MUNI
FSS

A unifying account of spurious multidimensionality in psychological questionnaires

Karel Rečka, David Elek

Department of Psychology, Faculty of Social Studies, Masaryk University, Brno, Czech Republic

Content

- Previous explanations of multidimensionality
- Our explanation of the phenomenon
- Empirical study:
  - Design and hypotheses
  - Results
  - Model vs. Item fit in detail
- Conclusions

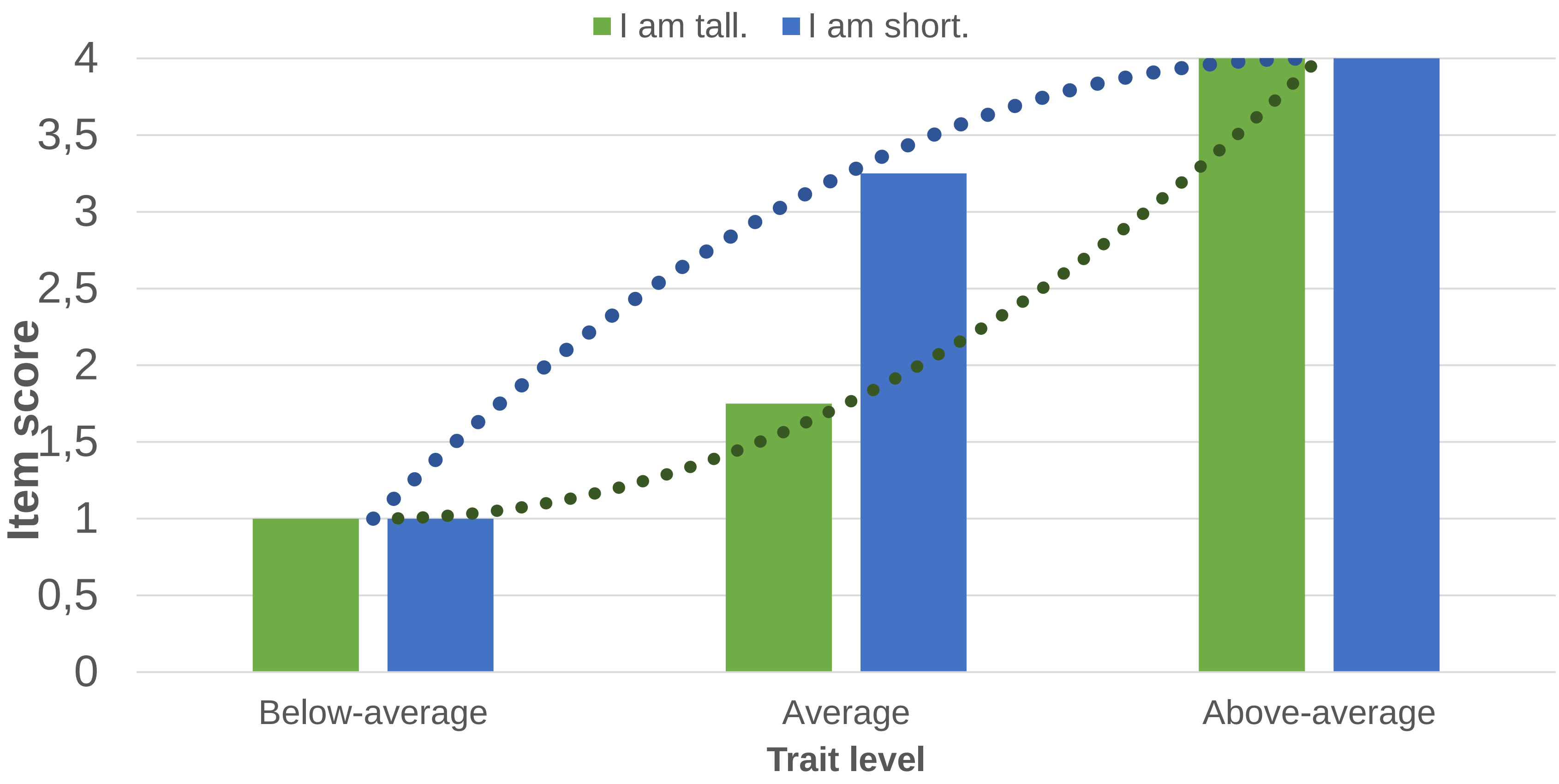
Motivation

- **Psychological questionnaires are rarely unidimensional**, especially when they contain both regular and reverse items.
- **Some authors dismiss reverse items** (multidimensionality contradicts theory, more complex models are necessary, lower reliability, confused respondents).
- Potential **benefits of reverse items**: implicit correction of response bias, reduction of monotony (higher engagement), better construct coverage (higher content validity).

Previous explanations

- Responses to reversed items are influenced by **construct-irrelevant factors to a greater extent or in a different direction than regular items**, such as acquiescence bias (Cronbach, 1942; 1950); social desirability (Krumpal, 2013; Paulhus, 1991; Rauch et al. , 2007); carelessness (Schmitt & Stults, 1985; Woods, 2006); or insufficient verbal ability (Marsh, 1996; Gnambs & Schroeders, 2020).
- More recently, Kam et al. (Kam et al., 2021; Kam & Meyer 2022) found that the relationship between the scores derived from regular and reverse items are **related in a nonlinear fashion**.
- Kam et al. argue that **the pattern of responses of "average" respondents to regular vs. reverse items is inconsistent** because they disagree with both regular and reverse items.

Item	Below-average	Average	Above-average
I am tall. (regular)	**Disagree** – – score **below** the midpoint	**Disagree** – score **below** the midpoint	**Agree** ++ score **above** the midpoint
I am short. (reversed)	**Agree** – – score **below** the midpoint	**Disagree** + score **above** the midpoint	**Disagree** ++ score **above** the midpoint



Older literature

- The notions of ***spurious multidimensionality*** appear in much older sources (Bernstein & Teng, 1989; Carroll, 1945; Ferguson, 1941).
- However, these authors framed the problem differently: **item difficulty**, together with their **ordinal** and **bounded** nature, affect the distribution of item responses (more difficult items are right skewed, easier items are left skewed).
- This affects the **strength of the correlations** between items, because the more the item distributions differ, the smaller the maximum correlation value can be.
- Regular items are usually more difficult than reverse items.

Our account

- What the previous authors describe is only a symptom.
- The true cause of spurious multidimensionality is a **misspecified relationship between a latent variable and its indicator (item response)**.
- In other words, the model implied relationships between a latent variable and its indicator(s) does match the empirical one.
- If the item response function is misspecified, items can **share a similar pattern of misfit/residuals**.
- If there are **multiple such shared patterns**, the unidimensional model will, by definition, show a poor fit to the data.
- Since **items share certain characteristics** (e.g., common response scale, difficulty), it is likely that the shared patterns of misfit/residuals emerge.

An empirical study

Instruments and design

- Three self-report inventories: Height Inventory, Weight Inventory, Age Inventory.
- Sample items: *I am taller than men of my age. I often need a stool to reach something other people would reach normally.*
- Two response scales: Likert (agree–disagree), item-specific (expanded item format).
- Two types of factor analysis: continuous (MLR) vs. ordinal (WLSMV).
- The participants also reported their height, weight, and age.
- For simplicity, we will focus on the **Height Inventory** with the traditional **Likert response scale** and **linear factor analysis** (that treats items as continuous, interval variables).

Research sample

- ***N* = 12,158 (49 % male)**.
- **Height ranged from 143 to 215 cm (*M* = 174.8, *SD* = 10.1).**
- Age ranged from 18 to 85 years (*M* = 36.5, *SD* = 13.8).
- Weight ranged from 40 to 172 kg (*M* = 81.0, *SD* = 19.6).
- BMI ranged from 14.2 to 59.1 kg/m^2 (*M* = 26.4, *SD* = 5.67).

Instruments and design

- Three self-report inventories: Height Inventory, Weight Inventory, Age Inventory.
- Sample items: *I am taller than men of my age. I often need a stool to reach something other people would reach normally.*
- Two response scales: Likert (agree–disagree), item-specific (expanded item format).
- Two types of factor analysis: continuous (MLR) vs. ordinal (WLSMV).
- The participants also reported their height, weight, and age.
- For simplicity, we will focus on the **Height Inventory** with the traditional **Likert response scale** and **linear factor analysis** (that treats items as continuous, interval variables).

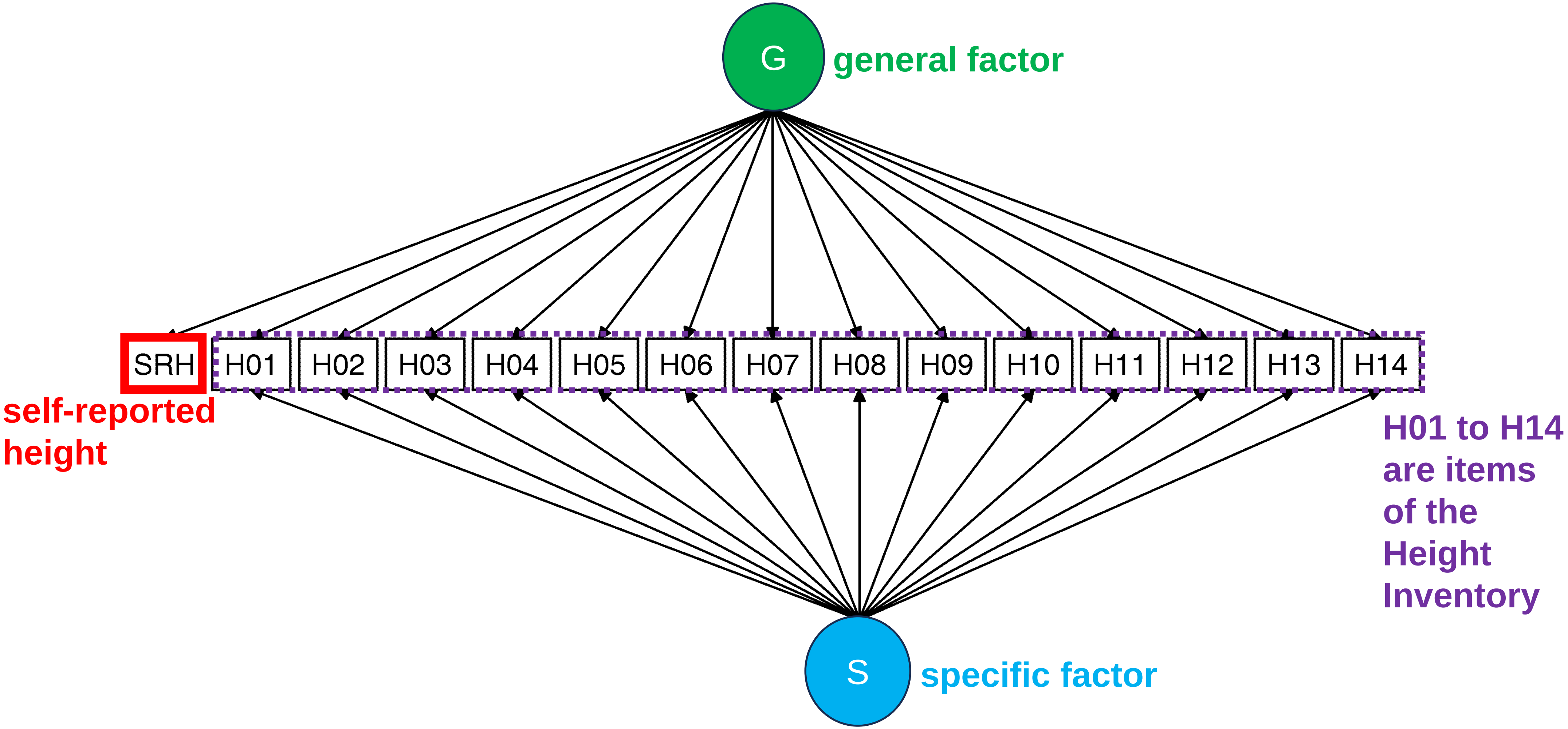
Aims and hypotheses

- Demonstrate that a **misspecified response function is a sufficient cause of spurious multidimensionality**.
- We expected:

1. More misfitting items to have stronger loadings on the specific factor.
2. The specific factor to still contain construct-relevant variance, that is, to be related to the general factor, but in a non-linear fashion.
3. The shape of their relationship to mirror the shared pattern of item misfit.

Instruments and design

- Three self-report inventories: Height Inventory, Weight Inventory, Age Inventory.
- Sample items: *I am taller than men of my age. I often need a stool to reach something other people would reach normally.*
- Two response scales: Likert (agree–disagree), item-specific (expanded item format).
- Two types of factor analysis: continuous (MLR) vs. ordinal (WLSMV).
- The participants also reported their height, weight, and age.
- For simplicity, we will focus on the **Height Inventory** with the traditional **Likert response scale** and **linear factor analysis** (that treats items as continuous, interval variables).


G
general factor
SRH
H01
H02
H03
H04
H05
H06
H07
H08
H09
H10
H11
H12
H13
H14
self-reported height
H01 to H14 are items of the Height Inventory
S
specific factor

Model fit

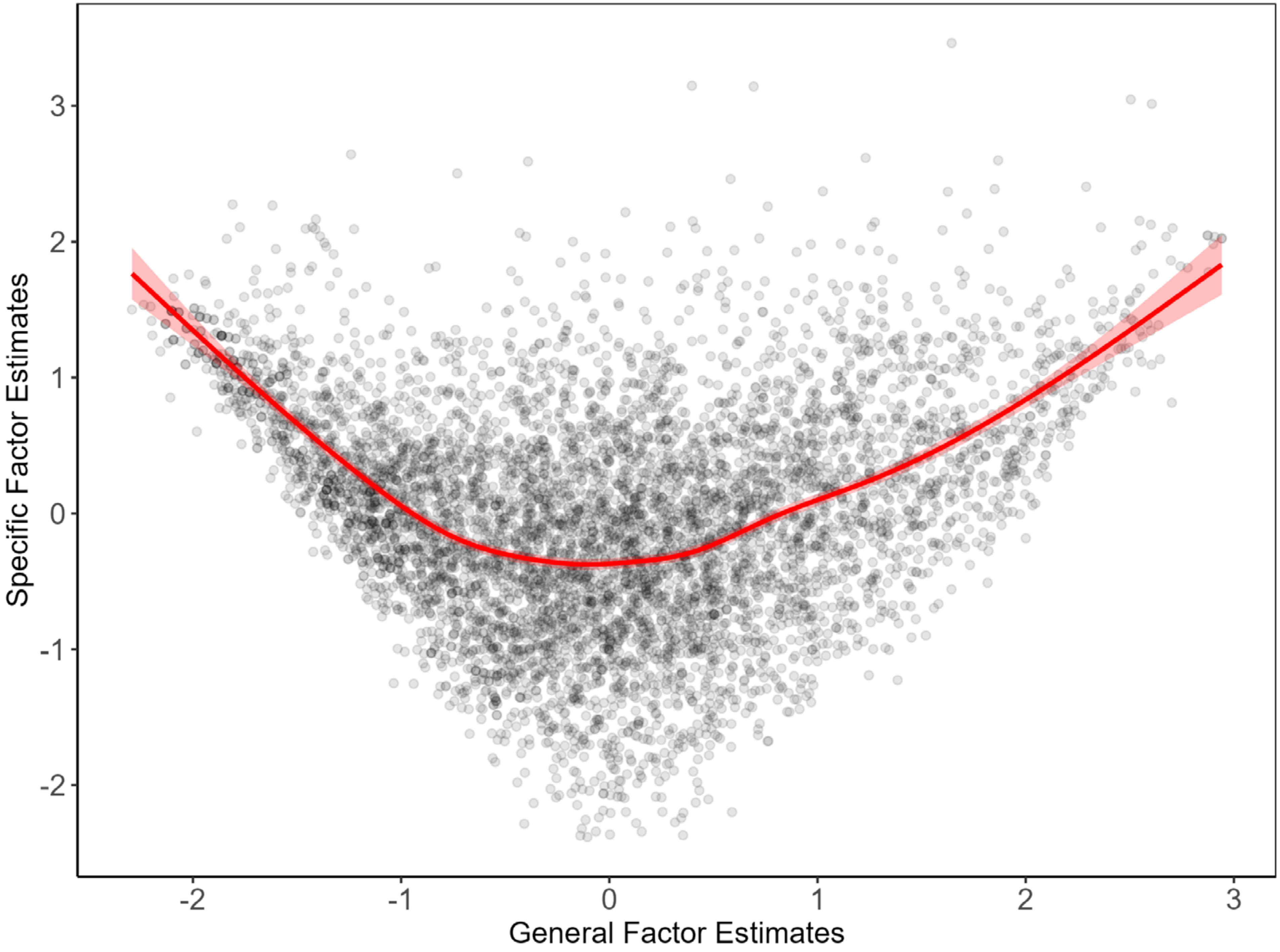
- **The unidimensional model showed a mediocre fit to the data**: $\chi^2(180) = 5280.9$ (unscaled 6623.1), $p < 0.001$, CFI = 0.917, TLI = 0.903, RMSEA = 0.077 (90% CI [0.075, 0.079]), SRMR = 0.049.
- **The bifactor i-1 model showed an excellent fit to the data**: $\chi^2(180) = 672.5$, (unscaled = 736,4) $p < 0.001$, CFI = 0.992, TLI = 0.989, RMSEA = 0.025 (90% CI [0.024, 0.027]), SRMR = 0.012.
- **The difference in fit was statistically significant**: $\chi^2(180) = 5085{,}6$, $p < 0.001$.

Item fit

- First, we have computed **factor scores estimates** for reach respondent**.**
- Second, we computed **model-predicted item scores** for each respondent and item.
- Then we computed "**empirical" item scores** using spline regression.
- The correlation between the model-predicted item scores and empirical item scores was used as a **measure of item fit**.
- As expected, the items with poor fit tended to have stronger loadings on the secondary factor.
- **The correlation between item fit and the loadings on the specific factor was strong**: Spearman's $\rho = -.70$,
  95% CI [–.86, –.41], $p < .001$.

Item fit

- First, we have computed **factor scores estimates** for reach respondent**.**
- Second, we computed **model-predicted item scores** for each respondent and item.
- Then we computed "**empirical" item scores** using spline regression.
- The correlation between the model-predicted item scores and empirical item scores was used as a **measure of item fit**.
- As expected, the items with poor fit tended to have stronger loadings on the secondary factor.
- **The correlation between item fit and the loadings on the specific factor was strong**: Spearman's $\rho = -.70$,
  95% CI [–.86, –.41], $p < .001$.

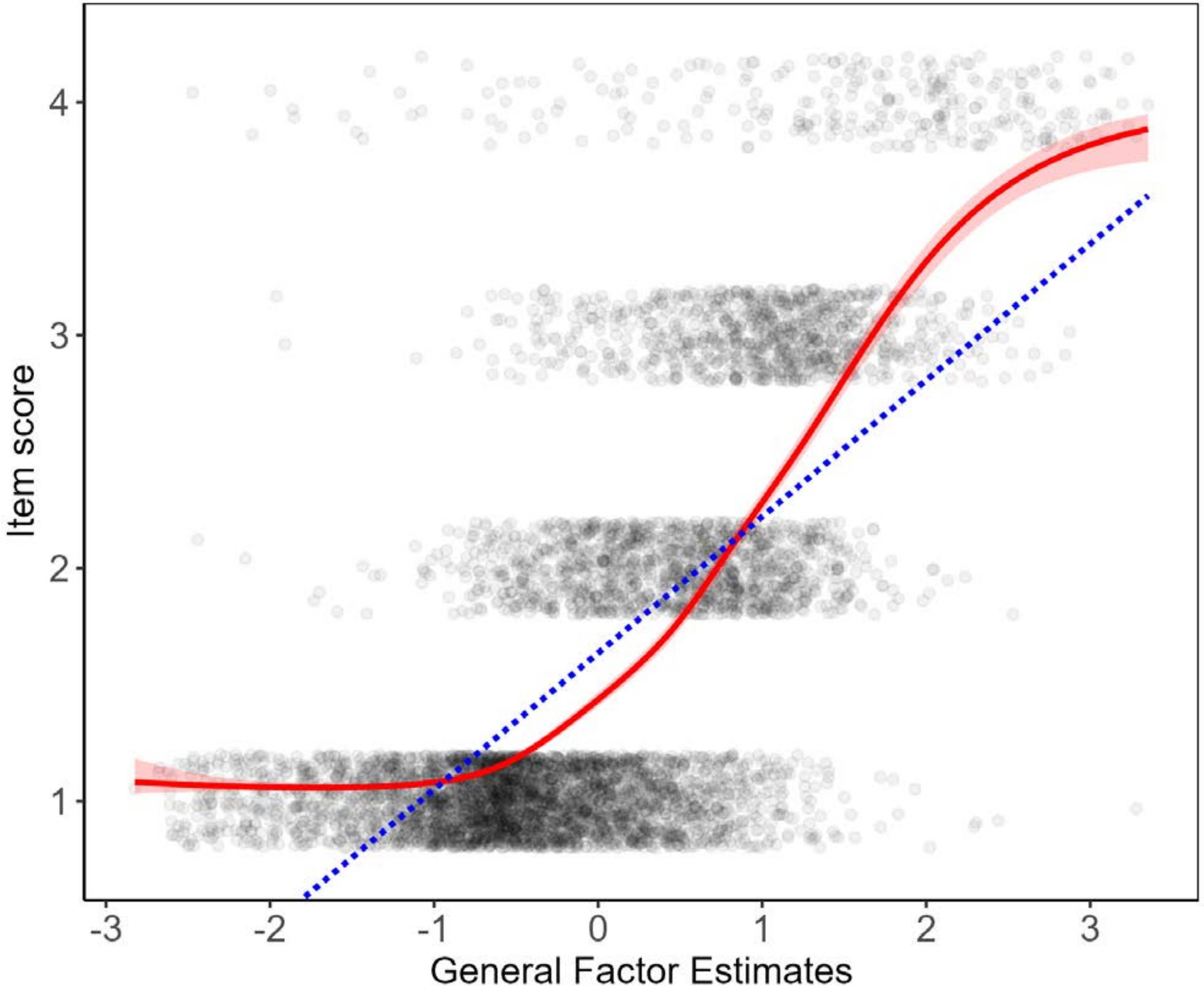


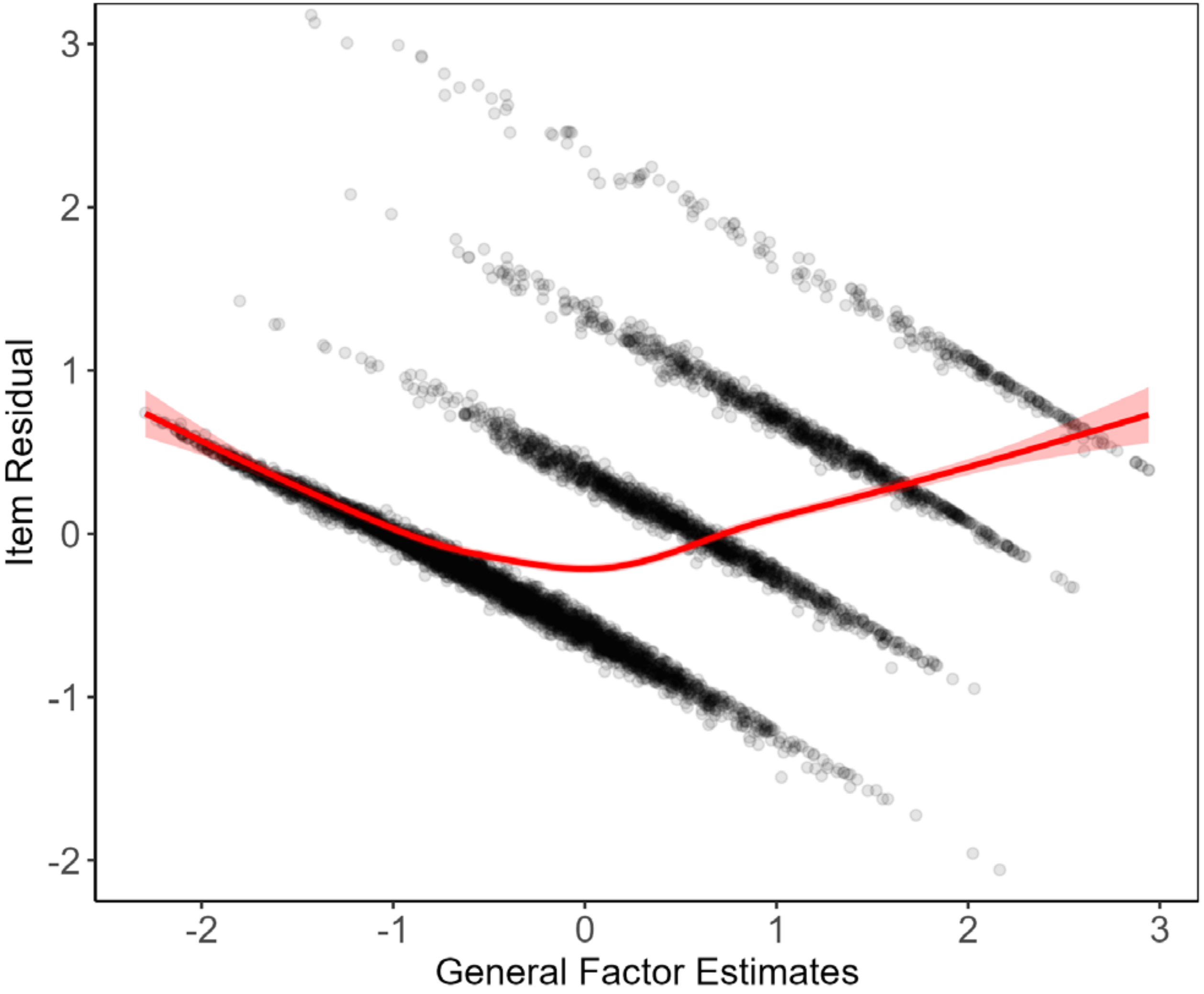
Cubic splines regression:
$R^2 = 0.28$

Linear regression with a quadratic term:
$R^2 = 0.27$

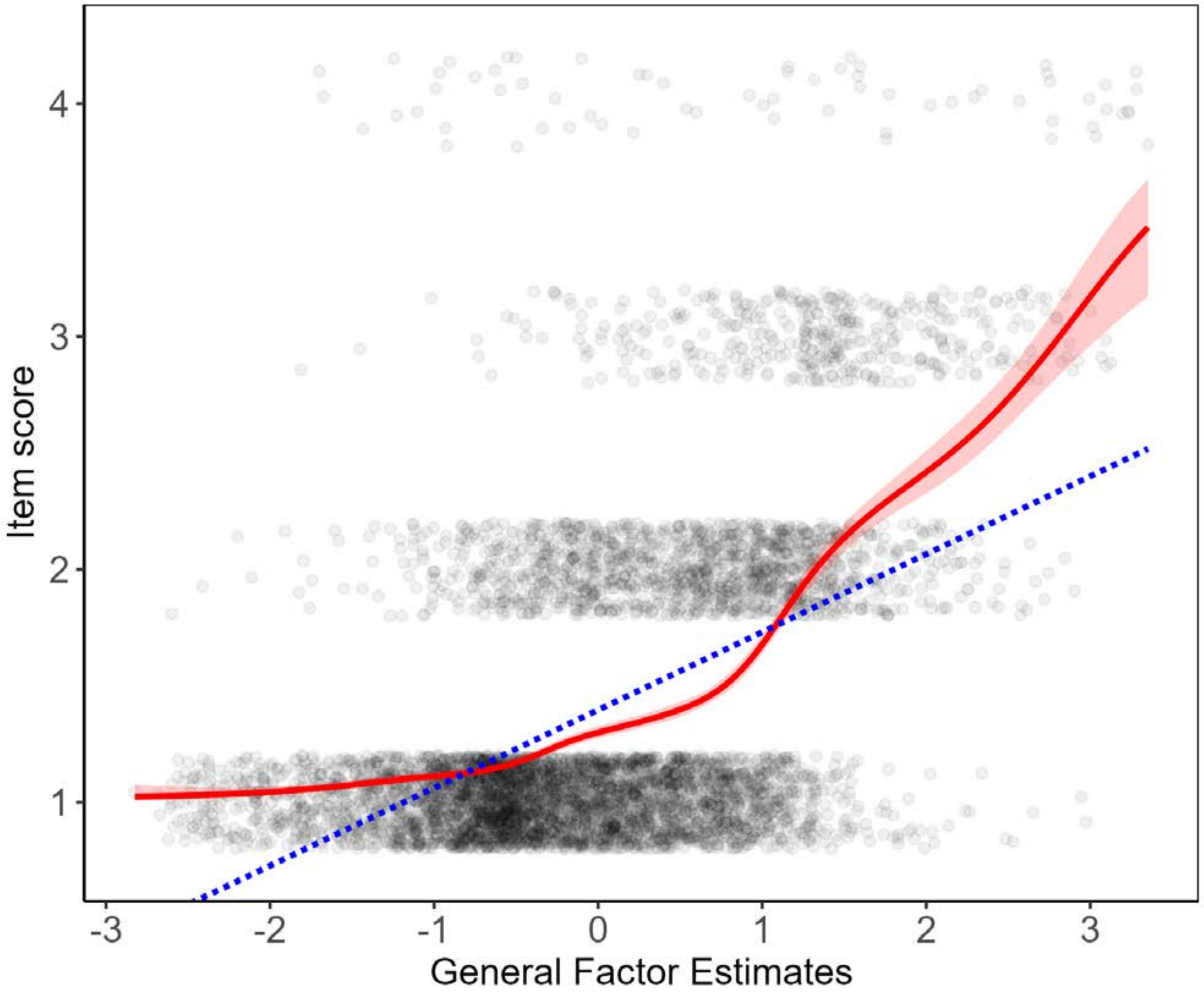
Squared Pearson correlation:
$r^2 < 0.01$

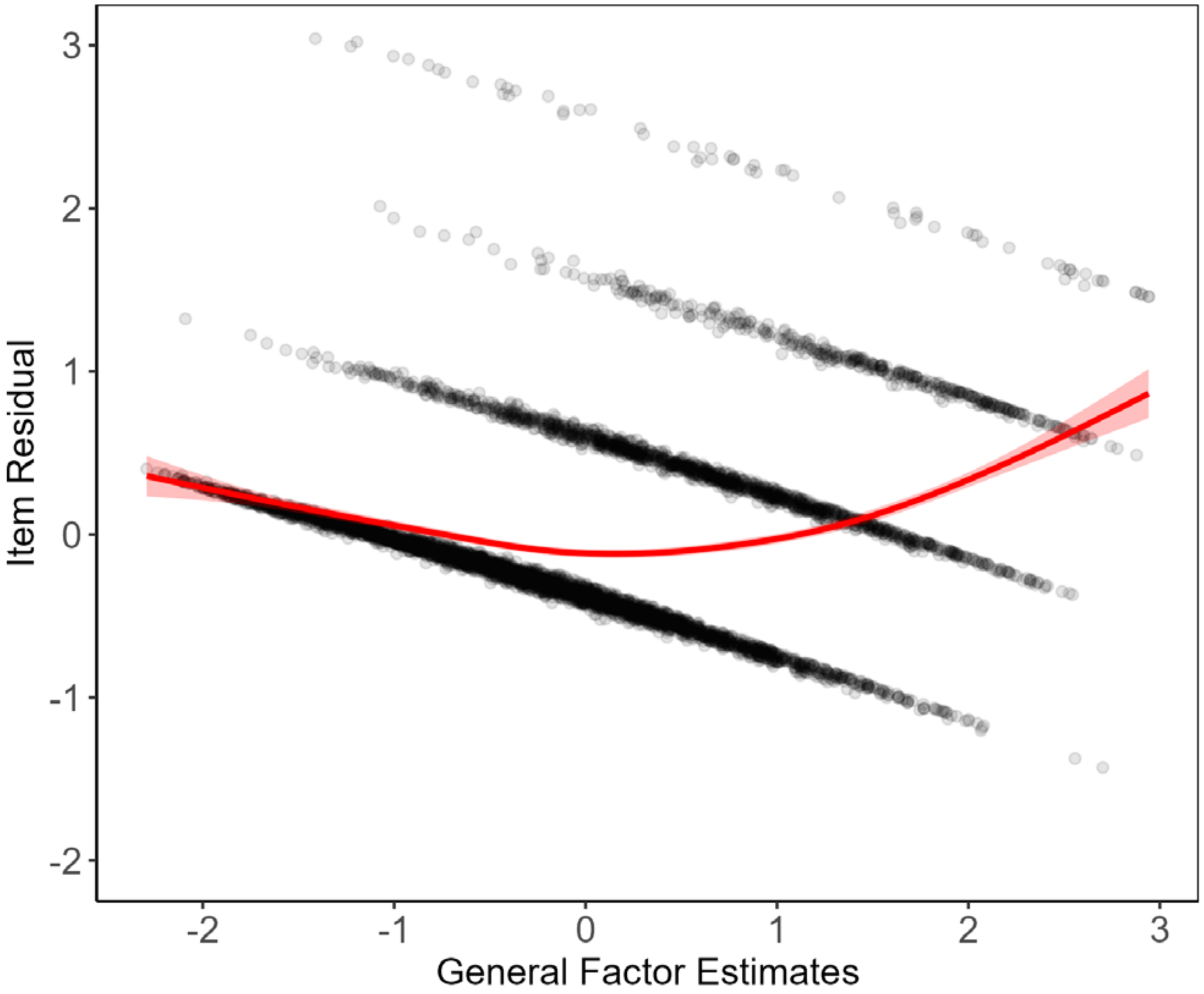
I am used to hearing comments about how tall I am.



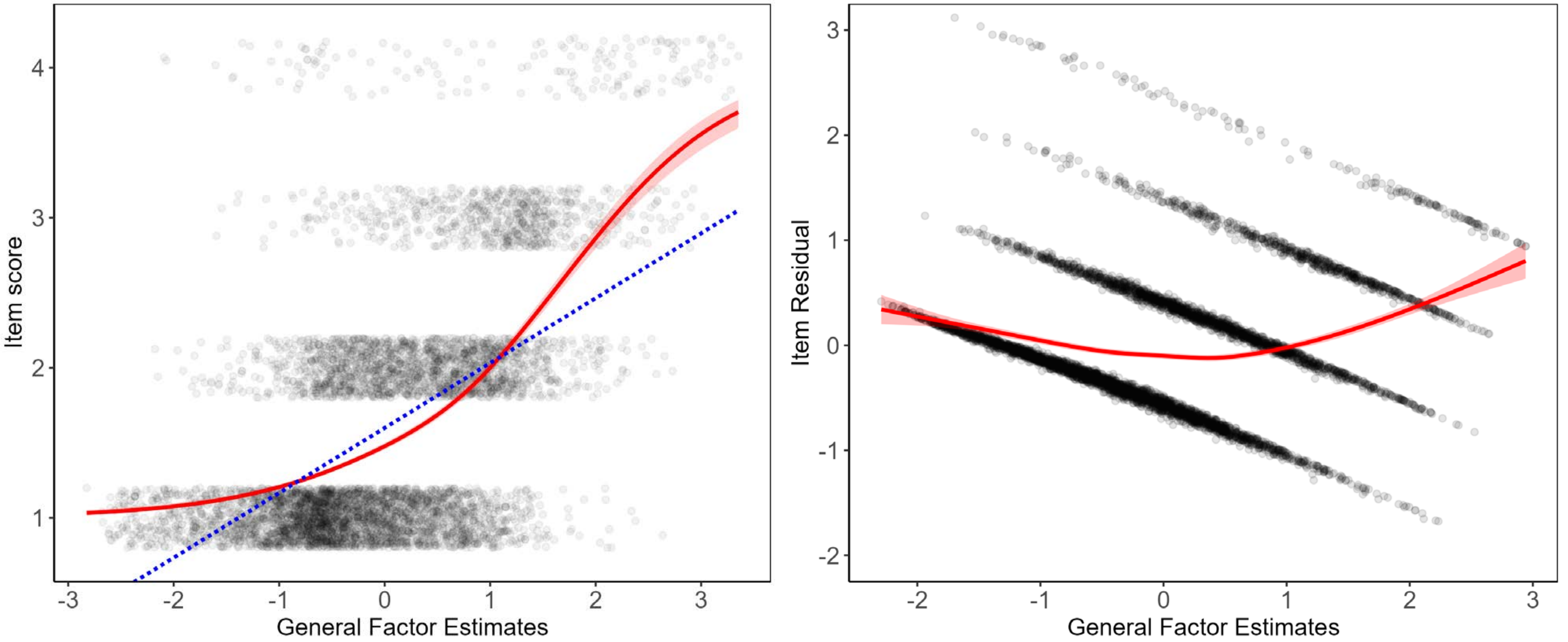


Ordinary beds are too short for me.

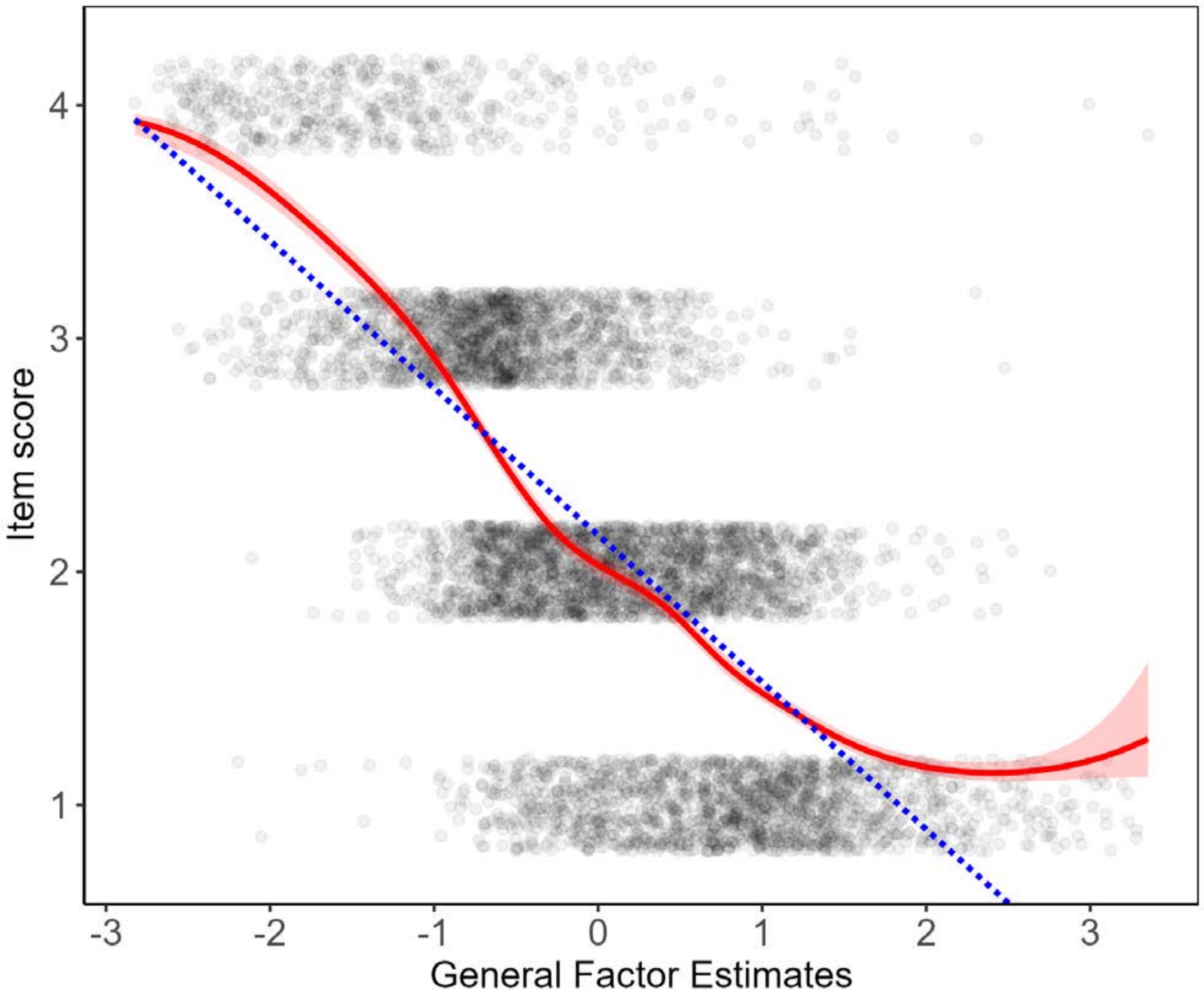


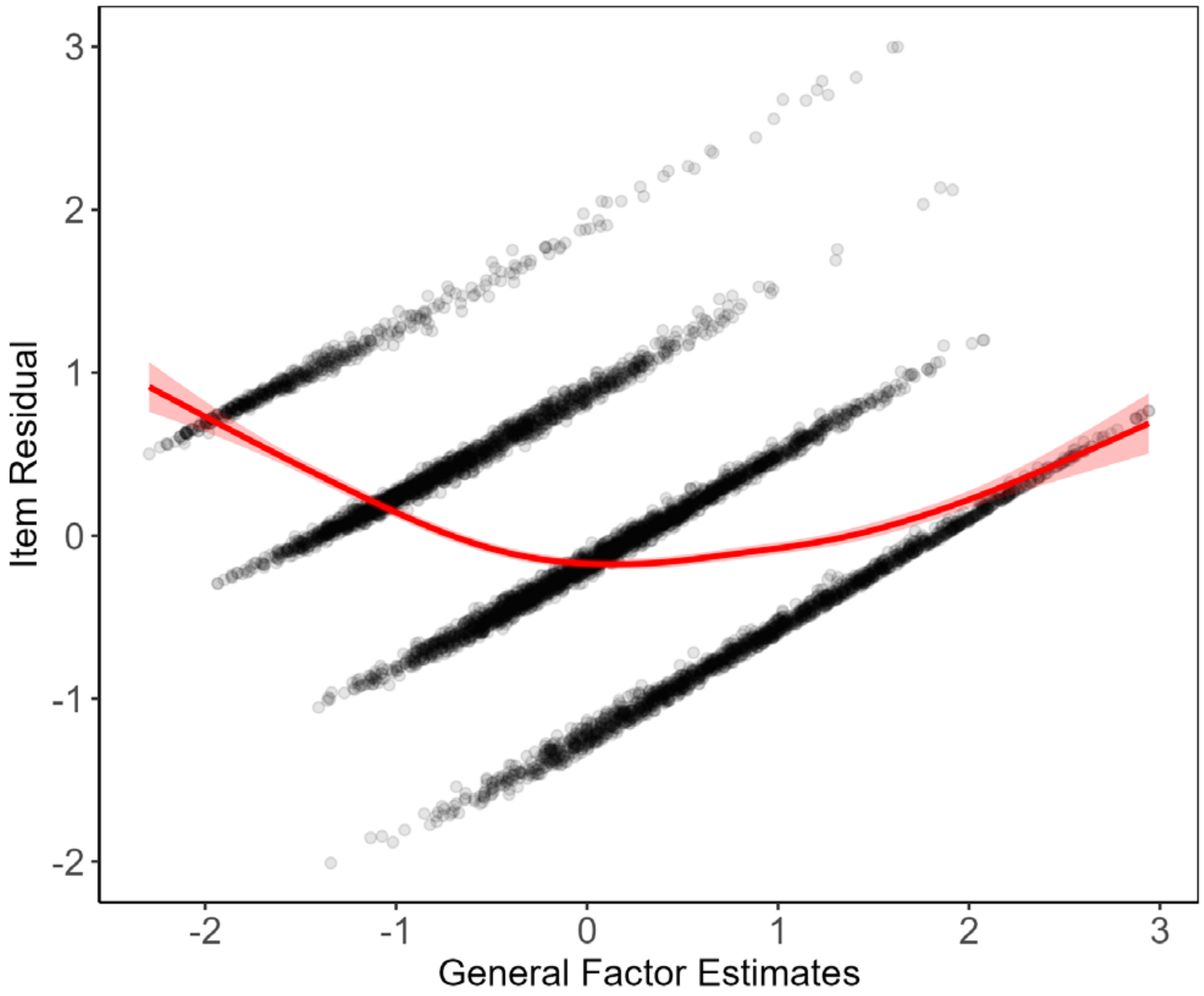


I often need to be careful to avoid bumping my head against a doorjamb or a low ceiling.

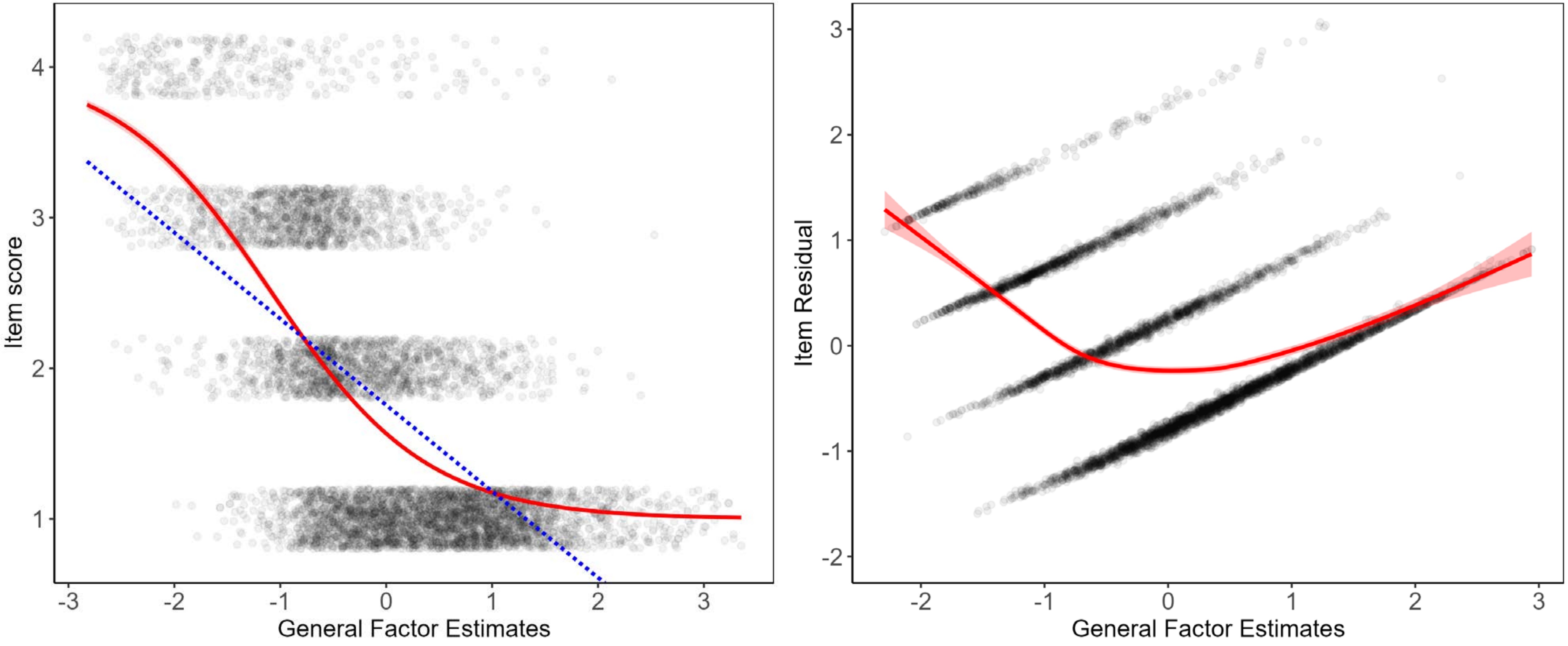


I often need a stool to reach something other people would reach without it. (reversed)





I could play a dwarf. (reversed)



Conclusions

- The results supported the hypothesis that a misspecified relationship between a latent variable and its indicator (item response) results in a shared pattern of misfit/residuals between items,
- In turn, this shared pattern results in a worse fit of a unidimensional model and the "emergence" of secondary factor(s).
- We know that these factors are "spurious" because they are nonlinearly related to the general factor and thus still contain construct-relevant variance.

Conclusions

- The results supported the hypothesis that a misspecified relationship between a latent variable and its indicator (item response) results in a shared pattern of misfit/residuals between items,
- In turn, this shared pattern results in a worse fit of a unidimensional model and the "emergence" of secondary factor(s).
- We know that these factors are "spurious" because they are nonlinearly related to the general factor and thus still contain construct-relevant variance.

Main takeaway

- In order to interpret the secondary factors as substantive, or content factors, it is first necessary to verify that the relationship between the latent variable and the items is not misspecified.
- Otherwise. there is a risk that the secondary factors are merely a statistical artifact.

What to do about it

- Thus, in the practical application of factor analysis, we recommend checking the following things to avoid interpreting spurious factors as substantive factors:

1. Is the relationship between the latent variable and the item specified correctly?
2. Do the items with the largest loadings on the secondary factor(s) share the same (or mirror-reversed) pattern of misfit/residuals (when plotted against the general factor)?
3. Is the primary factor strongly, but non-linearly related to the secondary factor(s). And if so, does the shape of the relationship mirror the pattern of residuals (from the previous step).

Funding

– This research was funded by the Grant Agency of the Czech Republic (project GA23-06924S)

Literature

- Bernstein, I. H., & Teng, G. (1989). Factoring items and factoring scales are different: Spurious evidence for multidimensionality due to item categorization. *Psychological Bulletin, 105*(3), 467–477. https://doi.org/10.1037/0033-2909.105.3.467
- Carroll, J. B. (1945). The effect of difficulty and chance success on correlations between items or between tests. *Psychometrika*, *10*(1), 1–19. https://doi.org/10.1007/bf02289789
- Cronbach, L. J. (1942). Studies of acquiescence as a factor in the true-false test. *Journal of Educational Psychology, 33*(6), 401–415. https://doi.org/10.1177/001316444600600405
- Cronbach, L. J. (1950). Further evidence on response sets and test design. *Educational and Psychological Measurement*, *10*(1), 3–31. https://doi.org/10.1177/001316445001000101
- Ferguson, G. A. (1941). The factorial interpretation of test difficulty. *Psychometrika*, *6*(5), 323–329. https://doi.org/10.1007/bf02288588
- Gnambs, T., & Schroeders, U. (2020). Cognitive abilities explain wording effects in the Rosenberg Self-Esteem Scale. *Assessment*, *2*, 404–418. https://doi.org/10.1177/1073191117746503

Literature

- Kam, C. C., & Meyer, J. P. (2022). Testing the nonlinearity assumption underlying the use of reverse-keyed items: A logical response perspective. *Assessment*, *0*(0), 1–21. https://doi.org/10.1177/10731911221106775
- Kam, C. C., Meyer, J. P., & Sun, S. (2021). Why do people agree with noth regular and reversed Items? A logical response perspective. *Assessment*, *28*(4), 1110–1124. https://doi.org/10.1177/10731911211001931
- Krumpal, I. (2013). Determinants of social desirability bias in sensitive surveys: a literature review. *Quality & Quantity*, *47*(4), 2025–2047. https://doi.org/10.1007/s11135-011-9640-9
- Marsh, H. W. (1996). Positive and negative global self-esteem: A substantively meaningful distinction or artifactors. *Journal of personality and social psychology*, *70*(4), 810–819. https://doi.org/10.1037//0022-3514.70.4.810
- Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson`, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of Personality and Social Psychological Attitudes* (pp. 17–59). Academic Press.
- Rauch, W. A., Schweizer, K., & Moosbrugger, H. (2007). Method effects due to social desirability as a parsimonious explanation of the deviation from unidimensionality in LOT-R scores. *Personality and Individual Differences*, *42*(8), 1597–1607. https://doi.org/10.1016/j.paid.2006.10.035