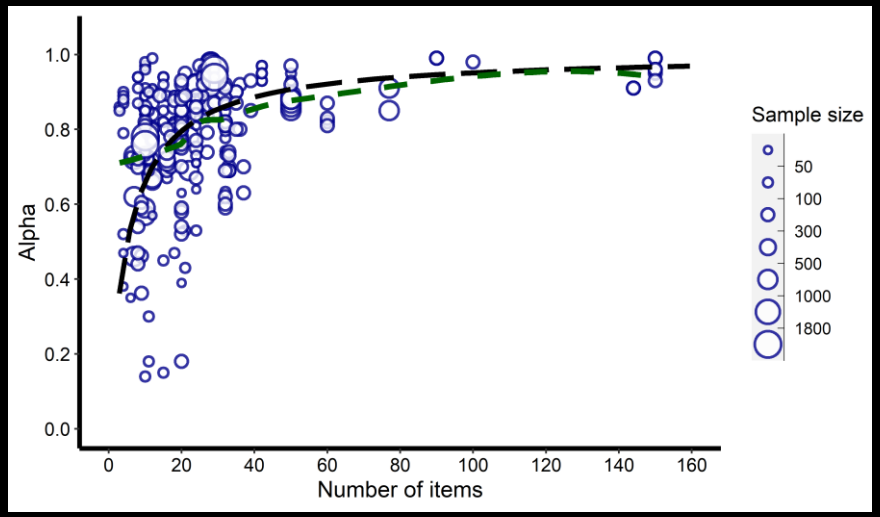
Beispiel:

Inhaltswissen

Meta-Analyse:

Mittlere Interkorrelation Wissensitems
r = .22

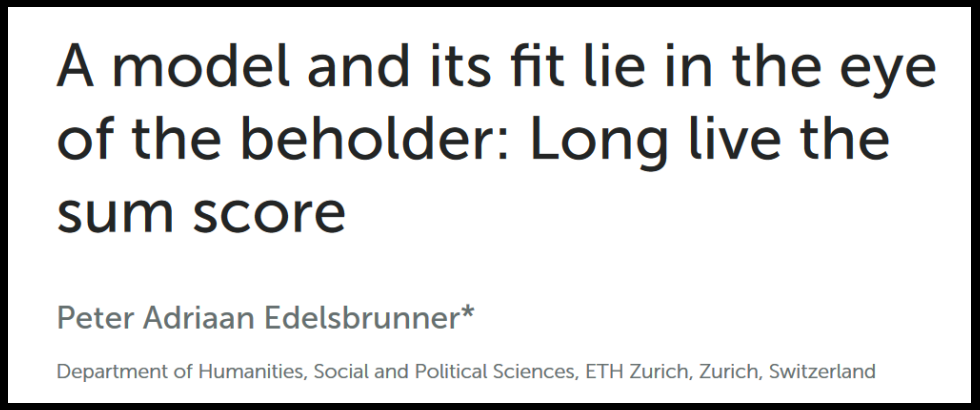




	CTT	Rasch	IRT	CFA	EFA	G-Theory	Netz-werk	Mokken	LCA/ LPA
Reliabilitätsschätzung	+	~	~	~	-	+	-	~	~
Dimensionalitätsprüfung	-	~	~	+	+	-	~	+	~
Globaler Fit	-	~	~	+	-	-	-	-	-
Item-/Personenfit		+	-	~		-	-		-
Bivariate Abhängigkeiten			~	~	~		+		
Nicht-Linearitäten/Subgruppen								~	+
Annahmenverletzung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

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

Danke

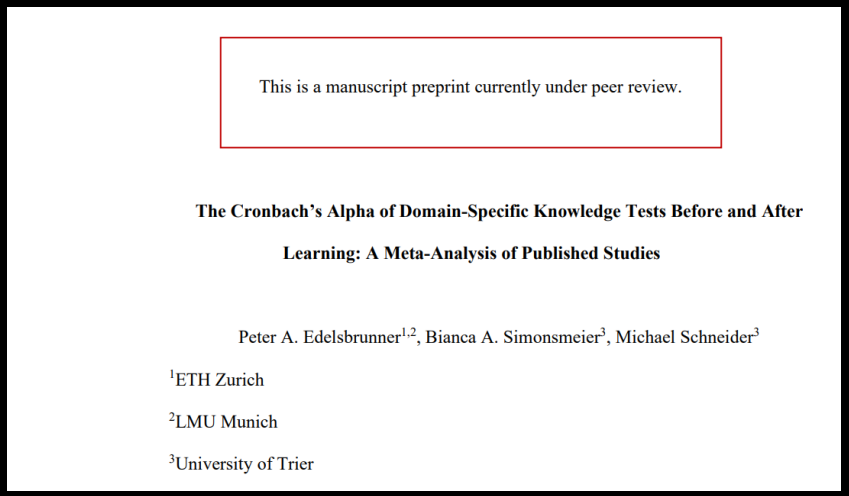


A model and its fit lie in the eye of the beholder: Long live the sum score

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

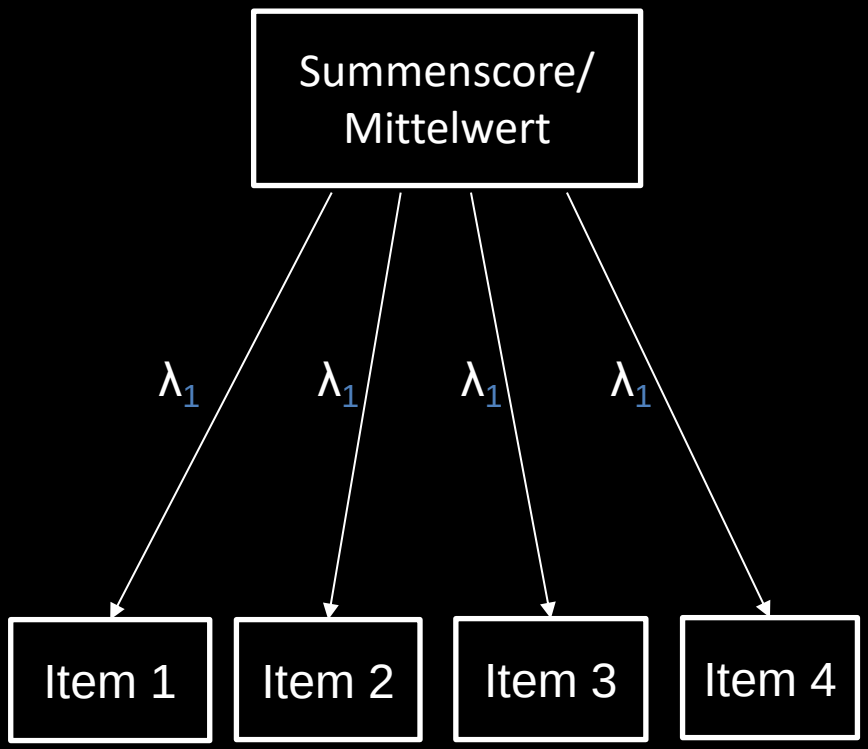
Daniel McNeish[1] · Melissa Gordon Wolf[2]

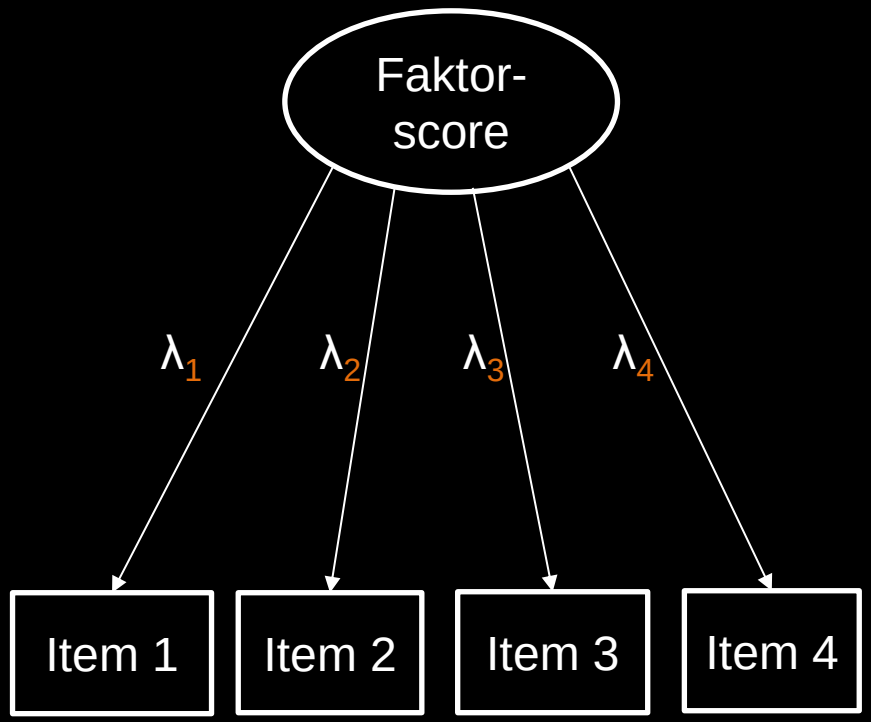
Published online: 22 April 2020

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:





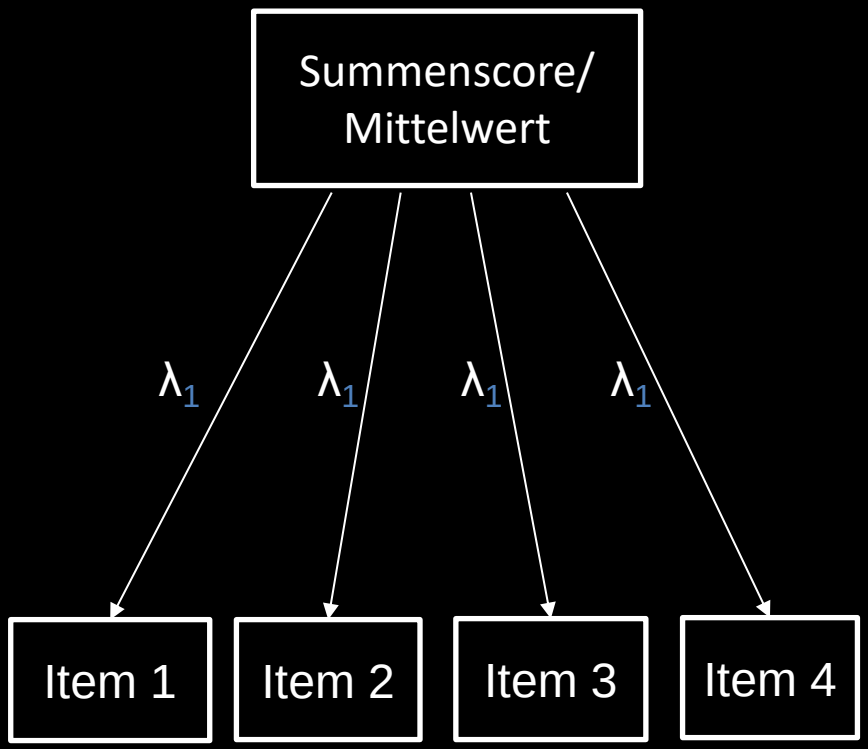
The sum score is a *constrained version* of factor analysis
(McNeish & Wolf, 2020)

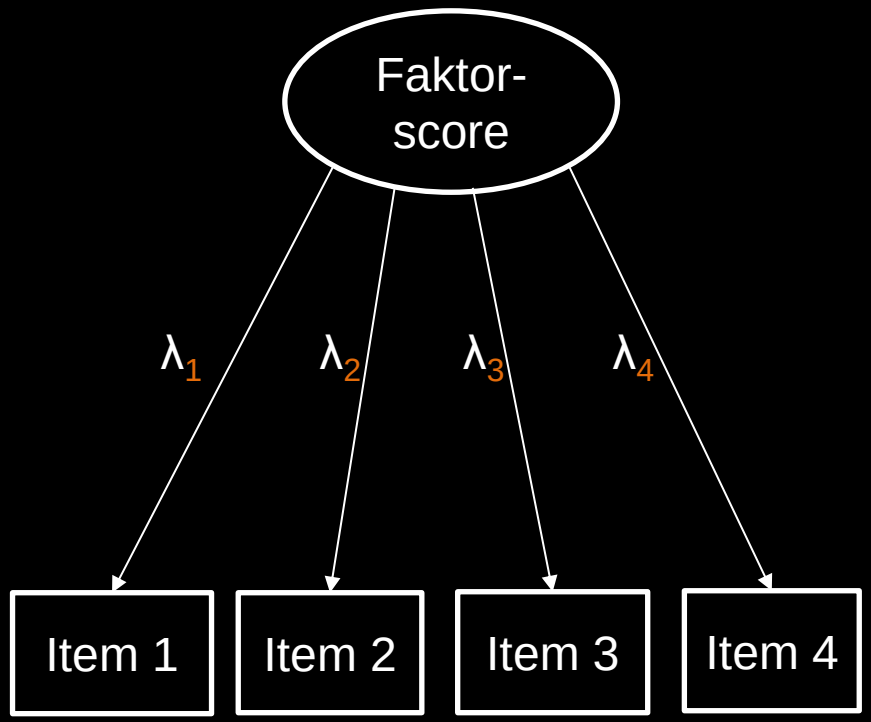
Alle Items werden *gleich gewichtet*

Faktorscore:
Nach Faktorladung gewichteter Wert aus allen Items

Entspricht implizit Faktorenanalyse mit gleichen Faktorladungen

Option 1 = Option 2:





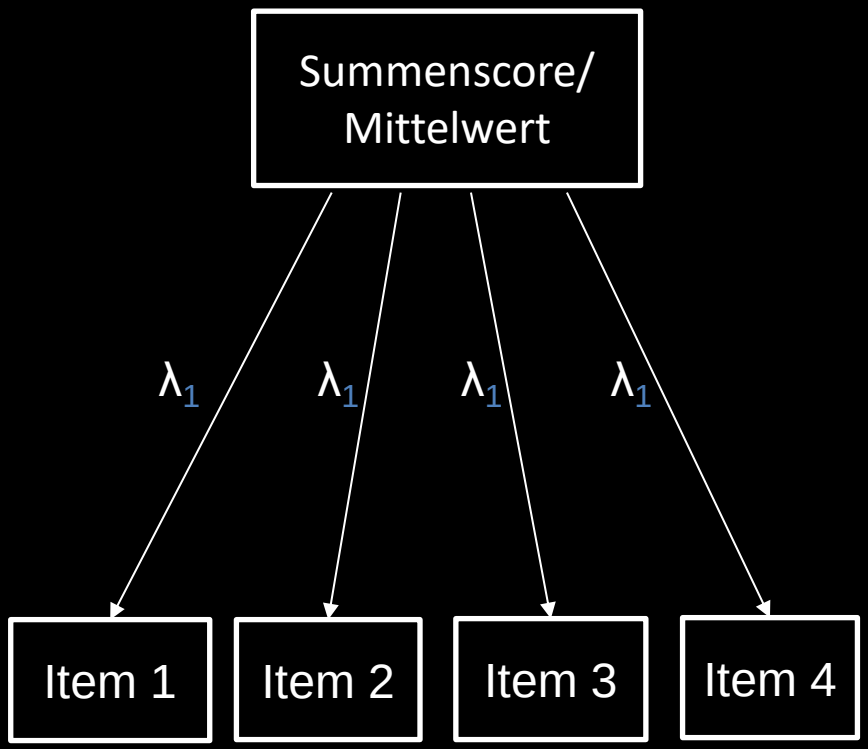
The sum score is a *constrained version* of factor analysis
(McNeish & Wolf, 2020)

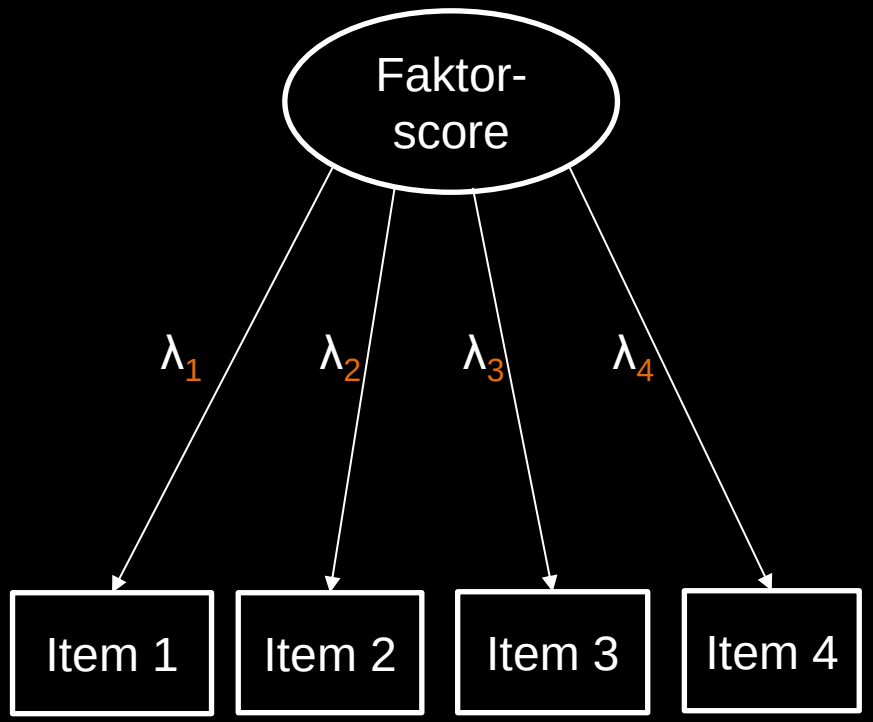
Alle Items werden *gleich gewichtet*

Entspricht implizit der Annahme gleicher Faktorladungen

Faktorscore:
Nach Faktorladung gewichteter Wert aus allen Items

Option 1 = Option 2:





Alle Items werden *gleich gewichtet*

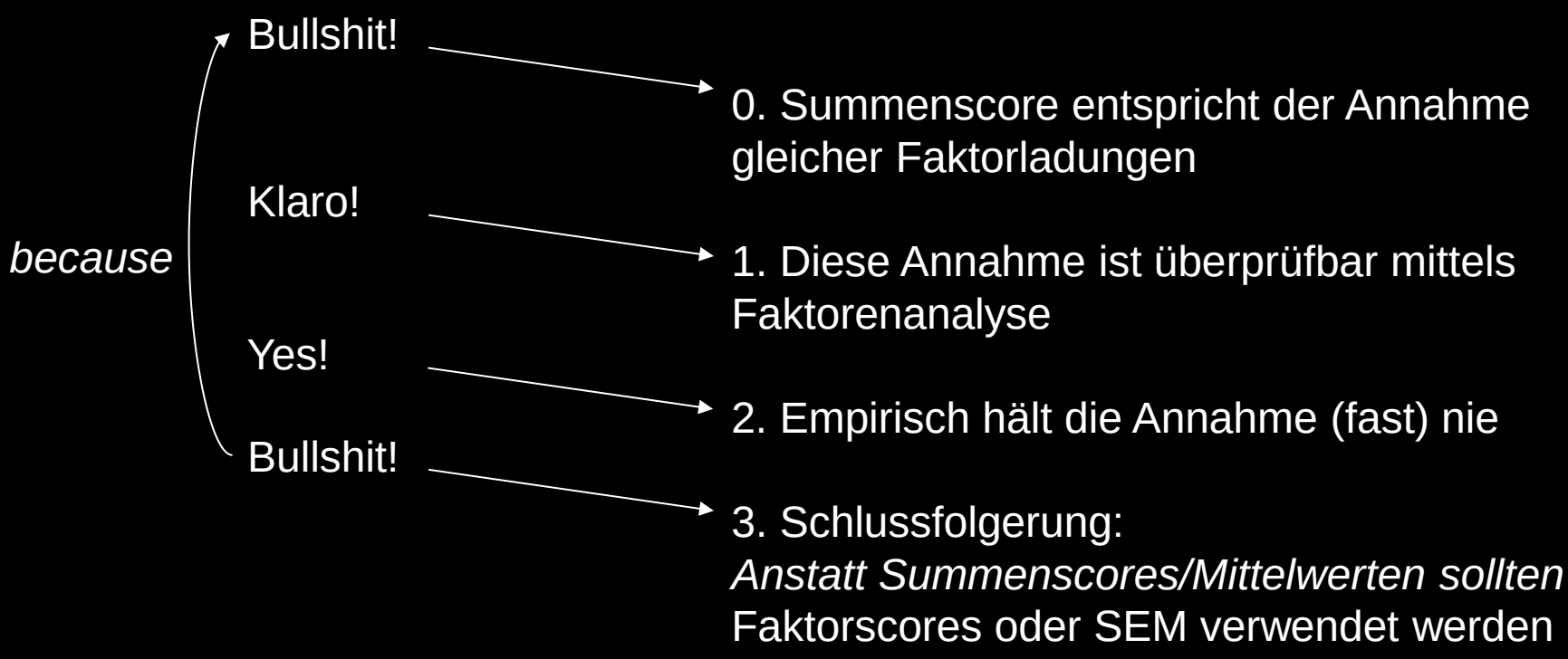
Entspricht implizit der *Annahme gleicher Faktorladungen*

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!