Modeling Knowledge

Peter Edelsbrunner

bit.ly/PeterE_presentations

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.

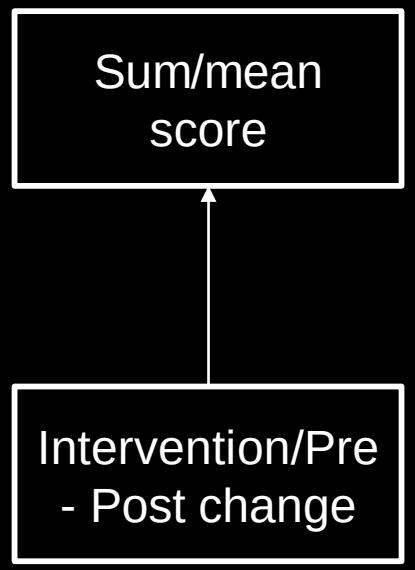
11. The following three balls roll on a horizontal plane:

- 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.

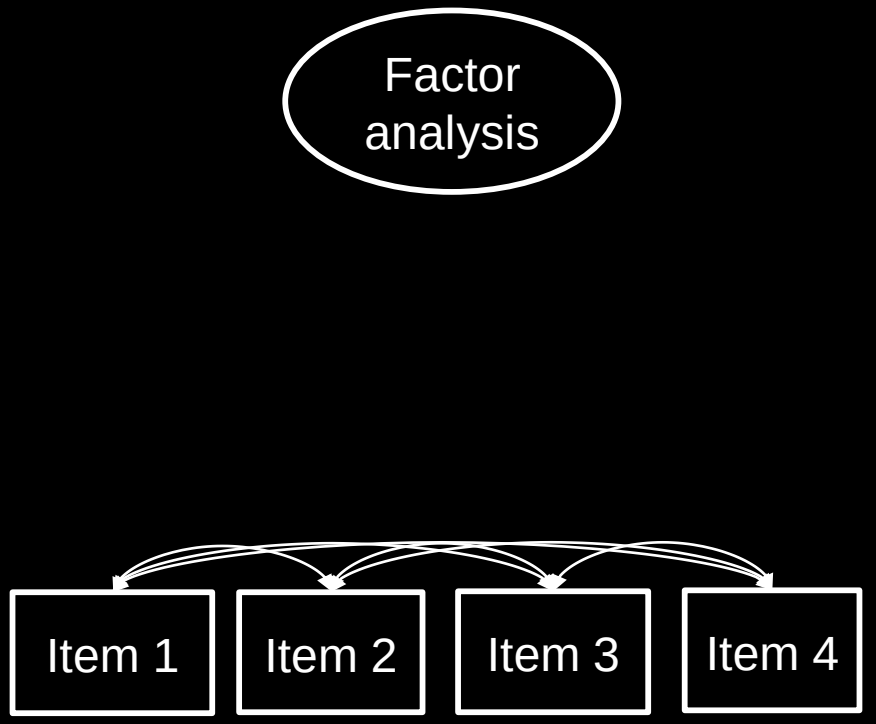
Which of the following statements are true?

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

Option 1:

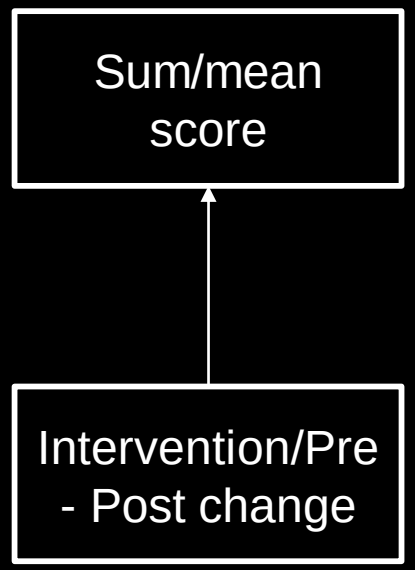


Option 2:

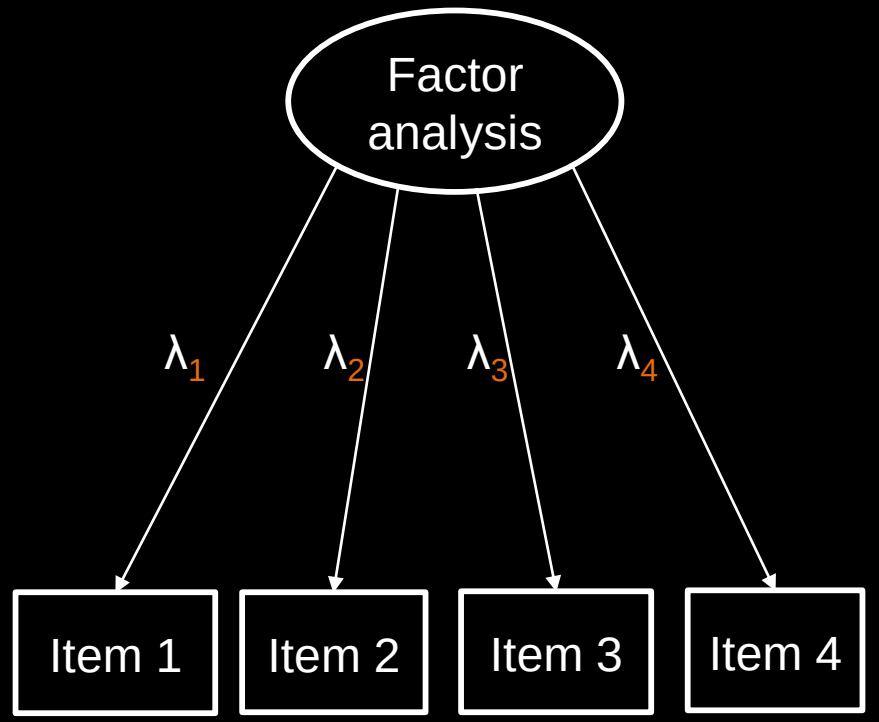


Determination of number of required *sources of common variation* (factors) to explain/model intercorrelations between items

Option 1:



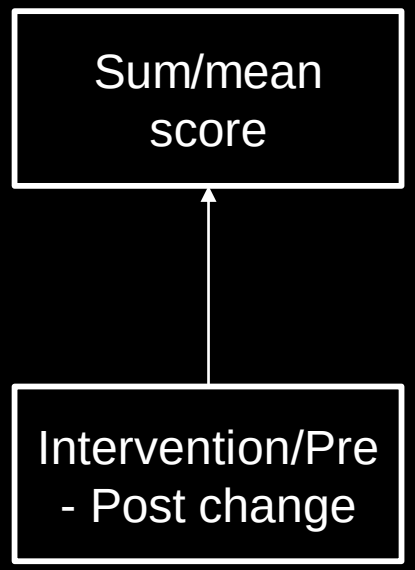
Option 2:



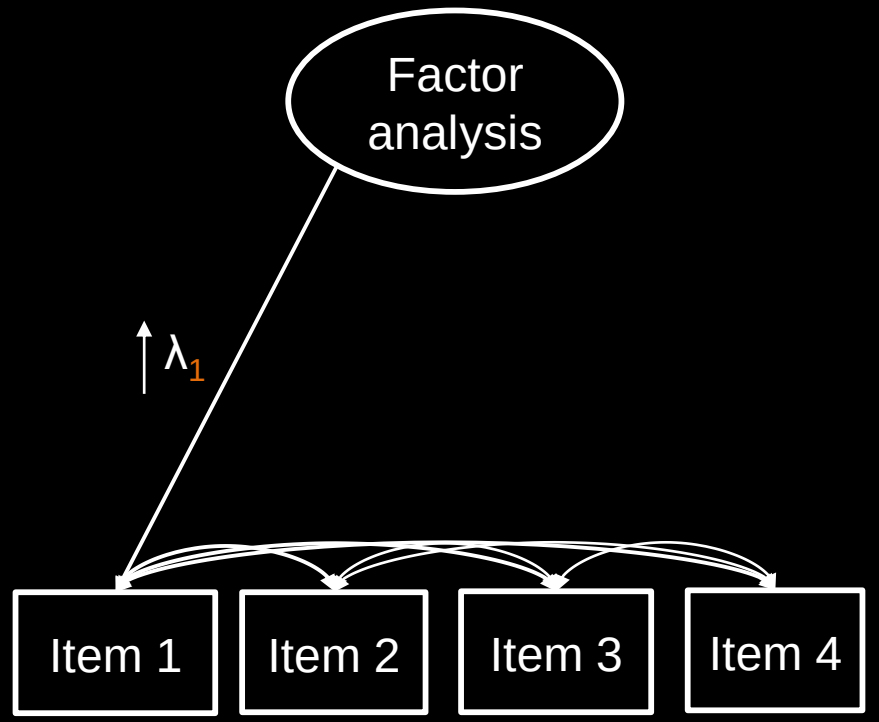
Determination of number of required *sources of common variation* (factors) to explain/model intercorrelations between items

Estimation of strength, by which the common variance goes into each item (factor loading λ_1-λ_4)

Option 1:



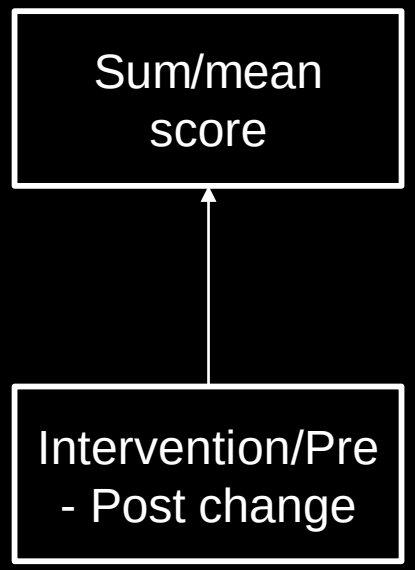
Option 2:



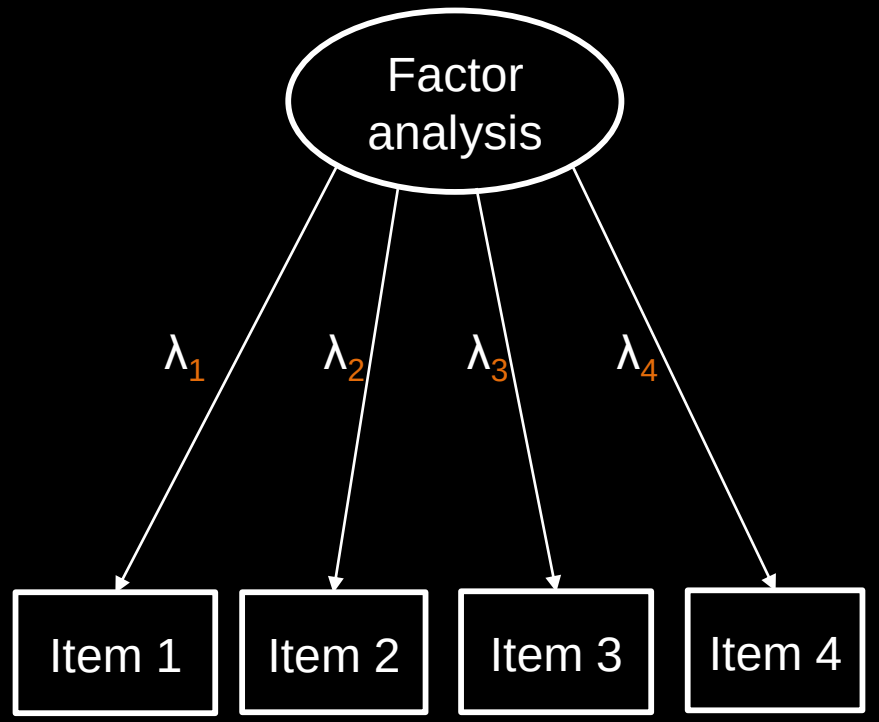
Items that show *high intercorrelations* with the other items receive *high factor loadings* (strong indicators of common construct)

Estimation of strength, by which the common variance goes into each item (factor loading λ_1-λ_4)

Option 1:



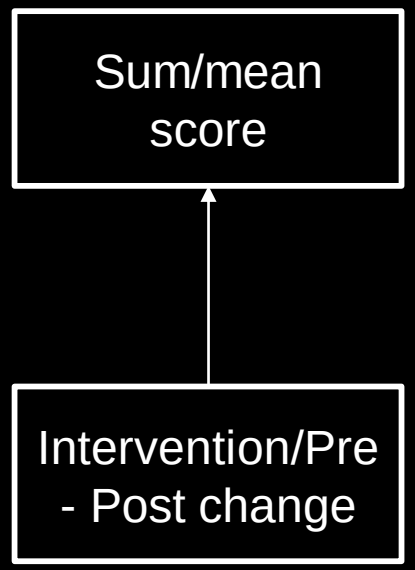
Option 2:



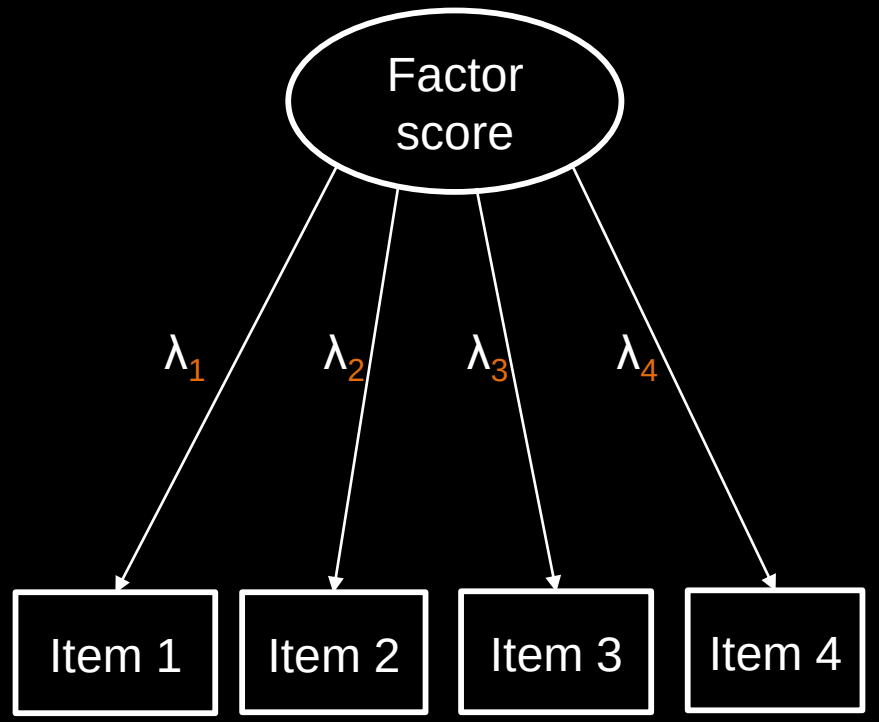
As *measurement model*:
Examine whether the *theoretically expected factor structure* (number & loadings) is present

As *scaling model*:
Build estimated factor scores from measurement model

Option 1:



Option 2:



Factor score:
Weighted value by factor loading across items

As *scaling model*:
Build estimated factor scores from measurement model

Option 1:

Sum/mean score

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

Avoiding bias from sum scores in growth estimat
An examination of IRT-based approaches to scor
longitudinal survey responses

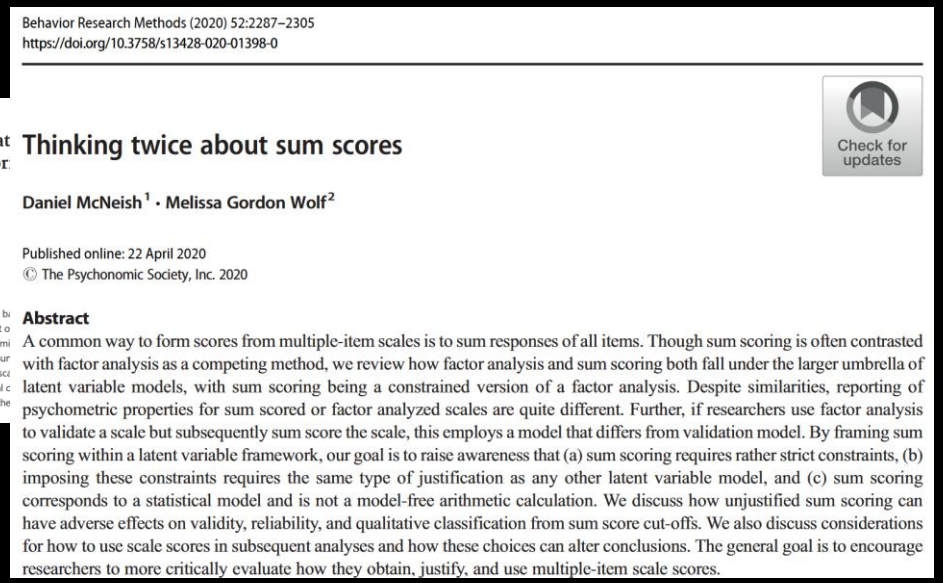
Megan Kuhfeld [1], James Soland [2]

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 b
survey data. Researchers use longitudinal growth modeling to understand the development o
students on psychological and social-emotional learning constructs across elementary and mi
school. In these designs, students are typically administered a consistent set of self-report sur
items across multiple school years, and growth is measured either based on sum scores or sca
scores produced based on item response theory (IRT) methods. Although there is great deal c
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

Check for updates

Thinking twice about sum scores

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

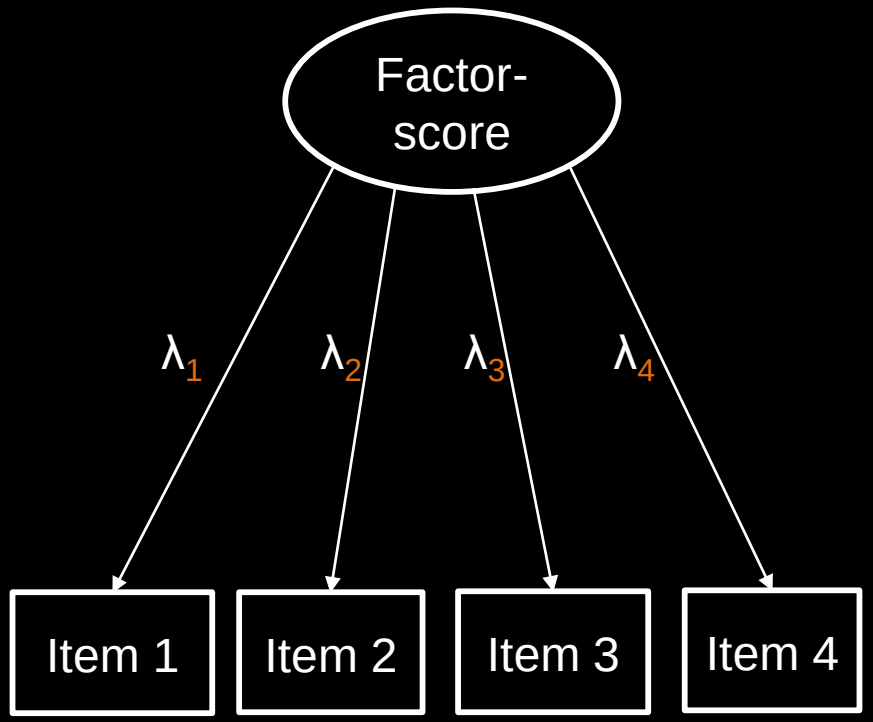
Published online: 22 April 2020

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.

Criticism of *Sum + Alpha*
(+ factor analysis) - approach

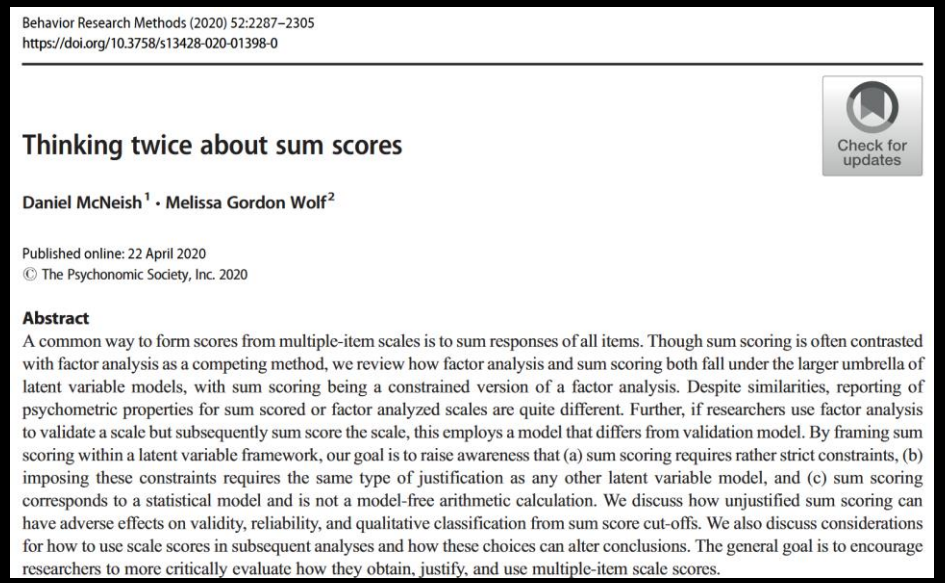
Option 2:



Factor score:
Weighted value by factor loading across items

Option 1:

Sum/mean score



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]

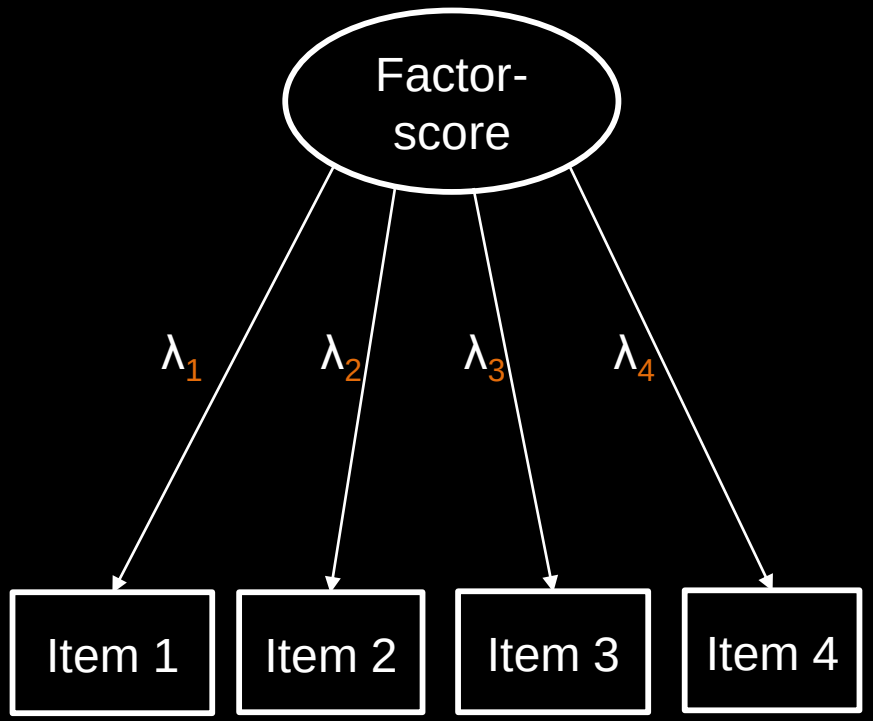
Published online: 22 April 2020

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.

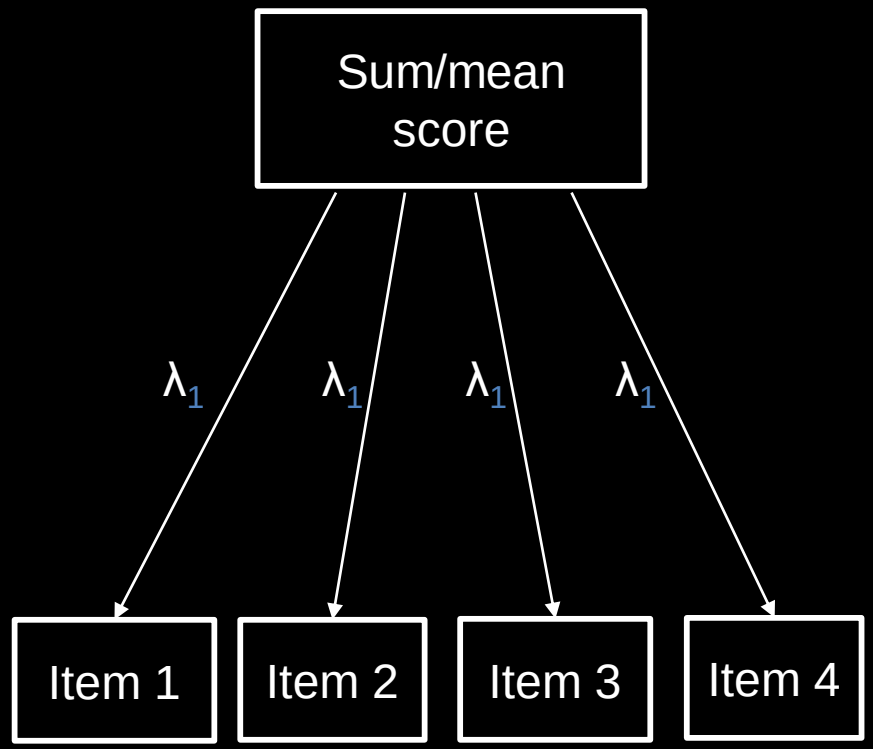
The sum score is a *constrained version* of factor analysis
(McNeish & Wolf, 2020)

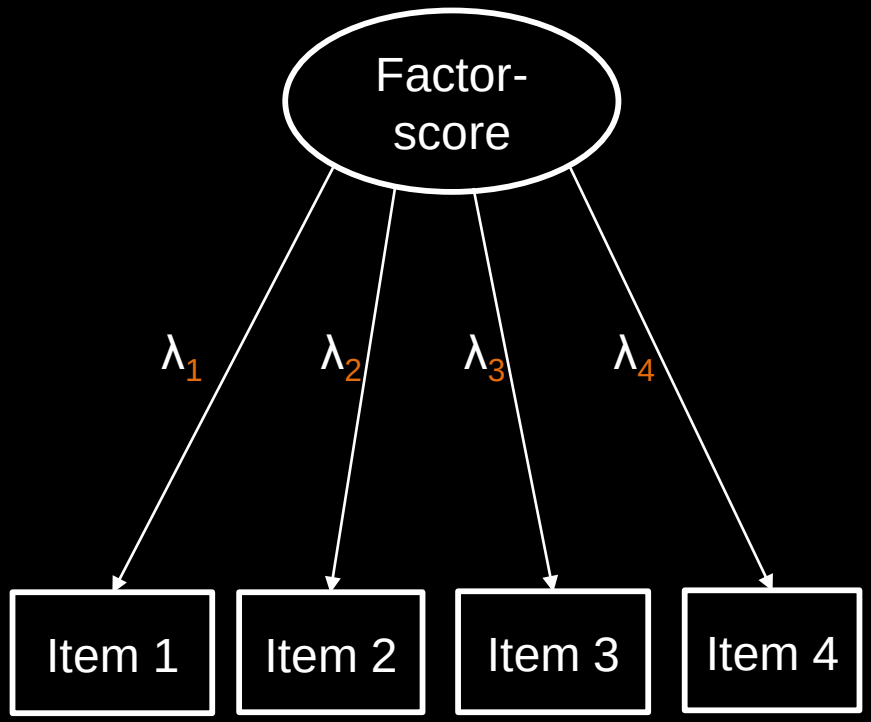
Option 2:



Factor score:
Weighted value by factor loading across items

Option 1 = Option 2:





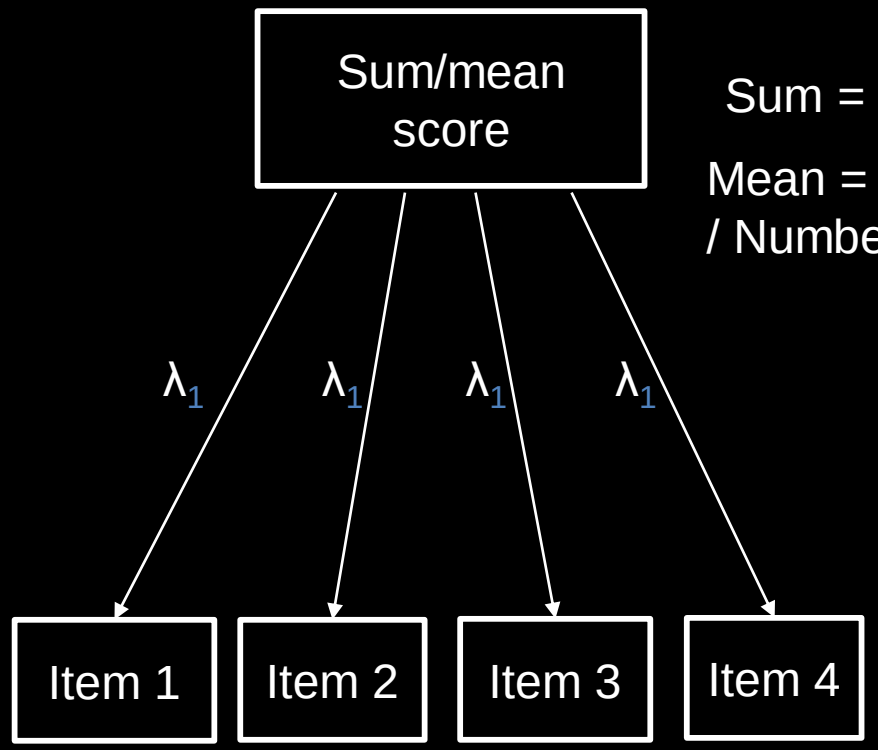
The sum score is a *constrained version* of factor analysis
(McNeish & Wolf, 2020)

Factor score:
Weighted value by factor loading across items

All items are *equally weighted*

Implies the assumption of equal (std.) factor loadings

Option 1



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

Mean = (1*Item 1 + 1*Item 2 + 1*Item 3 + 1*Item 4) / Number of items

0. Sum score implies the assumption of equal loadings

1. This assumption can be tested via factor analysis

2. Empirically this assumption is (almost) never given

3. Conclusion:
Instead of sum scores/means we should use factor scores/SEM
(IRT/person estimtes)

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

All items are *weighted equally*

Implies the assumption of equal (std.) factor loadings

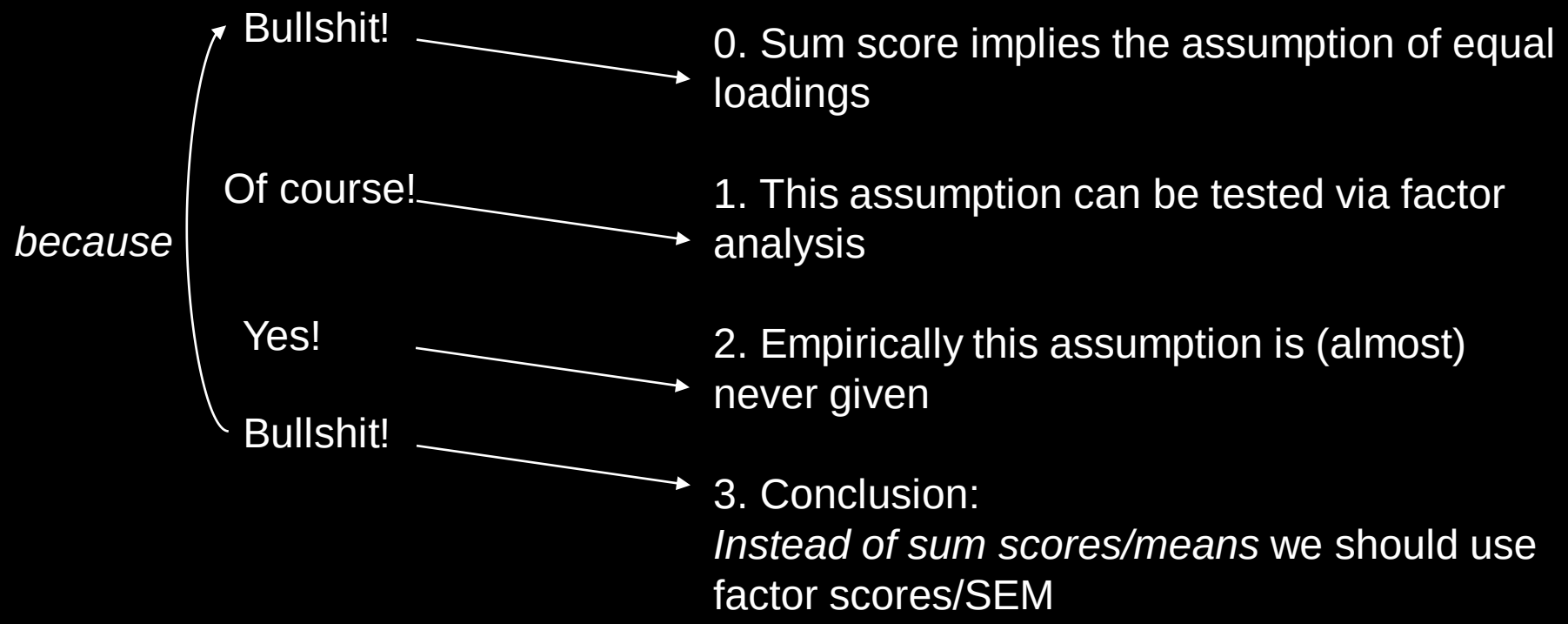
Should we generally use factor scores (or SEM/IRT) instead of sum/mean scores?

0. Sum score implies the assumption of equal loadings

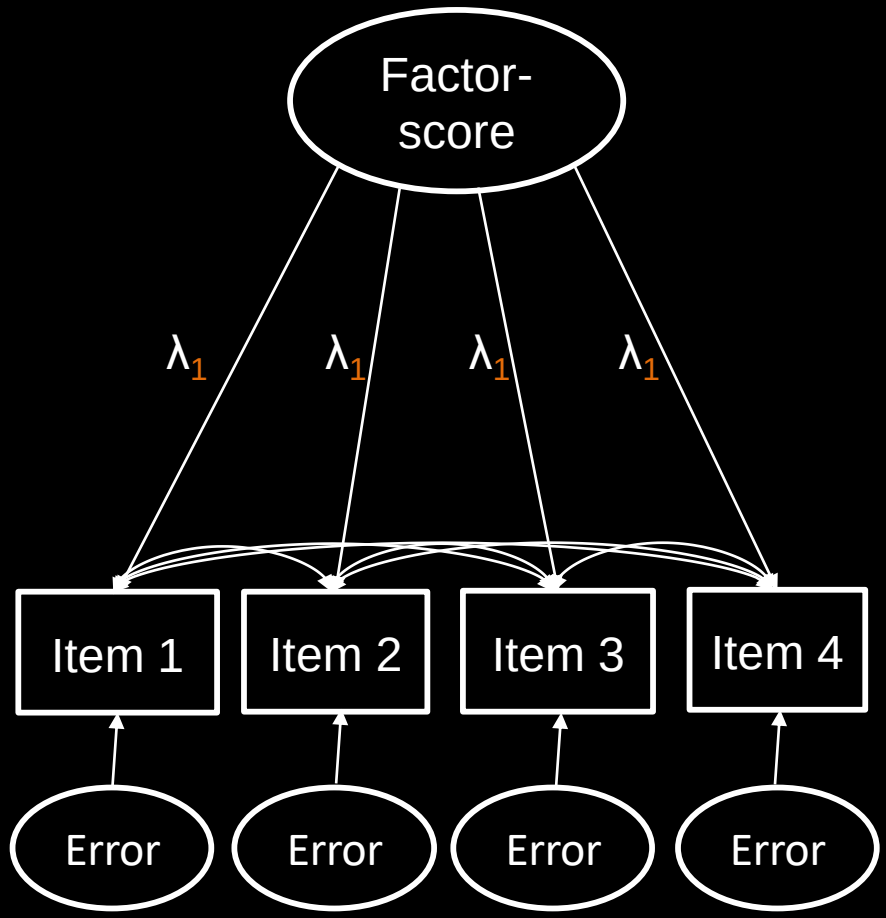
1. This assumption can be tested via factor analysis

2. Empirically this assumption is (almost) never given

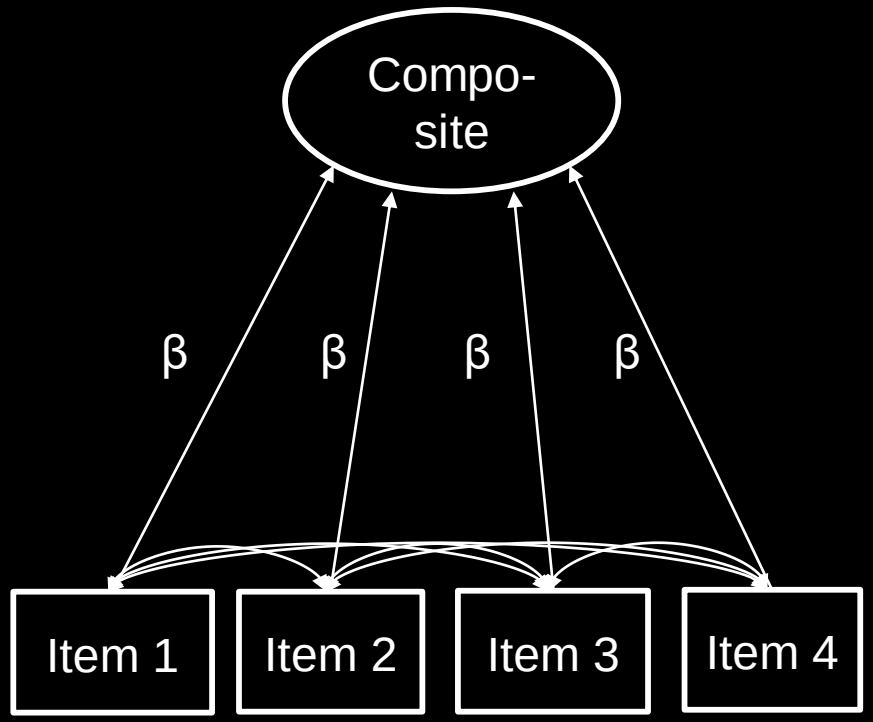
3. Conclusion:
Instead of sum scores/means we should use factor scores/SEM


Bullshit!
Of course!
Yes!
Bullshit!
because
0. Sum score implies the assumption of equal loadings
1. This assumption can be tested via factor analysis
2. Empirically this assumption is (almost) never given
3. Conclusion:
Instead of sum scores/means we should use factor scores/SEM

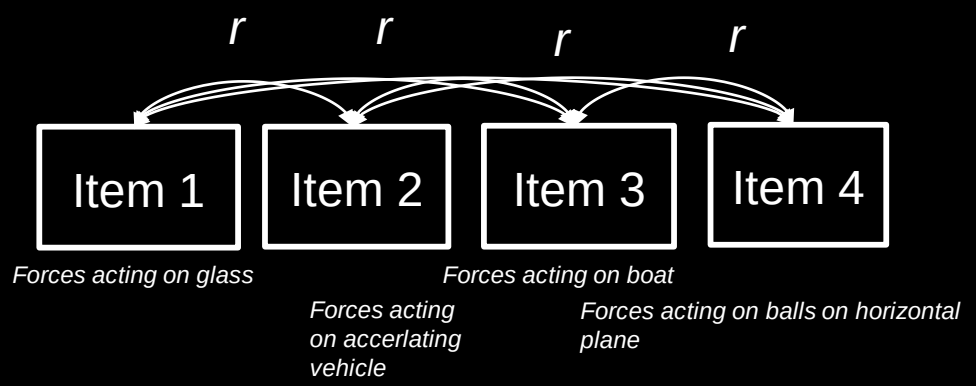
0. Sum score implies the assumption of **a factor model** with equal loadings



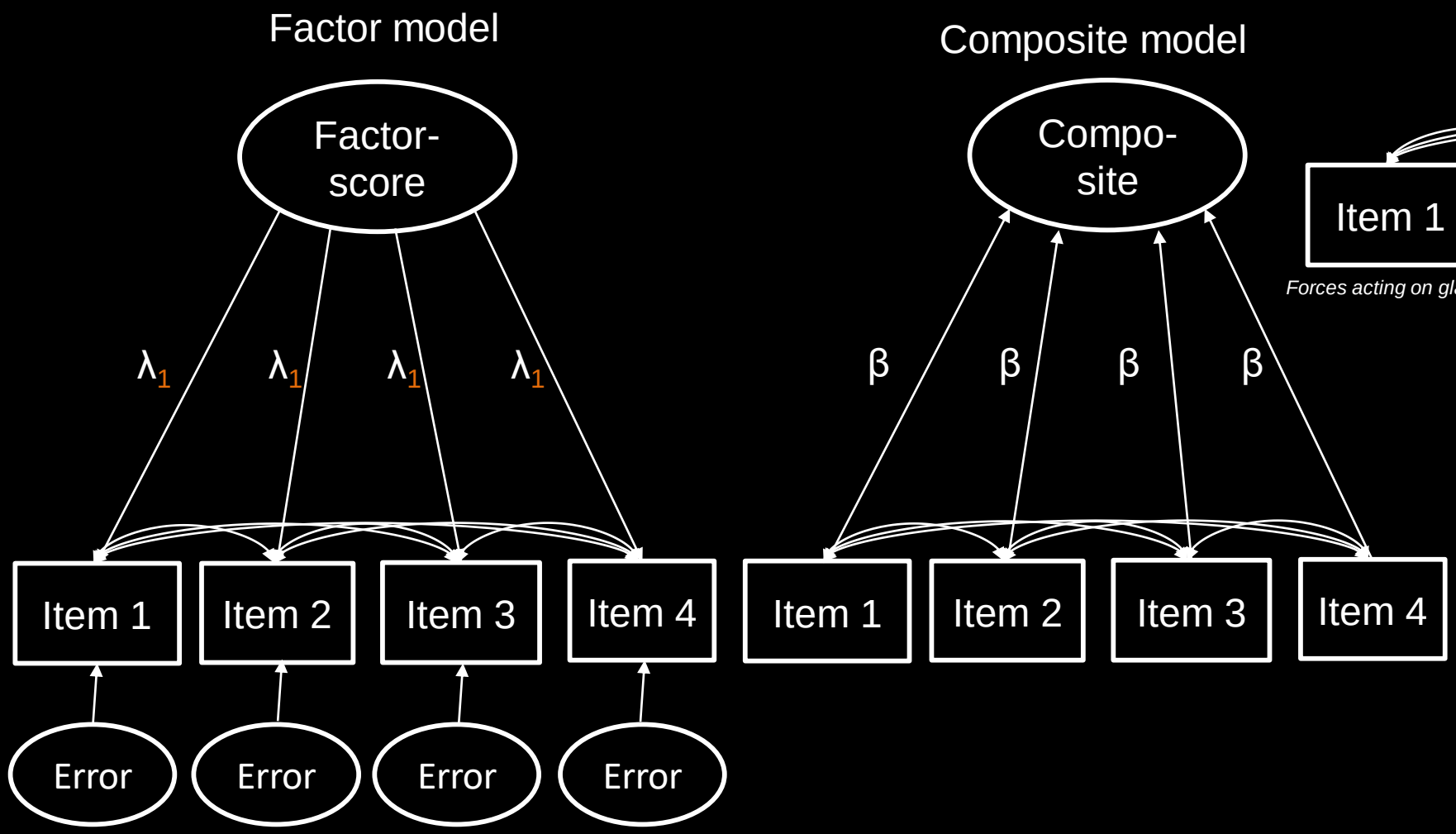
0. Sum score implies the assumption of **a composite model**

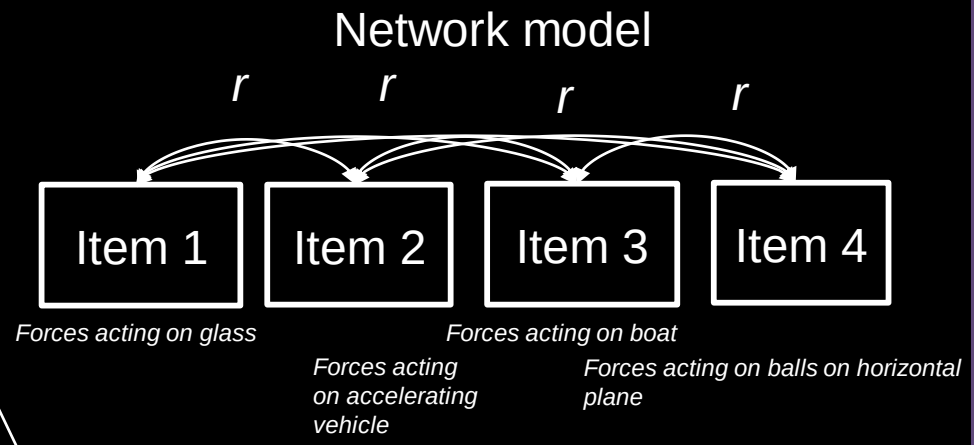


0. Sum score implies the assumption of **a network model**

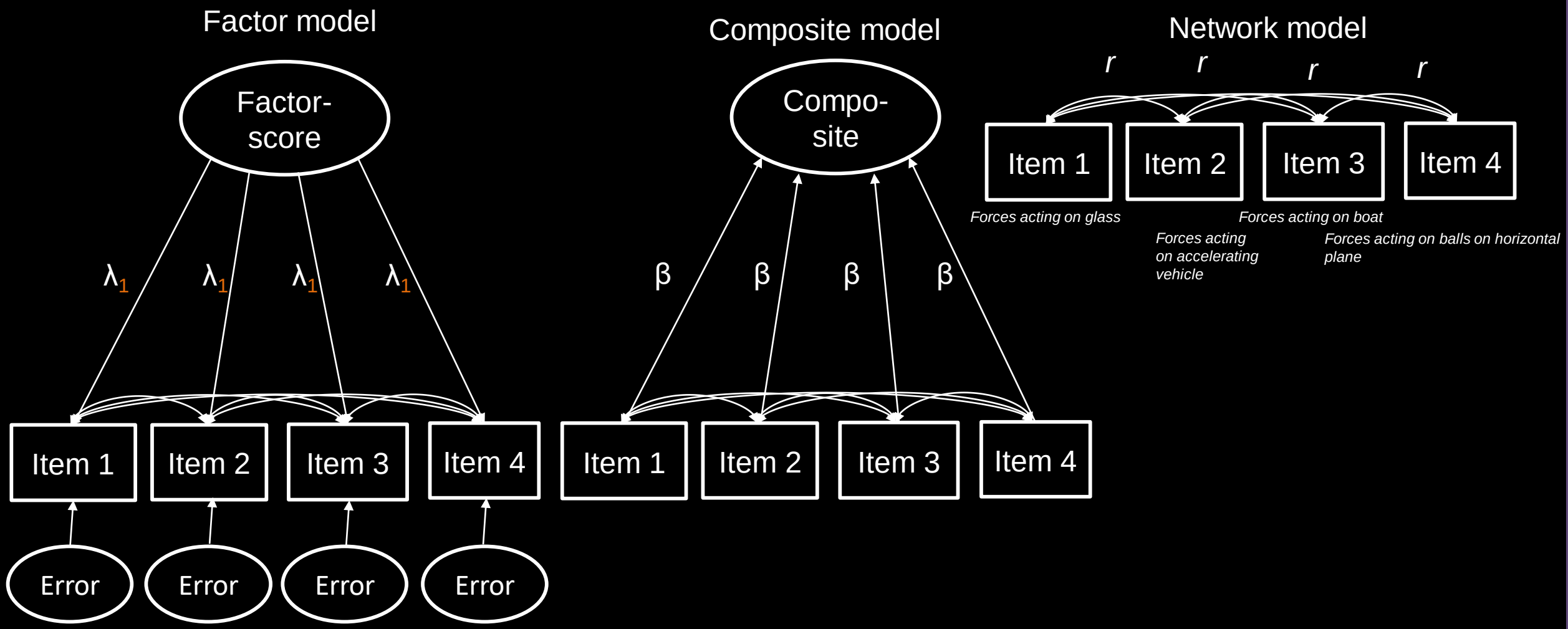


Empirically (almost) indistinguishable (Edelsbrunner, 2022 for references)

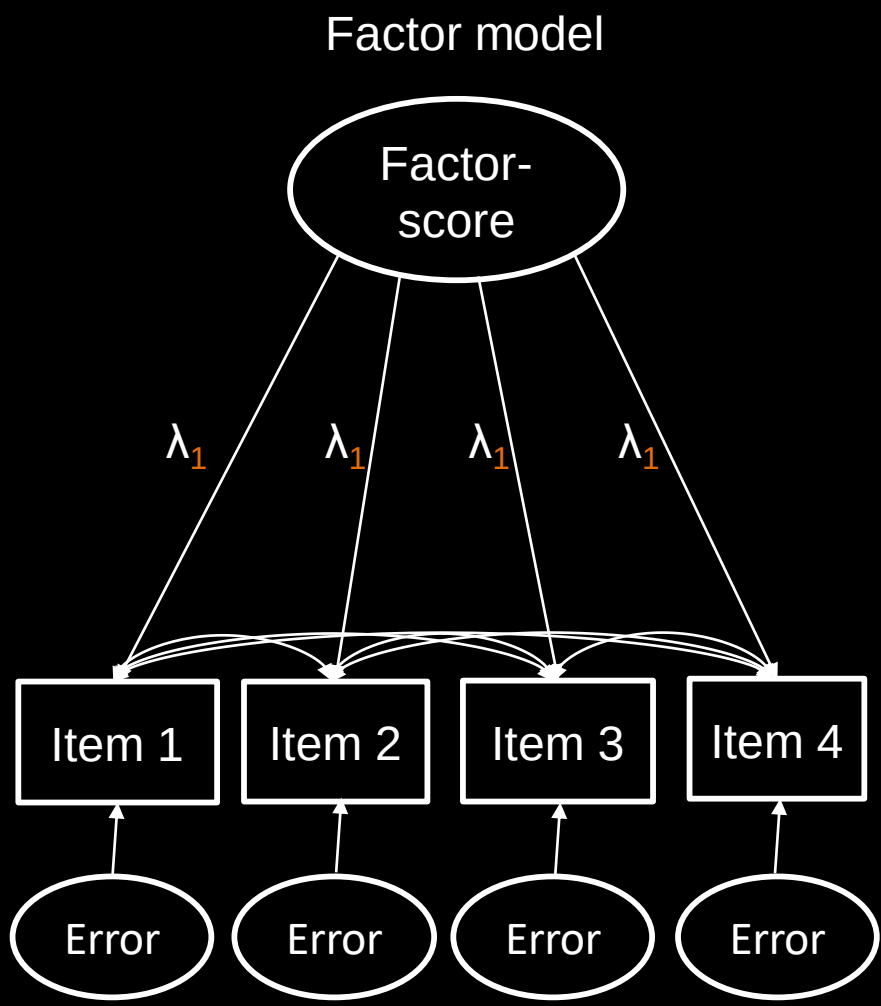


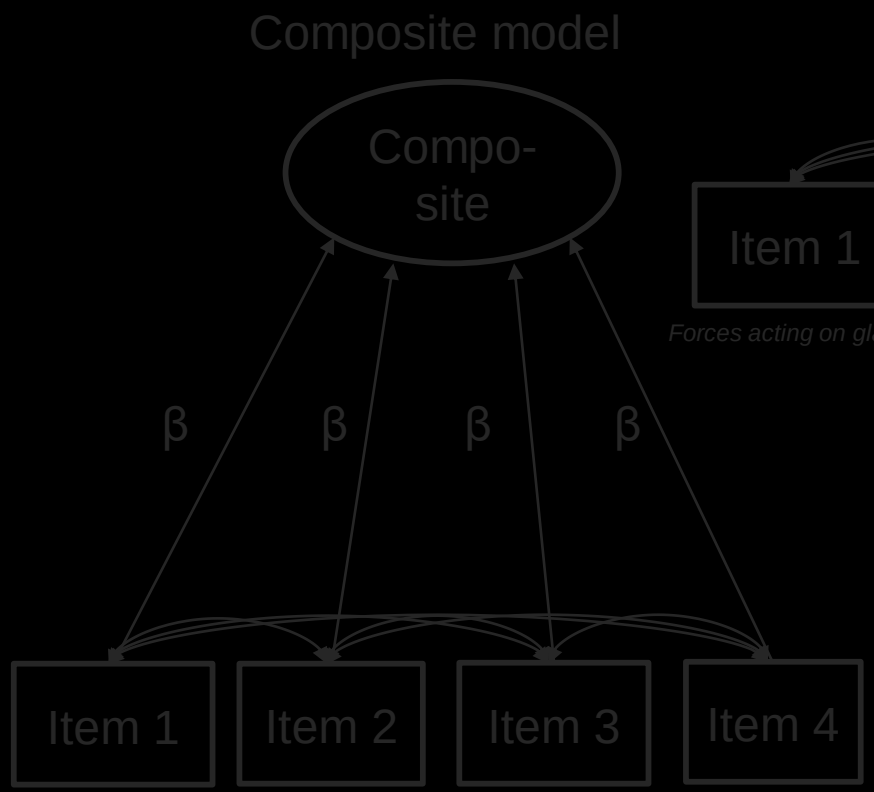


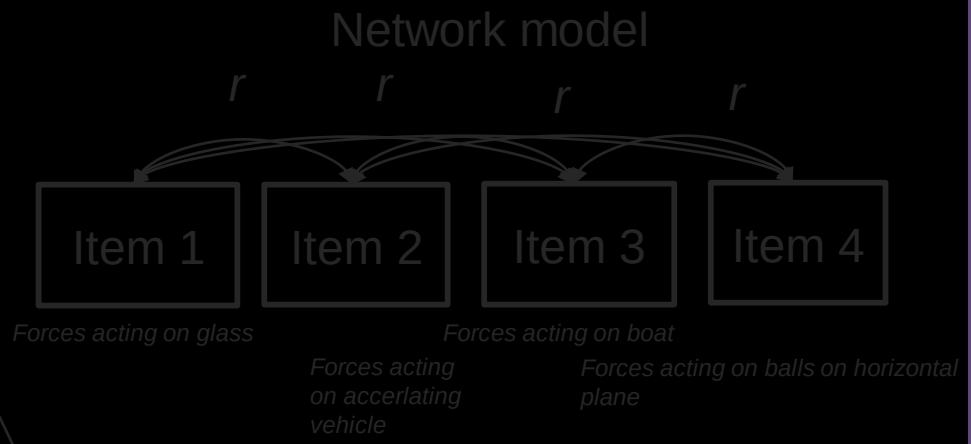
Theoretical distinction possible (Edelsbrunner, 2022)



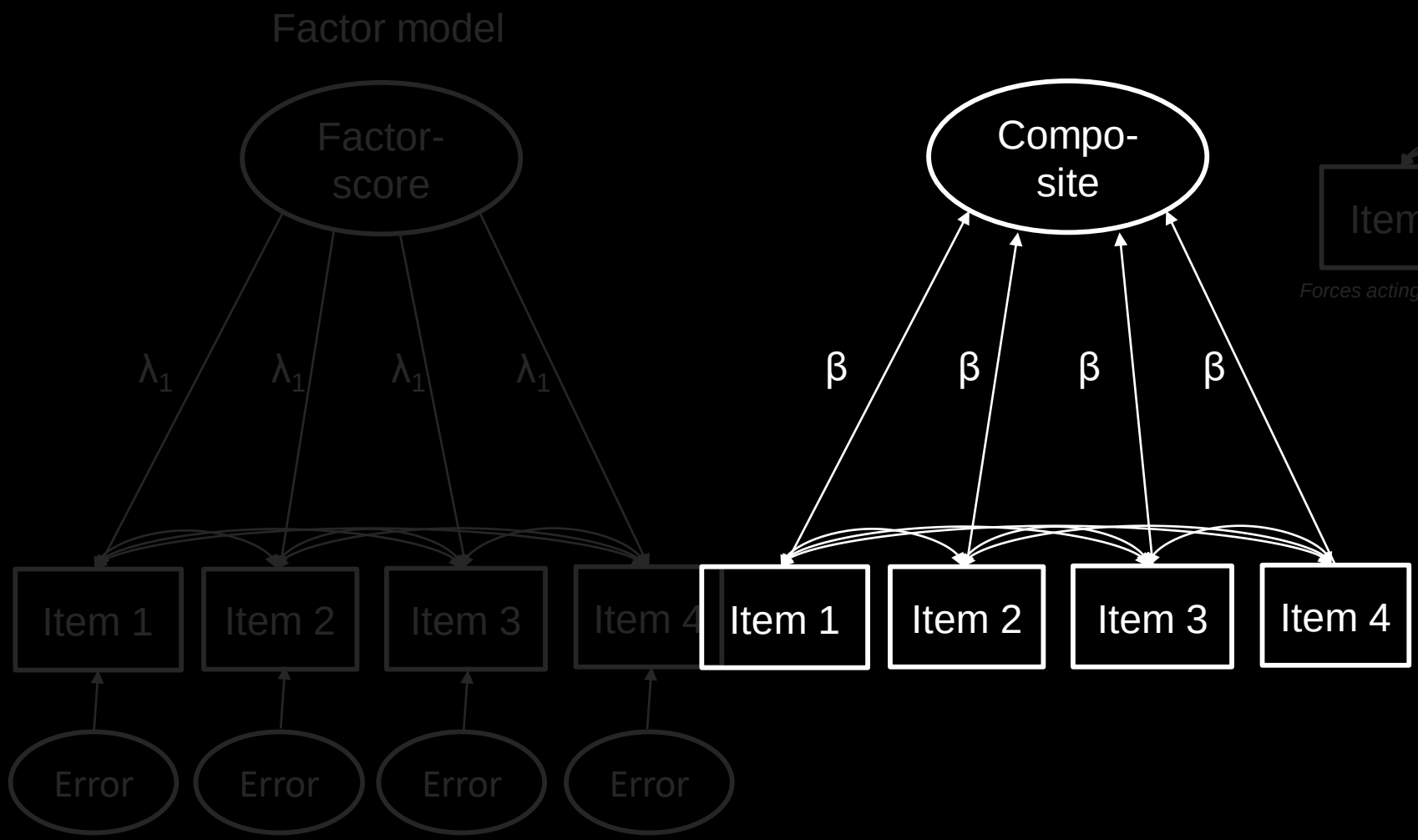
Correct if the items are *interchangeable indicators* of the same construct

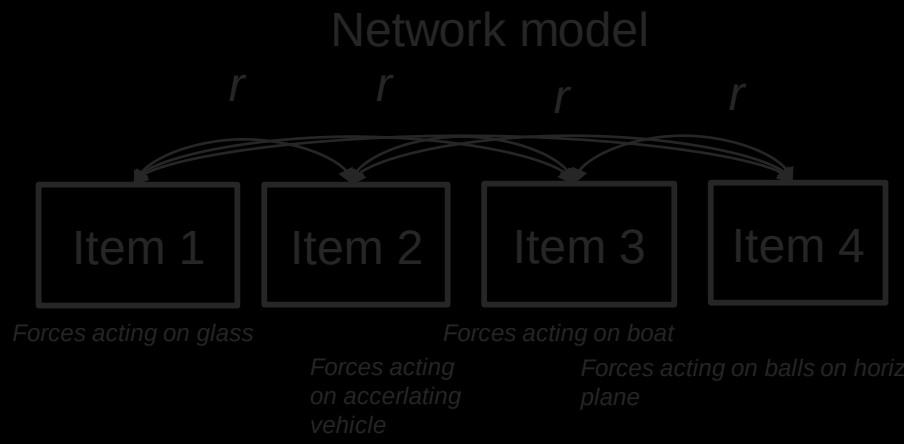




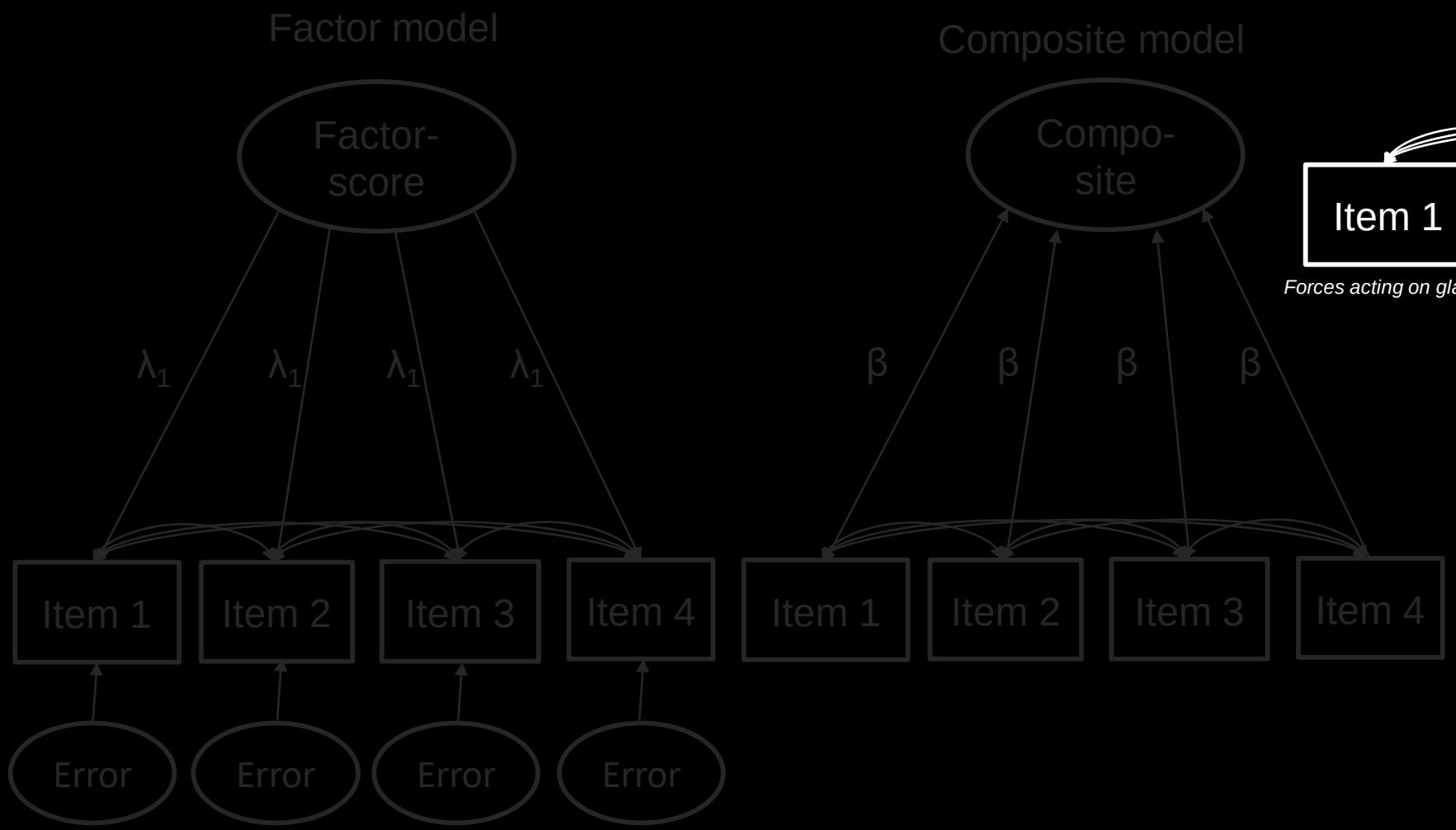


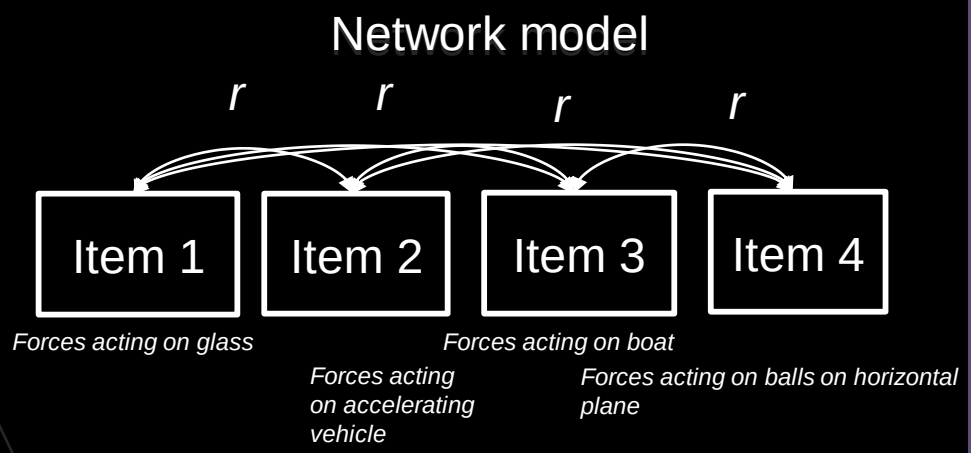
Correct if *each item adds* an *important part* of the construct



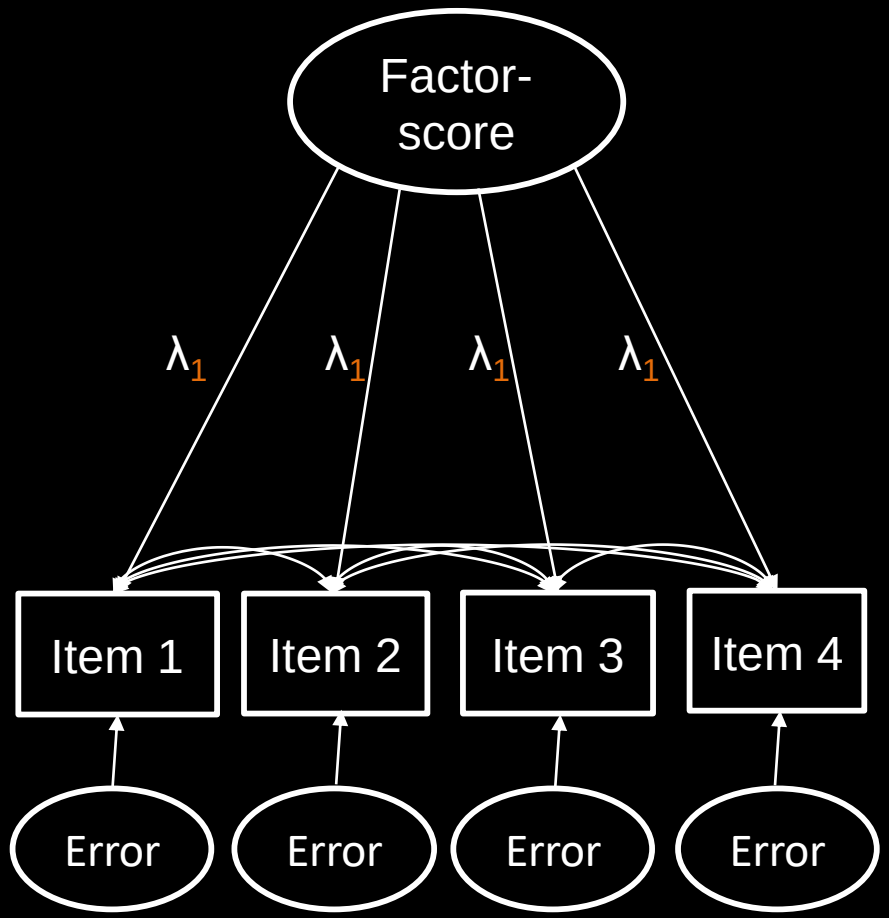


Correct if the *different parts* of the construct *interact*

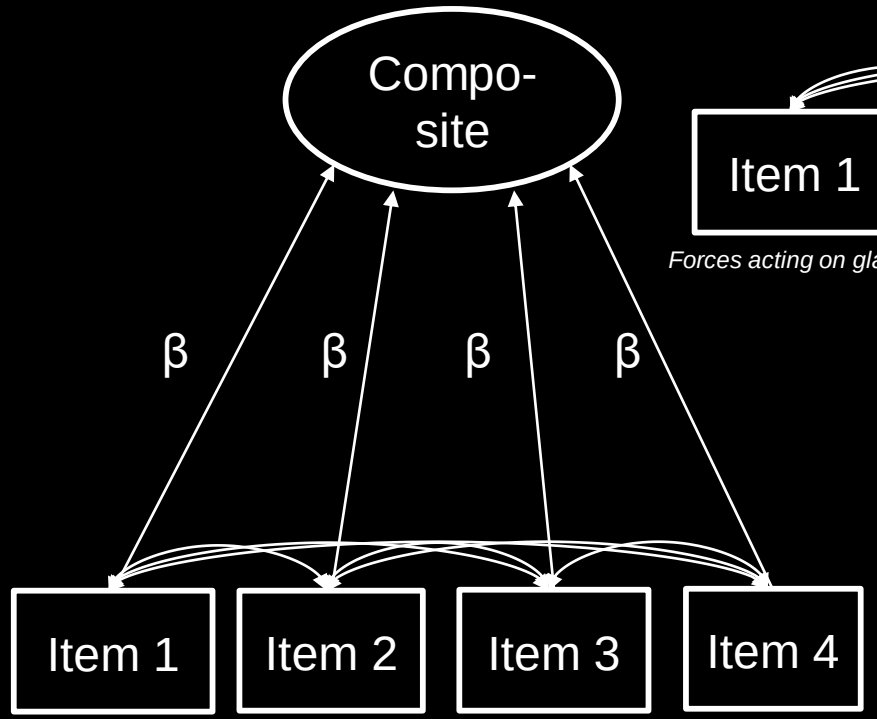


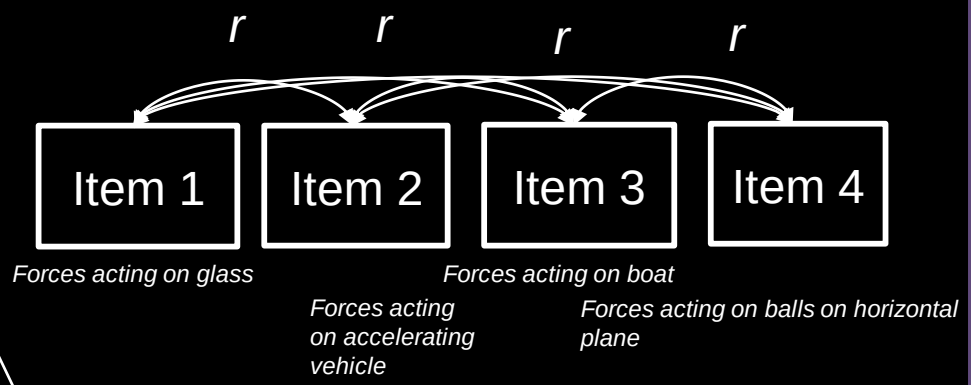


Valid representation of *specific concepts* (e.g., Newton's third law)



Valid representation of socially constructed concepts (e.g., Mechanics understanding, Math achievement, English comprehension, vocabulary)





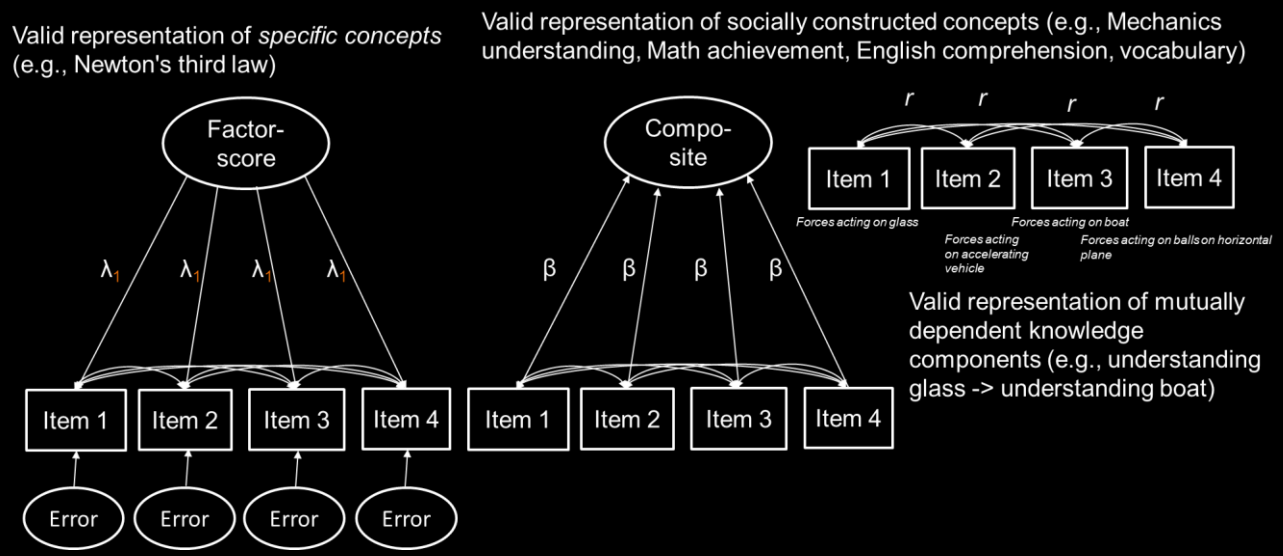
Valid representation of mutually dependent knowledge components (e.g., understanding glass -> understanding boat)

Implication 1:
The **theory** defines the model, **not** the **data**

Implication 2:
Should items intercorrelate?

High *internal consistency* is often asked for
e.g. Cronbach's Alpha (or Omega) > .70
As indicator of *reliability*
Taber 2018 (& Stadler et al., 2021)

For knowledge tests capturing composite/network constructs:
Inadequate



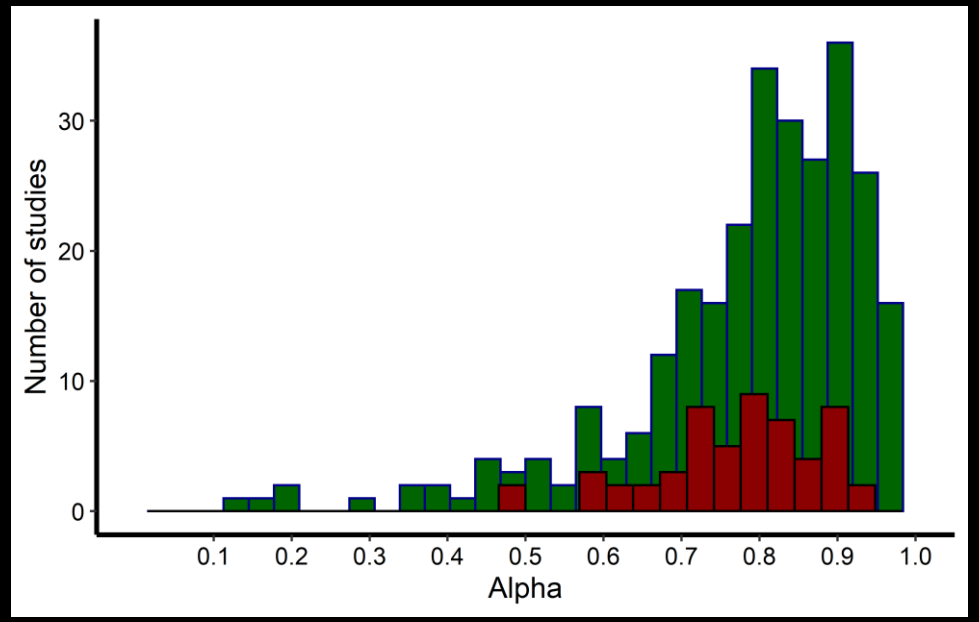
Meta-analysis:

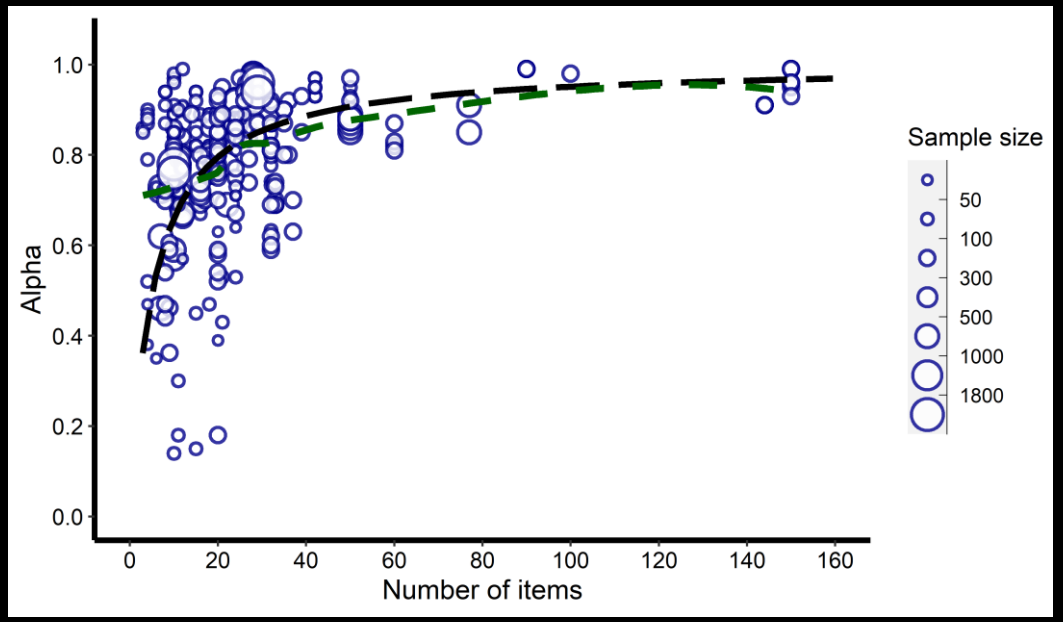
285 Alphas from 52 studies examining learning across multiple time points.

Results: Average Alpha = .85.
Very large heterogeneity (98%): Broad prediction intervals. Publication bias.

Higher in younger age, after instruction, during development, in Math and Languages vs. Sciences.

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

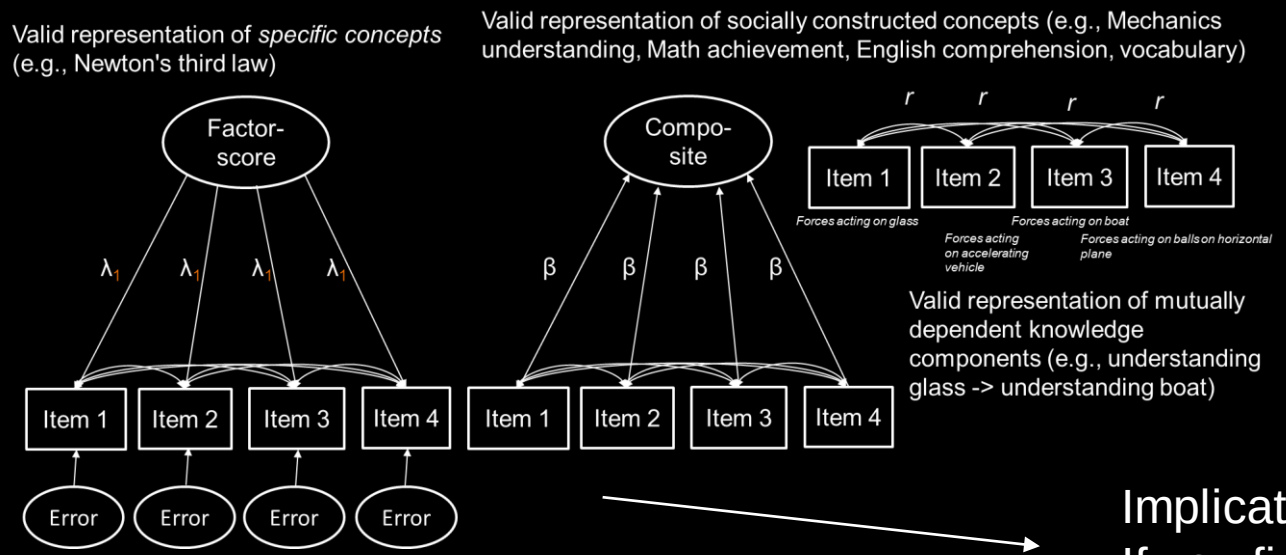




Implication 2:
Only expect (or demand) **high internal consistency** (i.e., item intercorrelations) **for** constructs well-represented by the **factor model**.

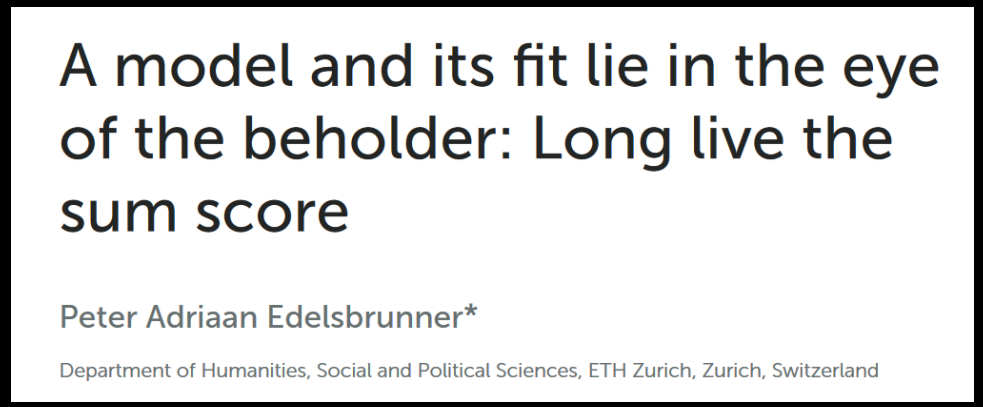
Implication 1:
The **theory** defines the model, **not** the **data**

Implication 3:
Deliver **other kinds** of **validity and reliability evidence** (e.g., retest/parallel test reliability, cognitive interviews, expert ratings).



Implication 4:
If you find a **strong model**,
It must make **strong assumptions**.

Thank you

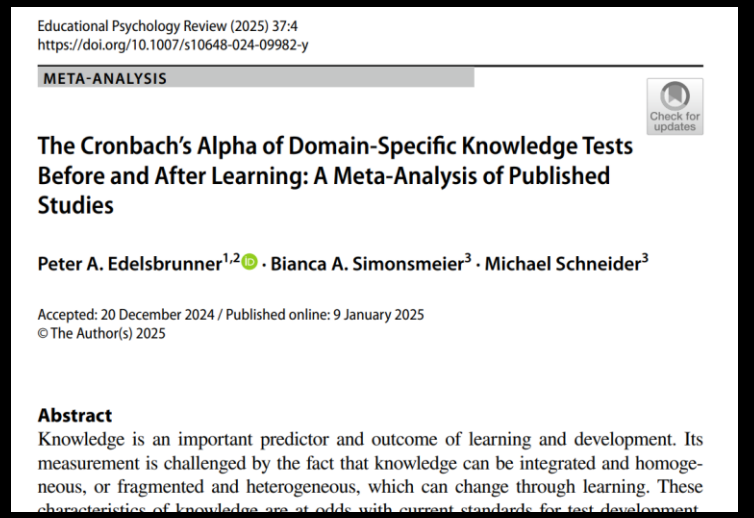


A model and its fit lie in the eye of the beholder: Long live the sum score

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 characteristics of knowledge are at odds with current standards for test development

https://osf.io/m8d7t/download

Get this presentation:

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