パフォーマンス評価におけるハロー効果の影響を考慮した項目反応モデル

An Item Response Theory Model Considering Halo Effects in Performance Assessment

北風陵汰，宇都雅輝
Ryota Kitakaze, Masaki Uto
電気通信大学
The University of Electro-Communications
Email: {kitakaze, uto}@ai.lab.uec.ac.jp

あらまし： 近年，パフォーマンス評価におけるハロー効果を考慮できる項目反応モデルが提案されている．先行研究では，受検者由来および評価者由来のハロー効果を個別に扱ってきた．しかし，実際の評価場面では，両者が同時に生じる可能性がある．また，従来手法ではハロー効果の有無を二値的に表現しているが，実際にはその影響は連続的であると考えられる．本研究では，受検者および評価者の双方に由来するハロー効果を同時に扱い，その誘発度を連続的に表現可能な項目反応モデルを提案する．
キーワード： 教育測定，パフォーマンス評価，ハロー効果，項目反応理論，多相ラッシュモデル

1　はじめに

近年，論理的思考力や問題解決力といった高次の能力を測定する方法として，パフォーマンス評価が注目されている．しかし，パフォーマンス評価では評点が人間評価者の主観に依存するため，評価の信頼性の確保が困難であるという課題が指摘されている [1]．このような課題に対処する方法の一つとして，評価者や項目の特性を考慮した能力測定が可能な項目反応理論（IRT: Item Response Theory）が知られている．具体的には，多相ラッシュモデル（MFRM: Many-Facet Rasch Model）と呼ばれる IRT モデル群が提案され，様々なパフォーマンス評価への適用が進められている [1]．

一方，パフォーマンス評価の信頼性を低下させる要因の一つとして，複数の項目に基づいて受検者を評価する際に，項目全体にわたって類似した評点が与えられる「ハロー効果」が知られている．近年，ハロー効果を明示的に扱う MFRM の拡張モデルが提案されている [2]．これらのモデルでは，ハロー効果の発生要因を受検者由来および評価者由来に区別し，各要因を個別に検出・考慮することが可能である．しかし，両者が同時に作用する状況は扱うことができず，また，ハロー効果の誘発度も二値的にしか表現できないという課題がある．

そこで本研究では，受検者および評価者の双方に由来するハロー効果を同時に扱い，その誘発度を連続的に表現可能な新たな項目反応モデルを提案する．提案モデルは，ハロー効果の柔軟な解釈を可能にするとともに，データ適合と能力測定精度の改善が期待できる．本研究では，シミュレーション実験および実データ分析を通して，提案モデルの有効性を検証する．

2　多相ラッシュモデル

本研究では，J 人の受検者を R 人の評価者が I 個の評価項目に基づいてそれぞれ K 段階で評価するパフォーマンス試験を想定する．このような多相データを扱う代表的な IRT モデルとして，MFRM が知られている．代表的な定式化では，評価者 r が項目 i における受検者 j のパフォーマンスに評点 k を与える確率 P_{ijrk} は次式で表される．

$$P_{ijrk} = \frac{\exp \sum_{m=1}^{k} [\theta_j - \beta_r - d_{im}]}{\sum_{l=1}^{K} \exp \sum_{m=1}^{l} [\theta_j - \beta_r - d_{im}]}.$$

ここで θ_j は受検者 j の能力，β_r は評価者 r の厳しさ，d_{im} は項目 i におけるステップパラメータを表す．

3　ハロー効果を考慮した従来モデル

一般に，ハロー効果は受検者由来と評価者由来に大別される．Jin et al.[2] は，これらを個別に考慮可能な MFRM の拡張モデルを提案している．

受検者由来のハロー効果を扱う既存モデル（MFRM-HS と呼ぶ）は次式で定義される．

$$P_{ijrk} = \frac{\exp \sum_{m=1}^{k} [\theta_j - \beta_r - (1-x_j)d_{im} - x_j d'_m]}{\sum_{l=1}^{K} \exp \sum_{m=1}^{l} [\theta_j - \beta_r - (1-x_j)d_{im} - x_j d'_m]}.$$

また，評価者由来のハロー効果を扱う既存モデル（MFRM-HR と呼ぶ）は次式で定義される．

$$P_{ijrk} = \frac{\exp \sum_{m=1}^{k} [\theta_j - \beta_r - (1-x_r)d_{im} - x_r d'_m]}{\sum_{l=1}^{K} \exp \sum_{m=1}^{l} [\theta_j - \beta_r - (1-x_r)d_{im} - x_r d'_m]}.$$

ここで d'_m は項目非依存のステップパラメータであり，x_j と x_r は受検者 j と評価者 r がハロー効果を誘発するときに 1，そうでないときに 0 をとる二値変数である．

これらの従来モデルでは，受検者由来および評価者由来のハロー効果が同時に存在する状況を扱えないことに加え，その誘発度を二値的にしか表現できないという課題がある．

4　提案手法

これらの課題に対処するために，本研究では受検者および評価者の双方に由来するハロー効果を同時に扱い，その強度を連続的に表現可能な MFRM の拡張モデルを

表 1　パラメータ推定精度の評価結果

J	I	R	θ_j	β_r	d_{im}	d'_m	y_j	y_r
100	5	40	0.120	0.092	0.110	0.041	0.302	0.293
40	5	40	0.145	0.142	0.174	0.048	0.321	0.297
100	5	10	0.218	0.114	0.224	0.052	0.318	0.335
100	10	40	0.100	0.084	0.110	0.030	0.303	0.280

表 2　ハロー効果を混入させた場合のモデル比較結果

データ生成	J	I	R	提案モデル	MFRM	MFRM-HS	MFRM-HR
提案モデル	100	10	40	**83,114.2**	85,703.6	85,306.1	84,210.1
	100	5	40	**41,994.2**	43,052.4	42,794.6	42,429.8
	40	5	40	16,895.0	17,195.8	17,138.9	**16,884.8**
	100	5	10	**10,484.1**	10,664.9	10,621.3	10,493.6
MFRM	100	10	40	**82,852.8**	85,574.0	85,167.7	84,034.0
	100	5	40	**41,948.7**	42,968.9	42,731.8	42,385.5
	40	5	40	**16,869.7**	17,180.9	17,133.8	16,877.5
	100	5	10	**10,498.3**	10,663.7	10,634.3	10,501.4
MFRM-HS	100	10	40	**82,937.5**	85,699.3	85,293.8	84,179.0
	100	5	40	**42,007.3**	43,072.4	42,816.8	42,457.9
	40	5	40	**16,834.0**	17,138.7	17,084.2	16,841.9
	100	5	10	**10,486.5**	10,667.4	10,644.4	10,491.3
MFRM-HR	100	10	40	**82,939.9**	85,626.7	85,224.2	84,024.9
	100	5	40	**41,979.8**	43,039.1	42,806.4	42,456.8
	40	5	40	**16,831.0**	17,136.0	17,085.2	16,851.9
	100	5	10	**10,458.0**	10,621.7	10,603.6	10,459.0

提案する．提案モデルでは P_{ijrk} を次式で定義する．

$$P_{ijrk} = \frac{\exp \sum_{m=1}^{k} \left[\theta_j - \beta_r - y_j y_r d_{im} - (1 - y_j y_r)\, d'_m\right]}{\sum_{l=1}^{K} \exp \sum_{m=1}^{l} \left[\theta_j - \beta_r - y_j y_r d_{im} - (1 - y_j y_r)\, d'_m\right]}.$$

ここで $y_j, y_r \in [0, 1]$ はそれぞれ受検者 j と評価者 r のハロー効果非誘発度を表す連続変数であり，値が小さいほどハローを強く誘発すると解釈する．受検者または評価者の双方がハロー効果を強く誘発する場合（$y_j, y_r \approx 0$）には，$y_j y_r$ が 0 に近づき，項目に依存しないステップパラメータ d'_m の影響が大きくなり，全項目が共通の評定尺度で評価されやすくなることを表現している．

5　シミュレーション実験

本節では，シミュレーション実験を通して，提案モデルの (i) パラメータ推定精度，(ii) ハロー効果存在下における頑健性，(iii) 能力推定精度を検証する．

まず，パラメータリカバリ実験により，提案モデルの各パラメータが適切に推定可能かを評価した．提案モデルから評点データを生成し，MCMC（Markov chain Monte Carlo）法により各パラメータを推定した後，推定値と真値との RMSE（Root Mean Squared Error）を算出した．表 1 に RMSE の平均値を示す．全体として先行研究 (1) と整合する合理的な傾向が確認され，提案モデルのパラメータ推定が適切に行えることが示された．

次に，ハロー効果が存在する状況における提案モデルの頑健性を，情報量基準によるモデル比較を通して評価した．提案モデル，MFRM，MFRM-HS，MFRM-HR の各モデルから生成したデータに対し，一部の受検者および評価者に人為的にハロー効果を混入させ，WAIC によるモデル比較を行った．表 2 に結果を示す．太字は各条件において最小の WAIC 値を示す．データ生成モデルによらず，ほぼすべての条件で提案モデルが最小の WAIC を示し，ハロー効果を含む評価状況に対して高い適合性を有することが確認された．

表 3　能力推定精度の評価

J	I	R	No-Halo	Only-HS	Only-HR	提案モデル
100	10	40	0.1248	0.1185	0.1191	**0.1172**
100	5	40	0.1325	0.1241	0.1242	**0.1223**
40	5	40	0.1342	0.1254	0.1262	**0.1243**
100	5	10	0.1978	0.1918	0.1916	**0.1902**

表 4　実データ実験におけるモデル比較結果

モデル	サイクリスト（調査 1）	サイクリスト（調査 2）	面接
MFRM	119,450.1	124,710.1	31,326.1
MFRM-HS	118,807.5	123,707.8	31,317.6
MFRM-HR	117,065.3	122,589.2	31,236.8
提案モデル	**115,713.2**	**120,327.3**	**30,478.2**

最後に，ハロー効果の考慮方法が受検者の能力推定精度に与える影響を比較した．提案モデルから生成したデータに対し，ハロー効果を考慮しない場合（No-Halo），受検者由来のみ考慮（Only-HS），評価者由来のみ考慮（Only-HR），双方を同時に考慮（提案モデル）の 4 条件で能力推定を行った．表 3 に示すように，提案モデルがすべての条件において最小の RMSE を示し，受検者および評価者由来のハロー効果を同時に考慮することの有効性が示唆された．

6　実データ実験

提案モデルの有効性を検証するため，実データを用いたモデル比較実験を行った．使用したデータは，サイクリスト評価データセット（調査 1・調査 2 の 2 種）と面接評価データセットの計 3 種類である．各データに対し，MFRM，MFRM-HS，MFRM-HR，提案モデルを比較対象として WAIC を用いてモデル比較を行った．

表 4 に結果を示す．太字は最小の WAIC 値を示す．提案モデルはすべてのデータで最小の WAIC を示し，実データにおいても最良のモデルとして選択された．

7　まとめと今後の課題

本研究では，パフォーマンス評価における代表的なバイアス要因であるハロー効果を扱う新たな IRT モデルを提案した．提案モデルは，受検者および評価者に由来するハロー効果を同時に考慮し，その誘発度を連続的に表現可能である点に特徴がある．提案モデルは，データ適合と能力推定精度を改善し，ハロー効果の柔軟な解釈を可能にした．今後は，教育評価をはじめとする多様な実データへの適用を通して，提案モデルの有効性をさらに検証していく．

参考文献

(1) Masaki Uto and Maomi Ueno. A generalized many-facet Rasch model and its Bayesian estimation using Hamiltonian Monte Carlo. *Behaviormetrika*, Vol. 47, No. 2, pp. 469–496, 2020.

(2) Kuan-Yu Jin and Thomas Eckes. When raters generalize: Examining sources of halo effects with mixture Rasch facets models. *Behavior Research Methods*, Vol. 57, No. 5, 2025.