統計数理(2024)
第 72 巻 第 1 号 121–146

特集「心理統計学の新展開」
[原著論文]

小サンプルサイズ下での認知診断モデルの推定精度の検討

──モデルの誤設定の影響と推定法の違いに着目して──

佐宗 駿[1]・岡 元紀[2]・宇佐美 慧[1]

(受付 2023 年 6 月 30 日;改訂 2024 年 2 月 28 日;採択 3 月 5 日)

要　　旨

認知診断モデル(cognitive diagnostic models; CDM)では,測定対象となる学習要素はアトリビュートと呼ばれ,学習者ごとに各アトリビュートの習得・未習得の状態が推定される.CDM による推定結果は,学校現場での形成的評価に有用であると示唆されてきた一方,未だ実践では十分に活用されていないという実態がある.その原因の一つとして,学校現場で想定される小サンプルサイズ下での CDM の推定精度および,CDM のモデル選択で利用される情報量規準の選択傾向に関する検討の不足が挙げられる.本研究では,このような小サンプルサイズの状況を想定したシミュレーションデザインのもと,CDM における一般化モデルに基づきデータを発生させ,その下位モデルにあたる様々なモデルの推定精度および情報量規準の選択傾向を検討した.主な結果として,(1)アトリビュート習得パタンおよび項目パラメタの推定精度の観点からは,最尤推定法およびベイズ推定法いずれの場合においても,各アトリビュートの習得が個別に正答確率に寄与する CRUM が,真のモデルと同程度もしくはそれに次ぐ水準の精度を全体として有していた.(2)サンプルサイズを増やすよりも,項目数を増やしアトリビュート数を減らすことが高い推定精度につながることが示唆された.(3)最尤推定法において AIC は CRUM を支持する割合が高く,BIC は最も倹約的なモデルを支持する割合が高かった.ベイズ推定法において WAIC は,真のデータ生成モデルもしくは,母数の数の意味でそれと同程度に複雑なモデルを支持する傾向が見られた.

キーワード:認知診断モデル,小サンプルサイズ,ベイズ推定法,情報量規準,形成的評価.

1. 問題と目的

1.1 診断的・形成的評価と認知診断モデル

学校現場において,テストにより学習者の学習上のつまずきを診断し,その結果を学習者の

[1] 東京大学 大学院教育学研究科:〒 113–0033 東京都文京区本郷 7–3–1

[2] Department of Statistics, London School of Economics and Political Science, Columbia House, Houghton Street, London, WC2A 2AE, UK.

Supplementary Information: 本論文の Supplement は https://www.ism.ac.jp/editsec/toukei/pdf/72-1-Supplement.pdf でもご覧いただけます.

表 1. 1 桁の整数の計算に関する Q 行列の例.

項目	A1：足し算	A2：引き算	A3：掛け算
$5+2$	1	0	0
9×3	0	0	1
⋮	⋮	⋮	⋮
$9-4+5$	1	1	0

表 2. 可能な全てのアトリビュート習得パタン.

クラス	A1：足し算	A2：引き算	A3：掛け算
1	0	0	0
2	1	0	0
3	0	1	0
4	0	0	1
5	1	1	0
6	1	0	1
7	0	1	1
8	1	1	1

学習改善および教師の指導改善に活かす，診断的・形成的評価の必要性は論を俟たない（e.g., 文部科学省, 2019）．当該単元で習得すべき学習要素のうち，どの要素が習得できており，どの要素が未習得であるのかといった学習者の習得状況を把握できれば，個別最適化された学習・指導改善に有用な診断情報となるだろう．学校現場では，実施したテストの結果として，合計点(総合得点)を用いてフィードバックすることが多い．しかし，合計点は学習者の順位づけや選抜に有用であるものの，どの学習要素につまずきを抱えているかについての診断的情報を直接与えるものではない（e.g., 池田, 2013）．

テスト理論の研究では，学習者の各学習要素の習得・未習得を推定できる統計モデルとして，認知診断モデル（cognitive diagnostic models, CDM; Rupp et al., 2010)が近年注目を浴びている．CDM では，測定の対象となる学習要素をアトリビュートと呼び，各項目の正答にどのアトリビュートが必要かを，行を項目，列をアトリビュートとした $\{0,1\}$ の 2 値変数を要素にもつ行列である Q 行列（Q-matrix; Tatsuoka, 1983)によって表現する．単純な例として，1 桁の整数の足し算・引き算・掛け算の遂行をアトリビュートとして，それらの習得状況の診断を目的とした表 1 の Q 行列を考える．アトリビュートには，「A1：足し算」・「A2：引き算」・「A3：掛け算」が設定されている．たとえば，項目「$9-4+5$」の正答には，「A1：足し算」・「A2：引き算」のアトリビュートが必要であるため，1 が割り振られており，一方で「A3：掛け算」のアトリビュートは不要であるため，0 が割り振られる．この Q 行列と解答データ行列をもとに各学習者は，表 2 に示した $2^3=8$ 通りの可能な全てのアトリビュート習得パタンのうち，いずれかに分類される．たとえば，ある学習者が表 2 におけるクラス 5 のアトリビュート習得パタンに分類された場合，足し算と引き算は習得しているが，掛け算は未習得であるとわかる．CDM では，全ての学習者のアトリビュート習得パタンをまとめた，行が学習者，列がアトリビュートの習得の有無を表す行列である，アトリビュート習得パタン行列が得られる（山口・岡田, 2017）．このような分類により，各学習者のつまずきをより詳細に診断できるため，学校や学級での学習・指導改善へとつながると考えられており，実践場面での CDM の活用が期待されている（Sessoms and Henson, 2018）．

1.2 教育現場での CDM 活用の現状

CDM は実践に活きる高い活用可能性を秘めているにもかかわらず，従来の研究では大規模なサンプルサイズ下での応用が中心であり，学校現場で想定される学級単位のような比較的サンプルサイズの小さい場面での応用が不十分であることが指摘されている（e.g., Henson, 2009; Sessoms and Henson, 2018）．たとえば，36 本の CDM の応用研究をレビューした Sessoms and Henson（2018）は，そのうち 61% の研究において，サンプルサイズが 1000 よりも大きかったことを報告している．

学校現場における数少ない実践的な活用事例として Uesaka et al.（2021）は，公立高校の 3 年生 1 学級 40 名を対象に実施された数学の定期テストに CDM を適用し，数学的問題解決過程において図表を活用する力である「図表活用力（diagrammatic competency）」を含む 5 つのアトリビュートの習得状況を推定した．その結果，「図表活用力」を習得している学習者の割合は半数程度であり，合計点の情報からは直接的に明らかにできなかった学習者のつまずきの傾向を明らかにした．Saso et al.（2023）は，公立中学校の 1 年生 3 学級 87 名を対象として，数学の理解度を診断するテストを開発・実施し，CDM の推定結果としてアトリビュート習得状況を学習者にフィードバックした．学習者への自由記述型の質問紙の結果，自身の学習上の強みと弱みを知ることができ，今後の学習改善に活きるという肯定的な反応を得たことを示している（佐宗 他, 2022）．Jang（2005）は，大学生・大学院生 27 名を対象に実施した英文読解テストに CDM を適用し，英文読解力に関するアトリビュートの習得状況を診断し，学習者へフィードバックした．そのフィードバックは，英文読解力における学習者自身の強みと弱みを把握できる点で，学習者に肯定的に受け止められたことを示している．このような実践的な活用事例は，小サンプルサイズの状況下も含め，教育現場での CDM の活用が学習者の学習改善および教師の指導改善に資する可能性を示唆するものであろう．

1.3 小サンプルサイズ下での推定精度の検討の必要性

CDM の実践的な活用事例の蓄積が，その学校現場での更なる活用事例を促すと考えられる一方で，小サンプルサイズでの活用事例が限られている理由の一つとして，学級単位のような比較的小さいサンプルサイズ下での CDM におけるアトリビュート習得パタンや後述する項目パラメタに関する推定精度の検討が限定的であることが考えられる．

CDM には，アトリビュート習得パタンと項目正答確率の関係をどのように表現するかに応じて，DINA（deterministic inputs, noisy “and” gate; Junker and Sijtsma, 2001）モデル，DINO（deterministic inputs, noisy “or” gate; Templin and Henson, 2006）モデル，RRUM（reduced reparameterized unified model; Hartz and Roussos, 2008），CRUM（compensatory reparameterized unified model; Hartz, 2002），LCDM（log-linear cognitive diagnostic model; Henson et al., 2009）といった様々なモデルが存在しており，これらのモデルの小サンプルサイズ下での推定精度に関するシミュレーション研究がこれまで部分的に行われてきた（Paulsen and Valdivia, 2021; Sen and Cohen, 2021; Oka and Okada, 2021; Hueying and Yang, 2019; Başokçu, 2014）．たとえば，アトリビュート習得パタンと項目パラメタの推定精度に関して Sen and Cohen（2021）は，サンプルサイズ $N = 50, 100, 200, 300, 400, 500, 1000, 5000$ の下で，DINA モデル，DINO モデル，CRUM，LCDM の 4 つの CDM を対象に，最尤推定法を用いて検討を行った．その結果，サンプルサイズが増えるほどアトリビュート習得パタンの推定精度が向上し，アトリビュート数が増えるほど同様の推定精度が低下することを示した．さらに，小サンプルサイズ下では，DINA モデルがそのほかのモデルと比べて，アトリビュート習得パタンと項目パラメタの推定精度が高いことを示した．

従来の小サンプルサイズを想定したシミュレーション研究では，最尤推定法が中心に検討さ

れてきた．しかし，ある項目に対して，全ての学習者が正答もしくは誤答してしまう完全解答パタン（perfect response pattern; Levy and Mislevy, 2016）が生じるリスクが他の条件が一定の下で小サンプルサイズの場合には高くなり，これが生じた場合，その項目における尤度が発散するため最尤推定法では推定が困難になる．この点を問題意識として Oka and Okada（2021）は，ベイズ推定法も考慮した小サンプルサイズ下での CDM の推定精度の検討を行った．具体的には，Oka and Okada（2021）は $N = 20, 40, 160$ の条件において，最尤推定法，パラメタの事前分布に無情報事前分布もしくは弱情報事前分布を設定したベイズ推定法，およびノンパラメトリック推定法に基づいて，DINA モデル，DINO モデル，RRUM，CRUM，LCDM の 5 つの CDM を用いて検討を行った．その結果，アトリビュート習得パタンの推定精度について，DINA モデル，DINO モデルについては，最尤推定法と比べてベイズ推定法が比較的優れた傾向にあるが，RRUM，CRUM，LCDM では，ベイズ推定法に比べて最尤推定法の方が優れた傾向にあることを示した．

しかし，Sen and Cohen（2021）や Oka and Okada（2021）をはじめとした従来の小サンプルサイズ下を想定したシミュレーション研究では，分析モデルとデータ生成モデルが同じ状況のもとで行われているという強い制約がある．一方，たとえば，CDM のうち倹約的なモデルの一つである DINA モデルの仮定に沿った解答プロセスを想定した診断テストおよび Q 行列を事前に設定したとしても，現実では，学習者が想定した解答プロセスに沿って項目に解答するとは限らず，むしろより複雑な解答プロセスも十分に想定される．そのため，たとえば CDM のクラスの中でも一般性の高いモデル（i.e., LCDM）に基づきデータを発生させ，その下位モデルにあたる様々な CDM を分析モデルとすることを通してモデルの誤設定の影響を考慮し，その上で推定精度を調べる方がより現実に即した検討となることが期待され，また上述の先行研究とは異なる選択結果が得られる可能性が大いに考えられる．

例外として，モデルの誤設定を考慮し，かつ小サンプルサイズ下を想定したシミュレーション研究として，Hueying and Yang（2019）がある．Hueying and Yang（2019）は，$N = 50, 75, 100, 200$ の条件下で，データ生成モデルを，LCDM と同様に CDM のクラスの中で一般性が高いモデルである G-DINA（generalized DINA; de la Torre, 2011）モデルとし，G-DINA モデルおよびその下位モデルである DINA モデルと DINO モデルを分析モデルとして，情報量規準である AIC（Akaike information criterion; Akaike, 1974）と BIC（Bayesian information criterion; Schwarz, 1978）のモデル選択の傾向を検討した．その結果，小サンプルサイズ下において AIC が真のモデルである G-DINA モデルをより高い割合で選択する傾向にあったことを示している．CDM ではこれまで様々な種類のモデルが提案され利用されているため，情報量規準を用いたモデル選択の作業は応用上重要である．一方で，Hueying and Yang（2019）は，情報量規準の選択傾向の検討に留まっており，モデルの誤設定の状況下でのアトリビュート習得パタンや項目パラメタの推定精度については検討していない．加えて，小サンプルサイズ下を想定してはいるものの，その設定値の最小は $N = 50$ であり，実際の学級サイズの分布からすれば依然高い設定値である．また，比較対象となった DINA モデルと DINO モデルはいずれも CDM のうち最も倹約的なモデルであり，検討されたモデルが限定的であったことも問題点として挙げられる．さらに，推定方法も最尤推定法に留まっており，ベイズ推定法において活用されてきた情報量規準である WAIC（widely applicable information criterion; Watanabe, 2010）の CDM の文脈における選択精度については未検討である．

1.4 本稿の目的と構成

以上を踏まえ，本研究の第一の目的は，学級場面を想定した小サンプルサイズ下における CDM のアトリビュート習得パタンおよび項目パラメタの推定精度を，より現実的と考えられ

るデータ生成モデルと分析モデルの誤設定の影響を踏まえながら，最尤推定法およびベイズ推定法に基づいて検討することである．参考として，アトリビュート習得パタンの推定精度については，小サンプルサイズ下での活用を意図して開発されているノンパラメトリック推定法も含めて検討を行う．第二の目的は，各情報量規準（AIC・BIC・WAIC）の，特に小サンプルサイズ下における CDM のモデル選択傾向や選択精度について明らかにすることである．

本論文の構成は次のとおりである．2 章では本研究で用いる CDM と情報量規準の概要をそれぞれ説明する．3 章ではシミュレーションの手続きならびに結果と考察を示す．最後に 4 章では本研究で得られた知見をまとめ，限界点と今後の展望について述べる．

2. 本研究で対象とする CDM と情報量規準の概要

2.1 CDM の概要

本節では本研究で扱う，DINA モデル，DINO モデル，RRUM，CRUM，LCDM の 5 つの CDM を説明する．これらのモデルは，従来の CDM の応用事例の中で実際に活用されており（e.g., Sessoms and Henson, 2018），後述するようにいずれも LCDM の下位モデルとして表現可能である．

添字 i $(1,2,\ldots,N)$ は学習者，j $(1,2,\ldots,J)$ は項目（テスト問題），k $(1,2,\ldots,K)$ はアトリビュート，l $(1,2,\ldots,L=2^K)$ はアトリビュート習得パタンのクラスをそれぞれ表す記号である．解答データ行列 $\boldsymbol{X}$ は，学習者 i の項目 j への解答 $x_{ij} \in \{0,1\}$（0：誤答，1：正答）を要素としてもつ $N \times J$ 行列である．学習者 i についての要素をまとめた $\boldsymbol{x}_i = (x_{i1},\ldots,x_{ij},\ldots,x_{iJ})^\top$ を用いると，$\boldsymbol{X} = (\boldsymbol{x}_1,\ldots,\boldsymbol{x}_i,\ldots,\boldsymbol{x}_N)^\top$ となる．Q 行列 $\boldsymbol{Q}$ は項目 j の正答にアトリビュート k が必要かどうかを示す $q_{jk} \in \{0,1\}$（0：不要，1：必要）を要素としてもつ $J \times K$ 行列である．項目 j についての要素をまとめた $\boldsymbol{q}_j = (q_{j1},\ldots,q_{jk},\ldots,q_{jK})^\top$ を用いると，$\boldsymbol{Q} = (\boldsymbol{q}_1,\ldots,\boldsymbol{q}_j,\ldots,\boldsymbol{q}_J)^\top$ となる．アトリビュート習得パタン行列 $\boldsymbol{A}$ は，学習者 i のアトリビュート習得パタンにおける k 番目のアトリビュートの習得の有無を表す $\alpha_{ik} \in \{0,1\}$（0：未習得，1：習得）を要素としてもつ $N \times K$ 行列である．学習者 i のアトリビュート習得パタン $\boldsymbol{\alpha}_i = (\alpha_{i1},\ldots,\alpha_{ik},\ldots,\alpha_{iK})^\top$ を用いると，$\boldsymbol{A} = (\boldsymbol{\alpha}_1,\ldots,\boldsymbol{\alpha}_i,\ldots,\boldsymbol{\alpha}_N)^\top$ となる．

CDM では，学習者 i の項目 j への正答確率 $p_{ij} = P(x_{ij}=1|\boldsymbol{\alpha}_i,\boldsymbol{\beta}_j;\boldsymbol{q}_j)$ を，それぞれのモデル上の仮定に応じて表現し，

$$\mathcal{L} = L(\boldsymbol{X}|\boldsymbol{\pi},\boldsymbol{B};\boldsymbol{Q}) = \prod_{i=1}^{N}\sum_{l=1}^{L}\pi_l \prod_{j=1}^{J} p_{ij}^{x_{ij}}(1-p_{ij})^{1-x_{ij}} \tag{2.1}$$

の周辺尤度関数をもとに，モデルパラメタである項目パラメタおよび混合比率パラメタの推定を行う．なお，各学習者が属するアトリビュート習得パタンの算出は，項目反応理論におけるスコアリングと同様の手続きを踏んで行われる（Kim and Nicewander, 1993）. ここで，$\boldsymbol{\beta}_j$ は項目 j に関する各モデルで推定される項目パラメタをまとめたベクトルであり，$\boldsymbol{B} = (\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_j,\ldots,\boldsymbol{\beta}_J)^\top$ である．$\boldsymbol{\pi}$ は L 個のクラスそれぞれのアトリビュート習得パタンへの混合（所属）比率をまとめたベクトルで $\boldsymbol{\pi} = (\pi_1,\ldots,\pi_l,\ldots,\pi_L)^\top$ であり，$\sum_{l=1}^{L}\pi_l = 1$ を満たす．また，セミコロン以降に表記した $\boldsymbol{Q}$ はそれが事前に設定された既知の値であることを意味する．

2.1.1 DINA モデル

DINA モデルは，CDM のうち最も倹約的なモデルの一つであり，項目の正答に必要なアトリビュートを全て習得している場合に限って，正答確率が高くなるような項目反応を仮定する．具体的には，学習者 i が項目 j の正答に必要なアトリビュートを全て習得している場合に

1(i.e., 正答)，そうでない場合は 0(i.e., 誤答)となる理想反応と呼ばれる指標，

$$\eta_{ij} = \prod_{k=1}^{K} \alpha_{ik}^{q_{jk}} \tag{2.2}$$

を導入する．ここで，$0^0 = 1$ と定義する．

実際の解答では，必要なアトリビュートを全て習得しているにもかかわらず誤答するケースや，必要なアトリビュートが全て揃っていないにもかかわらず正答するケースが想定される．これらを表現するための項目パラメタとして，前者のケースでは slip パラメタと呼ばれる s_j，後者については guessing パラメタと呼ばれる g_j を導入する．つまり，

$$s_j = P(x_{ij} = 0|\eta_{ij} = 1) \tag{2.3}$$

$$g_j = P(x_{ij} = 1|\eta_{ij} = 0) \tag{2.4}$$

という条件付き確率を表す項目パラメタである．これら η_{ij}，s_j，g_j の関数として，

$$p_{ij} = P(x_{ij} = 1|\boldsymbol{\alpha}_i, s_j, g_j; \boldsymbol{q}_j) = (1 - s_j - g_j)\eta_{ij} + g_j \tag{2.5}$$

と表現するのが DINA モデルである．

2.1.2 DINO モデル

DINO モデルは，DINA モデルと理想反応の仮定のみが異なる．項目の正答に必要なアトリビュートを少なくとも 1 つ習得している場合に正答確率が高くなり，それ以上のアトリビュートを習得している場合でも正答確率は変わらないことを仮定する．具体的には，学習者 i が項目 j の正答に必要なアトリビュートを少なくとも 1 つ習得している場合に 1(i.e., 正答)，そうでない場合は 0(i.e., 誤答)となる理想反応，

$$\omega_{ij} = 1 - \prod_{k=1}^{K} (1 - \alpha_{ik})^{q_{jk}} \tag{2.6}$$

を導入する．

DINA モデルと同様に，

$$s_j = P(x_{ij} = 0|\omega_{ij} = 1) \tag{2.7}$$

$$g_j = P(x_{ij} = 1|\omega_{ij} = 0) \tag{2.8}$$

という条件付き確率を表す項目パラメタ s_j，g_j を有する．これら ω_{ij}，s_j，g_j の関数として，p_{ij} を

$$p_{ij} = P(x_{ij} = 1|\boldsymbol{\alpha}_i, s_j, g_j; \boldsymbol{q}_j) = (1 - s_j - g_j)\omega_{ij} + g_j \tag{2.9}$$

と表現するのが DINO モデルである．

2.1.3 RRUM

DINA モデルでは，項目の正答に必要なアトリビュートを全て習得している場合とそうでない場合で正答確率が異なることを仮定していた．また，DINO モデルでは，必要なアトリビュートのうち少なくとも 1 つ習得している場合と，全て未習得の場合で正答確率が異なることを仮定していた．これに対して，正答に必要なアトリビュートがそれぞれ個別に正答確率に寄与すると仮定する方が現実的にはより自然と考えられるケースは多いだろう．このような仮定を表現するモデルの一つが RRUM である．RRUM では，項目パラメタとして，項目 j の正

答に必要な全てのアトリビュートを習得している場合の正答確率 τ_j および，アトリビュート k を習得していない場合に正答確率を低下させる罰則パラメタ $r_{jk}\,(0 < r_{jk} < 1)$ を用いて，

$$p_{ij} = P(x_{ij} = 1|\boldsymbol{\alpha}_i, \tau_j, r_{j1}, \ldots, r_{jK}; \boldsymbol{q}_j) = \tau_j \prod_{k=1}^{K} r_{jk}^{q_{jk}(1-\alpha_{ik})} \tag{2.10}$$

と表現する．ここで項目パラメタについて，項目の正答に必要なアトリビュートを全て習得している場合に誤答する確率 s_j は RRUM では $1-\tau_j$ に，また項目の正答に必要なアトリビュートを全て習得していないにも関わらず正答する確率 g_j は RRUM では $\tau_j \prod_{k=1}^{K} r_{jk}$ にそれぞれ対応している．

2.1.4 CRUM

RRUM と同様に，項目の正答に必要なアトリビュートが個別に正答確率に寄与することを仮定するのが CRUM である．具体的には，当て推量パラメタに相当する切片パラメタ λ_{j0} および，項目 j の正答に必要とされる各アトリビュートを習得している場合に正答確率を変化させる主効果パラメタ $\lambda_{j1}, \ldots, \lambda_{jK}$ を用いて項目正答確率は，

$$p_{ij} = P(x_{ij} = 1|\boldsymbol{\alpha}_i, \lambda_{j0}, \lambda_{j1}, \ldots, \lambda_{jK}; \boldsymbol{q}_j) = \frac{1}{1 + \exp(-(\lambda_{j0} + \sum_{k=1}^{K} \lambda_{jk}\alpha_{ik}q_{jk}))} \tag{2.11}$$

と表現される．ここで項目パラメタについて，項目の正答に必要なアトリビュートを全て習得している場合に誤答する確率 s_j は CRUM では $1 - \mathrm{logit}^{-1}(\lambda_{j0} + \sum_{k=1}^{K} \lambda_{jk}\alpha_{ik}q_{jk})$ に，また項目の正答に必要なアトリビュートを全て習得していないにも関わらず正答する確率 g_j は CRUM では $\mathrm{logit}^{-1}(\lambda_{j0})$ にそれぞれ対応している．

2.1.5 LCDM

ここまでの DINA モデル，DINO モデル，RRUM，CRUM を包含する一般化モデルが LCDM である．LCDM では，CRUM で仮定されている切片パラメタと主効果パラメタに加えて，全てのアトリビュートの組み合わせに関する交互作用効果パラメタを考える．たとえば，項目 j におけるアトリビュート k と $k'\,(k \neq k')$ の 1 次の交互作用効果を表すパラメタを $\lambda_{jkk'}$ と表現する．項目 j の切片パラメタ・主効果パラメタ・交互作用効果パラメタを用いて項目正答確率を，

$$\begin{aligned} p_{ij} &= P(x_{ij} = 1|\boldsymbol{\alpha}_i, \lambda_{j0}, \lambda_{j1}, \ldots, \lambda_{jK}, \lambda_{j11}, \ldots; \boldsymbol{q}_j) \\ &= \frac{1}{1 + \exp(-(\lambda_{j0} + \sum_{k=1}^{K} \lambda_{jk}\alpha_{ik}q_{jk} + \sum_{k=1}^{K-1} \sum_{k'>k} \lambda_{jkk'}\alpha_{ik}\alpha_{ik'}q_{jk}q_{jk'} + \cdots))} \end{aligned} \tag{2.12}$$

と表現する．このとき，項目の正答に必要なアトリビュートを全て習得している場合に誤答する確率 s_j は LCDM では $1 - \mathrm{logit}^{-1}(\lambda_{j0} + \sum_{k=1}^{K} \lambda_{jk}\alpha_{ik}q_{jk} + \sum_{k=1}^{K-1} \sum_{k'>k} \lambda_{jkk'}\alpha_{ik}\alpha_{ik'}q_{jk}q_{jk'} + \cdots)$ に，また項目の正答に必要なアトリビュートを全て習得していないにも関わらず正答する確率 g_j は LCDM では $\mathrm{logit}^{-1}(\lambda_{j0})$ にそれぞれ対応している．

LCDM は，パラメタに特定の制約を課すことで，DINA モデル，DINO モデル，RRUM，CRUM を下位モデルとして表現することが可能である．たとえば，交互作用効果パラメタを全て 0 と制約を置くことで CRUM を表現することができる．そのほかのモデルに関する制約の置き方の詳細は，Henson et al. (2009) を参照されたい．なお，LCDM は G-DINA モデルにロジットリンク関数を用いたモデルと等価であることが知られている（de la Torre, 2011）．

なお，CDM の分析では，あるアトリビュートを習得している場合，そのアトリビュートを習得していない場合と比べて，正答確率が同等以上になるという単調性制約（monotonic constraints; Henson et al., 2009）を置くことが多い．たとえば，DINA モデルと DINO モデルに

おいては，$0 \leq g_j \leq 1 - s_j \leq 1$ という不等式制約で表現される．これは理想反応が 1(i.e., 正答)であるときの正答確率(i.e., $1 - s_j$)が，理想反応が 0(i.e., 誤答)であるときの正答確率(i.e., g_j)を下回らないことを意味する．本研究においても，全ての CDM について，この制約を置くことのできる推定法である最尤推定法とベイズ推定法において，この制約を置いてシミュレーションを行った．

2.2 情報量規準

本研究では最尤推定法における情報量規準として，従来の CDM の応用研究で多く利用されている AIC と BIC（e.g., Sessoms and Henson, 2018），ベイズ推定法における情報量規準として，CDM の実践的な活用事例である佐宗 他（2023)で使用された WAIC を利用する．これらの定義式はそれぞれ，

$$\mathrm{AIC} = -2\ln \mathcal{L}_{max} + 2d \tag{2.13}$$

$$\mathrm{BIC} = -2\ln \mathcal{L}_{max} + d\ln n \tag{2.14}$$

$$\begin{aligned}\mathrm{WAIC} = &-2\sum_{i=1}^{N}\sum_{j=1}^{J} \ln E_{\boldsymbol{\alpha},\boldsymbol{\beta}|\boldsymbol{x}}[\mathcal{L}_c(x_{ij}|\boldsymbol{\alpha}_i,\boldsymbol{\beta}_j;\boldsymbol{q}_j)] \\ &+2\sum_{i=1}^{N}\sum_{j=1}^{J} Var_{\boldsymbol{\alpha},\boldsymbol{\beta}|\boldsymbol{x}}[\mathcal{L}_c(x_{ij}|\boldsymbol{\alpha}_i,\boldsymbol{\beta}_j;\boldsymbol{q}_j)]\end{aligned} \tag{2.15}$$

のようになる（Akaike, 1974; Schwarz, 1978; Merkle et al., 2019）．ただし，$\ln \mathcal{L}_{max}$ は式(2.1)で表される尤度関数の最大値に自然対数をとった最大対数尤度，d はモデルの自由パラメタ数を表し，$E_{\boldsymbol{\alpha},\boldsymbol{\beta}|\boldsymbol{x}}[\mathcal{L}_c(x_{ij}|\boldsymbol{\alpha}_i,\boldsymbol{\beta}_j;\boldsymbol{q}_j)]$ と $Var_{\boldsymbol{\alpha},\boldsymbol{\beta}|\boldsymbol{x}}[\mathcal{L}_c(x_{ij}|\boldsymbol{\alpha}_i,\boldsymbol{\beta}_j;\boldsymbol{q}_j)]$ はそれぞれ $\boldsymbol{\alpha},\boldsymbol{\beta}$ の事後分布に関する，条件付き尤度関数 $\mathcal{L}_c(x_{ij}|\boldsymbol{\alpha}_i,\boldsymbol{\beta}_j;\boldsymbol{q}_j)$ の期待値と分散を表す．d は具体的には，DINA モデルと DINO モデルでは $2J + 2^K - 1$ であり，項目 j の正答に必要とされるアトリビュートの総数を $K_j^* = \sum_{k=1}^{K} q_{jk}$ とすると，RRUM と CRUM では $\sum_{j=1}^{J} K_j^* + J + 2^K - 1$，LCDM では，$\sum_{j=1}^{J} 2^{K_j^*} + 2^K - 1$ である．AIC と WAIC は汎化損失，BIC は自由エネルギーの観点に基づいて，分析モデルの観測データへのあてはまりを示す相対的指標であり，その値が最小となるモデルが選択される（e.g., Vrieze, 2012; Watanabe, 2010; 浜田 他, 2019）．

なお正則性を満たし，かつサンプルサイズが大きい場合の AIC と BIC のモデル選択傾向の性質はよく知られている．具体的には，BIC は，(1)分析モデルの中に真のデータ生成モデルが含まれており，(2)真のデータ生成モデルのパラメタの数が一定で，かつ(3)パラメタの数が有限である場合において，サンプルサイズが大きくなるにつれて真のデータ生成モデルを選択する確率が 1 に収束する一致性を有する（Vrieze, 2012）．これに対して，AIC は自由パラメタの数を固定したままサンプルサイズが限りなく大きくなったとしてもこのような一致性は満たされない．漸近理論が適用できない小サンプルサイズ下では，たとえば BIC が真のモデルよりも自由パラメタ数の少ない単純なモデルを選択する傾向などが知られているものの，実際の学級サイズの下での種々の CDM のモデル選択の文脈におけるこれら情報量規準の選択傾向や選択精度については明らかになっていない．

3. シミュレーションの方法と結果・考察

まず，真の Q 行列とパラメタの真値をもとに，一般化モデルである LCDM を通して解答データ行列を生成した．生成された解答データと，真の Q 行列を用いて DINA モデル，DINO モデル，RRUM，CRUM, LCDM のそれぞれを当てはめ，アトリビュート習得パタンおよび項目パ

表 3．シミュレーションデザインの概要．

デザイン要因	水準数	各水準の内容
サンプルサイズ	3	$N = 20, 40, 80$
項目数	2	$J = 20, 40$
アトリビュート数	2	$K = 4, 5$
推定法	3	最尤推定法，ベイズ推定法，ノンパラメトリック推定法
分析モデル	5	DINA モデル, DINO モデル, RRUM, CRUM, LCDM

ラメタを推定した．この手順を各条件で 100 回繰り返し，得られたアトリビュート習得パタン $\boldsymbol{\alpha}_i$ および項目パラメタ s_j，g_j の推定値が真値をどれほど復元できているかを評価する．ただし，DINA モデルと DINO モデル以外の CDM である RRUM，CRUM，LCDM に関しては，項目パラメタ s_j，g_j をパラメタとして直接的には有さないが，第 2.1.3〜2.1.5 節で示したように，それぞれに対応する確率がモデル内の項目パラメタを用いて評価できるため，RRUM，CRUM についてはこの性質を利用して評価を行った．LCDM については，後述するように単調性制約を置くために Yamaguchi and Templin (2022)の定式化を利用したため，この定式化における s_j，g_j に対応するパラメタを評価した．本研究のシミュレーションデザインは，関連する先行研究を参考にしながら，学校現場における実際の活用場面を想定し，表 3 のように設定した．

3.1 シミュレーションデザインの設定

3.1.1 サンプルサイズ・項目数・アトリビュート数の設定

サンプルサイズ N は，実際の学校現場で CDM を活用した事例において Jang (2005)では 27 名，Uesaka et al. (2021)では 40 名，Saso et al. (2023)では 87 名を対象としていたことに加えて，「公立義務教育諸学校の学級編制及び教職員定数の標準に関する法律」において，学級編成基準として，1 学級当たりの標準とする生徒数は中学校では 40 人とされていることも考慮して，$N = 20, 40, 80$ の 3 条件とした．

項目数 J は，学校現場での CDM の活用事例の一つである佐宗 他 (2023)が対象とした，定期テストの項目数が 36 であったことから，$J = 40$ を上限とし，確認テストなど授業時間内の一部の時間を用いて実施される定期テストよりも項目数が少ない場面を想定するために，$J = 20$ も条件に含めた．アトリビュート数 K は，CDM の応用場面において $K = 4$ が最頻値であったこと (Sessoms and Henson, 2018)に加えて，学校現場で CDM を活用した事例である Uesaka et al. (2021) および Saso et al. (2023)では $K = 5$ であったことから，$K = 4, 5$ の 2 条件とした．

3.1.2 Q 行列の設定

シミュレーションデザインに従い，$(K, J) = (4, 20), (5, 20), (4, 40), (5, 40)$ の 4 つの Q 行列を設定した．Q 行列の設定においては，CDM のパラメタ推定における識別性(identifiability)の条件を満たすために，各アトリビュートが 1 つのみ寄与する項目を少なくとも 2 つ含むようにした (Xu, 2017)．まず，$(K, J) = (4, 20)$ の Q 行列 $\boldsymbol{Q}_{(4,20)}$ は，各アトリビュートが 1 つのみ寄与する 4 項目(i.e., 4 次の単位行列 $\boldsymbol{I}_4$)2 セット分に加えて，残りの 12 項目は 2 個もしくは 3 個のアトリビュートが寄与する全 ${}_4\mathrm{C}_2 + {}_4\mathrm{C}_3 = 10$ 通りと，さらに 4 つ全てのアトリビュートそれぞれが 20 項目を通して同じ回数ずつ測定されるように，2 個のアトリビュートが寄与する全 ${}_4\mathrm{C}_2 = 6$ 通りのうち 2 通りをランダムに選択して $\boldsymbol{Q}^*_{(4,20)}$ を設定し，合わせて $\boldsymbol{Q}_{(4,20)} = (\boldsymbol{I}_4, \boldsymbol{I}_4, \boldsymbol{Q}^*_{(4,20)})^\top$ とした．同様に，$(K, J) = (5, 20)$ の Q 行列 $\boldsymbol{Q}_{(5,20)}$ は，各アトリビュートが 1 つのみ寄与する 5 項目(i.e., $\boldsymbol{I}_5$)2 セット分に加えて，残りの 10 項目は 2 個のアトリビュートが寄与する全 ${}_5\mathrm{C}_2 = 10$ 通りのうち 5 通り，3 個のアトリビュートが寄与する全 ${}_5\mathrm{C}_3 = 10$ 通りのうち 5 通り

を 5 つ全てのアトリビュートそれぞれが 20 項目を通して同じ回数ずつ測定されるようにランダムに選択して $\boldsymbol{Q}^*_{(5,20)}$ を設定し，合わせて $\boldsymbol{Q}_{(5,20)} = (\boldsymbol{I}_5, \boldsymbol{I}_5, \boldsymbol{Q}^*_{(5,20)})^\top$ とした．これら $\boldsymbol{Q}_{(4,20)}$，$\boldsymbol{Q}_{(5,20)}$ を用いて，$(K, J) = (4, 40)$ の Q 行列 $\boldsymbol{Q}_{(4,40)}$，$(K, J) = (5, 40)$ の Q 行列 $\boldsymbol{Q}_{(5,40)}$ をそれぞれ $\boldsymbol{Q}_{(4,40)} = (\boldsymbol{I}_4, \boldsymbol{I}_4, \boldsymbol{Q}^*_{(4,20)}, \boldsymbol{I}_4, \boldsymbol{I}_4, \boldsymbol{Q}^*_{(4,20)})^\top$，$\boldsymbol{Q}_{(5,40)} = (\boldsymbol{I}_5, \boldsymbol{I}_5, \boldsymbol{Q}^*_{(5,20)}, \boldsymbol{I}_5, \boldsymbol{I}_5, \boldsymbol{Q}^*_{(5,20)})^\top$ と設定した．以上の設定に基づいたこれらの Q 行列は，Xu (2017) の定理 1 から，項目パラメタおよび混合比率パラメタの一致性を満たしている．具体的な Q 行列は Supplementary Information A1 を参照されたい．

3.1.3 項目パラメタの設定

CDM において，項目の識別力は $1 - s_j - g_j$ という指標で定義され，項目パラメタ s_j，g_j の値をもとに項目の質を捉えることができる (e.g., de la Torre, 2008)．つまり，s_j，g_j それぞれが小さくなるほど，その項目の識別力は高くなる．本研究では，学校現場で実施される定期テストや確認テストにおいてしばしば見られるように，ハイステークスかつ大規模な学力テストと比べて必ずしも項目作成に十分なコストがかけられず，その結果として項目の識別力が十分高い水準に至らない状況を想定する．具体的には，全てのアトリビュートが未習得である場合の正答確率が.20 (i.e., $g_j = .80$)，全て習得している場合の正答確率が.80 (i.e., $s_j = .20$) となるように項目パラメタの真値を設定した．また，各項目における主効果パラメタおよび交互作用効果パラメタの真値は，$s_j = g_j = .20$ の条件を満たすように，R の GDINA パッケージに含まれる simGDINA 関数の引数 gs.parm を設定することで，各生成データセットでランダムに設定された．

3.1.4 アトリビュート習得パタンの生成

各学習者の真のアトリビュート習得パタンは，Chiu and Douglas (2013) の多変量正規閾値モデルを用いて次のように生成した．まず，学習者 i の連続的な潜在特性値ベクトル $\boldsymbol{\theta}_i = (\theta_{i1}, \theta_{i2}, \ldots, \theta_{iK})^\top$ を，平均が 0 ベクトル，分散共分散行列が $\boldsymbol{\Sigma}$ である多変量正規分布から発生させる．ここで $\boldsymbol{\Sigma}$ は，対角成分が 1，非対角成分が .50 の K 次正方行列である．

得られた $\theta_{i1}, \ldots, \theta_{iK}$ を用いて，

$$\alpha_{ik} = \begin{cases} 1 & \text{if} \quad \theta_{ik} \geq \Phi^{-1}\left(\dfrac{k}{K+1}\right) \\ 0 & \text{(otherwise)} \end{cases} \tag{3.1}$$

の規則に基づいて，学習者 i のアトリビュート習得パタン $\boldsymbol{\alpha}_i = (\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{iK})^\top$ を生成する．ここで，Φ^{-1} は標準正規分布の累積密度関数の逆関数であり，その引数を $k/(K+1)$ とすることで各アトリビュートの習得難易度が異なるという一般的な状況を仮定している．つまり，k が大きいほどそのアトリビュートの習得難易度が高くなり，たとえば，$K = 4$ の状況では A1 から A4 の順に習得難易度が高くなっていくことを仮定している．なお，この生成方法では，すべての混合比率パラメタの真値が $\pi_l \in (0, 1)$ となり，いずれのアトリビュート習得パタン l についても $\pi_l = 0$ となる l は存在しないことも仮定している．

以上をもとに，解答データ行列を，R の GDINA パッケージに含まれる simGDINA 関数を用いて，各条件で 100 セットずつ生成した．

3.1.5 推定の実施と設定

ベイズ推定法では，R の R2jags パッケージ (Su and Yajima, 2021) と統計ソフトウェア JAGS (just another Gibbs sampler; Plummer et al., 2003) を用いて，MCMC 法 (Marchov chain Monte Carlo method) により実行した．MCMC の設定は，チェーン数が 4，イタレーション数が 10000,

バーンイン区間が 4000 として，各学習者のアトリビュート習得状況と項目パラメタの EAP (expected a posteriori)推定値を得た．なお，アトリビュート習得状況については，EAP 推定値を四捨五入した整数値を推定値として，以下では扱った．事前分布は，DINA モデル，DINO モデル，RRUM，CRUM については Zhan et al. (2019)および Oka and Okada (2021)の設定を参考に，単調性制約を満たすように設定した．LCDM については Yamaguchi and Templin (2022)の定式化を用いた上で，それぞれのパラメタに対して無情報事前分布を設定したのち，Yamaguchi and Templin (2022)で提案された単調性制約を満たすようなギブスサンプリングアルゴリズムを利用して推定を行った．事前分布の詳細は，Supplementary Information A2 を参照されたい．なお，マルコフ連鎖の収束について，実際のシミュレーションにおいて，全ての条件で，5 つのモデル (i.e., DINA モデル，DINO モデル，RRUM，CRUM，LCDM)のそれぞれ全ての量的パラメタについて Gelman-Rubin 統計量 $\hat{R}$ (Gelman et al., 2013)が 1.05，1.1 もしくは 1.2 未満であるかどうかを確認した．その結果，ほとんど全ての量的パラメタについて 1.05 もしくは 1.1 未満であり，全ての量的パラメタについて 1.2 未満であったことが確認されたため，マルコフ連鎖が収束したと判断した．具体的には，1.05 を上回るパラメタの回数が最大のパラメタにおいても，100 個の生成データセットに対する推定全体において $\hat{R} > 1.05$ を示したのは 18 個のみであった．各条件における，より詳細な収束状況は Supplementary Information A3 に掲載している．

最尤推定法では，R の GDINA パッケージ (Ma and de la Torre, 2020)の GDINA 関数を用いて，単調性制約を置いた上で EM(expectation-maximization)アルゴリズムを用いた周辺最尤推定法によりパラメタ推定を行った．EM アルゴリズムは，イタレーション数が 2000 に到達するもしくは，t 回目と $t-1$ 回目 $(t = 2, \ldots, 2000)$ の更新時のパラメタの値の差の絶対値が 10^{-4} を下回るまで反復計算を行った．なお，各学習者が属するアトリビュート習得パタンは，項目反応理論におけるスコアリング (Kim and Nicewander, 1993)と同様の手続きを踏んで，モデルパラメタである項目パラメタと混合比率パラメタの周辺最尤推定値を所与としたときのアトリビュート習得パタンの最尤推定値によって算出した (Huebner and Wang, 2011, 式(5))．

ノンパラメトリック推定法は，R の NPCD パッケージ (Zheng et al., 2019)の AlphaNP 関数を用い，この関数で実行可能な DINA モデルおよび DINO モデルに限定して推定を行った．ノンパラメトリック推定法では，学習者 i の解答ベクトル $\boldsymbol{x}_i$ と，DINA モデルもしくは DINO モデルで定義される，クラス l のアトリビュート習得パタンに対する理想反応ベクトル $\boldsymbol{\eta}^{(l)} = (\eta_1^{(l)}, \eta_2^{(l)}, \ldots, \eta_J^{(l)})^\top$ の距離 $d(\boldsymbol{x}_i, \boldsymbol{\eta}^{(l)})$ が最小となるアトリビュート習得パタンを推定値とする．本研究では，Chiu and Douglas (2013)でアトリビュート習得パタンの推定精度が高いとされた，

$$d(\boldsymbol{x}_i, \boldsymbol{\eta}^{(l)}) = \sum_{j=1}^{J} \frac{1}{\bar{p}_j(1-\bar{p}_j)} |x_j - \eta_j^{(l)}| \tag{3.2}$$

と表される，項目 j の正答率 $\bar{p}_j$ の分散の逆数を重みとした重み付きハミング距離を距離の指標とした．なお，ノンパラメトリック推定法ではこのように，最尤推定法およびベイズ推定法とは異なり，単調性制約を置く対象となる項目パラメタを有さない．この点を踏まえて，本推定法による結果の解釈は，以降，参考程度に留めることとする．

3.2 推定精度の指標

3.2.1 アトリビュート習得パタンの推定精度の指標

アトリビュート習得パタンの推定精度を評価する指標として，EACR(element-wise attribute classification rate)と PACR(pair-wise attribute classification rate)を用いた．EACR は，アトリ

ビュート習得パタンの要素ごとの一致率を表す指標で，m $(1,2,\ldots,M)$ 個目の生成データセットにおける i 番目の学習者のアトリビュート k の習得状況 α_{mik} の推定値 $\hat{\alpha}_{mik}$ と真値 α_{mik}^{true} の一致率を生成データセット数 M とアトリビュート数 K について平均した値である．PACR は，アトリビュート習得パタンの一致率を表す指標で，m 個目の生成データセットにおける整数値 $\hat{\alpha}_{mik}$ を要素にもつ学習者 i のアトリビュートの習得パタン $\boldsymbol{\alpha}_{mi}$ と真値 $\boldsymbol{\alpha}_{mi}^{true}$ の一致率を生成データセット数 M について平均した値である．これらの求めた一致率の生成データセットごとの変動を評価するためにそれぞれの標準偏差も算出した．

3.2.2 項目パラメタの推定精度の指標

項目パラメタの推定の正確性を評価するために，s_j と g_j の RMSE（root mean square error）と Bias を算出した．例として，slip パラメタ s_j の RMSE と Bias は，

$$Bias_{s_j} = \frac{1}{M}\sum_{m=1}^{M}(\hat{s}_{mj} - s_j^{true}) \tag{3.3}$$

$$RMSE_{s_j} = \sqrt{\frac{1}{M}\sum_{m=1}^{M}(\hat{s}_{mj} - s_j^{true})^2} \tag{3.4}$$

で求められる．ここで，$\hat{s}_{mj}$ は，m 個目の生成データセットにおける項目 j の slip パラメタの推定値，s_j^{true} は項目 j の slip パラメタの真値をそれぞれ表す．

なお，本研究で用いた R コードおよび JAGS コードは，https://osf.io/ajkrv/?view_only=124bd089ac3e4622923343eee6189658 で公開している．紙幅の都合上載せられなかった図表，および本章で示す図のカラー版についても，Supplementary Information として上記 URL に掲載した．

3.3 結果と考察

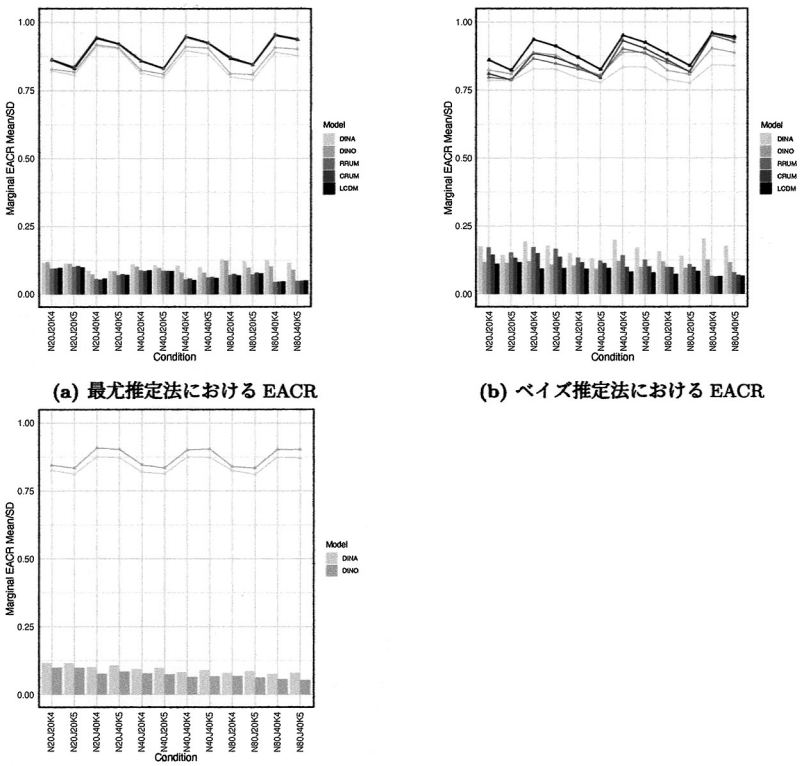
本節では，EACR と PACR の観点からアトリビュート習得パタンの推定精度の結果を示し，Bias と RMSE の観点から項目パラメタ，特に s_j，g_j の推定精度の結果を示す．さらに，各推定法で用いた情報量規準が選択したモデルの割合を示す．

なお，第 1 章で述べた通り，最尤推定法において完全解答パタンが生じた場合，推定が困難となる．各条件において推定法の間で同じ生成データセットが分析対象となるように，完全解答パタンが生じた生成データセットについてはデータセットを再生成し，各条件で完全解答パタンが生じていない 100 個の生成データセットを分析対象とした．

3.3.1 アトリビュート習得パタンの推定精度

図 1 は，推定法ごとの各条件における EACR およびその標準偏差，図 2 には，PACR およびその標準偏差を示した．横軸はシミュレーション条件を表しており，たとえば，N20J40K4 はサンプルサイズが 20，項目数が 40，アトリビュート数が 4 の条件を表している．全体的な傾向として，いずれの推定法においても，サンプルサイズが一定の条件では $(J,K) = (20,5)$ において EACR・PACR が最小となり，$(J,K) = (40,4)$ の条件で最大となった．これは，先行研究（e.g., Oka and Okada, 2021; Paulsen and Valdivia, 2021）で示された，アトリビュート数が少ないほど，また項目数が多いほどアトリビュート習得パタンの推定精度が向上するという結果が，モデルの誤設定が生じている場合にも観察されたものと解釈できる．

EACR と PACR の結果から，最尤推定法では，LCDM，RRUM および CRUM が同等に推定精度が高く，DINO モデルや DINA モデルのような倹約的なモデルの推定精度は相対的に低かった．ベイズ推定法においては，真のモデルである LCDM に続いて CRUM の推定精度



(a) 最尤推定法における EACR

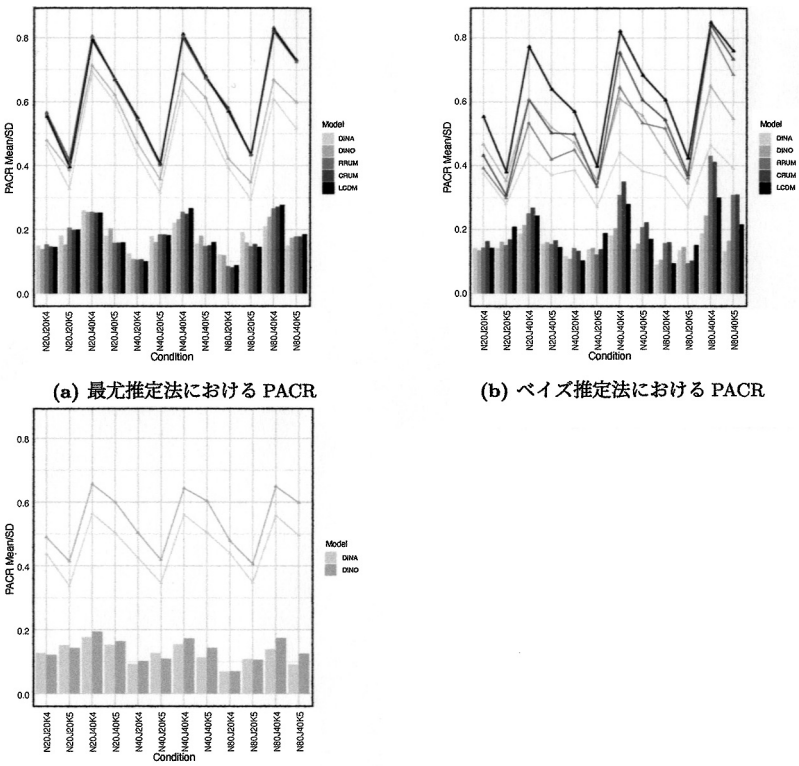
(b) ベイズ推定法における EACR

(c) ノンパラメトリック推定法における EACR

図 1. 推定法ごとの各条件における EACR（棒グラフ：標準偏差，折れ線グラフ：平均値）.

が高く，DINA モデルの推定精度が低い傾向にあった．ノンパラメトリック推定法においては，DINA モデルよりも DINO モデルの方が推定精度が高い傾向にあった．最尤推定法およびベイズ推定法において，真のデータ生成モデルである LCDM に比べてその下位モデルである CRUM が一貫してアトリビュート習得パタンの推定精度の水準が LCDM と同程度或いはそれに次いで高い傾向が見られた．その原因の一つとして，真のデータ生成モデルである LCDM が CRUM に近い状況になっていたことが考えられる．つまり，アトリビュートの主効果がアトリビュート間の交互作用効果よりも，相対的に大きく正答確率を変化させる，すなわち正答確率に与える主効果の影響が交互作用効果よりも大きい傾向にあったことが考えられる．具体的には，アトリビュートの主効果に基づく正答確率の増分の和が，交互作用効果パラメタの絶対値の和に比べて大きい傾向にあり，たとえば，$(N, J, K) = (20, 20, 4)$ の条件における 1 個目の生成データセットでは，A1 と A4 を測定する項目 11 では $d_0 = .20$, $d_1 = .12$, $d_4 = .26$, $d_{14} = .22$，A2 と A4 を測定する項目 13 では $d_0 = .20$, $d_2 = .27$, $d_4 = .48$, $d_{24} = -.15$ であった．ここで，d_0 はすべてのアトリビュートを習得していない場合の正答確率，d_i は i 番目のアトリビュートを習得している場合の主効果として生じる正答確率の増分，$d_{ij}(i \neq j)$ は i 番目と j 番目のアトリビュートを同時に習得している場合の 1 次の交互作用効果として生じる正答確率の増分を表している．

最尤推定法およびベイズ推定法において，DINA モデルの EACR および PACR の値が全体的に最も低かった．DINA モデルは DINO モデルと同様，最も倹約的なモデルである．自由パ



(a) 最尤推定法における PACR

(b) ベイズ推定法における PACR

(c) ノンパラメトリック推定法における PACR

図 2. 推定法ごとの各条件における PACR（棒グラフ：標準偏差，折れ線グラフ：平均値）.

ラメタ数が等しいにもかかわらず，DINA モデルの推定精度が低かった原因としては，DINA モデルと真のデータ生成モデルである LCDM の項目正答確率に対する仮定の違いが挙げられる．つまり，DINA モデルでは必要とされるアトリビュートを全て習得している場合は正答確率が高くなり，そうでない場合は正答確率が低下するという仮定をもつ一方で，LCDM では，習得しているアトリビュートの数に応じて，主効果パラメタおよび交互作用効果パラメタを通じて正答確率が高くなる．このような差異が，DINA モデルにおける推定精度を低下させる原因になったと考えられる．

このように，アトリビュートの主効果のみを考慮した CRUM の推定精度の水準が真のモデルである LCDM と同程度或いはそれに次いで高く，倹約的なモデルである DINA モデルの推定精度が相対的に低いことが明らかになった．これは，小サンプルサイズを想定したシミュレーション研究で示されてきた，倹約的なモデルである DINA モデルが相対的に推定精度が高いという結果（e.g., Sen and Cohen, 2021; Oka and Okada, 2021）と相反するものであり，CDM のクラスのうち最も一般的で，かつ現実の解答プロセスにより即していると考えられる LCDM をデータ生成モデルとして想定したことを反映した結果である．

追加の検討として，各シミュレーション要因の EACR および PACR への影響を検討するために，各要因を独立変数とし，EACR と PACR をそれぞれ従属変数とした分散分析を行った．その結果が表 4 である．なお，推定法の要因については，本研究で想定した 5 つのモデル全てに適用した最尤推定法とベイズ推定法に限定した．ここで，η^2 は各要因の効果に関する決定係

表 4. EACR・PACR を従属変数とした分散分析の結果.

		EACR				PACR			
	自由度	F 値	η^2	偏 η^2	p 値	F 値	η^2	偏 η^2	p 値
N	2	104.865	.017	.009	.000	91.872	.015	.008	.000
J	1	9461.978	.443	.404	.000	776.892	.395	.334	.000
K	1	512.099	.041	.022	.000	2086.586	.149	.090	.000
Model	4	4.007	.001	.001	.003	2.653	.001	.000	.031
Estimator	1	727.900	.058	.031	.000	72.815	.057	.031	.000
$N \times J$	2	16.774	.003	.001	.000	31.958	.005	.003	.000
$N \times K$	2	1.719	.000	.000	.179	.737	.000	.000	.478
$J \times K$	1	24.691	.002	.001	.000	29.067	.002	.001	.000
$N \times$ *Model*	8	9.970	.007	.003	.000	5.959	.004	.002	.000
$J \times$ *Model*	4	2.256	.001	.000	.061	1.137	.000	.000	.337
$K \times$ *Model*	4	.977	.000	.000	.419	.834	.000	.000	.503
$N \times$ *Estimator*	2	107.177	.018	.009	.000	102.208	.017	.009	.000
$J \times$ *Estimator*	1	42.403	.004	.002	.000	96.011	.008	.004	.000
$K \times$ *Estimator*	1	.697	.000	.000	.404	.051	.000	.000	.822
Model × *Estimator*	4	.999	.000	.000	.406	1.092	.000	.000	.358
$N \times J \times K$	2	2.155	.000	.000	.116	.516	.000	.000	.597
$N \times J \times$ *Model*	8	3.164	.002	.001	.001	2.722	.002	.001	.005
$N \times K \times$ *Model*	8	4.976	.003	.002	.000	3.346	.002	.001	.001
$J \times K \times$ *Model*	4	5.238	.002	.001	.000	2.963	.001	.001	.019
$N \times J \times$ *Estimator*	2	12.356	.002	.001	.000	23.207	.004	.002	.000
$N \times K \times$ *Estimator*	2	3.700	.001	.000	.025	2.967	.000	.000	.052
$J \times K \times$ *Estimator*	1	3.049	.000	.000	.081	.416	.000	.000	.519
$N \times$ *Model* × *Estimator*	8	.556	.000	.000	.814	.655	.000	.000	.732
$J \times$ *Model* × *Estimator*	4	.236	.000	.000	.918	.146	.000	.000	.965
$K \times$ *Model* × *Estimator*	4	1.130	.000	.000	.340	.530	.000	.000	.713
$N \times J \times K \times$ *Model*	8	2.374	.002	.001	.015	1.826	.001	.001	.067
$N \times J \times K \times$ *Estimator*	2	1.109	.000	.000	.330	.395	.000	.000	.674
$N \times J \times$ *Model* × *Estimator*	8	.354	.000	.000	.944	.323	.000	.000	.958
$N \times K \times$ *Model* × *Estimator*	8	.552	.000	.000	.818	.422	.000	.000	.909
$J \times K \times$ *Model* × *Estimator*	4	.671	.000	.000	.612	.331	.000	.000	.858
$N \times J \times K \times$ *Model* × *Estimator*	8	.306	.000	.000	.964	.179	.000	.000	.994

注）× は交互作用を表す．たとえば，$N \times J$ はサンプルサイズと項目数の 1 次の交互作用である．

数を示しており，「従属変数の分散が，それぞれの属性要因によってどれだけの割合説明されるか」を示す効果量の指標である（南風原, 2014）．また，偏 η^2 は，各要因の効果に関する偏決定係数を示しており，「その要因をモデルに含めることによって，その要因を含める前に説明できていなかった残差分散のうちの何 % を説明できたか」を示す効果量の指標である（南風原, 2014）．

表 4 から，EACR および PACR いずれを従属変数とした場合でも相対的に η^2 と偏 η^2 の値が最も大きかった要因が項目数である．項目数の η^2 と偏 η^2 の値は EACR の場合は η^2=.443，偏 η^2=.404 であり，PACR の場合は η^2=.395，偏 η^2=.334 であった．図 1，2 の結果も踏まえると，項目数を増やすことが，アトリビュート習得パタンの推定精度の向上に最も寄与する可能性が示唆された．

推定法の η^2 と偏 η^2 の値については，EACR の場合は η^2=.058，偏 η^2=.031 であり，PACR の場合は η^2=.057，偏 η^2=.031 であった．図 1，2 の結果も考慮すると，全体的な傾向として，本研究で設定した条件下では，最尤推定法はベイズ推定法よりも高い推定精度を有している

ことが示唆された．推定法間での差異が見られた原因を検討するため，追加の解析を行った．具体的には，最尤推定法とベイズ推定法における EACR の差の 2 乗和平均が最大となった DINA モデルに着目し，そのうち EACR の平均値の差異がこの 2 つの推定法間で最大であった $(N, J, K) = (40, 40, 4)$ の条件に焦点を当て，アトリビュートごとの真の値と要素の一致率を比較した．その結果，最尤推定法では A1 から A4 まで順にそれらの値は，.92, .92, .90, .85 であった一方で，ベイズ推定法では.93, .93, .88, .60 であり，難易度が最も高いアトリビュートである A4 について，最尤推定法と比べて相対的に一致率が低かった．この DINA モデルの推定結果の傾向は，そのほか全ての条件でも確認された．なお，真のモデルである LCDM のベイズ推定法では，全てのアトリビュートについて最尤推定法と同程度の推定精度を有していた．また，DINO モデルと RRUM は習得難易度が最も低いアトリビュートである A1 について，CRUM は DINA モデルと同様に習得難易度が最も高いアトリビュートである A4 について，最尤推定法と比べて相対的に一致率が低かった．これらの各アトリビュートの真値と要素の一致率をまとめた結果は Supplementary Information A4 に掲載している．

このような結果が得られた原因の一つとして，アトリビュート習得パタンの推定に直接関与するパラメタである，混合比率パラメタの推定精度の違いが挙げられる．たとえば，$(N, J, K) = (40, 40, 4)$ の条件における DINA モデルの解析で最尤推定法とベイズ推定法による差異が，最大であった生成データセットと最小となった生成データセットで比較した．その結果，A4 を習得しているアトリビュート習得パタンの混合比率パラメタの総和(i.e., $\sum_{l\in\{l;\alpha_{l4}=1\}} \pi_l$)は，差異が最大となった生成データセットでは，最尤推定法の場合.05 であったのに対してベイズ推定法では.68 であり，最尤推定法よりも，A4 の習得の有無について，チャンスレベルである.50 に近かった．一方で差異が最小となった生成データセットでは，同様の混合比率は最尤推定法では.73，ベイズ推定法では.69 であり推定法間で同程度の値を示していた．このような違いの原因として，生成データセット内での真のアトリビュート習得パタンにおける，A4 を習得しているアトリビュート習得パタンに属する学習者数の違いが考えられる．具体的には，たとえば，差異が最大となった生成データセットにおいては，A4 を習得しているアトリビュート習得パタンに属する学習者数が 40 名中 2 名であったのに対し，差異が最小となった生成データセットでは，40 名中 12 名であり，より半数に近い学習者が習得している状況であった．以下，A4 を習得しているアトリビュート習得パタンに属する学習者数を「当該アトリビュートの習得者数」とする．

当該アトリビュートの習得者数が少ない状況下で混合比率パラメタの推定精度がベイズ推定法において低下した原因として，各生成データセットにおけるアトリビュート習得パタンの事後分布の推定法間での違いに由来する，混合比率パラメタの値の更新方法の違いが挙げられる．つまり，最尤推定法では，EM アルゴリズムにおける M ステップにおいて，その時点における Q 関数を最大化する混合比率パラメタの値へと更新されるという意味において決定的であるのに対し，ベイズ推定法では，事前分布の情報を考慮した上で一つ前の MCMC 反復においてサンプルされたアトリビュート習得パタンの値に基づいて，確率分布からサンプリングされたパラメタへ更新されるという意味において確率的であるという違いである．

具体的には，最尤推定法の場合，混合比率パラメタは $\pi_l = \frac{\sum_{i=1}^{N} P(\boldsymbol{\alpha}_i=\boldsymbol{\alpha}_l|\boldsymbol{x}_i,\boldsymbol{B})}{N}$ で計算され，一つ前の反復におけるクラス l のアトリビュート習得パタンの混合比率パラメタを π_l^{old} とすれば，アトリビュート習得パタンの事後分布 $P(\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_l|\boldsymbol{x}_i, \boldsymbol{B})$ は $P(\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_l|\boldsymbol{x}_i, \boldsymbol{B}) = \frac{\pi_l^{\mathrm{old}} P(\boldsymbol{x}_i|\boldsymbol{B},\boldsymbol{\alpha}_i=\boldsymbol{\alpha}_l)}{\sum_{l=1}^{L} \pi_l^{\mathrm{old}} P(\boldsymbol{x}_i|\boldsymbol{B},\boldsymbol{\alpha}_i=\boldsymbol{\alpha}_l)}$ となる．ただし，$P(\boldsymbol{x}_i|\boldsymbol{B}, \boldsymbol{\alpha}_i = \boldsymbol{\alpha}_l)$ はクラス l のアトリビュート習得パタンを所与とした場合の学習者 i の項目反応ベクトル $\boldsymbol{x}_i$ における尤度である．これに対し，ベイズ推定法では $\tilde{\boldsymbol{N}} + \boldsymbol{\delta}_0$ をパラメタとするディリクレ分布 $Dirichlet(\tilde{\boldsymbol{N}} + \boldsymbol{\delta}_0)$ からサンプリ

ングされる．ここで，$\tilde{\boldsymbol{N}} = (\sum_{i=1}^{N} I(\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_1), \ldots, \sum_{i=1}^{N} I(\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_l), \ldots, \sum_{i=1}^{N} I(\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_L))^\top$，$\boldsymbol{\delta}_0$ は混合比率パラメタの事前分布のパラメタであり，$\tilde{\boldsymbol{N}}$ を構成する $\boldsymbol{\alpha}_i$ はカテゴリカル分布 $P(\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_l | \boldsymbol{x_i}, \boldsymbol{B}) = \prod_{l=1}^{L} P(\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_l | \boldsymbol{x_i}, \boldsymbol{B})^{I(\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_l)}$ からサンプリングされる．この時，アトリビュート習得パタンの事後分布 $P(\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_l | \boldsymbol{x}_i, \boldsymbol{B})$ は，一つ前の MCMC 反復における混合比率パラメタを $\boldsymbol{\pi}^{(t)}$ とすれば，$P(\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_l | \boldsymbol{x}_i, \boldsymbol{B}) = \frac{\pi_l^{(t)} P(\boldsymbol{x}_i | \boldsymbol{B}, \boldsymbol{\alpha}_i = \boldsymbol{\alpha}_l)}{\sum_{l=1}^{L} \pi_l^{(t)} P(\boldsymbol{x}_i | \boldsymbol{B}, \boldsymbol{\alpha}_i = \boldsymbol{\alpha}_l)}$ となる．なお，$\boldsymbol{\pi}^{(t)}$ は $Dirichlet(\tilde{\boldsymbol{N}}^{(t-1)} + \boldsymbol{\delta}_0)$ からサンプリングされた値である．ここで本研究では，混合比率パラメタの事前分布に対して無情報事前分布，つまり $\boldsymbol{\delta}_0$ として長さが L で要素が全て 1 のベクトルをパラメタとしたディリクレ分布を設定している．また通常，特に小サンプルサイズの下で当該アトリビュートの習得者数が 0 に近い場合，学習者全体における真のアトリビュート習得パタンの種類が，可能な全てのアトリビュート習得パタン数に比べて少なくなる．その結果として $Dirichlet(\tilde{\boldsymbol{N}} + \boldsymbol{\delta}_0)$ において，当該アトリビュートを習得しているアトリビュート習得パタンに対応する $\tilde{\boldsymbol{N}}$ の要素が 0 もしくはそれに近い値となるため，事後分布であるディリクレ分布において，それらのアトリビュート習得パタンに対応するパラメタの要素が，無情報事前分布におけるパラメタの要素とほぼ変化がなくなり，A4 を習得しているアトリビュート習得パタンの混合比率パラメタの総和(i.e., $\sum_{l \in \{l; \alpha_{l4}=1\}} \pi_l$)の推定値が，無情報事前分布上での A4 を習得しているアトリビュート習得パタンの混合比率パラメタの総和，つまりチャンスレベルである.50 に近い値になったと考えられる．

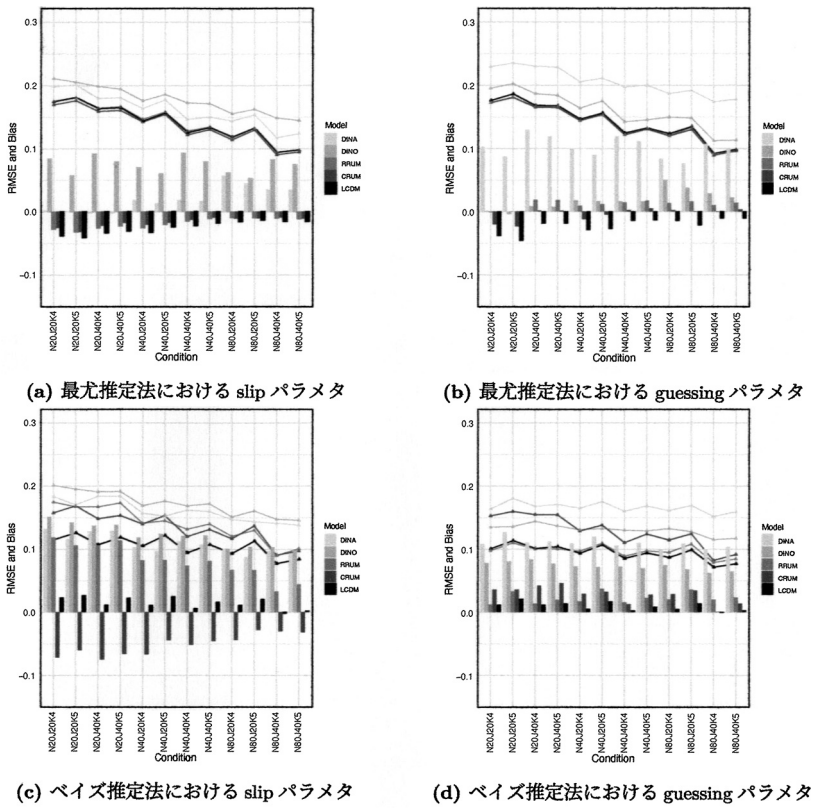
アトリビュート数の要因における η^2 と偏 η^2 は，PACR を従属変数とした場合に EACR を従属変数とした場合と比べて高い値を示した．図 1，2 における解釈も踏まえれば，アトリビュート数を減らすことが，PACR の推定精度を高めることを示唆する．これは，EACR がアトリビュート習得パタンの要素ごとの一致率を表す指標であったのに対して，PACR はアトリビュート習得パタンの一致率を表す指標であり，アトリビュート数が増えるにつれて可能なアトリビュート習得パタンの総数が指数関数的に増えることに起因すると考えられる．

そのほかの要因については，PACR を従属変数とした場合の，サンプルサイズと推定法の 1 次の交互作用効果における η^2=.018，偏 η^2=.017 が最大であり，項目数，アトリビュート数，推定法それぞれの主効果の要因と比べて相対的に小さかった．ここで，特にサンプルサイズの要因の η^2 と偏 η^2 の値が EACR では η^2=.017，偏 η^2=.009，PACR では η^2=.015，偏 η^2=.008 であったことは特筆すべき点である．つまり，学級の人数を想定した 20 名，40 名，80 名のいずれにおいても，アトリビュート習得パタンの推定精度に対して，必ずしも大きな影響を与えないということが示唆された．

以上をまとめると，まず表 1，2 から，EACR は.78 から.96 程度，PACR は.28 から.85 程度の範囲をもちシミュレーション条件間での差異が見られた．また最尤推定法およびベイズ推定法で EACR および PACR が最大の値を示した条件は，いずれも $(N, J, K) = (80, 40, 4)$ において CRUM が分析モデルの場合であった．さらに表 4 の結果から，推定精度の向上にはアトリビュート数を減らし，項目数を増やすことが有用であるとわかった．したがって，実践的な活用においてアトリビュートの設定を検討する際には，連関が高い 2 つ以上のアトリビュートは 1 つに統合することでアトリビュート数を不必要に増やすことを避けるといった工夫が特に効果的と考えられる．また，定期テストのように短期間に複数のテストが実施される場合に，いくつかの教科間で共通した教科横断的なアトリビュートが設定されている場合には，それらのテストを統合して CDM による分析を行うことで項目数を増やすといった工夫も考えられるであろう．

3.3.2 項目パラメタの推定精度

図 3 に，最尤推定法およびベイズ推定法における slip パラメタと guessing パラメタの Bias および RMSE の値を示した．全体的な傾向として，いずれの推定法においても，サンプルサイズが大きくなるほど，またアトリビュート数が少ないほど Bias と RMSE の値が小さくなる傾向が見られた．サンプルサイズが項目パラメタの推定精度に影響するという結果は，従来の結果（e.g., Oka and Okada, 2021; Paulsen and Valdivia, 2021）とも合致するものである．特に，最尤推定法において相対的に顕著にこの傾向が見られた原因の一つとして，境界問題（boundary problem; Maris, 1999, Yamaguchi, 2023）が考えられる．境界問題は，最尤推定法に特有の問題であり，項目パラメタの推定値が，真値と異なり，その項目パラメタの定義域の両極端の値を示す状況を指す．たとえば，DINA モデルおよび DINO モデルの項目パラメタ s_j，g_j の推定値がそれらの定義域の両極端の値に近い値を示す状況である．境界問題は，一部のアトリビュート習得パタンに属する学習者が（ほとんど）いない場合に生じやすいとされる（DeCarlo, 2011）．そのため，サンプルサイズが可能なアトリビュート習得パタンの総数を下回るような状況においては特に，境界問題が起こりやすいと考えられる．以上から，サンプルサイズが大きくなる，もしくはアトリビュート数が少なくなるにつれてアトリビュート習得パタンの総数（i.e., $K = 4$ では $2^4 = 16$ 通り，$K = 5$ では $2^5 = 32$ 通り）がサンプルサイズより下回るため，Bias と RMSE の値が相対的に小さくなったと考えられる．



(a) 最尤推定法における slip パラメタ　(b) 最尤推定法における guessing パラメタ

(c) ベイズ推定法における slip パラメタ　(d) ベイズ推定法における guessing パラメタ

図 3. slip パラメタと guessing パラメタの Bias と RMSE（棒グラフ：Bias，折れ線グラフ：RMSE）．

slip パラメタについては，真のデータ生成モデルである LCDM を除けば，全体的な傾向として DINO モデルにおいて最も Bias の絶対値および RMSE が大きく，RRUM もしくは CRUM において最も小さかった．DINO モデルでは，項目の正答に必要とされるアトリビュートを少なくとも一つ習得していれば理想反応が 1(i.e, 正答)となる一方で，真のデータ生成モデルである LCDM では，各アトリビュートの主効果および交互作用効果が項目正答確率に寄与する．この点を反映して，DINO モデルにおいて slip パラメタの Bias が正の方向に大きくなった，つまり slip パラメタが過大推定されたと考えられる．

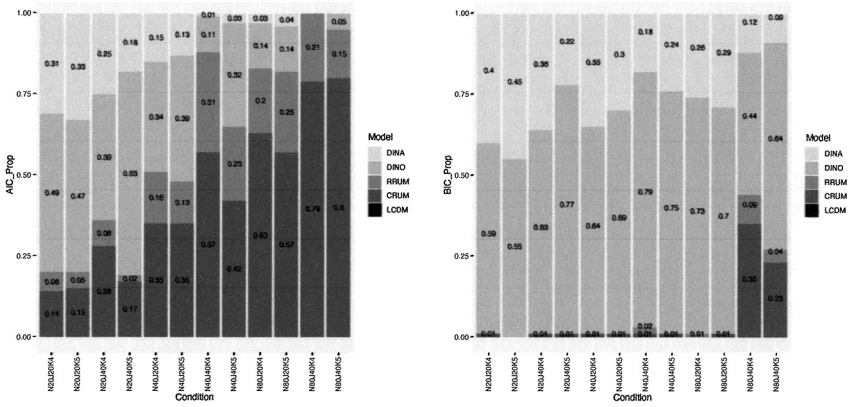
guessing パラメタに関して，DINA モデルが最も Bias の絶対値および RMSE が大きく，RRUM でそれらの値が最も小さかった．DINO モデルの場合と異なり，DINA モデルでは，項目の正答に必要とされるアトリビュートを全て習得している場合のみ理想反応が 1(i.e., 正答)となる．一方 LCDM では，仮に一つのアトリビュートのみを習得している場合でも，その主効果によって正答確率が高くなる．このような項目正答確率とアトリビュートの関係に関する仮定の違いを反映して，DINA モデルにおいて guessing パラメタの Bias が正の方向に大きくなった，つまり guessing パラメタが過大推定されたと考えられる．

さらに，推定法どうしを比較すると，ベイズ推定法において，特に slip パラメタの Bias が大きくなっていることが読み取れる．この点について，Oka and Okada (2021)では小サンプルサイズ下において CRUM と LCDM における項目パラメタの事後分布の分散が大きかったことを報告しており，本研究においてもこれに由来して，CRUM において slip パラメタの値に非線形変換する際に(i.e., $s_j = 1 - \text{logit}^{-1}(\lambda_{j0} + \sum_{k=1}^{K} \lambda_{jk}\alpha_{ik}q_{jk})$)，その値が 0 もしくは 1 に近い値に偏ったことが原因と考えられる．一方で，guessing パラメタへの変換においては，切片パラメタのみが影響するため(i.e., $g_j = \text{logit}^{-1}(\lambda_{j0})$)，このような傾向は見られなかったと考えられる．

以上をまとめると，まず，小サンプルサイズ下で項目パラメタの推定精度が高いのは DINA モデルまたは DINO モデルといった倹約的なモデルと従来されてきたが，より現実に即していると考えられるデータ生成モデルの下では，いずれの推定法においても DINA モデルでは guessing パラメタ，DINO モデルでは slip パラメタがそれぞれ過大推定される傾向にあることが明らかとなった．slip パラメタおよび，guessing パラメタは項目の質を事後的に捉える上で重要な値となるが，このような結果は実践上，DINA モデルや DINO モデルに基づいた項目パラメタの解釈を行う際の留意点となりうる新たな知見であろう．また表 3 から，slip パラメタおよび guessing パラメタについて最尤推定法では.09 から.24 程度，ベイズ推定法では.07 から.18 の RMSE が観察された．このうち最小の RMSE の値を示したのが，$(N, J, K) = (80, 40, 4)$ の条件において分析モデルが CRUM や LCDM の場合であった．とりわけ CRUM は，アトリビュート習得パタンの観点からもほかのモデルと比べて LCDM と同等もしくはそれに匹敵する程度の高い推定精度を示していた．LCDM は項目で測定されるアトリビュートごとの主効果に加えて，アトリビュート間の可能な全ての交互作用効果をパラメタとして導入している．一方で，この交互作用効果パラメタの推定値は，たとえば，最尤推定法において $N = 1000, 10000$ の状況であっても，十分に安定しないことが指摘されている（Kunina-Habenicht et al., 2012）．この点も踏まえれば，本シミュレーションで採用したようなより現実に即していると考えられるデータ生成モデルの下での CRUM の有用性が示唆されたといえよう．

3.3.3 情報量規準の選択傾向

最尤推定法において AIC と BIC がそれぞれ選択したモデルの割合を示したのが図 4 である．全体的な傾向として，AIC では，真のモデルに近い CRUM が選択される割合が最も高く，サンプルサイズが大きくなるほどその傾向はより顕著に見られた．一方で，BIC においては，本研究で検討した 5 つの CDM のうち自由パラメタ数 d が最小である DINA モデルもしくは，



(a) AIC **(b) BIC**

図 4. AIC・BIC が最小となったモデルの割合.

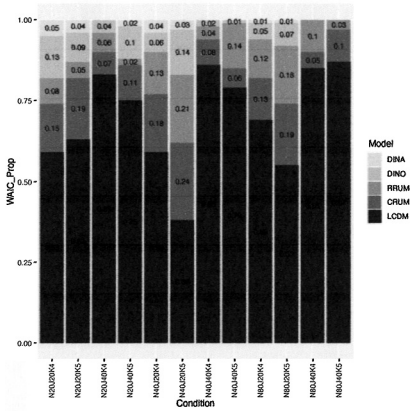


図 5. WAIC が最小となったモデルの割合.

DINO モデルが選択される傾向にあった．本シミュレーションで採用したようなより現実に即していると考えられるデータ生成モデルの下で，真のモデルに近い CRUM を選択した AIC の小サンプルサイズ下での有用性が示唆された．BIC は $N \geq 8$ において，AIC よりも一般に厳しい罰則を置くことになる．そのため，相対的に，サンプルサイズが $N = 20, 40, 80$ のいずれの条件においても BIC は AIC よりも倹約的なモデルを選択する傾向が見られたと考えられる．

次に，ベイズ推定法における情報量規準として WAIC が選択したモデルの割合を示したのが図 5 である．真のデータ生成モデルである LCDM および，RRUM と CRUM が選択される傾向にあった．図 4 において，WAIC と同じく汎化損失を最小とするモデルを選択する AIC よりも，LCDM を選択する傾向が見られた理由として，一般にベイズ推定法を階層構造をもつモデルに適用した場合に，仮により複雑な分析モデルを用いても汎化誤差が大きくなりにくいといった性質を反映していると考えられる（e.g., 渡辺, 2012）．

4. まとめ

4.1 本研究で得られた知見

本研究では，実際の教室場面での適用を想定した，CDM のアトリビュート習得パタンと項目パラメタの推定精度および，情報量規準のモデル選択傾向について検討した．特に，従来の小サンプルサイズを想定したシミュレーション研究において十分に検討されてこなかった，実際のデータ発生プロセスにより即していると考えられる一般性の高いモデルをデータ生成モデルとして設定し，下位モデルの推定精度と情報量規準のモデル選択傾向について，最尤推定法・ベイズ推定法・ノンパラメトリック推定法のそれぞれの推定法ごとに検討した．

結果として，最尤推定法およびベイズ推定法のいずれにおいても，真のデータ生成モデルである LCDM に比べて，その下位モデルである CRUM ではアトリビュート習得パタンの推定精度の水準が LCDM と同程度或いはそれに次いで高い傾向が見られた．さらに，情報量規準が選択するモデルについては，最尤推定法では AIC は CRUM を選択する割合が高く，BIC は最も倹約的なモデルである DINA モデルもしくは DINO モデルを選択する割合が高かった．ベイズ推定法における WAIC では，真のデータ生成モデルもしくは，CRUM もしくは RRUM といった真のデータ生成モデルに対して交互作用効果パラメタを 0 と制約したモデルが選択される割合が高かった．

CRUM の推定精度が真のデータ生成モデルである LCDM と同程度或いはそれに次いで高い傾向が見られたこと，および情報量規準 AIC において選択される傾向が見られたことから，今後の実践的活用では，一般的なデータ生成モデルが想定される状況において，従来のシミュレーション研究（e.g., Oka and Okada, 2021; Paulsen and Valdivia, 2021; Sen and Cohen, 2021）で推奨されてきた倹約的なモデルである DINA モデルではなく，より一般性の高いモデルである CRUM をむしろ基本的に利用していくべきであろう．

とりわけ，LCDM はアトリビュート間の全ての交互作用効果を項目パラメタに含んでおり，その数はアトリビュート数が増加に伴い組み合わせ論的爆発で増加することから，その設定および解釈が容易ではなくなっていく．一方で CRUM は項目パラメタとして，全てのアトリビュートを習得している場合の正答確率に対応する切片パラメタ（i.e., λ_{j0}）と項目の正答に必要とされるアトリビュートをそれぞれ習得している場合の正答確率の増分に対応する傾きパラメタ（i.e., λ_{jk}）のみをもつ．このようなパラメタは，実践的にも簡便に解釈が可能であるという特徴がある．たとえば，DINA モデルや DINO モデルでは，アトリビュートごとの正答確率への寄与の程度は項目パラメタから捉えることができなかったが，CRUM の推定結果をもとにすれば，各項目について，設定したアトリビュートのそれぞれがどの程度，正答確率に寄与しているかを捉えることができる．このような結果は，学校教師，より一般にテストの開発者がテストに含まれる項目の性質をより深く理解することができるようになり，今後のテスト改善にもより一層活きる可能性もあるだろう．

さらに，本研究で示した，モデルの誤設定の影響も踏まえた小サンプルサイズ下での情報量規準の選択傾向に関する知見は，今後，様々な選択肢がありうる CDM のモデル選択を行う際の参考になるだろう．たとえば，学校現場での CDM の活用時に，AIC は CRUM を選択している一方で，BIC は DINA モデルを選択しているといったケースが考えられる．実際のデータ発生プロセスとして DINA モデルや DINO モデルなどよりも一般性の高いモデルが想定される通常の状況では，小サンプルサイズ下では BIC は真のモデルよりもかなり倹約的な CDM を選択しやすい一方で，AIC は真のモデルやそれに近いモデルを選択する傾向にあるため，AIC の結果を踏まえたモデル選択が推奨されるだろう．

4.2 本研究の限界と今後の展望

4.2.1 その他の CDM の推定精度の検討の必要性

本研究で取り上げた CDM は，従来の応用研究で中心的に扱われてきた DINA モデル，DINO モデル，RRUM，CRUM，LCDM といった項目反応が 2 値で潜在変数が離散であるモデルであった．一方で，近年の CDM 研究の隆盛に伴い，CDM に属する様々な拡張モデルの提案がなされている（レビューとして，山口・岡田, 2017）．たとえば，複数時点における学習者のアトリビュート習得パタンの変化を捉えるための縦断的なモデル（レビューとして，Zhan, 2020），多枝選択式解答に適したモデル（e.g., de la Torre, 2009; Ozaki et al., 2020），複数のアトリビュートの情報を縮約する高次の連続的な潜在変数を仮定したモデル（e.g., de la Torre and Douglas, 2004），アトリビュートを連続的な潜在特性値として表現するモデル（Zhan et al., 2018; 丹・岡田, 2020）が挙げられる．このようなモデルの実践的な活用可能性が示唆されてはいるものの，学校現場を想定した小サンプルサイズ下における推定精度は必ずしも明らかにされていない．今後は，このようなモデルを含めた検討が必要だろう．

4.2.2 その他の推定方法との比較の必要性

本研究では，従来のシミュレーション研究の多くで用いられてきた最尤推定法だけでなく，ベイズ推定法および，小サンプルサイズ下での活用を意図したノンパラメトリック推定法の 3 つを取り上げた．ただし，ノンパラメトリック推定法によって推定を行ったモデルは，NPCD パッケージで実行可能な DINA モデルおよび DINO モデルに留まっていた．近年，ノンパラメトリック推定法に基づく新たな CDM の開発（e.g., Chiu et al., 2018, Wang et al., 2023）も行われており，たとえば，Chiu et al.（2018）は一般ノンパラメトリック推定法を提案している．今後は，このような推定法も含めた検討が必要だろう．

また本研究では，ベイズ推定法における事前分布として無情報事前分布，あるいは弱情報事前分布を設定した．一方で，ベイズ推定法の特徴として，事前の情報や分析者の知識・信念を事前分布の形で推定に組み込める点が挙げられる（Gelman et al., 2013）．つまり，学校の教師が普段の教育実践を通して経験的に有している情報を事前分布として設定することで，アトリビュート習得パタンおよび項目パラメタの推定精度を高めることができる可能性がある．たとえば，各アトリビュートの習得状況について，事前にどの程度の割合の学習者が習得しているという教師のみとりに基づいて，その情報を混合比率パラメタに対する事前分布として設定することが考えられる．また，学校の定期テストでは，前年度に当該単元を対象として実施されたテスト項目の一部を再び使用することもあり得る．その場合，前年度の推定結果を項目パラメタの事前分布を通して活用することも考えられる．

4.2.3 CDM の実践的活用に向けて

本研究で示した結果は，今後の学校現場での実践の参考になり得る一方で，そのほかの課題も考えられる．たとえば，山口・岡田（2017）が「Q 行列の設定は，文献調査や項目分析，複数の専門家の合議などを経て行われる時間と労力のかかるプロセスである」と述べているように，Q 行列の設定自体にコストを伴うことが考えられる．また，学校教師は必ずしも統計ソフトウェアに精通しているわけではないことを踏まえれば，CDM による分析を簡便に実施できない点も実践上の問題点として考えられる．

このような問題点を克服して CDM が活用されていくためには，心理統計学者を含む研究者が学校教師による実践的活用を支えていく必要がある．たとえば，Q 行列設定の方法を学校教師とともに協議していくことや，CDM による分析を学校教師がより簡便に実行できる Web ツールや GUI の開発が，実践的活用に向けた足場かけとして重要である．

心理統計学の分野では，CDM をはじめとして学校現場での活用を意識した学習・指導改善

に資する統計モデルの開発やその周辺の方法論的な研究が盛んに行われている．しかし，学校現場での活用事例はまだ限られたものであり，実際の方法論的研究の隆盛に比して少なからずギャップが生じている．このようなギャップを埋めるために今後，心理統計学者と学校教師の協働がますます重要となるであろう．

注．

本稿は第一著者の修士論文の一部を大幅に加筆・修正したものである．

謝　辞

本稿の執筆にあたり，2 名の査読者の先生方から多くの示唆に富むコメントをいただきました．ここに記して，深く御礼申し上げます．

参 考 文 献

Akaike, H. (1974). A new look at the statistical model identification, *IEEE Transactions on Automatic Control*, **19**(6), 716–723.

Başokçu, T. O. (2014). Classification accuracy effects of Q-matrix validation and sample size in DINA and G-DINA models, *Journal of Education and Practice*, **5**(6), 220–230.

Chiu, C.-Y. and Douglas, J. (2013). A nonparametric approach to cognitive diagnosis by proximity to ideal response patterns, *Journal of Classification*, **30**, 225–250.

Chiu, C.-Y., Sun, Y. and Bian, Y. (2018). Cognitive diagnosis for small educational programs: The general nonparametric classification method, *Psychometrika*, **83**, 355–375.

DeCarlo, L. T. (2011). On the analysis of fraction subtraction data: The DINA model, classification, latent class sizes, and the Q-matrix, *Applied Psychological Measurement*, **35**(1), 8–26.

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. (2009). A cognitive diagnosis model for cognitively based multiple-choice options, *Applied Psychological Measurement*, **33**(3), 163–183.

de la Torre, J. (2011). The generalized DINA model framework, *Psychometrika*, **76**, 179–199.

de la Torre, J. and Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis, *Psychometrika*, **69**(3), 333–353.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. and Rubin, D. B. (2013). *Bayesian Data Analysis*, CRC Press, Florida.

南風原朝和 (2014).『続・心理統計学の基礎 統合的理解を広げ深める』, 有斐閣, 東京.

浜田 宏, 石田 淳, 清水裕士 (2019).『社会科学のためのベイズ統計モデリング』, 朝倉書店, 東京.

Hartz, S. and Roussos, L. (2008). The fusion model for skills diagnosis: Blending theory with practicality, *ETS Research Report Series*, **2008**(2), i–57.

Hartz, S. M. (2002). A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality, Unpublished Doctoral Dissertation, University of Illinois at Urbana-Champaign, Illinois.

Henson, R. A. (2009). Diagnostic classification models: Thoughts and future directions, *Measurement: Interdisciplinary Research and Perspectives*, **7**, 34–36.

Henson, R. A., Templin, J. L. and Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables, *Psychometrika*, **74**, 191–210.

Huebner, A. and Wang, C. (2011). A note on comparing examinee classification methods for cognitive diagnosis models, *Educational and Psychological Measurement*, **71**(2), 407–419.

Hueying, T. and Yang, Y.-H. (2019). Improved performance of model fit indices with small sample sizes in cognitive diagnostic models, *International Journal of Assessment Tools in Education*, **6**(1), 154–169.

池田 央 (2013). テストの過去, 現在, そして未来の形を考える—その方向性と必要性—, 日本テスト学会誌, **9**(1), 1–14.

Jang, E. E. (2005). A validity narrative: Effects of reading skills diagnosis on teaching and learning in the context of NG TOEFL, Unpublished Doctoral Dissertation, University of Illinois at Urbana-Champaign, Illinois.

Junker, B. W. and Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory, *Applied Psychological Measurement*, **25**(3), 258–272.

Kim, J. K. and Nicewander, W. A. (1993). Ability estimation for conventional tests, *Psychometrika*, **58**, 587–599.

Kunina-Habenicht, O., Rupp, A. A. and Wilhelm, O. (2012). The impact of model misspecification on parameter estimation and item-fit assessment in log-linear diagnostic classification models, *Journal of Educational Measurement*, **49**(1), 59–81.

Levy, R. and Mislevy, R. J. (2016). *Bayesian Psychometric Modeling*, CRC Press, Florida.

Ma, W. and de la Torre, J. (2020). GDINA: An R package for cognitive diagnosis modeling, *Journal of Statistical Software*, **93**, 1–26.

Maris, E. (1999). Estimating multiple classification latent class models, *Psychometrika*, **64**, 187–212.

Merkle, E. C., Furr, D. and Rabe-Hesketh, S. (2019). Bayesian comparison of latent variable models: Conditional versus marginal likelihoods, *Psychometrika*, **84**, 802–829.

文部科学省 (2019). 学習評価のあり方ハンドブック 小・中学校編, https://www.nier.go.jp/kaihatsu/pdf/gakushuhyouka_R010613-01.pdf（最終アクセス日 2024 年 3 月 7 日）.

Oka, M. and Okada, K. (2021). Assessing the performance of diagnostic classification models in small sample contexts with different estimation methods, arXiv preprint, https://doi.org/10.48550/arXiv.2104.10975.

Ozaki, K., Sugawara, S. and Arai, N. (2020). Cognitive diagnosis models for estimation of misconceptions analyzing multiple-choice data, *Behaviormetrika*, **47**(1), 19–41.

Paulsen, J. and Valdivia, D. S. (2021). Examining cognitive diagnostic modeling in classroom assessment conditions, *The Journal of Experimental Education*, **90**(4), 916–933.

Plummer, M. et al. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling, *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*, **124**, 1–10, Vienna, Austria.

Rupp, A. A., Templin, J. and Henson, R. A. (2010). *Diagnostic Measurement*: *Theory*, *Methods*, *and Applications*, Guilford Press, New York.

佐宗 駿, 岡 元紀, 柴 里実, 植阪友理 (2022). 理解の深さの定量的評価とそのつまずきに応じた学習方略指導—認知診断モデルの実践的応用と生徒の反応—, 日本テスト学会第 20 回大会発表論文抄録集, 116–119.

Saso, S., Oka, M. and Uesaka, Y. (2023). Development of assessment tools for depth of understanding quantitatively with cognitive diagnostic models, *Advances in Information and Communication*: *Proceedings of the 2023 Future of Information and Communication Conference* (*FICC*), Volume 1, 766–774, Springer, Cham.

佐宗 駿, 岡 元紀, 植阪友理 (2023). 認知診断モデルを活用した理解の深さの診断と定期テストへの応用：定性的・定量的な Q 行列の設定とモデルの実践的有用性の検討, 認知科学, **30**(4), 515–530.

Schwarz, G. (1978). Estimating the dimension of a model, *The Annals of Statistics*, **6**(2), 461–464.

Sen, S. and Cohen, A. S. (2021). Sample size requirements for applying diagnostic classification models, *Frontiers in Psychology*, **11**, https://doi.org/10.3389/fpsyg.2020.621251.

Sessoms, J. and Henson, R. A. (2018). Applications of diagnostic classification models: A literature review and critical commentary, *Measurement*: *Interdisciplinary Research and Perspectives*, **16**(1), 1–17.

Su, Y.-S. and Yajima, M. (2021). R2jags: Using R to run 'JAGS', https://cran.r-project.org/web/

packages/R2jags/R2jags.pdf（最終アクセス日 2024 年 3 月 7 日）.
丹 亮人, 岡田謙介 (2020). 連続型の特性値をもつ補償型認知診断モデル, 日本テスト学会誌, **16**(1), 31–44.
Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory, *Journal of Educational Measurement*, **20**(4), 345–354.
Templin, J. L. and Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models, *Psychological Methods*, **11**(3), 287–305.
Uesaka, Y., Saso, S. and Akisawa, T. (2021). How can we statistically analyze the achievement of diagrammatic competency from high school regular tests?, *Diagrammatic Representation and Inference*: *12th International Conference, Diagrams 2021, Virtual, September 28–30, 2021, Proceedings 12*, 562–566, Springer, Cham.
Vrieze, S. I. (2012). Model selection and psychological theory: a discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC), *Psychological Methods*, **17**(2), 228–243.
Wang, D., Ma, W., Cai, Y. and Tu, D. (2023). A general nonparametric classification method for multiple strategies in cognitive diagnostic assessment, *Behavior Research Methods*, Advance online publication, https://doi.org/10.3758/s13428-023-02075-8.
Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory, *Journal of Machine Learning Research*, **11**(12), 3571–3594.
渡辺澄夫 (2012).『ベイズ統計の理論と方法』, コロナ社, 東京.
Xu, G. (2017). Identifiability of restricted latent class models with binary responses, *The Annals of Statistics*, **45**(2), 675–707.
Yamaguchi, K. (2023). On the boundary problems in diagnostic classification models, *Behaviormetrika*, **50**(1), 399–429.
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.
Yamaguchi, K. and Templin, J. (2022). A Gibbs sampling algorithm with monotonicity constraints for diagnostic classification models, *Journal of Classification*, **39**(1), 24–54.
Zhan, P. (2020). Longitudinal learning diagnosis: Minireview and future research directions, *Frontiers in Psychology*, **11**, https://doi.org/10.3389/fpsyg.2020.01185.
Zhan, P., Wang, W.-C., Jiao, H. and Bian, Y. (2018). Probabilistic-input, noisy conjunctive models for cognitive diagnosis, *Frontiers in Psychology*, **9**, https://doi.org/10.3389/fpsyg.2018.00997.
Zhan, P., Jiao, H., Man, K. and Wang, L. (2019). Using JAGS for Bayesian cognitive diagnosis modeling: A tutorial, *Journal of Educational and Behavioral Statistics*, **44**(4), 473–503.
Zheng, Y., Chiu, C. and Douglas, J. (2019). NPCD: Nonparametric methods for cognitive diagnosis, R package version, 1.0–11, https://cran.r-project.org/web/packages/NPCD/NPCD.pdf（最終アクセス日 2024 年 3 月 7 日）.

Examining Estimation Accuracy of Cognitive Diagnostic Models in Classroom Contexts —Focusing on the Impact of Model Misspecification and Different Estimation Methods—

Shun Saso[1], Motonori Oka[2] and Satoshi Usami[1]

[1]Faculty of Education, University of Tokyo
[2]Department of Statistics, London School of Economics and Political Science

Cognitive diagnostic models (CDMs) are promising psychometric models that can estimate students' mastery of specific learning elements referred to as attributes. Despite their potential to provide diagnostic information that aids students' learning and teachers' instruction, practical applications of CDMs in educational settings for classroom assessment have been severely limited. This limitation is partly due to an insufficient exploration of the estimation accuracy of CDMs and the behavior of model selection based on information criteria in classroom applications. In the present study, we investigated the accuracy of different estimation methods and evaluated the corresponding information criteria under a simulation design that resembles classroom contexts. Whereas previous works involved simulation studies in which a model fitted to data and the true model behind data generation were the same, the present study examines the effect of model misspecification. In particular, this study considers the log-linear cognitive diagnostic model (LCDM), which is one of the generalized CDM versions, as a model used to generate data for evaluating the performance of not only the LCDM but also its submodels, including the DINA (deterministic inputs, noisy "and" gate) model, DINO (deterministic inputs, noisy "or" gate) model, RRUM (reduced reparameterized unified model), and CRUM (compensatory reparameterized unified model). The key findings are summarized as follows: (1) The CRUM, which assumes that each attribute mastery independently affects item-correct probabilities, exhibited the highest estimation accuracy for attribute mastery patterns; its accuracy was comparable to or slightly less than that of the true-data-generating model in both the maximum likelihood estimation (MLE) method and the Bayesian estimation method. (2) An increase in the number of items and a decrease in the number of attributes improved the estimation accuracy of the attribute mastery patterns, whereas an increase in the sample size did not result in such an improvement. (3) In the MLE method, the Akaike information criterion (AIC) most frequently preferred the CRUM, whereas the Bayesian information criterion (BIC) showed a preference for the most parsimonious CDMs, implying the recommendation of model selection based on the AIC results. In Bayesian estimation, the widely applicable information criterion (WAIC) most frequently preferred CDMs that best approximated the true data-generating model.

Key words: Cognitive diagnostic models, small sample size, Bayesian estimation method, information criteria, formative assessment.