Cumulative *vs* unfolding IRT models

How practice may defy theory

Jorge Tendeiro, Rob Meijer

ECPA 2017, Lisbon

07 July 2017

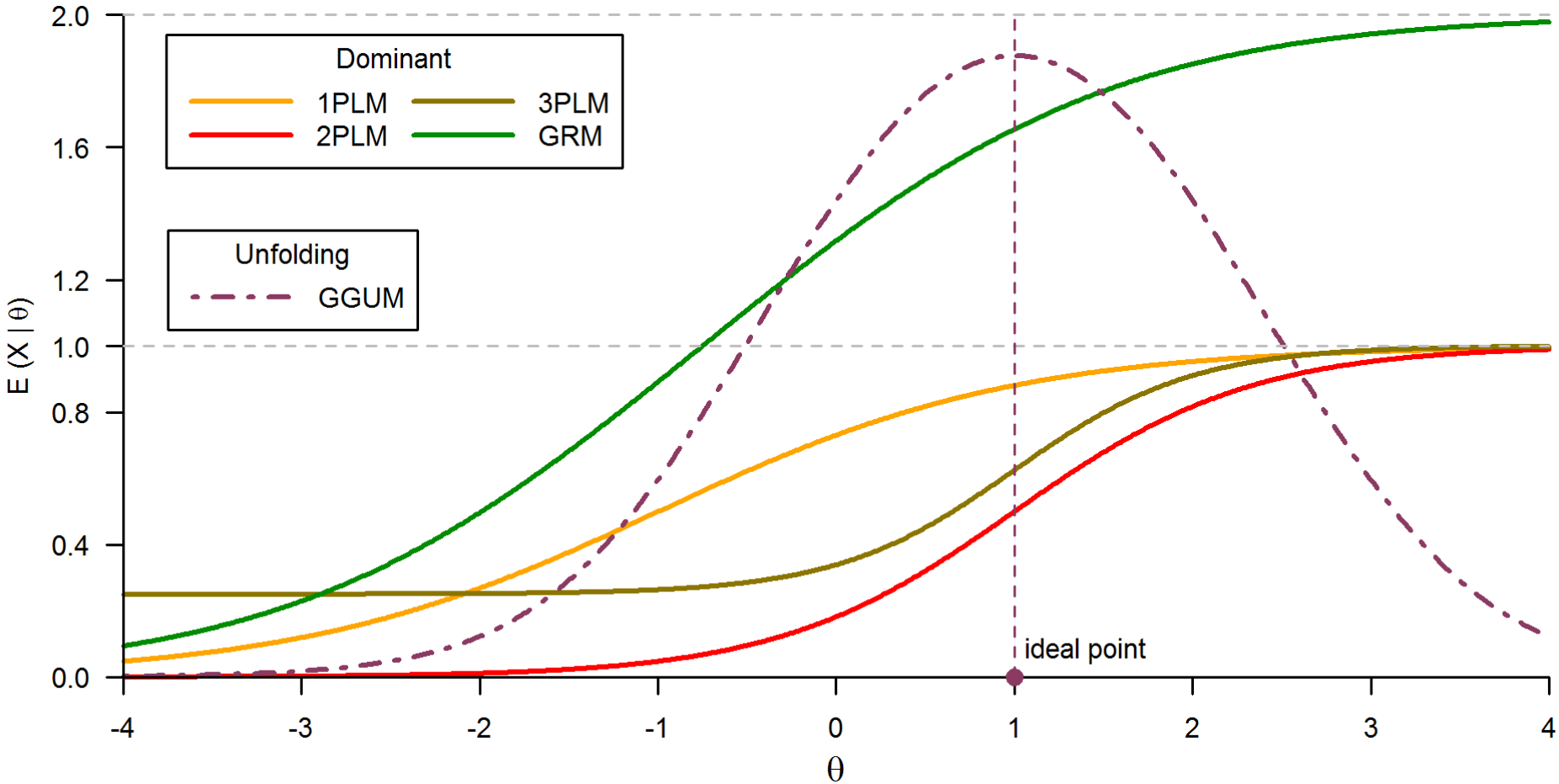
university of groningen

- IRT models are nowadays very popular psychometric tools.
- Several advantages of IRT (in comparison to CTT), e.g.:
  - Estimation of model parameters is not sample-dependent (up to a linear transformation).
  - SE of measurement may vary along the latent trait.
  - A broad variety of item/scale types are allowed.
- However, models need to be carefully chosen.
- Avoid fitting models just because they are 'typically prescribed.'
- My goal:
  - Show that, quite often, 'practice defies theory.'
  - I focus on dominant versus unfolding IRT models.

Two types of models (Coombs, 1964):

- Dominant – The probability of endorsing an item increases with θ.
  - Conceivable under maximum performance settings (e.g., cognitive tests).
  - It is the most common type of model.
  - 'The more the merrier.'

- Unfolding – The probability of endorsing an item increases as θ approaches the item's location ('ideal point').
  - Conceivable under typical performance settings (e.g., measuring preferences or attitudes).
  - It requires introspection:
    How typical is this item for me?
    How much do I agree with this item's statement?
  - 'The closer the merrier.'


Dominant
1PLM
2PLM
3PLM
GRM
Unfolding
GGUM
ideal point
E (X | θ)
θ
2.0
1.6
1.2
1.0
0.8
0.4
0.0
-4
-3
-2
-1
0
1
2
3
4

About the unfolding mechanism:

- What matters the most is the perceived distance between the item's statement and the person's standing.
- Conceivably, more of the latent trait may decrease the probability of endorsement.
- A person may disagree with an item statement's because she either believes too strongly in favor of it ('too far' to the right) or against it ('too far' to the left).

Altogether, this conceptualization of item endorsement is fundamentally distinct from the much more common dominant process.

When should we use dominant/unfolding models?

An attractive rule of thumb:

> *Use unfolding when assessing attitudes or preferences.*
> *(e.g., Drasgow, Chernyshenko, & Stark, 2010; Stark et al., 2006).*

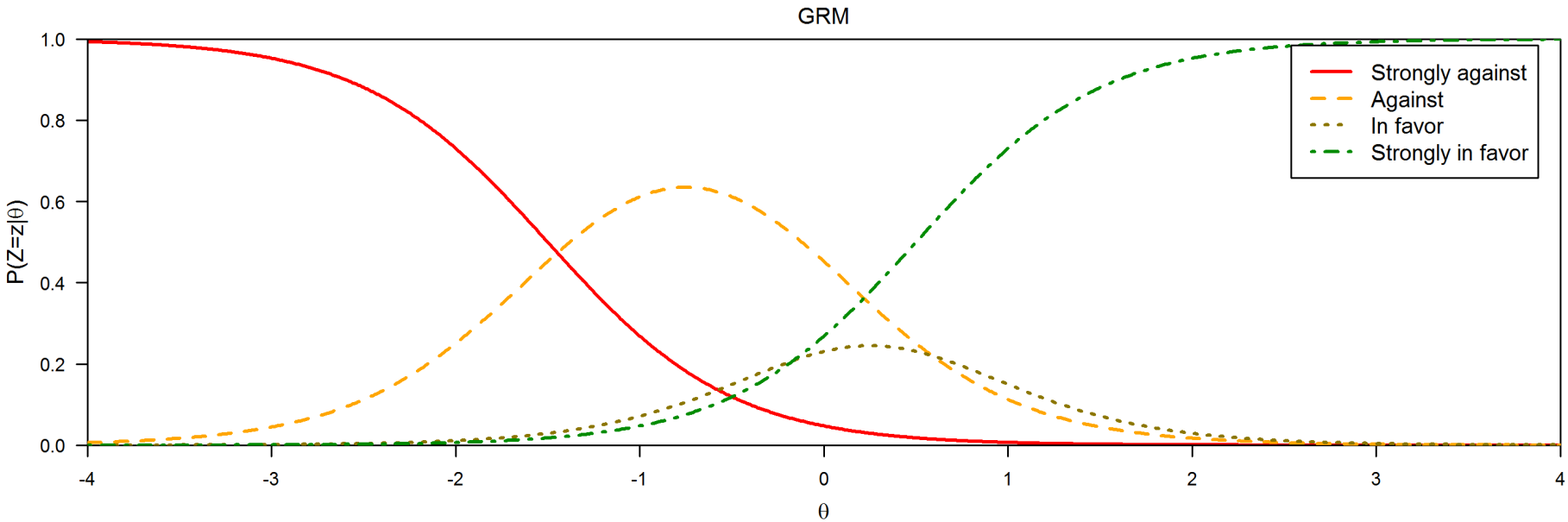
However, some research has suggested that unfolding models do not always fit when one would expect them to (e.g., Cho, Drasgow, & Cao, 2015; Huang & Mead, 2014; LaPalme et al., 2016).

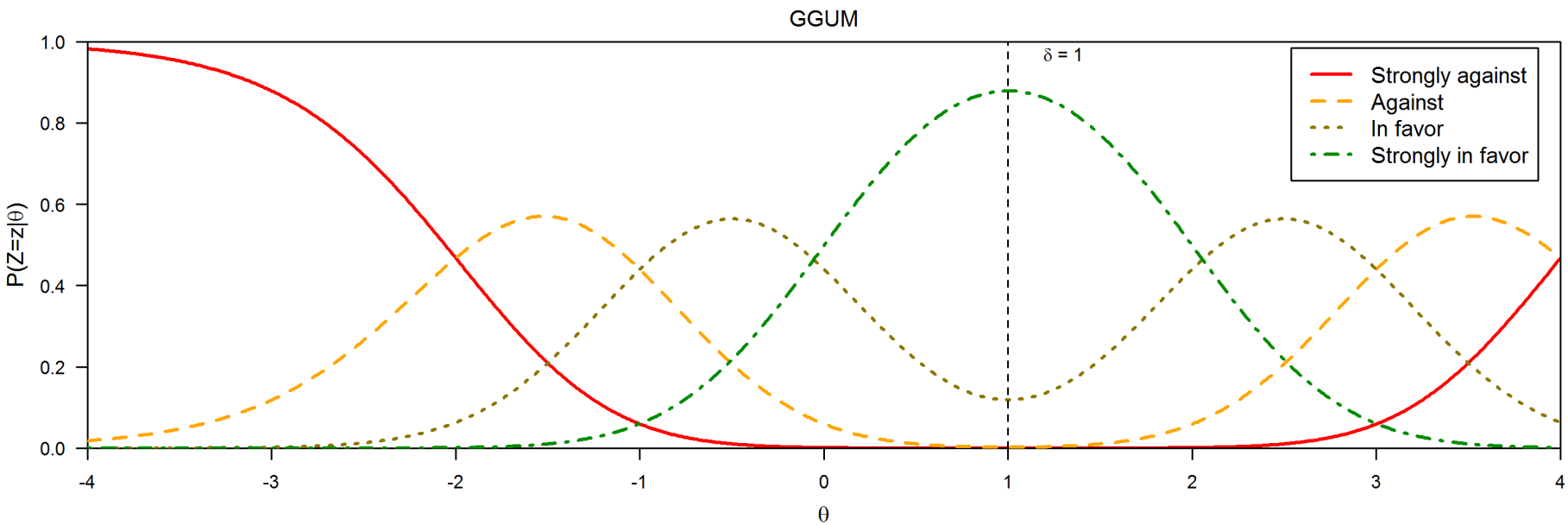
I fit both types of model to several datasets where unfolding would be conceivable.
Then I compared model fit.
My approach is purely empirical: Compare both types of models under various types of empirical data.
(No simulations for a change!)

Contributions to the current state of affairs:

- Use datasets with large samples (in some cases $N > 40,000$).
- Use original polytomous scores (no dichotomization).
- Overview over 27 data sets in total.
- Fit the GGUM (seemingly, the most popular unfolding model available) using dedicated R functions (to be available soon in CRAN) and compare to the GRM.
- Use various (relative) model fit measures.
- Use cross validation to compare predicted distributions.

IRT models used





Five (sets of) datasets were analyzed

1. Cattell's 16 personality factors test with items from the IPIP ($N = 49,159$)
   - Retrieved from `http://personality-testing.info`
   - Sixteen 5-point Likert subscales analyzed individually

Warmth	Liveliness	Vigilance	Openness to Change
Reasoning	Rule-Conscious.	Abstractedness	Self-Reliance
Emot. Stability	Social Boldness	Privateness	Perfectionism
Dominance	Sensitivity	Apprehension	Tension

   - Sample items:
     I'm not really interested in others. (Warmth)
     I'm hard to get to know. (Privateness)

2. Big Five personality test with items from the IPIP ($N = 19,719$)
   - Retrieved from `http://personality-testing.info`
   - Five 5-point Likert subscales analyzed individually.
   - Sample items:
     I feel comfortable around people. (Extraversion)
     I get upset easily. (Neuroticism)

3. Teacher-student interaction in secondary classrooms ($N = 1,196$)
   - Questionnaire on Teacher Interaction (QTI; van der Lans, 2017; Wubbels et al., 1985).
   - 5-point Likert scale (never, almost never, . . . , always).
   - Sample items:
     This teacher has a sense of humour.
     This teacher gives an insecure impression.

4. Undergraduates evaluation of a new digital testing system (Boeve et al., 2015; $N = 958$)
   - Combined sample (years 2014, 2015, 2016).
   - 4-point Likert scale (completely disagree, . . . , completely agree).
   - Sample items:
     It was clear how to start the exam.
     It was easy to navigate through the exam.
     (less 'unfolding'-like?)

5 STAR*D depression inventories.

- Hamilton Rating Scale (HRS-D_{17}); $N = 4,039$.
- Quick Inventory of Depressive Symptomatology (QIDS-C_{16}); $N = 3,921$.
  Three subscales from the QIDS were considered:
  - Patients displaying increasing appetite and weight.
  - Patients displaying decreasing appetite and weight.
  - Appetite and weight indicators removed.
- At baseline, rated by clinicians.
- Mix of 3- to 5-point Likert scale items.
- Sample items:
  - *Somatic anxiety*
    ('Absent', ..., 'Severe')
  - *Suicidal ideation*
    ('Does not think of suicide or death', ..., 'Thinks of suicide/death several times a day')

  (less 'unfolding'-like?)

The following three-step procedure was followed for each dataset:

1. Preprocess the data:
   - Remove response patterns with more than 50% missingness.
   - Remove response patterns with disagree-only answers (as per Roberts & Shim, 2008).
   - Merge response categories with low frequencies (typically < 20).

2. Fit both the GRM and the GGUM. Compare fit. (logLik, AIC, BIC, adjusted χ^2/df)

3. Perform cross-validation (Nreps $= 100$):
   - Estimate item parameters for both models in training sample.
   - Estimate person parameters for subjects in testing sample.
   - Compare observed distribution with expected distribution from testing sample. (bias, RMSD, Kullback-Leibler divergence)

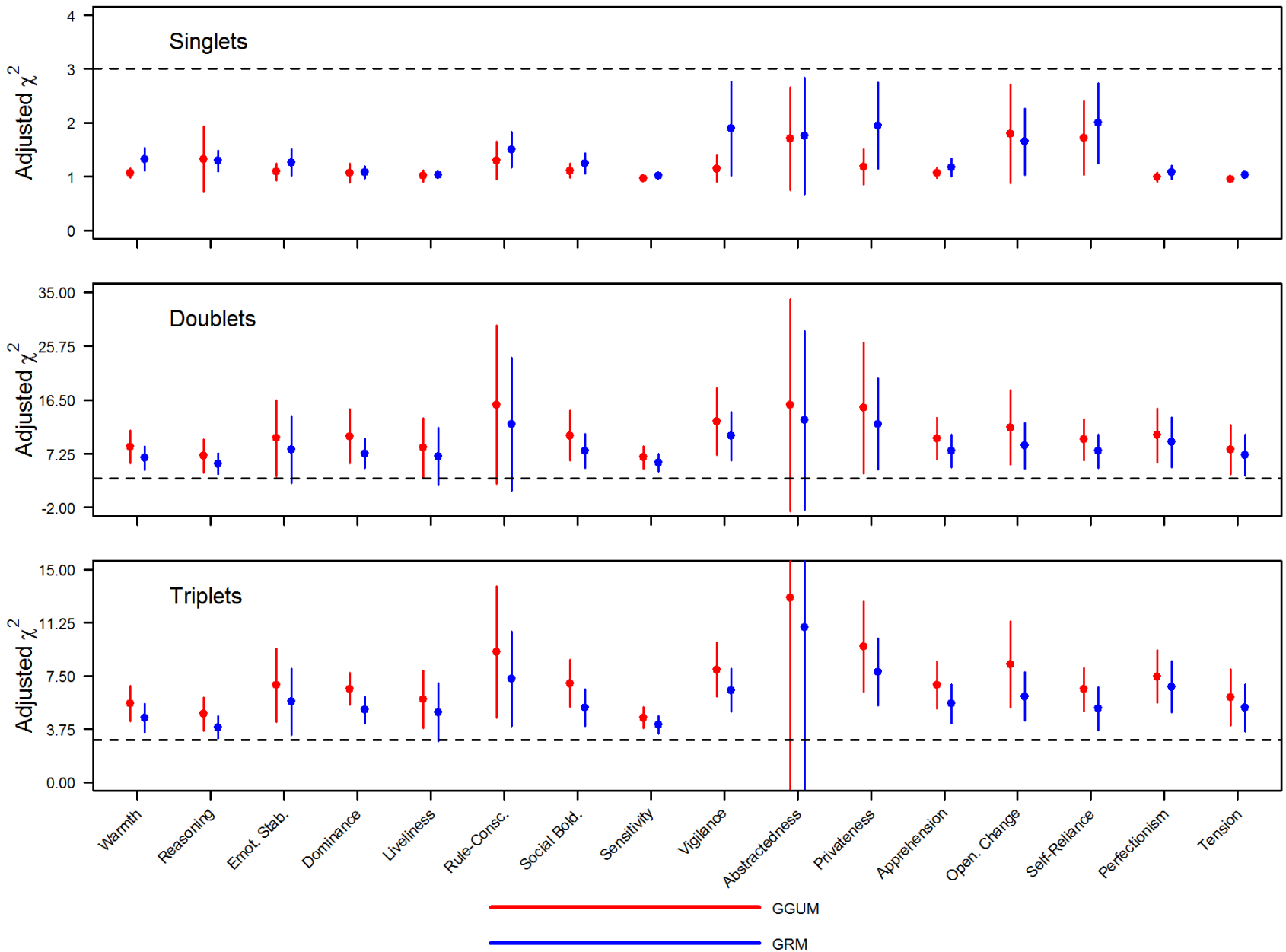
	Cattell (#16)	Big Five (#5)	QTI	Boeve	HRSD	QIDS (#3)
Removed? ($>$ 50% NAs)	✓	✓	✗	✓	✗	✗
Removed? (all-disagree)	✓	✓	✗	✓	✗	✓
Categories merged?	✗	✗	✓	✗	✓	✓
Items dropped?	✓ (1-2)	✗	✗	✗	✓ (1)	✓ (1)
AIC/BIC	GRM	GRM	GRM	GRM	GRM	GRM
Adjusted χ^2/df	GGUM (S.) GRM (D., T.)	GGUM (S.) GRM (D., T.)	GGUM	GRM	$\approx$	GRM ($\approx$)
Neutral items?	✗	✗	✓ (2)	✓ (1)	✗	✓ (1), ✗, ✗

Obs: S., D., T. = MODFIT's singles, doubles, triples.

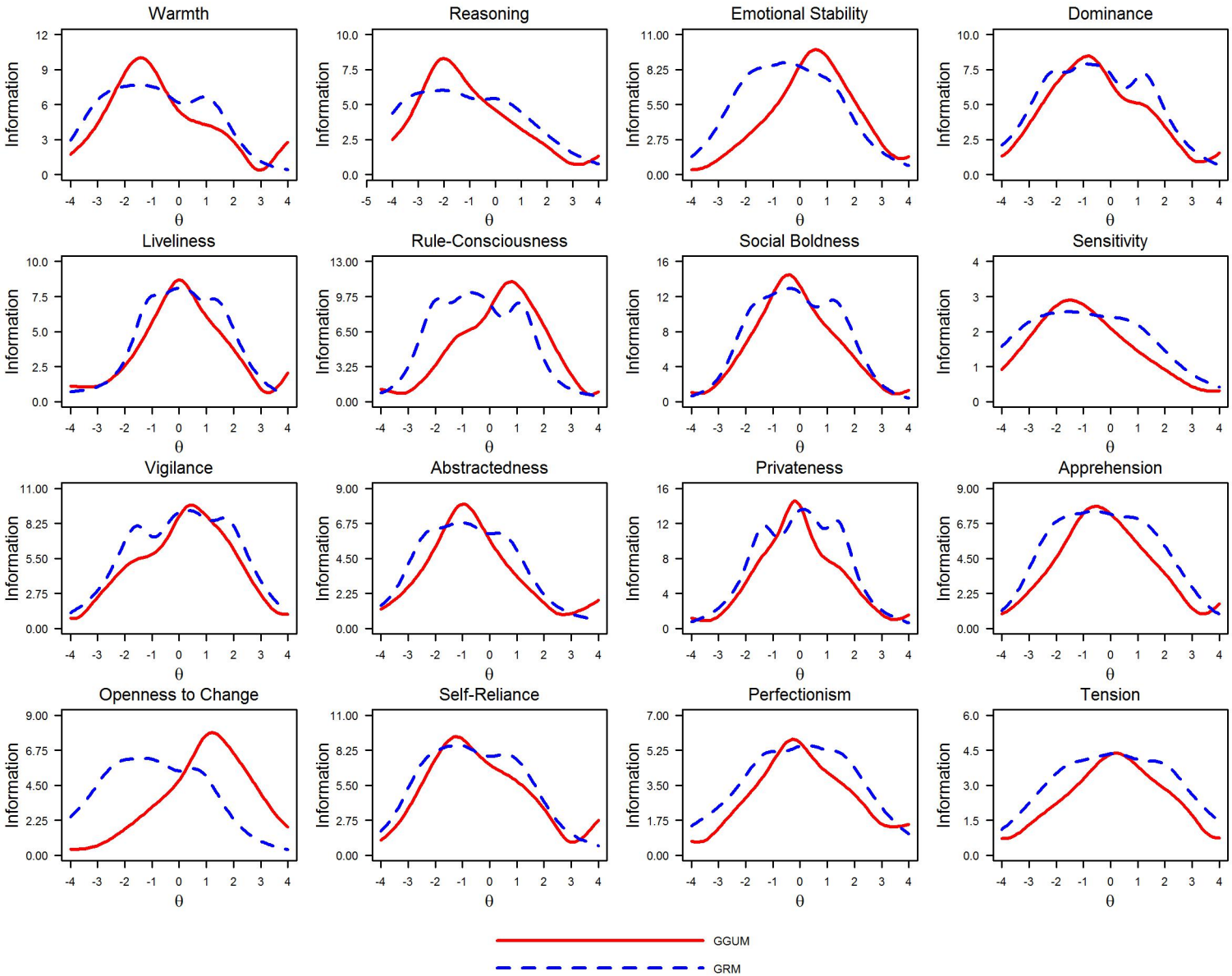
	Cattell (#16)	Big Five (#5)	QTI	Boeve	HRSD	QIDS (#3)
Bias	GGUM (negligible)	GGUM (negligible)	≈	GGUM (negligible)	≈	≈
RMSD	≈	≈	≈	≈	≈	≈
K-L divergence	GRM (most items)	GRM (most items)	GRM (negliglble)	≈	≈	≈

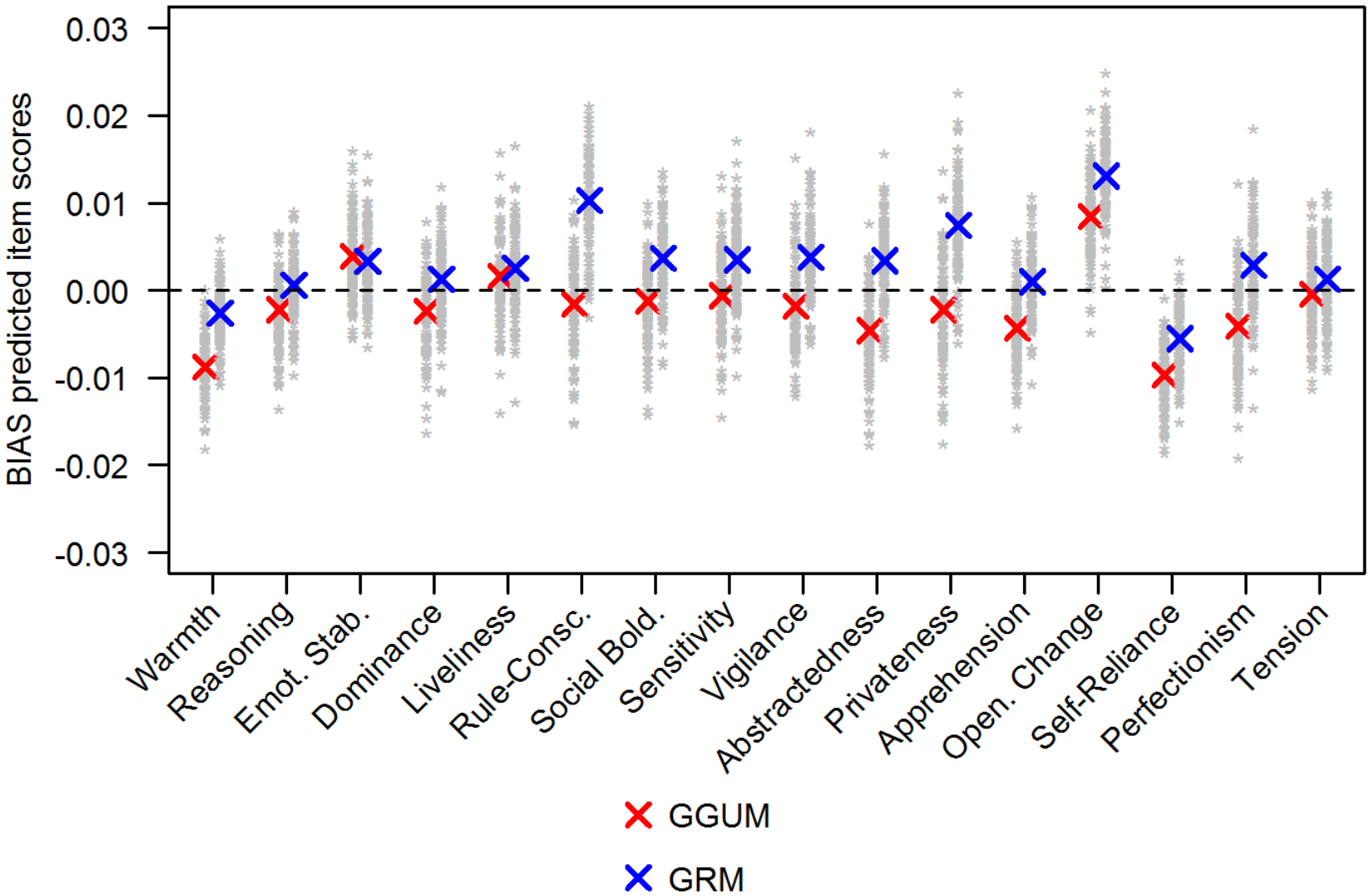
Obs: K-L = Kullback-Leibler.

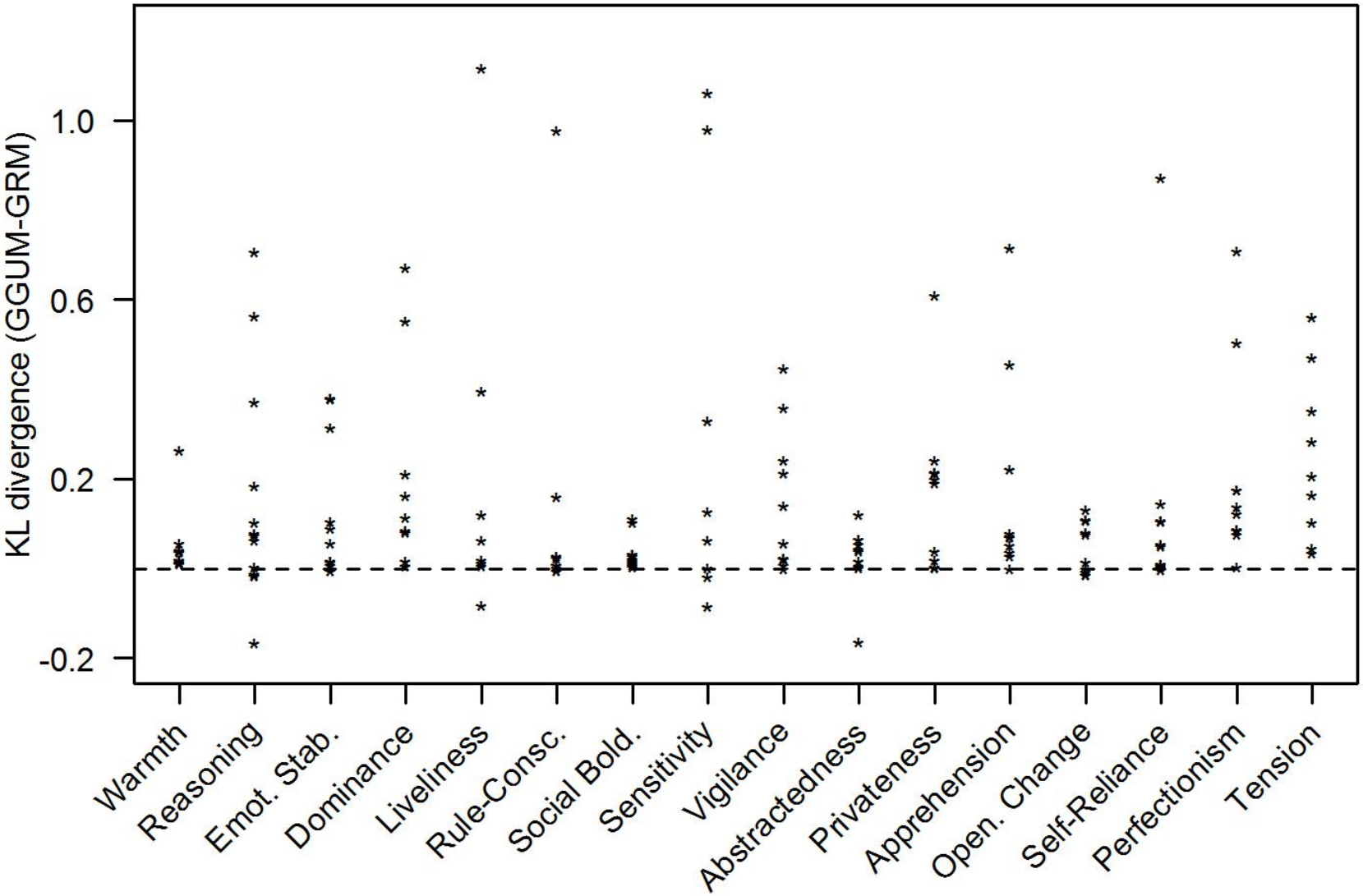
Cattell's 16 PF test [Model fit: Adjusted χ^2/df]



Cattell's 16 PF test [Model fit: Test information function]




BIAS predicted item scores
0.03
0.02
0.01
0.00
-0.01
-0.02
-0.03
Warmth
Reasoning
Emot. Stab.
Dominance
Liveliness
Rule-Consc.
Social Bold.
Sensitivity
Vigilance
Abstractedness
Privateness
Apprehension
Open. Change
Self-Reliance
Perfectionism
Tension
GGUM
GRM


KL divergence (GGUM-GRM)
1.0
0.6
0.2
-0.2
Warmth
Reasoning
Emot. Stab.
Dominance
Liveliness
Rule-Consc.
Social Bold.
Sensitivity
Vigilance
Abstractedness
Privateness
Apprehension
Open. Change
Self-Reliance
Perfectionism
Tension

- GGUM hardly ever fit better.
- Test information typically favored GRM over broader ranges of θ.
- Predicted distributions either favored the GRM (Cattell, Big Five) or were very similar between GGUM and GRM.
- Few neutral items (i.e., with $|\text{location}| < 2$; Ling et al. 2016) identified, thus unfolding not identified. This might indicate bias towards the GRM due to scale construction.

  (based on dominant procedures: Item-rest correlations, factor analysis, etc.).

- Interestingly:
  - Item discrimination parameters between GGUM and GRM highly correlated (at least $> .93$).
  - Latent person parameters extremely similar (r's close to 1, very small mean differences).
    $\longrightarrow$ If estimating the latent standing of persons is the main goal, then both models seem highly agreeable.

Altogether, fitting the GGUM did not seem to pay off in spite of the seemingly unfolding nature of many items.

- Including (sharper) ideal point items during scale construction is crucial for models such as the GGUM to shine.
- Without ideal point items, the GRM seems to be preferable (less parameters, better fit, simpler to interpret (?)).
- There are good reasons to consider including ideal point items when measuring attitudes or preferences (see Chernyshenko et al., 2007; Drasgow et al., 2010). Most currently used scales do not include such items.
- The choice of the most suitable psychometric model is a matter of the data at hand as well as of the typology of the administered scale.
- There are no written-in-stone rules. Proper modeling is an art!

Obrigado.

j.n.tendeiro@rug.nl