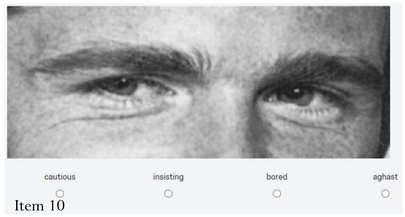
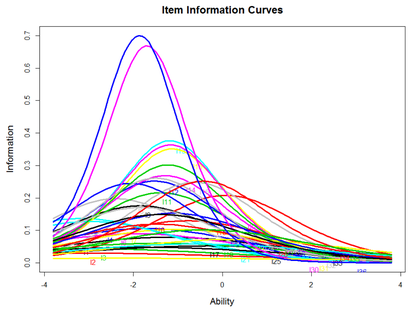
Item response theory analyses of the Reading the Mind in the Eyes Test

Jessica E. Black
University of Oklahoma



Item 10



Introduction

The Reading the Mind in the Eyes test (RMET; Baron-Cohen, Wheelwright, Hill, Raste, & Plumb, 2001) is used as a measure of theory of mind (Black & Barnes, 2015a, 2015b; Kidd & Castano, 2013; Panero et al., 2016), mentalizing (Kidd & Castano, 2016), empathy (Djikic, Oatley, & Moldoveanu, 2013), and/or perspective-taking (Bischoff and Peskin,2014), both in experimental and correlational designs, most recently related to the effects of fiction and nonfiction narratives. Some researchers categorize the items by valence (e.g., Chamorro-Premuzic & Ali, 2010) but most treat it as a unidimensional construct. To the best of our knowledge, only one study has been done to investigate the item properties: Preti, Vellante, and Petretto (2017) applied Item Response Theory (IRT) to an Italian sample (N = 200) and report evidence for a unidimensional structure, with a Rasch model as best fit.
The purpose of this research was to utilize IRT to analyze RMET item properties in an English speaking sample in order to (a) determine how many, if any, items could be dropped for a shorter instrument for use in future research, and (b) to estimate item parameters for use in rescoring and testing outcomes from past experiments.

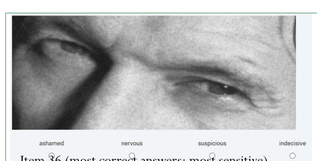
Method

Participants

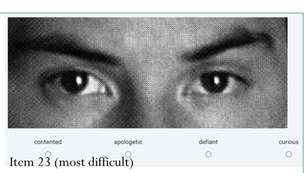
- Three labs contributed a total of 365 cases to this project
  - All data were from control groups (no manipulations)
  - N = 138 were OU undergraduates who completed an online survey in exchange for credit.
  - N = 31 were undergraduates recruited from the departmental pool at an East Coast university
  - N = 190 participated via Amazon.com's Mechanical Turk

Data Analyses

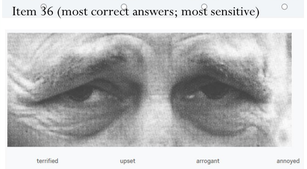
- IRTPRO and RStudio (ltm package) were used to estimate item parameters. 2PL analyses were run in both; parameters matched. R was used to test a Rasch model and obtain AIC/SBC statistics for model comparison. The full 3PL model did not converge without constraining guessing parameters. The 26- and 22-item tests converged without imposing contraints.
- Items were dropped successively based on item information curves, discrimination parameters, and fit.



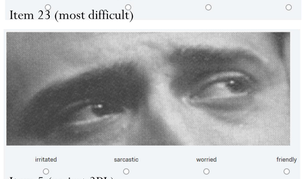
Item 36 (most correct answers; most sensitive)



Item 23 (most difficult)



Item 2 (second worst item; easiest 2PL)



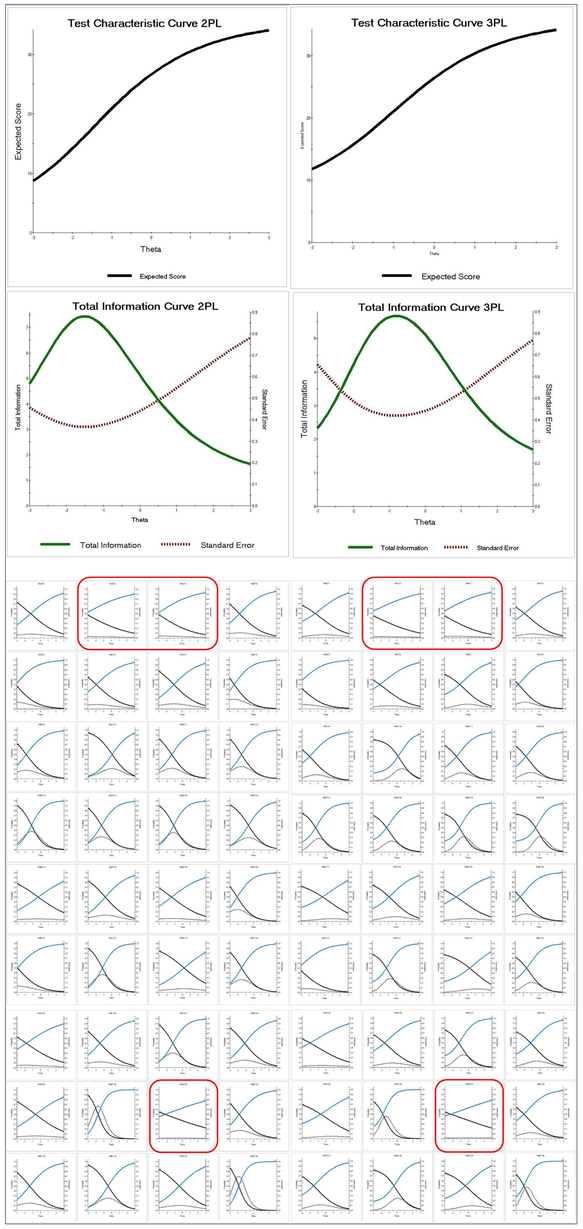
Item 5 (easiest 3PL)

Results

Table 1.
Item statistics from classic reliability analysis and two and three parameter IRT analyses.

					2PL				3PL		item fit	
Item	*P* value	*SD*	corrected item-total correlation	alpha if item deleted	α	*SE*	β	*SE*	α	β	χ^2	*p*
1	.66	.48	.22	.801	0.56	0.14	-1.24	0.34	0.61	-0.57	16.52	.685
2	.76	.43	.13	.804	0.35	0.14	-3.33	1.30	0.37	-2.32	18.97	.525
3	.78	.41	.17	.802	0.42	0.14	-3.18	1.07	0.43	-2.43	20.58	.362
4	.74	.44	.26	.800	0.65	0.15	-1.72	0.38	0.68	-1.19	19.59	.485
5	.89	.31	.23	.800	0.74	0.19	-3.15	0.71	0.71	-2.92	17.57	.352
6	.75	.44	.24	.800	0.60	0.15	-1.92	0.46	0.61	-1.41	28.57	.096
7	.65	.48	.24	.800	0.64	0.14	-1.06	0.27	0.75	-0.40	16.67	.613
8	.87	.34	.30	.798	0.89	0.19	-2.40	0.43	0.89	-2.11	17.85	.400
9	.80	.40	.31	.798	0.84	0.17	-1.85	0.33	0.85	-1.49	16.35	.635
10	.49	.50	.32	.797	0.91	0.17	0.08	0.13	1.15	0.46	13.60	.756
11	.75	.43	.35	.796	0.93	0.17	-1.38	0.23	0.96	-1.05	19.53	.362
12	.78	.41	.37	.796	1.00	0.18	-1.52	0.24	0.96	-1.35	33.88	**.027**
13	.76	.43	.43	.793	1.23	0.20	-1.18	0.17	1.20	-1.01	27.99	.062
14	.75	.43	.36	.796	1.04	0.18	-1.29	0.21	1.14	-0.90	37.48	**.007**
15	.75	.44	.40	.794	1.19	0.19	-1.13	0.17	1.36	-0.74	16.97	.526
16	.63	.48	.35	.796	1.00	0.17	-0.63	0.15	1.39	-0.08	23.30	.179
17	.53	.50	.18	.803	0.46	0.13	-0.31	0.25	0.55	0.57	28.16	.080
18	.63	.48	.28	.799	0.72	0.15	-0.80	0.21	0.73	-0.39	21.91	.288
19	.66	.48	.20	.802	0.49	0.13	-1.39	0.41	0.57	-0.52	22.04	.399
20	.85	.36	.36	.796	0.99	0.19	-2.06	0.33	0.92	-1.92	12.39	.869
21	.87	.33	.25	.800	0.72	0.18	-2.97	0.66	0.69	-2.71	23.96	.120
22	.76	.43	.42	.794	1.21	0.20	-1.18	0.17	1.25	-0.92	15.18	.650
23	.47	.50	.21	.802	0.54	0.14	0.21	0.21	0.64	0.87	19.59	.358
24	.78	.41	.36	.796	1.02	0.18	-1.50	0.23	1.10	-1.13	13.30	.774
25	.68	.47	.20	.802	0.44	0.13	-1.81	0.57	0.47	-0.98	24.43	.223
26	.72	.45	.28	.798	0.69	0.15	-1.46	0.32	0.71	-0.99	24.85	.207
27	.75	.43	.37	.795	1.10	0.18	-1.22	0.19	1.21	-0.86	18.05	.454
28	.69	.46	.31	.798	0.78	0.15	-1.13	0.24	0.85	-0.65	12.82	.848
29	.57	.50	.21	.802	0.48	0.13	-0.61	0.27	0.52	0.15	31.70	**.047**
30	.88	.33	.45	.794	1.64	0.27	-1.70	0.19	1.59	-1.58	19.75	.287
31	.63	.48	.12	.805	0.24	0.12	-2.19	1.19	0.29	-0.52	19.00	.523
32	.79	.41	.30	.798	0.83	0.17	-1.83	0.33	0.83	-1.50	27.14	.076
33	.72	.45	.30	.798	0.77	0.15	-1.39	0.28	0.80	-0.96	24.39	.225
34	.59	.49	.35	.796	1.00	0.17	-0.43	0.13	1.15	-0.06	19.12	.451
35	.56	.50	.24	.800	0.64	0.14	-0.40	0.19	0.72	0.18	21.27	.383
36	.90	.30	.45	.795	1.67	0.28	-1.86	0.2	1.63	-1.74	21.51	.089

<u>26 items</u>: 4, 5, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 18, 20, 22, 23, 24, 26, 27, 28, 30, 32, 33, 34, 35, 36
<u>22 items</u>: 4, 5, 7, 8, 9, 10, 11, 15, 16, 18, 20, 23, 24, 26, 27, 28, 30, 32, 33, 34, 35, 36



- 26 item version: 2 item fit statistics were significant, r_α = .80
- 22 item version: no fit problems, r_α = .76
- 2PL a significantly better fit than Rasch for all models, *p*s < .001.
- No difference between 2PL and 3PL models (full or reduced test), *p*s > .750; 2PL preferred for parsimony

Discussion

Many items were very poor; no discrimination between people of similar ability. Failure to find anything greater than a small effect of manipulation in past research may reflect poor measurement.

The RMET is an easy test; only two items had difficulty statistics above the mean.

- Sample of neurotypical adults… people are good at being social primates.
- Need for more challenging measure if we want to discriminate between high levels of theory of mind ability.

Limitations & future research

The 36-item test is probably not unidimensional. The single factor model was not a good fit for the full test; possibly I just eliminated items that should have been part of a second factor. A simple LRT in R suggested that a two-factor model would fit the data better (p < .001). Future research should include exploratory factor analyses.

We still don't know if this is primarily a vocabulary test.

Need to test for differential item functioning across different groups (e.g., gender, sample source).

Models tested.

Model	items	α	M_2	*df*	*p*	RMSEA	Log likelihood	AIC	SBC
Rasch	36	.803					-7050	14174	14318
2PL			822	594	< .001	.03	-6988	14120	14400
3PL			741	558	< .001	.03	-6996	14208	14629
2PL	26	.800	317	299	.229	.01	-4871	9846	10048
3PL			271	273	.516 <	.01	-4878	9912	10216
2PL	22	.758	243	209	.055	.02	-4170	8428	8600
3PL			205	187	.170	.02	-4175	8482	8740

Model fit statistics for single-factor solution, full and reduced RMET.

	Par	χ^2	*df*	Fit function	SRMSR	RMSEA	PCLOSE	TLI	AIC	SBC
36 items	107	1266	595	3.48	0.384	.056[.051, .060]	.015	0.430	1480	1897
26 item	78	376	299	1.03	**0.045**	**.027[.017, .035]**	**1.000**	**0.918**	532	836
22 item	66	285	209	**0.78**	0.046	.032[.022, .040]	.999	0.876	**417**	**674**

Note. N=365; *Par*: parameters in model. SRMSR: standardized root mean square residual. RMSEA: Root Mean Square Error of Approximation. PCLOSE: one-sided probability of close fit (RMSEA=.05). TLI: Tucker-Lewis Index. AIC: Akaike's Information Criterion. SBC: Schwarz's Bayesian information criterion.