評価者特性の時間変動を考慮した項目反応モデル

Item Response Theory Model Considering Rater Parameter Drift

林真由 *1，宇都雅輝 *1
Mayu Hayashi*1, Masaki Uto*1
*1 電気通信大学
*1The University of Electro-Communications
Email: {hayashi_mayu, uto}@ai.lab.uec.ac.jp

あらまし： パフォーマンス評価では，採点結果が評価者の特性に依存してしまい，評価の信頼性が低下する問題が知られている．この問題を解決する手法として，評価者バイアスの影響を取り除いて受検者の能力を推定できる項目反応モデルが近年多数提案されている．それらの既存モデルの多くは評価者の採点基準が採点過程で変化しないことを仮定している．しかし，この仮定は実際には成り立たない場合がある．この問題を解決するために，本研究では，評価者の厳しさの時間変化を推定できる新たな項目反応モデルを提案する．具体的には，一般化多相ラッシュモデルに時間区分ごとの評価者の厳しさを表すパラメータを付与したモデルを提案する．

キーワード： 項目反応理論，評価者バイアス，評価者ドリフト，教育測定

1　はじめに

近年，様々な評価場面においてパフォーマンス評価のニーズが高まっている．パフォーマンス評価では人間の評価者が採点を行うため，評価者の厳しさや一貫性などの特性差がバイアス要因となり，受検者の能力測定の信頼性が低下する問題が知られている．

このような問題を解決するアプローチの一つとして，評価者の特性差の影響を考慮して受検者の能力を推定できる項目反応理論（Item response theory: IRT）モデル [1] が近年多数提案されている．一方で，それらの既存モデルのほとんどは評価者の特性が採点過程で変化しないことを仮定している．しかし，多数の受検者を長時間かけて採点するような場合には，評価者の特性が採点の過程で変化する評価者ドリフト (Rater Drift) と呼ばれる現象がしばしば生じる．そのような評価者ドリフトを考慮できるモデルも提案されているが [2]，既存モデルでは評価者特性の時間変化を直線的にしか表現できないという問題点がある．

この問題を解決するために，本研究では，時間区分ごとの評価者の厳しさを推定できる新たな IRT モデルを提案する．提案モデルでは，既存モデルよりも柔軟に評価者ドリフトを表現でき，モデルの性能が改善すると期待できる．本研究では，シミュレーション実験と実データ実験を通して提案モデルの有効性を示す．

2　評価者特性を考慮した項目反応理論

評価者特性を最も柔軟に表現できる IRT モデルとしては，一般化多相ラッシュモデルが知られている [3]．このモデルでは，評価者 r が受検者 j のパフォーマンスにスコア k を与える確率を次式で定義する．

$$P_{jrk} = \frac{\exp\sum_{m=1}^{k}\{\alpha_r(\theta_j - \beta_r - d_{rm})\}}{\sum_{l=1}^{K}\exp\sum_{m=1}^{l}\{\alpha_r(\theta_j - \beta_r - d_{rm})\}} \quad (1)$$

ここで，α_r は評価者 r の一貫性，θ_j は受検者 j の能力，β_r は評価者 r の厳しさ，d_{rk} は評価者 r のスコア k に対する厳しさを表すステップパラメータである．モデルの識別性のために，$\sum_{k=2}^{K} d_{rk} = 0$，$d_{r1} = 0$ を仮定する．

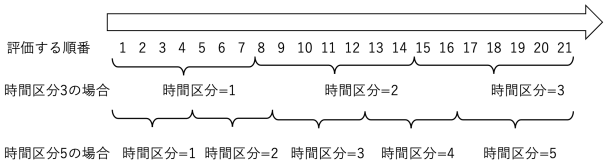

図 1　時間区分データの構成イメージ

一般化多相ラッシュモデルは評価者の特性が評価中に変化しないことを仮定しているが，1 章で述べたように，現実には評価者ドリフトという現象が起こる場合がある．評価者ドリフトを考慮できるモデルとして，評価者 r が時間区分 t で採点した受検者 j のパフォーマンスにスコア k を与える確率を次式で定義したモデルが提案されている [2]．

$$P_{jrtk} = \frac{\exp\sum_{m=1}^{k}(\theta_j - \beta_r - \pi_r t - d_m)}{\sum_{l=1}^{K}\exp\sum_{m=1}^{l}(\theta_j - \beta_r - \pi_r t - d_m)} \quad (2)$$

ここで，β_r は評価者 r の初期の厳しさ，π_r は評価者 r の厳しさの変化の傾きを表す．このモデルのパラメータ推定には，図 1 のように，各評価者の採点データを時系列順に並べ，それを複数の時間区分に分割して作成した時間区分情報付きの得点データを用いる．

このモデルは時間経過による評価者特性の変化を推定できるが，その変化は直線的にしか表現できない．この問題を解決するため，本研究では時間区分ごとの評価者の厳しさを推定できる新しい IRT モデルを提案する．

3　提案モデル

提案モデルでは，評価者 r が時間区分 t で採点した受検者 j のパフォーマンスにスコア k を与える確率 P_{jrtk} を次式で定義する．

$$P_{jrtk} = \frac{\exp\sum_{m=1}^{k}\alpha_r(\theta_j - \beta_{rt} - d_{rm})}{\sum_{l=1}^{K}\exp\sum_{m=1}^{l}\alpha_r(\theta_j - \beta_{rt} - d_{rm})} \quad (3)$$

$$\beta_{rt} \sim \mathrm{N}(\beta_{r,t-1}, \sigma),\ \beta_{r1} \sim \mathrm{N}(0,1),\ \sigma \sim \mathrm{LN}(-3,\ 0)$$

表 1　モデル比較の結果

	提案モデル	従来モデル	比較モデル 1	比較モデル 2	比較モデル 3
全ての評価者のデータ	**5027.951**	5361.581	5225.050	5104.463	5032.362
バイアスを強調した評価者を除外したデータ	**3134.700**	3279.444	3190.802	3154.331	3137.137

ここで，$\mathrm{N}(\mu, \sigma^2)$ と $\mathrm{LN}(\mu, \sigma^2)$ はそれぞれ平均 μ，標準偏差 σ^2 の正規分布と対数正規分布を表す．β_{rt} は評価者 r の時間区分 t における厳しさを表すパラメータであり，これにより時間区分ごとの評価者の厳しさを捉えることができる．また，提案モデルでは β_{rt} が直前の時間区分での厳しさ $\beta_{r,t-1}$ に依存して決まるように分布を設定している．ただし，一般に評価者の厳しさは直前の時間区分から大きくは変動しないと考えられるため，β_{rt} の分布の標準偏差 σ に対しては，その推定値が小さくなるような事前分布を採用している．さらに，提案モデルでは，より柔軟に評価者特性を表現するために，一般化多相ラッシュモデルで採用されている評価者の一貫性パラメータ α_r と各スコアに対する厳しさパラメータ d_{rk} も導入している．モデルの識別性のために，$\theta_j \sim \mathrm{N}(0, 1)$，$\prod_r \alpha_r = 1$，$d_{r1} = 0$，$\sum_{k=2}^{K} d_{rk} = 0$ を仮定する．

提案モデルのパラメータ推定手法にはマルコフ連鎖モンテカルロ法（Markov chain Monte Carlo methods : MCMC）を用いる．パラメータの事前分布は θ_j, $\log\alpha_r$, β_{rt}, $d_{rk} \sim \mathrm{N}(0,\ 1^2)$ とした．バーンインは 1000 とし，1000〜2000 時点までの 1000 サンプルを用いた．

4　実データ実験

本章では，実データ実験を通して提案モデルの有効性を評価する．本実験では，あるエッセイ課題に対する 134 名の解答を，16 名の評価者が 5 段階得点で採点したデータを使用する．評価者には，採点を 4 日に分け，日ごとに全体の 1/4 ずつ採点するように指示した．本実験では，これらの採点日ごとに時間区分 1，2，3，4 として扱う．なお，16 名の被験者のうち 6 人に対しては採点の際に，「日ごとに徐々に厳しくなるように採点せよ」などの指示を与え，人為的に評価者バイアスを発生させたデータも収集した．

このデータに対して提案モデルを適用して推定された β_{rt} の値を図 2 に示す．図では，縦軸が β_{rt} の値，横軸が時間区分，各線がそれぞれの評価者の β_{rt} の推定値を表す．また，図中では，例として，日ごとに徐々に厳しくするように指示した評価者の結果を赤色の線でハイライトしている．図の赤線から，この評価者が指示通りに採点をしており，その特性をモデルが適切に表現できていることがわかる．指示をした他の評価者についても，同様に，指示通りの傾向が表現できていたことを確認した．さらに，指示を与えていない評価者の中にも，評価者ドリフトが疑われる評価者が見受けられた．以上から，提案モデルが評価者ドリフトの傾向を適切に推定できていることがわかる．

次に，提案モデルの性能を評価するために，式 (2) の

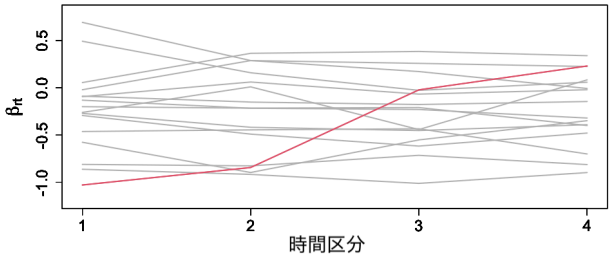

図 2　実データ実験による β_{rt} の推定結果

既存モデルとの情報量規準によるモデル比較を行った．さらに，既存モデルと提案モデルとの差分のうち，どの要因が有効であったかを確認するために，以下の 3 つのモデルとの比較を行った．

比較モデル 1:　d_{rk} を d_k に変更した提案モデル
比較モデル 2:　β_{rt} を $\beta_r - \pi_r t$ に変更した提案モデル
比較モデル 3:　α_r を削除した提案モデル

情報量規準には WAIC（Widely Applicable Information Criterion）を用いた．この基準は値が小さい方が適したモデルであることを示す．

表 1 に，全ての評価者のデータから求めた WAIC と，指示を与えてバイアスを強調した評価者を除外したデータから求めた WAIC を示す．表 1 から，いずれの場合でも提案モデルが最適なモデルとして選択されたことが確認できる．また，提案モデルと比較モデル 2 との比較から，時間区分ごとの厳しさパラメータを従来モデルの形式に変更すると性能が低下することがわかる．さらに，提案モデルで追加した α_r や d_{rk} を取り除いた比較モデル 1 や 3 も性能が低下していることが読み取れる．以上から，提案モデルの有効性が確認できる．

5　まとめ

本研究では，評価者の厳しさパラメータの時間変化を推定できる新しい IRT モデルを提案し，実データ実験を通してその有効性を評価した．なお，提案モデルは課題が一つの場合を想定しており，複数課題が出題される試験に適用できないため，今後は課題パラメータを追加したモデルも開発したい．

参考文献

[1] F.M. Lord. *Applications of item response theory to practical testing problems*. Erlbaum Associates, 1980.
[2] S.W. Raudenbush and A. S. Bryk. *Hierarchical linear models: Applications and data analysis methods*, Vol. 1. Sage Publications, 2001.
[3] Masaki Uto and Maomi Ueno. A generalized many-facet Rasch model and its Bayesian estimation using Hamiltonian Monte Carlo. *Behaviormetrika, Springer*, Vol. 47, No. 2, pp. 469–496, 2020.