多値アトリビュートにおける 認知診断の正確度へのQ行列の誤設定の影響 ——アトリビュートの階層性がある場合の検討——

丹 亮 人, 岡 田 謙 介

Impacts of Q-matrix Misspecification on the Accuracy of Polytomous-attribute Cognitive Diagnosis:
An Examination in the Presence of Attribute Hierarchy

Yoshito TAN and Kensuke OKADA

Cognitive diagnostic models (CDMs) are a class of statistical models that diagnose the mastery of respondents' cognitive traits, which are called attributes or skills. In the typical applications of CDMs, the Q-matrix, which represents which attributes are measured by each item, is specified by domain experts. In the case of dichotomous attributes, the impacts of Q-matrix misspecification on the classification accuracy have recently been studied; however, the case of polytomous attributes has not been reported. Therefore, in the present study, we examined how the difference between true and misspecified Q-matrix elements affects classification accuracy under four forms of attribute hierarchies. It was revealed that, in most conditions, larger difference between true and misspecified values resulted in lower classification accuracy. The impact of misspecification was the largest under the linear form of attribute hierarchy, which could be due to its smaller number of items that measure attribute levels. These results suggest that the number of items assigned to each attribute levels can be a key factor that affects the classification accuracy, especially when the degree of misspecification is large.

Key words: cognitive diagnostic models, Q-matrix misspecification, polytomous attributes, attribute hierarchy

キーワード: 認知診断モデル, Q行列の誤設定, 多値アトリビュート, アトリビュート階層性

1. 問題と目的

1.1. 認知診断モデルとQ行列の誤設定

教育学や心理学において, テストへの回答から回答者の認知的な特性の強み, 弱みについての情報をフィードバックするために特定の知識構造や処理スキルを測定するようにデザインされたアセスメントは認知診断アセスメント (cognitive diagnostic assess-

ment, CDA; Leighton & Gierl, 2007) と呼ばれる. CDA は, 回答者がフィードバックされた認知的な情報をその後の学習に活かすことができる点で重要である.

CDA において, 回答者の認知的な特性の有無を正確に診断することは主に統計学的な問題として考えられてきた. 潜在的な特性の有無を診断するための統計モデルにはいくつかの種類があるが, その中でも近年特に注目されてきているモデルに認知診断モデル (cognitive diagnostic models, CDM; von Davier & Lee, 2019; 山口・岡田, 2017) がある. CDM において, 潜在的特性はアトリビュートやスキルと呼ばれ, 測定されるアトリビュートはテスト項目ごとに定め

東京大学大学院教育学研究科
(Graduate School of Education, The University of Tokyo)

連絡先: 〒113-0033 東京都文京区本郷 7-3-1

E-mail: tan-yoshito753@g.ecc.u-tokyo.ac.jp

表 1. 算数の Q 行列の例

項目	足し算	掛け算	通分
32+15	1	0	0
9×8	0	1	0
2/10 + 1/2	1	1	1
4/21 + 5/33	1	1	1

られている。こうした項目とアトリビュートとの関係についての情報をモデルに取り入れる際には、行が項目、列がアトリビュートで表された行列を用いる。この行列は Q 行列と呼ばれ、CDM において重要な役割を担っている。表 1 に算数に関する項目の Q 行列の例を示した。この Q 行列から、1 行目の項目では、正答のためには足し算のアトリビュートが必要であると設定されていることがわかる。

CDM では、回答者が各アトリビュートを修得しているか否かについての理論的にあり得る全ての組み合わせを、アトリビュート修得パターン行列 (attribute mastery pattern matrix; 山口, 2019; Yamaguchi & Okada, 2020) と呼ばれる行列によって表現する。このアトリビュート修得パターン行列は、Q 行列で設定されたアトリビュートに基づいて構成される。表 2 に、表 1 と同じアトリビュートについてのアトリビュート修得パターン行列を示した。各回答者は、この行列における行が表すパタンのどれか 1 つをもつことになるため、この行列の各行がアトリビュート修得パターンと呼ばれる。

一般的に Q 行列は、測りたい内容領域の専門家が項目の内容から設定するなど、人の手によって予め定められる (専門家による Q 行列の作成例として、Jang, Kim, Vincett, Barron, & Russell, 2019)¹⁾。そのため、Q 行列の設定には主観的な作業が含まれ、専門家の間でも意見の食い違いが生じたり、誤った設定をしてしまったりする可能性は十分考えられる。そこで、Q 行列の誤設定をした際の診断の正確度への影響をシミュレーションにより確認する研究 (Rupp & Templin, 2008) や、Q 行列の誤設定をデータへの当てはまりを示す複数のモデル評価指標により発見する研究 (Sunbul & Asiret, 2018)、Q 行列の妥当性を測る指標を作成する研究 (de la Torre, 2008; de la Torre

表 2. アトリビュート修得パターン行列の例

パタン	足し算	掛け算	通分
1	0	0	0
2	1	0	0
3	0	1	0
4	0	0	1
5	1	1	0
6	0	1	1
7	1	0	1
8	1	1	1

& Chiu, 2016) などがこれまで行われてきている。これらの先行研究から得られた知見として、Q 行列の設定に誤りがある場合、アトリビュートの診断の正確度の低下や項目パラメータ推定値のバイアスにつながるということがわかっている。もしフィードバックされた情報にバイアスがある場合、回答者は誤りが含まれる情報をその後の学習に活かすこととなってしまうかねない。こうしたことから、現実のテスト開発において Q 行列の誤設定は深刻な問題である。

1.2. アトリビュートの階層性

表 2 のようなアトリビュート修得パターン行列を設定するだけでは、アトリビュート間の関係性はモデル上考慮されていない。一方で、アトリビュート間の関係性を明示的にモデルに取り入れられれば、アトリビュート間の関係の大きさを解釈できること、節約的なモデリングが可能になることといった利点が生じる。このとき、どのようにアトリビュート間の関係性をモデルに取り入れるとよいかということも CDM の応用上重要な検討事項の 1 つである。

とくにアトリビュート間に従属関係があると仮定できる場合には、アトリビュートの階層性 (attribute hierarchy) をモデルに取り入れることができる (Leighton, Gierl, & Hunka, 2004)。ここでいう階層性とは、複数の異なるアトリビュートが特定の順序性をもって積みあげられることで構造を成している状態のことである。階層性を Q 行列とアトリビュート修得パターン行列に取り入れることには、理論的にあり得ないアトリビュート修得パターンを予め除けるという利点がある。たとえば表 2 において、通分という

¹⁾ 近年、Q 行列も推定する方法の開発が盛んに行われてきている (Liu & Kang, 2019)。

アトリビュートを修得するためには、通分を行う際には必ず掛け算を用いることから、掛け算というアトリビュートを修得していることが必要条件だと考えることができる。このアトリビュートの階層性を考慮すると、表 2 のアトリビュート修得パターン行列から、「掛け算が未修得であるが通分が修得されている」という理論的・現実的にあり得ないパターンを予め除くことができる。これによってパターン 4, 7 は除外され、パターン数が 8 から 6 へと減る。Q 行列においても同様に、掛け算というアトリビュートを測らないにもかかわらず通分というアトリビュートを測るとするパターンは予め除外できる。

また、階層性を表す構造のタイプごとに診断の正確度が異なることが先行研究では示されている。山口 (2017) は階層性のタイプによって、Q 行列の誤設定による影響が異なることをシミュレーション研究で示した。こうしたことから、Q 行列の誤設定が診断の正確度に及ぼす影響について調べる際には、階層性のタイプごとに検討する必要があると考えられる。

1.3. 多値アトリビュートの Q 行列

表 1 の例で示したように、CDM においてアトリビュートの修得は二値で表現されることが多い。この仮定は、ある回答者がアトリビュートを修得しているか否かの診断のためにテストを運用する場合を想定している。しかし、1 つのアトリビュートに対して、その修得の程度に対応する多値のレベルを考えることが妥当な場面も少なからず存在する。たとえば表 1 の Q 行列において、測っているアトリビュートが同じである項目 3 と項目 4 を考える。この 2 つの項目は、桁数や通分の際の手続きの数が違うことから、項目 4 の方が項目 3 よりもレベルの高い問題と考えることが自然である。具体的な考え方の例として、通分に関して述べる。項目 3 では分母の最小公倍数が第 1 項の分母の値 (10) であるため、回答者は第 2 項だけ変換すれば通分を行うことができる。一方で、項目 4 では、分母の最小公倍数が第 1 項の分母 (21) と第 2 項の分母 (33) と異なるので、両方の項について分母を揃える変換を行う必要がある。この両項目の間の違いが正答確率に、誤差では説明できないほど大きな違いをもたらし、その違いが診断されることが望ましい場合には、通分のアトリビュートを 3 段階に分けることによって、Q 行列上でその違いを区別できるようになる。つまり、多値アトリビュートによって同一アトリ

表 3. 多値アトリビュートにより項目 3 と項目 4 のレベルの違いを取り入れた Q 行列の例

項目	足し算	掛け算	通分
32+15	1	0	0
9×8	0	1	0
2/10 + 1/2	1	1	1
4/21 + 5/33	1	2	2

ビュート内のレベルを設定すれば、項目 3 と項目 4 の困難度の違いが表現可能となる。そうした Q 行列の具体例を表 3 に示した。このようなアトリビュートの設定が診断の目的にかなったものであり、対象母集団の回答者が実際にそのような反応プロセスにより回答するのであれば、「項目 3 に正答し項目 4 には誤答する回答者」と「項目 3 と項目 4 の両方に正答する回答者」のアトリビュート修得レベルの違いを識別できるモデルは、これらを構造上識別できないモデルよりも、より適切だと考えられ、さらに実用上の価値もあるだろう。実際に Chen & de la Torre (2013) は、アトリビュートに 3 段階のレベルがあると想定される状況において、多値アトリビュートを仮定したモデルの方が二値アトリビュートを仮定したモデルよりも AIC, BIC, 診断の正確度において優れていたことを報告している。

しかし、多値アトリビュートを仮定した際の Q 行列の誤設定が診断の正確度に及ぼす影響については、著者らの知る限りではほとんど調べられていない。さらに、階層性を考慮した場合における多値アトリビュートの Q 行列の誤設定の影響に関しては全く調べられていない状況にある。階層性をもつ多値アトリビュートを考えるモデルは、先の例のように、① 1 つのテストで測定するアトリビュート間に何らかの従属関係があり、②測定するアトリビュートのレベルによって困難度が異なることが想定される場合には、応用領域を問わず利用することができる。したがって、様々な領域におけるこのような CDA に対して、本研究で扱う種類の CDM の応用可能性は大きいと考えられる。

多値アトリビュートにおける Q 行列は、二値アトリビュートの場合よりも、とりうる値が増える。たとえばアトリビュートの修得レベルが 0, 1, 2 の 3 段階で表されるとき、真値が 0 である Q 行列要素についてテスト開発者が設定する値は、正しく 0 と設定する場

合と誤って1と設定する場合だけでなく、誤って2と設定する場合も含む3ボタンが考えられる。この状況において、0を1に誤設定する場合よりも、0を2に誤設定する場合の方が、もし大幅に診断の正確度が悪くなるのであれば、多値アトリビュートのモデルを使用することはリスクが高すぎるかもしれない。なぜなら、生じる可能性があるテスト開発者が想定している程度を超えたバイアスが、回答者にフィードバックする情報に含まれてしまうかもしれないからである。こうした懸念から、多値アトリビュートのCDMにおいてもQ行列の誤設定の影響を調べる必要があると考えられる。

そこで本研究では、複数の階層性タイプのQ行列において多値アトリビュートを誤設定してしまった場合の、診断の正確度およびパラメタ推定への影響を調べる。特に、Q行列要素の真値と誤設定値の差が大きくなるほど診断の正確度が悪くなることが考えられるため、真値と誤設定値の差の大きさが診断の正確度へ及ぼす影響の程度を知ることを主な目的とする。また、こうした影響の程度が3節で説明するようなQ行列の階層性によりどのように異なるのかを知ることを2つめの目的とする。

本論文の構成は次の通りである。2節では本研究で用いる多値アトリビュートのCDMについて、3節では多値アトリビュートにおける階層性について述べる。4節ではQ行列の誤設定の影響を調べるためのシミュレーションの方法を述べ、5節ではその結果と考察を示す。最後に6節では総合考察と、将来の研究についての議論を行う。

2. 多値アトリビュートのCDM

アトリビュートの多段階の修得レベルをCDMにおける多値アトリビュートとして反映するための一般的な枠組みとして、OCACフレームワーク (ordered category attribute coding framework) が考案されている (Karelitz, 2004)。本研究では、OCACフレームワークに基づいて開発されたCDMの儉約的な拡張である、RPa-DINA (reparameterized polytomous attribute deterministic-input, noisy and gate; Zhan, Bian, & Wang, 2016; Zhan, Jiao, & Man, 2019) モデルを用いる。このモデルは次のように表現される。

CDMは、観測データの背後にカテゴリカルな潜在変数を考える、潜在クラスモデルの特殊な場合とし

て表現できる (e.g., 山口・岡田, 2017)。このとき、個々の潜在クラスは、それぞれ異なるアトリビュートの修得パターンに対応する。潜在クラスの具体的な表現および項目との関係は、以下で述べるようになる。まず、診断の対象となるアトリビュート数を K とし、可能なアトリビュート修得パターン数を C とする。また、アトリビュート k ($k = 1, 2, \dots, K$) の最大レベルを M_k とする。全てのアトリビュート修得パターンで構成される行列であるアトリビュート修得パターン行列を $\mathbf{A}$ とする。これは行がクラス c に対応し、 $\mathbf{A} = [\alpha_1, \dots, \alpha_c, \dots, \alpha_C]'$ である所与の行列である。また、ここでクラス c の K 個のアトリビュートについての修得パターンは $\alpha_c = [\alpha_{c1}, \dots, \alpha_{ck}, \dots, \alpha_{cK}]'$ であり、その要素は $\alpha_{ck} \in \{0, 1, \dots, m_k, \dots, M_k\}$ であるとする。つまり、 $\alpha_{ck} = m_k$ ならば、クラス c に属する回答者は、アトリビュート k をレベル m_k で修得している。本研究で扱うモデルでは、クラス数はアトリビュート数だけでなく、アトリビュートのレベル数やアトリビュートの階層性のタイプにも依存して決まる。項目とアトリビュートとの関係を表すQ行列は、要素 $q_{jk} \in \{0, 1, \dots, m_k, \dots, M_k\}$ において、1以上の値なら項目 j がアトリビュート k のその値のレベル m_k を測定することを、0ならアトリビュート k を測定しないことを意味する。項目 j についての要素をまとめて $\mathbf{q}_j = [q_{j1}, \dots, q_{jk}, \dots, q_{jK}]'$ とし、サイズ $J \times K$ のQ行列を $\mathbf{Q} = [\mathbf{q}_1, \dots, \mathbf{q}_j, \dots, \mathbf{q}_J]'$ で表す。また、クラス c に所属する回答者が項目 j に正答するために必要なアトリビュートについて、全て要求されるレベル以上である場合に1、そうでない場合に0となる指標

$$\eta_{cj} = \prod_{k=1}^K I(\alpha_{ck} \geq q_{jk})^{I(q_{jk} > 0)}, \quad (1)$$

を導入する。ここで、指標関数 $I(\cdot)$ は括弧内の条件式が正しければ1を、正しくなければ0を返す。この η_{cj} は、項目反応に誤差がない理想的な状況下で、クラス c に所属する回答者が項目 j に対して $\eta_{cj} = 1$ をとれば正答し、0をとれば誤答することを意味する。なお、これはクラス c と項目 j が所与のとき、アトリビュート修得パターン行列 $\mathbf{A}$ とQ行列 $\mathbf{Q}$ で設定された値から決まる定数である。そして、クラス c に所属する回答者が項目 j に正答する確率を

$$p_{cj} = g_j + (1 - s_j - g_j) \eta_{cj} \quad (2)$$

とする。ただし、 g_j は guessing パラメタとし、項目 j へ正答するために必要なアトリビュートをもたないにもかかわらずその項目に誤答する確率を、また s_j は slipping パラメタとし、項目 j に正答するために必要なアトリビュートをもっているにもかかわらずその項目に誤答する確率を表す項目パラメタである。これらをまとめて、項目パラメタベクトルを $\omega_j = (g_j, s_j)'$, $j = 1, 2, \dots, J$ とする。

次に、個人 i のクラス c への所属を表す潜在変数

$$z_{ic} = \begin{cases} 0, & \text{回答者 } i \text{ がクラス } c \text{ に所属しない場合.} \\ 1, & \text{回答者 } i \text{ がクラス } c \text{ に所属する場合.} \end{cases} \quad (3)$$

を導入する (山口, 2019)。また、回答者 i の項目 j への正誤反応を表す観測データを $Y_{ij} \in \{0, 1\}$, $i = 1, 2, \dots, N$ で表す。ただし、1 は正答、0 は誤答に対応する。潜在変数 z_{ic} のクラスをまとめたベクトルを $\mathbf{z}_i = [z_{i1}, \dots, z_{ic}, \dots, z_{iC}]'$ で表すとき、項目反応確率は

$$P(Y_{ij}|\omega_j, \mathbf{z}_i) = \prod_{c=1}^C \left[p_{cj}^{Y_{ij}} (1 - p_{cj})^{1-Y_{ij}} \right]^{z_{ic}} \quad (4)$$

とモデル化することができる。潜在変数 z_{ic} で条件付けることで各項目への反応が独立となることを仮定すると、項目反応をまとめたベクトル $\mathbf{Y}_i (= [Y_{i1}, \dots, Y_{ij}, \dots, Y_{iJ}]')$ を得る確率は

$$P(\mathbf{Y}_i|\Omega, \mathbf{z}_i) = \prod_{j=1}^J P(Y_{ij}|\omega_j, \mathbf{z}_i) \quad (5)$$

となる。ただし、ここで $\Omega = [\omega_1, \dots, \omega_j, \dots, \omega_J]'$ である。

さらに、 $\mathbf{z}_i$ の要素の母集団の比率であるアトリビュート修得パターン行列の混合比率を表すパラメタを π_c , $c = 1, 2, \dots, C$ とする (ただし、 $\sum_{c=1}^C \pi_c = 1$)。この分布はカテゴリカル分布

$$P(\mathbf{z}_i|\pi) = \prod_{c=1}^C \pi_c^{z_{ic}} \quad (6)$$

によって表すことができる。ここで $\pi = [\pi_1, \dots, \pi_c, \dots, \pi_C]'$ である。式 (4), (5), (6) から、観測変数である項目反応 $\mathbf{Y}_i$ と潜在変数 $\mathbf{z}_i$ の同時確率は

$$\begin{aligned} P(\mathbf{Y}_i, \mathbf{z}_i|\Omega, \pi) &= P(\mathbf{Y}_i|\Omega, \mathbf{z}_i) P(\mathbf{z}_i|\pi) \\ &= \prod_{c=1}^C \left[\prod_{j=1}^J p_{cj}^{Y_{ij}} (1 - p_{cj})^{1-Y_{ij}} \right]^{z_{ic}} \times \pi_c^{z_{ic}} \\ &= \prod_{c=1}^C \left[\pi_c \prod_{j=1}^J \left\{ p_{cj}^{Y_{ij}} (1 - p_{cj})^{1-Y_{ij}} \right\} \right]^{z_{ic}} \end{aligned} \quad (7)$$

であり、回答者の交換可能性の仮定のもとで、項目反応行列 $\mathbf{Y}$ と潜在変数行列 $\mathbf{Z}$ からなる完全データの尤度、

$$P(\mathbf{Y}, \mathbf{Z}|\Omega, \pi) = \prod_{i=1}^N P(\mathbf{Y}_i, \mathbf{z}_i|\Omega, \pi) \quad (8)$$

となる。

実際にデータを分析する上では、各回答者のクラスへの所属 $\mathbf{z}_i$ は未知である。このときの項目反応関数は

$$P(Y_{ij}|\omega_j, \pi_c) = \sum_{c=1}^C \pi_c \prod_{j=1}^J p_{cj}^{Y_{ij}} (1 - p_{cj})^{1-Y_{ij}} \quad (9)$$

という有限混合モデルによって表現される。これらの式に基づき、パラメタの最尤推定値が欲しい場合は DINA モデルと同様の EM アルゴリズム (たとえば、山口, 2019) を、事後確率が欲しい場合は適当な事前分布を用いて MCMC 推定 (たとえば、Zhan, Jiao, & Man, 2019) を行えば、項目パラメタ Ω と混合比率パラメタ π を推定できる。また、これらの方法により潜在変数 z_{ic} の事後確率を得る場合、(9) 式から

$$P(z_{ic}|\mathbf{Y}_i, \Omega, \pi) = \frac{\pi_c \prod_{j=1}^J p_{cj}^{Y_{ij}} (1 - p_{cj})^{1-Y_{ij}}}{\sum_{c=1}^C \pi_c \prod_{j=1}^J p_{cj}^{Y_{ij}} (1 - p_{cj})^{1-Y_{ij}}} \quad (10)$$

と求めることができる。各回答者のアトリビュート修得状況の推定のためには、 $z_{ic} = 1$ となるクラス c に対応するアトリビュート修得パターン行列 $\mathbf{A}$ の行 c において、アトリビュート k の値が示すレベルが回答者 i によって修得されているかどうかを表す変数 $\alpha_{ik} \in \{0, 1, \dots, m_k, \dots, M_k\}$ を導入する。これをアトリビュート修得状況パラメタと呼ぶこととする。MCMC 推定の場合、 α_{ik} を事後分布からサンプリングし、その代表値を各回答者についてのアトリビュート修得状況の点推定値とすることができる。なお、本研究では MCMC 法で推定を行った。

図 1. 本研究における直線型, 分岐型, 収束型におけるアトリビュートの階層性の模式図
※ A, B, C はそれぞれアトリビュートを表し, 数字はレベルを表す.

3. 多値アトリビュートにおける階層性

一般的なアトリビュート階層性のタイプは 4 つあり, それぞれ, 非構造 (unstructured) 型, 直線 (linear) 型, 分岐 (divergent) 型, 収束 (convergent) 型と呼ばれる (Leighton, Gierl, & Hunka, 2004; 実際のテストでの適用例として Tu, Wang, Cai, Douglas, & Chang, 2019). これらの階層性は Q 行列とアトリビュート修得パターン行列に組み込まれるが, 階層性のタイプと Q 行列が決まればアトリビュート修得パターン行列は自然に決まる (階層性を組み込んだ Q 行列の例としては, 4 節で示す表 4 を参照; 二値のアトリビュートの場合は山口, 2017 を参照). 以下では各階層性タイプについて説明する. 非構造型はその名の通り, アトリビュート間に従属関係がない場合である. 直線型は, 従属関係にあるアトリビュートが一直線上に連なっている場合である. 分岐型は, あるアトリビュートを修得すると, 次に修得可能となるアトリビュートが複数存在する場合である. そして収束型は, 複数のアトリビュートのうちどれか 1 つでも修得すれば次に 1 つのアトリビュートが修得可能となる場合である.

多値アトリビュートの場合, これら 4 つの階層性タイプに加えて, いくつかの設定の仕方が考えられる. 例として, アトリビュートが A, B, C の 3 つあり, アトリビュートの修得レベルが 0, 1, 2 の 3 段階ある状況を考える. この状況において, 1 つめの階層

から 2 つめの階層に上がるための条件として, 1 つめの階層のレベルを 2 まで完全に修得している必要があるか, それとも 1 まで修得していればよいかという 2 つの設定方法がありうる. こうした設定は, テストや項目の目的などによって決められるべきであり, 両方の設定における誤設定の影響が調べられることが望ましい. しかし, 本研究においては, 実験条件が複雑になりすぎて解釈が困難になることを避けるため, 階層を上げるためには 1 つ前の階層のアトリビュートの修得レベルが最大であることが必要であるという設定のもとで誤設定の影響を調べる. この設定における多値アトリビュートの直線型, 分岐型, 収束型の階層性の模式図を図 1 に示した.

4. シミュレーションの設定

Q 行列の誤設定の診断の正確度への影響を調べるための一般的なシミュレーションは, 以下のような手順で行われる. まず真の Q 行列とパラメタの真値からモデルを通してデータを生成し, そのデータに誤った設定の Q 行列を用いたモデルを適用することで各パラメタを推定する. これを複数回繰り返し, 得られたアトリビュート修得状況パラメタ α_{ik} の推定値が, 真値をどれほどリカバリできているかを評価するための指標を算出する. そしてこの評価指標を複数の条件間で比較する.

本研究において比較する条件は, 階層性について非

表 4. 各階層構造における Q 行列とその誤設定パターン

非構造型				直線型				収束型			
項目	A	B	C	項目	A	B	C	項目	A	B	C
1	0	0	1	1	1	0	0	1	1	$\begin{smallmatrix} 0 \rightarrow 1 (1) \\ 0 \rightarrow 2 (2) \end{smallmatrix}$	0
2	0	0	2	2	2	$\begin{smallmatrix} 0 \rightarrow 1 (1) \\ 0 \rightarrow 2 (2) \end{smallmatrix}$	0	2	2	0	$\begin{smallmatrix} 0 \rightarrow 1 (3) \\ 0 \rightarrow 2 (4) \end{smallmatrix}$
3	0	1	0	3	2	1	0	3	0	1	0
4	0	1	1	4	2	2	$\begin{smallmatrix} 0 \rightarrow 1 (3) \\ 0 \rightarrow 2 (4) \end{smallmatrix}$	4	0	2	0
5	0	1	2	5	2	2	1	5	1	1	0
6	0	2	0	6	2	2	2	6	1	2	0
7	0	2	1					7	2	1	0
8	0	2	2					8	2	2	0
9	1	0	0	分岐型				9	2	0	1
10	1	0	1	項目	A	B	C	10	2	0	2
11	1	0	2	1	1	0	0	11	0	2	1
12	1	1	0	2	2	$\begin{smallmatrix} 0 \rightarrow 1 (1) \\ 0 \rightarrow 2 (2) \end{smallmatrix}$	$\begin{smallmatrix} 0 \rightarrow 1 (3) \\ 0 \rightarrow 2 (4) \end{smallmatrix}$	12	0	2	2
13	1	1	1	3	2	1	0	13	2	1	1
14	1	1	2	4	2	2	0	14	2	1	2
15	1	2	0	5	2	0	1	15	1	2	1
16	1	2	1	6	2	0	2	16	1	2	2
17	1	2	2	7	2	1	1	17	2	2	1
18	2	$\begin{smallmatrix} 0 \rightarrow 1 (1) \\ 0 \rightarrow 2 (2) \end{smallmatrix}$	$\begin{smallmatrix} 0 \rightarrow 1 (3) \\ 0 \rightarrow 2 (4) \end{smallmatrix}$	8	2	2	1	18	2	2	2
19	2	0	1	9	2	1	2				
20	2	0	2	10	2	2	2				
21	2	1	0								
22	2	1	1								
23	2	1	2								
24	2	2	0								
25	2	2	1								
26	2	2	2								

※(1)と(3)が 0 を 1 に、(2)と(4)が 0 を 2 に誤設定したパターン

構造型、直線型、分岐型、収束型の 4 条件、項目セット数について 1 セットと 2 セットの 2 条件、Q 行列の誤設定条件については、のちに説明する 5 パターンを用意し、計 40 条件とした。ここでの項目セットとは、各階層性において項目が測定可能な全てのアトリビュートのパターンを備える Q 行列の項目群である。各階層性での 1 セット条件における Q 行列を、表 4 に誤設定する要素と共に表した。なお、アトリビュート修得の最大レベルについては、全条件で $M_k = 2$ とした。Q 行列の誤設定条件については、階層性に従い、真値が 0 の要素を 1 に誤設定するパターン（表 4 の (1) と (3)）と 2 に誤設定するパターン（表 4 の (2) と (4)）を、各階

層性条件で 2 つのアトリビュートに対して設定することで 4 パターンとし、さらに誤設定がなく真値と等しい設定の場合を加えて、計 5 パターンとした。なお、誤設定の条件を考えるにあたり、Sunbul & Asiret (2018) は 20%（15 項目中 3 項目）などと誤設定の割合を操作している。こうした設定は一つの方法であるが、誤設定 1 つにつきどれほどの影響があるのかを知るためには必ずしも適していない。本研究では、個々の誤設定の影響について調べるために、1 セットの項目数につき 1 つの要素にのみ誤設定があるという状況を考えることで、どれほど診断の正確度が悪くなるのかを検討した。

アトリビュート修得状況パラメタ α_{ik} の診断の正確度の評価指標としては、 r 回目の反復において回答者 i の k 番目のアトリビュートにおける事後平均値を四捨五入した整数値 $\hat{\alpha}_{ik}$ と真値 α_{ik} の一致率 AR_r (agreement rate) と、それを反復回数 R について平均した値である ACA (attribute classification accuracy)、反復回数ごとの一致率 AR_r のばらつきを知るための標準偏差 SD_{AR} 、そして今回のシミュレーションで見られた一致率の範囲を知るための最小値と最大値を算出した。 r 回目の反復における一致率 AR_r 、全反復回数による平均 ACA と標準偏差 SD_{AR} はそれぞれ、

$$AR_r = \frac{1}{NK} \sum_{i=1}^N \sum_{k=1}^K I\{\hat{\alpha}_{ik} = \alpha_{ik}\}, \quad (11)$$

$$ACA = \frac{1}{R} \sum_{r=1}^R AR_r, \quad (12)$$

$$SD_{AR} = \sqrt{\frac{1}{R} \sum_{r=1}^R (AR_r - ACA)^2} \quad (13)$$

と求められる。ここで、 $I\{\hat{\alpha}_{ik} = \alpha_{ik}\}$ は、回答者 i のアトリビュート k におけるアトリビュート修得状況パラメタの真値と推定値が一致していれば 1 を、一致していなければ 0 をとる指標変数である。

さらに詳しく診断の正確度を調べるため、アトリビュートごとのアトリビュート修得状況パラメタ α_k 、項目パラメタ s_j と g_j 、混合比率パラメタ π_c のバイアスと RMSE (root mean square error) を算出した。例として、項目 j の guessing パラメタ g_j のバイアスと RMSE は以下のように表される：

$$Bias = \frac{1}{R} \sum_{r=1}^R (\hat{g}_{jr} - g_j), \quad (14)$$

$$RMSE = \sqrt{\frac{1}{R} \sum_{r=1}^R (\hat{g}_{jr} - g_j)^2}. \quad (15)$$

項目パラメタの真値は $s_j = g_j = 0.2$ とし、混合比率パラメタ π_c の真値はパラメタが全て 1 のディリクレ分布から生成した。全条件でサンプルサイズ $N = 500$ 、アトリビュート数 $K = 3$ 、アトリビュート最大レベル $M_k = 2$ とした。二値アトリビュートの先行研究である山口 (2017) などから、本研究で最大の関心の対象であるアトリビュート修得状況パラメ

タ α_{ik} のパラメタリカバリにはサンプルサイズの影響があまり見られないことがわかっていて、本研究で使用するモデルでも同様に、アトリビュート修得状況パラメタ α_{ik} のパラメタリカバリにはサンプルサイズの影響があまり見られないことが確認されたため山口 (2017) が設定した最小のサンプルサイズである 500 に固定した。また、この 500 人の回答データは、2 節 (4) 式から乱数発生させた。シミュレーションは R (R Core Team, 2019) のパッケージ R2jags (Version 0.5-7; Su & Yajima, 2015) と JAGS (Plummer, 2017) により行った。MCMC の設定はイテレーションが 8000 回、バーンインが 4000 回、チェーン数 2 とし、シミュレーションの反復回数は $R = 100$ とした。また、推定の際の項目パラメタ s_j と g_j の事前分布は Beta(1,1)、混合比率パラメタ π_c の事前分布はパラメタが全て 1 のディリクレ分布とした。これらは各台上の一樣分布に対応する。

本研究で用いた R コードは <https://osf.io/e7pyz/> で公開する。次節で結果を示すが、紙幅の制限により本論文に載せられなかった結果は、電子付録として上記 URL であわせて掲載した。

5. 結果と考察

5.1. アトリビュート修得状況パラメタの診断の正確度

各条件における ACA と ACA の差を表 5 に示した。この表での ACA の差は、「真-(1)」と「真-(3)」の行が「真の Q 行列での ACA」と「0 を 1 に誤設定した Q 行列での ACA」との差、そして“(1)-(2)”と“(3)-(4)”の行が「0 を 1 に誤設定した Q 行列での ACA」と「0 を 2 に誤設定した Q 行列での ACA」との差である。もし前者の差と後者の差が同じ程度か、前者の差の方が大きければレベル数が多くなっても誤設定による影響の程度は予測しやすいと考えられる。なぜならアトリビュートのレベルを 1 つ誤設定するごとに、同じ程度かそれよりも小さい程度で診断の正確度も低くなることを示唆するからである。しかし、もし前者の差よりも後者の差がはるかに大きくなっているれば、誤設定の程度が大きくなるほど診断の正確度の悪化の程度が飛躍的に大きくなることを示唆し、テスト設計においても誤設定の影響の予測がしにくくなってしまうことが考えられる。

表 5 から、予想通り全ての階層性タイプにおいて、

表 5. ACA と Q 行列の誤設定パターンにおける ACA の差

非構造型						直線型					
ACA (%)			ACAの差			ACA (%)			ACAの差		
1セット	2セット		差	1セット	2セット	1セット	2セット		差	1セット	2セット
真	81.15	91.54				真	75.78	87.74			
(1)	79.70	90.53	真-(1)	1.45	1.01	(1)	74.83	85.54	真-(1)	0.95	2.20
(2)	79.52	89.92	(1)-(2)	0.18	0.61	(2)	72.54	82.25	(1)-(2)	2.28	5.50
(3)	79.90	90.30	真-(3)	1.25	1.24	(3)	74.16	83.94	真-(3)	1.61	3.80
(4)	79.28	89.78	(3)-(4)	0.62	0.52	(4)	71.58	79.76	(3)-(4)	2.59	4.18

分岐型						収束型					
ACA (%)			ACAの差			ACA (%)			ACAの差		
1セット	2セット		差	1セット	2セット	1セット	2セット		差	1セット	2セット
真	78.04	88.48				真	77.52	89.58			
(1)	75.06	85.88	真-(1)	2.97	2.60	(1)	75.42	87.33	真-(1)	2.10	2.25
(2)	75.27	84.67	(1)-(2)	-0.20	1.21	(2)	74.63	87.40	(1)-(2)	0.79	-0.07
(3)	75.17	86.32	真-(3)	2.87	2.16	(3)	76.29	87.68	真-(3)	1.23	1.91
(4)	74.90	84.65	(3)-(4)	0.26	1.67	(4)	74.03	86.03	(3)-(4)	2.26	1.65

真の Q 行列条件よりも誤設定要素がある Q 行列条件の方で ACA が低かった。ACA が低いほど平均的な診断の正確度が悪くなっていることを表すため、この結果から、誤設定された Q 行列要素が 1 つでもあることで診断の正確度が平均的に悪化することがわかる。また、項目数が 2 セットの場合に、1 セットの場合よりも ACA が全体的に高かった。これにより、単純に項目数が 2 倍になることで、診断の正確度が上がることがわかる。しかし、0→1 の誤設定と 0→2 の誤設定における ACA の違いに関しては階層性ごと、項目セット数ごとに異なる結果が得られた。

まず非構造型では、“真-(1)”と“真-(2)”の行からわかるように、両セット数条件で 0→1 の誤設定パターンでは真の設定パターンより 1 から 1.5 ポイントの間で ACA が低下しており、0→2 の誤設定パターンでは 0→1 の誤設定パターンより 0 から 1 ポイントの間で ACA が低下していた。つまり、真値と誤設定値との差が大きくなるほど、診断の正確度への影響は大きくなるものの、その影響の程度が飛躍的に大きくなるといったことは見られなかった。

次に直線型では、非構造型と同様に両セット数条件で一貫して、0→1 の誤設定より 0→2 の誤設定の方が真の Q 行列条件の ACA からより大きく低下していた。しかし、その低下の程度に関しては、「アトリビュートレベルを誤設定しなかった場合と 1 段階だけ

過大設定した場合の ACA の差」である“真-(1)”や“真-(3)”よりも、「1 段階過大設定した場合と 2 段階過大設定した場合の ACA の差」である“(1)-(2)”や“(3)-(4)”がより大きかった。このことから、1 段階ごとの誤設定の影響の程度が異なることにより、実際に Q 行列を作成する際には、診断の正確度が想定しにくくなることが考えられる。

また、項目数が 2 セットの場合には、1 セットの場合よりも ACA の値は高くなっていたものの、誤設定による診断の正確度への影響は全ての誤設定パターンで大きかった。これは、項目数が多くなることで、全体的な診断の正確度は上がっているものの、誤設定の影響も大きくなっていることを意味する。1 つのアトリビュートに含まれる誤設定項目の割合は変わっていないにもかかわらず、診断の正確度への影響が大きくなるということは、単純に複数の誤設定項目が含まれることが診断の正確度を大きく下げってしまう要因となることを示唆する。あるいは、後にも述べるように、アトリビュートのレベルを正しく測る項目数と誤って測る項目数が同じ場合に、項目数を単純に 2 倍に増やすことが、診断の正確度を大きく下げってしまう要因となることが示唆される。どちらの場合であっても、内容領域によっては、実際のテストで誤設定項目が複数あることも決して珍しくはないと考えられる。たとえ項目数を増やして全体的な診断の正確度を上げようとし

表 6. 一致率 AR の標準偏差と最小値と最大値

非構造型							直線型						
	1セット			2セット				1セット			2セット		
	SD	MIN (%)	MAX (%)	SD	MIN (%)	MAX (%)		SD	MIN (%)	MAX (%)	SD	MIN (%)	MAX (%)
真	2.41	74.87	87.73	1.52	87.67	95.53	真	4.01	60.87	87.20	1.94	83.13	94.40
(1)	2.85	71.87	85.53	1.90	85.53	94.20	(1)	4.31	64.47	83.73	3.57	72.07	91.87
(2)	2.44	73.93	86.20	1.96	84.73	94.47	(2)	5.56	55.20	82.87	4.93	53.60	90.20
(3)	2.93	70.67	86.20	2.05	81.87	94.67	(3)	4.52	58.07	86.67	4.74	69.73	95.53
(4)	2.49	72.80	84.60	2.23	82.47	94.00	(4)	4.60	58.73	85.47	5.89	61.73	92.80

分岐型							収束型						
	1セット			2セット				1セット			2セット		
	SD	MIN (%)	MAX (%)	SD	MIN (%)	MAX (%)		SD	MIN (%)	MAX (%)	SD	MIN (%)	MAX (%)
真	2.40	70.33	82.87	1.81	84.27	93.00	真	2.54	68.07	86.47	1.65	83.60	92.93
(1)	3.52	61.67	81.60	2.77	78.73	92.47	(1)	3.31	66.33	83.53	2.55	80.20	92.20
(2)	3.90	62.80	86.27	3.42	72.40	91.80	(2)	3.54	66.73	82.33	2.54	80.40	93.07
(3)	3.66	64.27	81.73	2.82	80.27	93.27	(3)	3.47	66.87	84.67	2.44	79.87	92.00
(4)	4.75	60.80	85.10	3.72	71.00	91.33	(4)	3.37	65.47	82.33	3.15	76.47	92.53

ても、それにより誤設定項目の数も増加する場合、その分だけ診断の正確度が下がってしまうことを念頭に置くべきだと思われる。このことに関しては、総合考察で再び議論する。

分岐型では、大半の条件で 0→1 の誤設定より 0→2 の誤設定の方が真の Q 行列条件の ACA からより大きく低下していた。しかし、1 セット条件ではアトリビュート B において“(1)–(2)”が負の値となった。これは誤設定の程度が大きい方が、小さい場合よりも、診断の正確度への影響がより小さいことを意味する。また、1 セット条件と 2 セット条件の比較においては、“真–(1)”と“真–(3)”では項目数が増えると診断の正確度の低下の程度は小さくなる一方で、“(1)–(2)”と“(3)–(4)”では項目数が増えると診断の正確度の低下の程度は大きくなった。この結果から、誤設定の程度が大きい場合は、誤設定の程度が小さい場合よりも、誤設定項目が 2 倍に増えたことによる影響がより大きくなったと考えられる。

最後に収束型では、分岐型同様に、大半の条件で 0→1 の誤設定より 0→2 の誤設定の方が真の Q 行列条件の ACA からより大きく低下していたものの、2 セット条件ではアトリビュート B において“(1)–(2)”が負の値となった。また、1 セット条件と 2 セット条件の比較においては、分岐型とは対照的に、“真–(1)”と“真–(3)”では項目数が増えると診断の正確度の低下の程度は小さくなる一方で、“(1)–(2)”と“(3)–(4)”では項目数が増えると診断の正確度の低下の程度は小

さくなった。これは、表 4 からわかるように、収束型の Q 行列では、アトリビュート B においてレベル 2 を測る項目数がアトリビュート C においてレベル 2 を測る項目数よりも多い一方で、レベル 1 を測る項目数はどちらのアトリビュートにおいても同じであることが原因だと考えられる。レベル 2 を測る項目数がアトリビュート B より C で少ないことは、1 セット条件において“真–(3)”より“(3)–(4)”の値の方が高かったことにも影響している可能性がある。

以上の結果から、現実のテストにおいても、特に直線型では、項目に設定するアトリビュートのレベルをより過大に誤設定するほど、診断の正確度がより大きく低下することが考えられる。しかし、ACA は平均的な影響であり、影響の程度にもばらつきが見られた。そこで一致率のばらつきとその範囲についての結果を次の項で示す。

5.2. 一致率のばらつきとその範囲

表 6 に反復回数ごとの一致率 AR_r のばらつきを評価するための標準偏差、そして R 回の反復における最小値と最大値を条件ごとに示した。

全体的な傾向として、真の Q 行列条件より誤設定条件で AR_r の標準偏差が大きかった。これは、同じ Q 行列であっても、データ生成と推定の際の確率的変動により、推定されるアトリビュート修得状況パラメタの値に変動があること、またその変動は誤設定がある場合に誤設定がない場合よりも大きくなることを意

味する。

さらに、真の Q 行列条件において、どの階層性タイプでも、項目数が 1 セットよりも 2 セットの方で標準偏差が小さかった。この傾向は誤設定のある Q 行列であっても大半の条件で同じだったが、直線型の階層性タイプでのみ例外が見られた。より詳しく述べると、アトリビュート C の Q 行列要素を誤設定したパターンである (3) と (4) においてのみ、2 セット条件の方で 1 セット条件よりも標準偏差が大きかった。この原因としては、アトリビュートを真に測る項目と誤って測る項目の相対的な数が考えられる。直線型のアトリビュート C の各レベルを測る項目数は、1 セット条件の真の Q 行列ではそれぞれ 1 つずつしかない。そして誤設定パターン (3) と (4) では、誤設定項目も 1 つあるため、真にアトリビュート C の該当レベルを測る項目の数と誤って測る項目の数と同じになる。この相対的な項目数の関係は項目セット数が増えても同じであるため、診断精度が改善しなかったと考えられる。

今回の研究では、誤設定する要素を項目 1 セットにつき 1 つしか設けなかった。しかし、先述の通り、実際のテストでは複数の Q 行列要素で誤設定が生じる可能性もある。もし 1 つのアトリビュート、および 1 つのレベルに対して少数の項目しか測っていない Q 行列を作成した場合、誤設定による診断精度の悪化には十分に気を付ける必要があるだろう。特に直線型では、階層が高くなるほどアトリビュートを測定する項目数が必然的に少なくなる。こうした階層性タイプの性質を十分理解したうえで Q 行列を作成するべきと考えられる。

5.3. 診断の正確度のさらなる検討

本研究では、誤設定があるときの診断の正確度への影響をより詳しく調べるために、各パラメタのバイアスと RMSE も算出した。このうち slipping パラメタに関しては、今回のような過大に誤設定する条件では、誤設定による診断の正確度への影響を受けないことが二値アトリビュートにおける先行研究で報告されており (山口, 2017)、今回も同様の結果が得られた。また、混合比率パラメタ π_c のバイアスと RMSE には誤設定の影響も見られたが、これは必ずしも通常的分析で主要な関心のあるパラメタではない。以上のことから、これら 2 点の結果については電子付録へ掲載するに留め、通常的应用においてより関心が高いと考えられる他のパラメタの結果について以下で述べる。

表 7. アトリビュート修得状況パラメタ α_k のバイアスと RMSE (直線型)

バイアス				バイアス			
1セット				2セット			
アトリビュート				アトリビュート			
	A	B	C		A	B	C
真	-.054	.012	.113	真	-.007	.011	.023
(1)	-.008	.059	.102	(1)	.010	.037	.022
(2)	-.019	.095	.118	(2)	-.015	.101	.050
(3)	-.054	.069	.193	(3)	-.005	.050	.155
(4)	-.032	.079	.328	(4)	-.010	.052	.374

RMSE				RMSE			
1セット				2セット			
アトリビュート				アトリビュート			
	A	B	C		A	B	C
真	.462	.551	.491	真	.337	.374	.347
(1)	.502	.562	.488	(1)	.403	.392	.344
(2)	.515	.584	.513	(2)	.432	.472	.362
(3)	.477	.573	.559	(3)	.340	.405	.475
(4)	.468	.578	.693	(4)	.341	.430	.729

※灰色の網掛けはアトリビュートの誤設定をした項目を、太字は各表において最も大きな値を示す。

アトリビュート修得状況パラメタ まずアトリビュートごとのアトリビュート修得状況パラメタ α_k のバイアスと RMSE の結果を表 7 に示した。全体的に類似した傾向が見られたため、ここでは最も誤設定の影響が大きかった直線型の結果のみを示した。

直線型以外の階層性条件も含む全体的な傾向としては、大半の条件、および大半のアトリビュートにおいて、バイアスは絶対値 0.1 未満であった。特に直線型の 1 セット条件以外では 0.05 未満がほとんどだった。直線型では、アトリビュート C で誤設定したパターンで大きく正にバイアスがあることが確認された。また、このバイアスの大きさは、0→1 の誤設定よりも 0→2 の誤設定においてより顕著に見られた。

RMSE でも同様の結果が見られ、2 セット条件では大半の条件、アトリビュートにおいて 0.5 未満だったにもかかわらず、直線型の誤設定パターン (4) では 0.729 という最も大きな値が得られた。これは、AR_r の平均や標準偏差の項で述べた通り、誤設定の程度の大きさだけでなく、真にこのアトリビュートのレベル 2 を測っている項目と誤って測っている項目が同数あることが原因だと考えられる。

また、1 セット条件と 2 セット条件の RMSE の比較においては、表 5 で示した ACA の増加に関連する結果が見られた。アトリビュートごとのアトリビュ

表 8. 分岐型の 1 セット条件における guessing パラメタ g_j のバイアスと RMSE

バイアス										
項目										
	1	2	3	4	5	6	7	8	9	10
真	.139	.071	-.001	-.001	.002	.000	.005	.004	.006	.004
(1)	.124	.351	.013	-.004	.017	.009	.002	.003	.000	.002
(2)	.156	.423	-.007	.000	.012	.004	.000	-.002	.004	-.004
(3)	.145	.344	.003	.000	.002	-.008	-.002	-.002	-.002	.000
(4)	.150	.427	.018	.002	-.009	.000	-.004	.000	-.003	-.001

RMSE										
項目										
	1	2	3	4	5	6	7	8	9	10
真	.200	.144	.068	.033	.077	.036	.030	.025	.023	.020
(1)	.181	.374	.065	.032	.075	.037	.037	.022	.023	.021
(2)	.206	.438	.063	.034	.066	.040	.035	.025	.024	.018
(3)	.190	.371	.065	.040	.054	.036	.029	.025	.022	.019
(4)	.193	.443	.082	.035	.067	.036	.041	.026	.022	.019

※灰色の網掛けはアトリビュートの誤設定をした項目を、太字は各表において最も大きな値を示す。

ト修得状況パラメタ α_k の RMSE の算出では、推定値として事後平均値を用いた。一方で、(8) 式で示した一致率 AR_r の算出で用いた推定値 $\hat{\alpha}_{ik}$ は、先述の通り事後平均値を四捨五入した整数値であった。アトリビュート修得状況パラメタの真値は乱数生成された 0, 1, 2 のいずれかの値であるため、真値と事後平均値の差が 0.5 未満であれば、事後平均値を四捨五入した値は真値と一致する。1 セット条件ではどの Q 行列設定、およびアトリビュートでも平均的な誤差の指標である RMSE が 0.5 に近い値か 0.5 を超えた RMSE が得られていたが、2 セット条件では 0.4 未満の RMSE が得られることが多く見られた。こうした RMSE の減少が表 5 で示した ACA の向上へとつながったと考えられる。Q 行列の診断の正確度を調べる研究では、ACA のみが調べられることが多いが、このように RMSE も算出することで、診断の正確度についてより詳しく調べることができる。

guessing パラメタ guessing パラメタのバイアスと RMSE については、最も誤設定の影響が大きかった分岐型の 1 セット条件の結果を表 8 に示し、後に説明する例外が見られた直線型の 1 セット条件の結果を表 9 に示した。なお、直線型、分岐型、収束型のうち、分岐型で最も guessing パラメタの RMSE が大き

表 9. 直線型の 1 セット条件における guessing パラメタ g_j のバイアスと RMSE

バイアス						
項目						
	1	2	3	4	5	6
真	.156	.071	.022	.012	-.001	-.005
(1)	.114	.152	.003	-.004	.005	.000
(2)	.119	.211	.003	.006	-.003	.001
(3)	.127	.051	.013	.025	.014	.001
(4)	.138	.032	-.012	.045	-.007	.001

RMSE						
項目						
	1	2	3	4	5	6
真	.198	.128	.096	.073	.030	.029
(1)	.190	.182	.088	.050	.039	.032
(2)	.178	.219	.058	.053	.037	.031
(3)	.183	.122	.078	.067	.051	.030
(4)	.209	.130	.078	.067	.044	.027

※灰色の網掛けは誤設定したアトリビュートを、太字は各表において最も大きな値を示す。

かったことは、二値アトリビュートの場合を調べた山口 (2017) の結果と一貫している。

まず全体的な傾向として、誤設定した項目のバイアスは正しい設定をした項目よりも正に大きいことが見られた。RMSE に関しては、直線型の 1 セット条件における誤設定パターン (3) と (4) 以外で、誤設定した項目の値が正しい設定をした項目の値よりも高かった。さらに、この傾向は $0 \rightarrow 1$ の誤設定条件よりも、 $0 \rightarrow 2$ の誤設定条件の方でより大きく見られた。直線型の 1 セット条件では、誤設定パターン (3) と (4) における RMSE で誤設定した項目は、正しい設定をした場合と比較して低い値が得られた。これも、真にアトリビュート C のレベルを測っている項目と誤って測っている項目の相対的な数が同じであることが原因かもしれない。

また、項目パラメタには、とりうる値の範囲が $[0,1]$ であり、かつ $0 \leq g_j + s_j \leq 1$ という制約がある。項目パラメタの値は、項目の質 (item quality) の指標としても捉えられ、これらの値が高いほど、項目が測定したいアトリビュートの推定が困難になると考えられる。もし項目パラメタのバイアスや RMSE の値が大きければ、本当は現実のテストに使用可能な項目であったとしても、使用可能ではないと判断されてしまう可能性が高くなるだろう。たとえば、分岐型の 1 セット条件の項目 2 においては、 $0 \rightarrow 2$ の誤設定をしたときにはバイアスが 0.4 を超えており、guessing の値は真値である 0.2 と合わせて 0.6 以上だと平均的に推定されていることになる。このような項目パラメタの推定値が得られた場合、その項目は実際のテストにおいては基本的に使用できないと判断されることになるだろう。つまり、たとえ項目パラメタの真の値が適切であっても、Q 行列を誤って設定してしまった場合には、本来利用可能な項目が利用されない可能性が高くなってしまう。

真の Q 行列設定条件においても、直線型、分岐型における 1 セット条件の項目 1 では、他の項目と比較してバイアスと RMSE が 0 からより大きく離れていた。これは、この項目がアトリビュート A のレベル 1 を測る唯一の項目であるからだと考えられる。ここでは示さなかったがアトリビュート A のレベル 1 を測る項目が複数ある 2 セット条件においては、バイアスと RMSE が 0 に近づいていたことも、この解釈を支持すると考えられる。

6. 総 合 考 察

本研究では、多値のアトリビュートにおいて Q 行列に誤設定がある場合に、誤設定の程度の大きさが診断の正確度に及ぼす影響の程度をシミュレーションにより複数の階層性タイプで調べた。全体的には、誤設定の程度が小さい場合よりも、大きい場合で診断の正確度への影響がより大きかった。しかし、その影響の程度は、階層性タイプ、項目数、誤設定するアトリビュートレベルの数により異なっていた。

現実のテストへの示唆として、アトリビュートのレベル数と項目数、そして Q 行列のデザインについて以下で議論する。まず、レベルが高いほど誤設定したときの影響が大きくなりやすいことを考慮して、不必要に多くのレベルを設けないようにするべきだと考えられる。

また、項目数は診断の正確度に大きな影響を与える要因であることが示された。したがって、同じアトリビュートであっても、多値アトリビュートでは各レベルに十分な項目数を用意することが望ましい。特に、直線型では、アトリビュートの階層が高いほど、測定する項目数が相対的に少なくなる傾向にあるため、最も階層が高いアトリビュートに誤設定がある可能性を考慮に入れて、項目数を用意するべきだと考えられる。もし誤設定項目の数が正しく設定されている項目の数と同じになってしまうと、推定値は正に偏りやすくなり、診断の正確度が下がってしまう。しかし、項目数を増やすことが必ずしも診断の正確度を高めるとは限らないことには注意すべきである。たとえ項目数を増やしても、誤設定項目数も増えてしまう場合には、誤設定した項目数に応じて診断の正確度が低下してしまい、むしろ全体的な診断の正確度が低下してしまう可能性がある。ただし、本研究では、複数の誤設定項目を含むパターンは限定的な条件設定でしか調べられていない。もし誤設定項目数が増加するほど、診断の正確度への影響が増加するかどうかについて、より詳しく調べたい場合は、異なる条件を設ける必要があるだろう。

さらに、Q 行列の誤設定のされ方は、本稿冒頭で述べた先行研究および本研究で検討した、アトリビュートのレベルを過大 (あるいは過少) に見積もるもの以外にもあり得るだろう。たとえば、テストで測るアトリビュートの数を過大 (あるいは過少) に見積もったり、アトリビュートの階層性のタイプを誤ったりする

ことなどが考えられる。Q 行列は認知診断モデルの核ともなる部分であるため、こうした種類の誤設定も含め、様々な観点から誤設定に対する頑健性が検討されていくことが望ましい。

階層性タイプ間の比較は、各アトリビュートを測る項目数を全ての階層性タイプで同じにすることで、より正確に行えると考えられる。しかし、これを統制することは今回の設定では不可能である。もし各アトリビュートを測る項目数を統制したいのであれば、他の Q 行列デザインを設定したり、構造上の仮定を緩めたりする必要があると思われる。今回の階層性とアトリビュートのレベルの組合せでは、アトリビュートのレベルを最大まで修得しなければ、次のアトリビュートを修得することはできないという仮定を置いていた。これは現実のテストでは比較的強い仮定だと考えられる。実際には、あるアトリビュートをある程度まで修得することで、他のアトリビュートが修得可能になる場合が多いだろう。また、全てのアトリビュートが同じ最大レベルをもつという仮定も現実には強い仮定だと思われる。実際には内容領域や項目の困難度に合わせてアトリビュートごとに適切なレベル数を設けることが望ましい。

また、他の Q 行列デザインとしては、二値アトリビュートにおける先行研究において、3 つの Q 行列デザインアプローチを階層性と組合せた場合の診断の正確度が調べられている (Liu, Huggins-Manley, & Bradshaw, 2017)。この 3 つの Q 行列デザインアプローチは多値アトリビュートの場合でも適用可能であり、誤設定の影響についても検討することができるため、必要に応じて将来の研究において調べられることが望まれる。

本研究では、Q 行列の誤設定が存在する場合における、認知診断の正確度への影響を検討した。しかしながら、バイアスや誤分類といった診断時の悪影響を避けるためには、Q 行列の作成時に、できるかぎり誤設定を回避できることが望ましい。そこで、Q 行列の作成に際して、誤設定をできる限り回避するために、どのような方法がとり得るのかを最後に考察する。1 節でも述べた通り、Q 行列は専門家の意見を参考に作成されることが多い。これは、「あるテスト項目の正解のためにはあるアトリビュートが必要である」という主張を支える根拠が、専門家の知識や知見に基づいて与えるためと考えられる。もちろん、こうした妥当性の根拠を与えるための手段は、専門家に依る方

法以外にもあり得る。テスト開発者が利用できる妥当性の根拠としては、既存の理論などから導く演繹的な方法だけでなく、経験的な手法による帰納的方法も利用できるだろう。たとえば、回答中の発話データや視線データ、反応時間データ、マウスのクリックやカーソルの軌道、ログデータなどを得ることは、回答者がどのようなアトリビュートを利用しながら回答を生成しているのかを知る手掛かりとなり得る (たとえば、Ercikan & Pellegrino, 2017; Zumbo & Hubley, 2017)。こうした考え方は、Q 行列に限らずテストの妥当性を検証する際に重要と考えられてきた (AERA, APA, & NCME, 2014)。現状、著者らの知る限りでは、Q 行列の作成のために上記のようなデータを用いた研究が多く行われているとは言い難い (発話データの使用例として、Jang, 2009)。IT 技術の発展により様々なデータの取得が容易になってきている現在、多様な観点から Q 行列を適切に設定するための検討がなされていくことが期待される。

謝 辞

本研究は JSPS 科研費 18H03612, 19H00616 の助成を受けたものです。

参 考 文 献

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, & NCME]. (2014). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.
- Chen, J. & de la Torre, J. (2013). A general cognitive diagnosis model for expert-defined polytomous attributes. *Applied Psychological Measurement*, **37**, 6, 419–437.
- de la Torre, J. (2008). An empirically based method of Q-matrix validation for the DINA model: Development and applications. *Journal of Educational Measurement*, **45**, 4, 343–362.
- de la Torre, J. & Chiu, C. (2016). A general method of empirical Q-matrix validation. *Psychometrika*, **81**, 2, 253–273.
- Ercikan, K. & Pellegrino, J. W. (Eds.). (2017). *Validation of Score Meaning for the Next Generation of Assessments: The Use of Response Processes*. New York, NY: Routledge.
- Jang, E. E. (2009). Demystifying a Q-matrix for making diagnostic inferences about L2 reading skills. *Language Assessment Quarterly*, **6**, 3, 210–238.

- doi: 10.1080/15434300903071817
- Jang, E. E., Kim, H., Vincett, M., Barron, C., & Russell, B. (2019). Improving IELTS reading test score interpretations and utilization through cognitive diagnosis model-based skill profiling. *IELTS Research Reports Online Series, No. 2*. British Council, Cambridge Assessment English and IDP: IELTS Australia. Available at <https://www.ielts.org/teaching-and-research/research-reports>
- Karelitz, T. M. (2004). *Ordered category attribute coding framework for cognitive assessments* (Unpublished doctoral dissertation). University of Illinois at Urbana-Champaign.
- Leighton, J. P., & Gierl, M. J. (Eds.). (2007). *Cognitive diagnostic assessment for education: Theory and applications*. New York, NY: Cambridge University Press.
- Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The attribute hierarchy method for cognitive assessment: A variation onatsuoka's rule space approach. *Journal of Educational Measurement, 41*, 205–237.
- Liu, J. & Kang, H. (2019). Q-matrix learning via latent variable selection and identifiability. In von Davier, M. & Lee, Y. (Eds.), *Handbook of diagnostic classification models- Models and model extensions, applications, software packages*. (pp.247–263) New York, NY: Springer.
- Liu, R., Huggins-Manley, A. C., & Bradshaw, L. (2017). The impact of Q-matrix design on diagnosis classification account for the presence of attribute hierarchies. *Educational and Psychological Measurement, 77*, 2, 220–240.
- Plummer, M. (2017). *JAGS Version 4.3.0 user manual*. Lyon, France. Retrieved from <http://sourceforge.net/projects/mcmc-jags/>
- R Core Team. (2019). *R: A language and environment for statistical computing*. Vienna, Austria: The R Foundation for Statistical Computing.
- Rupp, A. A. & Templin, J. (2008). The effects of Q-matrix misspecification on parameter estimates and classification accuracy in the DINA model. *Educational and Psychological Measurement, 68*, 1, 78–96.
- Su, Y. S., & Yajima, M. (2015). *R2jags: Using R to run "JAGS"* (R package version 0.5-7). Retrieved from <http://CRAN.R-project.org/package=R2jags>
- Sunbul, S. O. & Asiret, S. (2018). The effects of Q-matrix misspecification on item and model fit. *International Journal of Progressive Education, 14*, 5, 144–154. doi: 10.29329/ijpe.2018.157.12
- Tu, D., Wang, S., Cai, Y., Douglas, J. & Chang, H.-H. (2019). Cognitive diagnostic models with attribute hierarchies: Model estimation with a restricted Q-matrix design. *Applied Psychological Measurement, 43*, 4, 255–271.
- von Davier, M. & Lee, Y. (2019). Introduction: From latent classes to cognitive diagnostic models. In von Davier, M. & Lee, Y. (Eds.) *Handbook of diagnostic classification models: Models and model extensions, applications, software packages*. (pp.1–17) New York, NY: Springer.
- 山口一大 (2017). 認知診断モデルにおける Q 行列の誤設定が診断精度に与える影響—認知の階層構造を考慮した場合の検討—. *日本テスト学会誌, 13*, 17–32.
- Yamaguchi, K & Okada, K. Variational Bayes inference for the DINA model. *Journal of Educational and Behavioral Statistics*, Advance online publication. <https://doi.org/10.3102/1076998620911934>
- 山口一大・岡田謙介 (2017). 近年の認知診断モデルの展開. *行動計量学, 44*, 2, 181–198.
- 山口一大 (2019). DINA モデルの再定式化と推定アルゴリズムの直接的導出. *日本テスト学会誌, 15*, 21–43.
- Zhan, P., Bian, Y., & Wang, L. (2016). Factors affecting the classification accuracy of reparametrized diagnostic classification models for expert-defined polytomous attributes. *Acta Psychologica Sinica, 48*, 318–330.
- Zhan, P., Jiao, H., & Man, K. (2019). Using JAGS for Bayesian cognitive diagnosis modeling: A tutorial. *Journal of Educational and Behavioral Statistics, 44*, 4, 473–503.
- Zumbo, B. D., & Hubley, A. M. (Eds.). (2017). *Understanding and investigating response processes in validation research*. Cham, Switzerland: Springer International Publishing.

(2019年12月21日受理, 2020年4月29日採択)

(この間審査3回・審査期間合計85日)