

# Work in progress

last update 08/11/25

## Real dataset

### Correlation matrix (ITEM X FACTOR)

By definition, the factors identified through the CFA correspond to the dimensions measured by the [BESSI questionnaire](#); therefore, each factor is labeled according to its respective construct name.

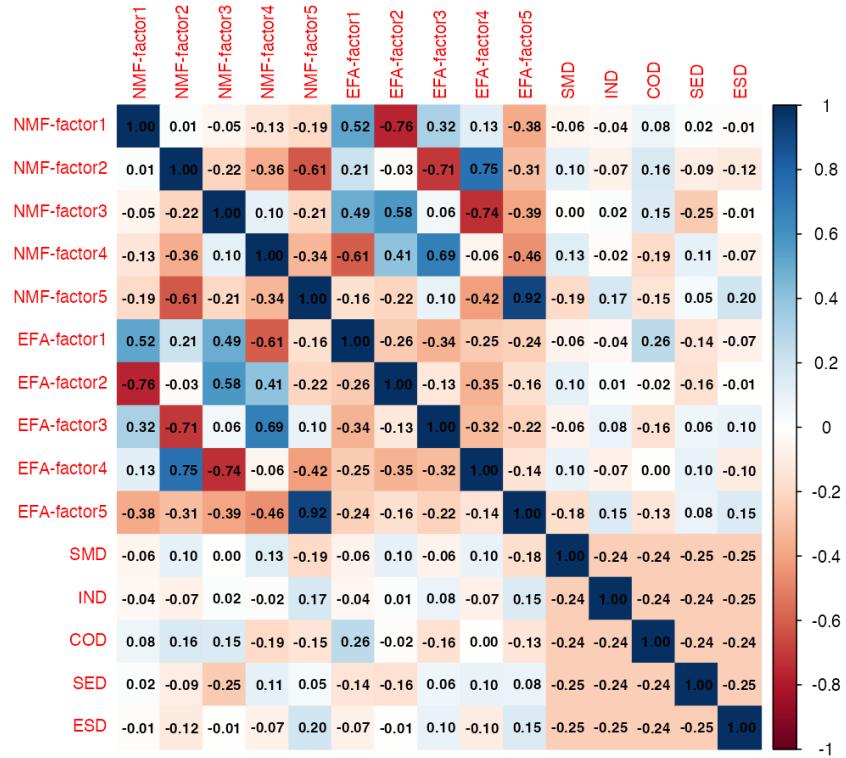


Figure 1: images/nmf-efa-cfa

We can observe that:

- NMF and EFA are well aligned: several factors extracted by NMF show strong correlations with two of the EFA factors.
- CFA does not correlate well with NMF and with EFA.

## Correlation matrix (UNIT X FACTOR)

Here we examine the correlations between the factor scores obtained from the three methods, to see whether a subject's profile derived from one technique resembles that produced by the others.

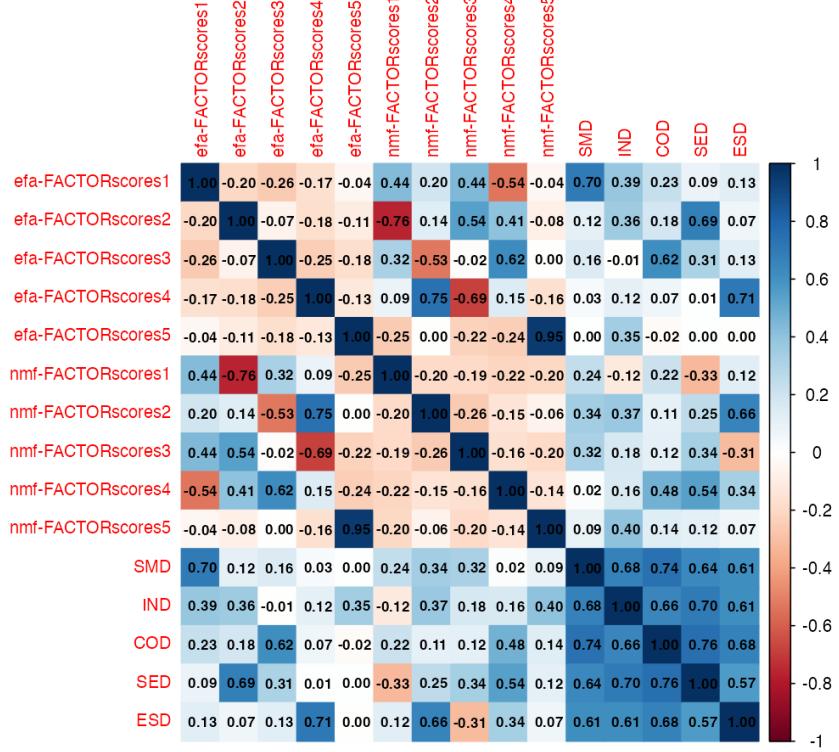


Figure 2: images/SCORESnmf-efa-cfa

We can observe that:

- CFA factors show strong correlations with each other, a pattern that is quite common in psychological data.
- EFA and CFA are reasonably aligned (except for the IND factor): we can clearly see pairs of latent factors that correlate strongly.
- EFA and NMF show a similar behavior to what we observed in the previous matrix. Two scenarios emerge: either one NMF factor correlates strongly with one EFA factor (e.g., NMF-Factor 1), or it correlates with two EFA factors but in opposite directions (e.g., NMF-Factor 2). Interestingly, this pattern mirrors what we saw in the correlation matrix of factor scores versus factor loadings, the same factors are involved in both cases.

```

# to test further rotations of the factors (below the ones tried together)
source("R/analysis.R")
plot(P[,3],efa_loadings[,5])

efa_loadings_diff <- efa_loadings[,4]-efa_loadings[,5]-efa_loadings[,3]
plot(P[,5],efa_loadings_diff)
cor(P[,5],efa_loadings_diff)

```

## Heatmap

In the paper by [Kim et al., 2025](#), both NMF and EFA are applied for different purposes, without, however, providing an in-depth discussion of the methodological rationale (how the two techniques differ and where they overlap). Nonetheless, the two approaches are used in a complementary way:

- EFA used disease **prevalence** data for **identifying patterns**;
- NMF used disease **incidence** data for **uncovering longitudinal patterns**.

In the paper, the heatmap is employed to visualize patterns of comorbidity (different diseases that fall within the same cluster).

Below we reproduce a similar heatmap for our dataset.

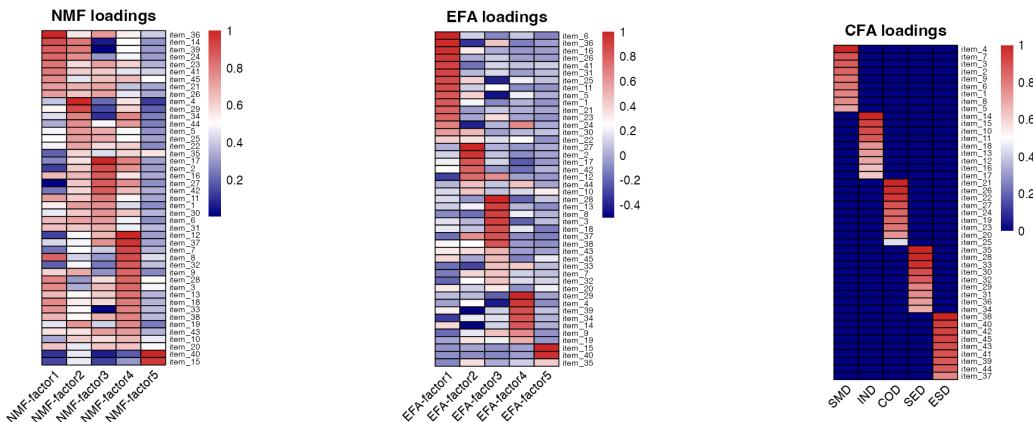


Figure 3: images/HEATMAPnmf-efa-cfa

It is evident that:

- in the CFA, the loadings are fixed to zero by design (as specified in the model definition)

- the data come from a questionnaire with well-defined dimensions, which the EFA successfully identifies
- NMF also distinguishes the dimensions (similar to EFA, it identifies roughly 6–7 key items per dimension), though in a less sharply defined manner

## Simulated data

The data-generating process follows the paradigmatic case for a confirmatory analysis, where a set of items is specifically designed to measure certain latent dimensions.

We simulated 1000 statistical units responding to a questionnaire consisting of 10 items: five assessing depression and five assessing anxiety (the two underlying latent constructs).

The data were simulated from binomial distributions to ensure positive values and to remain consistent with how such psychological phenomena are typically conceptualized.

### Correlation matrix (ITEM X FACTOR)

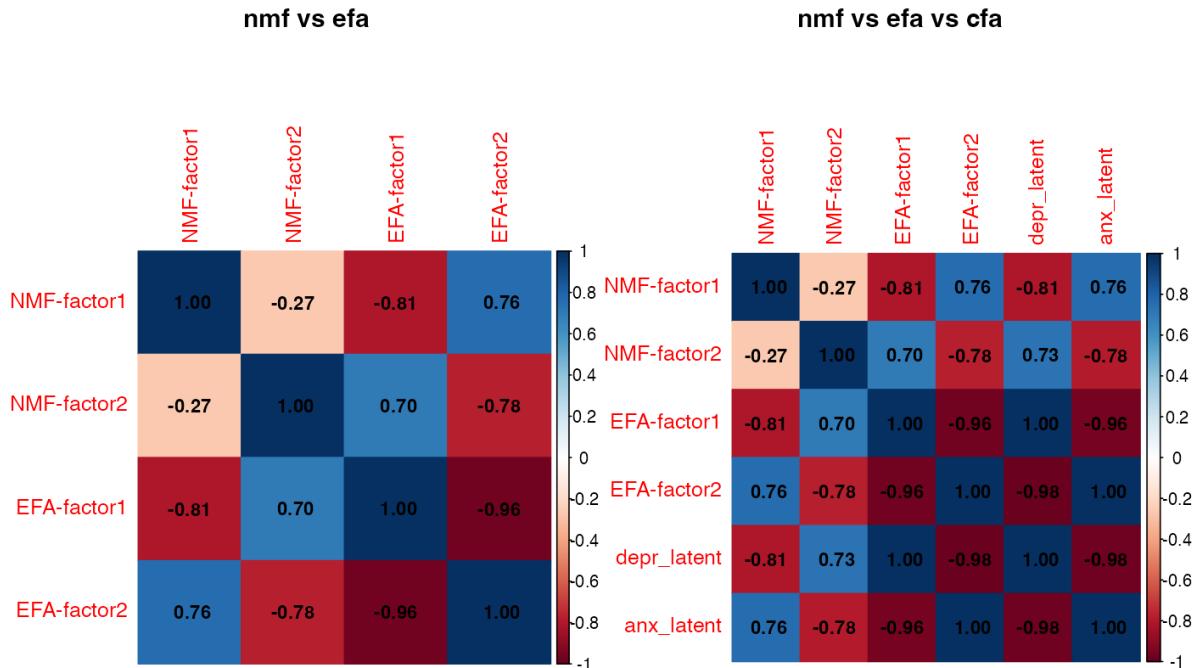


Figure 4: images/SIMnmf-efa-cfa

We know examine how the factors correlate *within* each technique:

- in the CFA, each factor correlates perfectly (1) with itself and shows a high correlation in the opposite direction with the other factor.
- The same pattern is observed in the EFA.
- In the NMF, each factor correlates perfectly (1) with itself and shows a correlation of -0.27 with the other factor, which interestingly resembles the 0.2 value set in the simulation.

### Correlation matrix (UNIT X FACTOR)

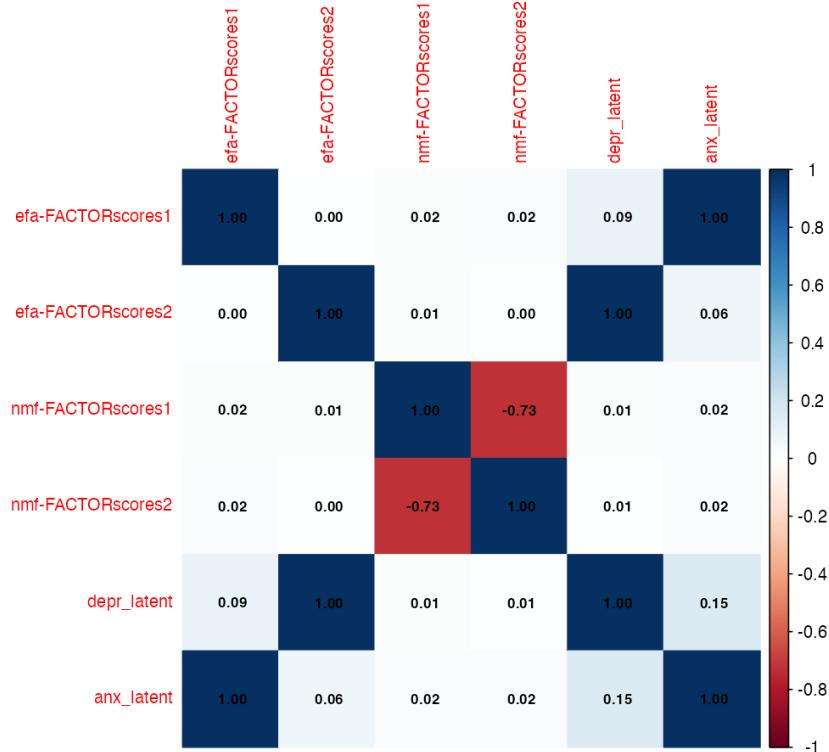


Figure 5: images/SIM-SCORESnmf-efa-cfa

We know examine how the factors correlate *within* each technique:

- in the CFA, each factor correlates perfectly (1) with itself and shows a correlation of 0.15 with the other factor, very close to the 0.20 value set in the simulation.
- in the EFA, this correlation is 0, which is still reasonably close to the simulated 0.20 value.
- in the NMF, however, the correlation between the two factors is entirely different from what we set in the simulation (-0.73 instead of 0.20).

It almost appears that in NMF the pattern is reversed compared to CFA and EFA when looking at factor correlations in the two conditions, additionally showing the correlation between factors specified in the simulation when correlating the factor loadings rather than the factor scores.