

段階反応モデルでの適応型テストの回答データを用いた新作項目の特性値推定

○小路 純寛 杉山 剛 藤田 彩子
株式会社リクルートキャリア 測定技術研究所

1 はじめに

適応型テストの運用においては項目プールの拡充は重要なテーマのひとつである。項目プールを拡充する方法としては、予備テストを行う方法のほかに、本番テストの中に新作項目を採点除外項目として紛れ込ませる形でデータ収集を行う方法がある。

予備テストを行う方法はコストがかかるほか、運用上の負荷も大きいという課題があるが、本番テストの中に新作項目を紛れ込ませる方法は、本番テストで得られた受検者の特性値 θ を新作項目の項目特性値推定に使用することで本番テストの項目との等化済みの項目特性値が算出されるという利点や、本番テストの運用の延長上で項目プールの拡充が可能になるという利点がある(杉山ら, 2016)。

項目反応理論を使用したテストにおいては2パラメータロジスティックモデル(以下2PLMと略す)を用いた2値採点が一般的であるが、より多くの情報量を得ることができるモデルとして段階反応モデル(Graded Response Model, 以下GRMと略す)、名義反応モデルなどを用いた多値採点も知られている(仁田ら, 2012)。

ここでGRMを用いて適応型で提供されている能力検査の本番テストに新作項目を紛れ込ませて回答データを収集するという状況を想定する。

新作項目のGRMでの項目特性値推定については仁田ら(2012)の提案手法がある。多肢択一式の客観式能力テストを用いて、多重対応分析による選択肢の序列づけと、MCMCによるGRMでの項目特性値推定をおこなっているが、この手法は出題が固定されたテストへの適用を前提としており、適応型出題のテストには適用できない。

また、適応型出題のテストにおける新作項目の項目特性値推定については杉山ら(2016)の提案手法があるが、この手法は2PLMを前提としたものであり、GRMへの単純拡張は難しい。

本研究では、仁田ら(2012)の提案手法を拡張して、GRMを用いた適応型テストに新作項目を紛れ込ませたときに、受検者の能力特性値 θ と当該項目の回答情報を使用して、選択肢の序列づけおよび項目特性値の推定を行う手法について、シミュレーションによる検証をおこなった。これにより、GRMを用いて適応型で提供されている能力検査において、本番テストに新作項目を紛れ込ませる方法でも項目特性値の推定が可能になると期待される。

2 方法

GRMでは受検者の選択肢を測定したい尺度に沿って順序づけする必要がある。仁田ら(2012)では、多重対応分析の結果と出題者の意図としての正答を比較しながら決めている。適応型出題のテストの場合は出題項目が受検者により異なるため多重対応分析を行うことはできない。一方で本番テストに紛れ込ませた新作項目では、本番テストで推定された受検者の能力特性値 θ が得られている。そこで多重対応分析の代わりに受検者の特性値 θ との相関が最大になるように選択肢の序列づけを行う。以下、本稿ではこうやって決定した選択肢の序列を項目情報カテゴリと呼ぶ。

次に項目特性値の推定であるが、仁田ら(2012)では、固定された出題順の全項目の項目情報カ

テゴリをもとに能力特性値 θ と項目特性値を MCMC により同時に推定している. 適応型の出題の場合は同じ出題順の受検者は事実上存在しないため, 同じ方法をとることはできない. そこで今回は仁田ら(2012)のモデルを拡張して, 1 問分の項目情報カテゴリと受検者の能力特性値 θ をもとに, MCMC により項目特性値だけを推定する.

本研究では以下の 2 つの仮説についてシミュレーションにより検証する.

仮説 1: 受検者の能力特性値 θ との相関が最大になるよう定めた選択肢の序列づけは, 各項目の項目情報カテゴリをよく再現する

仮説 2: 受検者の能力特性値 θ と項目情報カテゴリを用いることで, 項目特性値を精度よく推定できる

3 シミュレーション手順

4 肢の選択式テストを想定し, 以下のシミュレーションをおこなった.

Step1: 受検者の特性値 θ を標準正規分布 $N(0, 1)$ により発生させる

Step2: 項目特性値を a 値 ($0.1 \leq a \leq 1.0$), $b^*_1 \sim b^*_3$ ($-3.0 \leq b^*_1 \leq b^*_2 \leq b^*_3 \leq 3.0$) の範囲で変化させた項目を 100 項目用意する

Step3: 各受検者の各項目についての項目情報カテゴリごとの選択確率を GRM に基づいて計算する (項目情報カテゴリごとの選択確率の計算式は補足資料参照)

Step4: 各受検者の各項目についての項目情報カテゴリを Step3 で求めた確率を使って多項分布からランダムに発生させ, 受検者の選択する項目情報カテゴリとする. この項目情報カテゴリをそのまま受検者の選択肢番号とする.

Step5: 各項目について, 各受検者の特性値 θ との相関が最大になるような選択肢の序列づけをおこない, その結果と選択肢番号 (= 項目情報カテゴリ) とを比較する.

Step6: 各受検者の特性値 θ と Step4 で生成した項目情報カテゴリを用いて各項目の項目特性値を MCMC により推定し, Step2 で設定した項目特性値と比較する

なお Step1~Step4 はシミュレーションをおこなうための受検者, 項目, 回答データの準備であり, Step5 が選択肢の序列づけ, Step6 が項目特性値推定のシミュレーションである.

4 シミュレーション結果

受検者人数を 500 名~1 万名で変化させながら Step5 の比較をおこなった (表 1). 受検者人数が増えるにつれ, 選択肢の序列づけが項目情報カテゴリと一致していない項目の数が減っているが, 1 万名になっても不一致の項目はゼロにはならない. ただし不一致となるのは項目特性値の識別力 (a 値) が低い項目がほとんどであり, a 値が 0.4 以上の項目に絞るとほとんどの項目が一致する.

表1 項目情報カテゴリと一致していない項目の数

人数	500名	1000名	1500名	2000名	2500名	3000名	3500名	4000名	4500名	5000名	10000名
全項目 (100個)	15	13	11	10	7	10	10	10	6	7	6
a 値が0.4以上の項目 (70個)	3	1	0	2	0	0	1	1	0	0	0

つづいて, 受検者人数を 1000 名~3000 名で変化させながら Step6 の比較を行った (表 2). 表 2 のとおり, a 値の推定結果は各項目の項目特性値をよく再現している. また $b^*_1 \sim b^*_3$ については人数が増えるにつれて RMSE が改善することがわかる. 推定が必ずしもうまくいっていない項目も一

部あるが、これらの多くは a 値 0.4 未満(図中の▲)であった。a 値 0.4 未満の項目を除くと $b^*_1 \sim b^*_3$ の RMSE も大きく改善され、3000 名ではすべての項目特性値で RMSE が 0.1 未満となった。

表2 項目特性値推定結果と真値の比較

		相関係数			RMSE		
		1000名	2000名	3000名	1000名	2000名	3000名
全項目	a	0.990	0.995	0.996	0.04	0.03	0.02
	b^*_1	0.957	0.988	0.972	0.40	0.21	0.33
	b^*_2	0.978	0.992	0.992	0.24	0.15	0.15
	b^*_3	0.927	0.973	0.989	0.52	0.29	0.18
a値 0.4 以上	a	0.975	0.988	0.992	0.04	0.03	0.03
	b^*_1	0.993	0.996	0.998	0.15	0.12	0.09
	b^*_2	0.996	0.999	0.999	0.11	0.06	0.05
	b^*_3	0.993	0.995	0.998	0.14	0.12	0.08

項目特性値推定の環境
R3.3.1 rstan2.14.1 を使用
iter = 3000
warmup = 1000
chains = 4

図1 a値の真値と推定結果

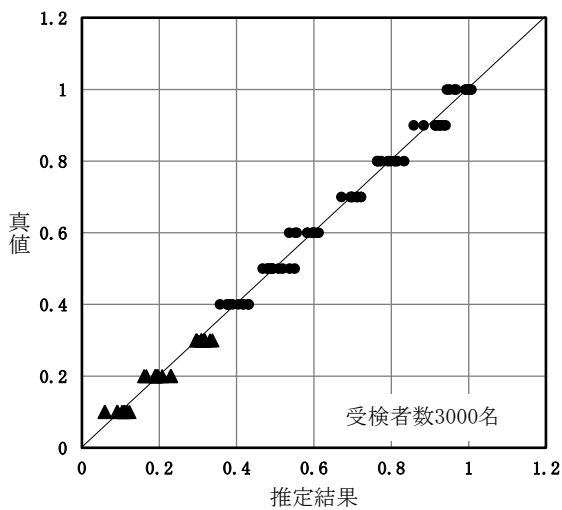


図2 b^*_1 の真値と推定結果

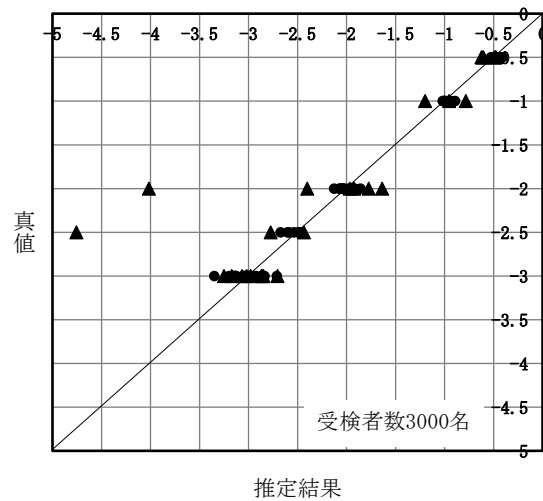


図3 b^*_2 の真値と推定結果

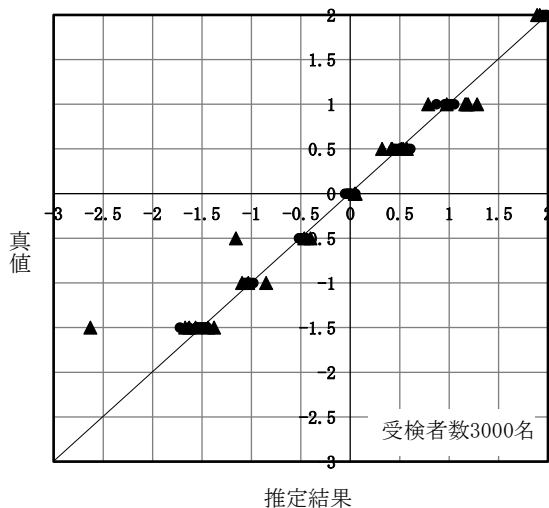
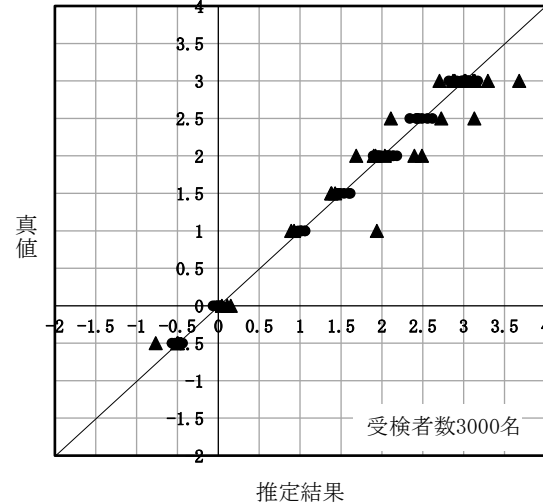


図4 b^*_3 の真値と推定結果



5 考察と今後に向けて

上記のシミュレーション結果から、仮説 1 については a 値が 0.4 以上の項目で 1000 名程度の十分な回答データがある、という条件のもとで支持された。a 値が低い項目については項目情報力

テゴリごとの選択確率の差が小さいため、正しい選択肢の順序付けをおこなうためにはもっと多くの受検者数が必要と考えられる。

仮説 2 についても、a 値が 0.4 以上の項目について、3000 名程度の十分な回答データがある、という条件のもとで支持された。a 値が低い項目については、a 値は精度よく推定できているものの、 $b_{i1}^* \sim b_{i3}^*$ の誤差が大きくなる。

仮説 1, 2 ともに a 値の低い項目については支持されない結果となったが、a 値の低い項目は通常本番テストに採用することはないため、実務上の問題はない。

本稿では、4 肢の項目を想定して、選択肢の序列づけと特性値の推定の手法についてシミュレーションで検証をおこない、a 値が一定以上の項目については実用的な手法であるということが確認された。しかしながら、受検者の特性値 θ との相関係数が最大になるような選択肢の序列づけについては選択肢数が増えていくと計算量が爆発的に増えてしまうことが懸念される(たとえば 10 肢だと 360 万通り)。そこで今後は、より効率的な選択肢の序列づけ手法について検討したい。

補足資料：GRM によるカテゴリごとの選択確率について

項目が K 個 ($0, \dots, k, \dots, K-1$) の項目情報カテゴリを持つとする。ある受検者の i 番目の項目への回答を、 u_i とし、潜在特性値 θ をもつ受検者が項目 i に $k \leq u_i$ と反応する確率を $P_{ik}^*(\theta)$ と定義する。 $k=1, \dots, k, \dots, K-1$ における $P_{ik}^*(\theta)$ を 2PLM によって表現すると、

$$P_{ik}^*(\theta) = \frac{1}{1 + \exp[-Da_i(\theta - b_{ik}^*)]}$$

となる。ここで $P_{i0}^*(\theta)$ と $P_{iK}^*(\theta)$ は

$$\begin{aligned} P_{i0}^*(\theta) &= 1 \\ P_{iK}^*(\theta) &= 0 \end{aligned}$$

と定義する。また境界特性曲線の位置母数 b_{ik}^* は項目情報カテゴリの順序を表現するため、

$$b_{i1}^* < b_{i2}^* < \dots < b_{ik}^* < \dots < b_{iK-1}^*$$

とすると、潜在特性値 θ をもつ受検者が項目 i にカテゴリ k と反応する確率は、

$$P(u_i = k|\theta) = P_{ik}(\theta) = P_{ik}^*(\theta) - P_{ik+1}^*(\theta)$$

と表現される (Samejima, 1969 ; 豊田, 2005)。

参考文献

- [1] 仁田光彦, 舛田博之 (2012) : 「客観式能力テストへの段階反応モデル適応についての検討」日本テスト学会第 10 回大会発表資料集
- [2] 杉山剛, 仁田光彦 (2016) : 「適応型テストの回答データを用いた新作項目の項目特性値推定」日本テスト学会第 14 回大会発表資料集
- [3] Samejima, F. (1969) : Estimation of latent ability using a response pattern of graded scores. Psychometrika, Monograph Supplement No. 17.
- [4] 豊田秀樹 (2005) : 「項目反応理論 [理論編] - テストの数理 -」 (第 2 章, p78-79)