Unfolding IRT models

Do they always fit when expected?

Jorge Tendeiro, Rob Meijer

IMPS 2017, Zürich

20 July 2017

Motivation

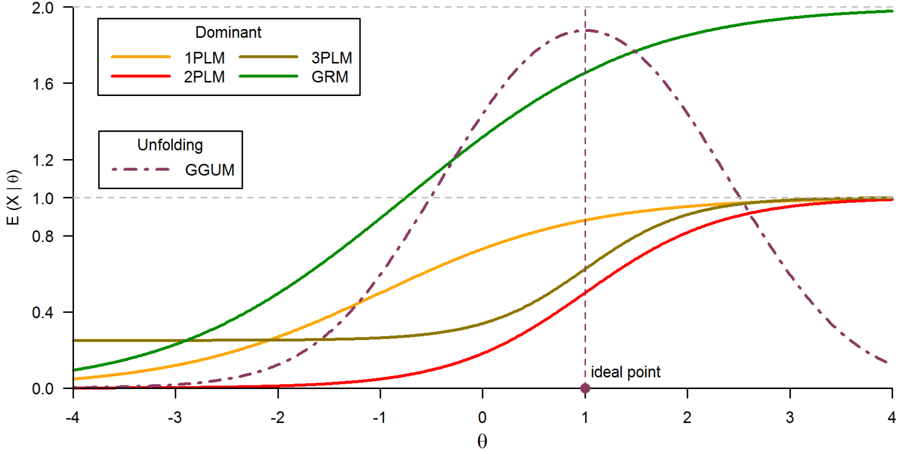
- IRT models are nowadays very popular psychometric tools.
- A multitude of IRT model options is available.
- However, choosing the most suitable model is not always obvious.
- Avoid fitting models just because they are ‘typically prescribed.’
- My goal:
  Show that the unfolding approach may not render the best fit even when one would expect so (e.g., Harvey, 2016).

Two types of 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.’

Two types of IRT models



Two types of IRT models

When should we use dominant/unfolding models?

Rule of thumb

Use unfolding when assessing attitudes or preferences.
(e.g., Carter et al. 2014; Drasgow et al., 2010; Stark et al., 2006).

Some results do seem to support this rule.
(e.g., Carter & Dalal, 2010, Carter et al., 2014; Ling et al., 2016; Tay et al., 2009)

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; Zampetakis et al., 2015)

Purpose of this study

I fit both types of model to several datasets where unfolding would be conceivable. Then I compared model fit.

Contributions to the current state of affairs:

- Use datasets with large samples.
- Use original polytomous scores (no dichotomization, unlike many available studies).

  (Cao et al., 2015; Chernyshenko et al., 2007; Stark et al., 2006; Tay et al., 2009)
- Fit the GGUM 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.

Here I focus on clinical data.

Data sets

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.

Data sets

HRS-D$_{17}$		QIDS-C$_{16}$		
Item	# AnsCat	Item	# AnsCat	
Initial insomnia	3	Sleep onset insomnia	4	
Middle insomnia	3	Mid-nocturnal insomnia	4	
Delayed insomnia	3	Early morning insomnia	4	
		~~Hypersomnia~~	~~4~~	
Depressed mood	$5 \longrightarrow 4$	Mood (sad)	$4 \longrightarrow 3$	
Psychic anxiety	5			
~~Loss of insight~~	~~3~~			
Appetite	3	Appetite (decreased)	4	OR
		Appetite (increased)	4	
Weight loss	3	Weight (decreased)	4	OR
		Weight (increased)	4	
Somatic anxiety	5			
Hypochondriasis	$5 \longrightarrow 4$			
		Concentration	4	
Guilt feelings, delusions	$5 \longrightarrow 4$	Outlook (self)	4	
Suicide	$5 \longrightarrow 4$	Suicidal ideation	$4 \longrightarrow 3$	
Work and interests	5	Involvement	$4 \longrightarrow 3^*$	
Somatic energy	3	Energy/fatigability	$4 \longrightarrow 3^*$	
Retardation	4	Psychomotor slowing	$4 \longrightarrow 3$	
Agitation	$5 \longrightarrow 4$	Psychomotor agitation	$4 \longrightarrow 3$	
Libido	3			

$N = 4,039$.

$N_{\text{dec}} = 2,533$.
$N_{\text{inc}} = 921$.
$N_{\text{red}} = 3,916$.
* Only for QIDS-C (inc).

Method

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)

General results – Model fit

Log-likelihood, information criteria

Data set	GRM			GGUM		
	LogLik	df	AIC	LogLik	df	AIC
HRS-D_{17}	−69389.5	60	138899.1	−69392.1	76	138936.2
QIDS-C_{16} (dec)	−36730.0	48	73556.0	−36767.0	61	73656.0
QIDS-C_{16} (inc)	−13172.9	46	26437.8	−13176.7	59	26471.3
QIDS-C_{16} (red)	−47219.1	40	94518.3	−47274.3	51	94650.6

GRM did better in all cases.

General results – Model fit

Adjusted χ^2/df

HRS-D	GRM					GGUM				
	< 1	1to2	2to3	> 3	M (SD)	< 1	1to2	2to3	> 3	M (SD)
Singles	16	0	0	0	0.35 (0.12)	16	0	0	0	0.27 (0.03)
Doubles	2	9	3	4	2.25 (1.51)	2	9	2	5	2.24 (1.47)
Triples	0	4	4	0	2.02 (0.52)	0	4	4	0	2.05 (0.51)

QIDS-C (dec)	GRM					GGUM				
	< 1	1to2	2to3	> 3	M (SD)	< 1	1to2	2to3	> 3	M (SD)
Singles	13	0	0	0	0.01 (0.02)	13	0	0	0	0.00 (0.00)
Doubles	3	1	8	3	2.38 (1.37)	3	2	5	5	2.45 (1.48)
Triples	0	2	4	1	2.31 (0.68)	0	2	4	1	2.35 (0.67)

Note. Similar results for QIDS-C (inc) and QIDS-C (red).

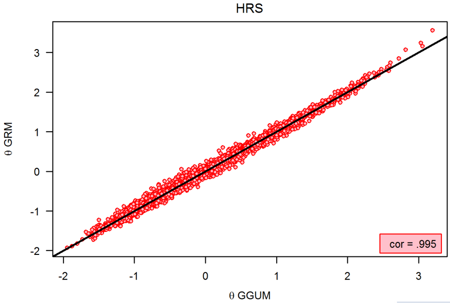
GRM and GGUM displayed similar adjusted χ^2/df fit statistics, with slight advantage for the GRM.

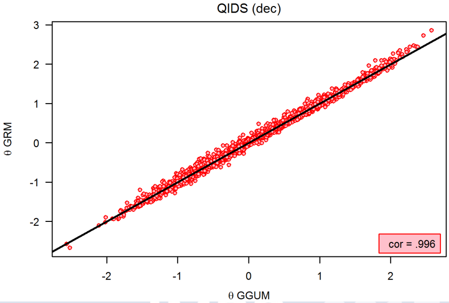
General results – Parameters

- Discrimination parameters highly correlated:

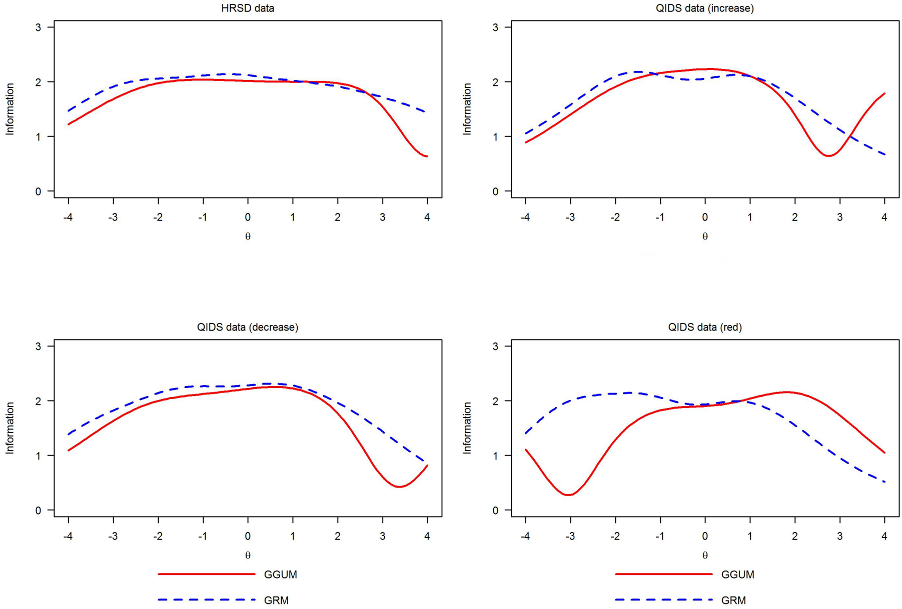
	HRS-D	QIDS-C (dec)	QIDS-C (inc)	QIDS-C (red)
r	.982	.928	.981	.981

- Only one neutral item (i.e., with $|\text{location}| < 2$; Ling et al. 2016) across all scales. Thus, unfolding not identified.
- Theta parameters highly congruent ($r > .99$).
  $\longrightarrow$ If estimating the latent standing of persons is the main goal, then both models seem highly agreeable.





General results – Test information function



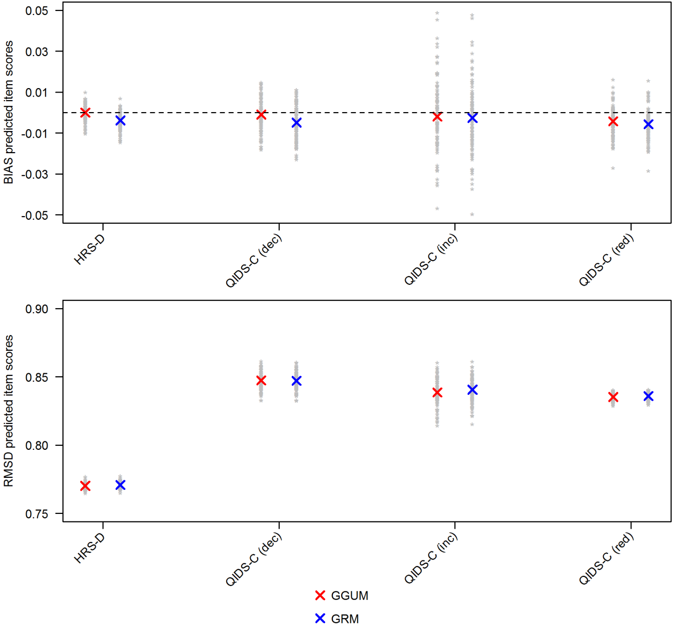
HRSD data

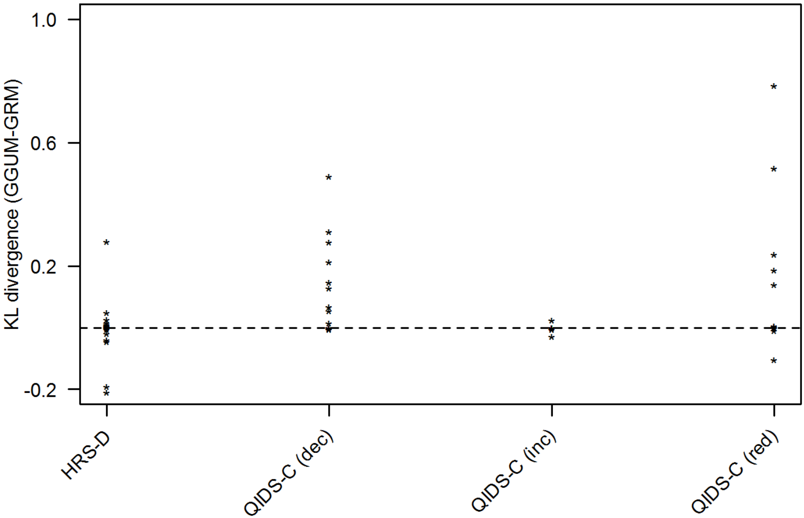
QIDS data (increase)

QIDS data (decrease)

QIDS data (red)

General results – Prediction bias, RMSD




KL divergence (GGUM-GRM)
1.0
0.6
0.2
-0.2
HRS-D
QIDS-C (dec)
QIDS-C (inc)
QIDS-C (red)

General results

- GGUM hardly ever fit better.
- Test information typically favored GRM over broader ranges of θ.
- Predicted distributions either favored the GRM or were very similar between GGUM and GRM.
- Few neutral items identified.
  This might indicate bias towards the GRM due to scale construction.

These results were replicated with other data sets

(Cattell's 16 PF, Big Five, questionnaire on teacher interaction, questionnaire on new digital system).

Discussion

- Including (sharper) ideal point items during scale construction is crucial for models such as the GGUM to shine (Cao et al., 2015; Dalal et al., 2014; Drasgow et al., 2010; Huang & Mead, 2014; Weekers & Meijer, 2008). This is unfortunately rare in practice (yet, see Chernyshenko et al., 2007)
- Without ideal point items, the GRM seems to be preferable (less parameters, better fit, simpler to interpret (?)).
- Actually, including neutral items may lead to convergence failure of dominant model estimation algorithms (Tay et al., 2011).
- 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.

Thank you.

j.n.tendeiro@rug.nl