Person Misfit and Person Reliability in Rating Scale Measures: The Role of Response Styles

Tongtong Zou, Daniel M. Bolt

Quantitative Methods,
Department of Educational Psychology

November 4, 2022

Table of Contents

Table of Contents

Person Fit

- **Person fit** methodology, also known as "**appropriateness measurement**", initially measures the degree of "unusualness" of an examinee's answer patterns (Levine & Drasgow, 1982)
- Commonly, the misfit for an individual test performance in relation to an IRT model, often likelihood based (Meijer & Sijtsma, 2001).

Person Fit: Interpretation & Reasons

- With binary items,non-fitting respondents often endorse more difficult (i.e., infrequently endorsed) items but fails to endorse easier (i.e., frequently endorsed) items;
- In rating scale measurement:
  - Under-fit: careless or effortless responding;
  - Over-fit: constantly selecting the exact same answer category (Curtis, 2004).

Person Fit: Example

- Example A: **Under-fitting** Binary Response Pattern
  Suppose Items 1 - 10 are ordered from easiest to the most difficult:

Item NO.	1	2	3	4	5	6	7	8	9	10
Fitting	1	1	1	0	1	0	0	1	0	0
Under-fitting	0	0	1	0	1	1	1	1	0	1

- Example B: **Over-fitting** in Rating Scale Measurement
  Suppose Items 1 - 10 are on Five point Likert-scale:

Item NO.	1	2	3	4	5	6	7	8	9	10
Fitting	2	3	3	2	2	2	2	1	4	3
Overfitting	4	4	4	4	4	4	4	4	4	4

Person Fit: IRT Model

- **Graded Response Model** (GRM; Samejima, 1969)

$$P(X_{ij} = k; \theta) = \frac{\exp[a_j(\theta - b_{j,k-1})]}{1 + \exp[a_j(\theta - b_{j,k-1})]} - \frac{\exp[a_j(\theta - b_{j,k})]}{1 + \exp[a_j(\theta - b_{j,k})]} \quad (1)$$

Person Fit: A Parametric Approach

Person Fit index l_0 based on ML estimation (Drasgow et al., 1985):

- Dichotomous item response model:

$$l_0 = \sum_{i=1}^{n} u_i \log P_i(\hat{\theta}_d) + (1 - u_i) \log Q_i(\hat{\theta}_d) \tag{2}$$

  u_i: 1, correct, 0, incorrect; $\hat{\theta}_d$: ML estimate of θ;

- Polytomous item response model:

$$l_{0,h} = \sum_{i=1}^{n} \sum_{j=1}^{A+1} \delta_j(v_i) \log P_{ij}(\hat{\theta}_d) \tag{3}$$

  In total $A + 1$ response categories, $\delta_j(v_i) = 1$ when category j is the score on item i, 0 otherwise.

Person Fit: A Parametric Approach

Standardized index l_z (Drasgow et al., 1985):

$$l_{z,h} = [l_{0,h} - E_h(\hat{\theta}_d)]/\sigma_h(\hat{\theta}_d) \tag{4}$$

- Asymptotically follows a standard normal distribution;
- The smaller the Z_h value, the greater the evidence for under-fit;

Person Reliability

- Trait Variability:
  - "Constant - θ" VS "Variable - θ " (Levine & Drasgow, 1983)
- Person variation parameter σ_d (Ferrando, 2009):

$$\Phi(\frac{\theta_d - \beta_{j,k-1}}{\sigma_d}) - \Phi(\frac{\theta_d - \beta_{j,k}}{\sigma_d}) \tag{5}$$

- **Person Reliability Index** γ_d:

$$\gamma_d = 1/\sigma_d \tag{6}$$

- Relation to person fit indices:
  - Strong positive association between l_z and γ_d (Ferrando,2004)

Response Style

- Definition: content-irrelevant stylistic tendencies in the use of rating scale categories, i.e. disproportionately over-/under- selection of categories, controlling for the latent trait.
- For five-point Likert-scale:
  - Extreme response style: high p_1, p_5 values;
  - Mid-point response style: high p_3 values;
  - No response style: uniform p_1, p_2, p_3, p_4, p_5 values.

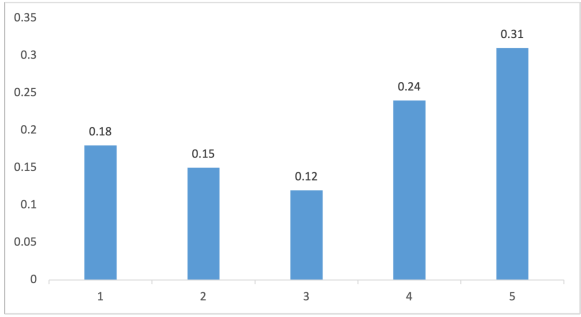
Present Study

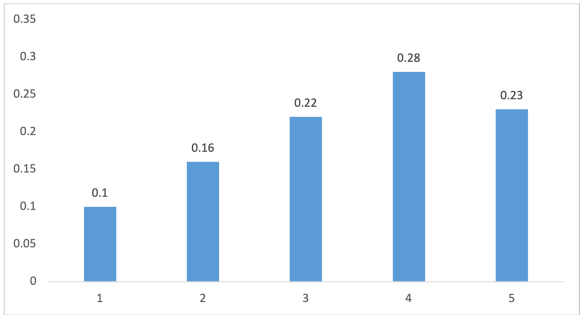
- A comparison between Person fit l_z and Person Reliability γ_d with real datasets
  - Polytomous, non-cognitive rating scale items;
  - *"Sensitivity to normative"* response style (Bolt & Johnson, 2009);

Table of Contents

Noncognitive Datasets (www.openpsychometrics.org)

- Machiavellianism Data
  - 20 items 1-5 rating scale
  - unidimensional
  - n = 5744
- Big Five Factor Markers Data
  - 50 items 1-5 rating scale
  - multidimensional (5 factors)
  - n = 5171
- Introversion-Extroversion Data
  - 91 items 1-5 rating scale
  - unidimensional
  - n = 7188





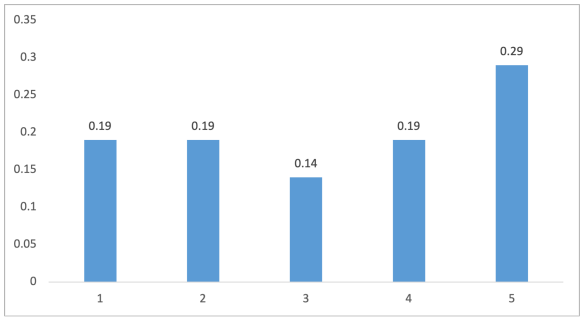


Table of Contents

Method

- Fit the Graded Response Model (GRM) to Empirical Datasets;

Method

- Fit the Graded Response Model (GRM) to Empirical Datasets;
- Estimate Person Fit l_z and Person Reliability γ_d Indices;

Method

- Fit the Graded Response Model (GRM) to Empirical Datasets;
- Estimate Person Fit l_z and Person Reliability γ_d Indices;
- Evaluate the Correlation Estimates Between Indices Across Data;

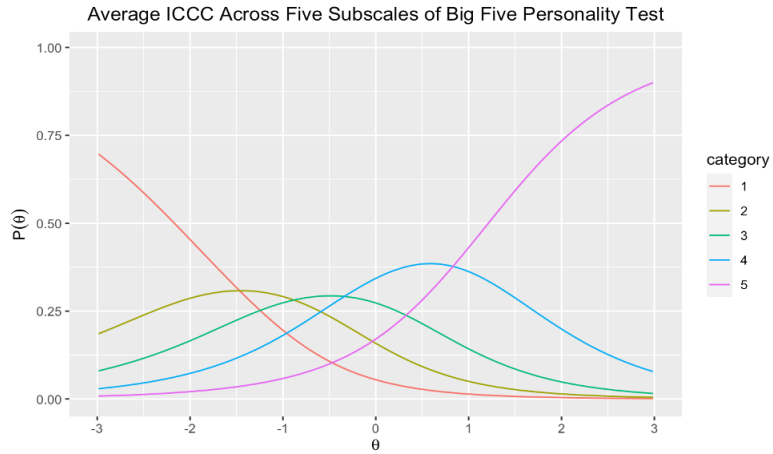
Method

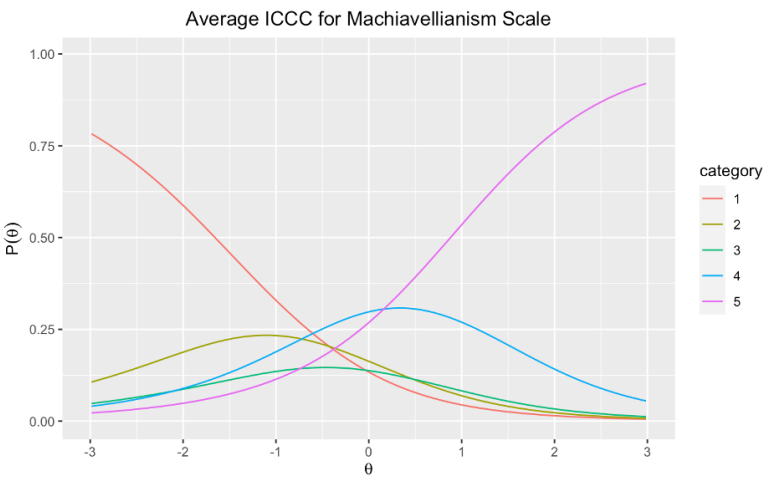
- Fit the Graded Response Model (GRM) to Empirical Datasets;
- Estimate Person Fit l_z and Person Reliability γ_d Indices;
- Evaluate the Correlation Estimates Between Indices Across Data;
- Fit Respondent-level Regression Models predicting l_z from γ_d and Response Style Indices.

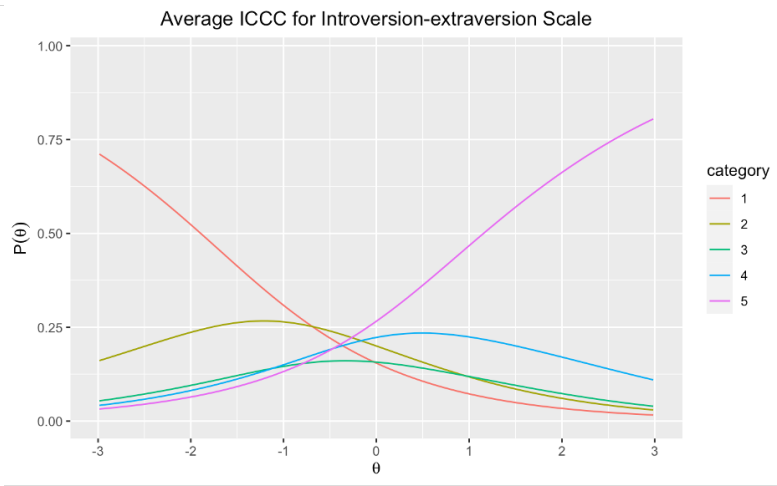
Method

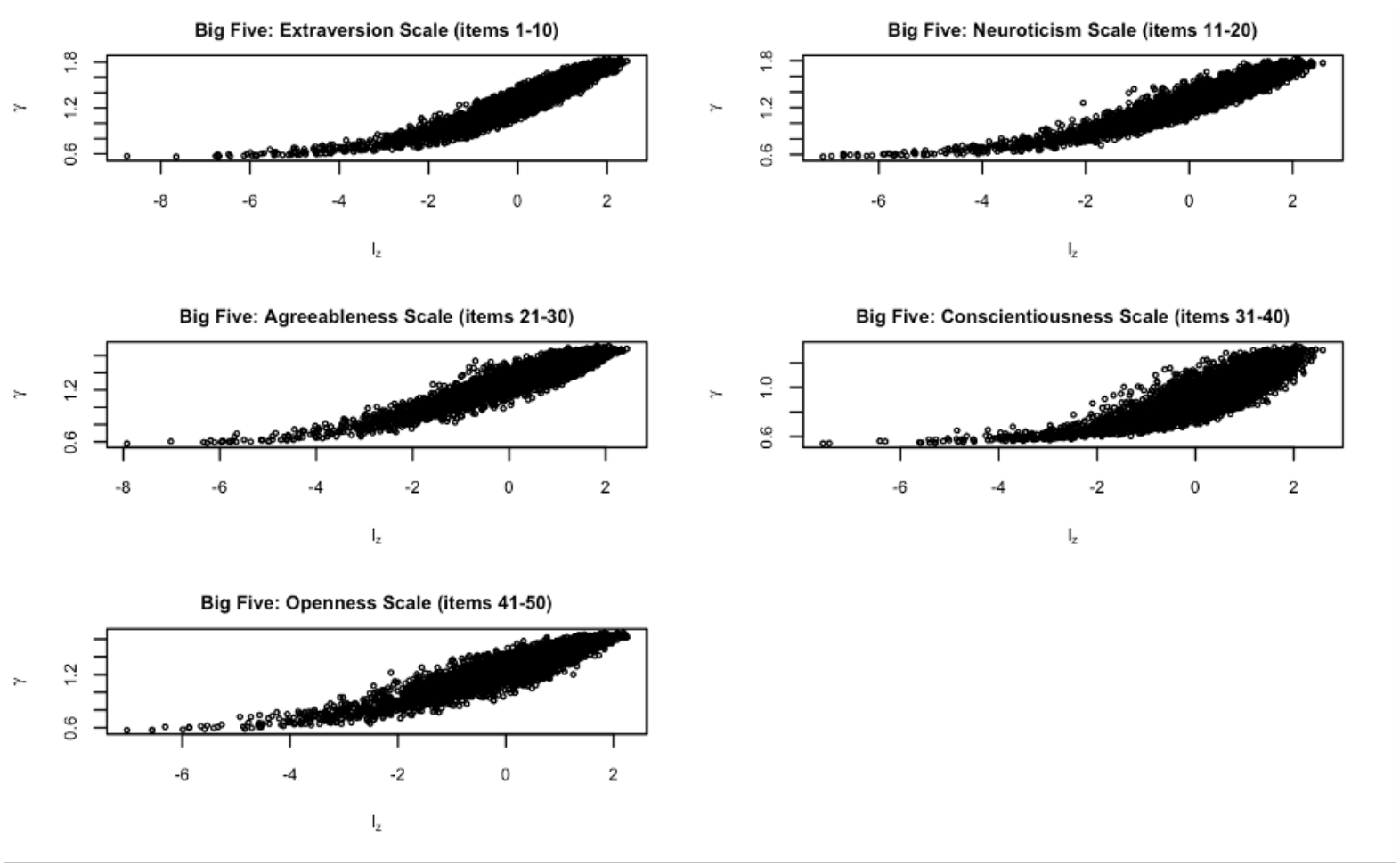
- Fit the Graded Response Model (GRM) to Empirical Datasets;
- Estimate Person Fit l_z and Person Reliability γ_d Indices;
- Evaluate the Correlation Estimates Between Indices Across Data;
- Fit Respondent-level Regression Models predicting l_z from γ_d and Response Style Indices.

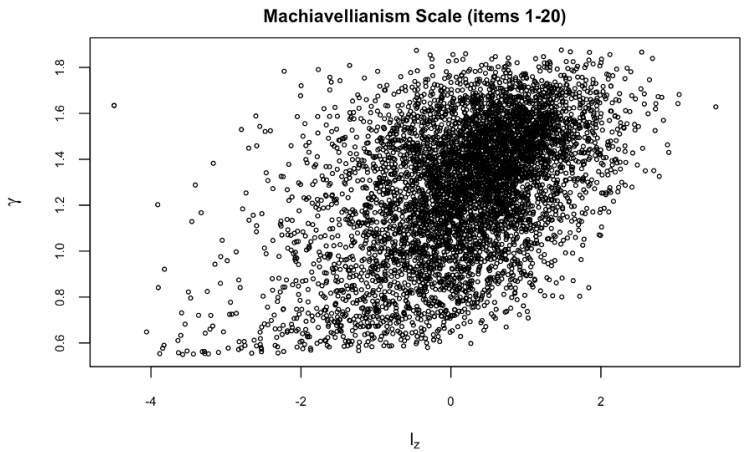
Table of Contents


Average ICCC Across Five Subscales of Big Five Personality Test
1.00
0.75
0.50
0.25
0.00
P(θ)
-3
-2
-1
0
1
2
3
θ
category
1
2
3
4
5


Average ICCC for Machiavellianism Scale
1.00
0.75
0.50
0.25
0.00
P(θ)
-3
-2
-1
0
1
2
3
θ
category
1
2
3
4
5


Average ICCC for Introversion-extraversion Scale
1.00
0.75
0.50
0.25
0.00
P(θ)
-3
-2
-1
0
1
2
3
θ
category
1
2
3
4
5


Big Five: Extraversion Scale (items 1-10)
γ
0.6
1.2
1.8
-8
-6
-4
-2
0
2
l_z
Big Five: Neuroticism Scale (items 11-20)
γ
0.6
1.2
1.8
-6
-4
-2
0
2
l_z
Big Five: Agreeableness Scale (items 21-30)
γ
0.6
1.2
-8
-6
-4
-2
0
2
l_z
Big Five: Conscientiousness Scale (items 31-40)
γ
0.6
1.0
-6
-4
-2
0
2
l_z
Big Five: Openness Scale (items 41-50)
γ
0.6
1.2
-6
-4
-2
0
2
l_z



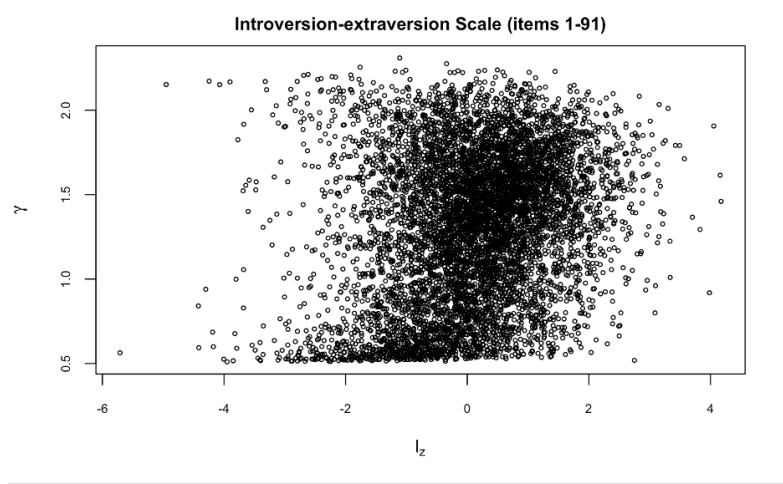


Table 1: Correlation Estimates between Person Fit l_z and Person Reliability γ

	Machiavellianism	Introversion/Extraversion	Big Five				
$\hat{r}$	0.49	0.23	0.93	0.94	0.93	0.85	0.92

Table 2: Forward Selection Regression Results Predicting l_z from γ, p_1, p_2, p_3, p_4, p_5

	Machiavellianism			Big Five			Introversion/Extroversion		
	est	s.e.	*p-value*	est	s.e.	*p-value*	est	s.e.	*p-value*
Intercept	-2.71	.05	<.001	-6.47	.04	<.001	-.44	.04	<.001
$\hat{\gamma}$	2.88	.04	<.001	5.41	.03	<.001	1.82	.04	<.001
p_1	1.15	.06	<.001						
p_2	-1.51	.08	<.001	-.18	.05	.003	-.62	.13	<.001
p_3	-5.49	.09	<.001	-.41	.04	<.001	-8.53	.15	<.001
p_4	-.42	.07	<.001				-2.96	.14	<.001
p_5				.28	.04	<.001			

Table 3: Example Respondents Displaying l_z Person Misfit, but High Person Reliability $\hat{\gamma}$

		Frequency of Category Selection							
	ID	Cat1	Cat2	Cat3	Cat4	Cat5	$\hat{\theta}$	l_z	$\hat{\gamma}$
Mach	1557	0	2	11	6	1	-0.17	-2.22	1.78
Mach	5458	1	3	12	3	1	-0.45	-2.00	1.72
IE	302	4	8	61	12	6	-0.19	-4.25	2.17
IE	5649	2	26	43	19	1	-0.38	-3.90	2.17

Table 4: Example Respondents Displaying l_z Person Fit, but Low Person Reliability $\hat{\gamma}$

		Frequency of Category Selection							
	ID	Cat1	Cat2	Cat3	Cat4	Cat5	$\hat{\theta}$	l_z	$\hat{\gamma}$
Mach	5207	8	1	0	4	7	-0.52	1.72	0.80
Mach	1978	7	1	0	5	7	-0.30	1.84	0.84
IE	3696	46	1	1	0	43	-0.24	2.75	0.52
IE	2282	41	13	2	10	25	-1.03	2.50	0.66

Table of Contents

Discussion

- Ferrando (2009) show high agreement between l_z and γ with binary items, by contrast we frequently see inconsistency between person fit l_z and person reliability γ due to response style heterogeneity in rating scale data:
  - High reliability $\hat{\gamma}$ but misfit by $\hat{l_z}$;
  - Low reliability $\hat{\gamma}$ but fit by $\hat{l_z}$.

Discussion

- Ferrando (2009) show high agreement between l_z and γ with binary items, by contrast we frequently see inconsistency between person fit l_z and person reliability γ due to response style heterogeneity in rating scale data:
  - High reliability $\hat{\gamma}$ but misfit by $\hat{l_z}$;
  - Low reliability $\hat{\gamma}$ but fit by $\hat{l_z}$.
- Normative aspects for the interpretation of response style;

Discussion

- Ferrando (2009) show high agreement between l_z and γ with binary items, by contrast we frequently see inconsistency between person fit l_z and person reliability γ due to response style heterogeneity in rating scale data:
  - High reliability $\hat{\gamma}$ but misfit by $\hat{l_z}$;
  - Low reliability $\hat{\gamma}$ but fit by $\hat{l_z}$.
- Normative aspects for the interpretation of response style;
- Simultaneous application of both person misfit and person reliability indices seems important for the evaluation of respondent-level validity;

Discussion

- Ferrando (2009) show high agreement between l_z and γ with binary items, by contrast we frequently see inconsistency between person fit l_z and person reliability γ due to response style heterogeneity in rating scale data:
  - High reliability $\hat{\gamma}$ but misfit by $\hat{l_z}$;
  - Low reliability $\hat{\gamma}$ but fit by $\hat{l_z}$.
- Normative aspects for the interpretation of response style;
- Simultaneous application of both person misfit and person reliability indices seems important for the evaluation of respondent-level validity;
- Alternative approach using response style models or different indices

Q & A

- Any Questions?

References I

Bolt, D. M. and Johnson, T. R. (2009).
Addressing score bias and differential item functioning due to individual differences in response style.
Applied Psychological Measurement, 33(5):335–352.

Curtis, D. D. (2004).
Person misfit in attitude surveys: Influences, impacts and implications.
International Education Journal, 5(2):125–143.

Ellis, J. L. and Van den Wollenberg, A. L. (1993).
Local homogeneity in latent trait models. a characterization of the homogeneous monotone irt model.
Psychometrika, 58(3):417–429.

Holland, P. W. (1990).
On the sampling theory roundations of item response theory models.
Psychometrika, 55(4):577–601.

References II


Levine, M. V. and Drasgow, F. (1982).
Appropriateness measurement: Review, critique and validating studies.
British Journal of Mathematical and Statistical Psychology, 35(1):42–56.


LEVINE, M. V. and DRASGOW, F. (1983).
Appropriateness measurement: Validating studies and variable ability models.
In *New horizons in testing*, pages 109–131. Elsevier.


Meijer, R. R. and Sijtsma, K. (2001).
Methodology review: Evaluating person fit.
Applied psychological measurement, 25(2):107–135.

References III

Ranger, J. (2013).
Modeling responses and response times in personality tests with rating scales.
Psychological Test and Assessment Modeling, 55(4):361.

Samejima, F. (1969).
Estimation of latent ability using a response pattern of graded scores.
Psychometrika monograph supplement.