Confirmatory factor analysis (CFA) for cognitive endpoint development: performance characteristics under misspecification, and implications for modeling and analysis

Leif Simmatis[1], Kinga Bernatowicz[1], Silvina Catuara[1], Libby Floden[1], Elisabeth Piault-Louis[1]

[1]Evinova AG

Background

- Cognition is complex and is unlikely to be represented adequately by a single instrument.
- Combining cognitive assessments in endpoints has been considered, but work is yet required to understand optimal combination of different components in a composite.
- Clinical development may benefit from methods that leverage knowledge-driven combinations of measures, to enhance targeting of endpoints to specific populations.
- We explored measurement combination using data simulation and confirmatory factor analysis (CFA) modeling, evaluating impacts of assumptions on modeling (e.g., recapture of underlying raw data properties).
- *Hypothesis*: CFA will be sensitive to the structure of the input data, evidenced by statistics such as root mean-square error of approximation (RMSEA).

Methodological issue

We address the need to understand the impact that underlying data generation processes have on the use of multiple measures in a composite measure. This has implications for developing targeted clinical trial endpoints.

Methods

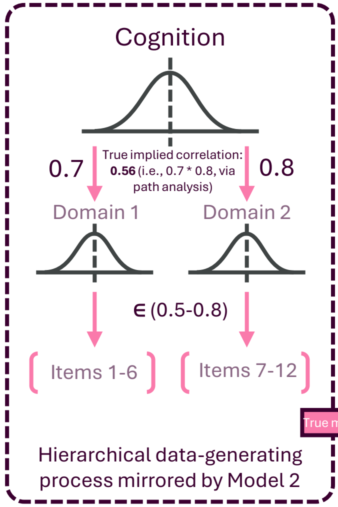
Data (n=500) were generated following a hierarchical structure. Simulated data were from each of 2 hypothetical domains that could represent different aspects of cognitive functioning, each loading onto 6 variables as might be observed in e.g., a PRO or clinRO instrument, and each being loaded onto a "general cognition" top-level latent factor.

5 Bayesian CFA models were developed in PyMC (ver. 5.26). Models were sampled using Markov Chain Monte-Carlo (MCMC) 1000x (+1000 tune samples) across 4 chains.

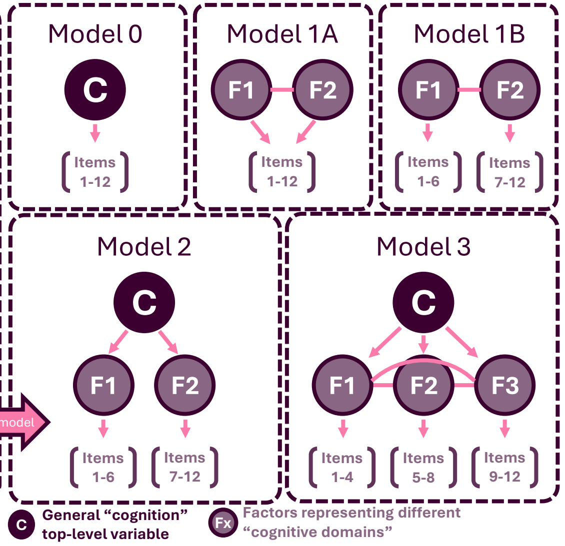
Models differed based on their structure and similarity to the "real" data – specifically in terms of their latent structure and number of levels. CFA models were evaluated as follows:

Name	Description	"Better" criterion
RMSEA	Root mean squared error of approximation	Lower
CFI	Cumulative fit index	Higher
TLI	Tucker-Lewis index	Higher
SRMR	Standardized root-mean square residual	Lower
WAIC	Watanabe–Akaike information criterion	Higher
PPP	Posterior predictive p-value	Closer to 0.5

Data-generation

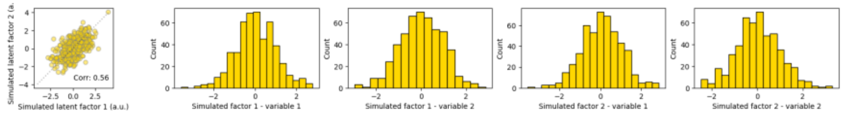


CFA modeling



Simulated data - examples

Pearson r of 0.56 expected based on path-tracing theory of common factors. Simulated variables are largely Gaussian.



Results

① **Model 2 generally performed the best**, especially in terms of conventional CFA metrics. Model 1B performed marginally better in Bayesian model-specific criteria.

Model	RMSEA	CFI	TLI	SRMR	WAIC	PPP
0	0.131 [0.130, 0.133]	0.769 [0.761, 0.775]	0.718 [0.708, 0.725]	0.102 [0.098, 0.109]	-7511.84	0.392
1A	0.059 [0.037, 0.093]	0.963 [0.912, 0.986]	0.940 [0.858, 0.978]	0.068 [0.098, 0.109]	-7102.51	0.546
1B	0.051 [0.027, 0.082]	0.963 [0.910, 0.990]	0.953 [0.888, 0.988]	0.073 [0.035, 0.130]	-7100.25	0.518
2	0.031 [0.023, 0.040]	0.988 [0.980, 0.993]	0.984 [0.974, 0.991]	0.037 [0.029, 0.052]	-7102.05	0.542
3	0.095 [0.092, 0.098]	0.892 [0.885, 0.898]	0.852 [0.841, 0.860]	0.091 [0.084, 0.101]	-7240.38	0.273

Table depicts metrics in the form mean [95% lower bound, 95% CI upper bound], except for WAIC and PPP, which are presented as point estimates (distributional summaries not possible for these metrics).

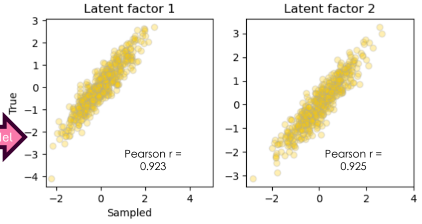
② **Model 2 (and 1B) were able to capture the structure of the original data best.** Estimated factor scores (i.e., per-individual scores) correlated strongly with the original data (r>0.9), demonstrating the recovery of original data patterns. Model 2 had the strongest overall correlations, although Model 1B approximated it closely.

	Latent factor 1			Latent factor 2		
Model	Est 1	Est 2	Est 3	Est 1	Est 2	Est 3
0	0.840			0.849		
1A	0.524	0.915		0.915	0.512	
1B	0.618	0.925		0.923	0.591	
2	0.620	0.925		0.923	0.594	
3	0.634	0.922	0.923	0.913	0.657	0.654

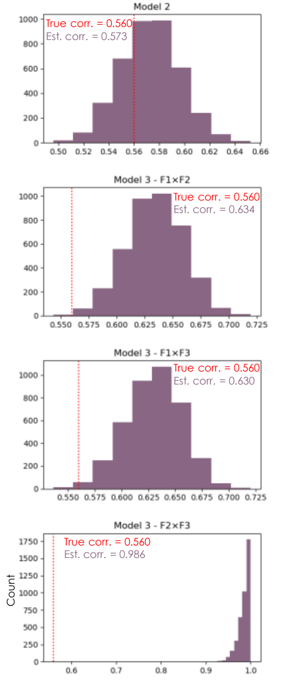
Table depicts correlations (Pearson r) between "latent factor" 1 and 2 (i.e., simulated data) and "estimated"/ sampled (Est.) factors from each model. Greyed out cells indicate non-applicability (e.g., Model 0 had only one estimated factor).

True model

Depiction of sampled (estimated) factor scores vs true (simulated) factor scores for Model 2 (latent factor 1 vs est. factor 2; latent factor 2 vs est. factor 1; cf. "Model 2" row of the adjacent table).



③ **Model 2 was able to capture the data-generating structure.** Model 2 inter-factor correlations were close to expected, none for Model 3 were close to the true value.



Conclusions

- CFA-based modeling is sensitive to underlying data structure. Model 2 – mirroring the original data structure – performed the best, followed by a structurally related model that did not have a general overlying cognitive factor (Model 1B). Inter-factor correlations were not recaptured when models were hierarchical in design, but substantially mis-specified (Model 3).
- Practically, this means that cognitive endpoints derived from multisource cognitive data require a good understanding of the theoretical relationship(s) between cognitive domains and assessment tools.
- Future work: Validation of knowledge-driven composite measure structures may enable enhanced targeting of endpoints in specific populations; additionally, iterative development of data-driven composites using exploratory factor analysis (EFA)combined with CFA to validate structure may be an important methodological combination towards creating composites that are data-driven.