

Rater stability and applicant pool quality across
successive applicant pools: A many-faceted Rasch
rating scale analysis

Peter D. MacMillan, Ph.D.
Colin A. Chasteauneuf, Ph.D.
University of Northern British Columbia
3333 University Way
Prince George, BC V2N 4Z9
(250) 960 5828 (tel.)
(250) 960 5536 (fax)
peterm@unbc.ca
chasteac@unbc.ca

Presented at the 2006
13th International Objective Measurement Workshop
Wednesday April 7, 2006
Symposium 2 – 2:15 p.m.
International House
University of California, Berkeley
Berkeley CA

- Within one small North American university a pool of applicants are judged using a series of common scales by raters from a pool of raters.
- Two raters pseudo-randomly selected from a pool of raters rate the applicants in the bundles of files that the raters are given.
- The applicants were placed within bundles by a pseudo-random process.
- All applicants are rated on the same set of items in a given year.

What's different about this rating system?

- Raters are faculty members, i.e., volunteer labor.
- Raters are relatively untrained and not subject to recruitment, removal, or retraining initiatives.
- Raters may/may not be involved in a given year.
- Raters score a relatively small number of files.
- Some items may change from year to year.

Why is this study of interest?

Table 1 *Scale Descriptions and their Abbreviations*

Abbreviation	Description
Content04A	From Personal Statement –originally a 10 point scale introduced in 2004
Content04B	that was to have assessed the content as it related to experiences related to teaching. This scale was dropped the following year.
Talent	From Personal Statement – <u>talents</u> related working with children
Skills	From Personal Statement – <u>skills</u> related working with children
Grammar	From Personal Statement – <u>grammar</u> of the written submission
Organization	From Personal Statement – <u>organization</u> of the written submission
ECY quantity	Experience with Children & Youth – <u>Quantity</u> of experience
ECY quality	Experience with Children & Youth – <u>Quality</u> of the experience
LR(1,2,3)ref	The appropriateness of the <u>referee</u> for each of the three letters (3 scales)
LR(1,2,3)cont	The appropriateness of the <u>content</u> for each of the three letters (3 scales)

- 2003 Skills and Talents were two separate 5-point scales.
- 2004 Skills/Talents; treated as Talents scale. Grammar/Organization – Organization.
- 2005 Skills & Talents a 10 point scale; split 2 5-point scales: Skills, Talents

The analysis employed the many-faceted version (Linacre, 1989) of the Rasch (1960 / 1992) model and the computer program FACETS (Linacre, 2004).

Specifically, the model used for this analysis of the data set is described as

$$\ln(P_{nmaik}/P_{nmai(k-1)}) = B_n - F_m - Y_a - D_i - C_k$$

where

P_{nmaik} = probability of applicant n obtaining score of k on item i from rater m in year a

B_n = quality of the file of applicant, $n = 1, 2, 3, \dots$

F_m = severity of the rater, $m = 1, 2, \dots$

Y_a = year of the rating, $a = 2003, 2004, 2005$

D_i = difficulty of item, $i = 1, 2, \dots 9$

C_k = category of the rating scale, $k = 1, 2, 3, 4, 5$

The “year” facet was only used to provide separate annual analysis and anchoring.

ALL THREE YEARS Table 8.1 Category Statistics.

Model = ?, ?, , , ?, R

DATA				QUALITY CONTROL			STEP		EXPECTATION		MOST	.5 Cumul.	Cat
Category	Counts	Cum.		Avge	Exp	Oufit	CALIBRATIONS		Measure at	PROBABLE	Probabil.	PEAK	
Score	Used	%	%	Meas	Meas	MnSq	Measure	S. E.	Category	-0.5	from	at	Prob
1	82	1%	1%	.11	-.14	1.3			(-2.74)		low	low	100%
2	318	3%	4%	.47	.32	1.2	-1.28	.12	-1.35	-2.09	-1.28	-1.74	33%
3	1760	19%	23%	.82	.89	.9	-1.12	.06	-.18	-.77	-1.12	-.86	47%
4	3768	40%	63%	1.53	1.53	1.0	.45	.03	1.29	.49	.45	.45	50%
5	3480	37%	100%	2.22	2.20	1.0	1.95	.02	(3.17)	2.33	1.95	2.11	100%
									(Mean)			(Modal)	(Median)

Figure 1 Part A. Scale structure for the five-point rating scales.

Probability Curves DATA FROM ALL THREE YEARS

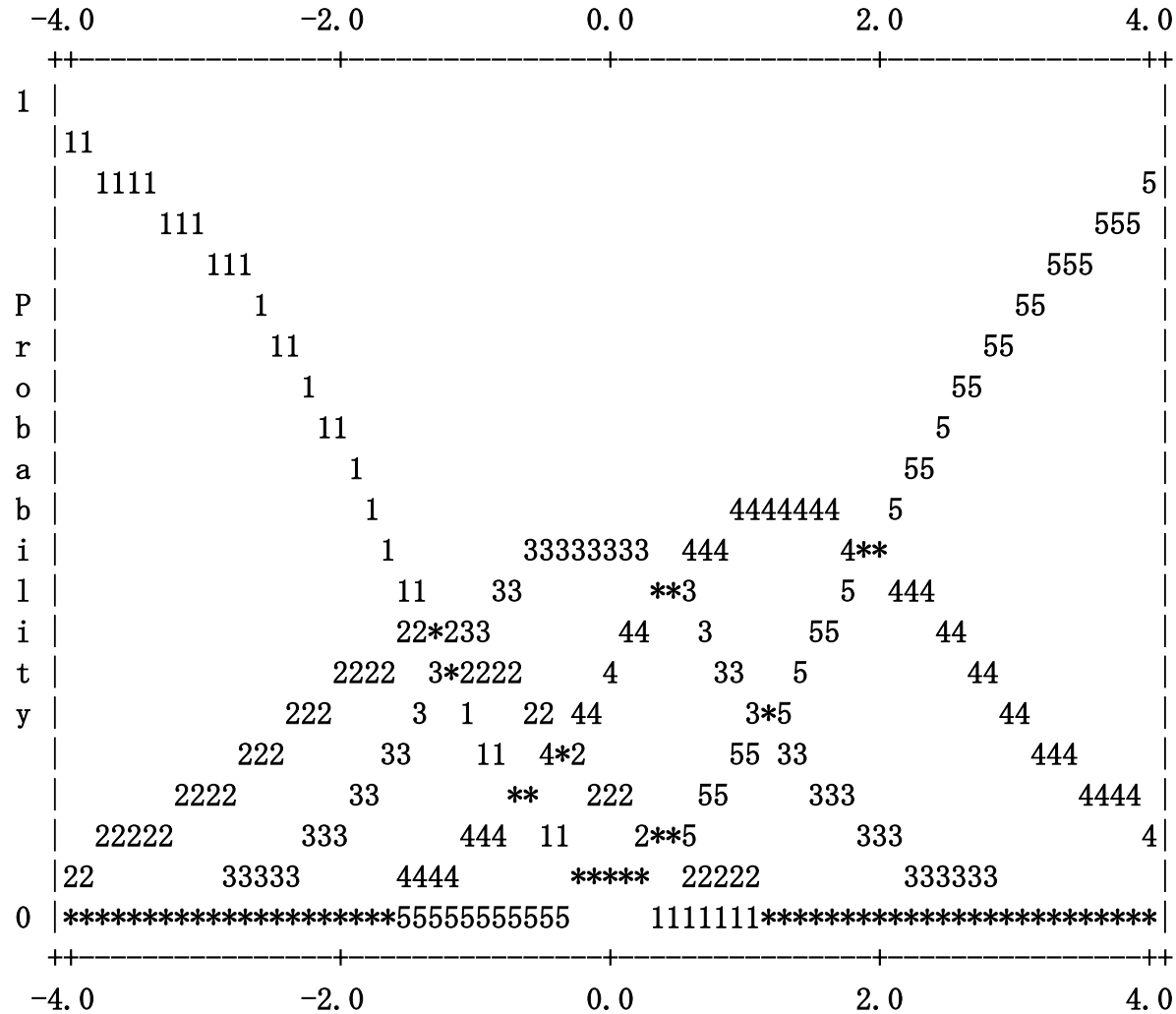


Figure 1 Part B. Scale structure for the five-point rating scales.

Table 2 *Stability of Items Across Years*

	All Three Years			2003			2004			2005		
Scale	Fair	Infit	Zstd	Fair	Infit	Zstd	Fair	Infit	Zstd	Fair	Infit	Zstd
Content04A	3.52	0.64	-4.8				3.58	0.67	-4.6			
Content04B	3.54	0.54	-6.4				3.60	0.57	-6.3			
Skill	4.03	0.77	-3.8	4.08	0.88	-1.3				3.97	0.72	-3.6
Talent	4.01	0.78	-4.6	3.94	1.07	0.7	4.14	0.85	-1.8	4.00	0.61	-4.2
Grammar	4.27	1.11	1.6	4.19	1.36	3.5				4.30	0.94	-0.1
Organization	4.24	0.89	-2.1	4.29	1.13	1.3	4.19	0.95	-0.6	4.30	0.71	-2.5
ECYQuality	4.26	0.78	-4.5	4.32	0.75	-2.8	4.36	0.85	-1.7	4.19	0.72	-3.0
ECYQuantity	4.23	0.98	-0.3	4.28	0.93	-0.7	4.38	1.13	1.4	4.13	0.86	-0.6
LR1Referee	4.30	1.30	5.1	4.21	1.07	0.7	4.44	1.57	5.3	4.27	1.28	1.5
LR1Content	4.03	0.94	-1.1	4.17	0.76	-2.8	4.05	1.06	0.7	3.97	0.89	-1.2
LR2Referee	4.24	1.28	4.8	4.22	0.97	-0.2	4.39	1.34	3.4	4.17	1.39	2.8
LR2Content	4.03	0.97	-0.6	4.10	0.78	-2.6	4.08	0.92	-1.0	3.98	1.06	0.6
LR3Referee	4.10	1.62	9.0	4.25	1.45	4.3	4.16	1.57	5.8	3.99	1.68	5.7
LR3Content	3.95	1.03	0.5	4.10	0.86	-1.6	3.93	1.02	0.2	3.91	1.07	0.9
N	392			110			132			150		
Separation	7.28			1.93			5.56			3.03		
Reliability	0.98			0.79			0.97			0.90		

Shaded items were anchored at their mean difficulties. Content04 items were dropped.

Table 3 Stability of Raters Across Years

	All Three Years			2003			2004			2005		
Rater	Fair	Infit	Zstd	Fair	Infit	Zstd	Fair	Infit	Zstd	Fair	Infit	Zstd
1	4.00	0.46	-5.8	4.15	0.48	-3.1	3.86	0.59	-3.8			
2	4.25	0.99	-0.4	4.46	0.92	-1.7	4.14	1.01	0.2	4.32	1.25	2.9
3	4.48	1.09	1.7				4.47	1.01	0.1	4.41	1.15	1.6
4	4.16	1.08	1.7				4.00	1.11	1.9	4.26	0.98	-0.2
5	4.12	0.89	-1.8	3.92	1.13	0.8	4.26	1.20	1.6	4.14	0.64	-4.7
6	3.98	1.05	0.7	4.13	1.48	5.9	3.75	0.84	-1.3			
7	4.02	0.67	-5.5	4.23	0.79	-3.7	4.29	0.79	-0.4			
8	3.53	1.62	8.6	4.22	1.78	6.7				3.44	1.37	4.5
9	4.26	0.93	-0.6	4.40	1.05	0.4						
11	3.50	0.45	-9.0	3.49	0.43	-9.0				3.76	0.53	-6.4
12	4.38	1.39	3.0	4.50	1.54	4.4						
14	3.91	1.99	5.7							3.98	1.69	4.1
15	3.76	1.16	1.6							3.85	0.97	-0.2
16	4.07	1.79	5.7							4.14	1.49	3.7
17	3.73	0.72	-3.3							3.79	0.59	-5.3
18	4.38	0.79	-1.8							4.43	0.73	-2.2
19	3.94	1.01	0.1							4.00	0.89	-1.0
20	4.39	1.08	1.0							4.43	0.98	-0.2
N	18			9			7			13		
Separation	5.89			5.20			2.45			5.69		
Reliability	0.97			0.96			0.86			0.97		

Table 4 *Stability of Raters*

	2003		2004		2005	
R	Severity	Fit	Severity	Fit	Severity	Fit
<u>2</u>	Very lenient	Good	Middle	Good	Mid-lenient	Good
<u>3</u>			Most lenient	Good	Most lenient	Good
<u>4</u>			Middle	Good	Middle	Good
<u>5</u>	Severe	Good	Middle	Good	Middle	Very Cramped
8	Middle	Highly erratic			Extremely severe	Highly erratic
<u>11</u>	Extreme severe	Extremely cramped			Severe	Very cramped

Raters 2, 3, 4, 5, and 11 were anchored at their combined mean severities.

Conclusions

- Most items were stable in their characteristics, difficulty and fit statistics, across the three years.
- Five of the six “core” raters demonstrated stability, severity and fit statistics, across years in spite of attempts at training.
- Anchoring procedures allowed the comparison of applicant pools across years in spite of variation in items used, raters who scored files, and variations in rater training procedures.

Discussion

- In agreement with what is known from studies of more structured rating systems, raters continue to maintain their specific traits thus requiring the use of a many-faceted Rasch analysis to produce fair assessment of all applicants.
- The familiarization with the results of these analyses and involvement of key Education faculty and staff members has resulted in a Rasch positive attitude. The net result is that Rasch analysis will become the accepted standard for producing suitability estimates for applicant selection