Researching response-scale format effects in questionnaires:

Using the Height Inventory

IMPS 2024 | July 16–19

Hynek Cígler, Stanislav Ježek, Karel Rečka

Psychology Research Institute, Masaryk University, Brno, Czechia

Measurement

What *we have*: **observations**.

- We decide about their nature and how we code them to obtain data.

What *we assume*: an **existence of quantitative attribute**.

- Theory decides about its quality and properties.

Measurement is then **linking** of the observations to the attribute through data.

Two cardinal principles:

- 1. Coordination (linking) function.
- 2. Calibration of the measure.

Measurement approaches in psychology

Problem: Psychological phenomena cannot be observed directly.
- Are they quantitative? What is their nature? How to establish the linking function?

Early attempts: Linking **sensory events** to external physical stimuli using a “law”.
- Deriving the scale from well established physical scale (e.g., pain as a function of a pressure).
- Psychophysics failed to provide an exact link – “Ferguson committee” (Ferguson et al., 1940).

Realism: **Latent trait models** (FA, IRT).
- An independently existing trait causes behavior (responses) to a questionnaire.
- Ontological claim: An existence of an attribute can only be assumed (e.g., Borsboom et al., 2003).
- Usually also distributional assumption, etc.

Operationalism: **Classical test theory**
- Attribute is operationalized using the measurement tool.
- Multiplication of entities: each questionnaire measures a different attribute (e.g., Fried, 2017).

Likert scale item format (LS)

The most common approach to measure personality and attitudes

LS as the scaling procedure vs. **LS as an item format**.

Many forms of Likert-like items.

- Number of response options (usually 4–7).
- Presence of the middle point.
- Extremity and actual labeling of verbal anchors.
- Presence of the reversed items.

These response format moderates the performance of the questionnaire.

- **Reliability** (internal consistency, test-retest).
- **Convergent validity** (may be biased by method factors if all the attributes are measured by LS).
- **Criterion validity** – usually only indirect criteria, as attributes cannot be measured independently.

We don’t have an objective criterion to study measurement performance in psychology.

Really?

Don’t we have any objective criterion to study measurement performance in psychology?


208.

Using human height as the attribute

Well, we have an objective criterion: Introducing the **Height Inventory** (HI)!

Example of the items:

- + A lot of trousers are too short for me.
- + I have an appropriate height for playing basketball or volleyball.
- + I must often be careful to avoid bumping my head against a doorjamb or a low ceiling.
- + At concerts, my stature usually obstructs other people’s views.
- - I have enough room for my legs when traveling by bus.
- - I could play a dwarf.
- - When I buy clothes, children’s sizes often fit me well.
- - When talking to other adults, I have to look upwards if I want to meet their eyes.
- ...

Using human height as the attribute

Measuring the physical rather than a psychological phenomenon using the same procedure.

- The idea is not entirely new and has repeatedly appeared (e.g., van der Linden, 2017).
- However, we propose to use the human height for researching measurement in psychology.

Advantages:

- Human height (length) exists, is quantitative, and people differ in it.
- We can measure it independently with high precision.
- It also solves the "duality problem" (distribution assumption of a trait vs. response link function shape).
- People even know their height with sufficient precision (*ICC* > .999; Rečka, 2018).

Disadvantages:

- HI items are "save" compared to common questionnaires – small non-response rate.
- Does it measure the "true height" or "psychological height"?

Isn’t it just a... “psychological height”?

No. We use questionnaires to measure extraversion, self-efficacy, aggression...
... not a “self perception” of extraversion, self-efficacy, or aggression.

Until psychological phenomena measured using questionnaires are not explicitly defined as “subjective self-efficacy”, we may measure height in the exactly same way.

- Actually, measuring height using a questionnaire needs weaker assumptions compared to measure of psychological phenomena.

Height Inventory

Original version (Rečka, 2018): 26 items (13 reverse scored).

- Open data with gender and self-reported height (`ShinyItemAnalysis::HeightInventory`).
- High internal consistency ($\omega_{tot} \approx .96$) and criterion validity ($r \approx .87$).

Shortened version (Hubatka et al., 2024): 11 items (6 reverse scored).

- Used in most of our studies.
- Still high internal consistency ($\omega_{tot} \approx .90$), test-retest reliability ($r \approx .94$), and criterion validity ($r \approx .85$).

More datasets are available so far.

- Different data collection designs, response formats, etc.

Additional attempts: Weight Inventory, Age Inventory.

- Performed slightly worse then height.

Height Inventory (26 items)

Almost **linear relationship** of sum scores with actual height.

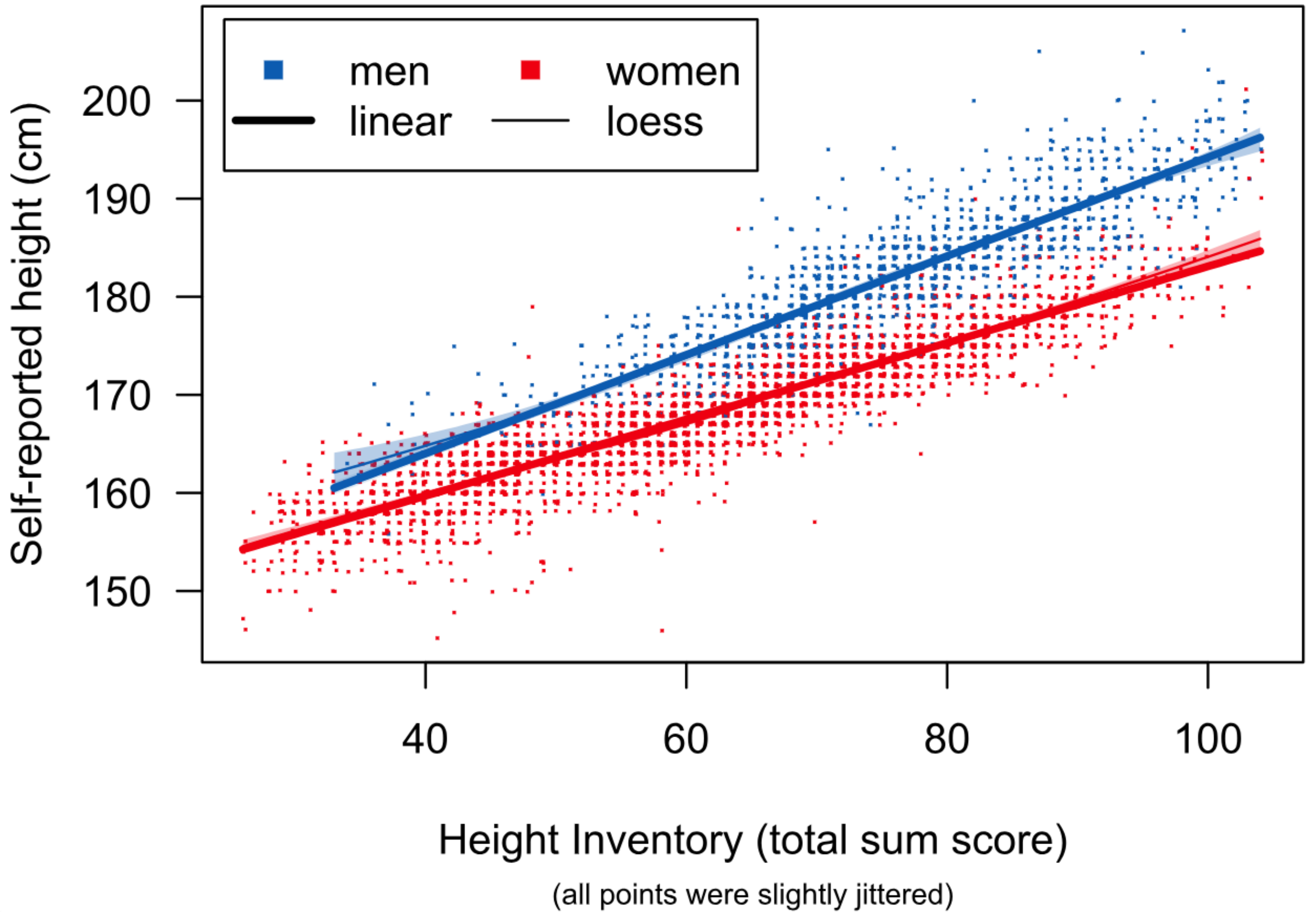
Different criterion validity across gender.

- **Predictive bias**.

Slightly non-invariant across gender.

- Especially the short version.
- Establishing invariance do not resolve predictive bias.

All the further analyses performed separately for men and women.



Identification of the latent trait

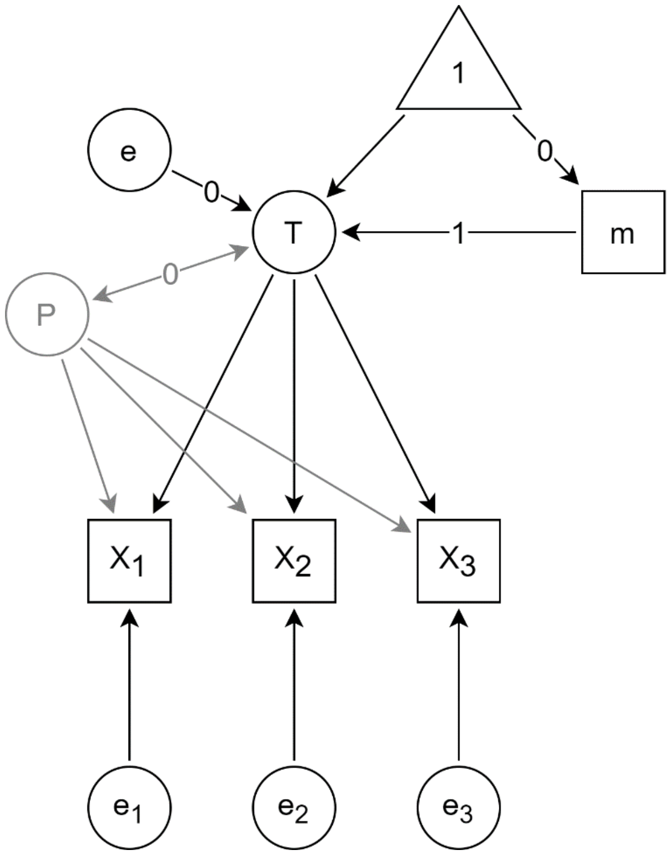
Left: height as the perfect cause of factor.

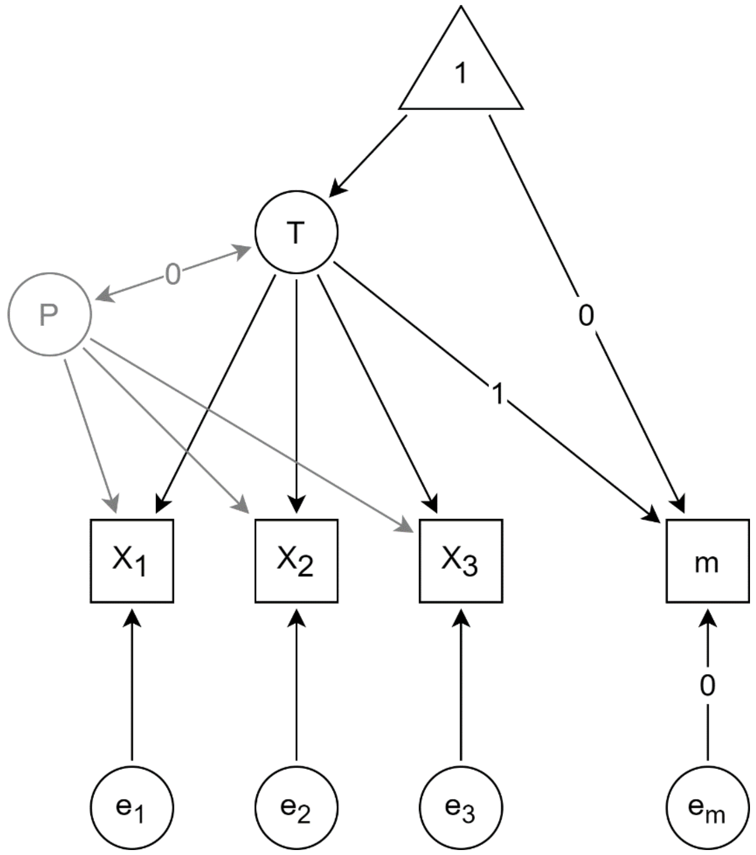
Right: height as the perfect indicator of factor.

Trait scale is in meters.

- T – latent variable identified as height.
- P – orthogonalized personal trait causing responses, but not related to the actual height.

Additional constraints on items are available.





Existing applications (examples)

Number of response options

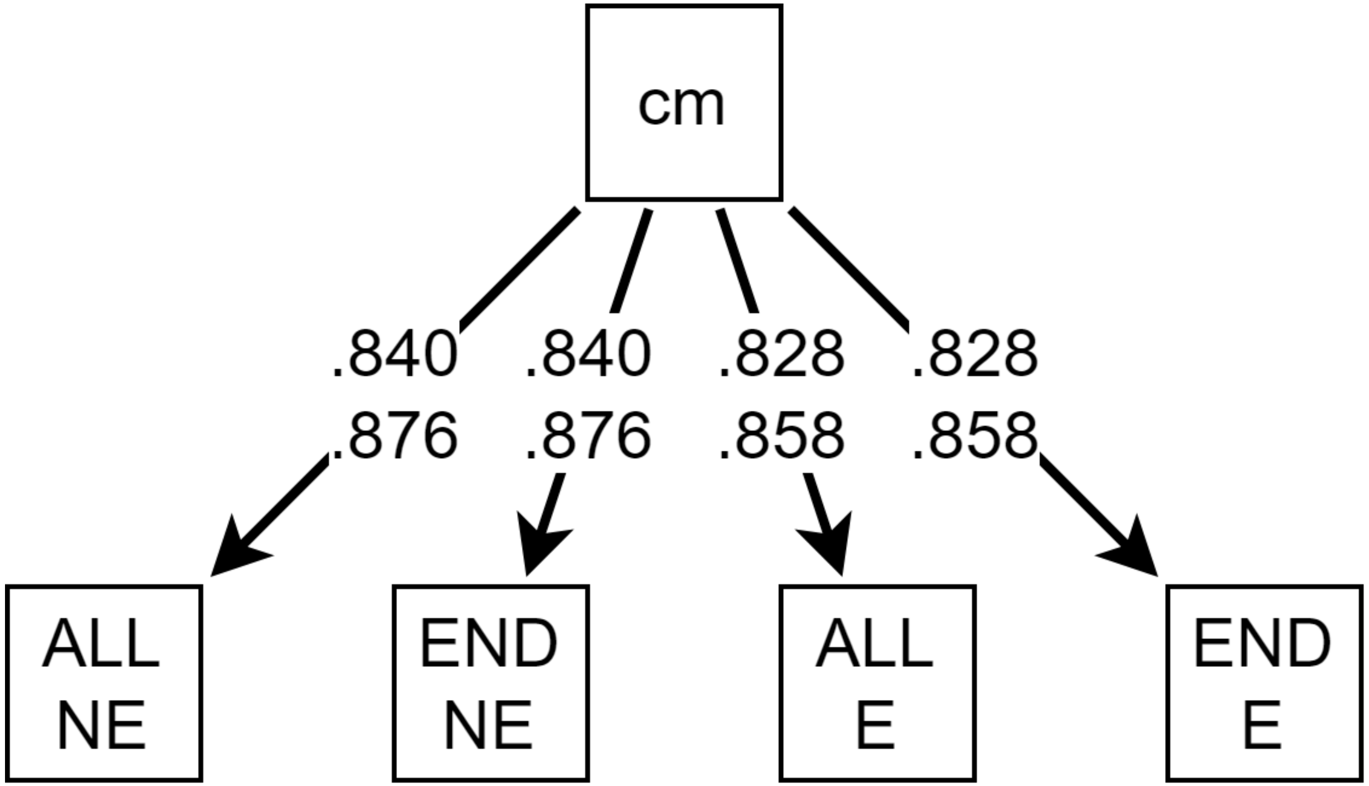
Extremity and presence of verbal labels

Comparison of Visual Analogue Scale with Likert scale

Intuitive vs. careful responding

Reverse-scored items

...



Verbal labels (five options)

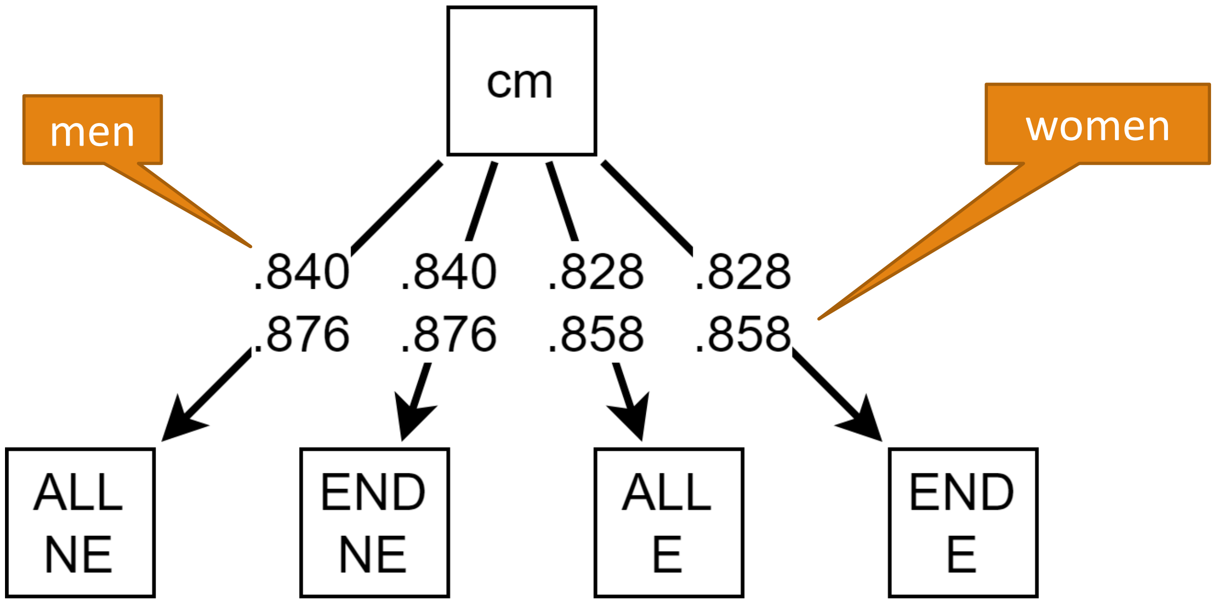
Presence manipulation:
- All options labeled (ALL)
- Only outer options labeled (END).

Extremity manipulation:
- Extreme condition (E): strongly agree/agree...
- Non-extreme condition (NE): agree/slightly agree...

Strict measurement invariance.

Negligible differences in reliability in favor for all, non-extreme labeled options.

Negligible differences in criterion validity.



Reliability: .904–.908 (men) / .935–.947 (women)

Model fit: $\chi^2(10) = 5.72$, $p = .838$.

Number of response options

Manipulation: 2, 6, or 10 options.

Strict measurement invariance

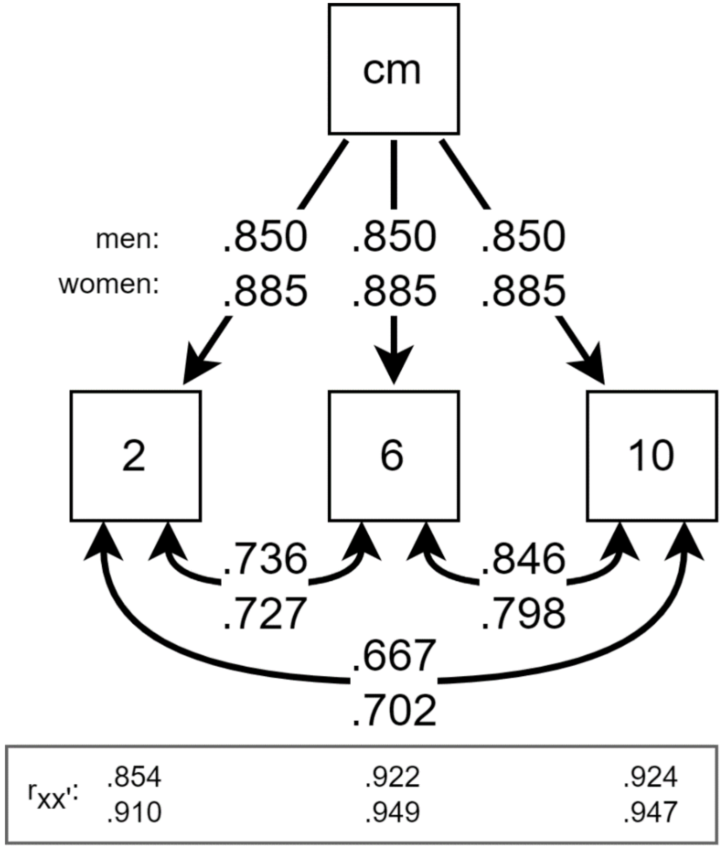
Strong differences in reliability (2 < 6 ≈ 10).

No differences in criterion validity, $\chi^2(12) = 11.21$, $p = .511$.

- However, higher residual correlation between 6 and 10 options than between 2 and 6 or 2 and 10.

Conclusion: **The increase of internal consistency is caused by systematic, but construct irrelevant variance**.

- Method factor (response style).
- Acquiescence bias? Extreme response style?
- Method factor thus might bias scale correlations.



Presentated at the IMPS 2024

Posters at IMPS (presented on Wednesday, 16:45):

- **Šragová**, Cígler, & Kalistová: Do visual analogue scales perform better than Likert-type scales?
- **Strojil** & Cígler: Video-administered questionnaire: Psychometric properties and comparison with a text-based format.

All the results are available at OSF repository:

- Cígler, H., Ježek, S., Rečka, K., Elek, D., Hubatka, P., Tancoš, M., & Šragová, E. (2024). *SCALING project.* https://doi.org/10.17605/OSF.IO/GQTA5

Using HI to identify latent trait

Example using reversed scored items

Using data available in `ShinyItemAnalasis::HeightInventory`

- *N* = 4885; about 68 % of women.

Ordinal confirmatory factor analysis.

- Estimated in `lavaan`.
- Scaled test and statistics (WLSMV).
- Theta parameterization.
- Pairwise missing data.
- Slight model improvement
  (items with residual covariances removed).
- Exploratory models estimated using ESEM syntax.
- Skewed geomin rotation if desirable.

Multiple-group analysis for men and women separately.

Factor analysis of HI

	x2	df	TLI	RMSEA [90% CI]	SRMR
1-factor	15836.5	378	.934	.129 [.128–.131]	.076
2-factors CFA	5200.9	372	.979	.073 [.071–.075]	.038
S-1 bifactor (specific factor for reversed items)	4317.8	354	.982	.068 [.066–.070]	.031
symmetric bifactor	3033.2	336	.987	.057 [.055–.059]	.024
2-factors EFA	**1641.0**	**338**	**.994**	**.040 [.038–.042]**	**.016**

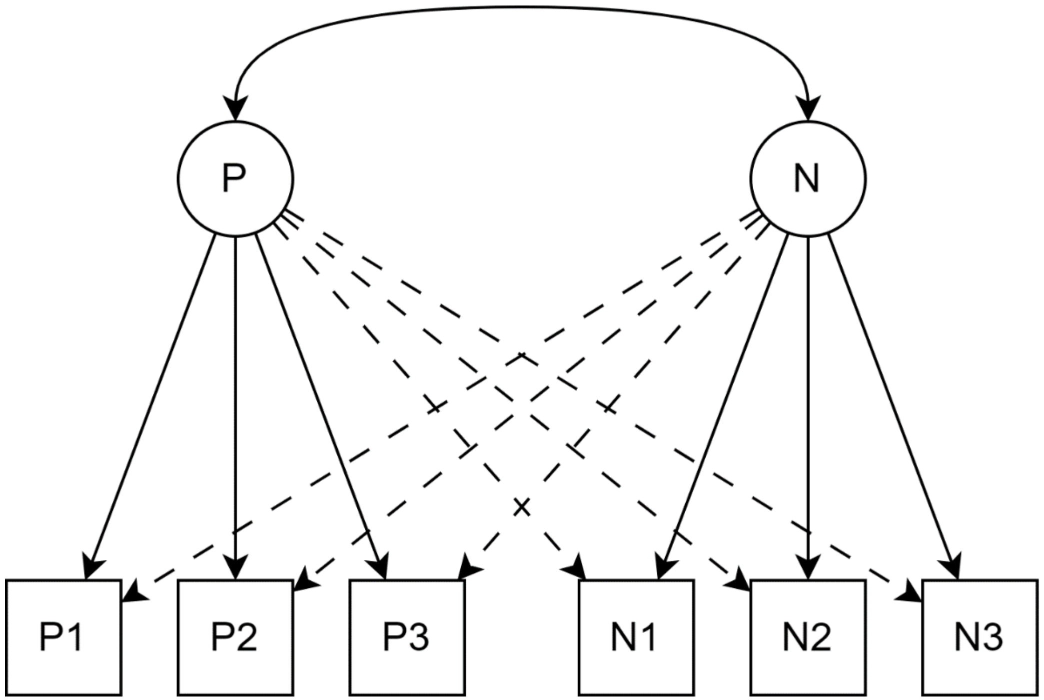
Latent trait identification

M0: EFA without manifest height.

Baseline model.

Strong latent correlation (geomin rotation):

- men: r = .630
- women: r = .739
- But not strong enough to be considered as the same construct.



Latent trait identification

M1: True height loads indicators directly.
Factors are unrelated to the "height construct".

Factors are (almost) uncorrelated (geomin):

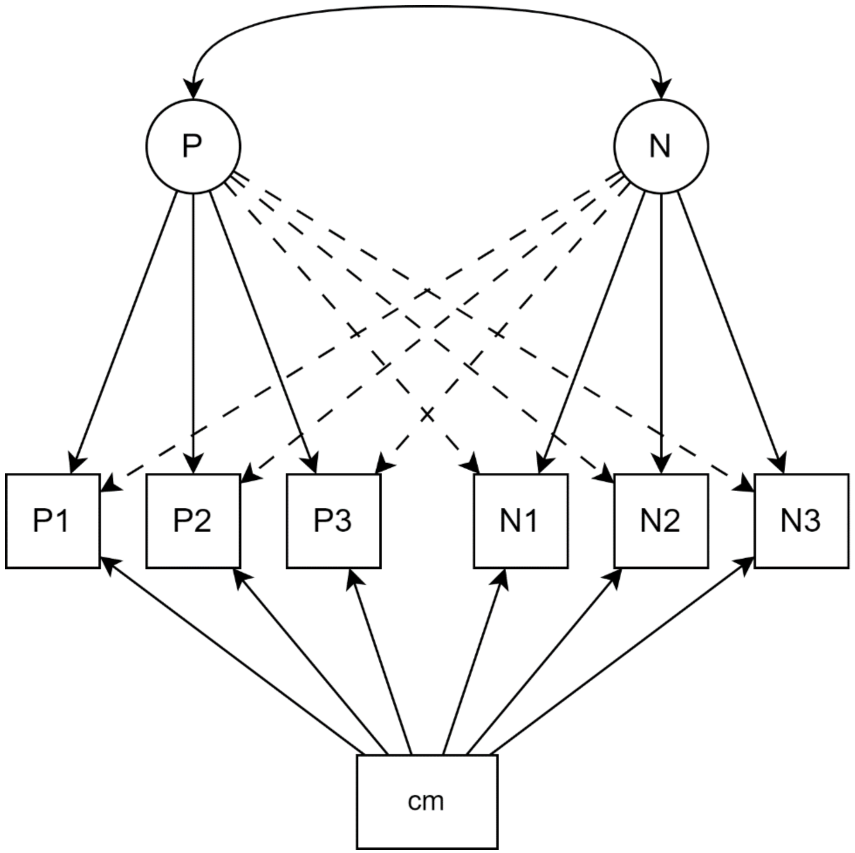
- men: $r = .151$, $p = .151$
- women: $r = -.203$, $p < .001$
- Factors mostly independent conditionally on height ($r_{PN|cm} \ll r_{PN}$).

Low factor scores (and reliability):

- men: M = .10 (SD = .22, min = -.34, max = .51)
- women: M = .12 (SD = .20, min = -.27, max = .38)
- There is no longer any positive and negative factor!

Standardized regression coefficients:

- men: M = .71 (SD = .12, min = .34, max = .84)
- women: M = .77 (SD = .08, min = .56, max = .87)



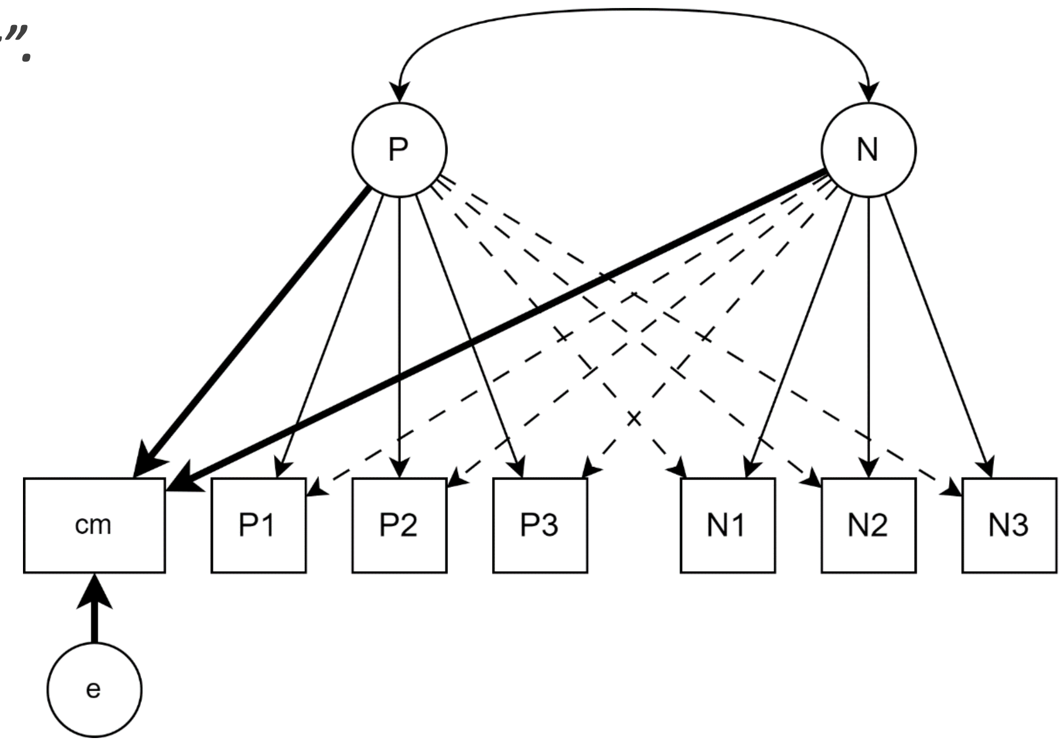
Latent trait identification

M2: The true height is loaded by both factors.
Both factors are together the "height construct".

Geomin rotation of factors:
- men: r = .676
- women: r = .743

Standardized factor loadings of the true height:
- men: $\lambda_P = .663, \lambda_N = .317$
- women: $\lambda_P = .485, \lambda_N = .502$

Explained variance of height:
- men: R^2 = .807
- women: R^2 = .848



Latent trait identification

M3: The true height is loaded by one factor.
Only the first factor represents the "height construct".

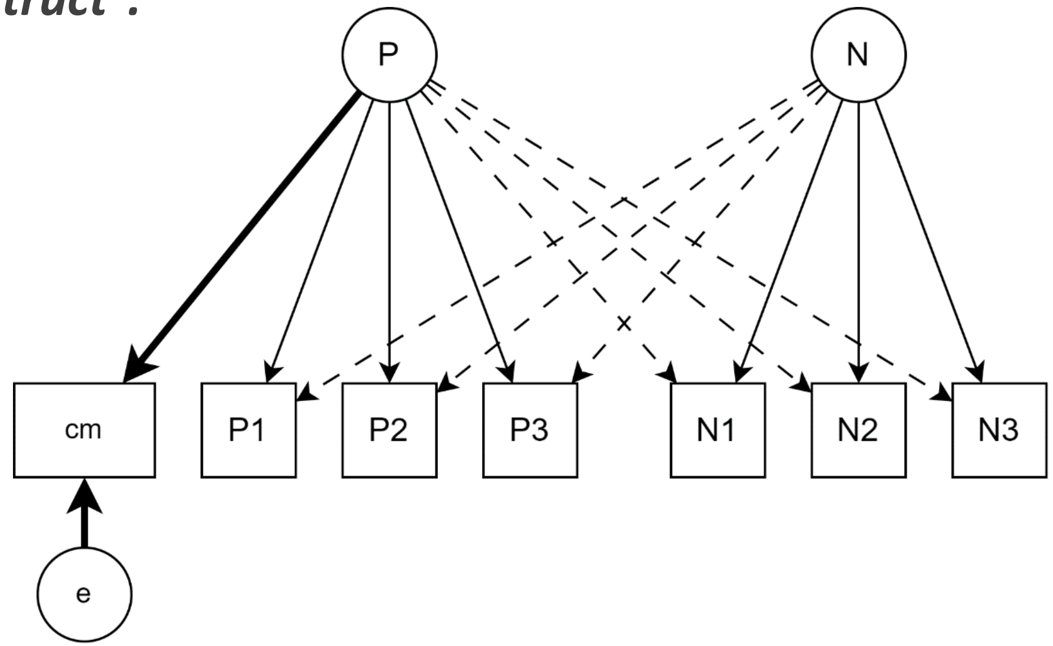
The factors are identified, so we don't use a rotation (r = 0).

Standardized factor loadings of the true height:

- men: $\lambda_P = .899, \lambda_N = 0$
- women: $\lambda_P = .921, \lambda_N = 0$

Explained variance of height:

- men: $R^2 = .807$
- women: $R^2 = .848$



Latent trait identification

M4: The true height is perfectly loaded by both factors.
Both factors are just identified as the "height construct".

Geomin rotation of factors:

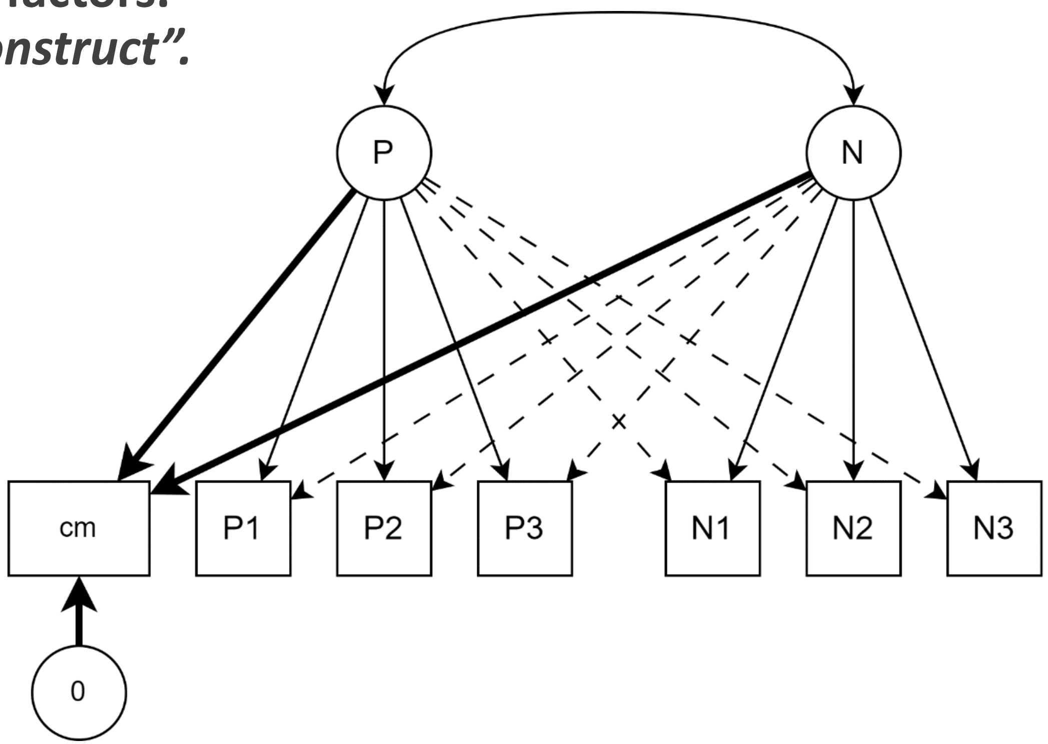
- men: r = .638
- women: r = .743

Standardized factor loadings of the true height:

- men: $\lambda_P = .339, \lambda_N = .749$
- women: $\lambda_P = .547, \lambda_N = .524$

Explained variance of height:

- men: $R^2 = 1$
- women: $R^2 = 1$



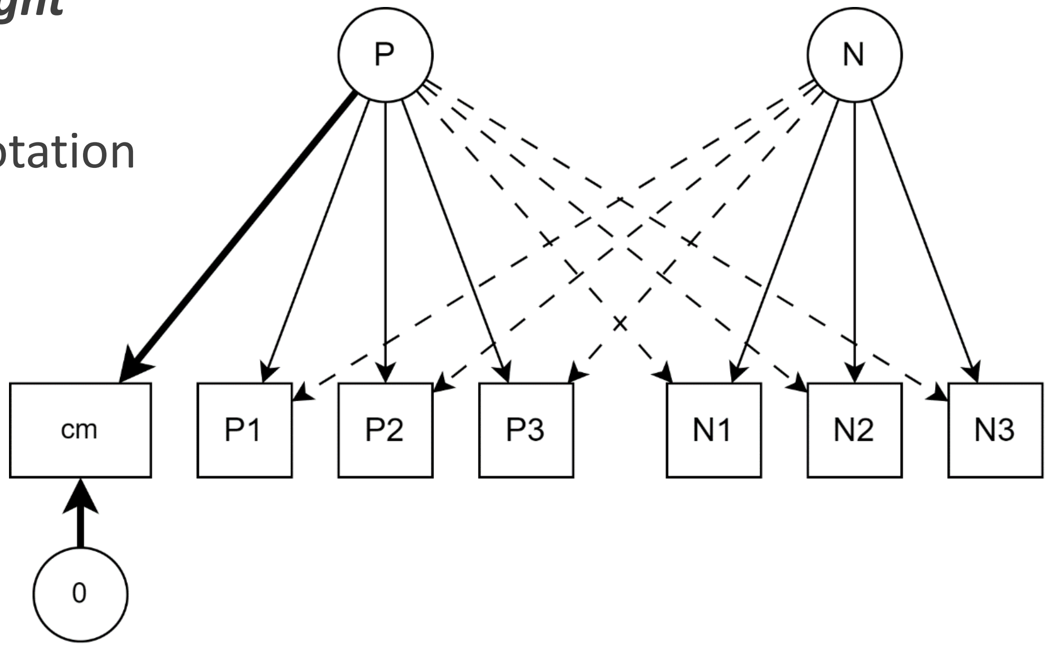
Latent trait identification

M5: The true height is perfectly loaded by one factor.
The first factor only is just identified as the "height construct".

The factors are identified, so we don't use any rotation (r = 0).

Standardized factor loadings of the true height:
- men: $\lambda_P = 1, \lambda_N = 0$
- women: $\lambda_P = 1, \lambda_N = 0$

Explained variance of height:
- men: R^2 = .807
- women: R^2 = .848



Latent trait identification

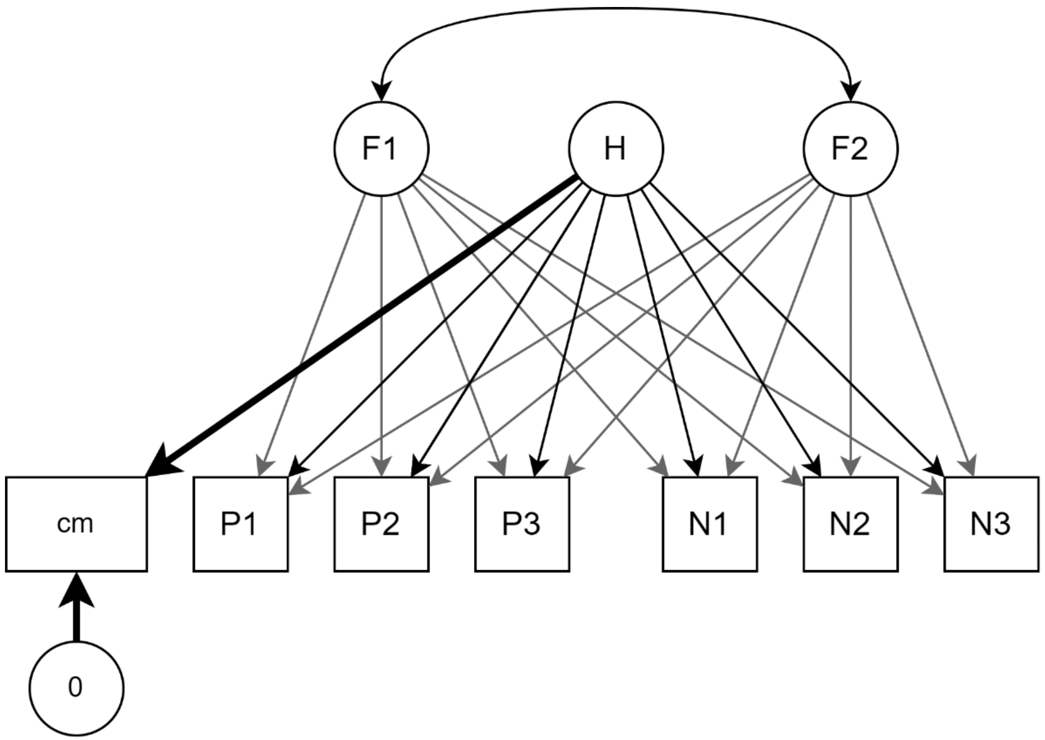
M6: Additional three-factors model.
The true height is perfectly loaded by one factor.

First factors are identified.
Method factors correlation:

- men: r = .151, p = .234
- women: r = -.203, p < .001

Standardized factor loadings of the true height:

- men: $\lambda_P = 1, \lambda_N = 0$
- women: $\lambda_P = 1, \lambda_N = 0$



Latent trait identification

	x2	df	TLI	RMSEA [90% CI]	SRMR
M1	1371.5	338	.994	.035 [.033–.037]	.013
M2, M3	1884.1	376	.993	.041 [.039–.042]	.016
M4, M5	2077.1	378	.992	.043 [.041–.045]	.016
M6	1371.5	338	.994	.035 [.033–.037]	.013

M1 – indicators loaded by the height

M2 – height loaded by both factor; M3 – height loaded by one factor

M4 – height perfectly loaded by both factor; M5 – height perfectly loaded by one factor

M6 – height perfectly loaded by one factor; two additional factors (3 in total).

Latent trait identification

If the latent trait is identified as the true height (model M5), then reliability should equal to R^2 of sum score and true height. This is not true:

- Men: ω_h = .929; but observed R^2 = .767.
- Women: ω_h = .959; but observed R^2 = .816.

However, it is true in the M3, estimating validity as $\hat{r}^2 = \omega_h \lambda_{cm}^2$:

- $\hat{r}^2 = \omega_h \lambda_{cm}^2 = .931 \cdot .899^2 = .752$; observed R^2 = .767.
- $\hat{r}^2 = \omega_h \lambda_{cm}^2 = .959 \cdot .921^2 = .813$; observed R^2 = .816.

Models M4, M5 introduce local misfit, despite the overall fit statistics are perfect.

  - M3: $\chi^2(376)$ = 1884.1, *TLI* = 0.993, *RMSEA* = 0.041 with ${}_{90\%}CI$ = [0.039, 0.042], SRMR = 0.016
  - M5: $\chi^2(378)$ = 2077.1, *TLI* = 0.992, *RMSEA* = 0.043 with ${}_{90\%}CI$ = [0.041, 0.045], *SRMR* = 0.016.

- Modification indices related to the height residual variance don't help to identify the problem. ($\chi_m^2(df = 1) = 24.1$, $\chi_f^2(df = 1) = 36.5$). Visual inspection of residual covariances does.

Latent trait identification: Conclusion

Factor score indetermination is not a problem.

Using rotation (or CFA with specific factors for positively and negatively scored items) biased the construct identification.

Using external criterion may identify a latent trait.

Common estimator failed to identify latent trait precisely as a height in restricted model, which was not obvious from the total fit statistics.

General, systematic, but height-irrelevant variance can be studied and used.

- It almost disappears in binary items!

Using the Height Inventory

References and more information:

- Cígler, H., Ježek, S., Rečka, K., Elek, D., Hubatka, P., Tancoš, M., & Šragová, E. *SCALING project.* https://doi.org/10.17605/OSF.IO/GQTA5
- The project is supported by the Czech Science Foundation (GA23-06924S).

Contact: Hynek Cígler

- mail: cigler@fss.muni.cz
- OSF: https://osf.io/t6ufg/
- GitHub: https://github.com/hynekcigler

Institute for Psychological Research
Faculty of Social Studies
Masaryk University, Czech Republic

Many thanks to several colleagues and our students, mainly:

- Standa Ježek, Karel Rečka, David Elek, Petra Hubatka, Martin Tancoš, Eva Šragová, Adam Strojil, Gabriela Kalistová, Petr Palíšek