Adjusting reviewer scores for a fairer assessment via Multi-Faceted Rasch Modelling

Dr Caterina Constantinescu

Measurement context

Data scientist @ The Data Lab

caterina.constatinescu@thedatalab.com

http://datapowered.io/

CaterinaC

- EdinbR organiser: @edinb_r EdinburghRusers
- As part of DataFest (2-week festival of data innovation, tens of events across Scotland)
- New DS conference **DataTech**, 14 March 2019: https://www.datafest.global/data-tech
- DataTech keynotes (3): Mine Çetinkaya-Rundel, Jared Lander, Debbie Bard
- 19 additional speakers (talks, lightning talks, poster pitches)

Aims & data

- 44 submissions
- 6 (blind) reviewers
- 7 items (with sub-indicators) judged against a written talk proposal: *Fit*, *Quality*, *Impact*, *Coverage*, *Novelty balance*, *Reproducibility*, *Clarity*
- 5-point Likert scales (4 steps: $0 \rightarrow 1$ | $1 \rightarrow 2$ | $2 \rightarrow 3$ | $3 \rightarrow 4$)
- Each submission marked by at least 2 markers - enough to create a 'linked' design (no need for a *complete* matrix)
- How to make sure rater bias is kept to a minimum?

Aims & data

- 44 submissions
- 6 (blind) reviewers
- 7 items (with sub-indicators) judged against a written talk proposal: *Fit*, *Quality*, *Impact*, *Coverage*, *Novelty balance*, *Reproducibility*, *Clarity*
- 5-point Likert scales (4 steps: $0 \rightarrow 1$ | $1 \rightarrow 2$ | $2 \rightarrow 3$ | $3 \rightarrow 4$)
- Each submission marked by at least 2 markers - enough to create a 'linked' design (no need for a *complete* matrix)
- How to make sure rater bias is kept to a minimum?

Rasch family of models

- Logistic test model - For dichotomous items
- Rating scale model - For items with the response choices throughout
- Partial credit model - Each item is allowed its own rating scale structure
- **Multi-Facet Rasch models** (*)

Basic idea:

Test performance = Person Ability - Item Difficulty - Rater Severity

Application to current data

```
library( TAM )
submissions <- na.exclude( submissions )

facets <- data.frame( rater = submissions$Name ) # Facet = Rater ID
pid <- as.vector( submissions$Title ) # Submission ID
resp <- submissions[ , item_names ] # Reviewer scores

formula <- ~ item*step+rater*step

M3 <- TAM::tam.mml.mfr( resp = resp,
                        facets = facets,
                        formulaA = formula,
                        pid = pid,
                        control = list( maxiter = 10000 ) )
```

Model	Formula	Params	Deviance	LogLik	AIC	BIC
1	item*step	29	1554.42	-777.21	1612	1663
2	item*step+rater	34	1486.67	-743.33	1555	1615
3	**item*step+rater*step**	49	**1431.85**	-715.93	**1530**	**1616**
4	item*step+rater*item	64	1441.65	-720.82	1570	1682
5	item*step+rater*item+rater*step	79	1395.46	-697.73	1553	1693
6	item*rater*step	169	1293.90	-646.95	1632	1930

Application to current data

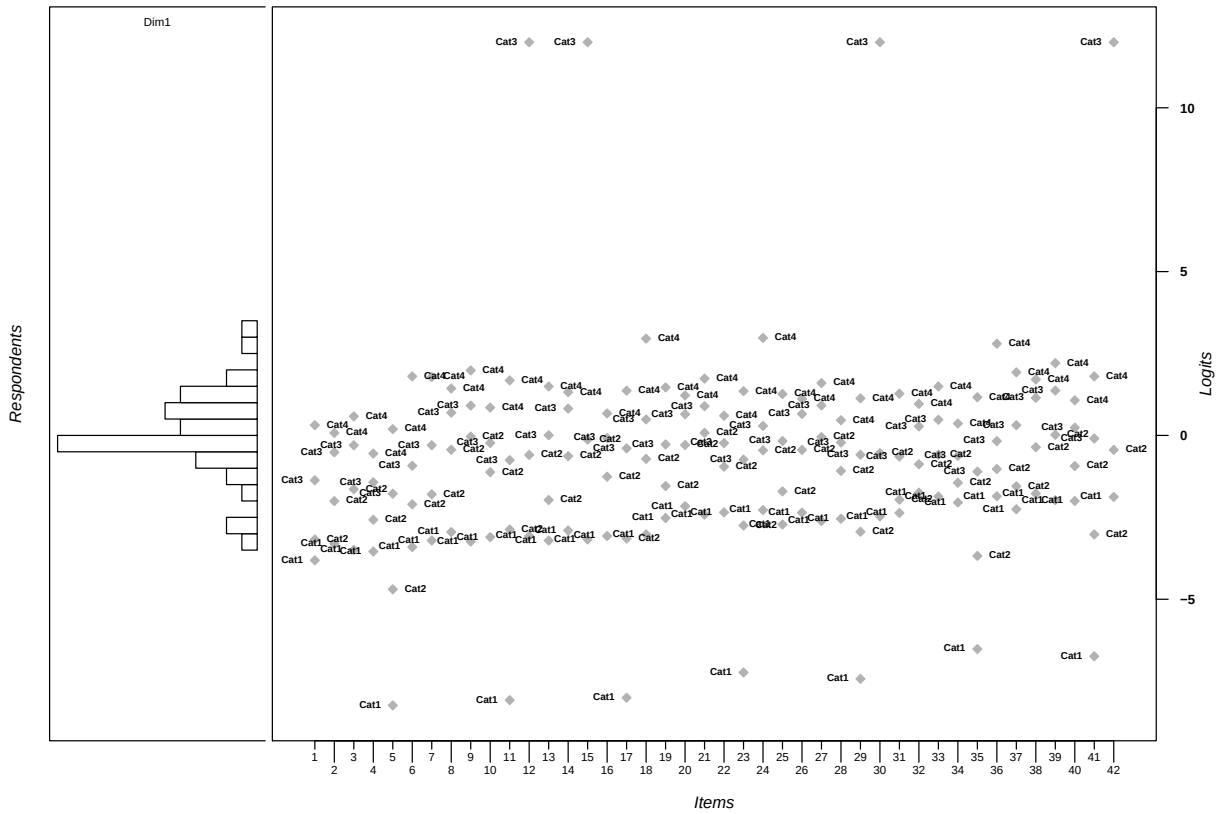
```
library( TAM )
submissions <- na.exclude( submissions )

facets <- data.frame( rater = submissions$Name ) # Facet = Rater ID
pid <- as.vector( submissions$Title ) # Submission ID
resp <- submissions[ , item_names ] # Reviewer scores

formula <- ~ item*step+rater*step

M3 <- TAM::tam.mml.mfr( resp = resp,
                        facets = facets,
                        formulaA = formula,
                        pid = pid,
                        control = list( maxiter = 10000 ) )
```

Model	Formula	Params	Deviance	LogLik	AIC	BIC
1	item*step	29	1554.42	-777.21	1612	1663
2	item*step+rater	34	1486.67	-743.33	1555	1615
3	**item*step+rater*step**	49	**1431.85**	-715.93	**1530**	**1616**
4	item*step+rater*item	64	1441.65	-720.82	1570	1682
5	item*step+rater*item+rater*step	79	1395.46	-697.73	1553	1693
6	item*rater*step	169	1293.90	-646.95	1632	1930

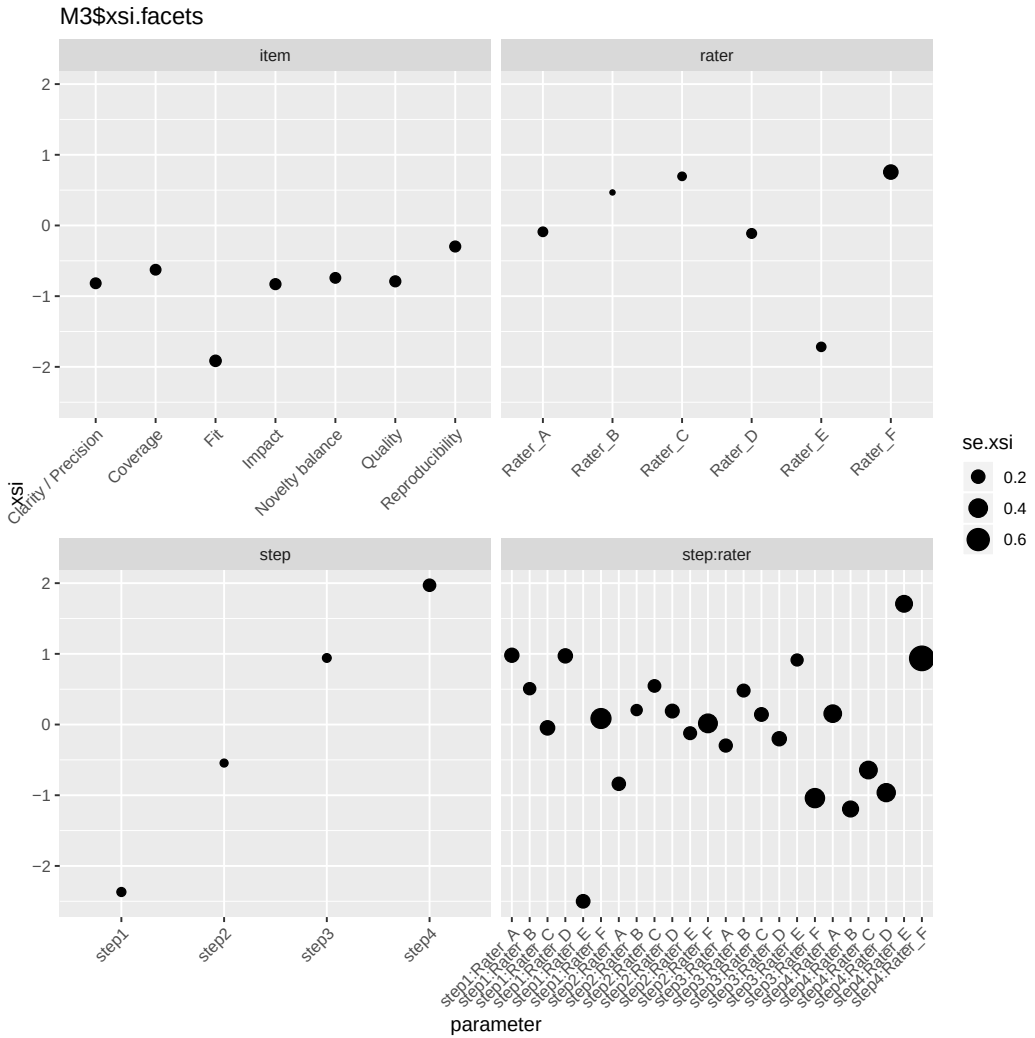
Application to current data

```
1   tam.threshold( M3 ) # Or steps
2
3   TAM::IRT.WrightMap( M3, show.thr.lab = TRUE, label.items = c(1:42),
4                       label.items.rows = 2, type = "WLE",
5                       main.title = "Wright Map: Thurstone item thresholds for 7 items x 6 raters" )
```

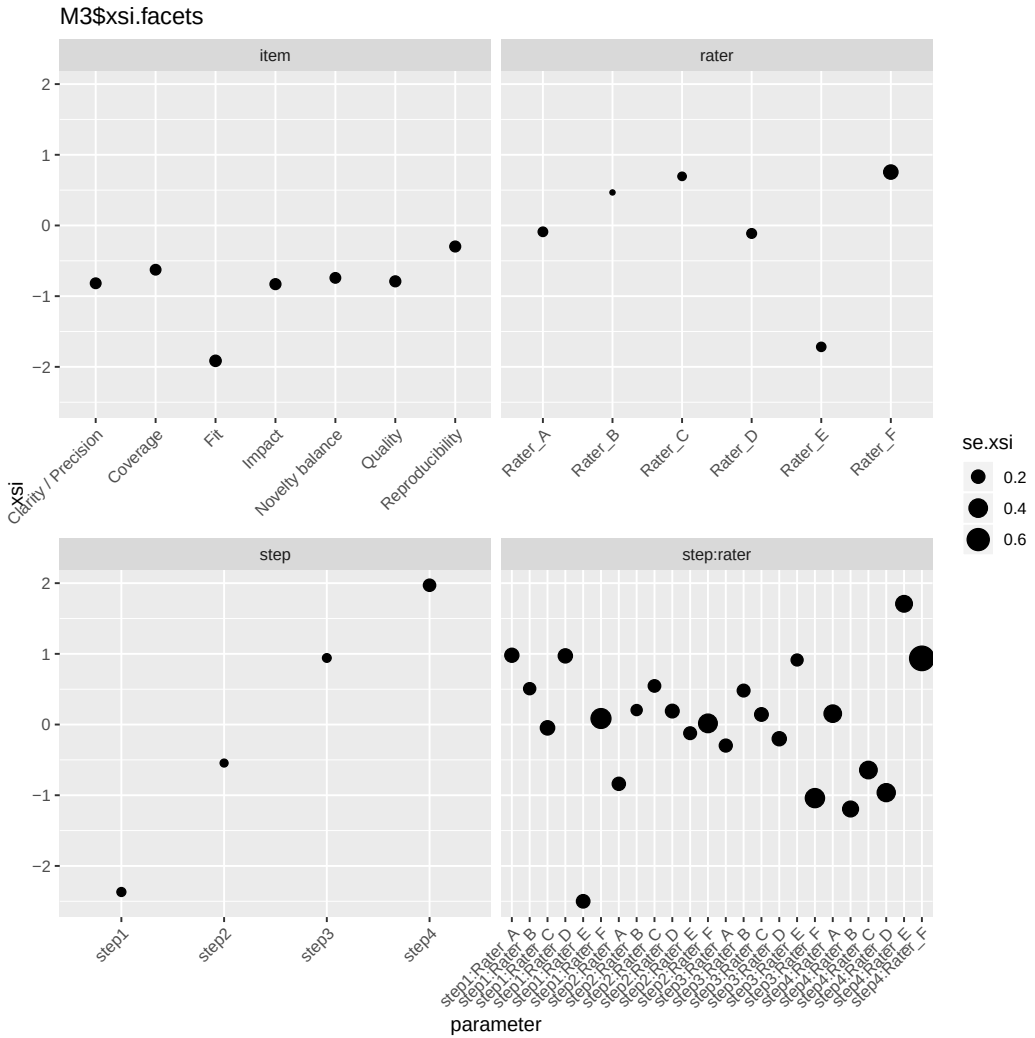
Wright Map: Thurstone item thresholds for 7 items x 6 raters



Application to current data



Application to current data



Application to current data

```
msq.itemfit( M3 )
tam.fit( M3 )
```

	item	fitgroup	Outfit	Outfit_t	Outfit_p	Infit	**Infit_t**	Infit_p
1	Fit-Rater_A	1	0.454	-1.051	0.293	0.515	-1.331	0.183
...	...	...	...	...	...	...	...	...
16	Impact-Rater_D	16	0.695	-0.364	0.716	0.591	-0.975	0.330
24	Coverage-Rater_F	24	5.393	5.060	0.0000	2.966	**3.069**	0.002
27	Novelty balance-Rater_C	27	1.182	0.568	0.570	1.164	0.548	0.584
39	Reproducibility-Rater_C	39	2.128	2.378	0.017	2.035	**2.252**	0.024
...	...	...	...	...	...	...	...	...

	parameter	Outfit	Outfit_t	Outfit_p	Outfit_pholm	Infit	**Infit_t**	Infit_p	Infit_pholm
1	Fit	0.876	-0.825	0.409	1	0.997	-0.009	0.993	1
2	Quality	0.679	-2.433	0.015	0.524	0.692	-2.325	0.020	0.684
3	Impact	0.726	-2.035	0.042	1	0.703	-2.230	0.026	0.850
4	Coverage	1.255	1.521	0.128	1	1.005	0.055	0.956	1
5	Novelty balance	0.949	-0.326	0.744	1	0.914	-0.576	0.565	1
6	Clarity / Precision	0.796	-1.391	0.164	1	0.748	-1.743	0.081	1
7	Reproducibility	1.209	1.305	0.192	1	1.254	1.563	0.118	1
...	...	...	...	...	...	...	...	...	...
12	Rater_A	2.059	6.881	0	0	2.449	8.825	0	0
13	Rater_B	1.604	5.609	0.00000	0.00000	1.729	6.606	0	0
14	Rater_C	1.956	6.721	0	0	2.030	7.186	0	0
15	Rater_D	1.811	4.950	0.00000	0.00004	1.883	5.307	0.00000	0.00001
16	Rater_E	2.525	10.644	0	0	2.399	9.958	0	0
17	Rater_F	1.191	1.358	0.175	1	1.929	5.625	0.00000	0.00000
...	...	...	...	...	...	...	...	...	...

Output

	Title	Format	Difficulty	Interactivity	Score	WLE	Error
1	A	Talk	Intermediate	Lecture-style (non-participatory)	80	3.271437676	0.5570065
2	B	Talk	Intermediate	Lecture-style (non-participatory)	82	2.888829500	0.6297298
3	C	Talk	Intermediate	Lecture-style (non-participatory)	48	1.653824238	0.3797413
4	D	Talk	Beginner	A few elements of interactivity occurring occasionall...	51	1.590337102	0.4096537
5	E	Lightning	Intermediate	Lecture-style (non-participatory)	47	1.316891415	0.3533718
6	F	Talk	Intermediate	Frequent feedback/participation requested from the ...	44	1.161409937	0.3407933
7	G	Talk	Beginner	Lecture-style (non-participatory)	70	1.085191235	0.2880148
8	H	Talk	Intermediate	Lecture-style (non-participatory)	43	1.050874337	0.3359556
9	I	Talk	Intermediate	Lecture-style (non-participatory)	45	1.024109801	0.3384065
10	J	Poster	Beginner	*NA*	44	0.916059290	0.3322050
11	K	Talk	Intermediate	Lecture-style (non-participatory)	39	0.784570693	0.3151142
12	L	Talk	Intermediate	Lecture-style (non-participatory)	42	0.708780464	0.3243062
13	M	Talk	Intermediate	Lecture-style (non-participatory)	38	0.609006127	0.3187580
14	N	Talk	Intermediate	Lecture-style (non-participatory)	40	0.593080607	0.3125103
15	O	Talk	Intermediate	A few elements of interactivity occurring occasionall...	39	0.517362579	0.3022782
16	P	Talk	Intermediate	Lecture-style (non-participatory)	37	0.413357465	0.3306307
17	Q	Talk	Intermediate	A few elements of interactivity occurring occasionall...	43	0.297549880	0.3394347
18	R	Lightning	Beginner	A few elements of interactivity occurring occasionall...	29	0.094852165	0.2914775
19	S	Talk	Beginner	Lecture-style (non-participatory)	35	0.017256455	0.2840717
20	T	Lightning	Intermediate	Lecture-style (non-participatory)	28	0.009785025	0.2949409
21	U	Talk	Intermediate	A few elements of interactivity occurring occasionall...	35	-0.003418413	0.3266766
22	V	Talk	Intermediate	Lecture-style (non-participatory)	34	-0.061238604	0.3320397
23	W	Lightning	Beginner	A few elements of interactivity occurring occasionall...	35	-0.095560531	0.3468514
24	X	Talk	Intermediate	Lecture-style (non-participatory)	33	-0.141372312	0.2832830

Rasch model assumptions & caveats

Key assumptions:

- Unidimensionality
- Local independence*

Caveats:

- Gender balance?
- Academia, industry, public sector balance?
- Items are given equal weights

Any questions?

Any questions?