統計数理 Vol. 72, No. 1

（通巻139号）

PROCEEDINGS OF THE INSTITUTE OF STATISTICAL MATHEMATICS

目　次

特集「心理統計学の新展開」

2024年6月

大学共同利用機関法人 情報・システム研究機構 統計数理研究所

〒190-8562 東京都立川市緑町 10-3　　電話 050-5533-8500（代）

本号の内容はすべて https://www.ism.ac.jp/editsec/toukei/ からダウンロードできます

ISSN 0912-6112

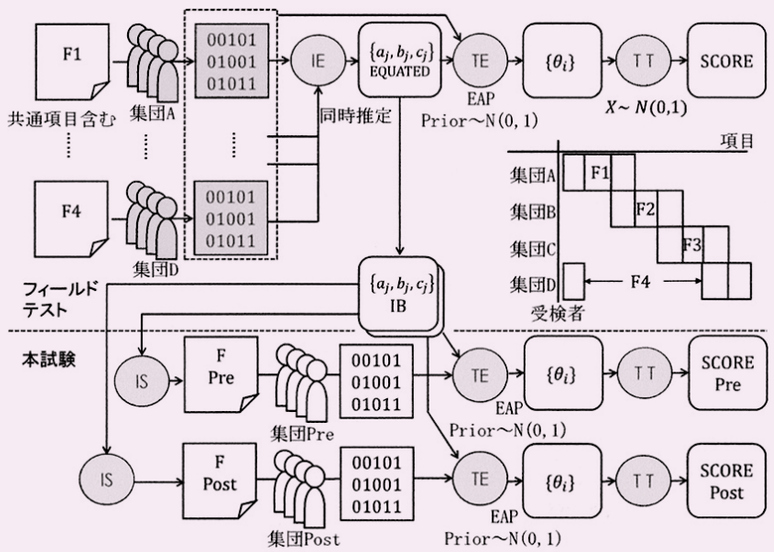
統計数理

PROCEEDINGS OF THE INSTITUTE OF STATISTICAL MATHEMATICS

第72巻 第1号 2024

統計数理 Vol. 72, No. 1

PROCEEDINGS OF THE INSTITUTE OF STATISTICAL MATHEMATICS



統計数理研究所

統計数理

（年 2 回発行）

「統計数理」は，統計数理研究所における研究成果を掲載する統計数理研究所「彙報」として 1953 年に歴史を始め，1985 年に誌名を変更し今の形となりました．現在は，統計数理研究所の研究活動に限らず，広く統計科学に関する投稿論文を掲載し，統計科学の深化と発展，そして統計科学を通じた社会への貢献を目指しています．

投稿を受け付けるのは，次の 6 種です．

a. 原著論文　　b. 総合報告　　c. 研究ノート
d. 研究詳解　　e. 統計ソフトウェア　　f. 研究資料

投稿された原稿は，編集委員会が選定・依頼した査読者の審査を経て，掲載の可否を決定します．投稿規程，執筆要項は，本誌最終頁をご参照ください．

また，上記以外にも統計科学に関して編集委員会が重要と認める内容について，編集委員会が原稿作成を依頼することがあります．

その他，「統計数理」に関するお問い合わせは，各編集委員にお願いします．

All communications relating to this publication should be addressed to associate editors of the Proceedings.

大学共同利用機関法人　情報・システム研究機構
統計数理研究所
〒190-8562 東京都立川市緑町 10-3　　電話 050-5533-8500（代）
https://www.ism.ac.jp/

印刷：笹氣出版印刷株式会社

表紙の図は本誌 73 ページを参照

PROCEEDINGS OF THE INSTITUTE OF STATISTICAL MATHEMATICS

Vol. 72, No. 1

Contents

June, 2024
Research Organization of Information and Systems
The Institute of Statistical Mathematics
10–3 Midori-cho, Tachikawa, Tokyo 190–8562, JAPAN

統計数理(2024)
第 72 巻 第 1 号 1–2

「特集・心理統計学の新展開」について

岡田 謙介[1]・持橋 大地[2]（オーガナイザー）

心理統計学は，人間の心と行動のメカニズムを探求する心理学において，データの収集，分析，解釈のために必要となる統計学的方法を研究する分野である．そのスコープに含まれる統計モデルは多岐にわたるが，とくに人の能力，態度，性格特性，精神的健康といった直接観測できない構成概念を，質問項目への解答・回答データに基づいて測定する方法を大きく発展させてきたことは，心理統計学を特徴づける要素である．経済現象の解明や予測のために用いる統計モデルを扱う研究分野は計量経済学(econometrics)であり，生物学的現象の定量的理解に関する統計モデルを扱う研究分野は計量生物学(biometrics, 生物測定学とも)である．これらと同様に，心理学的構成概念の測定のための理論と方法を提供する統計モデルを扱う研究分野は計量心理学(psychometrics, 心理測定学とも)と呼ばれる．

計量心理学の統計モデル研究は，Spearman (1904)による，各教科のテスト成績の背後に想定された一般知能因子を抽出する因子分析(factor analysis)法の開発に始まる．1950 年代からは，問題項目への正答・誤答のようなカテゴリカルな観測データに基づき，解答者の能力や特性を推定する項目反応理論(item response theory)モデルが発展し，様々な教育テストに応用されるようになった．また 1980 年代からは，解答者の知識習得状態を詳細に分析し，個々の学習や成長の進行状況を評価し，ニーズに合わせたフィードバックを提供するための診断分類モデル(diagnostic classification model，または認知診断モデル cognitive diagnostic model)の研究が進展した．

因子分析，項目反応理論，診断分類モデルに代表される，心理学的構成概念の測定のための統計モデルは測定モデル(measurement model)と総称される．心理学的構成概念を適切に定義し，測定することは，心理学研究を行う上で重要な課題であり，その実現には信頼性と妥当性の高い測定尺度の開発が求められる．ここで信頼性は測定が一貫性のある結果をもたらす程度を，妥当性は測定が実際に意図した構成概念を捉えている程度を指す．本特集号が対象とするのは，心理統計学の中でもとくに，こうした計量心理学の測定モデルである．

測定モデルは，教育評価，臨床診断，人事選考といった場面での社会的影響力の大きなテストや，日常的な場面での形成的評価や医療的スクリーニングのためのテストに応用されて，現代の社会や教育，医療・公衆衛生を支えている．心理測定をとりまく近年の変化として，コンピュータで実施されるテストの大幅な増加が挙げられる．たとえば文部科学省 CBT (computer-based testing)システム MEXCBT(メクビット)は，2024 年 2 月現在で公立小学校の 80% 超，公立中学校のほぼ全てに導入された．これによって，わが国でも義務教育段階から，コンピュータを用いて各種テストを実施することが当たり前になりつつある．自然言語処理や機械学習など，計量心理学の関連分野における技術の飛躍的進歩は，テストの作問や採点に大きな革新をもたらしている．一方で，デジタル技術や社会のネットワーク化の進展は，心理測

[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] 統計数理研究所：〒 190–8562 東京都立川市緑町 10–3

定における倫理やプライバシーといった新たな課題を浮き彫りにもしてきた．また，大規模言語モデルが知能と呼びうるような驚くべき性能を発揮することが明らかになった中で，各種言語モデルの「能力」をどのように測定・評価するかという新たな研究課題も生じており，テスト「解答者」は人だけでなく機械をも含むようになった．このように，測定モデル研究は，従来の枠組みを超えた新しい展開の時を迎えている．

こうした中で，本特集号では計量心理学の統計モデル研究における最近の発展を扱う 7 本の論文を掲載する．冒頭 3 篇は，情報技術と測定モデルの融合に焦点を当てたものである．加藤論文では，近年の技術的発展を踏まえた，コンピュータで実施されるテストの解答データを対象とする測定モデルの新たな展開が紹介され，議論される．江原論文では，問題項目が自然言語で記述されることを踏まえ，テキスト情報に基づいて推定される項目の困難度と項目反応理論との関わりが論じられる．また著者により開発された方法をテストデータに適用した結果が示される．渕本・植野論文では，e-Testing における項目の露出をコントロールしながら同程度の測定精度を実現するための方法が提案され，その性能を検証した数値実験結果が報告される．

続く 4 篇の論文は，測定モデルの重要な一側面に関する近年の研究動向をまとめたレビューを提供し，また実践例や数値実験結果を報告するものである．光永論文では，複数のテスト間の対応付けを行うための等化やリンキングの方法がレビューされ，学力調査における分析結果が報告される．国里・竹林論文では，米国国立衛生研究所（NIH）が主導する患者報告アウトカム尺度 PROMIS の尺度開発プロセスが解説され，わが国における課題が論じられる．山口論文では，診断分類モデルの各種推定法が網羅的にレビューされ，現代的な課題と今後の展望が論じられる．また佐宗・岡・宇佐美論文では，小サンプル状況下における推定法間の比較が数値実験によって行われ，議論される．

最後になるが，本特集号が編纂されるきっかけは，岡田が研究者として，持橋が領域アドバイザーとして参画する JST 戦略的創造研究推進事業（さきがけ）研究領域「信頼される AI の基盤技術」における出会いと議論にあった．同研究領域を発端として，このような貴重な機会に恵まれたことに，感謝の意を表したい．本特集号が心理統計学への理解や関心を広げ，その研究を一層前進させる一助となることを願ってやまない．

参 考 文 献

Spearman, C. (1904). ‘General intelligence,’ objectively determined and measured, *The American Journal of Psychology*, **15**(2), 201–293, https://doi.org/10.2307/1412107.

統計数理(2024)
第 72 巻 第 1 号 3–21

特集「心理統計学の新展開」
[総合報告]

技術強化型テスティングにおける測定モデルの考察と展望

加藤 健太郎†

(受付 2023 年 12 月 4 日；改訂 2024 年 3 月 7 日；採択 3 月 11 日)

要　　旨

コンピュータで実施するテスト(CBT)は教育アセスメントにおける主要なテスト実施形態となりつつあり，従来の紙筆式テスト(PBT)と比較して，コンピュータ上で実施することによる様々な利点，特に技術強化型項目(CBT だからこそ実現できるテスト問題)と，学習者の学習状態をよりよく測定する目的でのその利用に注目が集まっている．本稿では技術強化型項目から得られる様々なデータを心理測定学の立場から分析・モデリングする近年の試みや実践のレビューを行い，現在の課題や今後の研究の方向性について検討した．その結果，(a)TEIs で実現できる様々な解答形式の測定論的性質への影響，(b)プロセスデータの利用可能性に関する探索的検討，(c)TEIs を含む尺度と既存の測定尺度との等価性の検討，(d)プロセスデータを利用した解答過程の測定モデリングに関する研究が近年盛んに行われていることがわかった．今後の課題として，特にプロセスデータを用いた解答過程のモデリングについて，学習改善につながる有用な情報を提供していくことができるか，TEIs を含むテストの継続的・安定的な運用を実現できるかという観点で継続的な検討・改善が必要があると考えられた．

キーワード：コンピュータによるテスト, 項目反応理論, 技術強化型項目, プロセスデータ.

1. 問題と目的

近年の教育テストの動向は，測定内容の変化と測定方法の変化によって特徴づけられる(Kato, 2016)．測定内容の変化とは，学校教育において習得すべきものが，従来の教科・科目固有の知識から，現実的な問題解決の場面においてそれらを総合的に応用して問題を考察・解決できるかという，いわゆる「コンピテンシー」にシフトしており，教育テストが測定する対象もこれに対応して変化していることを指す．こうした背景には，DeSeCo のキー・コンピテンシー(Rychen and Salganik, 2003)や 21 世紀型スキルの指導とアセスメント (Care et al., 2018)，経済協力開発機構(OECD)が実施する PISA(Programme for International Student Assessment)といった，欧米における「これからの社会に必要とされるスキル」の定義と測定について再考する動きがあった．2020 年から順次導入されている日本の新しい学習指導要領において，「思考力・判断力・表現力」が資質・能力の要素の一つとして挙げられていることも，こうした世界的な潮流を反映していると言える．

もう一方の測定方法の変化の第一に挙げられるのは，従来の紙筆式(paper-based test [PBT])

† ベネッセ教育総合研究所：〒 206–8686 東京都多摩市落合 1–34

からコンピュータで実施するテスト(computer-based test [CBT])への移行が進んでいることである．例えば，日本国内においては，2019 年に始まった文部科学省の GIGA スクール構想の下で教育のデジタル化が進められ，これに並行して文部科学省 CBT システム(MEXCBT)が整備されつつある．文部科学省が実施する全国学力・学習状況調査や，大学入試センターが実施する大学入学共通テストといった公的な試験においても CBT の導入が検討されており(前者においては 2024 年度より順次導入が始まる)，今後は民間業者が提供するテストや日常の学習ツールも含めて CBT がさらに普及することが予想される．海外に目を向ければ，日本の全国学力・学習状況調査に相当する全米学力調査 NAEP(national assessment of educational progress)は早期に CBT の検討や試行を始めている(現在は digitally-based assessment と呼ばれている；https://nces.ed.gov/nationsreportcard/dba/)．また，先に挙げた OECD の PISA においても 2015 年より本格的に CBT への移行が進められている(こうした世界的な動きを受けて，日本でも CBT 導入の検討が加速したと言える)．

CBT の普及に伴い，単にテストをコンピュータ上で実施することにとどまらず，様々な情報技術がテストの開発・運用の各フェーズにおいて活用されるようになってきている．例えば，分寺 (2023)は，CBT に関する近年の研究トピックを(a)PBT と CBT の得点の比較可能性(モード効果)の検証，(b)適応型テスト(computerized adaptive testing [CAT])，(c)新しい形式のテスト項目，(d)オンライン試験における不正行為とその対策，(e)ログデータの活用，(f)(障がいを持つ受検者らへの)特別な配慮の 6 つに整理している．

Bennett (2015)は，現代における CBT の進化を 3 つの段階に分類・整理している．第 1 世代のテストは，従来の PBT をコンピュータ上で実施・受検できるように移植したものである．PBT 版の項目がそのまま画面に表示され，解答にはマウス操作やキーボードでのタイピングが必要となるものの，項目の形式自体は PBT と同じである．技術面では，項目反応理論(item response theory [IRT])による尺度構成に基づく，適応型の出題(adaptive testing)機能が搭載されたものもある．

第 2 世代のテストを特徴づけるのは，後で述べる技術強化型項目(technology-enhanced items [TEIs])，いわゆる「CBT ならではのテスト問題」の導入である．これにより，先に述べた測定内容の変化と合わせて，従来の形(第 1 世代やそれ以前の PBT)では難しかった能力やスキルの測定が可能になるという期待が高まっている．また，テスト項目の形式に加えて，テストの開発・運用のプロセスの中にも情報技術が積極的に導入される．本稿では扱わないが，近年では特に自然言語処理と機械学習(AI)を活用したテスト問題の自動生成や論述・口述解答の自動採点に注目が集まっており，例えば Jiao and Lissitz (2020)のような研究事例集でもこれらが主要なトピックとして取り上げられている．第 2 世代のテストに見られるような，CBT を前提として情報技術を積極的に取り入れたテストは技術強化型テスト(technology-enhanced tests)と呼ばれる．現在運用されている CBT の多くは第 1 世代であり，一部は第 2 世代に向けて進化しつつあると言われている．

これらに続く第 3 世代のテストを，Bennett (2015)は次世代アセスメント(next-generation assessment)と呼んでいる(本稿では，「テスト」と「アセスメント」という言葉を同義に用いる)．次世代アセスメントは，第 2 世代同様にテクノロジーの積極的な活用を前提としているが，オンラインの学習環境を前提として，学習の評価ではなく学習への貢献により大きな比重を置いて設計される点(assessment of learning から assessment for learning への転換)が強調される．例えば，米国教育省の報告書 (U.S. Department of Education, 2017, Sec. 4) では，学習の改善をゴールとして適切に技術を活用することが謳われており，その中で表 1 のような形で従来型のアセスメントと次世代アセスメントとの対比がなされている．

本稿では，大規模教育テストに見るこうした動向を踏まえて，能力測定の核となる測定モデ

表 1．従来型と次世代型のアセスメントの比較．

	従来型	次世代型
実施のタイミング	学習後	学習中
アクセシビリティ	限定的	誰でも
学習（テスト）の進行	定型	適応的
結果の返却	時間差あり	即時
項目の形式	汎用型	強化型

ルの現状と課題について考えたい．従来の PBT，あるいは第 1 世代の CBT においては，各テスト項目に対する解答を正答 1，誤答 0 の二値で採点し（項目得点），これらをテスト項目全体に渡って合計した正答数（素点）や正誤のパタンから推定される IRT の能力母数を受検者の能力の「測定値」あるいは「推定値」としてきた（cf. Lord, 1952）．しかし，CBT の特性を活かしたテスト項目（i.e., TEIs）の作成やデータ収集が可能となってきたことで，こうした従来の枠組みを超えた測定の可能性が開けてきている．本稿ではまず，TEIs の特徴を概観し，能力測定に関して TEIs がどのようなデータを提供し得るかを考察する．そして，これを踏まえた上で，TEIs の利用を念頭に置いて測定論の文脈で行われてきた分析やその中で応用・提案されてきた測定モデルのレビューを行う．

なお，本レビューでは以下の手順によって文献を収集した．検索には ERIC, PubMed, Google Scholar，および論文検索・閲覧サービスの DeepDyve を主に用いた．基本的な検索ワードとして “technology-[enhanced, based, rich]”，“[measurement, response] model”，“innovative [items, assessment]” の用語の組み合わせを用い，特定のモデルに関する検索結果を適宜追加した．得られた文献リストから，タイトルおよびアブストラクトを参照して，概ね 2020 年以降に出版された，本稿のテーマに合致すると思われる比較的新しい論文を主な検討の対象とした．

2. 技術強化型項目

2.1 技術強化型項目の定義と特徴

技術強化型項目（TEIs）は，広義には「従来の多枝選択式や解答構築式ではない，コンピュータで実施されるテスト項目」とされている（Bryant, 2017, p. 2）．より詳しい定義や呼称については研究者によってバリエーションがあるが（e.g., technology-based items, technology-enabled items, innovative items），主に (a) 解答形式，(b) 解くべき問題（タスク）の特徴，(c) 解答補助ツールの 3 つの観点から特徴づけることができる．

第一の特徴である TEIs で用いられる解答形式に関して，Jiang et al.（2021）は，2017 年の NAEP において採用された例として，タイピングによる記述解答の入力（文字だけでなく，数式エディタも含む），多枝多選択方式（複数選択可能な解答選択式；multiple-selection multiple-choice format），ドラッグ&ドロップ（解答要素を所定の箇所に配置する＝マッチングや順序付けなど），ドロップダウンによる穴埋め，グラフやイラストの領域指定による解答を挙げている．この他にも，言語テストでは音声入力（スピーキングテスト）が従来より用いられており（ただし，TEIs では記録方式がアナログからデジタルになる），さらに今後は，画像や動画などを含む新しい解答形式が出てくる可能性もある．Bryant（2017）は，解答の自由度（制約の強さ）の観点から，より制約が強い順に (a) 多枝選択式（multiple-choice；所与の選択枝から選ぶ），(b) 選択・同定（selection/identification；テキストの一部を指示する，グラフの領域を指定するなど，「選択枝」よりも対象が広いものを指す），(c) 要素の並べ替え（reordering/rearranging），(d) 代入・修正（substitution/correction；既にある要素を置き換えたり変更したりする），(e) 完成（completion；穴埋め形式など），(f) 構築（construction；記述・論述など），(g) プレゼンテー

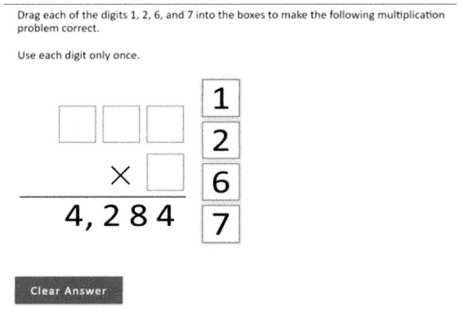


図 1. NAEP 数学テストにおけるドラッグ&ドロップ形式の項目例(U.S. Department of Education. Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress [NAEP], 2017 Mathematics Digitally-Based Assessment; https://www.nationsreportcard.gov/math_2017/sample-questions/?grade=8).

ション(presentation;意図やストーリーを持って行う表現)の 7 つのカテゴリで TEIs の解答形式の分類を試みている.

第二の TEIs のタスクの特徴として,特にコンピテンシーの測定において,より日常生活・社会生活に近い具体的な状況(文脈・シナリオ)を設定し,リアリティ(テストの分野では,特に真正性 [authenticity] と呼ばれる)の高いタスクやタスク素材を提示できるという点が挙げられる. Wools et al. (2019)は,TEIs に特徴的なタスクの形式をシミュレーション型(条件を様々に変えて実験を行い,結果を考察する),マルチメディア型(動画や音声の利用),ハイブリッド型(複数の形式の混合)の 3 種類に分類しているが,これらに共通しているのは,CBT 内で設定された現実を模した環境・状況の中で,受検者がインタラクティブなやりとりをしながら解答に至るタスクを実行していく点であろう. 例えば,CBT で実施された 2022 年の PISA の「数学的リテラシー」アセスメントでは,スプレッドシートを提示して計算式を考えさせたり得られた数値を考察する,地図を表示して所定の箇所をマウスでポイントし,ポイントした各点から目的地までの距離を表示させてその特徴を考察する,期間・利率・元金などを指定して積立貯金のシミュレーションを行わせる,といったタスクの例が示されている (OECD, 2023).

第三の特徴に関して,上述したタスクを実行する際の補助ツールとして計算機,デジタルメモ,デジタル定規,グラフ描画ツール等がデジタル環境内でシームレスに利用できることが挙げられる.

公開されている TEIs の例を図 1 および 2 に示す. 図 1 は NAEP の第 8 学年の数学の項目例で, Jiang et al. (2021)が分析に用いたものである. この項目におけるタスクは,筆算が成立するように 4 つの空欄のそれぞれに「1」「2」「6」「7」のいずれかの数字をドラッグ&ドロップすることである. ドラッグ&ドロップ形式は,大規模テストプログラムにおいて公開されている TEIs の中で最もよく使われている解答形式である (Russell and Moncaleano, 2019).

図 2 は, PISA 2012 で実施された創造的課題解決力(creative problem-solving)アセスメントにおける項目例で, Han et al. (2022)が分析に用いたものである. これは,鉄道駅の自動券売機で指定された条件に合う切符を購入するというタスクであり,画面の右側に示された架空の券売機のボタンをクリックして操作を行う. まず,乗車する鉄道網(地下鉄か広域鉄道か),料金種別(通常か割引か),切符の種類(1 日乗車券か片道か)を決めて,最後に「購入」ボタンを押して購入を完了する.

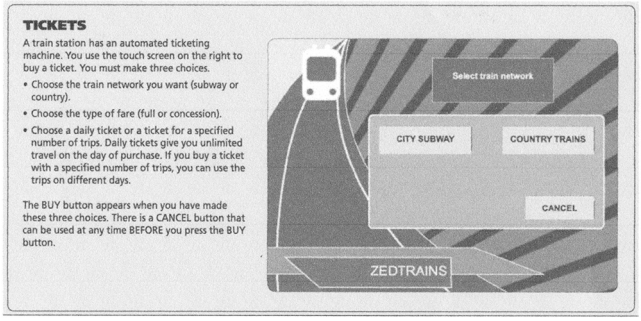


図 2. PISA 2012 における創造的問題解決の項目例（OECD, 2014, p. 39）.

2.2 技術強化型項目から得られるデータ

PBT・CBT のいずれにおいても，従来の能力測定において主に用いられてきたのはプロダクトデータ（product data）である．すなわち，受検者がテスト問題を解き，最終生成物として入力された解答（e.g., どの選択枝を選んだかや記述答案）が記録・採点され，それを能力測定のためのデータとしてきた．プロダクトデータを生成する解答形式は，客観的に正誤が決まる解答選択型（主に多枝選択式）と，基準にもとづく採点を要する解答生成型（穴埋め式や記述式）に大別される．TEIs においてもこの区別が当てはまるが，それぞれのタイプの中で，2.1 節で示したようにより多様な解答形式が実装できる．

一方，CBT の下では，解答の過程で生じる様々なデータ（プロセスデータ）を収集することができる．解答中に発生したイベントがタイムスタンプを伴って記録されたものはログファイル（log files）と呼ばれる（Provasnik, 2021）．プロセスデータ（process data）とは，主にログファイルから抽出・変換される情報から成るが，その中でも特に受検者のテスト項目への解答過程を反映する，測定したい構成概念に関連するデータを指す（Provasnik, 2021）．テスト中に CBT システム以外（e.g., 視線追跡，MRI，ビデオ監視など）から得られるデータもプロセスデータに含まれ得る．

Xiao et al. (2022)は，プロセスデータとして利用される情報は反応時間（response time）と反応行動（response behavior）に大別され，反応時間に関する測定モデリングの研究の蓄積に比して，反応行動の活用に関する研究は依然探索的な段階にあると述べている．反応行動データとして利用され得るものとしては，解答を入力・変更した順序（パタン）や回数，シミュレーション型タスクで設定した条件の値や実験順序・実験結果，補助ツールの使用や回数といったものが考えられる．例えば，図 1 のドラッグ&ドロップ形式であれば，どの数字（ソース）をどの空欄（ターゲット）にどのような順番でドロップしたかという反応列（response sequence）が得られる．図 2 の例では，鉄道網の選択 → 料金種別の選択 → 切符種別の選択 → 購入決定という一連の操作に関して，前に戻ってやり直すといった試行錯誤の行動も分析対象となり得る．こうした主としてマウス操作による反応の記録に加え，文字あるいは単語の入力を記録したキーボードの打鍵記録（keystroke log）も反応行動と考えることができる．

従来プロダクトデータのみにもとづいて算出していたテスト得点の推定精度や妥当性を高めることや，プロダクトデータだけでは把握できない，受検者の解答方略やより詳細な学習状態に関する情報を得るという目的において，プロセスデータの活用が注目・研究されている（e.g., 北條, 2023）．

個々の TEI から得られる多種多様な解答データ（ローデータ）は，次節で述べる測定モデルへの入力とするために，必要に応じて適当な「得点」に変換される．Luecht (2017)はローデータを

項目得点に変換する関数を採点評価器(scoring evaluator)と呼んだ．最も単純な採点評価器はパタンマッチングを行うものである．例えば，正誤の二値採点を行う場合には，解答と正答が一致していれば 1，そうでなければ 0 を返す関数として採点評価器が定義される．ここでの解答および正答は，多枝選択式であればそれぞれの選択枝番号，穴埋め式であれば記入された単語と正解となる単語といったものとなるだろう．後述する 3.3 節に，Betts et al. (2022)が研究対象とした，多枝多選択方式(複数選択可の多枝選択式)の解答に適用される採点方法の例を挙げているが，こうした場合には比較対象がベクトルあるいは集合で表されることもあり得る．

Betts et al. (2022)の例では多値採点も行っている．多値採点は，例えば解答と正答のパタンが一致する程度(個数など)に対して段階的に(一致度が高いほど高得点になるように)数値を割り当てるものである．記述式を中心とする構築型の解答に対しては多値採点が行われることが多いが，これは従来は人間の採点者が予め設定された採点基準を用いて行っていた．しかし，第 1 節で述べたように，近年では深層学習による自動採点が実用化され，これがそのまま採点評価器の役割を果たすようになってきている．

採点ルールの決定には，様々な条件を考慮する必要がある．まず，従来のテストと同様に，TEIs の設計や作問における指針となるのはそのテストで何を測定しようとしているかという意図である．したがって，どのような解答データを収集し(そのような仕組みを作り)，それらをどう採点するかは作問時にある程度想定されていることが望ましい(こうした点に関する知見を積み上げる目的で 3.3，3.4 節で紹介する探索的研究が行われていると言える)．また，測定モデリングを考慮するとき，基本的なモデル仮定である局所独立性(3.1，3.2 節参照)を担保するために複数の解答データをまとめて多値採点するなどの方策が取られることがある (Luecht, 2017)．

3. 技術強化型項目と測定モデル

3.1 項目反応理論モデル

従来の PBT(や第一世代の CBT)に用いられる標準的な測定理論は項目反応理論 (IRT; Lord, 1952)である．IRT における測定モデルは，テスト項目への反応(ここでは，採点評価器によって変換された項目得点を指すものとする)が生成されるメカニズムを表現する確率モデルであり，受検者 $i(i = 1, \ldots, N)$が持つ測定したい特性 θ_i(能力母数)と，その特性を測定する一連のテスト項目$(j = 1, 2, \ldots, J)$の特徴を表す母数 $\boldsymbol{\eta}_j$(項目母数)，そしてこれらの母数によって条件づけたときの受検者 i のテスト項目 j への反応 x_{ij} が生起する確率

$$P(X_{ij} = x_{ij}|\theta_i, \boldsymbol{\eta}_j) \tag{3.1}$$

を表現するものである．この確率を能力母数の関数と見なしたとき，特に項目反応関数(item response function)と呼ぶ．

実際に運用されているテストにおいて最もよく使用されているのは，基本モデルの 1 つとされる 3 母数ロジスティックモデル (Birnbaum, 1968)

$$P(X_{ij} = 1|\theta_i, \boldsymbol{\eta}_j) = c_j + \frac{1 - c_j}{1 + \exp(-a_j(\theta_i - b_j))} \tag{3.2}$$

や，上式で $c_j = 0$ と置いた 2 母数ロジスティックモデルである．ここで $X_{ij} \in \{0, 1\}$ は受検者 i が項目 j に正答すれば 1，誤答すれば 0 の値を取る二値変数である(従って，上式は項目 j への正答確率を表す)．また，$\boldsymbol{\eta}_j = \{a_j, b_j, c_j\}$ であり，それぞれ識別力，困難度，当て推量パラメタと呼ばれる．上記のモデルでは項目反応確率を左右する能力母数は単一の変数であることを仮定しており，これを一次元性(unidimensionality)の仮定と呼ぶ．

一揃いの J 個の項目に対して特定の反応パタン $\boldsymbol{x}_i = \{x_{i1}, \ldots, x_{iJ}\}$ が得られる確率は，能力母数と項目母数が所与のときに各項目に対する反応 X_{ij} は互いに独立であると仮定して，

$$P(\boldsymbol{X}_i = \boldsymbol{x}_i | \theta_i, \boldsymbol{\eta}) = \prod_{j=1}^{J} P(X_{ij} = x_{ij} | \theta_i, \boldsymbol{\eta}_j) \tag{3.3}$$

とされる．ただし，$\boldsymbol{\eta} = \{\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_J\}$ である．この，項目反応の条件付き独立性を局所独立性(local independence)の仮定と呼ぶ．

IRT の原点と言えるモデルとして，正規累積モデル（Lord, 1952)や上に挙げたロジスティックモデル（Birnbaum, 1968)，Rasch モデル（Rasch, 1980)が挙げられるだろう．これらの基本モデルは，いずれも正誤の二値反応を対象とし，能力母数に関する一次元性と項目反応の局所独立性を仮定するものである．受検者の特性と項目の特徴を分けて表現しておき，その上で，多数の項目を集めて予め共通の能力母数の尺度上にそれらのパラメタを推定・変換(等化)した項目プールを構築することで，難易度や測定精度を揃えた等質なテスト版を複数構成したり，受検者の能力レベルに応じて最適な項目を出題する(e.g., 適応型テストなど)といった状況でも，受検者の能力を同じ尺度上で比較可能な形で推定できるということが IRT の最大の利点である（加藤 他, 2014)．

こうした基本モデルに対しては，これまでに様々な方向で拡張がなされている（cf. van der Linden, 2016)．項目反応に関する拡張として，多値(3 つ以上の段階値やカテゴリ)の項目反応をモデル化した多値型 IRT モデル(polytomous IRT models)，連続型変数を扱うモデル(continuous IRT model; Samejima, 1973 など)がある．多値型 IRT モデルの中には，名義反応モデル（Bock., 1972)，段階反応モデル（Samejima, 1969)，部分得点モデル（Masters, 1982)，一般化部分得点モデル（Muraki, 1992)などがある．項目 j への解答が $X_{ij} \in \{0, 1, \ldots, K_j\}$ の $(K_j + 1)$ 段階で採点されるとすれば，一般化部分得点モデルの項目反応関数は，

$$P(X_{ij} = k | \theta_i, \boldsymbol{\eta}_j) = \frac{\exp(a_j \sum_{l=0}^{k} (\theta_i - b_{jl}))}{\sum_{m=0}^{K_j} \exp(a_j \sum_{l=0}^{m} (\theta_i - b_{jl}))} \tag{3.4}$$

と表される．a_j は(3.2)と同様に項目 j の識別力を表す母数である．b_{jk} は項目 j において $X_{ij} = k - 1$ の項目反応関数と $X_{ij} = k$ の項目反応関数の大小が入れ替わる θ 尺度上の点を表し，境界母数などと呼ばれる．ただし，(3.4)において便宜的に $\sum_{l=0}^{0} (\theta_i - b_{jl}) = 0$ とする．(3.4)で $a_j = 1$ と置けば部分得点モデルの項目反応関数が得られる．

能力母数を多次元に拡張したモデルとして，多次元 IRT モデル（multidimensional IRT models; Reckase, 2009)がある．多次元 IRT モデルは，正答確率のロジットが多次元の能力母数の線形結合で表される補償型(compensatory)と，正答確率が各能力母数のロジスティック関数の積で表される非補償型(non-conpensatory)の 2 種類に大別される．また，カテゴリカルな能力母数を扱うモデルとして，潜在クラスモデル（Lazarsfeld and Henry, 1968)がある．教育テストへの適用を念頭に置いたモデルとして，測定対象の概念やスキルの習得の有無を二値の潜在変数で表す状態習得モデル（state-mastery model; Macready and Dayton, 1977)や，その発展形として複数のスキルの習得パタンを推定する認知診断モデル（coginitive diagnostic models [CDMs]; Leighton and Gierl, 2007)が挙げられる．

その他にも，項目母数や能力母数に階層構造を仮定するモデル（hierarchical/multilevel IRT models; Fox, 2010)，項目特性曲線に単調増加性を仮定しない展開型 IRT モデル(unfolding IRT models; Andrich, 1988)など，項目反応や能力母数のタイプ以外の側面に関しても様々な拡張がなされている．

3.2 技術強化型項目の測定論的課題

Jiao et al. (2018)は，TEIs の特徴に起因する(IRT の適用を念頭に置いた)測定論的課題として，測定する特性(能力母数)の多次元性と，項目反応の局所依存性(local dependence)を挙げている．多次元性は，TEIs によって測定される能力・特性が多様な要素や側面を含み得ることに起因する．例えば，プロセスデータを考慮することによって，プロダクトデータだけでは捉えきれなかった側面が明らかになる可能性がある．ただし，これを元々意図していた測定の一部の要素として捉える(明示的に能力母数として扱う)か，別物(主たる測定に関する補助的な情報)と捉えるかは判断が分かれるところである．局所依存性は，測定論的には多次元性の問題と表裏一体の性質であるが，TEIs の文脈では主としてプロセスデータや，先に示した PISA の項目のように文脈を設定して一まとまりの素材に対して複数のタスクを要求するようなテストレット(大問)形式にまとめられた TEIs から生じるデータを扱う際に考慮すべき性質である．

また，現時点では TEIs から得られるデータ，特にプロセスデータからどのような情報が引き出せるのかについても十分に明らかになっているとは言えず，データを測定モデルの俎上に乗せる前の問題として，探索的な分析検討も行われている．以下，こうした観点から，TEIs によって生成されるデータを分析した研究や，測定モデルの適用を試みた研究を概観する．

3.3 探索的検討：解答形式

TEIs に関する探索的検討の研究の 1 つの分野として，TEIs ならではの解答形式を採用することによる，従来型のテストとの，あるいは TEIs 特有の解答形式(および採点方法)間の測定論的な性能比較(一種のモード効果の検討)がある．

解答形式の影響に焦点を当てた研究には，単純な項目・テスト統計量を比較するものと，IRT やその他のモデルを適用して項目母数や適合度を比較するものがある．前者の例として，Ponce et al. (2020)や Ponce et al. (2021)は，読解テストにおける「文の並べ替え」「図の整理」「cloze(穴埋め)テスト」といった解答形式の項目について，PBT 版と CBT 版(ドラッグ&ドロップ方式による解答)を比較し，テスト得点やその精度には差はないが，解答に要する時間はドラッグ&ドロップ方式の方が短いことを報告している．

IRT モデルを用いた比較検討の例をいくつか挙げると，Moon et al. (2020)は数学テストにおける多枝多選択方式(複数選択可で，当てはまる選択枝のボックスにチェックを入れる)とグリッド方式(各選択枝について「当てはまる」「当てはまらない」「わからない」などから 1 つの解答にチェックを入れる；チェックを入れるラジオボタンが選択枝を表側，解答を表頭としてグリッドの形に配置されたもの)の解答形式について，2 母数ロジスティックモデルおよび一般化部分得点モデルの母数推定値に生じた違いから，解答形式による項目反応への影響が無視できないことを報告している．また，Betts et al. (2022)は，「当てはまるものをすべて選びなさい」「当てはまるものを N 個選びなさい」という指示の下での多枝多選択方式の解答を多値採点する際の，個別の解答(選択)パタンの採点ルールの影響を検討している．具体的には，(a)二値法(選ばれた選択枝が正しい [選ばれるべき] 選択枝と完全一致するなら 1 点，そうでなければ 0 点)，(b)部分一致法(正しい選択枝のうち実際に選ばれたものの個数を得点とするが，正しくない選択枝が 1 つでも選ばれていたら 0 点)，(c)Ripkey 法(N 個選ぶという指示の場合に，N 個選んだうちの正しい選択枝の個数を得点とするが，N 個より多く選んだ場合には 0 点)，(d)加減法(選ばれた選択枝のうち，正しい選択枝の個数から正しくない選択枝の個数を引いたものを得点とする)，(e)多重真偽法(正しい選択枝を選んだ個数と，正しくない選択枝を選ばなかった個数の合計を得点とする)の 5 つの採点法によって得られる項目得点に部分得点モデルを適用し，困難度，モデル適合度，テスト情報量(測定精度)等の観点から項目およびテストの性能を比較した．その結果，多値採点，特に加減法と多重正誤法の二値採点に対する優位性

(e.g., 情報量の多さ)が示された.

Kim et al. (2022)は,第 1–12 学年対象の英語テストにおいて内容が等価な TEIs と多枝選択式項目を IRT モデルの特徴(識別力・困難度・情報量)を用いて比較した.その結果,TEIs の方が困難度が高くなる傾向があるが識別力には違いはないこと(その結果高学年・高能力層では TEIs の方が情報量が多くなる),TEIs の方がテスト時間が長くなり効率は下がることなどが示された.Ersan and Berry (2023)は,CBT で実施される数学の多枝選択式項目と,ホットスポット(画像の一部を指示する)・画像マッチ(図中の特定の場所に要素をドラッグ&ドロップする)・ドロップダウンリストによる文の穴埋め・テキスト記入による穴埋めの 4 種類の TEIs の性能を 2 あるいは 3 母数ロジスティックモデルを適用して比較し,テスト時間 1 分当たりの情報量は TEIs の方が高くなったことを報告している.解答時間や測定の効率性といった点に関しては,従来のテスト形式と比較して TEIs の解答時間が短い/長い,測定効率がよい/悪いといった一貫した結果は得られていない.インタラクティブで複雑なタスクを要求する TEIs では解答時間が長くなるのは当然であるし (e.g., Kim et al., 2022),PBT でのマーク式を CBT によるクリック式に改めれば解答時間はおそらく短縮されると考えられる (e.g., Ponce et al., 2021).

Eckerly et al. (2022)は,最初の項目への解答によって次に出題する項目を系統的に変える枝分かれ形式(branching format)の TEIs について,枝分かれを含む複数項目への解答パタンを多値採点して一般化部分得点モデルを適用したが,適合が悪かったことを報告している.多値採点に含まれた項目間で識別力が異なることが原因であるという考察がなされている.

こうした研究結果の蓄積は,従来型のテストで測定されている能力との等価性や,TEIs を用いて効率や精度に関してよりよい測定を追求するためのテストデザインを検討する際の参考になることが期待される.

3.4 探索的検討:プロセスデータ

プロセスデータが能力測定にどの程度・どのように寄与できるかということに関してはまだ十分に明らかになっているとは言えず (北條, 2023),従来のプロダクトデータによる測定を補足する,あるいはそれとは異なった側面を測定するものとして,プロセスデータから解答過程に関してどのような情報が得られるかを探索的に検討する研究が行われている.

Jiang et al. (2021)は,図 1 に示した NAEP 数学のドラッグ&ドロップ形式の項目のプロセスデータの分析を行っている.プロセスデータとして解答時のアクション(ソースをターゲットまでドラッグする/元の位置や別のターゲットに移す,解答をクリアするボタンを押す,などのマウス操作)とその順序,解答時間(総時間,最初に「止まった」タイミング,ドラッグ&ドロップ操作にかけた時間,最後に「止まった」タイミング)を収集し,これらのデータの分析から受検者の認知的・メタ認知的な思考プロセスや解答方略を推測し,解答データにもとづくテスト得点と比較することで,高得点者の解答方略がより効率的なものであることを示した.

Zhang and Andersson (2023)は,問題解決タスク(図 2 と同じく PISA 2012 で実施されたもの)のプロセスデータ(受検者の反応列と反応時間)を用いて,ネットワーク分析によって状態遷移(後述)を可視化・数値化して特徴を抽出した.用いられた特徴は(a)操作の多様性(可能な操作のうち実際に行った操作の割合),(b)辺の密度(既に行った操作の間を行き来する程度),(c)相互性(直前に行った操作を再度行う程度),(d)遷移性(ある操作に対して,その 1 つ先の操作を行ってからまたその直前の操作を行う程度),(e)外的-内的指標(一貫して正しい操作を行う程度),(f)平均時間(状態遷移にかかった平均時間),(g)時間の標準偏差(状態遷移にかかった時間のばらつき)の 7 つであり,問題解決の認知プロセス(問題の表象,計画・モニタリング,実行の 3 つ)を表現するものとしてネットワーク(特定の条件に当てはまる頂点や辺の数など)

および反応時間からそれぞれ定義された．これらの特徴に正規混合モデルを当てはめることで受検者のクラスタリングを行い，タスクへの取り組み方を反映する 6 つのグループ(低能力型，低努力型，適応型，行ったり来たり型，慎重型，試行錯誤型)を見出した．

キーストローク・ログに焦点を当てた研究として，Uto et al. (2020)がある．彼らは論述式の答案を入力する過程を記録し，そこから抽出した記入文字数，停止中，カーソル位置といった特徴量から時間区分ごとの状態(e.g., 思考中，入力中，推敲中など)を推測し，隠れマルコフモデルを用いて受検者ごとの状態遷移を分析した．状態遷移のパタンと，プロセスデータとは独立に採点した論述答案の採点結果と突き合わせることにより，答案の採点結果からだけではわからない「よい書き手」の筆記プロセスの特徴を明らかにする可能性が示唆された．

反応時間についてはそれ自体をテーマとして多数の先行研究が存在するが，プロセスデータ全般に対する期待と同様に，Molenaar (2015)は反応時間を考慮した測定モデリングの意図には，主に能力測定の精度改善と，解答過程や解答方略の推測の 2 つがあると述べている．前者の意図での測定モデリングについては，例えば van der Linden et al. (2010)がある．正誤採点されたプロダクトデータに対しては(3.2)の 3 母数ロジスティックモデルを当て，合わせて受検者 i の項目 j に対する反応時間 t_{ij} について，対数正規分布

$$p(t_{ij}|\tau_i,\alpha_j,\beta_j)=\frac{\alpha_j}{t_{ij}\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left[\alpha_j(\ln t_{ij}-(\beta_j-\tau_i))\right]^2\right\} \tag{3.5}$$

を仮定した．すなわち，受検者 i の項目 j に対する対数反応時間の平均が項目 j の強度(intensity)母数 β_j(正に大きいほど項目 j への平均的な反応時間は長くなる)と，受検者 i のスピード母数 τ_i(正に大きいほど受検者 i の平均的な反応時間は短くなる)の差で決まり，識別力母数 $\alpha_j>0$ が大きいほど実際の反応時間が平均値まわりに集中するという形のモデルとなっている． さらに θ と τ の間に二変量正規分布を仮定することで項目反応と反応時間に関わる能力母数どうしをリンクさせ，モデルの母数推定を改善しようとするものである．近年，解答過程に関心を持って行われた研究としては，Pohl et al. (2021)が挙げられる．彼女らは，PISA 2018 の数学的リテラシーアセスメントの項目反応と反応時間に加えて無解答フラグのデータを利用し，上記の能力母数とスピード母数を含む反応時間モデルに反応傾向母数(飛ばさずに解答を行う傾向)を追加したモデル (Ulitzsch et al., 2020)を適用した．そして，これら 3 種類の母数の推定値のプロファイルを国別に比較し，能力レベルが同程度でも，解答スピードや反応傾向(テストの受検態度や受検方略)に国ごとに大きな違いがあること，能力のみによらない質の違いを見ることの重要性を見出した．

視線追跡(eye tracking)に関しても様々な検討が行われている．Man and Harring (2019)は，テスト解答中の個人の関与度(engagement)を反映する指標として固視頻度(fixation counts)のモデリング(負の二項回帰モデル)を行った．Yaneva et al. (2022)は，視線追跡データから多枝選択式項目の解答過程の検討を行っている．多数の特徴量を用いて，機械学習によって正誤解答を予測(分類)する目の動きを検討し，問題文(幹)よりも選択枝を先に見るパタンが誤答に関連し(妥当性を脅かす望ましくない解答過程)，逆に幹 → 選択枝の順に見ていく(かつ，時間をかけて読む，自信を持って選択すると考えられる)パタンは正答に関連することが明らかになっている．Maddox et al. (2018)は，OECD 国際成人力調査 PIAAC(Programme for the International Assessment of Adult Competences)の解答中の注視とサッカード(眼の素早い動き)のデータを用いて，項目内容のどこに注目してどんな順番で読み取っていくのか，テスト解答中の個人の関与度，解答方略(テキストの読み方や検算の有無など)といった解答過程に関する詳細な情報を与えると主張した．また，Mayer et al. (2023)は視線追跡による問題解決過程の分析のレビューを行っている．実際のテスティング場面において装置を使って視線追跡を

行うことは，それが認知的・心理的な面で解答行動への負担になる可能性もあることから現実的であるとは言えないかもしれない．しかし，上記の研究からは解答過程に関する様々な示唆が得られており，視線追跡を用いた実験的研究がテストの妥当性を検証する，あるいは高めるために有用な情報(e.g., 受検者が，テスト開発者が意図した解答過程を実際に経て解答しているかどうか)をもたらすことが期待される．

その他，プロセスデータの利用に関する探索的研究に関しては，北條 (2023)が機械学習を用いた試みなどいくつかの例を挙げている．

3.5 既存の測定尺度との等価性・関係性の検討

CBT の過渡期にあって，既存のテストに TEIs を追加導入する際には，既存の測定尺度との等価性が問題となる (Luecht, 2017)．これは，既存のテストの運用や設計変更，妥当性(i.e., 本来そのテストで測定したい能力が測定できているか)にも関わる重要な論点である．

TEIs の導入そのものが，それまでになかった(それまでのテストで捉えられていなかった)次元を新たに反映するものである可能性もある．真正性を高めるという目的で TEIs が導入されることで，それまで測定できていなかった能力の側面が測定できるようになるのは，それが統計的に別の次元として見なされるかどうかは別として望ましいことである．一方で，TEIs のタスクの遂行に，本来測定したい能力とは関係のない機器の操作スキルなどが過度に求められるようであれば，これは測定の妥当性に対する脅威となる (Bryant, 2017; Luecht, 2017)．

また，TEIs でよく用いられるテストレット形式の項目群をどう扱うかという問題がある．この問題は PBT における課題としても存在したが，特に現実的な状況・文脈設定をする TEIs においてより顕在化しやすいものであると考えられている．テストレット形式では，1 つの状況・文脈・素材等に対して複数の項目がぶら下がる形になるため，設定された文脈に対する親和性や事前知識などが共通してそのテストレット中の項目への反応に影響する可能性がある．また，プロセスデータ(ある操作や解答自体が次の操作や解答に影響するような一連の反応行動)を従来のプロダクトデータの反応のように扱うとしたら，おそらく個人内でも(i.e., 能力母数で条件付けても)相互依存する可能性がある．これらは IRT における局所独立性からの逸脱につながる．

Kang et al. (2022)は，TEIs においてテストレット形式および多値採点が多く用いられることを念頭に，多値型データに対応したテストレットに適用可能な IRT モデルのパフォーマンスをシミュレーションデータによって比較した．比較対象とされたモデルは，テストレットを無視して各反応を一次元・局所独立として扱う一般化部分得点モデル(GPCM)，テストレット内の項目得点の合計点を算出し，各テストレットを 1 つの多値型項目と見なして分析するテストレット多値化モデル (TPIM; Wainer and Kiely, 1987)，能力母数に加えて受検者とテストレットの変量効果交互作用を組み込んだ変量効果テストレットモデル (RTM; Bradlow et al., 1999; 実質的には多次元 IRT モデルの特殊ケースと見なせる)，RTM の交互作用項を受検者に依存しないテストレット固有の固定効果とした固定効果テストレットモデル(FTM)の 4 つであり，GPCM，TPIM，FTM のそれぞれを真のモデルとしてサンプルサイズやテストレット効果の大きさ(i.e., テストレット内の局所依存性の強さ)を操作して項目反応データを生成し，これらの 4 つのモデルを当てはめて適合度等を比較することにより，現実場面における TEIs への適用可能性が検討された．また，Jiao et al. (2018)も，TPIM と同様に受検者とテストレットの変量効果交互作用を組み込んだ多次元非補償型テストレット IRT モデルを提案している．

Luecht (2017)は，TEIs によってもたらされ得る局所従属や新規の能力軸に起因する多次元性に関して，(a)既存の尺度をそのまま使ってよい(等価な尺度と見なしてよい)，(b)基本的に多次元であるが既存の尺度と TEIs による新規尺度の混合尺度(線形結合)とする，あるいは

(c)既存尺度と TEIs による尺度を別物として扱う(多次元)かどうかを，データにもとづく統計的な検証(次元性やモデル適合度のチェック)を行い，中長期的な運用を見据えたときの尺度の一般性・継続性，そして何よりも意図された測定が実現できているか(テスト得点にもとづいて測定したい能力に関する適切な推論ができるか)という妥当性の観点から慎重に判断すべきであると述べている．

3.6 解答過程のモデリング

近年，問題解決(problem-solving)のタスクを含む TEIs におけるプロセスデータを用いた測定モデリングの研究が盛んに行われている．問題解決タスクでは，初期状態(initial state)と目標状態(goal state；目指すべき最終的な状態の指示)が明確に与えられており，受検者は目標状態を達成するために種々のアクションを実行する．図 2 の例で言えば，初期状態 S_1 において取るアクションは乗車する鉄道網の選択であり，地下鉄か広域かを選ぶことによって次の状態 S_2 に遷移する．次は料金種別の選択であり，通常料金か割引料金かによって鉄道網と料金種別の特定の組み合わせとして状態 S_3 が決まるといったように，状態 S_t でアクションを起こすと次の状態 S_{t+1} に移る．これを繰り返して，最終状態 S_T(この例では「購入」を押した状態；ただし，最終状態が目標状態と一致するとは限らず，その場合にはタスクは失敗したということになる)に至るまでのアクションの列 $A_1, \ldots, A_T$ とその結果である状態の列 $S_1, \ldots, S_T$ が得られる(これらがプロセスデータを構成する)．初期状態と最終状態の間に取りうる様々な状態を中間状態(intermediate states)と呼ぶ．明確に定義された(well-defined な)問題解決タスクでは，あり得る状態の数や遷移を経て最終状態に至るパスの数は，無用なループなどを除けば有限であり，最小のアクションで目標状態に至る最良のパスが決まっている．例えば，図 2 の例ではあり得る中間状態の数は 22 個である (Han et al., 2022)．

こうしたプロセスデータに対して，Shu et al. (2017)はマルコフ IRT モデル(Markov IRT model)を適用している．同様のアイデアは Han et al. (2022)の逐次反応モデル(sequential response model [SRM])にも採用されている．Han et al. (2022)は，能力母数 θ_i を持つ受検者 i が，状態 $S_{i,t} = s_k$ のときに(何らかのアクションを起こして)状態 $S_{i,t+1} = s_{k'}$ に遷移する確率を

$$P(S_{i,t+1} = s_{k'} | S_{i,t} = s_k, \theta_i, \boldsymbol{\lambda}) = \frac{\exp(\lambda_{kk'} + I^+_{kk'}\theta_i)}{\sum_{h \in \boldsymbol{M}_j} \exp(\lambda_{kh} + I^+_{kh}\theta_i)} \tag{3.6}$$

のような多項ロジスティック関数とするモデルを立てた．ここで，$\lambda_{kk'}$ は状態 s_k から $s_{k'}$ への「遷移しやすさ」を表す母数，$\boldsymbol{M}_k$ は状態 s_k から遷移可能な状態の集合を表す．$I^+_{kk'}$ は，状態 s_k から $s_{k'}$ への遷移がより目標状態に近づく(正しい)ものであれば 1，そうでなければ -1 を取る，予め定められた指示関数である．従って，能力母数 θ_i が正に大きければ「正しい」状態に遷移する確率が高くなる．状態遷移の確率は直前の状態にしか依存しない，すなわち一次のマルコフ性が仮定されている．この仮定により，局所独立とは行かないまでも各受検者の状態列 $\boldsymbol{S}_i$ の同時確率分布(あるいは尤度関数)を直前の状態にのみ依拠する条件付き確率の積として効率的に表現することができる：

$$L(\boldsymbol{S}_i | \theta_i, \boldsymbol{\lambda}) = \prod_{t=1}^{T_i - 1} P(S_{i,t+1} | S_{i,t}, \theta_i, \boldsymbol{\lambda}) \tag{3.7}$$

なお，このモデルではどんなアクションを取るかではなく，アクションの内容に関わらずどの状態からどの状態に遷移したのかに着目している点が特徴である．

Han et al. (2022)は SRM を図 2 のタスクのデータに適用し，良好なモデルフィットが得られたとしている．また，SRM で推定した能力母数と，創造的思考力テスト全体から正誤デー

タを元に IRT で推定した能力母数を比較したところ，0.581 の相関が得られたとしている．前者は図 2 の単独のタスクに適用したモデルから得られた得点である．仮に両者が同一の能力を測定していると仮定し，やや乱暴ではあるが古典的テスト理論に基づく項目識別力（項目得点と，全項目得点を合計したテスト得点の間の相関係数）を模した値であると見なせば，それなりに高い値が出ていると言える（IRT 得点の分散の約 34% [$\approx 0.581^2$] を単独タスクの SRM 得点で説明できる）．ただし，同一の能力を測定しているという保証はなく，先述した等価性のさらなる検討が必要であると思われる．

Han and Wilson (2022) も協働的問題解決力を測定する類似の構造のタスクに対して，同様の考え方にもとづいてモデリングを行っている．遷移確率には Han et al. (2022) と同様に多項ロジスティック関数を採用しているが，彼らが扱ったタスクの性質上，同一の状態遷移が何回でも起こり得ることを考慮し，同じ遷移の繰り返し回数を反応データとして扱ったモデルとなっている（繰り返し数が多いほど，与えられる得点が低くなるように部分得点モデルの反応確率のロジットにおける θ_i の係数を設定している）点と，この遷移モデルを反応傾向が異なると考えられる 2 つの潜在クラスにネストした（i.e., 遷移確率が受検者が属するクラスによって異なる）混合部分得点モデル（mixture partial credit model）としている点が異なる．分析の結果得られた 2 つの潜在クラスは，ほぼタスク成功群と失敗群に対応することが確かめられた．

状態遷移のマルコフ性を仮定して，プロセスデータを利用して能力推定を行うモデルとしては，他にも Lamar (2018) のマルコフ決定過程測定モデル（Markov decision process measurement model）がある．遷移カーネルには名義反応モデルを用いており，状態遷移ではなくアクションの生起確率をモデリングしている点が異なる．また，Chen (2020) は，Han et al. (2022) が扱ったのと同じ図 2 のタスクのアクション列を点過程データと見なして，連続時間動的選択測定モデル（continuous-time dynamic choice measurement model）を提案している．

これらの一連の研究は，状態遷移の確率に能力母数を組み込むことで，より適切な状態遷移を繰り返す受検者の方が高い問題解決能力を示すというメカニズムを表現するものであり，今後の問題解決タスクのプロセスデータのモデル化に対して一定の方向性を示していると言える．なお，マルコフ性の仮定とは異なった方法でプロセスデータの個人内相互依存性を表現しようとする試みも行われている．Liu et al. (2018) はプロセスレベルと受検者レベルに分けて能力母数を設定し，プロセスレベルにおいては異なる問題解決ステップを選択する受検者群を区別するために，群ごとに異なるアクション実行確率を表現する混合 IRT モデルを提案した（マルチレベル混合 IRT モデル；multilevel mixture IRT model）．

上記の一連の手法とは（目的も含めて）異なるアプローチとして，Xiao et al. (2022) は，既存尺度上での能力推定の精度向上の目的で，プロセスデータそのものを用いず，縮約した情報を能力推定に利用する手法を提案している．Xiao et al. (2022) は，PIAAC の問題解決タスクのプロセスデータ（反応行動）から多次元尺度構成法（MDS）によって特徴抽出（feature extraction）を行い，さらにランダムフォレストを適用して特徴を選択した．MDS の適用に当たっては，受検者間の反応行動列の非類似度にもとづいて次元抽出を行った．また，特徴抽出は 7 つの課題解決タスクのプロセスデータを用いて同時並行で行われ，各タスク 3 つの特徴，すなわち合計 21 の特徴量が選ばれた．こうして選ばれた各受検者の特徴量 $f_{i1}, \ldots, f_{i21}$ を用いて，能力母数 θ_i の事前分布を

$$\theta_i \sim N(b_0 + b_1 f_{i1} + \cdots + b_{21} f_{i21}, \sigma^2) \tag{3.8}$$

と置き，これを各タスクへの解答データ（一般化部分得点モデルでモデル化）と合わせて能力母数（および回帰係数）をベイズ推定することで，θ_i の推定精度（事後分散）が大幅に改善した．このアイデアは，PISA においてアセスメントの得点を算出する際に用いられる潜在回帰（latent

regression）と同様のものである（https://www.oecd.org/pisa/data/pisa2018technicalreport/；ただし，PISA ではプロセスデータの特徴量ではなく，アセスメントと同時に行われるアンケート調査の情報を縮約した変数を用いている）．解答データに対する IRT モデルの部分では，解答データのみから推定した項目母数が用いられており，既存の能力尺度の意味付けを変えることなくプロセスデータの情報を利用して能力推定の精度を改善することが可能となっている．なお，認知診断モデルの文脈でも，プロセスデータ（マウスクリック&ドラッグ，反応時間）を利用することによって習得パタンの診断精度や項目パラメタの推定精度が向上したという報告がある（Liang et al., 2023）．

4. 現状と今後の課題

本稿では，新しい時代のアセスメントの特徴のひとつとなるであろう TEIs の利用について，決して網羅的とは言えないながら，能力測定・測定論の立場から実践・モデリング研究のレビューを行った．TEIs による CBT のアップデートは，その利用価値の探索とモデリングの試行錯誤が並行して行われている段階にあり，近年は以前にも増して関心が高まっているテーマであると言える．TEIs の利用価値に関する探索的検討は，様々な測定領域，様々な解答形式，（プロセス・プロダクトに関わらず）TEIs が生成する様々なデータの種類に関して，今後も継続的に行われることが望ましい．測定モデルに関しては，基本的（古典的）な二値型・多値型のロジスティックモデルや多次元 IRT モデルを基本として，それらが様々な形で応用されていることがわかった．

TEIs が生成し得るデータの中で，プロセスデータには特に注目が集まっていることがうかがわれる．マルコフ IRT モデルは，状態や遷移パスの数が有限（しかもかなり少ない数）である well-defined な問題解決タスクへの適用にとどまっているが，問題解決のみならず他の内容領域のタスクへの適用可能性の検討が望まれる．

その一方で，さらに掘り下げて検討していくべき課題もいくつかある．1 つ目はモデルの効用（利用可能性）についてである．まず，問題解決タスクにおけるプロセスデータのモデリングは解答過程のモデリングであって，それがどのような能力を測定しているのか（能力母数の解釈）についてはより理解を深めていく必要があるだろう．また，次世代の「学習のためのアセスメント」の実現のためには，特定のタスクの解答過程が明らかになること（類型化できること）だけでは不十分である．個々の受検者＝学習者にとって，学習の改善に資する有用なフィードバック（の元となる情報）を提供できるかという視点が必要である．

さらに挙げるとすれば，現状ではモデル設定上の制限により，モデルの適用を優先すると作成できるタスクの幅が狭くなってしまう可能性がある．これは，妥当性を制限することにもつながる．また，現状のほとんどのマルコフ IRT モデルは単一のタスクにおける解答過程のモデル化にとどまっている．通常のアセスメントでは，内容領域をなるべく網羅するために（かつ測定精度を上げるために）なるべく多くのタスクからなるべく多くの反応データを集められるようにテスト版が構成される．現状のマルコフ IRT モデルからも能力母数は推定できるが，単一のタスクから推定されたその数値が，想定している能力の推定値として十分な一般性を持つのかは疑問である．IRT による既存の能力推定とは別系統の測定として行うことも可能であるが，これはテストの利用目的によるであろう．

2 つ目の課題は TEIs を含むテストの運用についてである．測定モデリングやそれに付随する推定法の進展はめざましいものがある一方で，これらのモデルが実際のテスト運用に適うかどうかの検討は追いついていない．測定モデルは，受検者の状態を表す変数と項目の特徴を表す変数が与えられたときに解答が生成されるメカニズムを模すこと（i.e., 現象の記述）を基本と

しながらも，現実のテスト運用においてはテストの得点(能力母数の推定値)を算出するという機能を安定的・継続的に果たせることが重要である．そのためには，項目個別にモデルを立てるのではなく，多数の項目の挙動をなるべく統一的かつ単純に記述できるモデルを用いることが望ましい(一次元・局所独立の二値型 IRT モデルを安定的に運用していくだけでも並大抵のことではなく，既存尺度との関係性を重視する Luecht (2017)の視点には大いに共感できる)．マルコフ IRT モデルに限らず本稿で取り上げた他の多くのモデルは，かなり特定のタスクやデータに依存する側面が強く，上に述べた意味での実用に耐えうるのかは未知である．同じ文脈で，尺度の等化(equating)に関する研究が極めて少ない点も実用性を限定する大きな要素である．プロセスデータを加味して既存の能力測定を改善するという目的では，プロセスデータから抽出した特徴量を，既存尺度上の能力推定の際の事前分布の設定に利用するという Xiao et al. (2022)が提案したような手法は実用性が高いように思える．

「学習のためのアセスメント」の実現を念頭に置いたとき，TEIs がそのポテンシャルを十分に発揮するレベルで実用化されるまでには，解決すべき課題が数多く残されている．

参 考 文 献

Andrich, D. (1988). The application of an unfolding model of the PIRT type to the measurement of attitude, *Applied Psychological Measurement*, **12**, 33–51, https://doi.org/10.1177/014662168801200105.

Bennett, R. E. (2015). The changing nature of educational assessment, *Review of Research in Education*, **39**, 370–407, http://www.jstor.org/stable/44668662.

Betts, J., Muntean, W., Kim, D. and Kao, S.-C. (2022). Evaluating different scoring methods for multiple response items providing partial credit, *Educational and Psychological Measurement*, **82**(1), 151–176, https://doi.org/10.1177/0013164421994636.

Birnbaum, A. (1968). Some latent trait models, *Statistical Theories of Mental Test Scores* (eds. F. M. Lord and M. R. Novick), 397–424, Addison Wesley, Reading.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories, *Psychometrika*, **37**, 29–51, https://doi.org/10.1007/BF02291411.

Bradlow, E. T., Wainer, H. and Wang, X. (1999). A Bayesian random effects model for testlets, *Psychometrika*, **64**, 153–168, https://doi.org/10.1007/BF02294533.

Bryant, W. (2017). Developing a strategy for using technology-enhanced items in large-scale standardized tests, *Practical Assessment, Research & Evaluation*, **22**(1), 1–10, https://doi.org/10.7275/70yb-dj34.

分寺杏介 (2023). コンピュータを用いたアセスメントに関する研究トピックの整理と最新の動向, 日本テスト学会誌, **19**(1), 191–225, https://doi.org/10.24690/jart.19.1_191.

Care, E., Griffin, P. and Wilson, M. (eds.) (2018). *Assessment and Teaching of 21st Century Skills*: *Research and Applications*, Springer, Dordecht, https://doi.org/10.1007/978-3-319-65368-6.

Chen, Y. (2020). A continuous-time dynamic choice measurement model for problem-solving process data, *Psychometrika*, **85**, 1052–1075, https://doi.org/10.1007/s11336-020-09734-1.

Eckerly, C., Jia, Y. and Jewsbury, P. (2022). Technology-enhanced items and model–data misfit, Technical Report, RR-22-11, Educational Testing Service, Princeton, https://doi.org/10.1002/ets2.12353.

Ersan, O. and Berry, Y. (2023). Measurement efficiency for technology-enhanced and multiple-choice items in a K-12 mathematics accountability assessment, *Educational Measurement*: *Issues and Practice*, **42**(4), 19–32, https://doi.org/10.1111/emip.12580.

Fox, J.-P. (2010). Multilevel item response theory models, *Bayesian Item Response Modeling*: *Theory and Applications* (ed. J.-P. Fox), 141–191, Springer, New York, https://doi.org/10.1007/978-1-4419-0742-4_6.

Han, Y. and Wilson, M. (2022). Analyzing student response processes to evaluate success on a technology-

based problem-solving task, *Applied Measurement in Education*, **35**(1), 33–45, https://doi.org/10.1080/08957347.2022.2034821.

Han, Y., Liu, H. and Ji, F. (2022). A sequential response model for analyzing process data on technology-based problem-solving tasks, *Multivariate Behavioral Research*, **57**(6), 960–977, https://doi.org/10.1080/00273171.2021.1932403.

北條大樹 (2023). CBT 領域におけるプロセスデータ利活用研究の動向, 日本テスト学会誌, **19**(1), 177–190, https://doi.org/10.24690/jart.19.1_177.

Jiang, Y., Gong, T., Saldivia, L. E., Cayton-Hodges, G. and Agard, C. (2021). Using process data to understand problem-solving strategies and processes for drag-and-drop items in a large-scale mathematics assessment, *Large-scale Assessments in Education*, **9**(2), 1–31, https://doi.org/10.1186/s40536-021-00095-4.

Jiao, H. and Lissitz, R. W. (eds.) (2020). *Application of Artificial Intelligence to Assessment*, Information Age Publishing, Charlotte, https://www.infoagepub.com/products/Application-of-Artificial-Intelligence-to-Assessment.

Jiao, H., Lissitz, R. W. and Zhan, P. (2018). A noncompensatory testlet model for calibrating innovative items embedded in multiple contexts, *Technology Enhanced Innovative Assessment: Development, Modeling, and Scoring from an Interdisciplinary Perspective* (eds. H. Jiao and R. W. Lissitz), 117–137, Information Age Publishing, Charlotte, https://content.infoagepub.com/files/fm/p593e6be924373/9781681239316.pdf.

Kang, H.-A., Han, S., Kim, D. and Kao, S.-C. (2022). Polytomous testlet response models for technology-enhanced innovative items: Implications on model fit and trait inference, *Educational and Psychological Measurement*, **82**(4), 811–838, http://dx.doi.org/10.1177/00131644211032261.

Kato, K. (2016). Measurement issues in large-scale educational assessment, *Annual Review of Educational Psychology in Japan*, **55**, 146–164, https://doi.org/10.5926/arepj.55.148.

加藤健太郎, 山田剛史, 川端一光 (2014).『R による項目反応理論』, オーム社, 東京.

Kim, A. A., Tywoniw, R. L. and Chapman, M. (2022). Technology-enhanced items in grades 1–12 English language proficiency assessments, *Language Assessment Quarterly*, **19**(4), 343–367, https://doi.org/10.1080/15434303.2022.2039659.

Lamar, M. M. (2018). Markov decision process measurement model, *Psychometrika*, **83**, 67–88, https://doi.org/10.1007/s11336-017-9570-0.

Lazarsfeld, P. F. and Henry, N. W. (1968). *Latent Structure Analysis*, Houghton-Miffin, Boston.

Leighton, J. and Gierl, M. (eds.) (2007). *Cognitive Diagnostic Assessment for Education: Theory and Applications*, Cambridge University Press, New York, https://doi.org/10.1017/CBO9780511611186.

Liang, K., Tu, D. and Cai, Y. (2023). Using process data to improve classification accuracy of cognitive diagnosis model, *Multivariate Behavioral Research*, **58**(5), 969–987, https://doi.org/10.1080/00273171.2022.2157788.

Liu, H., Liu, Y. and Li, M. (2018). Analysis of process data of PISA 2012 computer-based problem solving: Application of the modified multilevel mixture IRT model, *Frontiers in Psychology*, **9**, 1–12, https://doi.org/10.3389/fpsyg.2018.01372.

Lord, F. M. (1952). A theory of test scores, *Psychometric Monograph*, No. 7, Psychometric Corporation, Richmond, http://www.psychometrika.org/journal/online/MN07.pdf.

Luecht, R. M. (2017). Calibrating technology-enhanced items, *Handbook of Item Response Theory, Volume 3: Applications* (ed. W. J. van der Linden), 87–103, Chapman and Hall/CRC, New York, https://doi.org/10.1201/9781315119144.

Macready, G. B. and Dayton, C. M. (1977). The use of probabilistic models in the assessment of mastery, *Journal of Educational Statistics*, **2**, 99–120, https://doi.org/10.2307/1164802.

Maddox, B., Bayliss, A. P., Fleming, P., Engelhardt, P. E., Edwards, S. G. and Borgonovi, F. (2018). Observing response processes with eye tracking in international large-scale assessments: Evidence from the OECD PIAAC assessment, *European Journal of Psychology of Education*, **33**(3), 543–558,

https://doi.org/10.1007/s10212-018-0380-2.

Man, K. and Harring, J. R. (2019). Negative binomial models for visual fixation counts on test items, *Educational and Psychological Measurement*, **79**(4), 617–635, https://doi.org/10.1177/0013164418824148.

Masters, G. N. (1982). A Rasch model for partial credit scoring, *Psychometrika*, **47**, 149–174, https://doi.org/10.1007/BF02296272.

Mayer, C. W., Rausch, A. and Seifried, J. (2023). Analysing domain-specific problem-solving processes within authentic computer-based learning and training environments by using eye-tracking: A scoping review, *Empirical Research in Vocational Education and Training*, **15**(1), 1–27, https://doi.org/10.1186/s40461-023-00140-2.

Molenaar, D. (2015). The value of response times in item response modeling, *Measurement*: *Interdisciplinary Research and Perspectives*, **13**, 177–181, http://dx.doi.org/10.1080/15366367.2015.1105073.

Moon, J. A., Sinharay, S., Keehner, M. and Katz, I. R. (2020). Investigating technology-enhanced item formats using cognitive and item response theory approaches, *International Journal of Testing*, **20**(2), 122–145, https://doi.org/10.1080/15305058.2019.1648270.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm, *Applied Psychological Measurement*, **16**(1), 59–71, https://doi.org/10.1002/j.2333-8504.1992.tb01436.x.

OECD (2014). *PISA 2012 Results*: *Creative Problem Solving*: *Students' Skills in Tackling Real-Life Problems* (*Volume V*), OECD Publishing, Paris, https://dx.doi.org/10.1787/9789264208070-en.

OECD (2023). *PISA 2022 Assessment and Analytical Framework*, OECD Publishing, Paris, https://doi.org/10.1787/dfe0bf9c-en.

Pohl, S., Ulitzsch, E. and von Davier, M. (2021). Reframing rankings in educational assessments, *Measurement*, **372**(6540), 338–340, https://doi.org/10.1126/science.abd3300.

Ponce, H. R., Mayer, R. E., Sitthiworachart, J. and López, M. J. (2020). Effects on response time and accuracy of technology-enhanced cloze tests: An eye-tracking study, *Educational Technology Research and Development*, **68**(5), 2033–2053, https://doi.org/10.1007/s11423-020-09740-1.

Ponce, H. R., Mayer, R. E. and Loyola, M. S. (2021). Effects on test performance and efficiency of technology-enhanced items: An analysis of drag-and-drop response interactions, *Journal of Educational Computing Research*, **59**(4), 713–739, https://doi.org/10.1177/0735633120969666.

Provasnik, S. (2021). Process data, the new frontier for assessment development: Rich new soil or a quixotic quest?, *Large-Scale Assessment in Education*, **9**, 1–17, https://doi.org/10.1186/s40536-020-00092-z.

Rasch, G. (1980). *Probabilistic Models for Some Intelligence and Attainment Tests*, expanded edition, The University of Chicago Press, Chicago.

Reckase, M. D. (2009). *Multidimensional Item Response Theory*, Springer, New York, https://doi.org/10.1007/978-0-387-89976-3.

Russell, M. and Moncaleano, S. (2019). Examining the use and construct fidelity of technology-enhanced items employed by K-12 testing programs, *Educational Assessment*, **24**(4), 286–304, https://doi.org/10.1080/10627197.2019.1670055.

Rychen, D. S. and Salganik, L. H. (eds.) (2003). *Key Competencies for a Successful Life and a Well-Functioning Society*, Hogrefe, Cambridge.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores, *Psychometrika*, **17**, 1–100, https://doi.org/10.1007/BF03372160.

Samejima, F. (1973). Homogeneous case of the continuous response model, *Psychometrika*, **38**, 203–219, https://doi.org/10.1007/BF02291114.

Shu, Z., Bergner, Y., Zhu, M., Hao, J. and von Davier, A. A. (2017). An item response theory analysis of problem-solving processes in scenario-based tasks, *Psychological Test and Assessment Modeling*, **59**, 109–131, https://www.psychologie-aktuell.com/fileadmin/download/ptam/1-2017_20170323/07_Shu.pdf.

Ulitzsch, E., von Davier, M. and Pohl, S. (2020). Using response times for joint modeling of response and omission behavior, *Multivariate Behavioral Research*, **55**(3), 425–453, https://doi.org/10.1080/00273171.2019.1643699.

U.S. Department of Education (2017). Reimagining the role of technology in education: 2017 national education technology plan update, Technical Report, U.S. Department of Education, Office of Educational Technology, Washington, D.C., https://tech.ed.gov/files/2017/01/NETP17.pdf.

Uto, M., Miyazawa, Y., Kato, Y., Nakajima, K. and Kuwata, H. (2020). Time- and learner-dependent hidden Markov model for writing process analysis using keystroke log data, *International Journal of Artificial Intelligence in Education*, **30**, 271–298, https://doi.org/10.1007/s40593-019-00189-9.

van der Linden, W. J. (ed.) (2016). *Handbook of Item Response Theory, Volume 1*: *Models*, Chapman and Hall/CRC, New York, https://doi.org/10.1201/9781315374512.

van der Linden, W. J., Entink, R. H. K. and Fox, J.-P. (2010). IRT parameter estimation with response times as collateral information, *Applied Psychological Measurement*, **34**(5), 327–347, https://doi.org/10.1177/0146621609349800.

Wainer, H. and Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets, *Journal of Educational Measurement*, **24**, 185–201, http://www.jstor.org/stable/1434630.

Wools, S., Molenaar, M. and Hopster-den Otter, D. (2019). The validity of technology enhanced assessments—Threats and opportunities, *Theoretical and Practical Advances in Computer-based Educational Measurement* (eds. B. P. Veldkamp and C. Sluijter), 3–19, Springer International Publishing, Cham, https://doi.org/10.1007/978-3-030-18480-3_1.

Xiao, Y., Veldkamp, B. and Liu, H. (2022). Combining process information and item response modeling to estimate problem-solving ability, *Educational Measurement*: *Issues and Practice*, **41**(2), 36–54, https://doi.org/10.1111/emip.12474.

Yaneva, V., Clauser, B. E., Morales, A. and Paniagua, M. (2022). Assessing the validity of test scores using response process data from an eye-tracking study: A new approach, *Advances in Health Sciences Education*, **27**(5), 1401–1422, https://doi.org/10.1007/s10459-022-10107-9.

Zhang, M. and Andersson, B. (2023). Identifying problem-solving solution patterns using network analysis of operation sequences and response times, *Educational Assessment*, **28**(3), 172–189, https://doi.org/10.1080/10627197.2023.2222585.

Review of Measurement Models for Technology-enhanced Testing

Kentaro Kato

Benesse Educational Research & Development Institute

As computer-based testing (CBT) has become a major mode in administrating educational assessments, its potential advantages over traditional paper-based testing (PBT) draw attention. Among them are the technology-enhanced items and their use for the purpose of improved measurement of learning status. Given this situation, this paper reviews recent attempts and practices in the context of exploratory analysis and psychometric modeling of various types of data generated from TEIs. The review identified the following research topics that are actively pursued: (a) effects of innovative response formats on the psychometric properties, (b) exploratory analysis of the utility of process data, (c) equivalence between the traditional measurement scale and the new scale involving TEIs, and (d) psychometric modeling of process data for revealing response processes. In terms of the last point, it was pointed out that further considerations will be necessary to produce measurement results more useful for learning and for sustainable and stable operation of tests that involve TEIs.

Key words: Computer-based testing, item response theory, technology-enhanced items, process data.

統計数理(2024)
第 72 巻 第 1 号 23–41

特集「心理統計学の新展開」
[研究詳解]

項目反応理論に基づく教育のための自然言語処理のモデル

江原 遥†

(受付 2023 年 7 月 4 日；改訂 2024 年 2 月 7 日；採択 2 月 19 日)

要　　旨

教育応用において，学習者の能力を測定したり項目の困難度など項目の特性を計測することは，学習支援システム等に幅広い応用がある教育応用の基礎タスクである．単に学習者が所与の項目に正答するかを予測するモデルではなく，学習者の能力や項目の特性を人間の教員が解釈できれば人間の教員が教育するときにも活用できる．統計分野においては，古くから項目反応理論等を用いて試験の学習者の回答パターンから解釈可能なパラメタを推定するアプローチが取られてきた．一方，項目の大部分は自然言語で記述されている．自然言語を解析する自然言語処理分野では，項目のテキストから困難度等の項目特性を抽出する研究に関心が持たれてきた．特に，単語頻度などの技術的に抽出が容易な特徴量から，項目の困難度の多くを説明可能な値を抽出できる語学学習支援などへの応用ではテキストからの困難度推定などの研究が盛んであった．そこで本稿では，テキストからの困難度の推定が項目反応理論とどのように関わりを持つのかについて，外国語の語彙学習支援や読解支援・可読性判定を中心に，様々な分野の研究を引用しながら説明する．そして近年，テキストの意味を考慮した解析で高精度を達成している自己教師あり学習や Transformer 等の手法を取り上げて詳説する．

キーワード：項目反応理論，自然言語処理，語学学習支援．

1. はじめに

学習者の能力に合わせた教育を行うためには，学習者に適合する教材を提示することが重要と考えられている．この「適合する」は，多くの場合，学習者の能力に合った「困難度(難しさ，難易度，難度)」を持つ教材を提示することに相当する．学習者の能力や教材の困難度を何らかの形で数値化して，適切な能力を持つ学習者に適切な教材を割り当てたい．

学習者の能力や教材の困難度をどのように計測したら良いだろうか？教科ごとに独自の尺度を作成することは比較が難しい．また，試験を通じて学習者の能力を計測するにしても，試験の配点を作問者が自由に決めて良いとすれば，必ずしも適切に能力を計測できているとは限らない．難しい項目(テストの問題や設問)の配点を小さくしすぎているなど，試験結果のデータに沿っていない状況が考えられる．

項目反応理論(item response theory，IRT)(Baker and Kim, 2004)は，試験結果のデータから，学習者の能力値や項目の困難度の両方を推定する手法の一つである．これらのパラメタは項目の自然文による内容等は一切参照せず，学習者集合の各項目に対するパターンだけから計

† 東京学芸大学 教育学部：〒 184–8501 東京都小金井市貫井北町 4–1–1

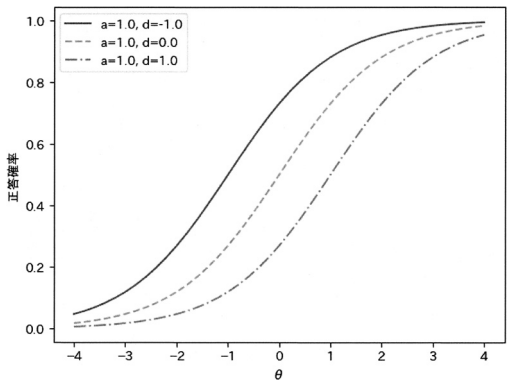


図 1. 困難度パラメタ d の異なる 3 項目に対する項目特性曲線.

測される．適切な項目集合を設定し，適切な等化(尺度調整)を行えば，異なる学習者集合間の能力を比較できることも知られている．

ここでは，IRT のモデルについて簡単に説明する．学習者の数を J，項目(項目, item)の数を I とする．簡単のため，学習者の添字(index)と学習者，項目の添字と項目を同一視する．例えば，i 番目の項目を単に i と書くことにする．y_{ij} は，学習者 j が項目 i に正答するとき 1，誤答であるとき 0 であるとする．試験結果データ $\{y_{ij}|i \in \{1,\ldots,I\}, j \in \{1,\ldots,J\}\}$ が与えられたとき，2 パラメタモデル(2PLM)では学習者 j が項目 i に正答する確率を次の式でモデル化する．

$$P(y_{ij} = 1|i, j) = \sigma(a_i(\theta_j - d_i)) \tag{1.1}$$

ここで，σ は $\sigma(x) = \frac{1}{1+\exp(-x)}$ で定義されるロジスティックシグモイド関数である．

σ は $(0, 1)$ を値域とする単調増加関数であり，$\sigma(0) = 0.5$ である．実数を $(0, 1)$ の範囲に射影し確率として扱うために用いられている．(1.1)において θ_j は能力パラメタ(ability parameter)と呼ばれ，学習者の能力を表すパラメタである．d_i は困難度パラメタ(difficulty parameter)と呼ばれ，項目の困難度を表すパラメタである．(1.1)より，θ_j が d_i を上回る時，学習者が正答する確率が誤答確率より高くなる．

能力値パラメタ θ に対して，(1.1)で定義される正答確率をプロットした図を「項目特性曲線」という．項目特性曲線を用いると，項目パラメタが正答確率にどの様な影響を及ぼすかを視覚的に理解しやすい．図 1 に，困難度パラメタを $d = -1.0, d = 0.0, d = 1.0$ と変えた 3 つの項目特性曲線を図示した．θ の値が同じでも，困難度の値が大きくなるほど曲線が右に平行移動することにより，学習者の項目に対する正答確率が低くなることがわかる．これは直感的には，困難度パラメタの値が大きい項目には学習者は正答しにくいということを表す．

a_i が大きいと $\theta_j - d_i$ の絶対値が小さくとも，$\theta_j - d_i$ の正負により，正答確率が 1 か 0 のどちらかに偏るようになる．言い換えると，a_i が大きい項目は $\theta_j - d_i$ の正負により，「学習者 j が設問 i に正答するか否か」がはっきりと分かるようになる．これは，設問 i が，a_i の値が大きいほど，能力値が高い(能力値パラメタが d_i 以上の)学習者と，能力値が低い(能力値パラメタが d_i 未満の)学習者を正確に見分けられることを表しているため「識別力」と呼ばれる．識別力の値が大きいことは，基本的には項目が良い性質を持っていることを表し，出題に適した項目として評価される．

図 2 に，識別力パラメタを $a = 0.6, a = 1.2, a = 1.8$ と変えた 3 つの項目特性曲線を図示した．困難度パラメタは $d = 0.5$ で固定した．困難度の値が同じでも，識別力の値が大きくなる

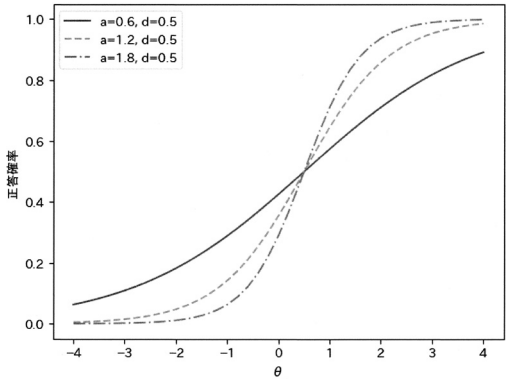


図 2. 識別力パラメタ a の異なる 3 項目に対する項目特性曲線.

ほど，曲線の傾きが大きくなることがわかる．これは，能力値パラメタが $d = 0.5$ 以上か否かで正答確率がそれぞれ 1 か 0 に偏ることを表す．すなわち，この例の中では識別力が最も大きい $a = 1.8$ のとき，学習者の能力が 0.5 以上であるかどうかが，この項目に学習者が正答するか否かではっきりと見分けられる．

(1.1)では，各項目パラメタに対して d_i と a_i の 2 つの項目パラメタがあることが分かる．そのため，これは 2PL モデルや 2 パラメタモデルなどと呼ばれる．全ての項目に対して $a_i = 1$ であるときを，特別に 1PL モデルや Rasch モデルと呼ぶ．

なお，IRT には多肢選択式の項目で分からなくても選択肢を無作為に選んで正答出来てしまう確率を考慮する 3 パラメタモデル(3PLM)が存在する．3PLM では，2PLM に比べて項目パラメタの高精度な推定に必要とされるサンプル数(学習者数)が一般に多い．

(1.1)のパラメタ推定の方法について述べる．IRT のパラメタ推定には，周辺化最尤推定(Marginalized Maximum Likelihood Estimation，MMLE)がよく用いられる．この方法では，能力値パラメタ θ_j について，(1.1)に基づいて定義される尤度を(数値的に)周辺化し周辺化尤度関数を最大にする項目パラメタを探索する．この MMLE は R 言語では “ltm” パッケージにガウス・エルミート求積法を使う方法が実装されている．その他，Hanson (2000)を実装した **pyirt** (https://github.com/17zuoye/pyirt)も Python 言語では用いられている．Hanson (2000)も，結局は能力値パラメタを周辺化していることに相当することが示されている (Hsu et al., 1999)．

本稿の構成について述べる．まず語彙学習支援において，テキストからの難度推定について説明する．次に，近年の教育データマイニングなどの手法を項目文(設問文，問題文)からの困難度パラメタ推定に着目して整理する．最後に直近の方法である Ehara (2022b)の「学習者トークン法」について説明する．

2. テキストからの難度推定

前節では IRT について簡単に説明し，特に学習者の正答/誤答の反応パターンから項目の困難度をどのように計算するかについて説明した．その中で項目の内容(項目のテキスト表現)などは一切参照されないことも説明した．しかし実際には，項目の困難度は項目文と密接な関係がある．IRT では，どの様な項目の困難度も，結局は一回学習者を集めて出題してみないとわからない．これは大変に手間であるので，項目文から直接難易度パラメタなどを推定できれば学習者を集めなくとも学習者に個別適合した適切な出題が可能かもしれない．より平たく言え

ば，項目文の中身を見て困難度を推定する課題を解きたい訳である．本節では，こうしたアプローチの先行研究を語学学習支援を中心に説明する．

2.1 応用言語学分野の研究

応用言語学(Applied Linguistics)は，平たく言えば第二言語獲得のための言語学である．第二言語獲得においては，特に，語彙獲得に膨大な労力が必要とされる．英語の場合，大学の授業程度の内容を英語で理解するためには，10,000 語程度かそれ以上の語を学習する必要があると言われている（Nation, 2006）．他の教科と異なり，量が膨大であるために，これだけの知識を学習者が持っているかどうかを網羅的に試験することは難しく，どのように試験すればよいかが研究されてきた．

多くの研究があるが簡単に述べる．まず，学習者の語彙量を計測するテストが開発された．単純に British National Corpus(BNC)（BNC Consortium, 2007)等の収録ジャンルの偏りがなるべく小さくなるように設計された一般コーパス(General Corpus)の単語頻度をまず求める．この単語頻度の降順に，単語頻度順位を用いて語の困難度を表す指標を作る．具体的には，1,000 語の困難度が同程度であること，学習者が単語頻度の多い順に語を記憶していることなど，強い仮定を置いた上で，例えば単語頻度の降順に 1,100 位の単語であれば語の困難度は 1,001 位〜2,000 位の水準であるので，レベル 2，などと考える．こうしてレベルが等しいと仮定した 1,000 語の中から 5 語程度を選定して，多肢選択式のテストを作成する．こうして作成されたテストは Beglar and Nation（2007)などで公開されている．学習者の語彙量を実際の教育と結びつける研究としては既知語率(lexical threshold)と呼ばれる指標が使われ，テキスト中の延べ語数(トークン数)換算で 95% 以上の語を知らなければ，読解力試験の点数が大きく下がることが知られている（Laufer and Ravenhorst-Kalovski, 2010）．

このように応用言語学分野では，独自の基準で語の困難度(一般コーパス中の単語頻度の降順の 1,000 語区切り)を計測する枠組みが作られた．こうして作られた語の困難度の枠組みを，IRT で検証する研究が後から行われた．特に Beglar（2010)は重要であり，試験で計測された語彙量の数値は多肢選択式の試験結果データを Rasch モデル(1PL モデル)にかけて算出される能力値パラメタとよく相関すること，$-\log$（単語頻度）の値が，語彙テストの困難度パラメタの値とよく相関することが報告された．これは英語の語彙テストという限られた設定ではあるものの，問うている項目の一般コーパスの $-\log$（単語頻度）の値を線形変換すれば，実際に語彙テストを実施しなくとも(学習者の反応パターン情報がなくとも)困難度パラメタの値を推定することが可能であることを示した点で重要である．

2.2 自然言語処理分野の語彙学習支援の研究

困難度パラメタの値を項目文のテキストから推定する研究は，Beglar（2010)とは独立に自然言語処理分野でも行われてきた．Ehara（2018)はその 1 つであり，単語テストの困難度パラメタの値を Support Vector Regression を用いて単語頻度から具体的に予測しその値について報告している．この様な，語彙を中心にした簡単な語彙テストの困難度パラメタ推定を，テキストから行う研究を実応用したのが，語学学習アプリとして有名な Duolingo(当時)のグループが行った Settles et al.（2020)の研究である．この研究では，Duolingo の大人数の反応データを用いて，困難度パラメタをテキスト中の様々な特徴量から予測することで，学習者がいないか極めて少ない項目に対してその困難度パラメタを推定し Computer Adaptive Testing(CAT)を行う手法を提案している．CAT は，IRT の困難度パラメタがわかっている場合に個々の学習者の能力パラメタを逐次的に推定する手法の総称であり，様々な手法が提案されている．

2.3 Deep Knowledge Tracing

自然言語処理とは独立に，教育 AI 分野で特に近年注目されている手法が Knowledge Tracing（知識追跡）と呼ばれる手法である．この手法は，IRT とは基本的にタスク設定が異なり，スマートフォン上の教育アプリや Massive Open Online Courses（MOOCs）のように，大人数の学習者が多数の項目に答えたログを用いた分析や予測を目的とする．基本的に，スパースな時系列データであり，学習者の正答/誤答に時間情報が紐づいている点が違いの 1 つである．学習者の能力値の向上をみることや，学習者の正答/誤答の予測精度などに大きな関心がある．ただし，こうしたアプリ上の項目は，通常，作問者などがどの様な知識を問うための項目であるのかをわかりやすくするため，各作問がどの様な能力（スキル）を問うているのか等の情報が与えられる設定が多い．

この Knowledge Tracing の設定で深層学習を用いて，従来手法より高精度を達成した手法が，Deep Knowledge Tracing（Piech et al., 2015）である．これは Long-Short Term Memory（LSTM）（Schmidhuber et al., 1997）を用いる手法である．Piech et al.（2015）は学習者を表すパラメタを陽に持たないモデルであり，学習者個別の予測をどのように行っているのかが直感的に分かりにくい．前述のように教育アプリなどでは，学習者の学習履歴（どの項目にいつ，正答/誤答したか）が残っているため，この回答した項目と正答/誤答のペアの時系列情報を，そのまま，個々の学習者の表現に用いていることが，Piech et al.（2015）のポイントの 1 つである．予測する際は，学習者がある時点までの反応パターンを与えた上で，その次の項目に対する反応を予測する形を取る．従って，例えば，全く同じ項目集合に対して全く同じ正答/誤答のパターンを示した学習者が 2 名いる場合には，Piech et al.（2015）の手法は両者を区別できない．しかし，反応パターンまで完全に一致することは殆ど起こらないので，実質的には個々の学習者に対して個別適応した予測が可能となる．

Piech et al.（2015）では予測精度の向上が中心であり，Knowledge Tracing のような時系列データを対象に IRT の能力値パラメタや困難度パラメタを推定する手法については提案されていなかった．Knowledge Tracing の設定のデータを対象に，学習者の能力値や項目の難易度といった解釈性を持ち，なおかつ高い精度を兼ね備えて持つ手法については，博士論文 Tsutsumi（2023）とその一連の著作が詳しい．

一方 Knowledge Tracing の設定では，項目の中のテキスト表現からテキストの困難度を抽出しようといった試みはあまりなされてこなかった．この理由は，おそらく教育アプリなどの設定では，項目の困難度や項目がどの様な能力を試験しようとしているのか（スキル，と表現されることが多い）といった情報は，作問者がある程度タグ付けした情報が入手できることが多いこと，数学やプログラミングなど論理的な推論を必要とする項目も多いためであろう．すなわち，スキルのタグ情報を有効活用した方が，項目文の本体から困難度パラメタの値を推定するより容易であったためであると推定される．

ただし，Knowledge Tracing の設定でも項目の中の自然文を活用しようという手法はいくつかある．1 つは，Relation-aware Knowledge Tracing（Pandey and Srivastava, 2020）という手法であり，項目の間の意味的近さの計測に自然文を活用している．直近ではプログラミング課題の練習システム（例えば，日本では Aizu Online Judge などが有名である）のログを対象に，回答が正答/誤答の様な 2 値ではなく，プログラムのコードやテキストなどオープンドメインな回答される設定で，Knowledge Tracing を行う手法が提案されている（Liu et al., 2022）．ここでも，結局，項目の項目を表現する自然文だけから困難度を算出するのではなく，項目文と学習者が書いたプログラムコードの情報を合わせて利用することで，項目の困難度情報がモデル中に暗に表現されているものと思われる．

その他 Knowledge Tracing については，近年では Abdelrahman et al.（2023）という良質な

サーベイ論文が発表されている．

2.4 外国語学習支援の学習者反応データセット

語彙テスト結果に関して，学習者反応を記録したデータセットにはどのようなものがあるのかを，ここで整理する．項目文を考慮した学習者反応予測を行いたい．そのためには項目文の文意を考慮することが重要であるような設定で試験を行い，その結果を記録したデータセットが必要となる．本節では，こうした語彙テスト結果データセットの先行研究を紹介する．ほぼ全ての公開されている英語学習者の語彙テスト結果データセットは，典型的な意味の語義についてのデータセットである．1 つは語学学習アプリ Duolingo 上の項目に対する回答データを用いた SLAM データセット（Settles, 2018）である．もう 1 つは多数の語学学習者に対して，文中のわからない語をアノテーションさせた複雑単語推定（Complex Word Identification, CWI）のデータセット（Yimam et al., 2018）である．

これらのデータセットの特徴として，各学習者は多くある項目のうちのごく一部にしか回答していないという点が挙げられる．言い換えると，学習者を行，項目を列とし，学習者の項目に対する回答内容を要素とする行列を考えた場合，これらのデータセットでは行列が疎になっている．IRT は，学習者の項目に対する回答内容から項目の困難度や学習者の能力値を推定を目標とするが，この推定のためには各学習者がなるべく多くの項目に回答している形式のデータセットが望ましい．またどちらのデータセットでも，文中の語に対する学習者の回答が記録されてはいるものの，項目について今回のデータセットのような語の通常の用例と意外と思われる用例といったようなアノテーションはされていない．さらに，Yimam et al.（2018）を含む CWI のデータセットでは，一般に提示された文に対して，学習者が難しいと感じた語が記録されているだけであり，学習者が実際にその語の意味を適切に理解しているかテストを通じた確認はしていない．すなわち意味は理解できたが難しいと感じてアノテーションした場合もあれば，単純に意味が分からなかった場合も含まれる．

IRT が適用しやすい多肢選択式の語彙テストを学習者に受験させたデータセットとしては，Beglar and Nation（2007）によるテストをクラウドソーシングで 100 人に受験させたデータセットが公開されている（Ehara, 2018）．語彙テスト結果の重要なデータセットとして，他に，個々の学習者に対して網羅的に語を知っているかどうかを自己申告式で回答してもらったデータセットがある．例えば，約 12,000 語を 15 人に対して自己申告式で回答させたデータセットが公開されている（Ehara et al., 2013）．また，最近，非公開ではあるが，NTT のグループが同種のデータセットを作成している例がある（藤田 他, 2023）．

2.5 可読性推定の研究

ここまで，項目文からの困難度推定の研究を語彙学習支援に関連する研究を中心に見てきたが，実は読みやすさ（可読性，リーダビリティ）をテキストから推定する研究も自然言語処理分野でほぼ独立に行われている．自然言語処理は人工知能の一分野という立ち位置もあり，人工知能は人間の知能を模倣するというのが典型的な研究の方向性であるため，「語学教師の読解教材に対する困難度の判断」を模倣させることで可読性を予測しようという一連の研究がある．つまり，こうした研究における「困難度」とは，語学教師などの人間の判断を記録したものであって，IRT の様な学習者の反応から求められるものではないという立ち位置である．このため，IRT を用いた語彙テストの研究と自然言語処理における可読性推定の研究は，ほぼ独立に発展してきたといってよい．このように，「困難度」という類似の概念を用いていても考え方に大きな違いがあるので，注意が必要である．

外国語学習における可読性推定の最近の研究としては，Vajjala and Lučić（2018）が複数の語

学教師に各テキストを実際に読解させて 3 段階でラベル付けした信頼性の高い可読性コーパスの研究がある．また，Martinc et al.（2021）の研究や Ehara（2022a）の研究によって，本稿でも用いている Bidirectional Encoder Representations from Transformers（BERT）（Devlin et al., 2019）等の近年の深層学習手法を用いた可読性推定の研究も行われている．

3. 学習者トークン法（Ehara, 2022b）の紹介

学習者トークン法は Ehara（2022b）が提案した方法である．項目文の文意を考慮して深層学習により学習者反応を予測し，更に，能力値や項目の困難度の抽出などもある程度可能な最近の手法であるので，以後，これについて詳解する．深層学習のアプローチで計算に多くの時間がかかる事前学習済モデルをそのまま用いて，微調整（fine-tuning）を行えるようにすることで，事前学習済モデルを有効活用できる点などに特徴がある．

前述のように，IRT（IRT）に基づくモデルは通常，学習者の回答パターンにのみ依存し，項目（項目）が自然文で書かれていても文意を理解しない．自然言語処理においては，近年，Transformer モデルに代表される，自己教師あり学習の枠組みを用いた大規模言語モデルが自然文理解で高い性能を示している（Devlin et al., 2019）．従って，項目文の理解に，これらの大規模言語モデルを用いたい．しかし，これらの言語モデルは，通常，言語のみをモデル化するため，学習者ごとに異なった判定を行うことができず，学習者反応の予測に用いることが難しい．これは，大規模言語モデルを用いた項目文の文意を考慮した学習者適応がそのままでは行えないことを示している．

Ehara（2022b）は，大規模言語モデルを項目文を考慮した学習者反応の予測問題に適用する簡便な方法を提案する論文である．項目文の文意を考慮した判定が行えているかを評価するため，外国語学習の語彙学習支援における多義語の各意味を知っているかを問う語彙テストデータセットを作成し，これを用いて提案手法を評価した．以後，理解を容易にするため，この問題に限定した用語を用いて提案手法の解説や評価を行うが，技術的には前述のように，幅広い問題に適用可能である．Ehara（2022b）の手法は，大規模な母語話者コーパスを事前学習に用いることで項目文の文脈を考慮することができる Transformer モデルの手法（BERT（Devlin et al., 2019）など）に基づく手法である．前述のように，Transformer モデルは，能力の考慮など，学習者によって異なる結果を予測する仕組みを通常持たない．Ehara（2022b）では，Transformer モデルを学習者反応予測問題に適用する手法をあわせて提案し，その予測性能が IRT による手法より高いことを示している．また，IRT の利点は学習者の能力値等を合わせて推定できる解釈性にあるが，Transformer モデルから IRT で推定した能力値とよく相関する値を抽出する手法も提案する．この研究で作成したデータセットは，今後著者の Web サイト（http://yoehara.com/）で公開を予定している．

4. 語彙テスト作成・データセット

語彙テスト作成・データセット作成は，著者が過去に語彙テスト結果データセット作成時の設定に準じて行った（Ehara, 2018）．データセットはクラウドソーシングサービス Lancers（https://lancers.co.jp/）から，2021 年 1 月に収集した．英語学習にある程度興味がある学習者を集めるため，過去に TOEIC を受験したことがある学習者のみ語彙テストを受けられると明記して，データを収集した．その結果，235 名の学習者から回答があった．Lancers の作業者は大部分日本語母語話者であるため，このデータセット中の学習者の母語は大部分日本語であると思われる．

まず，通常の語彙テストとしては Ehara（2018）と同様に，Vocabulary Size Test（VST）（Beglar

表 1. 実際の項目例.

It was a difficult period.
a) question
b) time
c) thing to do
d) book

表 2. 意外な意味を問う項目例.

She had a missed _____.
a) time
b) period
c) hour
d) duration

and Nation, 2007)を用いた．ただし VST は 100 問からなるのに対して，Ehara (2018)では，低頻度語に関する項目では Lancers 上のどの学習者もほとんどチャンスレートしか回答できていなかったことから，学習者の負担感を減らし的確な回答を集めやすくするため低頻度語 30 問を削った．すなわち，残り 70 問を通常の語彙テストとして用いた．この項目例を表 1 に示す．文中の単語に下線が引かれてあり，学習者は，この単語と交換した際に元の文と意味が最も近くなる選択肢を選ぶように求められる．この際，文法的情報から選択肢を絞れてしまわないように，選択肢は下線部と文字通り置き換えても正文となるように作られている．例えば表 1 であれば，複数形の選択肢が内容に配慮されている．

一方，学習者にとって意外であると思われる用例については，英語非母語話者 1 名が作問し，英語母語話者を含む静岡理工科大学の教員複数名に項目が英語の問題として成立しているか確認を取る方法で，作成した．この際，表 1 と同様の形式にして，“period” という単語について 2 つの項目があることが分かってしまうと，意外な語義については通常の語義以外の選択肢を選ぶことで，選択肢を絞り，意味を知らなくても回答できてしまう．そこで，Ehara (2022b) では，次の 2 つの工夫を行った．

(1) 意外な語義を問う項目については，下線部の意味について問う形にはせず，空欄を埋める形式の項目とした．これにより，意外な語義については正答を知らなければ，どの語についての項目であるのかもわからないようにした．
(2) 通常の語義についての項目を先に行ってしまうと，そこで出てきた単語と同じ語が正答であろう，という推測ができてしまう．そこで，意外な語義についての項目群を最初に行い，通常の語義についての項目群に移動したら，意外な語義についての項目群には戻れないようにした．

この 2 つの工夫を施した実際の項目例が表 2 である．“period” には通常の「期間」の他に「生理」という意味があり，これを問うている．学習者は，70 問の通常の用例の語彙テストの前に，表 2 のような項目を 13 問解くように求められる．ただし，先に解く表 2 の形式の選択肢が，表 1 の形式の項目に影響していないかどうかを後で確認できるよう，意外な語義ではあるが，通常の語義の項目群の側に対応する項目がない項目を 1 問設けた．これにより，対応する項目は 12 問となる．

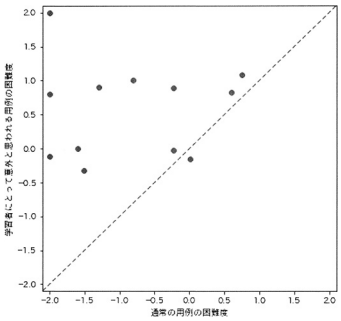


図 3. 各語の，通常の用例の困難度(横軸)と学習者にとって意外と思われる用例の困難度(縦軸)のプロット．各点は各語を表す．

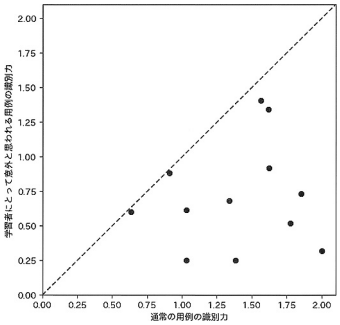


図 4. 各語の，通常の用例の識別力(横軸)と学習者にとって意外と思われる用例の識別力(縦軸)のプロット．各点は各語を表す．

4.1 作成したデータの IRT を用いた分析

前述のように，作成したデータセットでは，同じ語でも項目文の文意によって困難度が異なると思われる．本節では，まず IRT を用いて作成したデータセットを分析することで，実際に同じ語でも項目文の文意によって困難度が異なっていることを，困難度パラメタの値の違いを通じて確認する．

IRT の困難度・識別力の各パラメタを求めるには，Python 言語のライブラリである **pyirt** (https://github.com/17zuoye/pyirt)を用いた．これは，MMLE により IRT を行うライブラリである．前述のデータセットに対して，2PL モデルを用いて困難度と識別力パラメタを求めた．表 1 と表 2 のように，項目のペアが 12 組ある．通常の用例，学習者にとって意外と思われる用例の困難度パラメタを，それぞれ横軸，縦軸に表し，横軸と縦軸の縮尺・範囲を同一にプロットし図 3 に示した．各点は語を表す．

困難度の比較．図 3 の左下から右上まで，点線で対角線を示した．図 3 の横軸・縦軸とも困難度パラメタの値であり，この値が大きいほど難しいと判定される．そのため，この対角線より左上にある点は，通常の用例の困難度より，学習者にとって意外と思われる用例の困難度の方が，語彙テスト結果データからも学習者にとって回答が難しいと判定された語ということになる．今回は項目数が少ないので，図 3 の結果が偶然得られた可能性がどの程度あるか検証するため，横軸の値の列と縦軸の値の列で統計的検定を行った．Wilcoxon 検定の結果，縦軸の値の列が統計的有意に横軸の値の列より大きかった($p < 0.01$)．すなわち，縦軸の項目群の方が横軸の項目群より難しかったことが示唆される．

識別力の比較．識別力についても，図 3 と同様にプロットし，図 4 に示した．識別力は，直

観的には，高いほど，その項目で(他の項目で推定される)能力値が高い学習者と低い学習者を分けることができるという意味で，良問である度合いを表す．学習者にとって意外と思われる用例は，能力値が高い学習者でも知らないことがあり，低い学習者でも知っていることがあるため，通常の用例よりも識別力が低いと予想される．全ての語について，通常の用例の方が，意外と思われる用例よりも識別力が高いと推定されている．この結果も，Wilcoxon 検定の結果，統計的に有意であった($p < 0.01$)．

5. 学習者反応予測による評価

5.1 IRT による学習者反応予測

語の意外と思われる語義の困難度を典型的な語義の困難度で代替してしまうと，学習者が項目に正答/誤答するかを IRT で予測する際，どの程度の悪影響があるのだろうか？これを調べるために，次の実験を行った．まず，235 人の学習者を 135 人と 100 人に分ける(図 5)．意外と思われる語義の項目群(12 問)のパラメタについては前者の 135 人の学習者反応だけから，典型的な語義の項目群(70 問)のパラメタについては 235 人全員の学習者反応で推定する．この推定の際には，後者の 100 人 $\times$12 問，計 1,200 件の回答データは用いていないことに注意されたい．(1.1)より，推定された学習者の能力値 θ_j，語義の困難度 d_i を用い，$\theta_j > d_i$ であれば学習者 j が項目 i に正答，そうでなければ誤答と判定できる．項目 i の困難度パラメタとして，意外と思われる語義の 12 問の困難度パラメタを直接用いた場合と，対応する語の典型的な語義の困難度パラメタで代替した場合で，この 1,200 件の回答データの予測精度(accuracy)を比較した．予測精度の結果を表 3 に記す．その結果，直接用いた場合の予測精度は 64.4%，典型的な語義の困難度で代替した場合は 54.4% と，10 ポイントの差が出た．この差は，Wilcoxon 検定で $p < 0.01$ で有意であった．この結果から，学習者反応の予測における，語の語義ごとに困難度を推定することの重要性がわかる．より直接的に言い換えれば，この結果は，語の意外な用例の困難度を語の典型的な用例の困難度で置き換えると，学習者反応予測の精度が著しく低下することを示唆している．

5.2 Transformer モデルと IRT の性能比較

IRT を用いた手法は，学習者反応のみに依存し，項目文の意味などは全く考慮されていない．では，項目文の意味をも考慮した学習者反応予測を行うと，学習者反応のみを用いた IRT の手法より高精度に予測できるのだろうか？大規模言語モデルのうち，自然言語処理で文意を考慮した予測手法として近年多用される，BERT (Devlin et al., 2019)に代表される Transformer モデルと IRT の予測性能を比較した．

Transformer モデルは近年の深層転移学習による大規模言語モデルの代表的な手法であり，

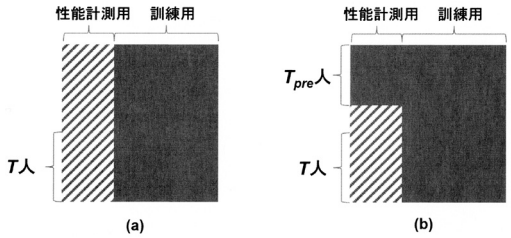


図 5. 実験設定．塗りつぶされた部分がパラメタ推定に使われる訓練データ．斜線部が性能計測に用いられるテストデータ．

表 3. 図 5 で，Ehara（2022b）のデータの「意外な語義」を性能計測用の項目セットに選び，$T_{pre} = 135$ の時の，斜線部のうち T 人の予測精度（accuracy）．

手法	精度
IRT (能力 - 235 人から推定した典型的な語義の困難度)	0.544
IRT (能力 - 135 人から推定した意外な語義の困難度)	0.644
Ehara（2022b）(bert-large-cased)	0.674 (**)
Ehara（2022b）(bert-base-cased)	**0.688** (**)
Ehara（2022b）(bert-base-uncased)	0.655
Ehara（2022b）(roberta-base)	0.681 (**)
Ehara（2022b）(albert-base-cased)	0.671 (*)

大量のラベルなしデータからの事前学習（pre-training）と，ラベル付きデータを用いた微調整（fine-tuning）という 2 種類の学習からなる．事前学習では，大量のラベルなしコーパスを用いて，当該言語の基本的な構造を学習し，入力文の言語としての自然さを計算可能にする．この過程は計算量が非常に大きいが，様々なタスクに対して汎用的に用いることができる．そこで，通常，事前学習は，**bert-large-cased** 等の英語版 Wikipedia 等を用いて訓練された **transformers**（https://github.com/huggingface/transformers/）の事前学習済モデルを用いる．事前学習済モデルの詳細情報，例えば事前学習に用いたコーパスなどの情報は https://huggingface.co/models に記載されている．多くのモデルは英語版 Wikipedia を使用している．

後段の微調整（fine-tuning）では，実際に目的とするタスクに合わせて，事前学習済モデルを追加訓練する．Ehara（2022b）のタスクにおいては，ラベルは IRT 同様，学習者が正答する場合 1，誤答する場合を 0 とする 2 値判別問題である．事前学習済モデルに項目文と学習者の両方を入力し，微調整を行いたいが，通常大規模言語モデルの微調整では言語しか入力として扱えないため，学習者の情報を入力することができない．そこで，次節に述べる方法でこの問題を解決する．

5.3 Ehara（2022b）の手法

Ehara（2022b）の手法を説明する．適応的学習者反応予測のタスク設定を図 5 に示す．テスト結果データは，学習者を行，項目を列とし学習者の項目に対する反応を要素とする．図 5 では，塗りつぶされた部分が訓練データ，斜線部が性能計測用のテストデータとなる．図 5(a) の設定では，完全に新しい項目に対する性能計測用の T 人の反応（正答するか/誤答するか）を高精度に予測することが目的である．(a) の設定では，性能計測用データ中の項目は訓練データにない全く新しい項目なので，項目文の情報を一切用いない IRT 等の手法では項目のパラメタ推定が行えず学習者反応予測ができない．一般には，項目を少人数の学習者に事前に受験してもらい項目の特性を事前に確かめるケースもある．このケースを含み (a) を一般化した設定を図 5(b) に示した．「少人数の学習者」の人数を T_{pre} とした．(a) は，(b) で $T_{pre} = 0$ の場合とみなせる．

大規模言語モデルに特殊なトークン（語）を加えて微調整を行い，様々な問題設定に対応させる手法は機械翻訳などの他タスクで知られておりライブラリ上で特殊なトークンを加える機能が用意されている．Ehara（2022b）では，この機能を利用することで，学習者に対応するトークン（学習者トークン）を作り，これを入力系列の最初に置くことによって判別を行う手法を提案した（図 6）．Ehara（2022b）では “It was a difficult period.” のような項目文中の語の語義を多肢選択式で問うデータセットを用いているため，図 6 もこの例にならっているが，語学学習

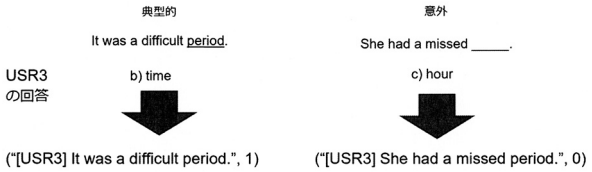


図 6. 学習者トークンの導入（Ehara, 2022b）.

支援に限らず，一般に項目文で示される項目を，指定した学習者が正答できるかを予測する識別器を構成している．例えば，学習者 ID が 3 番の学習者を表すトークン “[USR3]” を導入し，“[USR3] It was a difficult period.” が入力であれば，3 番の学習者が “It was a difficult period.” という文から成る項目に，正答するか否かを予測する問題に帰着させる．入力文はそのままで，入力文の直前に単純に学習者トークンを挿入する．導入するトークン数は学習者数と同数である．

重要な点として，Ehara（2022b）では，文中のどの語についての項目であるかという情報や誤答選択肢の情報は与えていない．すなわち，判別器は，表 1 のどの単語に下線が引かれているかや，表 1 や表 2 の正答以外の選択肢の情報を用いない．単純に項目文に学習者が正答できるか予測（識別）する予測器を構成していると解釈できる．これにより，Ehara（2022b）は，項目がテキストで構成されてさえいれば適用可能である．

Transformer モデルのその他の実験設定については多用される設定とした．判別には，**transformers** ライブラリの **AutoModelForSequenceClassification** を用いた．微調整の訓練には Adam 法（Kingma and Ba, 2015）を用い，バッチサイズは 32 とした．

6. 実験結果

BERT 他，Transformer モデルを用いた結果を，表 3 に示す．*は IRT の最高性能と比較して Wilcoxon 検定で統計的有意であることを表し，**は $p < 0.01$，*は $p < 0.05$ を表す．また提案手法の () 内は用いた事前学習済モデル名である．表 3 では，まず，学習者トークンを導入した提案手法が，IRT を用いた従来手法より高い性能を達成していることが分かる．この実験結果は，項目文の意味を考慮することで IRT より高精度な判別が行えることを示している．

次に，“roberta-base” は cased（大文字・小文字を区別するモデル）であるのに対し，“albert-base-v2” は uncased（大文字・小文字を区別しないモデル）である．この結果から，良い精度を得るためには “cased”，すなわち，大文字と小文字を区別して扱うモデルでなければならないことが示唆される．この理由は次のように推察される．この実験環境では各項目の項目文は短い文から構成されているため，モデルは大文字で始まる文の開始を認識する必要があるためであろう．

さらに，表 3 では **bert-base-cased** が最も高い性能を示した．より大きな事前学習済モデルである **bert-large-cased** よりも **bert-base-cased** が高い性能を示した理由として次のことが考えられる．学習者特性を表す学習者トークンの単語埋め込みベクトルは，今回作成した比較的小さい訓練データで訓練しているため，小さいモデルの方が微調整（fine-tuning）に適していた可能性がある．

また，Ehara（2022b）では，訓練データにない全く新しい項目に対する学習者反応予測を行う設定，すなわち $T_{pre} = 0$ の設定では実験していない．$T_{pre} = 0$ の設定で，本稿で新しく実験した結果を表 4，表 5 に示す．この設定では，学習者反応予測を行う「性能計測用」の項目につ

表 4. 図 5 で，Ehara (2022b) のデータの「意外な語義」を性能計測用の項目セットに選び，$T_{pre} = 0$ の時の，斜線部のうち T 人の予測精度 (accuracy)．すなわち，全く新しい，訓練データとも異質な項目に対する学習者反応を項目文を考慮することで予測する，最も過酷な設定での精度．

手法	精度
Ehara (2022b) (bert-base-cased)	0.601
Ehara (2022b) (bert-large-cased)	0.585
Ehara (2022b) (roberta-base)	**0.613**
Ehara (2022b) (roberta-large)	0.601

表 5. 図 5 で，Ehara (2022b) のデータの「典型的な語義」を性能計測用の項目セットに選び，$T_{pre} = 0$ の時の，斜線部のうち T 人の予測精度 (accuracy)．すなわち，全く新しいが，訓練データとは同質な項目に対する学習者反応を，項目文を考慮することで予測する，表 4 よりは過酷ではない設定での精度．

手法	精度
Ehara (2022b) (bert-base-cased)	0.632
Ehara (2022b) (roberta-base)	0.613
Ehara (2022b) (roberta-large)	**0.662**

いては回答した学習者がいないため，IRT では項目の項目パラメタを求められず性能を比較できないため，IRT を表から除外した．

表 4 は，表 3 とは $T_{pre} = 0$ であること以外は同一の設定である．ただし，Ehara (2022b) は，外国語の語彙テストを題材に，訓練データに典型的な語義の項目，性能計測用のテストデータに意外な語義の項目を置くというように，訓練データとテストデータの性質が違うものを計測している．

この設定は一般的とは言い難いので，より一般的に訓練データも性能計測用のテストデータの方も典型的な語義を問う同質の項目 (ただし，もちろん項目集合自体は完全に disjoint である) として，計測しなおしたものが表 5 である．

データが小規模であるためか，どのデータセットでも **bert-base-cased** で比較的高い精度が達成できていること，最も過酷な設定である表 4 では表 3 より全体的に精度が大きく低くなっていること，表 5 は，表 4 ほどには過酷ではない設定のため，全体的により高い精度が達成されていることがわかる．

7. 解釈性—学習者トークンからの能力値抽出

IRT は学習者の能力パラメタを持つことにより，学習者の特性について解釈しやすい．一方，Transformer モデルでは学習者の特性は学習者トークンに対する単語埋め込みベクトルという多次元の形で表現されており，そのままでは直感的な解釈が難しい．しかし，Transformer モデルは個人化判別問題で高精度を達成しているので，学習者トークンの単語埋め込みベクトルの中に能力値の情報が含まれていると考えられる．

微調整後の **bert-large-cased** の場合の学習者トークンに対する単語埋め込みベクトルのみを集めた．すなわち，学習者の人数分の単語埋め込みベクトルの集合がある．このベクトル集合に対して主成分分析を行い，その第一主成分得点と IRT の能力値パラメタを比較した (図 7)．各点は学習者を表す．IRT の能力値パラメタの算出には，Python の **pyirt** ライブラリを用い

た．両者は相関係数 0.72 という強い相関を示した($p < 0.01$)．これにより，提案手法を用いた場合でも，能力値は学習者トークンの第一主成分得点として容易に抽出できることが分かった．これにより，提案手法は文意を考慮することにより IRT より高い精度を達成しながら，IRT と同様に「能力値を取り出せる」という高い解釈性を持つことが示された．

7.1 IRT パラメタの推定手法に依らないことの確認

IRT の能力値パラメタは，データが同じでも，どの推定用ソフトウェアを用いるかによって，多少の違いが生じることが知られている．前節では，図 7 では **pyirt** ライブラリを用いたが，この推定用ソフトウェアの特性によって統計的有意性が生じた可能性もある．

そこで，確認のため，同じデータを，**pyirt** とはプログラミング言語も異なる全く独立の実装である R 言語の **ltm** パッケージ(https://cran.r-project.org/web/packages/ltm/index.html)を用いて推定した．これは教育心理学の標準的な教科書で使用されているソフトウェアである（Paek and Cole, 2019）．その結果を図 8 に示す．この場合も，目で見て取れる相関があり，実際に相関係数は 0.72 で，やはり統計的有意性を示した($p < 0.01$)．従って，IRT の能力値パラメタを算出するソフトウェアによらず学習者トークンから能力値が抽出できることが示された．

pyirt も **ltm** も，MMLE を用いていると記されている．ただし，同じ MMLE でも，周辺化する際の数値積分の方法が異なっており，**ltm** ではデフォルトでは 15 点からなるガウス=エルミート求積を使用していることが，ドキュメント(**ltm** の Reference manual である ltm.pdf)に記載されている．一方, **pyirt** の側では Hanson (2000)を使用していると記載されているだけで, 求積方法についてはドキュメント(https://github.com/17zuoye/pyirt/wiki/Model-Specification)に書かれていない．Hanson (2000)は，学習者の能力値パラメタを連続変数とみなしても実装

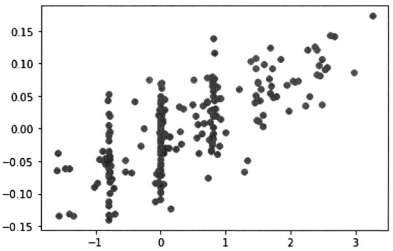


図 7. IRT の能力パラメタ(横軸，**pyirt** によって算出)と，学習者トークンの単語埋め込みベクトルの第一主成分得点(縦軸)．

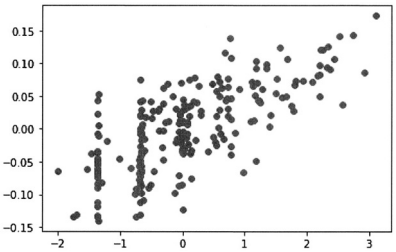


図 8. R 言語の **ltm** パッケージによって算出した IRT の能力パラメタ(横軸)と学習者トークン埋込ベクトルの第一主成分(縦軸)の関係．

上は有限の点で数値積分することになるため，学習者の能力値パラメタがはじめから決められた離散値しか取らないと仮定することで MMLE と EM アルゴリズムの関係性を整理し直した未出版のノートである．Hanson（2000）の手法は，記法が異なるだけで手法としては MMLE と同一であることが，その中で述べられている．Hanson（2000）の記述も数値積分の手法によらない記法となっているため，Hanson（2000）の手法を使用している事だけからでは **pyirt** がどのような数値積分を内部で行っているのかはわからない．そこで，筆者が **pyirt** のソースコードを読んで確認した限りでは，**pyirt** はデフォルトでは能力値パラメタについて $[-4, 4]$ の区間を 10 等分して離散値として扱い，単純な区分求積法を使っているようである．つまり，積分点は単純に区間の等分であって，多くの学習者が該当する能力値の区間でも細かく積分点が取られることはない．図 7 の縦の筋は，この荒い数値積分のためではないかと推察される．また **ltm** では，学習者の能力値パラメタの算出は Expected A Posteriori（EAP）で算出されている．**pyirt** ではドキュメンテーションに能力値パラメタ計算法についての明確な記載が無いが，著者がソースコードを読んだ限りでは，こちらも EAP で能力値を算出しているようである．

学習者トークンの埋め込みに第一主成分得点には能力値が含まれていることが示されたが，第二以降の主成分得点には能力値との相関はあるのだろうか？この疑問について調べるために，第二主成分得点についても能力値との相関係数を計算したが，統計的に有意な相関は得られなかった．このことから，学習者の能力値は各学習者の学習者トークン埋め込みベクトルの第一主成分得点にのみ保持されていることがわかる．

8. 予測確率を用いた困難度の抽出

Ehara（2022b）の手法では，マスク言語モデルである BERT 本体には，困難度を直接表現するパラメタは存在しない．また，パラメタも膨大であるため，fine-tune されたモデルから直接困難度パラメタを抽出することは難しい．しかし，このモデルは学習者個別の予測が可能である．そこで，同じ T 人の集合に対する各項目の正答確率を予測させた上で，その確率の平均値と T 人をかけあわせて，平均で何人正答者がいると予測されているかを計算することで，困難度を表現する値自体は抽出可能と考えられる．実際に図 9 に，これを行った結果を示した．縦軸は，前述の方法による予測正答者数である．IRT を用いて図 5 の全てのデータが見えている状態で推定した項目の困難度パラメタの値が横軸である．相関係数は，図 9 では 0.78（$p < 0.01$）であり，BERT モデルから IRT の困難度と統計的有意に相関する困難度パラメタの推定値を取り出せていることが分かる．

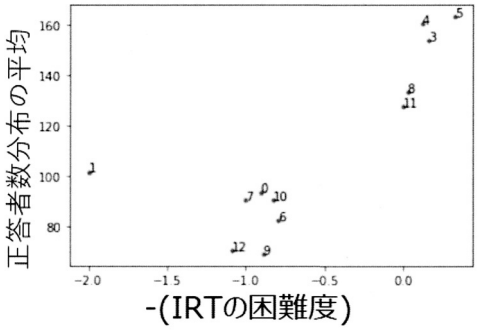


図 9. 予測正答者数と −(IRT の困難度).

9. 議論

表 1 と表 2 はどちらも多肢選択式であるが，表 1 が下線部の同義表現を選択肢から探させる項目，表 2 が空欄補充形式になっていることの影響について議論する．まず，「空欄補充」(fill-in-the-blank)といった場合，空欄を埋める文字列を学習者自身が思いつかないといけない形式の項目を意味する事も多くある．この意味での空欄補充であれば，「一般論としては空欄補充のほうが難易度が高い」といえる．

一方，今回の設問は，空欄に当てはまるものの選択肢が与えられているため，あくまで多肢選択式の出題方法の 1 つとして，項目文中の下線部の同義表現を探す形式(下線部形式)と空欄補充形式との間で困難度に差があるかどうかが議論の焦点になる．結論としては，著者の知る限り，多肢選択式の項目において項目文の下線部形式と空欄補充形式の差が，他の要因に比べて大きく困難度に影響することを示した既存研究はない．むしろ，下線部形式か空欄補充形式かの差よりも他の要因の方が困難度に大きく影響することが示されている．熊澤 (2010)では，言語テストにおける，多肢選択式の出題方法の 1 つとしての空欄補充形式と他の形式の困難度等の比較を，英語学習者 608 名を対象として行っている．熊澤 (2010)では，下線部形式を扱ってはいないものの，多肢選択式の空欄補充形式として，本稿のような 1 文中の空欄補充と複数文からなる短文会話の中の空欄を補充する「クローズ」という形式の 2 つが比較されている．多相ラッシュモデルを扱うソフトウェアである FACETS を用いて分析したところ，通常の空欄補充の困難度が −0.10 であり，クローズの空欄補充の困難度が 0.36 と報告されており，同じ空欄補充形式でも大きな違いが出ることが示されている．熊澤 (2010)の分類では，下線部形式と同様，「最初に提示される選択肢ではないテキストに空欄がなく」，「選択肢として英文が並べられ」，「1 つだけ正しい英文を選ぶ」という形式の項目は，提示された和文と同じ意味の英文を選択肢から選ぶ「和文英訳」しかないが，この形式の項目の困難度は −0.08 と，通常の空欄補充形式ともっとも近い値を取っている．

また，多肢選択式の分類では，下線部形式と空欄補充形式の様な項目文による分類の他に，選択肢による分類もあり得る．池上 (2015)の分類では，本稿の下線部形式も空欄補充形式も，どちらも「通常の択一式」に分類される．一方，「(他の選択肢に)正答なし」が選択肢に入る「「正答なし」を含む択一式」や，複数の選択肢が正答になり得る「複数選択式」があり，この順番に困難度が高くなることが報告されている．

以上のように，下線部形式か空欄補充形式か以外の要因の方が困難度に影響を及ぼすことが報告されており，本稿では，空欄の有無の違いはあるとはいえ，どちらも択一式の形式の項目を扱っていることから，空欄を用いる形式以外の要因が困難度に影響していると考えることが妥当であると思われる．すなわち，最初に設定した，通常の意味に関する項目か，意外な意味に関する項目であるかの差異が困難度により大きく影響していると推察される．

10. おわりに

本稿では，外国語の語彙学習支援を中心に，IRT に注目して，特に項目のテキスト表現から項目の困難度パラメタを推定する手法についてレビューを行った．そして，直近の BERT などのマスク言語モデルの事前学習済みモデルをそのまま活用して，能力値や困難度を抽出することの可能な手法を Ehara (2022b)に基づき，詳細に説明した．

今後のこの分野の研究の方向性について述べる．本稿では，語彙学習支援を中心に述べ，実験もこれに関して記した．しかし，もちろん，教育応用は語彙学習支援だけではない．大規模言語モデル(Large Language Models, LLMs)の性能が近年，急速に向上しており，特に ChatGPT (https://chat.openai.com/)等の大規模言語モデルを用いれば，項目文を解釈して，項目の困難

度を推定することはかなり可能であると思われる．実際，ChatGPT を用いて，外国語学習の項目で，困難度を数値指定できることについては，著者はある程度確かめている．外国語学習以外の応用に広げていることが今後の課題となろう．また，困難度が関心の中心であったが，大規模言語モデルから識別力の値を抽出する手法については，著者の知る限りほとんど報告がない．これも，著者はある程度予備実験を行っているが，今後のおもしろい研究課題になると考えられる．

参 考 文 献

Abdelrahman, G., Wang, Q. and Nunes, B. (2023). Knowledge tracing: A survey, *ACM Computing Surveys*, **55**(11), 1–37.

Baker, F. B. and Kim, S.-H. (2004). *Item Response Theory*: *Parameter Estimation Techniques*, CRC Press, Boca Raton, Florida, USA.

Beglar, D. (2010). A Rasch-based validation of the vocabulary size test, *Language Testing*, **27**(1), 101–118, https://doi.org/10.1177/0265532209340194.

Beglar, D. and Nation, P. (2007). A vocabulary size test, *The Language Teacher*, **31**(7), 9–13.

BNC Consortium (2007). The British National Corpus, XML Edition, Oxford Text Archive, http://hdl.handle.net/20.500.14106/2554（最終アクセス日 2024 年 2 月 29 日）.

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*, 4171–4186.

Ehara, Y. (2018). Building an English vocabulary knowledge dataset of Japanese English-as-a-second-language learners using crowdsourcing, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, 484–488.

Ehara, Y. (2022a). Analyzing readability of scientific abstracts for ESL learners, *Companion Proceedings 12th International Conference on Learning Analytics & Knowledge* (*LAK22*), 86–88.

Ehara, Y. (2022b). No meaning left unlearned: Predicting learners' knowledge of atypical meanings of words from vocabulary tests for their typical meanings, *Proceedings of Educational Data Mining* (*short paper*), 50.

Ehara, Y., Shimizu, N., Ninomiya, T. and Nakagawa, H. (2013). Personalized reading support for second-language web documents, *ACM Transactions on Intelligent Systems and Technology*, **4**(2), 31:1–31:19, http://doi.acm.org/10.1145/2438653.2438666.

藤田早苗, 小林哲生, 服部正嗣 (2023). 日本語母語話者を対象とした英語の語彙数調査, Technical Report, No. 3, NTT コミュニケーション科学基礎研究所, 京都.

Hanson, B. (2000). IRT Parameter Estimation using the EM Algorithm, http://www.openirt.com/b-a-h/papers/note9801.pdf（最終アクセス日：2024 年 2 月 29 日）.

Hsu, Y., Ackerman, T. A. and Fan, M. (1999). The relationship between the Bock-Aitkin procedure and the EM algorithm for IRT model estimation, *ACT Research Report Series*, **99**(7), 1–35.

池上真人 (2015). 択一式と複数選択式の多肢選択文法問題の比較研究, 四国英語教育学会紀要, **35**, 15–24, https://doi.org/10.32276/seles.35.0_15.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization, *Proceedings of the 3rd International Conference on Learning Representations*, https://doi.org/10.48550/arXiv.1412.6980.

熊澤孝昭 (2010). 多肢選択式項目の項目形式が文法テストパフォーマンスに与える影響について, *JALT Journal*, **32**(2), 169–184.

Laufer, B. and Ravenhorst-Kalovski, G. C. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension, *Reading in a Foreign Language*, **22**(1), 15–30, https://eric.ed.gov/?id=EJ887873.

Liu, N., Wang, Z., Baraniuk, R. and Lan, A. (2022). Open-ended knowledge tracing for computer science education, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 3849–3862, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, https://aclanthology.org/2022.emnlp-main.254.

Martinc, M., Pollak, S. and Robnik-Šikonja, M. (2021). Supervised and unsupervised neural approaches to text readability, *Computational Linguistics*, **47**(1), 141–179.

Nation, I. (2006). How large a vocabulary is needed for reading and listening?, *Canadian Modern Language Review*, **63**(1), 59–82.

Paek, I. and Cole, K. (2019). *Using R for Item Response Theory Model Applications*, Routledge, Oxfordshire, UK.

Pandey, S. and Srivastava, J. (2020). RKT: relation-aware self-attention for knowledge tracing, *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 1205–1214.

Piech, C., Bassen, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L. J. and Sohl-Dickstein, J. (2015). Deep knowledge tracing, *Advances in Neural Information Processing Systems*, **28**, 505–513.

Schmidhuber, J., Hochreiter, S. et al. (1997). Long short-term memory, *Neural Computation*, **9**(8), 1735–1780.

Settles, B. (2018). Data for the 2018 Duolingo Shared Task on Second Language Acquisition Modeling (SLAM), Harvard Dataverse, V4, https://doi.org/10.7910/DVN/8SWHNO.

Settles, B., LaFlair, G. T. and Hagiwara, M. (2020). Machine learning–driven language assessment, *Transactions of the Association for Computational Linguistics*, **8**, 247–263, https://aclanthology.org/2020.tacl-1.17.

Tsutsumi, E. (2023). Item response theory based on deep learning with independent student and item networks, Ph.D. Thesis, Graduate School of Informatics and Engineering, The University of Electro-Communications, Tokyo.

Vajjala, S. and Lučić, I. (2018). OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification, *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 297–304, Association for Computational Linguistics, New Orleans, Louisiana, https://www.aclweb.org/anthology/W18-0535.

Yimam, S. M., Biemann, C., Malmasi, S., Paetzold, G., Specia, L., Štajner, S., Tack, A. and Zampieri, M. (2018). A Report on the complex word identification shared task 2018, *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 66–78, Association for Computational Linguistics, New Orleans, Louisiana, https://doi.org/10.18653/v1/W18-0507.

Interpretable Natural Language Processing Models for Educational Applications

Yo Ehara

Faculty of Education, Tokyo Gakugei University

In education applications, measuring the ability of a learner or the characteristics of a question, such as the difficulty of a question, is a fundamental task that has broad utility in learning support systems and other applications. If human teachers can interpret the abilities of examinees and the characteristics of questions rather than simply using models to predict whether subjects will answer a given question correctly, the abilities and characteristics can also be used by human teachers when teaching. In statistics, item response theory (IRT) has been used to estimate interpretable parameters from the response patterns of test takers although IRT does not use the natural language text of each question. By contrast, in natural language processing, there has been interest in research to extract item characteristics such as difficulty from the text of questions. In particular, research on difficulty estimation from text has been active in applications such as language learning support, where values that can explain much of the difficulty of a question can be extracted from easy-to-extract features, such as word frequency. In the present paper, we explain how difficulty estimation from text is related to IRT, introducing research in various fields in addition to statistics, with particular focus on learning vocabulary and reading in second languages. We then discuss recent neural methods such as self-supervised learning and Transformers, which have achieved high accuracy in recent years in analyses that consider the meaning of text.

Key words: Item response theory, natural language processing, second language acquisition.

統計数理(2024)
第 72 巻 第 1 号 43–59

特集「心理統計学の新展開」
[原著論文]

項目露出ペナルティを用いた整数計画法による自動並行テスト構成

渕本 壱真[†]・植野 真臣[†]

(受付 2023 年 6 月 26 日；改訂 12 月 27 日；採択 2024 年 1 月 24 日)

要　　旨

e-Testing の特徴は異なる問題で構成されるが同一精度の測定を実現できるテストの自動構成であり，その重要な課題は可能な限り多くのテストを生成することである．自動テスト構成手法は多数存在するが，整数計画法を用いた最大クリークが現在最も多くのテストを高い測定精度で生成できることが報告されている．しかし，この手法は，テスト間に項目の重複を許すため，項目の出題頻度に偏りを生じさせ，テストの信頼性を低下させる．この問題を解決するために，本研究では整数計画法の目的関数に露出数を所与としたロジスティック関数による以下の 2 種類のペナルティ，(1)ロジスティック関数による決定論的ペナルティ，(2)ロジスティック関数による確率論的ペナルティ，を提案する．数値実験により，提案手法はテスト数を減らすことなく露出数の偏りを減らすことを示す．

キーワード：自動テスト構成，項目反応理論，e-Testing，項目露出問題，整数計画法．

1. はじめに

e-Testing とは，異なる問題で構成されるが，同一精度の測定を実現できるコンピュータテストのことである（Ueno, 2021; Ueno et al., 2021）．e-Testing を用いることで，同一能力の受検者が異なるテストを受検しても同一得点となる保証がある．そのために，受検者が同一精度で複数回受検が可能となる．我が国においても医療系大学間共用試験や情報処理技術者試験などが e-Testing で行われている．また，大学入学試験や公務員試験での導入も検討されている（植野, 2023）．

歴史的には，e-Testing のアイデアは Lord and Novick (1968)が同じ真の得点を測定する 2 つのテストについて並行テストと定義したことから始まる．しかし，並行テストは古典的テスト理論における仮定であり，このようなテストの実現は困難であった．そのため，Samejima (1977)は項目反応理論(Item Response Theory: IRT)（例えば，van der Linden, 2017)を用いて，並行テストの概念を拡張した．具体的には，項目反応理論におけるテスト情報量の逆数が受検者能力推定値の漸近分散に収束することを用いて，この値が等価なテストを弱並行テストとして定義した．e-Testing の普及に伴い，この弱並行テストの概念に基づいた自動テスト構成手法が数多く提案されている（Boekkooi-Timminga, 1990; Armstrong et al., 1994, 1998; van der Linden and Adema, 1998; van der Linden, 2005; Songmuang and Ueno, 2011; Ishii et al., 2013, 2014; Ishii and Ueno, 2017; Fuchimoto et al., 2022）．

[†] 電気通信大学 大学院情報理工学研究科：〒 182–8585 東京都調布市調布ヶ丘 1–5–1

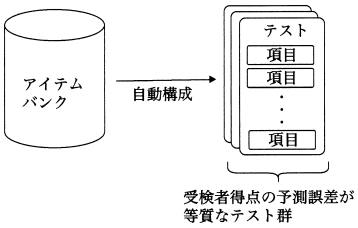


図 1. アイテムバンクからの自動テスト構成.

一般に，e-Testing ではテストの管理方法としてアイテムバンク方式が用いられる．アイテムバンクとは出題する問題(以降，項目と呼ぶ)の出題分野や統計データ等を格納しているデータベースのことである．自動テスト構成では所望のテストの性質を満たす項目の組合せをアイテムバンクから計算機により探索する．

図 1 は自動テスト構成手法の概念図である．一般に，自動テスト構成ではアイテムバンクから互いに受検者得点の予測測定誤差が等質となるように異なる項目の組合せを列挙する．これにより，同一能力の受検者が異なるテストを受けても同一の得点となることが保証される．

自動テスト構成手法で最も有名な Big Shadow Test method(BST 法)では混合整数計画法を用いて，目標とする受検者の測定誤差からの差異が最小となるテストを貪欲法で逐次的に構成する（van der Linden, 2005)．しかし，BST 法は逐次的に誤差が小さいテストから構成するため，テスト数の増加につれて，受検者の予測測定誤差が大きくなる問題がある．そのため，BST 法はテスト数を十分に確保できず，実用的でない．例えば，年間 20 万人以上受検する情報処理技術者試験の一区分「IT パスポート試験」(独立行政法人情報処理推進機構, 2023)では全国 47 都道府県に設置された複数の会場で月に数回開催されている．また，年間 1 万人以上受検する医療系大学間共用試験（公益社団法人医療系大学間共用試験実施評価機構, 2023)では各大学ごとにカリキュラム(臨床実習開始時期)に応じて異なる時期にテストを受検する必要がある．さらに，これらの試験ではカンニング等の不正防止のために，同一試験会場であっても受検者ごとに異なるテストを出題している．したがって，受検者数以上のテストを用意しなければ同じ項目で構成されるテストが再出題され，テストの信頼性低下につながる（Wainer, 2000)．

この問題を解決するために，Ishii et al. (2013)はその当時，最も多くのテストを構成する最大クリーク法を提案した．この手法は自動テスト構成をグラフ上で定義される最大クリーク問題に帰着させる．具体的には，与えられたアイテムバンク・テスト構成条件で構成可能な全てのテストを頂点集合とし 2 つのテストが等質かつ共通する項目の数が一定数以下である場合に，頂点(テスト)間に辺を引いたグラフからクリークと呼ばれる任意の 2 頂点が隣接している最大の部分グラフ構造を探索することで自動テスト構成を行う．

この手法は理論的に最大数のテスト構成を保証するが，構成可能な全てのテストを頂点とするグラフ構造は組合せ爆発的に大きくなるため，最大クリークを探索することやグラフ構造をメモリ上に保存することは困難である．そのため，Ishii et al. (2014)はグラフ全域から部分グラフをランダムに抽出し，ここから最大クリーク探索を繰り返すことによりグラフ全体の最大クリークを近似的に探索する手法を提案した．本手法により，当時の既存研究よりも 10〜100 倍以上多くのテストを構成できるようになっている．

しかし，最大クリーク探索は最先端の最大クリーク探索手法（Tomita et al., 2016; Li et al., 2017)を用いても，$O(|V|^2)$($|V|$ はグラフの頂点数)の空間計算量を少なくとも必要とするため，(著者らの計算機環境で)最大で 10 万のテストを構成することが限界であった．そこで，Ishii

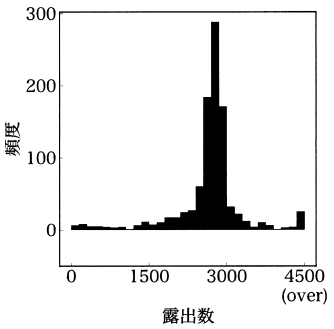


図 2. 露出数のヒストグラムの例.

and Ueno (2017)は探索中のクリークの全頂点と隣接する頂点を整数計画法を用いて逐次的に探索することで，計算に必要な空間計算量を $O(|V|)$ へ減少させる手法を提案した．これにより，10 万を超えるテストを構成できるようにした．ただし，整数計画法の時間計算量が $O(2^n)$ (n はアイテムバンクの項目数)と大きく，テスト構成数の改善は僅かなものであった．

整数計画法の計算時間を改善するために，Fuchimoto et al. (2022)は探索中のクリークの全頂点と隣接する頂点を並列探索する Hybrid Maximum Clique Algorithm Using Parallel Integer Programming method(HMCAPIP 法)を提案した．本手法により，最も時間を要した整数計画法による頂点探索を並列化することで探索時間を大幅に減少できた．具体的には多くの条件下で探索時間を非並列化時の約 25% 程度に抑えられた (Fuchimoto et al., 2022)．さらに，並列化探索で得られた頂点を整数計画法の目的関数の下限値として用いることで分枝限定法の効果により探索を高速化した．結果として，最大 7 日間で当時の最新研究の約 2.7 倍にあたる約 27 万のテスト生成を実現している (Fuchimoto et al., 2022)．

しかし，これらの手法はテスト間に重複を許すことで各項目の出題回数(以降，露出数と呼ぶ)に偏りが生じる．図 2 は Synthetic Personality Inventory (SPI)試験 (Recruit, 2023)で実際に用いられていたアイテムバンクを用いて従来手法 (Fuchimoto et al., 2022)によりテスト構成したときの各項目の露出数のヒストグラムを示したものである．図 2 より，露出数のばらつきが大きいことがわかる．特に，4500 回以上出題される項目やほとんど出題されない項目が存在する．このような露出数の偏りはさまざまな弊害を生じる．最も深刻な問題は，露出数の大きい項目が受検者間で共有され受検対策されてしまい，その項目の信頼性を低下させることである (Wainer, 2000)．また，作問にかかるコストは多大であり，露出の低い項目は可能な限り出題することが望ましい．そのため，露出数は可能な限り一様であることが理想的である．

露出数の偏りを軽減するために，Ishii and Ueno (2015)は最大クリーク法と整数計画法を用いた手法によりテストを構成し，その中から最も露出率が小さいテスト構成を選択するアルゴリズムを提案した．具体的には，探索した全てのテスト群を候補として保存しておき，最後に候補中で最も露出率が小さいテスト群を出力する．これによって，従来手法よりも最大露出率を約 0.1%〜4% 軽減できた．しかし，これらの手法は露出の偏りを軽減する定式化がされていないため，その改善に限界がある．

その他に，適応型テストでは数多くの露出制御手法が提案されている (Hetter and Sympson, 1997; van der Linden and Reese, 1998; Stocking and Lewis, 1998, 2000; van der Linden and Veldkamp, 2004; van der Linden, 2017; van der Linden and Choi, 2020)．ここで，適応型テストとは受検者の能力を逐次的に推定しながら，その能力に応じて測定精度が最も高い項目を出題す

るテスト形式の一つである．最も代表的な手法として，van der Linden and Reese (1998)は最大露出数の上限を制約式に与えた整数計画問題を用いて項目を選択する手法を提案した．この手法は最大露出数を厳密に制御できるが，受検者の能力推定精度が低下する．この問題を緩和した手法として，確率的アプローチが提案されている（Hetter and Sympson, 1997; van der Linden and Reese, 1998; Stocking and Lewis, 1998, 2000; van der Linden and Veldkamp, 2004; van der Linden, 2017; van der Linden and Choi, 2020）．例えば，Sympson-Hetter 法では各項目の露出数に応じた出題確率をシミュレーション実験により求め，その確率に応じて出題を行う（Hetter and Sympson, 1997; Stocking and Lewis, 1998, 2000）．さらに，シミュレーション実験不要な手法として，van der Linden らは各項目の出題可能な確率を意味する適格確率を定義し，受検者ごとに適格確率に従ってアイテムバンクからその項目を除くことで露出数を制御する手法を提案した（van der Linden and Reese, 1998; van der Linden and Veldkamp, 2004; van der Linden, 2017; van der Linden and Choi, 2020）．しかし，これらの手法（Hetter and Sympson, 1997; van der Linden and Reese, 1998; Stocking and Lewis, 1998, 2000; van der Linden and Veldkamp, 2004; van der Linden, 2017; van der Linden and Choi, 2020)は最大露出率の抑制のみを目的としており，露出数を一様に近づけることは限定的である．そのため，これらのアイデアを並行テストに適用することは可能かもしれないが，本論文が目標としている露出数を一様に近づけ各項目の露出数を減少させることは難しい．

露出数の偏りを防ぐために，本研究では露出数を所与としたロジスティック関数による 2 種類のペナルティ項を HMCAPIP 法における整数計画問題の目的関数に追加することを提案する．1 つ目はこのロジスティック関数を用いた決定論的ペナルティ項である．この決定論的ペナルティ項は露出数に応じた負の重みを常に各項目の決定変数に与える．2 つ目はロジスティック関数を用いた確率論的ペナルティ項である．確率論的ペナルティ項は数理計画法の Big-M 法（Williams, 1990)に基づいて，露出数に応じた確率により，大きな負の重みを各項目の決定変数に与える．

本論文では，提案手法の有効性をシミュレーション及び実データを用いて示した．具体的には，従来手法と比較して，テスト構成数を減少させることなく，露出数の偏りを抑制できることを示した．

2. 項目反応理論

項目反応理論は受検者の項目への正答確率をモデル化し，異なる項目から構成されるテストを受けた受検者の能力を同一尺度上で評価できる．

一般的に，テストでは項目 $i(=1,\ldots,n)$ に対する受検者 $j(=1,\ldots,m)$ の反応 u_{ij} を以下のように表す．

$$u_{i,j}=\begin{cases}1 & i \text{ 番目の項目に受検者 } j \text{ が正答}\\ 0 & \text{それ以外}\end{cases}$$

本論文では，項目反応理論の中で最もよく使われている 2 母数ロジスティックモデル(2-Parameter Logistic Model:2PLM)を用いる．このモデルでは，能力値 $\theta_j \in (-\infty,\infty)$ を持つ受検者 j が項目 i に正答する確率 $p_i(\theta_j)$ を以下のように定義する．

$$\begin{aligned} p_i(\theta_j) &\equiv p(u_{ij}=1|\theta_j) \\ &= \frac{1}{1+\exp(-1.7a_i(\theta_j-b_i))} \end{aligned} \tag{2.1}$$

ここで，$a_i \in [0,\infty), b_i \in (-\infty,\infty)$ はそれぞれ i 番目の項目の識別力パラメータ，難易度パ

ラメータと呼ばれる項目パラメータである．

IRT では項目 i において，式(2.1)を用いて計算したフィッシャー情報量を項目情報量 $I_i(\theta)$ (Item Information)と呼び，以下のように表す．

$$I_i(\theta) = 1.7^2 a_i^2 p_i(\theta)(1 - p_i(\theta)) \tag{2.2}$$

また，テストに含まれる項目の項目情報量の総和をテスト情報量と呼び，以下のように表す．

$$I(\theta) = \sum_{i \in T} I_i(\theta) \tag{2.3}$$

ここで，T はテストに含まれる項目の集合である．このテスト情報量の逆数が受検者能力推定値の漸近分散に収束する（Lord and Novick, 1968）．Samejima（1977)は，このテスト情報量が等価なテストを弱並行テストとして定義した．ただし，この時点の弱並行テストはテスト項目数やカテゴリ区分などの制約を一切課していない．多くの自動テスト構成手法（Boekkooi-Timminga, 1990; Armstrong et al., 1994, 1998; van der Linden and Adema, 1998; van der Linden, 2005; Songmuang and Ueno, 2011; Ishii et al., 2013, 2014; Ishii and Ueno, 2017; Fuchimoto et al., 2022）がこの弱並行テストの定義に基づいている．

全ての受検者の能力値 $\theta_j \in (-\infty, \infty)$ について，テスト情報量が等価なテストを生成するのは困難である．そのため，実際にはテスト情報量における受検者の能力値 θ_k を $\theta_k = \{\theta_1, \theta_2, \ldots, \theta_K\}$ のように幾つかの点でサンプリングし，離散的に扱う．例えば，van der Linden（2005)は混合整数計画問題を用いて，サンプリングした各点 θ_k ごとにテスト情報量の目標値を設定し，その目標値との誤差の和が最小となるテスト情報量を持つテストを逐次的に構成している．しかし，この手法は誤差が小さいテストから逐次的に構成するため，テスト数が増加するに従って，受検者の予測測定誤差が大きくなる問題がある．

この問題を解決するために，Ishii et al.（2014)は θ_k におけるテスト情報量の上限・下限制約（UB(θ_k)，LB(θ_k))を設定し，全ての制約を満たすテストを受検者得点の予測測定誤差が等質であるとした．これにより，受検者の予測測定誤差の許容範囲を事前に設定できるため，テスト数が増加するに従って，受検者の予測測定誤差が大きくなる問題が解消される．図 3 は，表 1 に示した並行テストのテスト情報量への上限下限の例である．図中の #1〜#4 は構成テストの情報量関数である．#1, #2 は共に制約を満たしており等質である．一方で，#3, #4 は制約を満たしておらず等質でない．

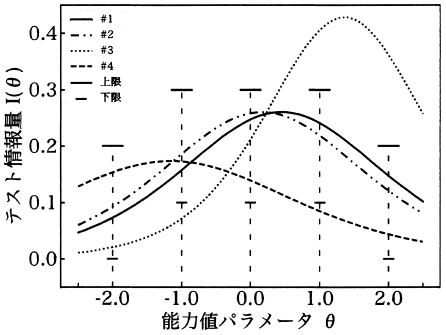


図 3. テスト情報量への上限下限の例.

表 1. テスト情報量制約の例.

テスト情報量 $I(\theta)$（下限値/上限値）				
$\theta = -2.0$	$\theta = -1.0$	$\theta = 0.0$	$\theta = 1.0$	$\theta = 2.0$
0.0/0.2	0.1/0.3	0.1/0.3	0.1/0.3	0.0/0.2

3. 自動テスト構成アルゴリズム

本章では，現在最も多くのテストを生成可能な最大クリーク法を用いた自動テスト構成手法を紹介する．

3.1 最大クリーク法を用いた自動テスト構成

Ishii et al. (2013)はテスト構成をグラフ上で定義される最大クリーク問題に帰着することで，厳密に最大数のテストを構成する手法を提案した．ここで，クリークとは任意の 2 頂点が隣接している部分グラフである．具体的には，無向グラフ $G = (V, E)$ を頂点の有限集合 V と辺の集合を E としたとき，最大クリーク法は次のように定式化できる．

$$\textbf{variables} \quad C \subseteq V \tag{3.1}$$

$$\textbf{maximize} \quad |C| \tag{3.2}$$

$$\textbf{subject to} \quad \forall v, \forall w \in C, \{v, w\} \in E \tag{3.3}$$

$\{v, w\} \in E$ は頂点 v, w に辺が引かれていることを意味する．

最大クリーク法によるテスト構成は図 4 のように，テスト候補を以下のグラフ構造とみなし，その中から最大クリーク探索を行うことで，テスト構成する．

（頂点） 与えられたアイテムバンクからテストの構成条件を満たす全てのテストを頂点とする．

（辺） 2 つのテスト候補の共通する項目数が一定値以下（以降，OC と呼ぶ）の場合，その 2 つの頂点（テスト）間に辺を引く．

このグラフの任意の頂点はテスト構成条件を満たしている．さらに，クリーク中の任意の 2 頂点は隣接しており，OC を満たす．したがって，このクリーク中の頂点に対応するテストはそれぞれ等質であり，その中でも最大クリークは理論的に最大数を保証したテスト群となる．

最大クリーク法のアルゴリズムを以下に示す．ただし，詳細なアルゴリズムについては Ishii et al. (2013, 2014)を参照されたい．

(1) アイテムバンクの項目の全ての組合せから，OC 以外の条件を満たすテストの組合せを探索木（Ishii et al., 2013, 2014）を用いて全て列挙する．
(2) (1) で列挙したテストをそれぞれ頂点とみなし，2 つのテストの共通する項目数が OC 以下の場合，その頂点間に辺を引く．
(3) (1)，(2) で生成されたグラフから最大クリーク探索（例えば，Tomita et al., 2016 や Li et al., 2017）を行う．

この手法は厳密に最大数のテストを構成できるアルゴリズムであるが，時間，空間計算量が $O(2^{0.19171|V|})$，$O(|V|^2)$ と計算コストが高い（Nakanishi and Tomita, 2008）．特に，自動テスト構成ではグラフの頂点の総数 $|V|$ はアイテムバンクの項目数に対して，組合せ爆発的に増加する．そのため，実際のアイテムバンクは数千項目から構成されるため，最大クリーク探索を厳密に行うことが困難である．

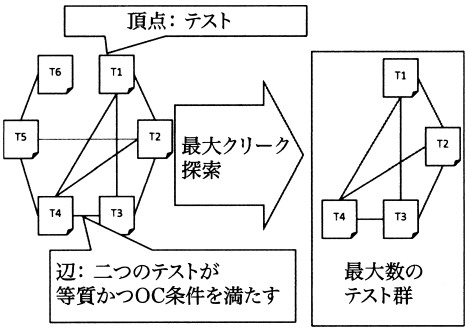


図 4. 最大クリーク法の概要図.

これらの計算コストを緩和するため，Ishii et al.（2014）は最大クリーク探索を行う近似アルゴリズム Random Maximum Clique Problem method（RndMCP 法）を提案した．最大クリーク法の問題点は，テスト構成数が増加するとグラフの探索空間が莫大となることである．そのため，RndMCP 法ではテスト候補グラフ全体から部分グラフをランダムに抽出し，このグラフから最大クリーク探索を繰り返す．これにより，グラフ全体の最大クリークを近似的に探索する．RndMCP 法は従来手法（van der Linden, 2005; Sun et al., 2008; Songmuang and Ueno, 2011）と比較して，10〜1000 倍以上多くのテストを構成できた．

3.2 整数計画法を用いた最大クリークアルゴリズム

RndMCP 法は空間計算量が大きく，10 万程度のテスト構成が上限であった．RndMCP 法の空間計算量を緩和するために，Fuchimoto et al.（2022）は整数計画法を用いた二段階並列探索手法である Hybrid Maximum Clique Algorithm Using Parallel Integer Programming method（HMCAPIP 法）を提案した．HMCAPIP 法の第一段階目は RndMCP 法でメモリが許す限り多くのテストを生成する．HMCAPIP 法の第二段階目は空間計算量が小さいが，時間計算量が大きい整数計画法による探索に切り替え，さらに多くのテストを生成する．具体的には，第一段階目に求めたクリークの全頂点と接続する頂点を以下の整数計画法により求める．

$$\textbf{maximize} \quad \sum_{i=1}^{n} \lambda_i x_i, \tag{3.4}$$

$$\textbf{subject to} \quad \sum_{i=1}^{n} = L, \tag{3.5}$$

$$\sum_{i=1}^{n} X_{i,t} x_i \leq \mathrm{OC} \qquad (t = 1, 2, \ldots, |C|), \tag{3.6}$$

$$LB_{\theta_k} \leq I(\theta_k) \leq UB_{\theta_k} \qquad (k = 1, 2, \ldots, K). \tag{3.7}$$

ここで，x_i は項目 i を選択する場合に 1，選択しない場合に 0 を示す決定変数，$X_{i,t}$ はクリーク中の t 番目のテストに項目 i が含まれる場合に 1, 含まれない場合に 0 を示す定数，L はテストの長さを表す定数である．また，λ_i は互いに独立な $[0, 1]$ の連続一様分布からの乱数であり，整数計画法を解く度にリサンプリングする．この乱数により，実行可能解の中からランダムにテストを生成できる．

さらに，Fuchimoto et al.（2022）は整数計画法の計算時間を緩和するために，頂点探索を並列化した．これにより，同一計算時間内において，HMCAPIP 法は最大 7 日間で当時の最新研

究の約 2.7 倍にあたる約 27 万のテスト生成可能としている（Fuchimoto et al., 2022）．

4. 提案手法

現在，HMCAPIP 法は世界で最も多くのテストを生成できる．しかし，HMCAPIP 法はテスト間に項目の重複を許すため各項目の露出数に偏りが生じる．この露出数の偏りはさまざまな弊害を生じる．最も深刻な問題は，露出数の大きい項目が受検者間で共有され受検対策されてしまい，その項目の信頼性を低下させることである（Wainer, 2000）．また，作問にかかるコストは多大であり，露出の低い項目は可能な限り出題することが望ましい．そのため，露出数は可能な限り一様であることが理想的である．

HMCAPIP 法においても，整数計画法の目的関数の決定変数に乱数の重み付け λ_i を与えることで，露出数の偏りを軽減している．しかし，各項目の特性(識別力パラメータや難易度パラメータ)に依存して，実行可能解中に既に露出数の偏りが生じており，λ_i による改善には限界がある．

この問題を緩和するために，本研究では既に生成したクリークの露出数を所与としたロジスティック関数によるペナルティを HMCAPIP 法における整数計画問題の目的関数に追加する．具体的には，以下の 2 種類のペナルティ，(1)ロジスティック関数による決定論的ペナルティ，(2)ロジスティック関数による確率論的ペナルティを提案し，露出数の標準偏差と最大露出率を従来手法よりも減少させる．

はじめに，現時点で生成したテスト群 U における，項目 i の露出数 IE_i を以下のように定義する．

$$IE_i = \sum_{t=1}^{|U|} X_{i,t}. \tag{4.1}$$

露出数 IE_i はテスト構成数に応じて増加する．そのため，提案手法ではテスト構成数に依らず露出数を共通尺度上で扱えるように，標準化露出数 z_i を以下のように定義する．

$$z_i = \frac{IE_i - IE_\mu}{IE_\sigma}. \tag{4.2}$$

ただし，IE_μ と IE_σ はそれぞれ全項目の露出数の平均値と標準偏差を示す．標準化露出数 z_i が標準正規分布に従うと仮定すると，その累積分布関数は次のロジスティック項目露出関数 $f(z_i)$ で近似できる．

$$f(z_i) = \frac{1}{1 + \exp^{-z_i}}. \tag{4.3}$$

ロジスティック項目露出関数 $f(z_i)$ は [0,1] の範囲における露出数の度合いとして解釈できる．また，ロジスティック項目露出関数 $f(z_i)$ は決定変数の重み λ_i の範囲に対応する．この性質を用いて，提案手法では HMCAPIP 法の目的関数に対する決定論的・確率論的ペナルティを 2 つ提案する．

4.1 決定論的ペナルティ

ロジスティック項目露出関数を $[0, 1]$ の度合いとして解釈すると，乱数の重み λ_i の値域と対応する．そのため，決定論的ペナルティではロジスティック項目露出関数を単純な決定変数のペナルティの重みとして扱うことを考える．具体的には以下の目的関数を最適化し，テストを生成する．本研究ではこの手法を提案手法(Det)と呼ぶ．

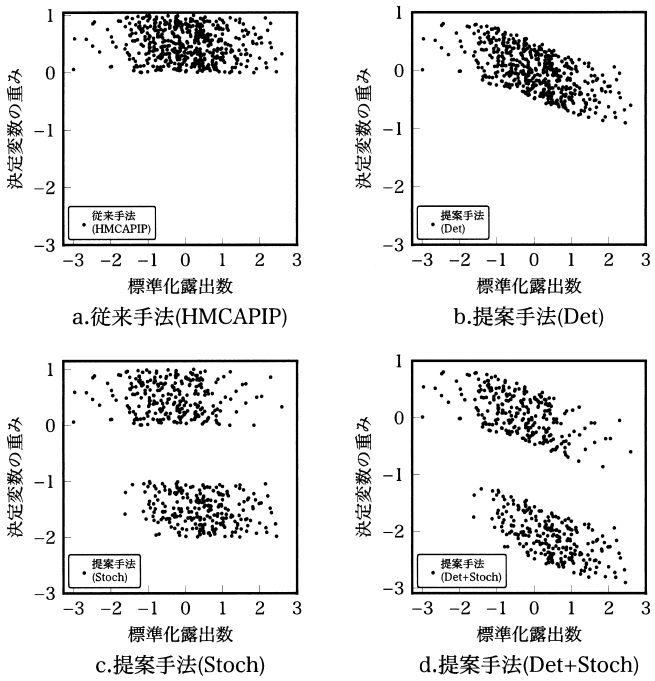


図 5. 標準化露出数と決定変数の重みの散布図.

$$\textbf{maximize} \qquad \sum_{i=1}^{n} \left(\lambda_i - \frac{1}{1+\exp^{-z_i}} \right) x_i. \tag{4.4}$$

数式(4.4)の意味を説明するために，図 5(a)及び図 5(b)に従来手法と提案手法(Det)における目的関数の決定変数の重み λ_i と $\lambda_i - \frac{1}{1+\exp^{-z_i}}$ をそれぞれ取り出し 500 個プロットした．ただし，標準化露出数は標準正規分布からの乱数，λ_i は互いに独立な $[0,1]$ の連続一様分布からの乱数としてデータを発生させた．図 5(a)より，従来手法では露出数を全く考慮しておらず偏りが生じる．一方で，図 5(b)より，提案手法(Det)では目的関数の決定変数の重みはロジスティック関数に従い，各項目の標準化露出数が小さい(大きい)ほど大きく(小さく)なる．これにより，この目的関数の最適解は現時点で露出数の高い項目の選択を抑制し，露出数の低い項目の選択を促進できる．したがって，提案手法(Det)では露出数の標準偏差を抑制したテスト構成が期待できる．

4.2 確率論的ペナルティ

提案手法(Det)では露出数の高い項目についても乱数の値によって，決定変数の重みが大きくなる．そのため，提案手法(Det)では露出数の高い項目に対するペナルティが小さい．本章では露出数の高い項目に対して，数理計画法の Big-M 法 (Williams, 1990)に基づいた重いペナルティを与えることを考える．Big-M 法は決定変数の重みの上界よりも大きなペナルティ M を与えることで，不必要な変数選択の個数を抑制できる定式化として知られている．例えば，提案手法において j 番目の決定変数 x_j の変数選択を可能な限り避けたいとする．このとき，以下の目的関数の値は x_j を含む実行可能解より含まない実行可能解の方が必ず大きな値を取る．したがって，決定変数 x_j を含む実行可能解は含まない実行可能解が存在しない場合を除き選

択されないことが保証される．

$$\textbf{maximize} \qquad \sum_{i=1}^{n} \lambda_i x_i - M x_j \qquad \left(\sum_{i=1}^{n} \lambda_i x_i < M \right). \tag{4.5}$$

この Big-M 法を用いて，確率論的ペナルティではロジスティック項目露出関数を確率とみなし，この確率に従ってペナルティ M を与える．具体的には以下の目的関数を最適化し，テストを生成する．本研究ではこの手法を提案手法(Stoch)と呼ぶ．

$$\textbf{maximize} \qquad \sum_{i=1}^{n} (\lambda_i - M_{i,p}) x_i, \tag{4.6}$$

$$M_{i,p} = \begin{cases} M & \eta_i < \frac{1}{1+\exp^{-z_i}}, \\ 0 & \text{otherwise.} \end{cases}$$

ここで，η_i は互いに独立な $[0,1]$ の連続一様分布からの乱数であり，整数計画法を解く度にリサンプリングする．

数式(4.6)の意味を説明するために，図 5(a)及び図 5(b)と同様の手順で，図 5(c)に従来手法と提案手法(Det)における目的関数の決定変数の重み $\lambda_i - M_{i,p}$ を取り出し 500 個プロットした．ただし，ペナルティの値は λ_i の最大値よりも大きい $M = 2$ とした．図 5(c)より，提案手法(Stoch)の決定変数の重みはペナルティが与えられない場合と与えられる場合でそれぞれ決定変数の値域が $[0,1]$ と $[0-M, 1-M]$ の 2 群に分離される．また，ペナルティ M が与えられる項目はロジスティック項目露出関数に従うため，標準化露出数が相対的に大きな(小さな)項目ほど，ペナルティが与えられる(与えられない)確率が高いことがわかる．結果として，この目的関数の最適解は極端に露出数の高い項目(最大露出数)の選択を避けることができる．一方で，提案手法(Stoch)では標準化露出数が極端に大きい項目のみを抑制するので，提案手法(Det)ほど露出数の標準偏差を抑えられない可能性がある．

4.3 決定論的・確率論的ペナルティ

提案手法(Det)及び提案手法(Stoch)はそれぞれ露出数の標準偏差と最大露出率を抑える効果が期待できる．これらの手法は独立して決定変数の重みにペナルティを与えられるため，単純な提案手法(Det)と提案手法(Stoch)の組合せにより，露出数の標準偏差と最大露出率を両方抑えることが可能かもしれない．具体的には以下の目的関数を最適化し，テストを生成する．この手法を提案手法(Det+Stoch)と呼ぶ．

$$\textbf{maximize} \qquad \sum_{i=1}^{n} \left(\lambda_i - \frac{1}{1+\exp^{-z_i}} - M_{i,p} \right) x_i, \tag{4.7}$$

$$M_{i,p} = \begin{cases} M & \eta_i < \frac{1}{1+\exp^{-z_i}}, \\ 0 & \text{otherwise.} \end{cases}$$

数式(4.7)の意味を説明するために，図 5(a)～(c)と同様の手順で，図 5(d)に提案手法(Det+Stoch)における目的関数の決定変数の重み $\lambda_i - \frac{1}{1+\exp^{-z_i}} - M_{i,p}$ を取り出し 500 個プロットした．提案手法(Det+Stoch)では各項目の露出数が小さい(大きい)ほど，大きく(小さく)なる．これにより，この目的関数の最適解は現時点で露出数の高い項目の選択を抑制し，露出数の低い項目の選択を促進できる．さらに，提案手法(Det+Stoch)ではペナルティ M を与えられると目的関数の決定変数の重みが極端に小さな値となり，標準化露出数の相対的に大きな項目ほど，ペナルティが与えられることがわかる．これにより，この目的関数の最適解は露出数

の高い項目ほど不必要な選択を避けることができる．

5. 評価実験

提案手法が従来手法よりもテスト構成数を減少させず，露出数の偏りを減少させることをシミュレーション実験により示す．具体的には従来手法(MCALIE 法 (Ishii and Ueno, 2015)及び HMCAPIP 法 (Fuchimoto et al., 2022))とテスト構成数，最大露出率及び露出数の標準偏差を比較した．ここで，最大露出率及び露出数の標準偏差は生成したテスト群 U における，項目 i の露出数 IE_i(式(4.1))と全項目の露出数 IE_i の平均値 IE_μ を用いて以下の通り計算した．

$$\text{最大露出率} = \frac{\max\{IE_1, IE_2, \ldots, IE_n\}}{|U|} \tag{5.1}$$

$$\text{露出数の標準偏差} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(IE_i - IE_\mu)^2} \tag{5.2}$$

また，各手法の計算時間は最大 24 時間とし，MCALIE 法と HMCAPIP 法のパラメータはそれぞれ，Ishii and Ueno (2015)と Fuchimoto et al. (2022)が実施した条件と同様である．アイテムバンクにはシミュレーション及び Synthetic Personality Inventory (SPI)試験 (Recruit, 2023) で実際に用いられていたデータを用いた．シミュレーションアイテムバンクは 2000 の項目を持ち，各項目の識別力パラメータを $a \sim U(0,1)$，難易度パラメータを $b \sim N(0,1^2)$ として発生させた．実データアイテムバンクの詳細は表 2 の通りである．

本研究では，これらのアイテムバンクから表 3 のテスト情報量制約を満たす 25 項目 ($L = 25$) のテスト構成を行った．また，OC の値は 1〜10 と 1 つづつ変化させて評価実験を行った．これらの条件は Fuchimoto et al. (2022)の評価実験と同様であり，実際に運用されている e-Testing の規模におけるテスト構成を模倣している．

結果を表 4 に示す．はじめに，提案手法は最大露出率および露出数の標準偏差を従来手法よりも抑えることができた．また，一部の条件において，従来手法(MCALIE)が露出数の標準偏差を抑えたが，他の手法に比べてテスト構成数が小さく，受検者数が 10 万人を超えるような試験では実用的でない．提案手法ごとに分析すると，提案手法(Det)は露出数の標準偏差を小さ

表 2. 実アイテムバンクの詳細.

アイテムバンクサイズ	識別力パラメータ a			難易度パラメータ b		
	範囲	平均	標準偏差	範囲	平均	標準偏差
87	0.15 ∼ 0.67	0.35	0.134	−2.09 ∼ 4.55	0.73	1.625
93	0.19 ∼ 0.69	0.43	0.122	−3.92 ∼ 3.61	−0.79	1.196
104	0.13 ∼ 1.10	0.59	0.213	−0.18 ∼ 4.55	1.50	1.188
141	0.24 ∼ 1.09	0.64	0.155	−1.41 ∼ 3.91	0.60	0.855
158	0.15 ∼ 3.08	0.44	0.255	−4.00 ∼ 4.00	−1.12	1.434
175	0.12 ∼ 0.93	0.39	0.139	−2.93 ∼ 3.12	−0.25	1.113
220	0.16 ∼ 0.92	0.46	0.155	−4.00 ∼ 2.82	−1.28	1.098
Total: 978	0.12 ∼ 3.08	0.46	0.198	−4.00 ∼ 4.55	−0.22	1.572

表 3. テスト情報量制約.

テスト情報量 $I(\theta)$ (下限値/上限値)				
$\theta = -2.0$	$\theta = -1.0$	$\theta = 0.0$	$\theta = 1.0$	$\theta = 2.0$
2.0/2.4	3.2/3.6	3.2/3.6	3.2/3.6	2.0/2.4

表 4. 従来手法と提案手法のテスト構成数および露出数.

アイテムバンクサイズ	OC	従来手法 (MCALIE) テスト構成数	最大露出率	露出数の標準偏差	従来手法 (HMCAPIP) テスト構成数	最大露出率	露出数の標準偏差	提案手法 (Det) テスト構成数	最大露出率	露出数の標準偏差	提案手法 (Stoch) テスト構成数	最大露出率	露出数の標準偏差	提案手法 (Det+Stoch) テスト構成数	最大露出率	露出数の標準偏差
2000 シミュレーション	1	1117	2.1	**4.3**	1169	2.1	4.5	1221	2.0	4.4	**1271**	2.0	4.6	1210	**1.9**	4.4
	2	7602	2.0	**27.6**	10354	2.0	37.7	10335	**1.7**	**34.5**	**10655**	1.9	37.7	10519	**1.7**	35.6
	3	27884	2.0	**104.6**	45313	2.0	168.0	38306	**1.6**	**125.1**	**52032**	1.7	184.5	43174	**1.6**	142.2
	4	62154	3.4	**276.7**	96166	3.0	401.1	97426	**1.6**	317.9	103962	**1.6**	374.6	**105152**	**1.6**	345.9
	5	90894	3.9	423.0	121260	3.5	537.7	120480	**1.6**	**393.1**	131656	**1.6**	476.0	**130896**	**1.6**	430.4
	6	94014	3.8	438.4	118804	3.6	533.8	**132223**	**1.6**	431.4	128816	**1.6**	465.8	117392	**1.6**	**386.1**
	7	93972	3.9	437.8	118007	3.7	530.2	120674	**1.6**	**393.7**	**131675**	**1.6**	476.0	122334	**1.6**	402.3
	8	93978	3.9	438.1	118967	3.7	534.5	123197	**1.6**	**402.0**	**129231**	**1.6**	469.2	122474	**1.6**	402.7
	9	93906	3.9	437.4	121893	3.6	545.9	124855	**1.6**	**407.4**	**128231**	**1.6**	463.5	119888	**1.6**	**394.3**
	10	94071	3.9	437.8	120728	3.7	539.7	123211	**1.6**	402.0	**132118**	**1.6**	477.7	121358	**1.6**	**399.1**
978 実データ	1	272	5.2	1.5	287	4.9	1.3	246	4.5	1.3	295	**4.4**	**1.2**	**301**	4.7	**1.2**
	2	1456	5.2	5.9	1415	5.2	5.8	1510	5.2	4.9	1511	4.7	5.0	**1574**	**4.6**	**4.6**
	3	7577	4.4	31.1	8856	5.4	34.7	8893	5.6	18.2	**9278**	3.9	25.7	8816	**3.8**	**16.7**
	4	27897	4.8	133.4	**32152**	4.8	153.8	26323	6.2	43.5	31254	3.3	81.8	24922	**3.0**	**35.6**
	5	51244	9.6	364.0	68395	8.7	452.0	65864	6.5	**109.8**	**73226**	3.1	198.1	62712	**3.0**	**86.4**
	6	85273	14.8	775.6	94341	13.9	809.4	100802	8.2	201.0	**102101**	3.1	280.3	93857	**3.0**	**124.2**
	7	93889	15.6	889.8	105266	14.8	945.4	100266	8.2	202.0	**108423**	**3.1**	298.6	99555	**3.1**	**128.2**
	8	94439	15.7	900.3	106071	14.8	952.9	102036	8.3	208.0	**106080**	**3.1**	295.4	99900	**3.1**	**128.1**
	9	94477	15.5	899.6	**106677**	14.7	968.7	105865	8.3	217.3	106538	**3.1**	293.3	100024	**3.1**	**128.4**
	10	94472	15.8	902.8	106195	14.7	950.9	105448	8.3	216.3	**111786**	**3.1**	307.7	100981	**3.1**	**129.8**

くできていることがわかる．これは，提案手法(Det)がロジスティック項目露出関数を用いた決定論的な重みを与えることで，露出数の低い項目ほど選ばれやすく，露出数の高い項目ほど選ばれにくくなるためである．一方で，提案手法(Stoch)は最大露出率を抑制できている．この理由は，提案手法(Stoch)が露出数の極端に高い項目に Big-M 法に基づいた重みを与えるので，その時点で露出数の高い項目を選択しないことにある．さらに，提案手法(Det+Stoch)は提案手法(Det)と提案手法(Stoch)の長所を持ち，最大露出率および露出数の標準偏差が最も抑えられる傾向にあった．

また，驚くべきことに，露出数を制御しているにも関わらず提案手法は従来手法(HMCAPIP)と同等以上のテストを構成できた．この理由は，露出数を制御しながらテスト構成する提案手法の方が従来手法よりも局所解(極大クリーク)に陥りにくく，多くのテストを構成できたと推測される．特に，従来手法では露出率の高い項目が存在することで，重複項目数の条件を満たす項目の組合せがテスト数を増加させるほど少なくなってしまい，局所解に陥るのかもしれない．実際に，最大露出率を抑えるためのペナルティだけを持つ提案手法(Stoch)が最もテスト構成数を増加させる傾向があった．そのため，提案手法(Stoch)が決定論的ペナルティを持つ 2 つの提案手法と比較して，最大露出率を抑えつつランダム性を持っており，テスト構成数を増加させていると推測できる．

次に，各手法の露出数の分布を分析するために，図 6 は(アイテムバンクサイズ，重複項目数条件)=(978,10)の条件における各手法の露出数のヒストグラムを表したものである．図 6(a)より，従来手法(HMCAPIP)は露出数の偏りが大きく，4500 回以上出題される項目が多数存在する．一方で，図 6(d)より，提案手法(Det+Stoch)では露出数の偏りがほとんどない．ただし，図 6(b)より，提案手法(Det)は 4500 回以上出題された項目が存在し，図 6(c)より，提案手法(Stoch)はほとんど出題されない項目が存在する．そのため，いずれかのペナルティだけでは，露出数の偏り改善に限界があった．

提案手法(Det)と提案手法(Stoch)で一部の項目に偏りが生じた原因は各項目のパラメータの特性値によるものと考えられる．特に，識別力パラメータの値が高い項目は情報量が高く，実行可能解に含まれる割合が大きいと推測される．そこで，(アイテムバンクサイズ，重複項目

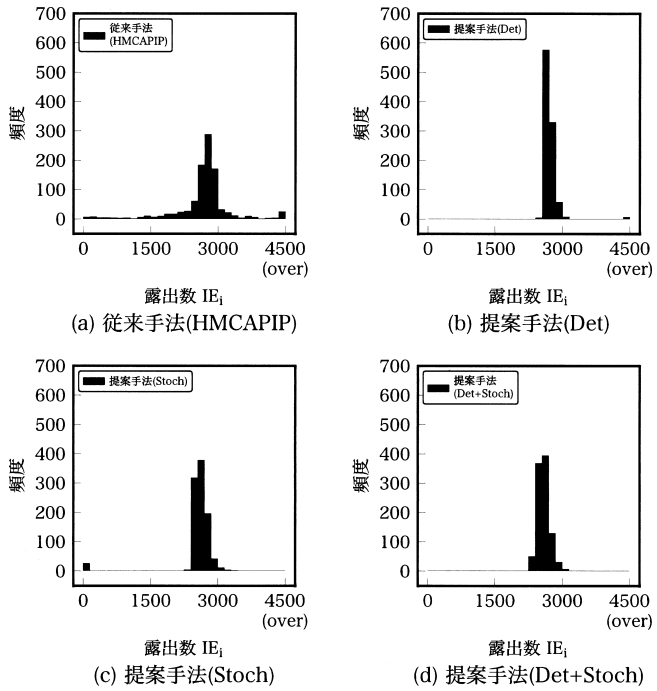


図 6. 露出数のヒストグラム(実データアイテムバンク，OC=10)．

数条件)=(978,10)の条件について，各手法における各項目の識別力パラメータとその項目の露出数を図 7 にプロットした．

図 7(a)より，従来手法は識別力パラメータの値が大きいほど，露出数の偏りが大きい．これは識別力パラメータの値が大きいほど情報量が高く，実行可能解に含まれる割合が大きいと考えられる．露出の高い項目は受検対策される恐れがあり，信頼性の低下につながる（Wainer, 2000)．特に，識別力の高い項目は作問が難しく貴重なため，過度な出題を避ける必要がある．

次に，提案手法ごとに分析すると，図 7(b)より，提案手法(Det)は 8000 回以上出題された識別力の高い項目が一つだけ存在する．この項目は過度な出題を避ける必要がある．また，図 7(c)より，提案手法(Stoch)はほとんど出題されない項目が複数存在する．これらの項目には識別力の高いものも含まれており，可能な限り活用することが望ましい．一方で，図 7(d)より，提案手法(Det+Stoch)は一様分布に近く理想的な露出数の分布にできた．

最後に，表 4 より，提案手法(Det)はシミュレーションアイテムバンクに限り，提案手法(Stoch)や提案手法(Det+Stoch)と同等に最大露出率を抑えられた．そこで，二つのアイテムバンクにおけるデータ特性との関連を分析するために，図 8 へ識別力パラメータのヒストグラムをプロットした．シミュレーションアイテムバンクは一様分布からデータを発生させたため，識別力の高い項目から低い項目まで均一に存在する．一方で，実データアイテムバンクは識別力の高い項目が少ない．したがって，提案手法(Det)は実データのように識別力パラメータの高い項目が不足している場合に最大露出率を抑制できないと考えられる．より各手法の特徴を分析するために，今後は項目パラメータの分布や他のテスト条件を変えた場合に，各手法が露出数の分布に与える影響を調査する．

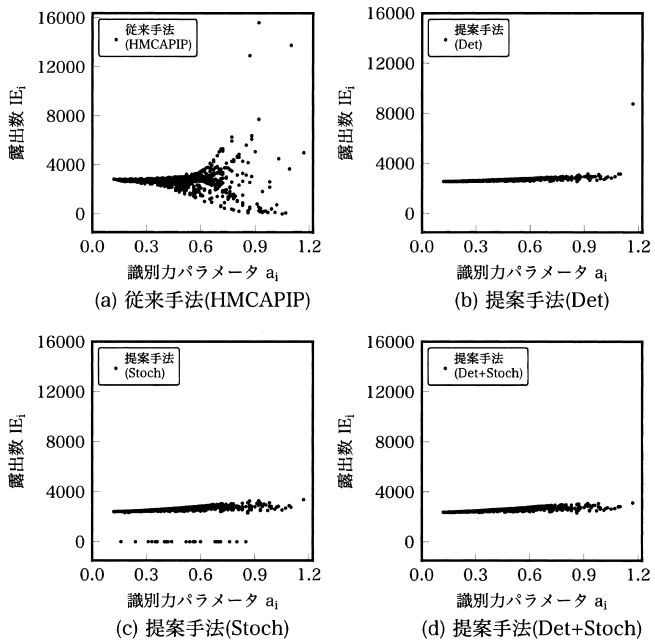


図 7. 識別力パラメータ a_i と露出数の散布図(実データアイテムバンク，OC=10)．

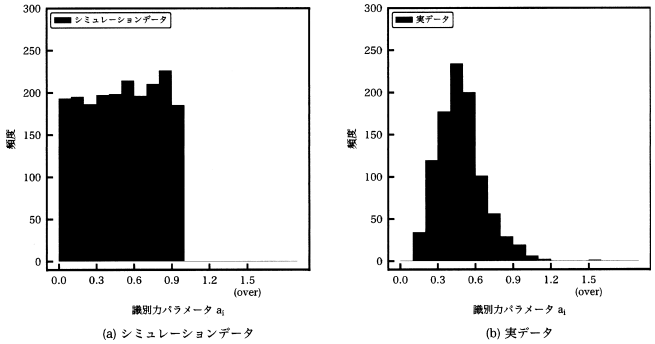


図 8. 識別力パラメータ a_i のヒストグラム．

6. むすび

本研究では，HCMAPIP 法における整数計画法の目的関数に露出数を所与としたロジスティック関数による 2 つのペナルティ項を追加した．提案手法ではこのロジスティック関数による決定論的・確率論的ペナルティ項を提案した．シミュレーション及び実データを用いた実験により，提案手法はテスト数を減らすことなく露出数の偏りを減らすことができた．これにより，e-Testing の信頼性向上が期待できる．

近年，測定精度を減少させずに問題数を減少できる技術として適応型テストが知られている(例えば，van der Linden, 2017; Kishida et al., 2023)．ここで，適応型テストとは受検者の能力を逐次的に推定しながら，その能力に応じて測定精度が最も高い項目を出題するテスト形式の一つである．しかし，同一能力の受検者には同一の項目群が出題される傾向にあり，特定の項

目群が頻繁に露出する傾向がある．また，各受検者の等質性が保証されていない．

この問題を解決する手法として，最大クリーク法による自動並行テスト構成技術（Fuchimoto et al., 2022）を用いた，二段階等質適応型テストが提案されている（Kishida et al., 2023）．この手法は従来の適応型と同等の測定精度を持ったまま問題数を減少でき，等質であるが露出数の分布が一様となるように出題できる．本研究の提案手法を二段階等質適応型テストに用いることで，露出数をより一様にできる可能性がある．

最後に，欧米では日本の共通テストに匹敵する Scholastic Aptitude Test（SAT）が全面的に e-Testing に移行する．また，我が国では 2024 年に電気通信大学において e-Testing による入試が開始される予定である（植野, 2023）．しかし，入学試験などで用いられる思考力を問う項目は所要時間が長く，十分な項目数を出題できずに測定精度が著しく低下する問題がある．この問題を解決するために，今後は各項目の所要時間を考慮し，制限時間内で等質であるが露出数の分布が一様となる自動並行テスト構成手法や二段階等質適応型テストへの拡張を検討する．これにより，e-Testing が制限時間内に各受検者の思考力を評価できることを目指す．

謝　辞

本研究の一部は科学研究費補助金（基盤研究（S），代表：植野真臣）「信頼性向上を持続する e テスティング・プラットフォームの開発」（19H05663）の助成を受けた．

参 考 文 献

Armstrong, R. D., Jones, D. H. and Wang, Z. (1994). Automated parallel test construction using classical test theory, *Journal of Educational Statistics*, **19**(1), 73–90.

Armstrong, R. D., Jones, D. H. and Kunce, C. S. (1998). IRT test assembly using network-flow programming, *Applied Psychological Measurement*, **22**(3), 237–247.

Boekkooi-Timminga, E. (1990). The construction of parallel tests from IRT-based item banks, *Journal of Educational Statistics*, **15**(2), 129–145.

独立行政法人情報処理推進機構 (2023). 【IT パスポート試験】情報処理推進機構, https://www3.jitec.ipa.go.jp/JitesCbt/index.html（最終アクセス日 2023 年 12 月 27 日）.

Fuchimoto, K., Ishii, T. and Ueno, M. (2022). Hybrid maximum clique algorithm using parallel integer programming for uniform test assembly, *IEEE Transactions on Learning Technologies*, **15**(2), 252–264.

Hetter, R. D. and Sympson, J. B. (1997). Item exposure control in CAT-ASVAB, *Computerized Adaptive Testing*: *From Inquiry to Operation* (eds. W. A. Sands, B. K. Waters and J. R. McBride), 141–144, American Psychological Association, Washington, D.C.

Ishii, T. and Ueno, M. (2015). Clique algorithm to minimize item exposure for uniform test forms assembly, *International Conference on Artificial Intelligence in Education*, 638–641, Springer, Switzerland.

Ishii, T. and Ueno, M. (2017). Algorithm for uniform test assembly using a maximum clique problem and integer programming, *International Conference on Artificial Intelligence in Education*, 102–112, Springer, Switzerland.

Ishii, T., Songmuang, P. and Ueno, M. (2013). Maximum clique algorithm for uniform test forms, *The 16th International Conference on Artificial Intelligence in Education*, 451–462.

Ishii, T., Songmuang, P. and Ueno, M. (2014). Maximum clique algorithm and its approximation for uniform test form assembly, *IEEE Transactions on Learning Technologies*, **7**(1), 83–95.

Kishida, W., Fuchimoto, K., Miyazawa, Y. and Ueno, M. (2023). Item difficulty constrained uniform adaptive testing, *International Conference on Artificial Intelligence in Education*, 568–573, Springer, Switzerland.

公益社団法人医療系大学間共用試験実施評価機構 (2023). 2023 年(令和 5 年)度 第 21 版 共用試験ガイドブック, https://www.cato.or.jp/e-book/21/index.html (最終アクセス日 2023 年 12 月 27 日).

Li, C.-M., Jiang, H. and Manyà, F. (2017). On minimization of the number of branches in branch-and-bound algorithms for the maximum clique problem, *Computers & Operations Research*, **84**, 1–15.

Lord, F. and Novick, M. (1968). *Statistical Theories of Mental Test Scores*, Addison-Wesley, Boston.

Nakanishi, H. and Tomita, E. (2008). An $O(2^{0.19171n})$-time and polynomial-space algorithm for finding a maximum clique, *IPSJ SIG Technical Report*, **2008**(6), 15–22.

Recruit (2023). Synthetic Personality Inventory (SPI), https://www.spi.recruit.co.jp/ (最終アクセス日 2023 年 12 月 27 日).

Samejima, F. (1977). Weakly parallel tests in latent trait theory with some criticisms of classical test theory, *Psychometrika*, **42**(2), 193–198.

Songmuang, P. and Ueno, M. (2011). Bees algorithm for construction of multiple test forms in e-Testing, *IEEE Transactions on Learning Technologies*, **4**, 209–221.

Stocking, M. L. and Lewis, C. (1998). Controlling item exposure conditional on ability in computerized adaptive testing, *Journal of Educational and Behavioral Statistics*, **23**(1), 57–75.

Stocking, M. L. and Lewis, C. (2000). Methods of controlling the exposure of items in CAT, *Computerized Adaptive Testing*: *Theory and Practice*, 163–182, Springer, New York City.

Sun, K.-T., Chen, Y.-J., Tsai, S.-Y. and Cheng, C.-F. (2008). Creating IRT-based parallel test forms using the genetic algorithm method, *Applied Measurement in Education*, **2**(21), 141–161.

Tomita, E., Yoshida, K., Hatta, T., Nagao, A., Ito, H. and Wakatsuki, M. (2016). A much faster branch-and-bound algorithm for finding a maximum clique, *International Workshop on Frontiers in Algorithmics*, 215–226, Springer, Cham.

Ueno, M. (2021). Ai based e-testing as a common yardstick for measuring human abilities, *2021 18th International Joint Conference on Computer Science and Software Engineering* (*JCSSE*), 1–5.

植野真臣 (2023). CBT の最前線, 情報処理, **64**(5), e1–e6.

Ueno, M., Fuchimoto, K. and Tsutsumi, E. (2021). E-testing from artificial intelligence approach, *Behaviormetrika*, **48**(2), 409–424.

van der Linden, W. J. (2005). *Linear Models for Optimal Test Design*, Springer, New York City.

van der Linden, W. J. (2017). *Handbook of Item Response Theory*: *Volume 3*: *Applications*, CRC Press, Florida.

van der Linden, W. J. and Adema, J. J. (1998). Simultaneous assembly of multiple test forms, *Journal of Educational Measurement*, **35**(3), 185–198.

van der Linden, W. J. and Choi, S. W. (2020). Improving item-exposure control in adaptive testing, *Journal of Educational Measurement*, **57**(3), 405–422.

van der Linden, W. J. and Reese, L. M. (1998). A model for optimal constrained adaptive testing, *Applied Psychological Measurement*, **22**(3), 259–270.

van der Linden, W. J. and Veldkamp, B. P. (2004). Constraining item exposure in computerized adaptive testing with shadow tests, *Journal of Educational and Behavioral Statistics*, **29**(3), 273–291.

Wainer, H. (2000). *CATS*: *Whither and Whence*, Educational Testing Service, New Jersey.

Williams, H. P. (1990). *Model Building in Mathematical Programming*, Wiley, New York City.

Automated Parallel Test Assembly Using Integer Programming with Item Exposure Penalties

Kazuma Fuchimoto and Maomi Ueno

Graduate School of Informatics and Engineering, The University of Electro-Communications

One feature of e-Testing is the automated test assembly of parallel test forms, by which each form has equivalent measurement accuracy, but with a different set of items. Unfortunately, the automated test assembly often causes a bias of item exposure. This difficulty of bias decreases the item and test reliability. To resolve this difficulty, this study examines a formulated test assembly problem as the objective function of integer programming with two logistic item exposure penalties: a deterministic penalty of logistic item exposure; and a stochastic penalty with logistic item exposure based on the Big-M method, which is a standard technique in mathematical programming. Numerical experiments demonstrate that the proposed methods reduce the bias of item exposure without decreasing the number of tests.

Key words: Automated test assembly, item response theory, e-Testing, item exposure problem, integer programming.

統計数理(2024)
第 72 巻 第 1 号 61–78

特集「心理統計学の新展開」
[原著論文]

心理尺度の統計的共通化：等化とリンキングの方法と実践

光永 悠彦†

(受付 2023 年 6 月 30 日；改訂 2024 年 2 月 13 日；採択 2 月 19 日)

要　　旨

人間の能力の一側面を測る目的で行われるテストは，能力の指標となる値を「得点」として得るための仕組みである．異なる回に行われるテストで共通の意味をもつ得点を返すための仕組みとして「等化」や「リンキング」と呼ばれる方法が提案されてきた．「等化」は複数のテストがそれぞれ同一の構成概念を測っていることがわかっている場合に，テスト得点の尺度をそれらの間で共通化する操作を指すが，「リンキング」は一次元性といった制約が少ない場合の共通尺度化手法を総称する概念である．本稿ではまず複数のテスト版による共通尺度化の手法についてその概略を述べる．そのうえで，継続して公平なテストを実施し続けることができるようなテスト計画に，等化の手法がどのように用いられるかについて，実践例を挙げながら説明する．またこれらの説明に際し，具体的な等化の手続きを可視化するためのブロックダイヤグラムについて触れ，従来紹介される機会が少なかった大規模調査の等化の実践について，本稿では学力調査の例を用いて説明する．

キーワード：テスト理論，項目反応理論，大規模テスト，教育測定．

1. はじめに

1.1 テスト得点の尺度化と項目反応理論(IRT)

人間のもつ能力の一側面，たとえば「語学能力」といった要素を測定するために，その要素を測定するための多数の問題(以下「項目」と表記する)を測定対象となる者(以下「受検者」と表記する)に出題し，正答・非正答といった反応を収集したうえで，測定のための尺度を構成し，「得点」として表示することが行われる．このような特定の能力を測定するための仕組みは，一般的に「テスト」と呼ばれている．テストで測られる要素は，心理学的な「構成概念」であるとされ，テストを実施する目的に応じて構築される．

同じ，あるいは類似した構成概念を測定するテストが複数回行われ，異なる内容のテスト版(項目をひとまとめにしたもの)を出題しつつ，それらの得点を比較可能とするような共通尺度上に乗せることが，多くのテストで行われている．全国学力・学習状況調査の「経年変化分析調査」もその一つであるが，共通尺度化の手法は公開されている(文部科学省, 2022)．この仕組みを実現するためには，項目ごとに，受検者の能力分布によらない困難度，すなわち標本依存性のない困難度(標本依存性については山田, 2014 を参照)を見いだす必要がある．標本依存性のない指標を見いだすための理論的枠組みの一つが項目反応理論(item response theory, IRT)

† 名古屋大学 大学院教育発達科学研究科：〒 464–8601 名古屋市千種区不老町

である(詳細は後述する).

1.2 共通尺度化の必要性とその方法

しかしながら,複数の受検機会で得られた異なるテストの結果を IRT により別々に尺度化した結果は,そのままでは共通の尺度上で表されているとはいえないため,何らかの手続きにより共通の尺度に変換する必要がある.

Holland and Dorans (2006)は,得点を比較可能とするための手続きをリンキング(linking)と総称している.リンキングは,その目的に応じて(a)予測(predicting),(b)尺度調整(scale aligning),(c)テストの等化(test equating)の 3 種類に分けられている.2 種類のテスト X と Y において,(a)は X の得点を用いて Y の得点を予測する方法であり,(b)は X と Y の間で「得点の換算表」を作成するための変換法を指す.また(c)は「相互に得点が変換可能(interexchangeable)」となるような変換法を指す.(b)と(c)はいずれも,複数のテストの間で共通の意味をもつ得点の尺度を得る変換であることから,「共通尺度化」のための変換法であるといえるが,(b)では(c)のように厳密な意味で相互に変換可能な共通尺度ではなく,あくまで「比較が可能」であるという場合の変換を含むことに注意が必要である.一方で(c)の場合は「等化」と呼ばれ,(b)に比べてより厳しい制約が課されている.具体的には「測定される構成概念が同一であること」「相互に信頼性が等しいこと」「等化の対称性(4.2.1 項で述べる)があること」「能力が同一の受検者において,素点の条件付き分布が等しいこと」「母集団に依存しないこと」といった「等化の前提」が満たされる必要がある(川端, 2014; 光永, 2017).

複数のテストにおいて「測定される概念が同一で,かつ信頼性や困難度の分布が同一であるような関係」が見られる場合,それらのテストは「平行テスト(parallel test)」であると呼ばれる.等化される尺度が平行テストを構成するとみなすには,等化の前提のうち「信頼性が同等」であることと「素点の条件付き分布が等しい」ことが必要である.等化の操作によって,調整対象となる尺度が平行テストに近くなるが,等化前の尺度間で信頼性や困難度が著しく異なる場合には,等化後の尺度を平行テストの関係に近づけることが困難となり,等化の精度にも影響する(川端, 2014).そのため,等化前の尺度においても信頼性や困難度がなるべく同等であることが,等化の前提の一つとなる.また「母集団に依存しないこと」とは,異なる母集団(たとえば性別の違いなど)ごとに等化を行って尺度間の得点換算表を作成した場合,得点の対応関係が集団間で一致することを指す(川端, 2014).

学力調査の実践でしばしば行われる,複数の異なる学年間で共通の尺度を得る操作は,測定される構成概念が学年の間で完全に同一であることは考えにくいため,等化ではなく尺度調整の範疇に入る.しかし,構成概念が学年間で相互に類似しているテスト版を用いて,IRT を応用した等化の手法を援用して共通尺度化が行われている.このような学年間の共通尺度化は「垂直尺度化(vertical scaling)」と呼ばれ,さまざまな手法やテストデザイン(Carlson, 2011; Kolen and Brennan, 2014, pp. 431–435)が提案されている.それに対して,学年の違いのような能力差を仮定せず,テスト版で構成される尺度が等化の前提を満たしていると想定されている等化は「水平等化」と呼ばれている.

1.3 本論の目的

以上の背景を踏まえ,本稿では最初に,リンキングのなかで(b)に相当する「尺度調整」の手法について述べる.次に,IRT による尺度化がなされた能力値や困難度指標の性質について述べ,IRT に基づく等化を行う理論的枠組みについて説明する.

また本稿では,等化の手順を図示する「ブロックダイヤグラム」による表現(光永, 2022)を紹介し,等化の手続きを可視化することにより,大規模テストにおける等化の手続きをわかりや

すく説明する方法について述べる．

2. 等パーセンタイル法による尺度調整

実際のテストでは調整対象となるテスト版が必ずしも同一の構成概念を測定しているとはいえない場面も多く存在する．たとえばある大学の入試において，「理科」という教科の下に「物理」「化学」「生物」「地学」の 4 科目のテストを受検者が選択でき，どの科目の得点も「理科」という教科の得点と見なして合否を決定する，というような場合である．このような場合，完全な意味で得点が等しい意味をもっているとはみなされないが，教科「理科」における素養の程度を共通尺度上で序列化する，あるいは大まかな形で「理科」の素養について検討するという目的においては，尺度調整の手法を用いることで，変換後の得点のもつ意味が限定された共通尺度化を行うことが可能である．本章ではそのような「尺度調整」の方法について述べる．

複数のテストの得点尺度を共通尺度に乗せる方法として，同じパーセンタイルにあたる受検者は同じ能力をもっているという前提で尺度調整する方法がある．この手法は「等パーセンタイル法」と呼ばれている．一方で線形変換による尺度調整を考えることもできる．

ある 100 点満点のテストを 100 人からなる集団に提示して，素点を得たとする．このとき 80 点をとった受検者が最高点であったとするなら，80 点より小さな素点をとった受検者が全体の 100 パーセントであると表現できる．同様に 75 点をとった受検者についても，75 点より小さな素点であった受検者が全体の 95 パーセントである，というように表現できる．素点データを小さな順に並べた際，小さなほうから数えて全体の $100p$ パーセントにあたる素点を「$100p$ パーセンタイル」と呼ぶ．また，素点データからすべての受検者についてパーセンタイルを求めたものを「累積相対度数分布」と定義する．

集団 A と集団 B がランダム等価グループである場合（4.1 節を参照）で，それぞれが異なる内容の 100 点満点のテスト A 及びテスト B を受検している場合を考える．集団 A を基準にして，集団 A の得点 x_A を確率変数と考え，この尺度上で集団 B の得点 x_B（確率変数）を表すような等パーセンタイル法の尺度調整は以下のように行われる．集団 A，集団 B における素点の累積相対度数分布をそれぞれ $F_A(\cdot)$，$F_B(\cdot)$ とおく．また，集団 B の累積相対度数を与えた場合に，パーセンタイルを返す関数（すなわち $F_B(\cdot)$ の逆変換）を $F_B^{-1}(\cdot)$ と表すと，等パーセンタイル法による変換の結果得られる「集団 A の尺度上で表現された集団 B の得点 $x_{B\text{-}A}$」は

$$x_{B\text{-}A} = F_B^{-1}(F_A(x_A)) \tag{2.1}$$

で求められる．

線形変換による方法は平均と標準偏差を基準に合わせる変換であったが，等パーセンタイル法による変換は基準となる素点の分布に他のテストの素点分布を「分布の形状が全て同じになるように」合わせる変換である（証明は前川, 2018 を参照）．実際のテストにおいては素点が離散的な値（100 点満点のテストでは 0 点から 100 点までの 101 通りの値）を取る場合が多いため，F_A や F_B について，連続的な値をとるようなスムージングを行う．この方法については高次関数を当てはめる方法や移動平均をとる方法などが提案されている（Kolen and Brennan, 2014; 前川, 1999）．

尺度調整は等化と比べて仮定が少ない手法であると述べたが，等パーセンタイル法は素点を用いた変換であり，同じパーセンタイルに位置する受検者が同じ能力をもつという別の強い仮定が必要である．素点を用いた尺度調整の手法とは別に，IRT による TCC（4.2.2 項を参照）の関係性を用いて尺度調整を行う手法が提案されており，IRT true score equating と呼ばれている（Lord and Wingersky, 1984）．共通尺度化を行う前提としてどのような仮定をおくかによっ

て，使用する等化・尺度調整の手法も適切に選択する必要がある．

3. IRT による尺度化と等化

3.1 IRT に基づく尺度化

受検者から得られた正答・非正答のデータを「反応データ」と呼ぶことにする．以下，反応データとして「正答」「非正答」のいずれかのカテゴリが観測され，正答カテゴリが「1」，非正答カテゴリが「0」で表される場合を考える．i 番目の受検者$(i = 1, 2, \ldots, N)$において j 番目の項目$(j = 1, 2, \ldots, J)$に対する反応を u_{ij} と表記する．受検者 i における正答数の合計 $u_i = \sum_{j=1}^{J} u_{ij}$ は受検者ごとの「素点」と呼ばれ，項目 j における平均正答数 $u_j = 1/N \sum_{i=1}^{N} u_{ij}$ は項目ごとの「通過率」と呼ばれるが，前述の通り，これらの指標は標本依存性があるため，そのままでは適切な解釈ができない．

IRT では，受検者の正答・非正答が一次元の潜在特性値 θ で説明できるという前提をおいたうえで，項目 j に正答する確率 $P_j(\theta)$ を θ の関数として表現したものを項目反応関数(item response function, IRF)と考える．u_{ij} が二値であるとき，IRF としては「1 パラメタ・ロジスティックモデル」，「2 パラメタ・ロジスティックモデル」又は「3 パラメタ・ロジスティックモデル」のいずれかが仮定される場合が多い(それぞれ 1PLM, 2PLM 及び 3PLM と略記する)．3PLM の式は

$$P_j(\theta) = c_j + (1 - c_j)\frac{1}{1 + \exp(-Da_j(\theta - b_j))} \tag{3.1}$$

であり，項目 j において，a_j は「識別力」，b_j は「困難度」，c_j は「下方漸近線」パラメタとそれぞれ呼ばれている(それぞれのパラメタの意味については山田, 2014 や光永, 2017 を参照)．これらを総称して「項目パラメタ」と呼ぶ．ただし $-\infty < \theta < \infty$，$a_j > 0$，$-\infty < b_j < \infty$，$0 \leq c_j \leq 1$ である．3PLM の IRF において，全ての項目について $c_j = 0$ とおいた場合が 2PLM，$c_j = 0$ かつ $a_j = a$(ただし $a > 0$)とおいた場合が 1PLM である．また D は「尺度要素」と呼ばれる定数で，$D = 1.702$ とおくことで，IRF のロジスティック曲線が正規累積曲線の良い近似となることが知られている．

項目パラメタの推定においては，尤度関数として

$$L(\boldsymbol{\xi}, \boldsymbol{\theta}|\boldsymbol{u}) = \prod_{i=1}^{N}\prod_{j=1}^{J} P_j(\theta_i)^{u_{ij}} Q_j(\theta_i)^{1-u_{ij}} \tag{3.2}$$

を考える(この方法は「同時最大尤度法」と呼ばれる)．ただし $\boldsymbol{\xi}$ は項目パラメタ全体を要素にもつベクトル(たとえば 3PLM の場合は $\boldsymbol{\xi} = (\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})$ となる．ただし $\boldsymbol{a} = (a_1, a_2, \ldots, a_J)'$，$\boldsymbol{b} = (b_1, b_2, \ldots, b_J)'$，$\boldsymbol{c} = (c_1, c_2, \ldots, c_J)'$ を表す)，$\boldsymbol{\theta}$ は θ_i をベクトル化したもの，$\boldsymbol{u}$ は u_{ij} を要素にもつ $N \times J$ の行列，$Q_j(\theta_i)$ は $1 - P_j(\theta_i)$，すなわち非正答確率を表す．ただし実用上は $\boldsymbol{\xi}$ と $\boldsymbol{\theta}$ を同時に推定するのではなく，$\boldsymbol{\xi}$ のみを推定したい場合が多い．たとえば 5.1 節で説明する「項目バンク」の構築にあたっては，項目ごとの $\boldsymbol{\xi}$ の推定結果を記録しておけば十分である．この場合は(3.2)式の尤度関数に代えて，$\boldsymbol{\theta}$ を周辺化した尤度関数を

$$L_m(\boldsymbol{\xi}|\boldsymbol{u}) = \prod_{i=1}^{N}\int_{-\infty}^{\infty} L(\boldsymbol{\xi}, \theta_i|\boldsymbol{u}) g(\theta_i|\boldsymbol{\tau}) d\theta_i \tag{3.3}$$

とおく．この方法は「周辺最大尤度法(marginal maximum likelihood method, MML 法)」(Bock and Lieberman, 1970)と呼ばれる．ここで $g(\theta_i|\boldsymbol{\tau})$ は θ に関する事前分布であり，$\boldsymbol{\tau}$ は事前分布のハイパーパラメタを表すが，多くの実用場面においては事前分布として標準正規分布を仮定

することが多い．MML では(3.3)式の尤度の最大化を目指す際，各受検者について尤度関数の期待値をとることにより，θ に関する項を消去できる．実際の数値計算においては L_m の対数をとったものを最大化し，積分を和に置き換える．また最適化手法としては EM アルゴリズム(Dempster et al., 1977)を用いる方法（Bock and Aitkin, 1981)が提案されている．

$\boldsymbol{\xi}$ がすでに推定されている場合は，その推定値 $\hat{\boldsymbol{\xi}}$ と受検者 i の各項目に対する反応データ $u_j = (u_{i1}, u_{i2}, \ldots, u_{iJ})$ から θ_i を推定できる．最尤法による推定の場合は

$$L_\theta(\boldsymbol{\theta}|\boldsymbol{u}) = \prod_{j=1}^{J} P_j(\theta_i)^{u_j} Q_j(\theta_i)^{1-u_j} \tag{3.4}$$

が尤度関数であるが，実際の計算では L_θ の対数をとったものを最大化する．また最尤法では素点の全問正答・非正答に対する θ が推定できないが，θ に事前分布を仮定した EAP(Expected A Posteriori)法を用いることで，回避できる（詳細は村木, 2011, pp. 84–85 を参照．ただし，適切な事前分布の指定が必要である)．

テストの実践においては，$\boldsymbol{u}$ に欠測がある場合が多く見られる．この場合は，多くの最尤推定の場合と同様に，$\boldsymbol{u}$ で観測されている部分のみを用いて尤度を求める手法が用いられる．後述する「同時推定」による等化の手法において，欠測値を含む $\boldsymbol{u}$ に対する項目パラメタ推定が役立つ．

3.2 多母集団を仮定した場合の尺度化

前節で説明した尤度関数の最大化によって，単一の事前分布を仮定した場合(すなわち，θ の母集団分布が一つであることを仮定した場合)に，単一のテスト版を受検した受検者から得られた u_{ij} から項目パラメタ $\boldsymbol{\xi}$ を推定できる．一方で，u_{ij} が複数の異なる能力値母集団分布をもつ受検者集団(グループ)から観測されており，受検者ごとに所属グループが与えられている場合に，複数の集団ごとに異なる θ の母集団分布を考慮した形で項目パラメタを推定する方法が提案されている．

このような多母集団 IRT モデルは，尤度関数を多母集団に拡張する方法（Mislevy, 1984; Bock and Zimowski, 1997; Bock and Gibbons, 2021, pp. 104–105)と，θ を従属変数とした潜在回帰モデルに基づく方法（Adams et al., 1997; Adams and Wu, 2007)が提案されている．以下，これらについて説明する．

3.2.1 IRT の尤度関数を多母集団に拡張する方法

この方法では，複数の母集団の違いを表現するために(3.3)の尤度関数を次のように書き換える．

$$L_{mm}(\boldsymbol{\xi}|\boldsymbol{u}_g) = \prod_{i=1}^{N} \int_{-\infty}^{\infty} L(\boldsymbol{\xi}, \theta_i|\boldsymbol{u}_g) g_k(\theta_i|\boldsymbol{\eta}) d\theta_i \tag{3.5}$$

ここで $g_k(\theta_i|\boldsymbol{\eta})$ は k 番目の集団 $(k = 1, 2, \ldots, K)$ における母集団分布の確率密度関数であり，$\boldsymbol{\eta}$ は複数の母集団分布におけるハイパーパラメタをまとめたものである．またデータとしては $\boldsymbol{u}$ に代えて $\boldsymbol{u}_g$ を用いる．これは集団 k に属する i 番目の受検者における j 番目の項目に対する反応 u_{ijk} を行列としたものを表す．

このモデルでは(3.5)式のように，「複数の母集団を特徴づける $\boldsymbol{\eta}$ が集団ごとに異なる値をとる」という仮定をモデル上で表現できる．たとえば集団 1 から集団 3 までの 3 つの質的に異なる集団に対して同一の内容のテスト版を出題したデータを用いた分析では，いずれの母集団も互いに異なる平均 μ_g と標準偏差 σ_g となる正規分布を仮定し，かつ集団 1 は標準正規分布と

なるように固定した場合，$\boldsymbol{\eta} = ((0,1)', (\mu_2, \sigma_2)', (\mu_3, \sigma_3)')$ となるような要素をもつ行列に含まれるパラメタを推定することとなる．ここで集団 1 の分布を固定したのは，集団 1 が規準集団(後述)であることを表現するためである．

3.2.2 潜在回帰モデルによる方法

この方法では，(3.3)式による項目パラメタの尤度の最大化にあたり，受検者の能力値 θ に関して下記のような構造をもつモデルを仮定する．

$$\boldsymbol{\theta} = \boldsymbol{Y}'\boldsymbol{\beta} + E, \ \ E \sim N(0, \sigma^2) \tag{3.6}$$

ここで $\boldsymbol{Y}$ は質的に異なる集団の違いを表す変数，$\boldsymbol{\beta}$ はこれに対応する回帰係数，E は残差であり，このモデルを用いた場合は $\boldsymbol{\xi}$ だけではなく $\boldsymbol{\beta}$ 及び σ も推定する．

Adams et al. (1997)は，この手法が IRT の尤度関数を多母集団に拡張する手法に比べて，(1)結果の解釈が回帰係数 $\boldsymbol{\beta}$ の形でできるため，テスト得点と外的基準(生徒の居住地域や社会経済的地位など)との関連を検討する場合に有利である，すなわち先に θ を推定し，次にそれを用いて回帰分析をするという二段階推定を避けることができる，(2) $\boldsymbol{Y}$ による構造化により $\boldsymbol{\xi}$ や $\boldsymbol{\theta}$ の推定精度が高まる（Mislevy, 1987)ことを指摘している．

4. 異なるテスト間における尺度の等化方法

ここからは，複数のテストが同一の構成概念を測定し，それぞれの信頼性が等しい場合において，それぞれのテストから IRT により項目パラメタを推定することで，複数のテスト間で共通の意味をもつ尺度を得るための手続きについて述べる．前述の通り，この手続きは「テストの等化(test equating)」と呼ばれているが，等化に先立ち，それぞれのテストにおいて一次元性が確認され，かつ測定している構成概念が同一である状況を考える．説明の都合上，等化の対象となるテスト版が 2 種類(テスト A とテスト B)あり，テスト A 受検者を基準としてテスト B 受検者の θ の尺度を等化する場面を考える．ここで尺度の基準となる集団を「規準集団(reference group, base group)」，規準集団に合わせられる集団を「焦点集団(focal group)」と呼ぶ．

4.1 等化に必要な共通要素

等化の対象となる 2 種類のテストは，何らかの共通要素を含むように計画して行われる．共通要素としては「共通項目デザイン」「共通受検者デザイン」の 2 通りがある．図 1 に，2 種類の

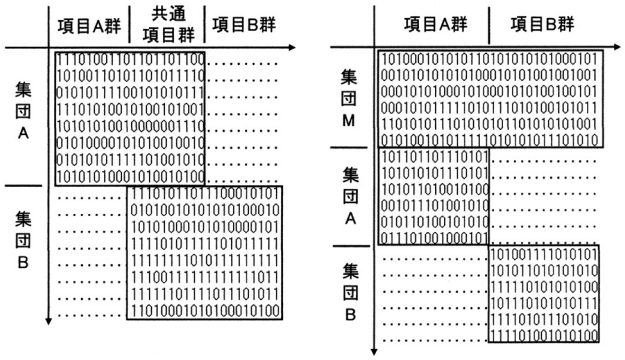


図 1. 共通項目デザインと共通受検者デザイン．

テストデザインの例を記した．共通受検者デザインでは「集団 M」を介して共通尺度が得られることとなるが，集団 M としては得点を返すことを目的とせず，等化の手掛かりを得るために，テスト実施者が募集した集団が充てられる場合がある．このような受検者を「モニター受検者」と呼ぶ．

上記のテストデザインにおいては，集団 A，集団 B 及び集団 M について，受検者の θ の分布が同一の母集団から得られたものと仮定できるか否かが問題となる．集団 A と集団 B の θ の分布が同一の母集団からのものであるとはいえない場合，これらの集団は別々の母集団からサンプリングされていると考え，IRT の分析においてはこれらの非等質性を考慮した分析が必要となる．Kolen and Brennan (2014) では図 1 左に示す共通項目デザインを共通項目不等価グループデザイン (common-item non-equivalent groups design) と呼んでいる．一方で，異なる集団に属する受検者が同一の母集団からランダムに抽出されていると見なせる場合はそれらの集団がランダム等価グループ (random equivalent groups) であると呼ぶ．たとえば単一の母集団に属するとみなせる 1 群から受検番号の偶奇によって 2 群を構成する場合などが，それに該当する．

IRT を用いた等化の方法として，個別のテスト版について項目パラメタを推定した後，パラメタ推定値を線形等化する方法があり，separate calibration と呼ばれている（Arai and Mayekawa, 2011; Lee and Ban, 2010; Karkee et al., 2003）が，テスト版ごとに項目パラメタを推定することから「個別推定」とも呼ばれる．またそれとは別に同時尺度調整法 (concurrent calibration) と呼ばれる方法もあるが，この手法は一度に複数のテスト版の項目パラメタを同時に推定することから「同時推定」とも呼ばれる．あわせて，「固定項目パラメタ法 (fixed common item parameters)」と呼ばれる方法が提案されている．以下，これらの手法について述べる．

4.2 IRT による等化 (1) 個別推定

異なるテスト版 X と Y があり，それぞれのテスト版を 2 集団が解答し，尺度が構成されている場面を考える．2 集団いずれも θ の平均は 0，標準偏差は 1 であると仮定すると，X と Y はそれぞれ固有の尺度上で項目パラメタが算出されていると解釈できる．ここでテスト版 X と Y を別の集団 Z に提示して解答させた場合，Z に含まれる任意の受検者 i の能力値を θ_i と表すなら，i がテスト版 X を受検した場合の能力値は θ_{iX}，テスト版 Y を受検した場合は θ_{iY} と表せるが，θ_{iX} と θ_{iY} を標準化した値は常に相等しくなる．すなわち，

$$\frac{\theta_{iX} - \mu_{\theta_X}}{\sigma_{\theta_X}} = \frac{\theta_{iY} - \mu_{\theta_Y}}{\sigma_{\theta_Y}} \tag{4.1}$$

の関係性が常に成り立つ．ただし，μ_{θ_X} 及び σ_{θ_X} は θ_{iX} の母平均及び母標準偏差，μ_{θ_Y} 及び σ_{θ_Y} は θ_{iY} の母平均及び母標準偏差を表す．

テスト版 Y を基準にして，テスト版 X の尺度をテスト版 Y の尺度上で表現する関係式を導くためには，(4.1) 式を θ_{iY} について解けばよい．すなわち，

$$\theta_{iY} = \frac{\sigma_{\theta_Y}}{\sigma_{\theta_X}}\theta_{iX} + \mu_{\theta_Y} - \mu_{\theta_X}\frac{\sigma_{\theta_Y}}{\sigma_{\theta_X}} \tag{4.2}$$

となり，θ_{iX} から θ_{iY} への変換式は

$$\theta_{\theta_{iY}} = K\theta_{iX} + L \tag{4.3}$$

$$K = \frac{\sigma_{\theta_Y}}{\sigma_{\theta_X}} \tag{4.4}$$

$$L = \mu_{\theta_Y} - K\mu_{\theta_X} \tag{4.5}$$

となる．観測されたデータから変換式を求めるには，μ と σ をそれぞれ実際の推定値の平均及び標準偏差に置き換える．

このように，2 集団の間での θ の変換式は線形変換の形で表すことができるが，変換のための係数 K 及び L は「等化係数」と呼ばれる．なお，Y の尺度値を従属変数，X の尺度値を独立変数とする単回帰分析における回帰係数（傾き）の最小二乗解は(4.4)式の分母を θ_X と θ_Y の共分散で置き換えた式であり，等化係数の式とは一致しないことに注意が必要である．等化係数を用いて，異なる 2 集団から得られた同一項目に対する項目パラメタを変換する式を求めるには，以下に述べる複数の方法が提案されている．代表的なものを以下に挙げるが，詳細な説明は川端 (2014)や服部 (2006)も参考にされたい．

4.2.1 Mean/Sigma 法及び Mean/Mean 法

Mean/Sigma 法 (Marco, 1977)では，(4.3)式の関係性が，θ と同様に項目困難度 b_j にも当てはまることを利用する．すなわち，2 集団の尺度 X と Y について，

$$K = \frac{\sigma_{b_Y}}{\sigma_{b_X}} \tag{4.6}$$

$$L = \mu_{b_Y} - K\mu_{b_X} \tag{4.7}$$

を用いて，

$$a^*_{jX} = a_{jX}/K \tag{4.8}$$

$$b^*_{jX} = Kb_{jX} + L \tag{4.9}$$

$$c^*_{jX} = c_{jX} \tag{4.10}$$

として，「尺度 Y 上で表された変換後の尺度 X の項目パラメタ」$(a^*_{jX}, b^*_{jX}, c^*_{jX})$ を求めることができる．ただし，μ_{b_X} 及び μ_{b_Y} は集団 X 及び集団 Y から得られた b_j の母平均，σ_{b_X} 及び σ_{b_Y} は集団 X 及び集団 Y から得られた b_j の母標準偏差を表し，等化係数の算出の際にはデータから得られた b_j の平均と標準偏差を用いる．

Mean/Sigma 法は，項目パラメタのうち b_j の情報のみを手がかりに等化係数を求める方法であるが，a_j も加味した方法が Mean/Mean 法 (Loyd and Hoover, 1980)である．(4.8)式及び(4.9)式において，左辺を尺度 Y から得られた共通項目におけるパラメタ値，右辺を尺度 X から得られた共通項目におけるパラメタ値とみなし，それぞれ平均をとると，下記の関係式が得られる．

$$M(\hat{a}_{jY}) = M(\hat{a}_{jX})/K \tag{4.11}$$

$$M(\hat{b}_{jY}) = KM(\hat{b}_{jX}) + L \tag{4.12}$$

ただし $M(\hat{a}_{jY})$ 及び $M(\hat{b}_{jY})$ は共通項目における尺度 Y から得られた a_j 及び b_j の推定値の平均値であり，$M(\hat{a}_{jY})$ 及び $M(\hat{b}_{jY})$ は同様に尺度 X から得られた a_j 及び b_j の推定値の平均値である．この関係式を K と L について解くと

$$K = \frac{M(\hat{a}_{jX})}{M(\hat{a}_{jY})} \tag{4.13}$$

$$L = M(\hat{b}_{jY}) - KM(\hat{b}_{jX}) \tag{4.14}$$

が得られる．これが Mean/Mean 法による等化係数である．ただし，K の推定にあたって，a_j の算術平均の代わりに幾何平均を用いた方法(Mean/Geometric Mean 法)も提案されている (Mislevy and Bock, 1990; 村木, 2011)．

ここまでは，尺度 Y を基準にして尺度 X の尺度を等化する際の等化係数 K, L について述べてきたが，同様の手順で尺度 X を基準にして尺度 Y の尺度を等化する場合の等化係数も求めることができる．そのようにして求めた等化係数を K' 及び L' とすると，

$$K' = 1/K \tag{4.15}$$

$$L' = -L/K \tag{4.16}$$

という関係性がある．等化係数の推定値がこれらの関係性を満たす場合，その等化係数の推定方法は「等化の対称性がある」と呼ばれる．Mean/Sigma 法及び Mean/Mean 法は，等化の対称性を満たすことが知られている．

Mean/Sigma 法や Mean/Mean 法は，いずれも複数のテスト版に共通して出現する項目の b_j は互いに等しいと考えて等化が行われるため，「困難度等化法」と呼ばれることがある．また項目パラメタの標準偏差の情報までを用いて等化係数を求めているため，「モーメント法」とも呼ばれている．

4.2.2 Haebara 法及び Stocking-Lord 法

困難度等化法とは異なるアプローチとして，IRF の形を等化前と等化後で一致させるように等化係数を求める方法が提案されている．主な手法として Haebara 法（Haebara, 1980）及び Stocking-Lord 法（Stocking and Lord, 1983）がある．IRF を合わせるこれらの手法は，モーメント法に比べて安定した推定結果を得られる傾向がある（Ogasawara, 2000）．

尺度 X について，等化係数 K, L を用いて変換した後の項目パラメタ $a^*_{jX}, b^*_{jX}, c^*_{jX}$ を用いて表現された IRF を $P^*_j(\theta_{iX})$，基準となる尺度 Y の IRF を $P_j(\theta_{iY})$ とおくと，もし完全に等しい尺度に等化された場合，ある能力値 θ_i をもつ受検者において，$P_j(\theta_{iY}) = P^*_j(\theta_{iY})$ が成り立つ．しかし，実際の等化場面ではこのような完全な変換を行うための等化係数を求めるのではなく，これに近似するような等化係数を求めることとなる．Haebara 法では，$P_j(\theta_{iY})$ と $P^*_j(\theta_{iY})$ のずれの大きさを

$$Q_H = \sum_{s=1}^{S}\sum_{j=1}^{J}[P_j(\theta_{iY}) - P^*_j(\theta_{iY})]^2 h(\theta_{iY}) + \sum_{s=1}^{S}\sum_{j=1}^{J}[P_j(\theta_{iX}) - P^{**}_j(\theta_{iX})]^2 h(\theta_{iX}) \tag{4.17}$$

と定義し，Q_H を最小化する K 及び L を求める．ここで(4.17)式の右辺第 2 項は等化の対称性を保証するために必要であり，$P^{**}_j(\theta_{iX})$ は尺度 X を基準にして尺度 Y を等化した際の逆変換の結果得られた IRF を表す．また $h(\theta_{iX})$ 及び $h(\theta_{iY})$ はそれぞれ尺度 X と尺度 Y における能力値 θ_i の確率密度関数であり，IRF と同様に S 個の求積点 $(s = 1, 2, \ldots, S)$ に等分されているものとする．

Stocking-Lord 法では，テスト特性曲線(test characteristic curve, TCC)の差の最小化を目指す．TCC はあるテスト版に含まれる項目すべてについて ICC を合計したもの，すなわち $\sum_{j=1}^{J} P_j(\theta)$ で定義される．これを用いて，ずれの大きさの指標を

$$Q_{SL} = \sum_{s=1}^{S}\left(\left[\sum_{j=1}^{J} P_j(\theta_{iY}) - \sum_{j=1}^{J} P^*_j(\theta_{iY})\right]^2 h(\theta_{iY}) + \left[\sum_{j=1}^{J} P_j(\theta_{iX}) - \sum_{j=1}^{J} P^{**}_j(\theta_{iX})\right]^2 h(\theta_{iX})\right) \tag{4.18}$$

と定義し，Q_{SL} を最小化する K と L を求める．Stocking and Lord（1983）では(4.18)式ではなく右辺第 1 項のみが示されており，等化の対称性が欠如しているが，Kim and Kolen（2003）では等化の対称性が保証されており，ここではその式を紹介した．

これまで説明した手法については，等化される基準となるテスト版が 1 つであるのに対し，等化により尺度が変換されるテスト版が 1 つだけの場合であったが，実際には複数のテスト版

を同時に等化したい場合がある．上記の Q_H や Q_{SL} を「基準関数」と考えた場合，複数のテスト版を同時に等化する基準関数の定義と，最小化のために必要な「Q の等化係数による 1 階・2 階微分」があれば，複数テスト版の等化が可能となるが，これらの導出はかなり複雑である．詳細は Battauz (2017) を参照のこと．また，Arai and Mayekawa (2011) は別個に複数のテスト版を同時に等化するための手法を提案しているが，交互最小二乗法を用いることで，複雑な微分の計算を不要にしている．

4.3 IRT による等化(2)同時推定

個別推定による方法では，個別のテスト版から求められた項目パラメタの推定値のみが，等化の手掛かりとなっており，複数のテスト版における正誤データの完全情報による推定となっていない．そのため，あるテスト版に割り当てられた集団の人数が，他の集団に比べて少ない場合において，その集団に提示された共通項目の等化の精度(K や L の推定値の標準誤差)が低下するおそれがある．

「同時推定」は，複数のテスト版に含まれる正誤データすべてを用いて，集団間で共通の能力尺度を仮定し，この上で項目パラメタを推定する方法である．すなわち，3.2 節で述べた手法を用いて，多母集団を仮定した IRT モデルにより集団間で共通の項目パラメタ及び母集団における能力値分布を各集団において推定する．

共通項目を含む 2 種のテスト版 X と Y を等化する場合においては，(1) テスト版 X と Y を受検した全員について，同じ項目への反応が同じ列になるように(図 1 左のように)データを整形する，(2) このデータに対して多母集団 IRT モデルを適用する，という方法をとる．もし集団 X を基準にして集団 Y の尺度を等化する場合は，集団 X の θ の平均を 0 と固定したモデルを適用すればよい．

4.4 IRT による等化(3)固定項目パラメタ法

図 1 左の状況において，集団 A を基準となる集団と定義し，集団 B の尺度を集団 A に等化する場合を考える．集団 A に出題した全ての項目群について項目パラメタを推定すると，基準となる集団の能力尺度上で共通項目群の項目パラメタが推定される．次に，集団 B に出題した全ての項目群について項目パラメタを推定するが，共通項目群の項目パラメタについては，集団 A から推定された値であると固定するような制約を入れたうえで，残りの項目群の項目パラメタを推定する．この方法により，集団 B において推定された項目パラメタは，集団 A を基準とした場合の値と解釈できる．この手法により等化する方法を「固定項目パラメタ法」と呼ぶ．

固定項目パラメタ法は，Kim and Kolen (2016) や Kim (2006) にその方法が解説されているが，これらで紹介されている方法は 3.2.1 項で説明した多母集団 IRT モデルとは異なる尤度関数の最大化を目指している．詳細は Woodruff and Hanson (1996) を参照のこと．

4.5 θ の尺度得点への変換

IRT によるテストは，受検者の能力値は -0.5 や 2.6 といったように $-\infty \leq \theta \leq +\infty$ の範囲で 0 を中心とした範囲の値で表示される．しかしこの値をそのまま受検者の能力を表す値として公表すると，能力値の概念になじみのない者にとって親切であるとは言えない．そのため，θ を適当な範囲の値に変換する．変換された値はテストの得点の一種であると考えることができるため「尺度得点」と呼ばれる．

θ から尺度得点への変換方法は，線形変換による方法と，非線形変換による方法がある．線形変換による方法は，変換後の尺度得点を X とおくと，

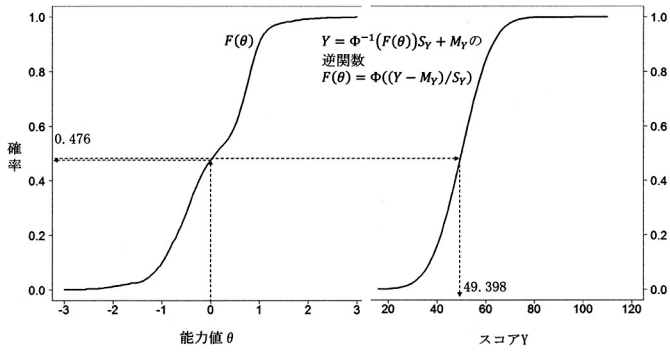


図 2. 正規分布していない θ を正規分布に従う尺度得点 Y に変換する方法（光永, 2017, p.137）.

$$X = \frac{\theta - M(\theta)}{S(\theta)} \times S_X + M_X \tag{4.19}$$

で求められる．ただし $M(\theta)$，$S(\theta)$ はそれぞれ θ の平均と標準偏差，M_X と S_X は変換後の X の平均と標準偏差であり，この変換によって X の平均が M_X，標準偏差 S_X となるような尺度得点が得られる．テスト実施者は解釈がしやすい尺度得点の範囲になるように，M_X と S_X を事前に定めておく必要がある．この手法においては，θ の分布が正規分布に近い場合は，変換後の分布も正規分布に近いものが得られ，尺度得点が「偏差値」のように解釈可能である．

一方で，θ の分布が正規分布しない場合も現実には多くありうる．その場合に θ から尺度得点 Y へ変換する方法としては，θ 及び Y を連続的な確率変数であると考え，(4.19)式に代わる変換の式を

$$Y = \Phi^{-1}(F(\theta))S_Y + M_Y \tag{4.20}$$

と定義する．ただし $F(\theta)$ は θ を引数にとる累積相対度数を返す関数を表し，Φ^{-1} は標準正規分布の累積分布の逆関数を表す．これにより，θ の尺度は平均 M_Y，標準偏差 S_Y の正規分布に従う Y の尺度に変換される．このような非線形変換の詳細は図 2 を参照のこと．図 2 の左は，正規分布であるとは限らない θ を $F(\theta)$ により累積相対度数化することを表し，右は対応するパーセンタイルを(4.20)式により Y に変換することを表している．この方法は等パーセンタイル法を応用している．

変換後の尺度得点の M_Y 及び S_Y を定めるのではなく，尺度得点の範囲について，たとえば下限 Y_L から上限 Y_U の値を取ると決められている場合は，θ の最小値を変換した結果が Y_L に，また θ の最大値を変換した結果が Y_U にそれぞれ変換されるように，M_Y 及び S_Y を定めればよい．たとえば $M_Y = (Y_U - Y_L)/2$ とおき，θ を変換した後の尺度得点の範囲が Y_L と Y_U の範囲に収まるような S_Y を探索的に定めることとなる．

5. 大規模テストにおける等化

5.1 大規模テストで求められる要件と項目バンク

年に複数回，複数の会場において実施されるテストは，複数種類のテスト版を用意したうえで，それらのどのテスト版を受検しても返される得点の意味が同じになるようなテストとして設計される．受検者はテストを複数回受検してもよいため，これら複数種類のテスト版に対し

て同一の受検者が解答することとなる．このようなテストにおいては公平性の確保や結果の妥当な解釈のために，(1)同一受検者に同一の項目を 2 度以上提示しない，(2)複数種類のテスト版は，項目パラメタの分布がなるべく同等となるようなものとする，といった要件が課される．

これらの仕様を満たすために，あらかじめ出題候補となる項目のリストを作成しておき，項目特性を推定しておいたうえで，出題履歴を記録する仕組みが求められる．このようなデータベースを「項目バンク(item bank)」と呼ぶ．

5.2 等化手続きのブロックダイヤグラムによる図示

大規模テストを実施するためのテストデザインを説明するために，手続きの流れを図で示すブロックダイヤグラムを用いると，データ処理の手続きを明確に記述できるため便利である．光永 (2022)では図 3 に示すアイコンを等化手続きの計算の操作に対応させ，ブロックダイヤグラムにより等化の手続きを可視化することを行っている．本稿もこの形式により大規模テストの等化手続きを説明する．

図 3 の(1)は受検者集団に対してテスト版を提示し，正誤データを得る手続きを表す．実際の分析においてはテスト版に含まれる共通項目の情報に基づいて，分析対象となる正誤データが抽出・加工されるが，正誤データを整形する手続きもこのアイコンで表される．(2)は正誤データから項目パラメタを推定する手続きであり，複数の集団から得られた複数のテスト版を出題して，集団で共通の項目パラメタを推定する操作(同時推定)を表す．(3)は項目パラメタが既知の項目群を集団に出題して得られた正誤データから，θ を推定する手続きを表す．(4)は 4.5 節で説明した θ を尺度得点に変換する手続きを表す．(5)は規準集団から得られた項目パラメタの尺度上に，他の集団から得られた項目パラメタを等化する手続きであり，等化係数を用いた個別推定の手続きを示す．(6)は項目パラメタ既知の項目群を記録しておくデータベースを「項目バンク」と定め，そこから項目を引用してテスト版を作成する操作を表す．

これらのアイコンを組み合わせて，図 1 左の共通項目デザインの場合に等化を行う手順を記したものを図 4 に示す．図 4 上と下を比べると，同時推定では等化係数の算出を行わずに等化を行えるが，すべてのデータを項目単位でまとめるステップ(図では点線の枠囲みで記した)が必要であることが示されている．

5.3 大規模テストにおけるテストデザイン例

図 5 は，ある教授学習法の効果を検討する目的で計画されたテストデザインの例である．生

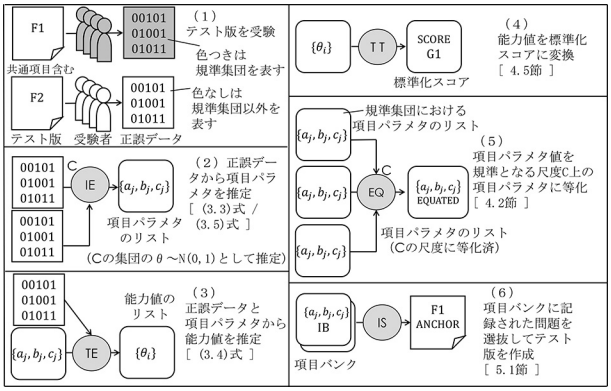


図 3. 等化にあたってデータ処理の流れを図示するためのアイコン（光永, 2022 を一部改変）.

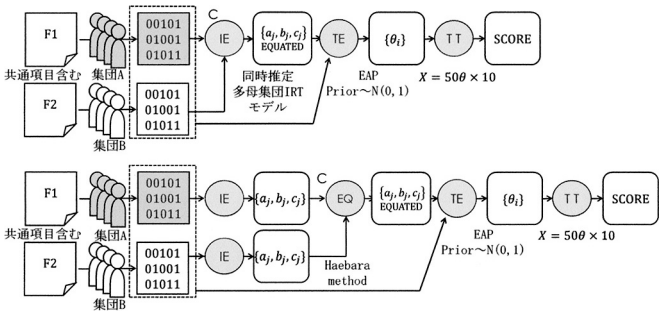


図 4. 共通項目デザインによる等化のブロックダイヤグラム．上は同時推定，下は個別推定の場合を表す．

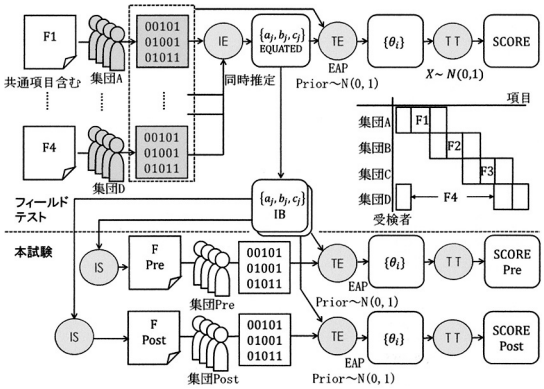


図 5. 事前テストと事後テストの間で比較可能な得点を得るテストのためのブロックダイヤグラム．フィールドテストの重複テスト分冊法による項目割り付け図とともに記した．

徒はある年度において，効果を検証したい学習法による授業を受けるが，それに先立ち年度当初に事前テスト(F Pre と表記している)を受検する．その後，一連の授業の履修が終わった年度末に事後テスト(F Post と表記している)を受検する．この 2 回のテストは，互いに同じ項目を含まない形で実施されるべきものであり，また互いに等質な項目パラメタをもつテスト版であることが求められる．

事前・事後テストを実施するのに先立ち，項目パラメタを明らかにするためのテストを実施する(受検者に成績を返す目的ではないテストを「フィールドテスト」と呼ぶ)．モニター受検者を 4 群に分け，図 5 右中のように 4 種類のテスト版を分冊の形で出題する(重複テスト分冊法)．これにより，受検者一人当たりに出題される項目数を減らすことができる．4 群をランダム等価(random equivalent)と考えて単一の θ の母集団を仮定して同時推定することで，これら 4 群全体を一つの規準集団とおくことができる．

本試験として事前・事後テストを実施するに先立ち，項目バンクから 2 種類のテスト版を作成して出題し，項目バンク上に記録された規準集団上での項目パラメタを用いて θ を算出する．最後に尺度得点へ変換すれば，それらは相互に同じ意味をもつようになる．

6. 大規模テストにおける等化の例

6.1 テストデザイン及び調査の実施

学習指導要領の範囲において教科「算数・数学」の学力を測るための項目を，小学 2 年生から中学 3 年生向けにそれぞれ作成し，一部に学年間の共通項目を含む形で各学年向けのテスト版を作成した．

テストデザインとしては，図 5 の点線より上で示したフィールドテストのブロックダイヤグラムと同様で，テスト版の数が 8 種類存在し，小 6 の受検者集団を規準集団とおいた．項目は全部で 100 項目あり，図 6 上に示すように，各テスト版に一つ下の学年向けの項目を共通項目として含むようにした．

受検者は小 2 が 10,212 名，小 3 が 10,616 名，小 4 が 10,217 名，小 5 が 10,855 名，小 6 が 11,088 名，中 1 が 8,691 名，中 2 が 8,520 名，中 3 が 8,438 名であった．

はじめに，学年ごとのテスト版それぞれについて，尺度の一次元性の仮定が満たされているかを検討した．項目間の四分相関係数(tetrachoric correlation coefficient)行列から固有値を求めたところ，すべての学年のテスト版で第 1 固有値が突出して高い傾向が見られたため，尺度の一次元性が高いと判断した．

次に個別推定と同時推定の両方を行い，結果を比較した．固定項目パラメタ法については，モデル上で固定すべきパラメタの指定が煩雑になることから，本稿では行っていない．本事例では θ の値と外的基準との関連を検討することがないため，同時推定においては，能力値の母集団が学年ごとに異なると仮定した多母集団 IRT モデルを用いて，小学 6 年生の集団を規準集団と考え，規準集団の能力値分布を平均 0，標準偏差 1 と固定してパラメタを推定した．また個別推定では本来，規準集団上の尺度に一つひとつのテスト尺度を等化していく手続きが必要であるが，調整対象となるテスト版の数が 7 種類と比較的多数であったため，Battauz (2017) で複数のテスト版を同時に等化するように拡張された Haebara 法により等化した．

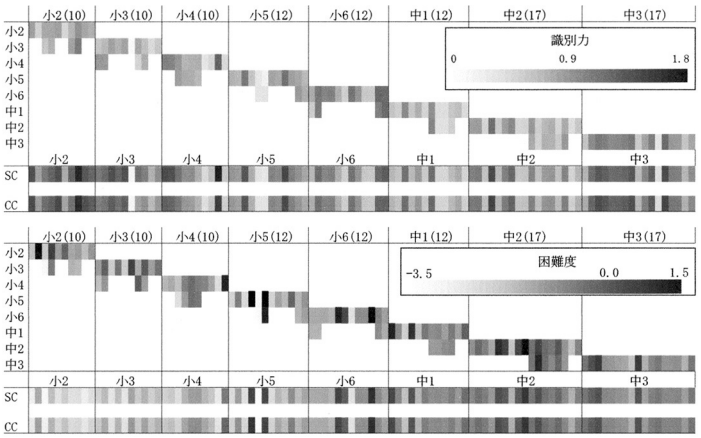


図 6. 上から，個別のテスト版の出題状況及び項目ごとに推定された等化前の識別力パラメタ，個別推定により等化した識別力パラメタ(SC)，同時推定により等化した識別力パラメタ(CC)，個別のテスト版から推定された困難度パラメタ，個別推定により等化した困難度パラメタ(SC)，同時推定により等化した困難度パラメタ(CC)．それぞれのセルの色は，識別力パラメタの値の大小を表している．( )内の数字は，当該学年向けに出題された項目数を示す．

6.2 項目パラメタの垂直尺度化結果

等化前と等化後の項目パラメタの値を色の濃さの形で記したものを図 6 に示した．学年ごとに別々に推定された項目パラメタは，共通項目について 2 通りの値が推定されるが，これらの共通項目のパラメタ値を手がかりに等化された結果，共通尺度上での値が求まるさまが見て取れる．

6.3 垂直尺度化における実践上の注意点

図 6 より，個別推定と同時推定で得られた等化後の項目パラメタ値は識別力・困難度とも，ほとんど同等の結果となった．本事例のように，尺度の一次元性が高い場合で，学年それぞれで測定される構成概念が類似している場合においては，等化方法間で結果に大きな差が見られなかったものの，等化の前提が十分満たされていない場合における等化や垂直尺度化にあたっては，結果に差が生じる可能性がある．個別推定と同時推定の比較については Arai and Mayekawa (2011)や Karkee et al. (2003)を参照のこと．また垂直尺度化を行うためのテストデザインの検討にあたっては，共通項目の数が問題となりやすいが，共通項目の識別力が小さくなったり，困難度が極端な範囲に偏ると等化が不安定になりやすい．垂直尺度化の結果が理論に近いものとなるためには，項目数の検討だけではなく，共通項目のパラメタ値にも留意が必要である．そのため，共通項目を選抜する目的でフィールドテストを実施することも行われている．

本事例では各学年において十分な児童生徒数が存在しており，また学年内における学力の範囲も幅広い傾向が見られた．しかし実践上，一部の学年において児童生徒の数が極端に少ない(たとえば数百名程度の)場合や尺度の一次元性が高いとはいえない場合は，個別推定を行うと，最初の段階で個別に推定された項目パラメタ値の推定精度が下がり，その値が等化に用いられるため，等化後の結果が理論通りにならない場合が予想される．同時推定では学年ごとに個別のパラメタ推定を行わないため，この問題が回避できる可能性があるが，受検者数が多数となった場合，パラメタ推定に多くの計算資源を要するという問題がある．実践の上ではこれらの問題点に注意しつつ，複数の等化方法を比較検討することがほとんどである．

7. おわりに

本稿では心理尺度を比較可能にする手続きとして「等化」を中心にその方法を述べ，実際の大規模テストや学力調査に応用する具体的方法を説明した．学力調査や入試，資格試験といったような場面で，多くの受検者において統一された尺度上で能力を表現することが求められるが，そのような仕組みを提供するためには等化の考え方による能力尺度の共通尺度化が不可欠である．

本稿で述べたブロックダイヤグラムによる記述は，過去にどのような等化を行ったかを記録する手段としても有用である．あわせて，ブロックダイヤグラムに書かれた等化手続きを自動化する仕組みも考えることができよう．作業手順の可視化のみならず，複数ある等化手法の比較検討を行いやすくすることで，大規模テストの開発がより効率的に進められることが期待できる．

謝　辞

分析例に用いたデータを提供していただいた，横浜市教育委員会教育課程推進室の皆様に感謝申し上げます．

参 考 文 献

Adams, R.J. and Wu, M.L. (2007). The mixed-coefficients multinomial logit model: A generalized form of the Rasch model, *Multivariate and Mixture Distribution Rasch Models*: *Extensions and Applications* (eds. M. von Davier and C.H. Carstensen), 55–76, Springer, New York, https://doi.org/10.1007/9780387498393_4.

Adams, R.J., Wilson, M. and Wu, M. (1997). Multilevel item response models: An approach to errors in variables regression, *Journal of Educational and Behavioral Statistics*, **22**(1), 47–76, https://doi.org/10.3102/10769986022001047.

Arai, S. and Mayekawa, S. (2011). A comparison of equating methods and linking designs for developing an item pool under item response theory, *Behaviormetrika*, **38**(1), 1–16, https://doi.org/10.2333/bhmk.38.1.

Battauz, M. (2017). Multiple equating of separate IRT calibrations, *Psychometrika*, **82**(3), 610–636, https://doi.org/10.1007/s11336-016-9517-x.

Bock, R.D. and Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm, *Psychometrika*, **46**(4), 443–459, https://doi.org/10.1007/BF02293801.

Bock, R.D. and Gibbons, R.D. (2021), *Item Response Theory*, Wiley, Hoboken, New Jersey.

Bock, R.D. and Lieberman, M. (1970). Fitting a response model for n dichotomously scored items, *Psychometrika*, **35**(2), 179–197, https://doi.org/10.1007/BF02291262.

Bock, R.D. and Zimowski, M.F. (1997). Multiple group IRT, *Handbook of Modern Item Response Theory* (eds. W.J. van der Linden and R.K. Hambleton), 433–448, Springer, New York.

Carlson, J.E. (2011). Statistical models for vertical scaling, *Statistical Models for Test Equating, Scaling and Linking* (ed. A.A. von Davier), 59–70, Springer, New York.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion), *Journal of the Royal Statistical Society, Series B*, **39**, 1–38.

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method, *Japanese Psychological Research*, **22**(3), 144–149, https://doi.org/10.4992/psycholres1954.22.144.

服部 環 (2006). 共通項目計画に基づくテストの等化, 筑波大学心理学研究, **31**, 19–29.

Holland, P.W. and Dorans, N.J. (2006). Linking and Equating, *Educational Measurement*, 4th ed. (ed. R. L. Brennan), 187–220, American Council on Education and Praeger Publishers, Westport, Connecticut.

Karkee, T., Lewis, D.M., Hoskens, M., Yao, L. and Haug, C. (2003). Separate versus concurrent calibration methods in vertical scaling, Paper presented at the Annual Meeting of the National Council on Measurement in Education (Chicago, Illinois), http://files.eric.ed.gov/fulltext/ED478167.pdf（最終閲覧日：2023 年 9 月 30 日）.

川端一光 (2014). 等化,『R による項目反応理論』(加藤健太郎, 山田剛史, 川端一光 著), 243–281, オーム社, 東京.

Kim, S. (2006). A comparative study of IRT fixed parameter calibration methods, *Journal of Educational Measurement*, **43**(4), 353–381, https://doi.org/10.1111/j.1745-3984.2006.00021.x.

Kim, S. and Kolen, M.J. (2003). *POLYST*: *A Computer Program for Polytomous IRT Scale Transformation Version 1.0*, The University of Iowa, Iowa City, Iowa.

Kim, S. and Kolen, M.J. (2016). Multiple group IRT fixed-parameter estimation for maintaining an established ability scale, Center for Advanced Studies in Measurement and Assessment (CASMA) Research Report, No. 49, 1–20.

Kolen, M.J. and Brennan, R.L. (2014), *Test Equating, Scaling and Linking*, 3rd ed., Springer, New York.

Lee, W. and Ban, J. (2010). A comparison of IRT linking procedures, *Applied Measurement in Education*, **23**, 23–48, https://doi.org/10.1080/08957340903423537.

Lord, F.M. and Wingersky, M.S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings", *Applied Psychological Measurement*, **8**, 453–461, https://doi.org/10.1177/014662168400800409.

Loyd, B.H. and Hoover, H.D. (1980). Vertical equating using the Rasch model, *Journal of Educational Measurement*, **17**(3), 179–193, https://doi.org/10.1111/j.1745-3984.1980.tb00825.x.

前川眞一 (1999). 得点調整の方法について,『大学入試データの解析』(柳井晴夫, 前川眞一 編), 88–109, 現代数学社, 京都.

前川眞一 (2018). テスト得点を同じ物差しにのせる—対応づけと QQ プロット—, 統計, **69**(8), 8–15.

Marco, G.L. (1977). Item characteristic curve solutions to three intractable testing problems, *Journal of Educational Measurement*, **14**(2), 139–160, https://doi.org/10.1111/j.1745-3984.1977.tb00033.x.

Mislevy, R.J. (1984). Estimating latent distributions, *Psychometrika*, **49**(3), 359–381, https://doi.org/10.1007/BF02306026.

Mislevy, R.J. (1987). Exploiting auxiliary information about examinees in the estimation of item parameters, *Applied Psychological Measurement*, **11**(1), 81–91, https://doi.org/10.1177/014662168701100106.

Mislevy, R.J. and Bock, R.D. (1990). *BILOG3. Item Analysis and Test Scoring with Binary Logistic Models*, 2nd ed., Scientific Software International, Mooresville, Indiana.

光永悠彦 (2017).『テストは何を測るのか　項目反応理論の考え方』, ナカニシヤ出版, 京都.

光永悠彦 (2022). IRT を用いた標準化テストを入試で活用する,『テストは何のためにあるのか　項目反応理論から入試制度を考える』(光永悠彦 編著), 154–226, ナカニシヤ出版, 京都.

文部科学省 (2022). 令和 3 年度『全国学力・学習状況調査』経年変化分析調査テクニカルレポート, 文部科学省, 東京, https://www.mext.go.jp/kaigisiryo/content/20220325-mxt_chousa02-000021553_5.pdf (最終閲覧日：2023 年 6 月 8 日).

村木英治 (2011).『項目反応理論』, シリーズ〈行動計量の科学〉8, 朝倉書店, 東京.

Ogasawara, H. (2000). Asymptotic standard errors of IRT equating coefficients using moments, *Economic Review, Otaru University of Commerce*, **51**(1), 1–23.

Stocking, M.L. and Lord, F.M. (1983). Developing a common metric in item response theory, *Applied Psychological Measurement*, **7**(2), 201–210, https://doi.org/10.1177/014662168300700208.

Woodruff, D.J. and Hanson, B.A. (1996). Estimation of item response models using the EM algorithm for finite mixtures, ACT Research Report 96-6, ACT Incorporated, Iowa City, Iowa.

山田剛史 (2014). 項目反応理論入門,『R による項目反応理論』(加藤健太郎, 山田剛史, 川端一光 著), 2–20, オーム社, 東京.

How to Obtain a Common Scale for Psychological Concept: Methods and Practices of Equating and Linking

Haruhiko Mitsunaga

Graduate School of Education and Human Development, Nagoya University

Testing programs can reveal examinees' scores that reflect their ability or competency. To align multiple scales of different test administrations and obtain a common scale among these tests, researchers have proposed various equating or linking procedure. An equating method might be applied when multiple tests measure the same latent scale, whereas a linking method can be considered a collective term for using a common scale that measures the latent scale under weak constraints such as unidimensionality. This article illustrates a method of equating or linking. In practice, a testing program that ensures equity requires the specification that multiple tests should be administered longitudinally. We propose a block diagram notation to visualize a test design and equating procedure and describe an example of a large-scale assessment that equates the scales of multiple grades.

Key words: Test theory, item response theory, large-scale assessment, educational measurement.

統計数理(2024)
第 72 巻 第 1 号 79–92

特集「心理統計学の新展開」
［総合報告］

項目反応理論を用いた症状評価項目バンクの現状と今後の課題

国里 愛彦[1]・竹林 由武[2]

(受付 2023 年 6 月 30 日；改訂 10 月 20 日；採択 12 月 6 日)

要　　旨

患者報告アウトカムは，質問紙などを用いて患者から直接得られる健康状態に関する報告であり，臨床試験のアウトカムとしての利用も多い．患者報告アウトカム尺度は多数開発されているが，PROMIS®(Patient-Reported Outcomes Measurement Information System)では，項目反応理論によるコンピュータ適応型テストを想定した項目バンクの開発が行われている．コンピュータ適応型テストにより，測定精度を保ったまま質問数を減らすことができ，回答者の負担を減らすことができる．PROMIS®では，項目プールの開発，項目の心理測定学的検討，妥当性の検討の順番で尺度開発をすすめ，その開発にあたって科学的に妥当な基準を設定している．具体的には，(1)対象の構成概念の定義，(2)項目の構成，(3)項目プールの構築，(4)項目バンクの性質の特定，(5)検査のフォーマット，(6)妥当性，(7)信頼性，(8)解釈可能性，(9)翻訳と文化適応の 9 つの基準である．本論文では，PROMIS®での尺度開発プロセスについて解説をするとともに，日本国内において患者報告アウトカムの項目バンクを開発する上で必要となることについて議論した．

キーワード：患者報告アウトカム，項目反応理論，項目バンク，PROMIS®，COSMIN．

1. はじめに

心理統計学の臨床領域への応用として，患者報告アウトカム(patient-reported outcome [PRO])の開発と運用を挙げることができる．アメリカ食品医薬品局(Food and Drug Administration [FDA])によると，患者報告アウトカムは「患者から直接得られる患者の健康状態に関する全ての報告であり，患者の回答に対して臨床医や他の誰の解釈も介さないもの」とされる (Food and Drug Administration, 2009)．患者報告アウトカムの測定方法には，面接による測定方法もあるが，多くの場合は自己記入式質問紙による測定方法が用いられる．FDA が指針を出してから，臨床試験における患者報告アウトカムの活用が増え，その開発と運用も活発化している．日本国内でも患者報告アウトカムに対する関心は高くなってきており，「患者報告アウトカム(Patient-Reported Outcome: PRO)使用についてのガイダンス集」も作成されている (下妻 他, 2023)．

患者報告アウトカムの運用方法として，項目反応理論(item response theory [IRT])を用いたコンピュータ適応型テスト(computerized adaptive test [CAT])がある．紙とペンを用いて実施

[1] 専修大学 人間科学部：〒 214–8580 神奈川県川崎市多摩区東三田 2–1–1
[2] 福島県立医科大学 医学部：〒 960–1295 福島県福島市光が丘 1 番地

される質問紙とは異なり，コンピュータ適応型テストでは，コンピュータを用いることで参加者の反応に応じて呈示する項目を調整する．コンピュータ適応型テストを用いると，測定精度を保ったまま質問数を減らすことができ，回答者の負担を減らすことができる．コンピュータ適応型テストを用いて患者報告アウトカムを測定する場合，多数の尺度項目から構成される項目バンクを作成する必要がある．コンピュータ適応型テストを想定した患者報告アウトカムの項目バンクの開発においては，PROMIS®(Patient-Reported Outcomes Measurement Information System)の取り組みが参考になる．そこで，本論文では，PROMIS®で提案されている患者報告アウトカム尺度の項目バンク開発について整理し，国内における患者報告アウトカム尺度の項目バンクを開発する際の課題について議論する．

2. PROMIS®とは

PROMIS®は，米国国立衛生研究所(National Institutes of Health [NIH])が 2002 年に打ち出した医学研究のためのロードマップに従って，2004 年に NIH 主導の多施設共同研究グループとして立ち上げられたプロジェクトである（Cella et al., 2007）．Duke 大学，Stanford 大学，Stony Brook 大学，North Carolina 大学，Pittsburgh 大学，Washington 大学などの 6 つの研究拠点と統計調整センター(statistical coordinating center [SCC])を中心に構成されており，国家的なプロジェクトになる．PROMIS®は，様々な慢性疾患の症状や健康状態を測定するための項目バンクを構築・検証し，臨床現場で効率的に利用できるようにすることを目的としている（Cella et al., 2007）．その項目バンクの活用では，項目反応理論を用いたコンピュータ適応型テストを用いる．PROMIS®のウェブサイト(https://www.healthmeasures.net/explore-measurement-systems/promis)では，実際のコンピュータ適応型テストのデモも用意されており，ブラウザ上で回答に対する推定結果をその場で確認することができる．

PROMIS®では，プロジェクトの最初の段階で，慢性疾患にかかわる様々な領域(がん，リウマチ，精神保健など)の専門家，臨床試験の専門家，製薬業界などのステークホルダーを集めた専門家パネルを構成した（Cella et al., 2007）．そして，専門家パネルは PROMIS®がターゲットとするドメインを定め，その概念枠組みについても検討した（Cella et al., 2007）．初期の PROMIS®のドメインは，身体機能，疲労，疼痛，感情的苦痛，社会役割参加の 5 つであったが，現在は，図 1 にあるように身体的健康，精神的健康，社会的健康に大きく分けた上で，それぞれにプロフィールドメインと追加ドメインが置かれている[1)]．これらのドメインを測定する項目バンクの開発と検証については，PROMIS®が作成している検査開発と妥当化のための科学的基準（PROMIS, 2013)に従って説明する．

3. PROMIS®における患者報告式アウトカム尺度開発

PROMIS®では，項目プールの開発，項目の心理測定学的検討，妥当性の検討の順番で尺度開発を進める(図 2)．PROMIS®では，検査の開発と妥当化のための科学的基準（PROMIS, 2013）を定めており，(1)対象の構成概念の定義，(2)項目の構成，(3)項目プールの構築，(4)項目バンクの性質の特定，(5)検査のフォーマット，(6)妥当性，(7)信頼性，(8)解釈可能性，(9)翻訳と文化適応の 9 つの基準がある．以下ではそれぞれの基準について解説する．

3.1 (1)対象の構成概念の定義

PROMIS®の各ドメインで測定される構成概念は，明確に定義される必要がある（PROMIS, 2013）．構成概念の定義にあたり，まずは既存の文献レビューを行って，理論的・実証的観点から検討を行う．Pittsburgh 大学の PROMIS®グループは，包括的な文献検索方法の開発のた

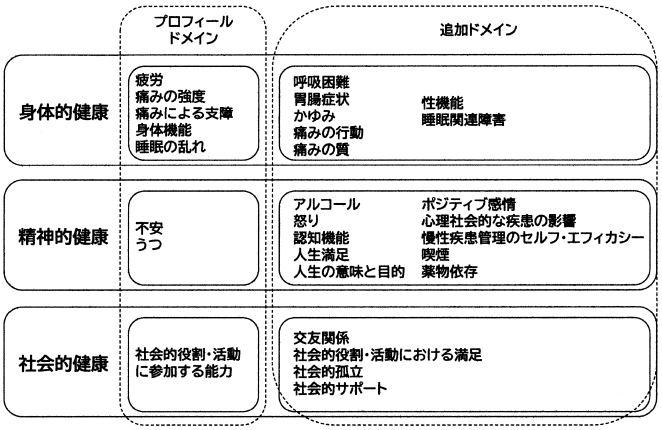


図 1. PROMIS®の成人用ドメインの構成.

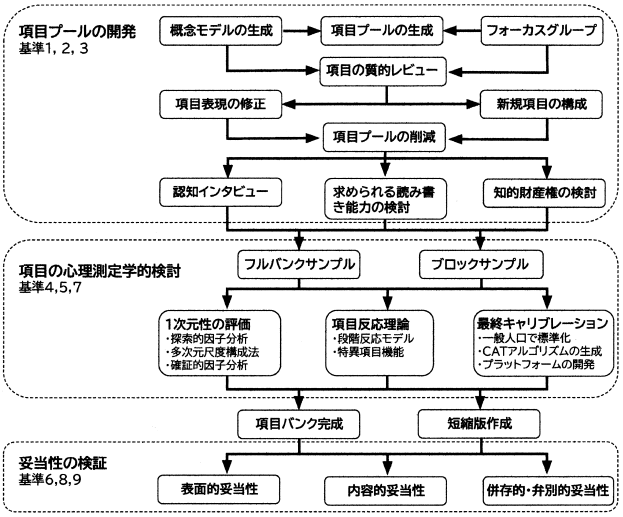


図 2. PROMIS®開発フロー．Pilkonis et al. (2011)の Figure 1 を参考に一部改変.

めに，図書館司書と協働している（Klem et al., 2009）．図書館司書が研究チームに加わることで，より網羅的に既存の文献を見つけることができ，その概念の研究領域での位置づけなどについての情報を得ることができる[2)]．

対象の構成概念は，その構成概念や測定の専門家，臨床家，患者，ユーザー，関連するステークホルダーによって，適切な質的研究法によるレビューが行われる必要がある（PROMIS, 2013）．まず，専門家パネルが暫定的な構成概念の定義について精査する．その具体的な実施方法としては，修正型デルファイ法が推奨されている（PROMIS, 2013）．修正型デルファイ法では，匿名調査を用いて専門家間のコンセンサスを得ることができる．以降の項目作成時のフォーカスグループでのフィードバック，得られたデータの統計解析(因子分析やモデル適合度)，項目作成後の専門家パネルの精査などを通して，構成概念の改訂が行われる．

3.2 (2)項目の構成

PROMIS®の項目を作成にするにあたり，測定する概念に関する既存の尺度と項目を包括的に文献検索して収集する（Klem et al., 2009）．収集した項目は，構成概念ごとにグループ化する「分配(binning)」を行う（DeWalt et al., 2007）．これにより，構成概念を捉えるのに必要な項目と冗長な項目を特定することができる．さらに，ドメインの定義にあわない項目や冗長な項目などを除外する「より分け(winnowing)」を行う（DeWalt et al., 2007）．「分配」も「より分け」も 1 名ではなく 2 名以上の研究者で取り組む．「分配」や「より分け」によって整理された項目に対して，文法的な正しさ，読みやすさ，翻訳可能性の観点から改訂する（DeWalt et al., 2007）．読みやすさに関しては，12 歳程度の読み書き能力があれば理解できる項目にすること，曖昧な表現は避けること，スラングは避けてシンプルな表現にすること，1 つの項目に複数の質問が含まれるダブルバーレルは避けることなどに留意する（DeWalt et al., 2007）．また，米国での使用を想定しているため，多様な文化的背景をもった回答者が回答しやすいように，他言語への翻訳可能な項目になるように調整される．具体的には，特定の国，文化圏，社会階層のみで通用する表現などは避ける．

項目の改訂は，専門家によってなされるだけでなく，認知インタビューを通しても行われる（PROMIS, 2013）．認知インタビューは非専門家や患者を対象に，項目が意図したような意味で理解されているのか，項目は理解しやすいか，どのように回答するのかを検討するために行われる（DeWalt et al., 2007）．PROMIS®では，各項目に対して，最低 5 名の参加者による認知インタビューが行われ，途中で大きな改訂がなされた場合はさらに 3 名から 5 名を追加して精査する（DeWalt et al., 2007）．実施方法には，参加者に項目を読んで回答するように依頼し，その後で，面接によって各項目の理解度や分かりにくい表現などを確認する回顧的プロービング(retrospective probing)法を用いる．認知インタビューでは参加者の多様性を保証するために，高校を卒業していない者，読み書きの水準が 15 歳以下の者，認知障害のある者などを含めたり，地域や人種的な多様性にも配慮する（DeWalt et al., 2007）．

3.3 (3)項目プールの構築

項目の収集・改訂が進み項目プール候補ができてきたら，項目プールが対象の構成概念を漏れなく捉えられているか検討する（PROMIS, 2013）．心理学や精神医学では，これまで様々な患者報告アウトカム尺度を開発しているが，同じ構成概念であっても，項目内容が異なることがある．例えば，主要な抑うつ症状尺度に対する調査では 52 個の抑うつ症状がリストアップされたが，1 つの尺度で測定しているのはそれらの内の最大 40% しかないという指摘もある（Fried et al., 2022）．既存の 1 つの尺度で測定できる範囲は限定的になる．項目プールが対象の構成概念を漏れなく捉えられているか検討する方法には，(1)構成概念の各側面をリスト化して対応する項目があるか検討する，(2)患者を対象としたフォーカスグループによって検討するの 2 種類がある．PROMIS®でのフォーカスグループは，3 つ以上のグループで，各グループが 6 から 12 名の患者，ファシリテーター，記録係で構成される（PROMIS, 2013）．フォーカスをあてたグループにするために，ある一定の基準を満たした患者に参加を依頼する．フォーカスグループでは，議論方法は構造化されておらず，自由な発言を促される．フォーカスグループを通して，患者の目線から，提示された項目が構成概念をまんべんなく測定できているか検討され，必要に応じて新規項目の追加や修正がなされる．

3.4 (4)項目バンクの性質の特定

項目バンクが作成できたら，項目の心理測定学的性質を回答者の代表的サンプルに基づいて決定する（PROMIS, 2013）．2006 年から 2007 年にかけて実施された PROMIS®の第 1 波調査

では，米国の一般人口と複数の患者集団からデータを収集している（Cella et al., 2010）．この調査では，世論調査会社を使って 21133 名からデータを収集した．また，米国国勢調査を参考にしてサンプルマッチング法を用いて代表性を担保している．次に，収集したデータから得られる各項目の心理測定学的性質を検討する．この検討を通して，項目バンクには優れた心理測定学的性質をもつ項目が残される．まず，項目反応理論の適用の前に，各項目に対して項目分析を実施する．項目分析では，(1)反応頻度，平均値，標準偏差，得点範囲，尖度，歪度などの記述統計量の検討，(2)項目間相関行列，項目-尺度相関，項目減による Cronbach の α 係数の変化などの内的整合性に関する検討が行われる（PROMIS, 2013）．

PROMIS®では項目反応理論を用いるが，項目反応理論適用の前提条件(一次元性，局所独立性，単調性)を満たすかどうかを確認する．一次元性の確認では，最初に 1 因子モデルの確証的因子分析による検討を行い，適合度の評価を行う（PROMIS, 2013）．なお，患者報告アウトカム尺度の回答は順序データと考えられるので，因子分析にあたってポリコリック相関係数を用いる．もし，確証的因子分析の適合度が低い場合は，探索的因子分析によって 1 因子構造が得られるか確認する（PROMIS, 2013）．探索的因子分析で 1 次元性を確認する目安の一つとして，第一因子の分散説明率が 20% 以上でかつ第二因子との分散説明率の比が 4 以上であることが挙げられる．下位因子を想定した上で全項目を説明する一般因子を想定する双因子(bifactor)モデルを検証することが適切な場合も存在する．双因子モデルで一般因子を想定することの合理性を判断するために共通分散説明率(explained common variance [ECV])が利用できる（Reise et al., 2013）．局所独立性は，個人の特性で条件づけられた際の項目間の反応が独立していることを指す．具体的には，ある項目が他の項目の反応に依存するものになってないか検討したり，1 因子モデルの確証的因子分析の残差相関行列から検討したりすることができる（PROMIS, 2013）．単調性は，健康に関する特性がより高いほど，健康に関する各項目への反応はより高くなることを指す．具体的には，合計得点から項目得点を引いた得点に条件づけられた項目平均得点のプロットや，ノンパラメトリック項目反応理論モデル(例えば，Mokken 尺度分析)によって検討することができる．Mokken 尺度分析で単調性を検討する場合には，各項目と尺度項目全体について推定されるスケーラビリティ係数(scalability coefficients)H を基準とすることができる（Mokken, 1971）．一次元性，局所独立性，単調性は，項目反応理論の適用の条件となるため，項目バンク内の項目がこれらを満たすか確認してから，項目反応理論モデルを適用する必要がある．また，一次元性の検討などを通して，項目反応理論モデルの適用の前に項目プールから項目を減らし，整理することもできる（Reeve et al., 2007）．

PROMIS®では，一次元性が仮定できる多肢選択式項目が基本になるので，段階反応モデル(Samejima, 1969)が用いられる．段階反応モデルで推定されるパラメータは，個人の特性値 θ，各項目の識別力 a と困難度 b になる．特性値 θ，識別力 a，困難度 b の下で，反応がカテゴリー k(多肢選択の 1 つ)である確率が，以下の式を用いて計算される（PROMIS, 2013）．

$$P\left(X_i = k \mid \theta, b_i, a_i\right) = \frac{1}{1+\exp\left[-a_i\left(\theta - b_{i,k-1}\right)\right]} - \frac{1}{1+\exp\left[-a_i\left(\theta - b_{i,k}\right)\right]}$$

項目反応理論モデルに対してもモデル適合度の評価を行う．モデル適合度の評価には，観測反応頻度と期待反応頻度の比較や適合度指標(χ^2，尤度比 G^2)を用いる．項目反応理論モデルの推定結果は，カテゴリー反応曲線や項目情報曲線から解釈をする．項目反応理論の観点から項目を検討し，性能が低い項目については，専門家パネルが内容のレビューをした上で項目バンクに残すかどうか判断する（Hansen et al., 2014）．性能は低いが，臨床的な関連性が高い項目は，保持されたり，修正されたりする．また，PROMIS®で用いる基本的な項目反応理論モデルでは 1 因子構造を想定しているが，実際の項目プールの構造は 1 因子構造とは限らない．そ

のため，多次元項目反応理論モデルを用いることも適宜検討される（PROMIS, 2013）．

PROMIS®では，多様な背景をもった患者の患者報告アウトカム尺度を開発するため，特定の集団において他の集団とは異なる機能をもった項目を特定する必要がある．このような特定の集団（ジェンダー，年齢，疾患，教育，人種など）に対して項目が特異的な機能を有する性質を特異項目機能（differential item functioning [DIF]）と呼ぶ．特異項目機能のある項目があると，特定の集団に関して過小もしくは過大評価をしてしまうため，現状では PROMIS®では特異項目機能のある項目は項目プールから除外する．特異項目機能については，事前に仮説を設定した上で検討する．特異項目機能の特定は，項目反応理論に基づいて尤度比検定や項目反応理論には基づかない順序ロジスティック回帰，あるいは構造方程式モデリングの多重指標多重原因（Multiple indicators multiple causes [MIMIC]）モデルや確証的因子分析モデルの多母集団同時分析を用いた方法が代表的である（Teresi et al., 2021）．例えば，MIMIC モデルを用いた方法では，項目と測定概念を反映する潜在変数からなる確証的因子分析モデルに集団の属性を反映する項目を加え，属性項目から潜在変数とすべての尺度項目にパスを追加する．そこから修正指数等を参照し，属性項目が尺度項目に有意な影響性を示す項目を特異項目機能を示す項目と判断する．そして，特異項目機能の大きさと影響を考慮して，特異項目機能のある項目を項目プールから除外する．なお，項目プールから除外せずに，特異項目機能も考慮した上で推定することも考えられるが，現状では採用されていない．

PROMIS®の項目プール全体には多数の項目が含まれているので，参加者に全ての項目に回答を求めるのは難しい．例えば，PROMIS®の第 1 波調査では，項目プール全体で 1000 項目以上が含まれており，1 名が全ての項目に回答するのは難しかった（Cella et al., 2010）．そこで，フルバンクデザインとブロックデザインの 2 種類を実施し，1 人の回答者が回答するのは 150 項目程度に限定して調査を実施している．フルバンクデザインは，参加者に PROMIS®の特定のドメイン内の全ての項目に回答を求める調査デザインである．これによって項目反応理論の適用にかかわる一次元性の評価とドメイン内の項目のキャリブレーションが可能となる．また，既存の尺度も一緒にデータ収集することで，既存の尺度との等化も可能となる（Thompson et al., 2017）．例えば，既存の抑うつ尺度である Patient Health Questionnaire(PHQ)-9 と PROMIS®のうつ尺度でキャリブレーションを実施した研究から，9 項目の PHQ-9 よりも PROMIS®の適応型テストの 3 項目の方が測定誤差が小さいことが示されている（Gibbons et al., 2011）．ブロックデザインでは，複数のドメインの一部の項目に回答をすることで，ドメイン間の関係の評価が可能となる．ブロックデザインでは，項目のブロックごとに一般集団から 900 名以上，500 名の慢性疾患患者を含むように調査が設計された（Cella et al., 2010）．項目バンクには項目に関するパラメータが明らかになった項目が収載され，コンピュータ適応型テストで利用される．

3.5 (5)検査のフォーマット

PROMIS®の項目バンクの活用法としては，参加者の反応に応じて質問を提示するコンピュータ適応型テストや，予め定められた項目から構成される項目バンクの短縮版尺度などがある．それぞれの検査形式は，使用目的や項目バンクの性質に合わせて設定される（PROMIS, 2013）．また，検査として使う上では，各尺度の検査としての適切な性質を示し，回答者の負担についても検討を行う．回答者の負担は回答にかかる時間や回答数などが含まれる．パソコン上での回答や紙とペンによる回答などの異なる実施方法での比較可能性についても検討する．

PROMIS®等で用いられるコンピュータ適応型テストの基本的なアルゴリズムを図 3 に示す．コンピュータ適応型テストは，項目反応理論を通じてキャリブレーションされた項目バンクから特定の能力値（θ）を任意に選択し，その能力値に適した項目の選択と提示が行われる．回答者

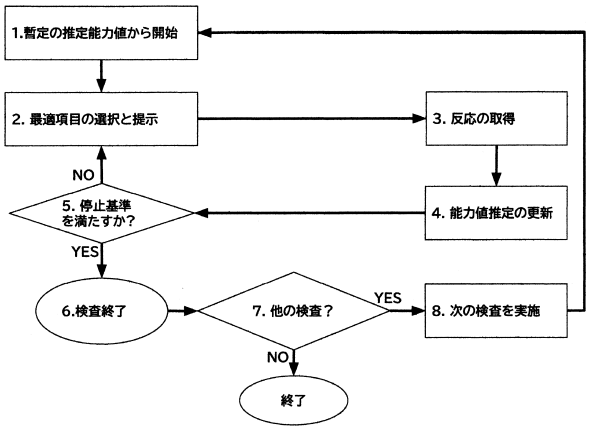


図 3. 適応型テストの基本アルゴリズム．İnce Araci and Tan (2022)を参考に一部改変.

の θ に関する事前情報がない場合には，母集団の θ の平均値を割り当て，それに適した項目の提示から開始する，あるいは項目バンクに含まれる最初の項目を提示するといった方法が一般的である．最初に提示した項目への反応に基づいて θ の推定を更新する．θ の推定は，最尤法またはベイズ推定による期待事後推定量(Expected a Posteriori Estimator [EAP])や最大事後確率推定量(Maximum a Posteriori Estimator [MAP])がよく用いられる（Lord and Novick, 2008; Segall, 1996）．能力値 θ の更新後，次の質問項目を提示する必要性を停止基準に従って判断する．停止基準には標準誤差がよく用いられる．標準誤差が一定の水準まで小さくなった時点で項目の追加提示を止め，その時点の能力値を求めることで精度の高い推定値が得られる．能力値の更新後に次に提示する項目を選択する方法には，A ルール，E ルール，D ルール，T ルール，W ルール，Kullback-Leibler 情報量規準(KL)，連続エントロピー法などがある（İnce Araci and Tan, 2022）．これらの規準は最終的な能力値推定における規準値を最大化または最小化するという点で共通しているが，規準値の定義が異なる．例えば，KL 法と連続エントロピー法は，それぞれ事後期待 KL 情報を最大化し，期待連続エントロピーを最小化する（Mulder and van der Linden, 2009）．

3.6 (6)妥当性

PROMIS®の妥当性検討では，基準関連妥当性，構成概念妥当性，内容的妥当性，反応性を検討する（PROMIS, 2013）．基準関連妥当性では，基準となる尺度を定めて，その尺度との関連の強さについて事前に仮説を立てて検証する．PROMIS®の場合，基準としては既存の尺度や診断などのゴールドスタンダードとなる基準が挙げられる．構成概念妥当性では，測定する構成概念とその他の構成概念との間の関係について事前に仮説を立てて検証する（Cella et al., 2010）．具体的には，測定する構成概念と関連する構成概念との関係について検証する収束的妥当性，測定する構成概念と関連しない構成概念との関係について検証する弁別的妥当性について検討する．内容的妥当性の検討は項目の作成の段階でも行われてきているが，項目反応理論によるキャリブレーション後にも検討する（Riley et al., 2010）．特に，項目反応理論の適用などに際して項目の絞り込みがなされている場合に，当初測定する予定だったドメインと項目バンクの内容に乖離が生じる可能性がある．内容的妥当性では，最初に定義したドメインと作成された検査との間の一貫性について検討する．

反応性は構成概念の時間的な変化に関する妥当性であり，患者の変化を尺度が捉えることが

できているのかを検討する（PROMIS, 2013）．そのため，ある程度の変化が期待できる時間間隔でデータを測定し，変化が期待できるグループと安定していると期待できるグループで比較したり，医師や患者による変化に関する評定のような外部アンカーとの関連を検討したりする．例えば，PROMIS®のうつ尺度に関する反応性を調べた研究では（Pilkonis et al., 2014），うつ病患者の治療初期から 3 ヶ月後までのフォローアップ調査を行って，治療に伴う症状変化を測定した．その結果，PROMIS®のうつ尺度は，各測定時点の既存の尺度の PHQ-9 と CES-D（Center for Epidemiological Studies Depression）と相関しており，変化に関する収束的妥当性を示した．また，患者による全般的改善の評価（「非常に改善」から「悪化」で評価）をアンカーとして，そのアンカーと PROMIS®のうつ尺度の変化が対応することも確認している．

3.7 （7）信頼性

PROMIS®では，対象構成概念の信頼性と再検査信頼性を検討する（PROMIS, 2013）．対象構成概念の信頼性については，古典的テスト理論の場合は尺度全体の Cronbach の α 係数や測定誤差から検討できる．項目反応理論の場合は特性値を横軸においてテスト情報量をプロットしたテスト情報曲線を描くことができ，構成概念の特性値の全範囲における信頼性を検討することができる．対象構成概念の再検査信頼性については，2 時点の間での測定値の相関を検討する．その際には，2 時点の間で対象集団に大きな変化はなく安定していることが想定できるようにデータ収集する必要がある．

3.8 （8）解釈可能性

PROMIS®では回答から連続的な得点が得られるが，その得点を臨床的な意味として解釈しやすい質的なカテゴリーに分けることで解釈可能性を高めることができる．PROMIS®では，最小限の重要な差（minimally important difference [MID]）を示すことが推奨されている．最小限の重要な差の計算方法としては，尺度の測定誤差などの分布に基づく方法と患者や医療者による変化に関する外的アンカーに基づく方法がある（宮崎, 2015）．なお，時間的な変化に関しては，最小限の重要な差ではなく，最小限の重要な変化（minimally important change [MIC]）と呼ぶことがある（Mokkink et al., 2010）．

妥当性，信頼性，解釈可能性については，PROMIS®の基準やその付録も参考になるが，それらの検討方法は COSMIN（COnsensus-based Standards for the selection of health Measurement INstruments）が詳しい．COSMIN では，患者報告アウトカムの選択・評価に関するガイドラインを作成しており（Mokkink et al., 2010; Prinsen et al., 2018; 佐藤・土屋, 2022），それらは尺度開発においても有用である．COSMIN のガイドラインは PROMIS®の基準と内容的に重複する部分もあるが，COSMIN は格付けのための明確な規格を持ったチェックリストを提供するため，患者報告アウトカム尺度を開発する場合にも活用することができる．

3.9 （9）翻訳と文化適応

PROMIS®は，英語で項目バンクを作成しているが，PROMIS®の項目バンクを他言語に翻訳し，活用することもできる．PROMIS®では，教示，項目，選択肢について厳密な翻訳過程を通して他言語への翻訳が行われる（PROMIS, 2013）．具体的には，英語から他言語への順翻訳と翻訳した他言語から英語への逆翻訳をおこなって，他言語への翻訳によって項目の意味が異なってないか確認する．順翻訳と逆翻訳に関しては，繰り返し実施して，適切な翻訳になるように調整し，そのプロセスにおいてはバイリンガルの専門家による検討も推奨される（PROMIS, 2013）．また，逆翻訳を想定して翻訳された項目は，読みにくかったり，分かりにくかったりすることもあるため，認知インタビューによる検討を通してその文化にあった項目

を開発することが重要になる．なお，PROMIS®では，尺度翻訳にかかわる最小限の要件を設定している（PROMIS, 2014）．この最小限の要件としては，リリース前の要件（適切な翻訳，特異項目機能を含む予備検討）とリリース後の要件（項目バンクのキャリブレーション，その言語における参照得点，信頼性，妥当性，反応性，解釈可能性）が設定されている．

4. PROMIS®の検査成熟モデル

上記の 1 から 9 の基準にしたがって PROMIS®の項目は作成されるが，PROMIS®では検査としての成熟モデルを設定し，各検査がどの開発段階か分かるようにしている（PROMIS, 2012）．具体的には，ステージ 1（開発：概念化と項目プール），ステージ 2（開発：キャリブレーション），ステージ 3（一般公開：キャリブレーションと予備的な妥当性検討），ステージ 4（成熟化：反応性と拡張），ステージ 5（完全に成熟したユーザーサポート）の 5 ステージからなる．ここまで説明をした基準にもとづいて，ステージ 1 からステージ 3 まで開発し，一般公開する．その後，ステージ 4 において，異なる臨床集団に対して実施したり，反応性や解釈可能性について検討し，より検査を使いやすくする．最後のステージ 5 は，様々な集団で検討され解釈も可能となった状態であり，ユーザーが利用しやすいように採点と解釈マニュアルなどを作成する．PROMIS®の検査成熟モデルを用いることで，患者報告アウトカム尺度の開発を，プロダクトの開発プロセスのように実施することができ，プロジェクトの進行確認などにも有用と思われる．

5. PROMIS®の現状と今後の課題

PROMIS®は，項目反応理論を用いたコンピュータ適応型テストをベースにした患者報告アウトカム尺度の開発を行っている．患者報告アウトカム尺度や精神症状を測定する尺度は，紙とペンを用いたものが多く，臨床現場で実施する場合は採点などを行う必要があった．また，毎回同じ項目を用いることによって患者があまり深く考えずに回答することもあるかもしれず，項目数などの負担も大きい．一方で PROMIS®は，コンピュータ適応型テストによって，患者の反応に合わせて項目を提示し反応を取得するため，患者の負担は小さくなり採点も自動化されるため実施者の負担も小さくなる．また，実施や採点に対する負担だけでなく，項目反応理論に基づく推定値は古典的テスト理論に基づく得点よりも精度が高くなる．例えば，慢性的な腰痛や首の痛みに関する治療上の変化について，古典的テスト理論に基づくものよりも，PROMIS®の項目反応理論に基づく変化の推定のほうが正確になるとされる（Hays et al., 2021）．PROMIS®のような患者報告アウトカムの項目バンクの作成は，臨床的にも有用であり，今後も利用の拡大が予想できる．

当初，PROMIS®の項目は英語で作成されたが，米国内の英語以外の話者（スペイン語と中国語）も使えるように他言語への翻訳も進められている．また，米国内だけでなく，米国以外の国で活用するための翻訳も行われている．ノルウェー語版 PROMIS®（Rimehaug et al., 2022）のように翻訳と検証プロセスを論文化しているものもあるが，それ以外にも PROMIS®のウェブサイトから多くの言語に翻訳されていることが確認できる．痛み，うつ，不安，疲労，身体活動，睡眠，イライラなどの PROMIS®尺度については，既に日本語訳されていることも分かる．

PROMIS®を日本でも活用することができれば，臨床研究と臨床実践の両面への多大な貢献が期待できる．しかし，英語とスペイン語については無料で利用できるが，その他の言語に関しては手数料が必要とされる（PROMIS, 2023）．そのため，日本語版は項目などがオープンにはなっていない．なお，研究利用の場合は，手数料が免除される可能性もあるとされている．患者報告アウトカムの開発や翻訳には多くの研究資金が投入されていることを考えると手数料や利用料がかかることは尤もであるかもしれない．しかし，手数料や利用料によって

PROMIS®の活用に制限がかかる可能性もある．

手数料の問題だけでなく，患者報告アウトカムの測定は使われる言語に依存することを考慮すると，日本でも PROMIS®のような患者報告アウトカムの項目バンクを開発する必要があるかもしれない．日本語の特性も考慮した項目バンクの作成は患者にとって切実なアウトカムの測定につながることが期待できる．もし，日本において患者報告アウトカムの項目バンクを開発する場合，(1)方法論的な厳密性を担保した開発フレームの設定，(2)教示，項目，選択肢，推定された項目パラメータ，解釈のための情報のオープン化，(3)図書館情報学研究者，患者，臨床実践家，統計学者，臨床研究者からなるプロジェクトチームが必要と考える．

まず，日本において患者報告アウトカムの項目バンクを開発するために，方法論的な厳密性を担保した開発フレームを設定する必要がある．海外で開発された項目バンクの翻訳ではなく日本で項目バンクを開発する場合は，方法論的な厳密性に基づく等価性を担保することが必要になる．既に PROMIS®や COSMIN において，患者報告アウトカム尺度開発の指針は示されている．開発にあたっては，PROMIS®が行ったようなドメインの枠組みの検討のように，大枠の検討から初めて，それぞれのドメインの開発をすることが重要になる．その点において PROMIS®は参考になるが，文化特異的なドメインが含まれる可能性もある．PROMIS®の検査成熟モデルにあるような，開発の進捗が分かりやすいような枠組みも採用し，プロジェクトが順調に進むような工夫も必要となる．

次に，教示，項目，選択肢，推定された項目パラメータ，解釈のための情報はオープン化し，研究利用から実践まで幅広く利用できるようにする必要がある．そのためにも PROMIS®において行われている尺度の項目のライセンスのチェックなども必要になる．また，オープン化することで，利用者の項目への曝露が増えてしまう可能性もある．項目バンクは随時項目を追加しながら持続的に発展するようなシステムにできると良いだろう．項目バンクのオープン化に関しては，参加者に日に複数回質問を行う経験サンプリングに関する項目リポジトリが参考になるかもしれない（Kirtley et al., 2019）．ESM Item Repository（https://www.esmitemrepository.com/）では，経験サンプリングで利用できる項目が各種メタ情報とともに複数の言語で利用可能になっている．

最後に，項目バンクの作成にあたっては，図書館情報学研究者，患者，臨床実践家，統計学者，臨床研究者からなるプロジェクトチームが必要となる．そのためには，異なる専門性や立場の者が集まってプロジェクトに持続的に取り組むための仕組みが必要となる．2〜3 年で終わるプロジェクトではなく長期間にわたり，メンバーが入れ替わりつつ取り組むための枠組みや組織が必要となるだろう．

上記のように，日本での患者報告アウトカムの項目バンク開発には，多くの困難があると予想される．困難はあるものの，項目バンク開発は患者のケア向上につながり，さらなる研究利用にもつながることが期待できる．本論文が，日本における患者報告アウトカムの項目バンク開発の一助になれば幸いである．

謝　辞

本研究は JSPS 科研費 JP20K20870，JP20H00625 の助成を受けたものです．

注．

[1] ドメイン構成，ドメインに含まれる構成概念，項目数などの詳細は，PROMIS®のウェブサイトに掲載されている．例えば，2023 年の段階では，成人用 PROMIS®のうつ尺度の項目プールには 28 個の項目が含まれる．また，英語版 PROMIS®については，具体的な項

目もウェブサイトから入手することができる．

[2] 欧米の研究機関などでは，図書館情報学などの学位のある者が図書館司書を担っている．彼らは文献検索などについて専門的知識を有しており，系統的レビューなどをサポートすることができる．日本国内においては，図書館情報学の専門家を研究チームに加えることが対応すると思われる．

参 考 文 献

Cella, D., Yount, S., Rothrock, N., Gershon, R., Cook, K., Reeve, B., Ader, D., Fries, J. F., Bruce, B., Rose, M. and PROMIS Cooperative Group (2007). The Patient-Reported Outcomes Measurement Information System (PROMIS): Progress of an NIH roadmap cooperative group during its first two years, *Medical Care*, **45**(5 Suppl 1), S3–S11.

Cella, D., Riley, W., Stone, A., Rothrock, N., Reeve, B., Yount, S., Amtmann, D., Bode, R., Buysse, D., Choi, S., Cook, K., Devellis, R., DeWalt, D., Fries, J. F., Gershon, R., Hahn, E. A., Lai, J.-S., Pilkonis, P., Revicki, D., Rose, M., Weinfurt, K., Hays, R. and PROMIS Cooperative Group (2010). The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005–2008, *Journal of Clinical Epidemiology*, **63**(11), 1179–1194.

DeWalt, D. A., Rothrock, N., Yount, S., Stone, A. A. and PROMIS Cooperative Group (2007). Evaluation of item candidates: The PROMIS qualitative item review, *Medical Care*, **45**(5 Suppl 1), S12–S21.

Food and Drug Administration (2009). Guidance for industry: Patient-reported outcome measures: Use in medical product development to support labeling claims, https://www.fda.gov/regulatory-information/search-fda-guidance-documents/patient-reported-outcome-measures-use-medical-product-development-support-labeling-claims（最終アクセス日 2023 年 6 月 26 日）.

Fried, E. I., Flake, J. K. and Robinaugh, D. J. (2022). Revisiting the theoretical and methodological foundations of depression measurement, *Nature Reviews Psychology*, **1**(6), 358–368.

Gibbons, L. E., Feldman, B. J., Crane, H. M., Mugavero, M., Willig, J. H., Patrick, D., Schumacher, J., Saag, M., Kitahata, M. M. and Crane, P. K. (2011). Migrating from a legacy fixed-format measure to CAT administration: Calibrating the PHQ-9 to the PROMIS depression measures, *Quality of Life Research*, **20**(9), 1349–1357.

Hansen, M., Cai, L., Stucky, B. D., Tucker, J. S., Shadel, W. G. and Edelen, M. O. (2014). Methodology for developing and evaluating the PROMIS smoking item banks, *Nicotine and Tobacco Research*, **16**(Suppl 3), S175–S189.

Hays, R. D., Spritzer, K. L. and Reise, S. P. (2021). Using item response theory to identify responders to treatment: Examples with the Patient-Reported Outcomes Measurement Information System (PROMIS®) physical function scale and emotional distress composite, *Psychometrika*, **86**(3), 781–792.

İnce Araci, F. G. and Tan, Ş. (2022). Multidimensional computerized adaptive testing simulations in R, *International Journal of Assessment Tools in Education*, **9**(1), 118–137.

Kirtley, O. J., Hiekkaranta, A. P., Kunkels, Y. K., Verhoeven, D., Van Nierop, M. and Myin-Germeys, I. (2019). The Experience Sampling Method (ESM) item repository, http://dx.doi.org/10.17605/OSF.IO/KG376.

Klem, M., Saghafi, E., Abromitis, R., Stover, A., Dew, M. A. and Pilkonis, P. (2009). Building PROMIS item banks: Librarians as co-investigators, *Quality of Life Research*, **18**(7), 881–888.

Lord, F. M. and Novick, M. R. (2008). *Statistical Theories of Mental Test Scores*, Information Age Publishing, Charlotte, North Carolina.

宮崎貴久子 (2015). QOL 評価の臨床的意味：Minimally Important Difference（臨床における最小重要差：MID）, 行動医学研究, **21**(1), 8–11.

Mokken, R. J. (1971). *A Theory and Procedure of Scale Analysis with Applications in Political Research*, De Gruyter Mouton, Hague, Netherlands.

Mokkink, L. B., Terwee, C. B., Knol, D. L., Stratford, P. W., Alonso, J., Patrick, D. L., Bouter, L. M. and de Vet, H. C. (2010). The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: A clarification of its content, *BMC Medical Research Methodology*, **10**, 22.

Mulder, J. and van der Linden, W. J. (2009). Multidimensional adaptive testing with optimal design criteria for item selection, *Psychometrika*, **74**(2), 273–296.

Pilkonis, P. A., Choi, S. W., Reise, S. P., Stover, A. M., Riley, W. T., Cella, D. and PROMIS Cooperative Group (2011). Item banks for measuring emotional distress from the Patient-Reported Outcomes Measurement Information System (PROMIS®): Depression, anxiety, and anger, *Assessment*, **18**(3), 263–283.

Pilkonis, P. A., Yu, L., Dodds, N. E., Johnston, K. L., Maihoefer, C. C. and Lawrence, S. M. (2014). Validation of the depression item bank from the Patient-Reported Outcomes Measurement Information System (PROMIS) in a three-month observational study, *Journal of Psychiatric Research*, **56**, 112–119.

Prinsen, C. A. C., Mokkink, L. B., Bouter, L. M., Alonso, J., Patrick, D. L., de Vet, H. C. W. and Terwee, C. B. (2018). COSMIN guideline for systematic reviews of patient-reported outcome measures, *Quality of Life Research*, **27**(5), 1147–1157.

PROMIS (2012). PROMIS® Instrument Maturity Model, https://staging.healthmeasures.net/images/PROMIS/PROMISStandards_Vers_2_0_MaturityModelOnly_508.pdf（最終アクセス日 2023 年 6 月 26 日）.

PROMIS (2013). PROMIS® Instrument Development and Validation Scientific Standards Version 2.0, https://www.healthmeasures.net/images/PROMIS/PROMISStandards_Vers2.0_Final.pdf（最終アクセス日 2023 年 6 月 26 日）.

PROMIS (2014). Minimum Requirements for the Release of PROMIS Instruments after Translation and Recommendations for Further Psychometric Evaluation, https://staging.healthmeasures.net/images/PROMIS/Standards_for_release_of_PROMIS_instruments_after_translation_v8.pdf（最終アクセス日 2023 年 6 月 26 日）.

PROMIS (2023). Available Translations, https://www.healthmeasures.net/explore-measurement-systems/promis/intro-to-promis/available-translations（最終アクセス日 2023 年 6 月 20 日）.

Reeve, B. B., Hays, R. D., Bjorner, J. B., Cook, K. F., Crane, P. K., Teresi, J. A., Thissen, D., Revicki, D. A., Weiss, D. J., Hambleton, R. K., Liu, H., Gershon, R., Reise, S. P., Lai, J.-S., Cella, D. and PROMIS Cooperative Group (2007). Psychometric evaluation and calibration of health-related quality of life item banks: Plans for the Patient-Reported Outcomes Measurement Information System (PROMIS), *Medical Care*, **45**(5 Suppl 1), S22–S31.

Reise, S. P., Bonifay, W. E. and Haviland, M. G. (2013). Scoring and modeling psychological measures in the presence of multidimensionality, *Journal of Personality Assessment*, **95**(2), 129–140.

Riley, W. T., Rothrock, N., Bruce, B., Christodolou, C., Cook, K., Hahn, E. A. and Cella, D. (2010). Patient-reported outcomes measurement information system (PROMIS) domain names and definitions revisions: Further evaluation of content validity in IRT-derived item banks, *Quality of Life Research*, **19**(9), 1311–1321.

Rimehaug, S. A., Kaat, A. J., Nordvik, J. E., Klokkerud, M. and Robinson, H. S. (2022). Psychometric properties of the PROMIS-57 questionnaire, Norwegian version, *Quality of Life Research*, **31**(1), 269–280.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores, *Psychometrika*, **34**(Suppl 1), 1–97.

佐藤秀樹, 土屋政雄 (2022). 尺度研究における COSMIN ガイドラインの動向, 認知行動療法研究, **48**(2), 123–134.

Segall, D. O. (1996). Multidimensional adaptive testing, *Psychometrika*, **61**(2), 331–354.

下妻晃二郎, 鈴鴨よしみ, 宮崎貴久子, 内藤真理子, 中島貴子, 川口崇, 山口拓洋, 齋藤信也, 兼安貴子, 星野絵里, 小嶋智美, 堺琴美, 白岩健, 宮路天平, 森脇健介 (2023). 厚生労働省科学研究班開発患者報告アウトカム(Patient-Reported Outcome:PRO)使用についてのガイダンス集, https://www.lifescience.co.jp/pro/index.html (最終アクセス日 2023 年 6 月 20 日).

Teresi, J. A., Wang, C., Kleinman, M., Jones, R. N. and Weiss, D. J. (2021). Differential item functioning analyses of the Patient-Reported Outcomes Measurement Information System (PROMIS®) measures: Methods, challenges, advances, and future directions, *Psychometrika*, **86**(3), 674–711.

Thompson, N. R., Lapin, B. R. and Katzan, I. L. (2017). Mapping PROMIS global health items to EuroQol (EQ-5D) utility scores using linear and equipercentile equating, *PharmacoEconomics*, **35**(11), 1167–1176.

Current Status and Future Directions of Patient-Reported Outcome Item Banks Using Item Response Theory

Yoshihiko Kunisato[1] and Yoshitake Takebayashi[2]

[1]School of Human Sciences, Senshu University
[2]School of Medicine, Fukushima Medical University

Patient-reported outcomes are health status reports obtained directly from patients through methods such as questionnaires and are often used as outcomes in clinical trials. Although numerous patient-reported outcome measures have been developed based on classical test theory, the Patient-Reported Outcomes Measurement Information System (PROMIS®) has been developed an item bank for computerized adaptive tests based on item response theory. The computerized adaptive test enables the number of questions to be reduced while maintaining measurement precision, thereby lessening respondent burden. PROMIS® advances scale development in the order of item pool development, psychometric testing of items, and check of validity, setting scientific standards for scale development. The scientific standards involves (1) definition of target construct, (2) composition of individual items, (3) item pool construction, (4) determination of item bank properties, (5) testing and instrument formats, (6) validity, (7) reliability, (8) interpretability, and (9) language translation and cultural adaptation. In this paper, we discuss the scale development process in PROMIS® and deliberate the requirements when developing patient-reported outcome item banks in Japan.

Key words: Patient-reported outcome, item response theory, item bank, PROMIS®, COSMIN.

統計数理(2024)
第 72 巻 第 1 号 93–119

特集「心理統計学の新展開」
[研究詳解]

近年の診断分類モデルの推定法の展開

山口 一大†

(受付 2023 年 6 月 23 日；改訂 12 月 7 日；採択 12 月 12 日)

要　　旨

診断分類モデルあるいは認知診断モデルは学力テストの解答に必要な一連の認知能力(アトリビュート)を想定し，テスト受験者をアトリビュート習得の有無のパタンに分類する統計モデルである．診断分類モデルは，教育テスト分析において有用なツールであり，応用も広がりつつあるものの，パラメタ推定についての包括的な日本語のレビューは未だ存在しない．そこで本稿では診断分類モデルのパラメタ推定法についての展開を展望し，診断分類モデルの応用の促進と理論的発展に資することを目指した．レビューの結果，最尤推定法，ベイズ推定法，ノンパラメトリック推定法の 3 種類の方法での発展がみられた．最尤推定法では正則化法の利用，ベイズ推定法では変分ベイズといった比較的新しい方法も用いられていた．レビューの結果を踏まえて，診断分類モデルの推定法について残された問題と今後の展望について議論した．

キーワード：診断分類モデル，認知診断モデル，パラメタ推定法．

1. はじめに

学力テストからテスト解答者の強みや弱み推定することができる統計分析モデルとして，診断分類モデル(diagnostic classification models [DCMs])あるいは認知診断モデル(cognitive diagnostic models [CDMs])と呼ばれるモデル (Rupp et al., 2010; von Davier and Lee, 2019)への注目が集まっている．DCM では，例えば中学生における「数学能力」という包括的な構成概念を分解し，「計算力」，「概念理解」，「図形操作力」，「論理力」などといったテストの正答に必要なより細かい認知要素(アトリビュート；attribute)を想定することで，個々のテスト解答者がどのアトリビュートを習得しているのかを推定することができる(応用例として，鈴木 他, 2015 などが参考になる)．このように，DCM はテスト解答者がどのような認知的なつまづきを抱えているのかを，学力テストへの項目反応パタンから推定し，個別のテスト解答者への学習の診断情報の提供を可能にするモデルである．

DCM は，統計モデルとしては制約付き潜在クラスモデル(restricted latent class models; Rupp and Templin, 2008b)として考えることができる．DCM では，アトリビュートの習得の有無の組み合わせによって定義されるアトリビュート習得パタンが潜在クラスに対応する．和文で参照できる DCM のレビューとして，山口・岡田 (2017)があり，DCM の主要なモデルについては網羅的に解説されている．

しかし，山口・岡田 (2017)では，DCM の推定法の詳細は記載されていない．この他，山口 (2019)は DCM の中でも最も基本的な DINA モデル(deterministic input noisy “AND”-gate model; e.g., Junker and Sijtsma, 2001)の推定法を示しているが，特定の DCM の特定の推

† 筑波大学 人間系心理学域：〒 305–0006 茨城県つくば市天王台 1–1–1

定法について述べているにすぎず，記述が限定的であると言わざるを得ない．DCM についての成書（e.g., Rupp et al., 2010; von Davier and Lee, 2019）においても，DCM のパラメタ推定について包括的にレビューされていない．DCM を含めた潜在変数モデルにおいては，一般にモデルの特定のみならずパラメタ推定法の選択も重要であるものの，包括的に DCM のパラメタ推定法をレビューした日本語で参照できる文献がないという現状がある．

DCM のパラメタ推定法をレビューした論文がない理由として，DCM 定式化の複雑さがあると考えられる．DCM は，アトリビュートを多次元のカテゴリカル変数とみなすか，アトリビュート習得パタンを潜在クラスとみなすかという定式化の任意性があり，結果の解釈をしやすい定式化と推定を行いやすい定式化に隔たりがある．すなわち，推定法を統合的に理解する際にわかりやすい定式化とモデルを応用していく際に解釈しやすい定式化が一致していないということである．このような DCM の定式化の複雑さから，DCM の推定方法のレビューが行われてこなかった可能性がある．この結果として，DCM の専門家以外にはパラメタの推定法を統合的に理解することが難しい現状があると考えられる．本稿の意義は，DCM の 2 種類の定式化とその関係を示し，現在提案されている DCM の推定法を概観し，それぞれの推定法の特色について整理することにある．

上記を踏まえて，本稿ではパラメタ推定法を定式化しやすい定式化として潜在クラスモデルベースのものを採用し，一般性の高い DCM がどのように制約つき潜在クラスモデルとして定式化できるのかを示す．その上で，多数の推定法の異同や，DCM に特有のノンパラメトリック推定法や単調性制約を満足するような推定法についても合わせてレビューする．これにより，DCM の特徴を，推定法の観点からも理解できると期待できる．DCM の推定法をレビューすることにより，DCM の応用のみならず DCM の理論的研究の発展に寄与することを本稿の目的とする．なお，本研究では Q 行列と呼ばれる要素が既知である確証的な DCM のパラメタ推定について扱い，近年盛んに研究がなされている Q 行列もデータから推定を行う探索的 DCM（e.g., Culpepper, 2019）についての詳細は扱わないこととした．

以下に本論文の構成を示す．第 2 章では，潜在クラスモデルベースの診断分類モデルの定式化および単調性制約（monotonicity constraints）について説明する．本稿では，最も一般性の高い DCM の一つである対数線形認知診断モデル（log linear cognitive diagnostic model [LCDM]; Henson et al., 2009) から，潜在クラスモデルベースの診断分類モデルの定式化を行う．LCDM のモデルパラメタは DCM の応用に際して理解しやすいパラメタであり，DCM というモデル族の特徴を把握しやすいと考えられる．なお，本稿では von Davier and Lee (2019) などの最新の成書に準じて，CDM ではなく DCM という用語を用いるが，LCDM はすでに定着したモデルの名称であるため本稿でもその名称を変更せずに用いる． 第 3 章では，最尤推定法に関係する推定法を概観する．特に，Expectation-Maximization（EM）アルゴリズムによる推定アルゴリズムおよび単調性制約を満足するための方法について概観する．更に，正則化推定法（regularized estimation method）についても示す．第 4 章ではベイズ推定法について概観する．古典的な MAP（maximum a posteriori）推定のみならず，MCMC による事後分布の近似法と変分ベイズによる事後分布の近似法についても述べる．第 5 章ではノンパラメトリック推定法について示す．第 6 章では，本稿の推定法のレビューにもとづき，DCM の推定法の選択について議論する．最後の第 7 章では，レビューした推定法から DCM のパラメタ推定について残された問題と今後の研究について議論を行う．

2. 診断分類モデルの定式化

2.1 記号の準備と測定モデル

本節では項目反応が 2 値である DCM の基礎的な定式化を行う．まず，あるテスト解答者 $i = 1, 2, \ldots, I \in \mathbb{N}$ の項目 $j = 1, 2, \ldots, J \in \mathbb{N}$ への反応を $X_{ij} \in \{0, 1\}$ とする．ここで，$X_{ij} = 1$ は正答反応，$X_{ij} = 0$ を誤答反応とする．DCM においては，テスト全体で $K \in \mathbb{N}$ 個のアトリビュート，すなわち問題の正答に必要な認知要素を想定する．アトリビュートは $\alpha_k \in \{0, 1\} (k = 1, 2, \ldots, K)$ であり，$\alpha_k = 1$ は k 番目のアトリビュートを習得している状態，$\alpha_k = 0$ は未習得の状態を意味する．項目反応およびアトリビュートは多値を想定することもある．K 個のアトリビュート習得の有無を並べたベクトル $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_K)^\top \in \{0, 1\}^K$ をアトリビュート習得パタンと呼ぶ．アトリビュート数が K であることから，ありうるアトリビュートの習得パタンの数は 2^K であり，アトリビュート習得パタンを添え字 $l = 1, 2, \ldots, L = 2^K$ によって区別して $\boldsymbol{\alpha}_l$ と書き，その要素を α_{lk} とする．また，テスト解答者 i のアトリビュート習得パタンを $\boldsymbol{\alpha}_i$ とする．

DCM においては，項目 j の正答に関係するアトリビュートを示す長さ K のベクトル $\boldsymbol{q}_j = (q_{j1}, q_{j2}, \ldots, q_{jK})^\top \in \{0, 1\}^K \backslash \mathbf{0}_K$ (ただし，$\mathbf{0}_K$ は 0 を K 個並べたベクトル) を $\boldsymbol{q}$ ベクトルと呼び，その k 番目の要素 $q_{jk} \in \{0, 1\}$ は $q_{jk} = 1$ ならば項目 j の正答にアトリビュート k が必要，$q_{jk} = 0$ であれば不要であることを示す．また，$\boldsymbol{q}_j^\top$ を第 j 行に持ち，サイズが $J \times K$ である行列 $\mathrm{Q} = (\boldsymbol{q}_1^\top, \boldsymbol{q}_2^\top, \ldots, \boldsymbol{q}_J^\top)^\top$ を Q 行列 (Tatsuoka, 1985) と呼ぶ．

この $\boldsymbol{q}$ ベクトル，アトリビュート習得パタンに加え，項目パラメタベクトル $\boldsymbol{\lambda}_j = (\lambda_{j0}, \lambda_{j1}, \ldots, \lambda_{j1\ldots K})^\top$ を用いて，LCDM の項目反応関数は，

$$P(X_{ij} = 1 | \boldsymbol{\lambda}_j, \boldsymbol{q}_j, \boldsymbol{\alpha}_i = \boldsymbol{\alpha}_l) = \frac{\exp(f(\boldsymbol{\lambda}_j, \boldsymbol{q}_j, \boldsymbol{\alpha}_i))}{1 + \exp(f(\boldsymbol{\lambda}_j, \boldsymbol{q}_j, \boldsymbol{\alpha}_i))} \tag{2.1}$$

と表される．ここで，$f(\boldsymbol{\lambda}_j, \boldsymbol{q}_j, \boldsymbol{\alpha}_i)$ は

$$\begin{aligned} f(\boldsymbol{\lambda}_j, \boldsymbol{q}_j, \boldsymbol{\alpha}_i) &= \log \frac{P(X_{ij} = 1 | \boldsymbol{\lambda}_j, \boldsymbol{q}_j, \boldsymbol{\alpha}_i = \boldsymbol{\alpha}_l)}{1 - P(X_{ij} = 1 | \boldsymbol{\lambda}_j, \boldsymbol{q}_j, \boldsymbol{\alpha}_i = \boldsymbol{\alpha}_l)} \\ &= \lambda_{j0} + \sum_{k=1}^{K} \lambda_{jk} q_{jk} \alpha_{ik} + \sum_{k=1}^{K} \sum_{k'<k} \lambda_{jkk'} q_{jk} q_{jk'} \alpha_{ik} \alpha_{ik'} + \cdots + \lambda_{j1\ldots K} \prod_{k=1}^{K} q_{jk} \alpha_{ik} \end{aligned} \tag{2.2}$$

と表される関数である．LCDM の項目パラメタベクトル $\boldsymbol{\lambda}_j$ は以下のパラメタを含んでいる．まず，λ_{j0} は切片パラメタであり，切片パラメタは当該の項目に必要なアトリビュートを全く習得していない場合のベースラインの正答確率を規定している．次に，λ_{jk} はアトリビュート k の主効果を表し，あるアトリビュートの正答確率への単独の効果を規定するパラメタである．さらに，$\lambda_{jkk'}$ はアトリビュート k と $k' (\neq k)$ の間の 1 次の交互作用パラメタおよび $\lambda_{j1\ldots K}$ は K 個のアトリビュートの間の $K - 1$ 次の交互作用パラメタであり，複数のアトリビュートを同時に習得している場合の効果を表している．実際には，すべての項目について 2^K 個のパラメタが想定されるわけではなく，$q_{jk} = 0$ を含む項の λ パラメタは推定できないため，項目 j のパラメタの個数は $2^{\sum_{k=1}^{K} q_{jk}} \leq 2^K$ である．

例として，$\boldsymbol{q}_j = (1, 1, 0)^\top, K = 3$ という項目に対する，LCDM の項目反応関数を具体的に書くことにする．まず，式 (2.2) から，LCDM の $f(\boldsymbol{\lambda}_j, \boldsymbol{q}_j, \boldsymbol{\alpha}_i)$ 関数は

$$f(\boldsymbol{\lambda}_j, \boldsymbol{q}_j, \boldsymbol{\alpha}_i) = \lambda_{j0} + \sum_{k=1}^{3} \lambda_{jk} q_{jk} \alpha_{ik} + \sum_{k=1}^{3} \sum_{k'<k} \lambda_{jkk'} q_{jk} q_{jk'} \alpha_{ik} \alpha_{ik'} + \lambda_{j123} \prod_{k=1}^{3} q_{jk} \alpha_{ik} \tag{2.3}$$

である．ここに，$q_{j1} = q_{j2} = 1$ および $q_{j3} = 0$ を代入すると，項目反応関数は

$$P(X_{ij}=1|\boldsymbol{\lambda}_j,\boldsymbol{q}_j=(1,1,0)^\top,\boldsymbol{\alpha}_i=\boldsymbol{\alpha}_l)=\frac{\exp(\lambda_{j0}+\lambda_{j1}\alpha_{i1}+\lambda_{j2}\alpha_{i2}+\lambda_{j12}\alpha_{i1}\alpha_{i2})}{1+\exp(\lambda_{j0}+\lambda_{j1}\alpha_{i1}+\lambda_{j2}\alpha_{i2}+\lambda_{j12}\alpha_{i1}\alpha_{i2})} \tag{2.4}$$

と縮約した形式で記述できる．この項目 j の項目パラメタの個数は $2^2=4$ となる．

さらに，ここで LCDM のモデルパラメタの数値例を示す．Templin and Hoffman (2013) は The Examination for the Certificate of Proficiency in English (ECPE) という英語力を測定するテストに LDCM を適用した例を示している．ECPE データには 28 の多枝選択項目が含まれており，アトリビュート数は $K=3$ で，1. 形態的統語的規則 (morphosyntactic rules)，2. 一貫性規則 (cohesive rules)，3. 語彙規則 (lexical rules) が想定されていた．また，分析のためのサンプルサイズは 2,922 であった．ECPE データの 1 番目の項目には，形態的統語的規則および一貫性規則が必要とされているため，$\boldsymbol{q}_1=(1,1,0)^\top$ であった．Templin and Hoffman (2013) の Table 1 から，項目 1 の項目パラメタの推定値は，$\lambda_{10}=0.835$，$\lambda_{11}=0.000$，$\lambda_{12}=0.600$ および $\lambda_{112}=1.222$ であった．これらの結果から，例えば，$\lambda_{11}=0.000$ であるため，形態的統語的規則の単独の習得は項目の正答に寄与しないこと，$\lambda_{112}=1.222$ であることから形態的統語的規則と一貫性規則のどちらも習得していることによる交互作用効果が項目の正答に寄与しているといったことがわかる．

LCDM は項目パラメタに制約をかけることによって，さまざまな DCM のサブモデルを表現できる．例えば，最も基本的な DCM である DINA モデルは切片と当該項目で測定しているアトリビュートの添字集合 $\mathcal{K}=\{k;q_{jk}=1,k=1,2,\ldots,K\}$ についての最高次の交互作用パラメタのみを含めて，

$$f(\boldsymbol{\lambda}_j,\boldsymbol{q}_j,\boldsymbol{\alpha}_i)=\lambda_{j0}+\lambda_{j\mathcal{K}}\prod_{k\in\mathcal{K}}\alpha_{ik} \tag{2.5}$$

と表される．DINA モデルでは，2 つの項目パラメタを

$$g_j=\frac{\exp(\lambda_{j0})}{1+\exp(\lambda_{j0})}, \tag{2.6}$$

$$1-s_j=\frac{\exp(\lambda_{j0}+\lambda_{j\mathcal{K}}\prod_{k\in\mathcal{K}}\alpha_{ik})}{1+\exp(\lambda_{j0}+\lambda_{j\mathcal{K}}\prod_{k\in\mathcal{K}}\alpha_{ik})} \tag{2.7}$$

と表す．ここで，g_j は当て推量 (guessing) パラメタ，s_j はスリップ (slipping) パラメタと呼ばれる．すなわち，g_j は項目 j の正答に必要なアトリビュートのうち，少なくとも一つが未習得であるアトリビュート習得パタンのテスト解答者が当該項目に正答する確率であり，s_j は項目 j の正答に必要なアトリビュートをすべて習得しているアトリビュート習得パタンのテスト解答者が当該項目に誤答する確率である． また，切片と主効果項のみを含めたモデルは compensatory reduced unified model (C-RUM; Rupp et al., 2010) と呼ばれ，

$$f(\boldsymbol{\lambda}_j,\boldsymbol{q}_j,\boldsymbol{\alpha}_i)=\lambda_{j0}+\sum_{k=1}^{K}\lambda_{jk}q_{jk}\alpha_{ik} \tag{2.8}$$

と表される．

さて，LCDM においては，$\boldsymbol{\lambda}$ がモデルパラメタ，$\boldsymbol{\alpha}$ がカテゴリカルで多次元の潜在変数とみなすことができる．この意味では，LDCM はで連続的な多次元の潜在変数を持つ多次元項目反応理論モデル (multidimensional item response theory models) と関係があるモデルとしても考えることができる．しかし，$\boldsymbol{\lambda}$ ではなく，アトリビュート習得パタンごとの項目反応確率をモデルのパラメタとみなすことで，本稿の主眼であるパラメタ推定法を表現しやすくなる．さきほどの $\boldsymbol{q}_j=(1,1,0)^\top$ という項目について，4 つの (条件付き) 正答反応確率パラメタを $0\leq\theta_{jh}\leq 1(h=1,2,3,4)$ と書くとすると，

$$\theta_{j1} = \frac{\exp(\lambda_{j0})}{1+\exp(\lambda_{j0})}, \tag{2.9}$$

$$\theta_{j2} = \frac{\exp(\lambda_{j0}+\lambda_{j1})}{1+\exp(\lambda_{j0}+\lambda_{j1})}, \tag{2.10}$$

$$\theta_{j3} = \frac{\exp(\lambda_{j0}+\lambda_{j2})}{1+\exp(\lambda_{j0}+\lambda_{j2})}, \tag{2.11}$$

$$\theta_{j4} = \frac{\exp(\lambda_{j0}+\lambda_{j1}+\lambda_{j2}+\lambda_{j12})}{1+\exp(\lambda_{j0}+\lambda_{j1}+\lambda_{j2}+\lambda_{j12})} \tag{2.12}$$

となる．ここで，θ_{j1} はアトリビュート習得パタン $(0,0,0)^\top$ と $(0,0,1)^\top$ の正答確率に対応するパラメタと考えることができる．同様に，$\theta_{j2},\theta_{j3},\theta_{j4}$ はそれぞれ，$(1,0,0)^\top$ と $(1,0,1)^\top$，$(0,1,0)^\top$ と $(0,1,1)^\top$，$(1,1,0)^\top$ と $(1,1,1)^\top$ に対応する正答確率パラメタを表している．このことから，DCM はアトリビュート習得パタンを潜在クラスとみなした制約付き潜在クラスモデルと考えることができる．

上記の例からわかるように，一般的な DCM において，$q_{jk}=0$ は項目 j においてアトリビュート k の習得の有無の情報を区別することができないことを意味している．この意味で，$\boldsymbol{q}$ ベクトルはアトリビュート習得パタンを項目特有のアトリビュート習得パタンに縮約する役割を持っていると考えることができる．ここで，$\boldsymbol{q}$ ベクトルが区別できるアトリビュート習得パタンを $\boldsymbol{\alpha}^*_{jh}$ と書くことにする．$\boldsymbol{q}_j=(1,1,0)^\top$ という $\boldsymbol{q}$ ベクトルから，$\boldsymbol{\alpha}^*_{j1}=(0,0,*)^\top$，$\boldsymbol{\alpha}^*_{j2}=(0,1,*)^\top$，$\boldsymbol{\alpha}^*_{j3}=(1,0,*)^\top$，$\boldsymbol{\alpha}^*_{j4}=(1,1,*)^\top$ という 4 種類のアトリビュート習得パタンが区別できることがわかる．“$*$” は，そのアトリビュートの習得の有無を無視することを意味する．ここから，ある項目 j において同じ正答確率を持つパタンを示す

$$\mathcal{I}(\boldsymbol{\alpha}^*_{jh}=\boldsymbol{\alpha}_l)=\begin{cases}1 & \alpha^*_{jhk}=\alpha_{lk},\forall k\in\{k;q_{jk}=1,k=1,\ldots,K\}\\ 0 & \text{Otherwise}\end{cases} \tag{2.13}$$

という指示関数 $\mathcal{I}(\cdot)$ を導入する．この記号と条件付き正答確率パラメタ $\boldsymbol{\theta}_j=\{\theta_{jh}\}_{h=1}^{H_j}, H_j=2^{\sum_{k=1}^K q_{jk}}$ をもちいることで，DCM の条件付き尤度関数は

$$P(X_{ij}=x_{ij}|\boldsymbol{\theta}_j,\boldsymbol{\alpha}_i=\boldsymbol{\alpha}_l)=\prod_{h=1}^{H_j}\prod_{l=1}^{L}[\theta_{jh}^{x_{ij}}(1-\theta_{jh})^{1-x_{ij}}]^{\mathcal{I}(\boldsymbol{\alpha}^*_{jh}=\boldsymbol{\alpha}_l)\mathcal{I}(\boldsymbol{\alpha}_i=\boldsymbol{\alpha}_l)} \tag{2.14}$$

と書くことができる．ただし，$\mathcal{I}(\boldsymbol{\alpha}_i=\boldsymbol{\alpha}_l)$ は $\boldsymbol{\alpha}_i$ と $\boldsymbol{\alpha}_l$ の要素がすべて等しい場合に 1，そうでなければ 0 をとる関数として，先に導入した指示関数とは引数の違いで区別する．式(2.14)から，テスト解答者 i の項目反応パタンベクトル $\boldsymbol{X}_i=(X_{i1},X_{i2},\ldots,X_{iJ})^\top$ の実現値 $\boldsymbol{x}_i=(x_{i1},x_{i2},\ldots,x_{iJ})^\top$ を観測したときのパラメタの尤度は

$$P(\boldsymbol{X}_i=\boldsymbol{x}_i|\boldsymbol{\Theta},\boldsymbol{\alpha}_i=\boldsymbol{\alpha}_l)=\prod_{j=1}^{J}\prod_{h=1}^{H_j}\prod_{l=1}^{L}[\theta_{jh}^{x_{ij}}(1-\theta_{jh})^{1-x_{ij}}]^{\mathcal{I}(\boldsymbol{\alpha}^*_{jh}=\boldsymbol{\alpha}_l)\mathcal{I}(\boldsymbol{\alpha}_i=\boldsymbol{\alpha}_l)} \tag{2.15}$$

となる．ただし，$\boldsymbol{\Theta}=\{\boldsymbol{\theta}_j\}_{j=1}^J$ は項目パラメタの集合である．さらに，$\boldsymbol{\alpha}_i$ が混合比率パラメタ $\boldsymbol{\pi}=(\pi_1,\pi_2,\ldots,\pi_L)^\top$ (ただし，$0\le\pi_l\le1,\sum_l\pi_l=1$)であるカテゴリカル分布

$$P(\boldsymbol{\alpha}_i=\boldsymbol{\alpha}_l|\boldsymbol{\pi})=\prod_{l=1}^{L}\pi_l^{\mathcal{I}(\boldsymbol{\alpha}_i=\boldsymbol{\alpha}_l)} \tag{2.16}$$

に従うとする．式(2.15)および(2.16)から，テスト解答者 i の項目反応パタンベクトルを得たときの周辺尤度は

$$P(\boldsymbol{X}_i=\boldsymbol{x}_i|\boldsymbol{\Theta},\boldsymbol{\pi})=\sum_{l=1}^{L}P(\boldsymbol{X}_i=\boldsymbol{x}_i|\boldsymbol{\Theta},\boldsymbol{\alpha}_i=\boldsymbol{\alpha}_l)P(\boldsymbol{\alpha}_i=\boldsymbol{\alpha}_l|\boldsymbol{\pi}) \tag{2.17}$$
$$=\sum_{l=1}^{L}\left\{\pi_l\prod_{j=1}^{J}\prod_{h=1}^{H_j}[\theta_{jh}^{x_{ij}}(1-\theta_{jh})^{1-x_{ij}}]^{\mathcal{I}(\boldsymbol{\alpha}_{jh}^*=\boldsymbol{\alpha}_l)}\right\}$$

と書くことができる．さらに，すべてのテスト解答者が独立にサンプリングされているとし，項目反応行列を $\mathrm{X}=(\boldsymbol{x}_1^\top,\boldsymbol{x}_2^\top,\ldots,\boldsymbol{x}_I^\top)^\top$ とすると，周辺尤度関数は

$$L(\boldsymbol{\Theta},\boldsymbol{\pi})=P(\mathrm{X}|\boldsymbol{\Theta},\boldsymbol{\pi})=\prod_{i=1}^{I}P(\boldsymbol{X}_i=\boldsymbol{x}_i|\boldsymbol{\Theta},\boldsymbol{\pi}) \tag{2.18}$$
$$=\prod_{i=1}^{I}\sum_{l=1}^{L}\left\{\pi_l\prod_{j=1}^{J}\prod_{h=1}^{H_j}[\theta_{jh}^{x_{ij}}(1-\theta_{jh})^{1-x_{ij}}]^{\mathcal{I}(\boldsymbol{\alpha}_{jh}^*=\boldsymbol{\alpha}_l)}\right\}$$

となる．DCM のモデルパラメタの推定は上記の周辺尤度関数を用いる方法が基本となる．

具体的な推定法について記述を行う前に，DCM を利用する際の一般的な注意点を述べる．まず，Q 行列はテストで測定したいアトリビュートについての専門知識と項目の分析から設定される．ただし，Q 行列の誤特定は，アトリビュート習得パタンの推定にバイアスを生じさせることが知られているため（e.g., Rupp and Templin, 2008a），慎重に設定する必要がある．そのため，項目反応データから当該のテストの Q 行列をデータから推定する方法も近年活発に開発されている（e.g., Chen et al., 2018）．さらに，Q 行列の設定はパラメタの識別性（identification; Xu and Shang, 2018）を規定するため重要である．また，アトリビュートの間の習得の順序関係を表すアトリビュート階層構造（attribute hierarchy structure; e.g., Leighton et al., 2004）を想定する場合には，アトリビュート習得パタンの数は 2^K よりも少なくなる．本稿のモデルの定式化はアトリビュートの間の関係を特段仮定しないものであった．これをもとに，モデルのパラメタ $\{\boldsymbol{\Theta},\boldsymbol{\pi}\}$ の推定法を示す．

2.2 単調性制約

単調性制約は，あるアトリビュート習得パタンにおける項目の正答確率よりも，そのアトリビュート習得パタンで未習得のアトリビュートを習得している別の習得パタンの正答確率のほうが高いか少なくとも等しい，という制約である．前述の $\boldsymbol{q}_j=(1,1,0)^\top$ という項目のパラメタの具体的な単調性制約は以下のようになる．式(2.9)から(2.12)で表される正答確率は，それぞれ $\boldsymbol{\alpha}_{j1}^*=(0,0,*)^\top$，$\boldsymbol{\alpha}_{j2}^*=(1,0,*)^\top$，$\boldsymbol{\alpha}_{j3}^*=(0,1,*)^\top$ および $\boldsymbol{\alpha}_{j4}^*=(1,1,*)^\top$ という習得パタンに対する正答確率であるため

$$0\le\theta_{j1}\le\theta_{j2}\le\theta_{j4}\le 1,$$
$$0\le\theta_{j1}\le\theta_{j3}\le\theta_{j4}\le 1$$

という順序関係を満足する必要があり，このパラメタの順序関係が単調性制約である．単調性制約から，当該の項目の正答に関連するアトリビュートを一つも習得していないパタン（ここでは $\boldsymbol{\alpha}_{j1}^*=(0,0,*)^\top$）は最も低い正答確率であり，すべてのアトリビュートを習得しているパタン（$\boldsymbol{\alpha}_{j4}^*=(1,1,*)^\top$）は最も高い正答確率になることがわかる．

単調性を形式的に表現するために，アトリビュート習得パタンの「順序」を以下のように定義する．すべての $k=1,\ldots,K$ に対して，$\alpha_{lk}\ge\alpha_{l'k}$ であれば $\boldsymbol{\alpha}_l\succeq\boldsymbol{\alpha}_{l'}$ とし，$\boldsymbol{\alpha}_l\succeq\boldsymbol{\alpha}_{l'}$ を満足し，かつ，いくつかの k に対して $\alpha_{lk}>\alpha_{l'k}$ であれば $\boldsymbol{\alpha}_l\succ\boldsymbol{\alpha}_{l'}$ とする．$\boldsymbol{\alpha}_{jh}^*$ についても同様に順序を考えることとする．このとき，上記の $\boldsymbol{q}_j=(1,1,0)^\top$ に対して

$$\boldsymbol{\alpha}^*_{j4} \succ \boldsymbol{\alpha}^*_{j2} \succ \boldsymbol{\alpha}^*_{j1},$$
$$\boldsymbol{\alpha}^*_{j4} \succ \boldsymbol{\alpha}^*_{j3} \succ \boldsymbol{\alpha}^*_{j1}$$

である．この表記を用いることで，単調性制約は $\boldsymbol{\alpha}^*_{jh} \succ \boldsymbol{\alpha}^*_{jh'}$ であれば，$\theta_{jh} \geq \theta_{jh'}$ を満足することと表現できる．

上記の単調性制約はやや強い関係であり，理論研究においては Xu（2017）や Xu and Shang（2018）による

$$\max_{\boldsymbol{\alpha};\boldsymbol{\alpha}\succeq\boldsymbol{q}_j} \theta_{j,\boldsymbol{\alpha}} = \min_{\boldsymbol{\alpha};\boldsymbol{\alpha}\succeq\boldsymbol{q}_j} \theta_{j,\boldsymbol{\alpha}} \geq \theta_{j,\boldsymbol{\alpha}'} \geq \theta_{j,\mathbf{0}_K}, \forall \boldsymbol{\alpha}' \not\succeq \boldsymbol{q}_j, \tag{2.19}$$

という単調性の定義を用いることもある．この定義においては，ある項目 j で測定されているアトリビュートをすべて測定しているアトリビュート習得パタンの正答確率が全て同じで最も高く，また，当該項目の正答に関係するアトリビュートを一つも習得していないパタンが最も低い正答確率になり，それ以外のアトリビュート習得パタン（$\boldsymbol{\alpha}'$ であらわされる）の正答確率はそれらの 2 つの正答確率の間であるという関係を表している．ここで，$\boldsymbol{\alpha}'$ に対応する複数のアトリビュート習得パタンの間の正答確率の間には，大小関係を想定していない点に注意が必要である．後述の単調性を満足するためのアルゴリズムは，前述のより厳しい単調性を満足するものであるが，こちらのややゆるい単調性を満足するように変更することも容易にできる．

3. 最尤推定法

本章では，モデルの尤度を用いた推定法として，同時最尤推定法（joint likelihood estimation method），周辺最尤推定法（marginal likelihood estimation method），正則化推定法を示す．周辺最尤推定法については，EM アルゴリズムを用いた方法を示し，パラメタの推定の標準誤差の推定法および単調性制約を満足する推定値を得るための方法も示す．

3.1 同時最尤推定法

DCM においては，アトリビュート習得パタン $\boldsymbol{\alpha}_i$ が潜在変数であり，テスト解答者のアトリビュート習得パタンベクトルが観測されていれば，$\mathcal{I}(\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_l)$ が計算でき，完全データの尤度を構成できる．後述の周辺最尤推定法は，尤度から潜在変数を周辺化して消去するため，アトリビュート数が大きくなったりサンプルサイズが大きくなった場合に計算の負荷が大きくなる傾向がある．一方，同時最尤推定法は単純である完全データの尤度を用いるため，周辺最尤推定法と比較して推定の計算アルゴリズムが簡単になり，効率的な推定が可能である．しかし，同時最尤推定法は一般にパラメタ推定値に一致性がないことから，潜在変数モデルにおいてはあまり用いられてきていない方法である．Chiu et al.（2016）は，外部で推定された一致性のあるアトリビュート習得パタンを初期値として用いることにより，同時最尤推定でも一致性のある項目パラメタを得ることができることを証明し，この問題を解決した．これにより，効率的な計算アルゴリズムを活用して統計的に望ましい性質をもった推定値を得ることができるため，大規模データやアトリビュート数が多い状況で DCM を利用しやすくなったと考えられる．

同時最尤推定法の具体的な方法を以下に示す．式（2.16），（2.15）から，完全データの尤度は

$$L(\boldsymbol{\mathcal{A}}, \boldsymbol{\Theta}, \boldsymbol{\pi}) = P(\mathrm{X}, \boldsymbol{\mathcal{A}} | \boldsymbol{\Theta}, \boldsymbol{\pi}) = \prod_{i=1}^{I} \prod_{l=1}^{L} \left\{ \pi_l \prod_{j=1}^{J} \prod_{h=1}^{H_j} [\theta_{jh}^{x_{ij}} (1-\theta_{jh})^{1-x_{ij}}]^{\mathcal{I}(\boldsymbol{\alpha}^*_{jh} = \boldsymbol{\alpha}_l)} \right\}^{\mathcal{I}(\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_l)} \tag{3.1}$$

である．ただし，I 人分のアトリビュート習得パタンの集合を $\boldsymbol{\mathcal{A}} = \{\boldsymbol{\alpha}_i\}_{i=1}^{I}$ とした．同時最尤

推定法は

$$\{\hat{\boldsymbol{\mathcal{A}}}, \hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\pi}}\} = \underset{\boldsymbol{\mathcal{A}}, \boldsymbol{\Theta}, \boldsymbol{\pi}}{\operatorname{argmax}} \{\log L(\boldsymbol{\mathcal{A}}, \boldsymbol{\Theta}, \boldsymbol{\pi})\} \tag{3.2}$$

と式(3.1)の完全データの尤度あるいはその対数尤度を最大化する潜在変数 $\hat{\boldsymbol{\mathcal{A}}}$ およびモデルパラメタ $\{\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\pi}}\}$ を推定値とする方法である．

同時最尤推定法においては，モデルパラメタとテスト解答者パラメタの更新を反復する方法で完全データの尤度を最大化することで，パラメタの推定値を得ることができる．より具体的には，まず後述のノンパラメトリック推定法で得られたアトリビュート習得パタンの一致推定値を初期値とし，アトリビュート習得パタンを既知としたもとで，項目ごとの尤度(式(2.14)をテスト解答者について積をとることで得られる尤度)を最大化する項目パラメタを求める．また，混合比率パラメタは式(2.16)を用いて更新する．さらに，これらのパラメタを既知として，式(2.15)を最大化するアトリビュート習得パタンを得る．これらの項目パラメタとアトリビュート習得パタンの更新を適当な収束基準を満足するまで反復する，というのが同時最尤推定法のパラメタ推定アルゴリズムである．Chiu et al. (2016)は，アトリビュート習得パタンが既知の状況で，LCDM パラメタを切片パラメタから高次の交互作用に向けて再帰的に更新する比較的単純なパラメタ更新規則を示した．

上記のパラメタ更新アルゴリズムにおいては，モデルパラメタの更新とテスト解答者パラメタの更新の際に，項目ごとあるいはテスト解答者ごとのパラメタに注目した更新が可能であり，更新の並列化が可能となる．このように，効率的なパラメタの更新アルゴリズムを利用することができる点が同時最尤推定法の利点である．Chiu et al. (2016)は，シミュレーションにより，同時最尤推定法のほうが周辺最尤推定法よりも高速にパラメタ推定が完了することと，同時最尤推定法の推定値が一致的であることを確認した．ただし，Chiu et al. (2016)の推定アルゴリズムはパラメタの単調性を満足するものではない．しかし，その推定アルゴリズムにおいて，項目パラメタの更新を行う際に，後述する周辺最尤推定法で述べる EM アルゴリズムの M ステップにおいて単調性を満足するための最適化法が利用できる．

3.2 周辺最尤推定法

周辺最尤推定法は，欠測データを含むデータ分析や潜在変数モデルのパラメタ推定法として標準的に用いられる方法であり，DCM のパラメタ推定においても基礎的な方法として位置づけられる．式(2.18)の周辺尤度と式(3.1)の完全データの尤度の間には

$$L(\boldsymbol{\Theta}, \boldsymbol{\pi}) = \sum_{\boldsymbol{\mathcal{A}}} L(\boldsymbol{\mathcal{A}}, \boldsymbol{\Theta}, \boldsymbol{\pi}) \tag{3.3}$$

という関係が成り立つ．ここで $\sum_{\boldsymbol{\mathcal{A}}}$ は $\boldsymbol{\mathcal{A}}$ が取りうるすべてのパタンについての和を取ることを意味する．実際には，局所独立の仮定があるため，各テスト解答者について L 個のアトリビュート習得パタンごとに式(2.14)を用いて式(2.15)を評価することで，周辺尤度の値を評価する．

周辺最尤推定法は上記の周辺尤度を用いて，

$$\{\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\pi}}\} = \underset{\boldsymbol{\Theta}, \boldsymbol{\pi}}{\operatorname{argmax}} \{\log L(\boldsymbol{\Theta}, \boldsymbol{\pi})\} = \underset{\boldsymbol{\Theta}, \boldsymbol{\pi}}{\operatorname{argmax}} \left\{ \log \sum_{\boldsymbol{\mathcal{A}}} L(\boldsymbol{\mathcal{A}}, \boldsymbol{\Theta}, \boldsymbol{\pi}) \right\} \tag{3.4}$$

と周辺尤度を最大化するモデルパラメタを推定量とする方法である．周辺最尤推定法におけるモデルパラメタは，最尤推定法の一般論から一致的である．周辺最尤推定法においてはアトリビュート習得パタンが周辺化されているため，モデルパラメタの推定値 $\{\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\pi}}\}$ を用いて事後

的に推定される．

このため，アトリビュート習得パタン $\hat{\boldsymbol{\mathcal{A}}} = \{\hat{\boldsymbol{\alpha}}_i\}_{i=1}^{I}$ の推定法にはいくつかの選択肢がある．まず，式(2.15)の条件付き尤度を用いた方法として，テスト解答者 i の項目反応パタン $\boldsymbol{x}_i$ から

$$\hat{\boldsymbol{\alpha}}_i^{ML} = \underset{\boldsymbol{\alpha}_i}{\operatorname{argmax}}\{P(\boldsymbol{X}_i = \boldsymbol{x}_i|\hat{\boldsymbol{\Theta}}, \boldsymbol{\alpha}_i = \boldsymbol{\alpha}_l)\} \tag{3.5}$$

はアトリビュート習得パタンの最尤推定値である．

アトリビュート習得パタンの事後確率を

$$\hat{\boldsymbol{\alpha}}_i^{MAP} = \underset{\boldsymbol{\alpha}_i}{\operatorname{argmax}}\{P(\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_l|\boldsymbol{x}_i, \hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\pi}})\} \tag{3.6}$$

と最大化するアトリビュート習得パタンが事後確率最大化(MAP)推定値である．ここで，$P(\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_l|\boldsymbol{x}_i, \hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\pi}})$ はテスト解答者 i のアトリビュート習得パタン $\boldsymbol{\alpha}_i$ が $\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_l$ である事後確率であり，

$$\begin{aligned} P(\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_l|\boldsymbol{x}_i, \hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\pi}}) &= \frac{P(\boldsymbol{X}_i = \boldsymbol{x}_i|\hat{\boldsymbol{\Theta}}, \boldsymbol{\alpha}_i = \boldsymbol{\alpha}_l)P(\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_l|\hat{\boldsymbol{\pi}})}{\sum_{l'=1}^{L} P(\boldsymbol{X}_i = \boldsymbol{x}_i|\hat{\boldsymbol{\Theta}}, \boldsymbol{\alpha}_i = \boldsymbol{\alpha}_{l'})P(\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_{l'}|\hat{\boldsymbol{\pi}})} \\ &= \frac{\hat{\pi}_l \prod_{j=1}^{J}\prod_{h=1}^{H_j}[\hat{\theta}_{jh}^{x_{ij}}(1-\hat{\theta}_{jh})^{1-x_{ij}}]^{\mathcal{I}(\boldsymbol{\alpha}_{jh}^{*}=\boldsymbol{\alpha}_l)}}{\sum_{l'=1}^{L}\{\hat{\pi}_{l'}\prod_{j=1}^{J}\prod_{h=1}^{H_j}[\hat{\theta}_{jh}^{x_{ij}}(1-\hat{\theta}_{jh})^{1-x_{ij}}]^{\mathcal{I}(\boldsymbol{\alpha}_{jh}^{*}=\boldsymbol{\alpha}_{l'})}\}} \end{aligned} \tag{3.7}$$

と計算される．期待事後推定(expected-a-posteriori[EAP])法は，各テスト解答者の各アトリビュート習得の式(3.7)による事後確率を用いた事後期待値を用いる．アトリビュート習得の事後期待値は

$$\begin{aligned} \mathbb{E}_{P(\boldsymbol{\alpha}_i|\boldsymbol{x}_i,\hat{\boldsymbol{\Theta}},\hat{\boldsymbol{\pi}})}[\alpha_{ik}] &= \sum_{l=1}^{L}\alpha_{lk}P(\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_l|\boldsymbol{x}_i, \hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\pi}}) \\ &= P(\alpha_{ik} = \alpha_{lk}|\boldsymbol{x}_i, \hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\pi}}) \end{aligned} \tag{3.8}$$

である．ここで，$\mathbb{E}_{P(\boldsymbol{\alpha}_i|\boldsymbol{x}_i,\hat{\boldsymbol{\Theta}},\hat{\boldsymbol{\pi}})}[\cdot]$ はアトリビュート習得パタンの事後確率による期待値である．これはアトリビュート習得確率としても解釈でき，アトリビュート習得確率を DCM の結果として報告することもある．ここから，テスト解答者 i の k 番目のアトリビュート習得の EAP 推定値は，

$$\alpha_{ik}^{EAP} = \begin{cases} 1 & P(\alpha_{ik} = \alpha_{lk}|\boldsymbol{x}_i, \hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\pi}}) \geq 0.5 \\ 0 & P(\alpha_{ik} = \alpha_{lk}|\boldsymbol{x}_i, \hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\pi}}) < 0.5 \end{cases} \tag{3.9}$$

であり，アトリビュート習得パタンの EAP は $\boldsymbol{\alpha}_i^{EAP} = (\alpha_{i1}^{EAP}, \ldots, \alpha_{ik}^{EAP}, \ldots, \alpha_{iK}^{EAP})^{\top}$ である．EAP 推定を行う場合には，閾値を設定する必要があるが，ここでは 0.5 とした．

3.2.1 EM アルゴリズム

周辺尤度の最大化を実行するためのアルゴリズムとして，EM アルゴリズムが利用できる．EM アルゴリズムを用いた周辺尤度の最大化による最尤推定値を得る方法を MML-EM (marginal maximum likelihood with EM algorithm)と呼ぶ．EM アルゴリズムは，まずモデルパラメタの適当な初期値 $\{\boldsymbol{\Theta}^{(0)}, \boldsymbol{\pi}^{(0)}\}$ を定め，以下の E-step(expectation step)と M-step(maximization step)を反復して，対数周辺尤度を最大化するパラメタ値を求める．本節では上付き添字 $(t), (t = 1, 2, \ldots, T)$ をアルゴリズムの反復時点を表すものとする．

まず，すべてのテスト解答者 i について，アトリビュート習得パタン $\boldsymbol{\alpha}_i$ が $\boldsymbol{\alpha}_l$ である事後分布による期待値 $\gamma_{il}^{(t)} = \mathbb{E}_{P(\boldsymbol{\alpha}_i|\boldsymbol{x}_i,\boldsymbol{\Theta}^{(t-1)},\boldsymbol{\pi}^{(t-1)})}[\mathcal{I}(\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_l)]$ を求める．これは，式(3.7)で計算さ

れる事後確率に等しい．

次に式(3.1)の完全データの尤度を用いて，t 回目の反復において M-step で最適化する関数 $Q(\{\boldsymbol{\Theta},\boldsymbol{\pi}\}|\{\boldsymbol{\Theta}^{(t-1)},\boldsymbol{\pi}^{(t-1)}\})$ は

$$\begin{aligned} Q(\{\boldsymbol{\Theta},\boldsymbol{\pi}\}|\{\boldsymbol{\Theta}^{(t)},\boldsymbol{\pi}^{(t)}\}) &= \mathbb{E}_{P(\boldsymbol{\mathcal{A}}|\boldsymbol{X},\boldsymbol{\Theta}^{(t)},\boldsymbol{\pi}^{(t)})}[\log L(\boldsymbol{\mathcal{A}},\boldsymbol{\Theta},\boldsymbol{\pi})] \\ &= \sum_{i=1}^{I}\sum_{l=1}^{L}\left[\gamma_{il}^{(t)}\left\{\log\pi_l + \sum_{j=1}^{J}\sum_{h=1}^{H_j}\mathcal{I}(\boldsymbol{\alpha}_{jh}^{*}=\boldsymbol{\alpha}_l)(x_{ij}\log\theta_{jh}+(1-x_{ij})\log(1-\theta_{jh}))\right\}\right] \end{aligned} \tag{3.10}$$

と計算される．ここで，$\mathbb{E}_{P(\boldsymbol{\mathcal{A}}|\boldsymbol{X},\boldsymbol{\Theta}^{(t-1)},\boldsymbol{\pi}^{(t-1)})}[\cdot]$ はデータ X 及びパラメタ $\boldsymbol{\Theta}^{(t-1)},\boldsymbol{\pi}^{(t-1)}$ を所与としたときの，アトリビュート習得パタン集合 $\boldsymbol{\mathcal{A}}$ の事後確率による期待値である．

$Q(\{\boldsymbol{\Theta},\boldsymbol{\pi}\}|\{\boldsymbol{\Theta}^{(t-1)},\boldsymbol{\pi}^{(t-1)}\})$ を各パラメタについて偏微分し，$\sum_l \pi_l = 1$ という条件に注意しつつ，0 とおいた方程式を解くことで，M-step における $\theta_{jh}^{(t)}$ と $\pi_l^{(t)}$ の更新式

$$\theta_{jh}^{(t)} = \frac{\sum_{i=1}^{I}\sum_{l=1}^{L}\mathcal{I}(\boldsymbol{\alpha}_{jh}^{*}=\boldsymbol{\alpha}_l)\gamma_{il}x_{ij}}{\sum_{i=1}^{I}\sum_{l=1}^{L}\mathcal{I}(\boldsymbol{\alpha}_{jh}^{*}=\boldsymbol{\alpha}_l)\gamma_{il}}, \tag{3.11}$$

$$\pi_l^{(t)} = \frac{\sum_{i=1}^{I}\gamma_{il}}{I} \tag{3.12}$$

を得る．更新されたパラメタを用いて，E-step として新たに $\gamma_{il}^{(t+1)}$ を計算することで Q 関数を更新し，さらにモデルパラメタを更新するというステップを適当な収束基準を満足するまで繰り返す．なお，この更新式においては，正答確率パラメタの単調性制約を満たしていない点に注意が必要である．

3.2.2 単調性制約を満足するためのアルゴリズム

Hong et al. (2016)は単調性制約を満足するためのアルゴリズムとして，Upward アルゴリズムと Downward アルゴリズムの 2 つを提案している．これは，EM アルゴリズムなどの最適化の過程において単調性を犯す推定値が得られてしまった場合に，それらを補正するアルゴリズムである．まず，Upward アルゴリズムは，M-step の際に，$\boldsymbol{\alpha}_{jh}^{*} \succ \boldsymbol{\alpha}_{jh'}^{*}$ に対して $\theta_{jh} < \theta_{jh'}$ となった場合に，θ_{jh} を

$$\theta_{jh}^{(t)} = \max\{\theta_{jh'}^{(t)}|\boldsymbol{\alpha}_{jh}^{*} \succ \boldsymbol{\alpha}_{jh'}^{*}\} \tag{3.13}$$

とする操作を導入することで単調性を満足するように推定を行う．Downward アルゴリズムは，逆に，$\boldsymbol{\alpha}_{jh}^{*} \succ \boldsymbol{\alpha}_{jh'}^{*}$ に対して $\theta_{jh} < \theta_{jh'}$ となった場合に，$\theta_{jh'}$ を

$$\theta_{jh'}^{(t)} = \min\{\theta_{jh}^{(t)}|\boldsymbol{\alpha}_{jh}^{*} \succ \boldsymbol{\alpha}_{jh'}^{*}\} \tag{3.14}$$

とする操作を用いる．

この他，Ma and Jiang (2021)は単調性制約を満足するために，制約を表す行列 $\boldsymbol{C}$ を導入し，

$$\boldsymbol{C}\boldsymbol{\theta}_j = \begin{pmatrix} -1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ -1 & 0 & 0 & 1 \\ 0 & -1 & 0 & 1 \\ 0 & 0 & -1 & 1 \end{pmatrix} \times \begin{pmatrix} \theta_{j1} \\ \theta_{j2} \\ \theta_{j3} \\ \theta_{j4} \end{pmatrix} \geq \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} = \mathbf{0}_{H_j} \tag{3.15}$$

というパラメタの差の非負値制約を課して，M-step の項目パラメタについての最適化を実行する方法を提案している．項目反応確率への直接的な単調性制約は，LCDM パラメタにおけるパラメタ制約に変換することができる．LCDM およびその下位モデルについての制約の具体的

な表現は Rupp et al. (2010)で詳細に解説されている．さらに，Rupp et al. (2010)や Templin and Hoffman (2013)では一般的な潜在変数モデルのパラメタ推定が可能な Mplus というソフトウェアを用いて単調性制約を満足する推定値を得るためモデルの記述方法を示している．

3.2.3 標準誤差の推定

標準誤差はフィッシャー情報行列を用いて推定される．Philipp et al. (2018)はフィッシャー情報行列 $\mathcal{I}_{\boldsymbol{\vartheta}}$ (ただし，$\boldsymbol{\vartheta}$ は $\vartheta_p \in \{\boldsymbol{\Theta}, \boldsymbol{\pi}\}, p = 1, \ldots, P$ を要素とする長さが P のベクトル)の逆行列により，パラメタ推定値 $\hat{\boldsymbol{\vartheta}}$ の漸近分散共分散行列 $V_{\boldsymbol{\vartheta}} = \mathcal{I}_{\boldsymbol{\vartheta}}^{-1}$ を推定する方法を示した．パラメタの推定値の漸近標準誤差は，$V_{\boldsymbol{\vartheta}}$ の対角要素の平方根により推定される．さて，式(2.17)の尤度から得られるスコア関数 $\psi(\boldsymbol{\vartheta}; \boldsymbol{x}) = \left(\frac{\partial \log(P(\boldsymbol{X}=\boldsymbol{x}|\boldsymbol{\vartheta}))}{\vartheta_1}, \ldots, \frac{\partial \log(P(\boldsymbol{X}=\boldsymbol{x}|\boldsymbol{\vartheta}))}{\vartheta_p}, \ldots, \frac{\partial \log(P(\boldsymbol{X}=\boldsymbol{x}|\boldsymbol{\vartheta}))}{\vartheta_P}\right)^\top$ から，フィッシャー情報行列は

$$\mathcal{I}_{\boldsymbol{\vartheta}} = E(\psi(\boldsymbol{\vartheta}; \boldsymbol{x})\psi(\boldsymbol{\vartheta}; \boldsymbol{x})^\top) \tag{3.16}$$

と定義される．ここで，期待値は式(2.17)の同時確率を用いて，すべての項目反応パタンについて取る．

このフィッシャー情報行列の推定法として，対数尤度の外積を用いた方法とヘシアンを用いた方法がよく用いられる．勾配の外積推定量は

$$\mathcal{I}_{\boldsymbol{\vartheta}} \approx \frac{1}{I}\sum_{i=1}^{I} \psi(\boldsymbol{\vartheta}; \boldsymbol{x}_i)\psi(\boldsymbol{\vartheta}; \boldsymbol{x}_i)^\top \Big|_{\boldsymbol{\vartheta}=\hat{\boldsymbol{\vartheta}}} = \mathcal{I}_{\boldsymbol{\vartheta},OPG} \tag{3.17}$$

であり，対数尤度の負のヘシアンによる近似は

$$\mathcal{I}_{\boldsymbol{\vartheta}} \approx -\frac{1}{I}\sum_{i=1}^{I} \frac{\partial^2 \log(P(\boldsymbol{X} = \boldsymbol{x}_i|\boldsymbol{\vartheta}))}{\partial\boldsymbol{\vartheta}\partial\boldsymbol{\vartheta}^\top}\Big|_{\boldsymbol{\vartheta}=\hat{\boldsymbol{\vartheta}}} = \mathcal{I}_{\boldsymbol{\vartheta},Hess} \tag{3.18}$$

である．これらの具体的な要素の計算方法については，Philipp et al. (2018)で示されている．また，フィッシャー情報行列として，構造パラメタ($\boldsymbol{\pi}$)を考慮しない項目レベルでの情報行列が de la Torre (2009)と de la Torre (2011) によって提案されている．さらに，Liu et al. (2019)はモデルを誤設定したもとでのサンドイッチ型の漸近分散共分散行列を得る方法として，フィッシャー情報行列を

$$\mathcal{I}_{\boldsymbol{\vartheta}} \approx \mathcal{I}_{\boldsymbol{\vartheta},SW} = \mathcal{I}_{\boldsymbol{\vartheta},Hess}^{-1}\mathcal{I}_{\boldsymbol{\vartheta},OPG}\mathcal{I}_{\boldsymbol{\vartheta},Hess}^{-1}/I \tag{3.19}$$

とし，シミュレーションによりサンドイッチ型フィッシャー情報行列を用いた標準誤差の頑健性を示した．

また，Yamaguchi (2023b)は EM アルゴリズムの枠組みを活用した supplemental EM アルゴリズム (Cai, 2008; Meng and Rubin, 1991)を活用し，完全データのフィッシャー情報行列から観測フィッシャー情報行列を得る方法を提案した．Supplemental EM アルゴリズムはパラメタ推定後に追加の EM 計算が必要であり，計算量が増えるものの，完全データの情報行列は観測データの情報行列よりも計算がしやすく，実装が容易である．Supplemental EM アルゴリズムには，推定されたパラメタ $\hat{\boldsymbol{\vartheta}}$ の漸近分散共分散行列が

$$\mathcal{I}_{\boldsymbol{\vartheta},Obs}^{-1}|_{\boldsymbol{\vartheta}=\hat{\boldsymbol{\vartheta}}} = \mathcal{I}_{\boldsymbol{\vartheta},Comp}^{-1}[\boldsymbol{I}_P - \Delta(\boldsymbol{\vartheta})]^{-1}|_{\boldsymbol{\vartheta}=\hat{\boldsymbol{\vartheta}}} \tag{3.20}$$

と書くことができることを利用する．ここで，$\mathcal{I}_{\boldsymbol{\vartheta},Obs}$ は観測データの情報行列，$\mathcal{I}_{\boldsymbol{\vartheta},Comp}$ は完全データの情報行列であり，$\boldsymbol{I}_P$ は $P \times P$ の単位行列，そして $\Delta(\boldsymbol{\vartheta})$ は EM 写像のヤコビアンである．$\Delta(\boldsymbol{\vartheta})$ の要素 $d_{ij}, i, j = 1, \ldots, P$ は 1 回の EM サイクルの出力の j 番目の要素 $M_j(\boldsymbol{\vartheta})$ を用いて，

$$d_{ij} = \frac{\partial M_j(\boldsymbol{\vartheta})}{\partial \vartheta_i} \tag{3.21}$$

と表すことができ，EM アルゴリズムを用いた数値微分により求めることができる（Jamshidian and Jennrich, 2000）．

3.3 正則化推定法

DCM のモデルパラメタは，Q 行列と項目反応関数を適切に仮定することで，潜在クラスとしてのアトリビュート習得パタンやアトリビュートの項目への影響やアトリビュート間の交互作用効果としての項目パラメタといったような解釈しやすいものになっていた．しかし，DCM は，場合によっては制約が強すぎてデータに適合しにくいことや，数多くの DCM からどのモデルを選択すればよいのか明確ではないことが問題となる．その一方，制約がない探索的な潜在クラスモデルは比較的制約が緩いためデータにモデル適合しやすいものの，解釈性が高いパラメタ推定値が得られるとは限らない．また，DCM を用いるためには Q 行列を適切に設定したり，項目反応関数を設定する必要があり，その設定はかならずしも容易ではない．実際，前述のように Q 行列に誤設定があることによって項目パラメタの推定値にバイアスがかかるといった問題が生じる．こうしたことから，制約ができるだけ少なく，DCM 的な推定結果が得られるモデルが応用上は使いやすいと考えられる．

このような目標を達成するために，Chen et al. (2017) は部分的に併合された潜在クラス (partially merged latent class) という考え方を提案し，DCM 的な解釈性の高いパラメタ推定と探索的潜在クラスモデルの柔軟性をもつ正則化潜在クラス分析 (regularized latent class analysis) を提案した．DCM の項目反応関数 (式 (2.1)) からは，アトリビュート習得パタンで条件づけた正答確率が同じになる複数のパタンがあることがわかる．これは，言い換えると，潜在クラスの一部が併合されていると見ることができるため，DCM は正則化潜在クラスモデルとみなすことができる．このようなモデルのパラメタ推定には正則化推定法が利用できる．正則化推定法はスパースモデリングなどの文脈で頻繁に利用される方法であり，(対数) 周辺尤度の後に正則化項を追加した目的関数を最適化するパラメタを推定値とする推定法である．

Chen et al. (2017) はあるクラス l の項目 j に対する正答反応確率を θ_{jl} とする通常の潜在クラスモデルの周辺尤度

$$L(\boldsymbol{\Theta}, \boldsymbol{\pi}) = \prod_{i=1}^{I} \sum_{l=1}^{L} \left\{ \pi_l \prod_{j=1}^{J} \theta_{jl}^{x_{ij}} (1-\theta_{jl})^{1-x_{ij}} \right\} \tag{3.22}$$

に対して項目反応関数に対する正則化項 $\kappa_\xi(\boldsymbol{\Theta})$ を仮定し

$$\{\hat{\boldsymbol{\Theta}}^\xi, \hat{\boldsymbol{\pi}}^\xi\} = \underset{\boldsymbol{\Theta}, \boldsymbol{\pi}}{\operatorname{argmax}} \{\log L(\boldsymbol{\Theta}, \boldsymbol{\pi}) - I\kappa_\xi(\boldsymbol{\Theta})\} \tag{3.23}$$

をパラメタ推定値とする方法を開発した．正則化項 $\kappa_\xi(\boldsymbol{\Theta})$ により，同じ正答確率を示すクラスを推定することができる．DCM は前述のように，ある項目に対して同じ正答確率を示すアトリビュート習得パタンを想定するため，上記の正則化推定法により，通常の潜在クラスモデルから DCM と同じような推定を得ることができる．

正則化項 $\kappa_\xi(\boldsymbol{\Theta})$ は $\kappa_\xi(\boldsymbol{\Theta}) = \sum_{j=1}^{J} p_\xi(\boldsymbol{\theta}_j), \boldsymbol{\theta}_j = (\theta_{j1}, \ldots, \theta_{jL})$ と項目ごとの正則化項 $\kappa_\xi(\boldsymbol{\theta}_j)$ の和で表される．さらに，スパース性を決定するハイパーパラメタ p_ξ と項目反応確率パラメタの順序統計量 $\theta_{j(1)} \leq \theta_{j(2)} \leq \cdots \leq \theta_{j(L)}$ を用いて，項目ごとの正則化項は

$$p_\xi(\boldsymbol{\theta}_j) = \sum_{l=1}^{L-1} p_\xi^{SCAD}(\theta_{j(l+1)} - \theta_{j(l)}) \tag{3.24}$$

と表される．ここで，p_ξ^{SCAD} は

$$p_\xi^{SCAD}(x) = \begin{cases} \xi|x| & \text{if } |x| \leq \xi \\ -\left(\frac{|x|^2 - 2a\xi|x| + \xi^2}{2(a-1)}\right) & \text{if } \xi < |x| < a\xi \\ \frac{(a+1)^2\xi^2}{2} & \text{if } |x| \geq a\xi \end{cases} \tag{3.25}$$

と定義される．また，スパース性を決めるパラメタの決定方法も Chen et al.（2017）で示されている．適当な条件のもと，この正則化項を用いた推定により，一致的なパラメタ推定とモデル選択ができることが示されている（Chen et al., 2017, pp. 668–670）．

また，正則化推定法はアトリビュート階層構造の推定（Wang and Lu, 2021; Ma et al., 2023b）や縦断的 DCM（e.g., Yamaguchi and Martinez, 2024）におけるアトリビュート習得パタンの遷移で表される学習軌跡の推定（Wang, 2021），Q 行列の推定（Chen et al., 2020, 2015; Xu and Shang, 2018）にも活用されているように，DCM におけるパラメタ推定・構造推定のための標準的な方法として定着しつつある．

4. ベイズ推定法

本章では DCM におけるパラメタのベイズ推定法を示す．本章で扱うベイズ推定法は，モデルパラメタに事前分布を仮定し，事後分布やその近似分布などからパラメタの推定値や予測分布を得る方法として広く捉える．

4.1 MAP 推定

周辺尤度に加えパラメタの事前分布を考慮し，点推定を得る方法は MAP 推定法として知られている．MAP 推定法は式(2.18)に事前分布の対数を加えた目的関数を最適化するパラメタを推定値とする方法であり，

$$\{\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\pi}}\} = \underset{\boldsymbol{\Theta},\boldsymbol{\pi}}{\operatorname{argmax}}\{\log L(\boldsymbol{\Theta}, \boldsymbol{\pi}) + \log(p(\boldsymbol{\Theta})) + \log(p(\boldsymbol{\pi}))\} \tag{4.1}$$

とパラメタの推定値を得る方法である．ここで，$\log(p(\boldsymbol{\Theta}))$ と $\log(p(\boldsymbol{\pi}))$ はそれぞれ項目パラメタと混合比率パラメタの対数事前確率密度である．例えば，θ_{jh} の事前分布としてパラメタが $a_{jh}^0 > 0, b_{jh}^0 > 0$ であるベータ分布，$\boldsymbol{\pi}$ の事前分布としてパラメタが $\boldsymbol{\delta}^0 = (\delta_1^0, \ldots, \delta_L^0), \delta_l^0 > 0, \forall l$ であるディリクレ分布を仮定することができる．具体的な 2 つの事前確率密度は

$$p(\boldsymbol{\Theta}) = \prod_{j=1}^{J}\prod_{h=1}^{H_j} \frac{\Gamma(a_{jh}^0)\Gamma(b_{jh}^0)}{\Gamma(a_{jh}^0 + b_{jh}^0)} \theta_{jh}^{a_{jh}^0 - 1}(1-\theta_{jh})^{b_{jh}^0 - 1}, \tag{4.2}$$

$$p(\boldsymbol{\pi}) = \frac{\Gamma(\sum_{l=1}^{L}\delta_l^0)}{\prod_{l=1}^{L}\Gamma(\delta_l^0)} \prod_{l=1}^{L} \pi_l^{\delta_l^0 - 1} \tag{4.3}$$

である．

この事前分布の設定のもと，MAP 推定のための EM アルゴリズムの Q 関数は，式(3.10)に対数事前分布の定数項を除いたものを足し，

$$\begin{aligned} &Q_{MAP}(\{\boldsymbol{\Theta}, \boldsymbol{\pi}\}|\{\boldsymbol{\Theta}^{(t)}, \boldsymbol{\pi}^{(t)}\}) \\ &= Q(\{\boldsymbol{\Theta}, \boldsymbol{\pi}\}) + \sum_{j=1}^{J}\sum_{h=1}^{H_j}\{(a_{jh}-1)\log(\theta_{jh}) + (b_{jh}-1)\log(1-\theta_{jh})\} + \sum_{l=1}^{L}(\delta_l - 1)\log(\pi_l) \end{aligned} \tag{4.4}$$

となる．M-step ではこの $Q_{MAP}(\{\boldsymbol{\Theta},\boldsymbol{\pi}\}|\{\boldsymbol{\Theta}^{(t)},\boldsymbol{\pi}^{(t)}\})$ を $\{\boldsymbol{\Theta},\boldsymbol{\pi}\}$ について最大化する．このときの更新式は，周辺最尤推定法の場合の導出と同様に導出でき，

$$\theta_{jh} = \frac{\sum_{i=1}^{I}\sum_{l=1}^{L}\mathcal{I}(\boldsymbol{\alpha}^{*}_{jh}=\boldsymbol{\alpha}_l)\gamma_{il}x_{ij} + a^0_{jh} - 1}{\sum_{i=1}^{I}\sum_{l=1}^{L}\mathcal{I}(\boldsymbol{\alpha}^{*}_{jh}=\boldsymbol{\alpha}_l)\gamma_{il} + a^0_{jh} + b^0_{jh} - 2}, \tag{4.5}$$

$$\pi_l = \frac{\sum_{i=1}^{I}\gamma_{il} + \delta^0_l - 1}{I + \sum_{l=1}^{L}\delta^0_l - L} \tag{4.6}$$

となる．ただし，この更新式は周辺最尤推定法のための EM アルゴリズムと同様にパラメタの単調性制約を課していない点に注意が必要である．最尤推定の場合と同様に，MAP 推定の M-step においても，単調性制約を課した推定を行うこともできる．

4.2 マルコフ連鎖モンテカルロ法

MAP 推定法は最尤推定法と同様に点推定値を得る方法であった．一方，マルコフ連鎖モンテカルロ（Markov chain Monte Carlo [MCMC]）法を用いて，パラメタの事後分布を近似することで，点推定値だけではなく推定の不確かさを事後標準偏差などの統計量によって評価することができる．MCMC 法は MAP 推定や後述の変分ベイズ推定法よりも計算負荷が高いものの，単調性制約などモデルの制約も柔軟に取り込んだ推定が比較的容易に実行できるという利点がある．本節では単調性を満足するギブスサンプリング法（Liu and Johnson, 2019; Yamaguchi and Templin, 2022b）を示す．事前分布は，MAP 推定の際に用いたものと同様の分布を仮定する．

まず，上付き添字 $(t), (t=1,\ldots,T)$ は MCMC サンプリングのイテレーション番号を表すとする．$\boldsymbol{x}_i, \boldsymbol{\Theta}^{(t-1)}, \boldsymbol{\pi}^{(t-1)}$ が得られているもとでのテスト解答者のアトリビュート習得パタン $\boldsymbol{\alpha}_i^{(t)}$ は，完全条件付き確率が

$$\begin{aligned} &P(\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_l|\boldsymbol{x}_i, \boldsymbol{\Theta}^{(t-1)}, \boldsymbol{\pi}^{(t-1)}) \\ &= \frac{\pi_l^{(t-1)}\prod_{j=1}^{J}\prod_{h=1}^{H_j}[(\theta_{jh}^{(t-1)})^{x_{ij}}(1-\theta_{jh}^{(t-1)})^{1-x_{ij}}]^{\mathcal{I}(\boldsymbol{\alpha}^{*}_{jh}=\boldsymbol{\alpha}_l)}}{\sum_{l'=1}^{L}\{\pi_{l'}^{(t-1)}\prod_{j=1}^{J}\prod_{h=1}^{H_j}[[(\theta_{jh}^{(t-1)})^{x_{ij}}(1-\theta_{jh}^{\hat{(t-1)}})^{1-x_{ij}}]^{\mathcal{I}(\boldsymbol{\alpha}^{*}_{jh}=\boldsymbol{\alpha}_{l'})}\}} \end{aligned} \tag{4.7}$$

であるカテゴリカル分布からのサンプリングにより生成する．

次に，混合比率パラメタ $\boldsymbol{\pi}^{(t)}$ の完全条件付き事後分布はパラメタが

$$\delta_l^{*,(t)} = \sum_{i=1}^{I}\mathcal{I}(\boldsymbol{a}_i^{(t)}=\boldsymbol{a}_l)(1-x_{ij}) + \delta^0_l \tag{4.8}$$

であるディリクレ分布であり，この分布からサンプリングを行う．

最後に，項目パラメタ $\theta_{jh}^{(t)}$ の完全条件付き事後分布は，パラメタが

$$\begin{cases} a^{*,(t)}_{jh} = \sum_{i=1}^{I}\sum_{l=1}^{L}\mathcal{I}(\boldsymbol{a}^{*}_{jh}=\boldsymbol{a}_l)\mathcal{I}(\boldsymbol{a}_i^{(t)}=\boldsymbol{a}_l)x_{ij} + a^0_{jh} \\ b^{*,(t)}_{jh} = \sum_{i=1}^{I}\sum_{l=1}^{L}\mathcal{I}(\boldsymbol{a}^{*}_{jh}=\boldsymbol{a}_l)\mathcal{I}(\boldsymbol{a}_i^{(t)}=\boldsymbol{a}_l)(1-x_{ij}) + b^0_{jh} \end{cases} \tag{4.9}$$

であるベータ分布である．ただし，項目パラメタのサンプリングの際には，単調性制約を満足するために，切断ベータ分布を用いた逆関数法を利用したアルゴリズム用いる（e.g., Hoijtink, 1998; Laudy et al., 2004）．上限と下限が Upp と Low である切断ベータ分布から乱数をサンプリングする逆関数法は以下の 2 つの手続きを踏む．まず，下限 0 から上限 1 の一様分布に従う変数 u をサンプリングし，次にパラメタが a, b のベータ分布の分位点 $Low + u \times (1 - Upp - Low)$ を計算する．この分位点が，目標とする切断ベータ分布からのランダムサンプリングになっており，この手続きを項目パラメタの単調性制約を満足するために利用する．

項目パラメタ θ_{jh} をサンプリングする際には，以下の手続きを踏む．まず，$\{h|\sum_k \alpha^*_{hjk}=0\}$ に対して，下限 0, 上限 $\min\{\theta^{(t-1)}_{jh'}|\boldsymbol{\alpha}^*_{h'j} \succ \boldsymbol{\alpha}^*_{hj}\}$ で，パラメタが $a^{*,(t)}_{jh}, b^{*,(t)}_{jh}$ の切断ベータ分布から MCMC サンプルをサンプリングする．次に，$\{h|\sum_k q_{jk} > \sum_k \alpha^*_{hjk} > 0\}$ に対して，$\sum_k \alpha^*_{hjk} = 1, 2, \ldots, \sum_k q_{jk} - 1$ の順に，下限 $\max\{\theta^{(t)}_{jh}|\boldsymbol{\alpha}^*_{hj} \succ \boldsymbol{\alpha}^*_{h'j}\}$, 上限 $\min\{\theta^{(t-1)}_{jh'}|\boldsymbol{\alpha}^*_{h'j} \succ \boldsymbol{\alpha}^*_{hj}\}$ で，パラメタが $a^{*,(t)}_{jh}, b^{*,(t)}_{jh}$ の切断ベータ分布から MCMC サンプルをサンプリングする．最後に，$\{h|\alpha^*_{hjk} = \sum_k q_{jk}\}$ に対して，下限 $\max\{\theta^{(t)}_{jh}|\boldsymbol{\alpha}^*_{hj} \succ \boldsymbol{\alpha}^*_{h'j}\}$, 上限 1 で，パラメタが $a^{*,(t)}_{jh}, b^{*,(t)}_{jh}$ の切断ベータ分布から MCMC サンプルをサンプリングする．以上の手続きを，項目パラメタおよび混合比率パラメタの初期値 $\boldsymbol{\Theta}^{(0)}, \boldsymbol{\pi}^{(0)}$ を定めたもとで，予め決めた回数 T だけ反復することで，MCMC サンプルを得る．

この他，後述のノンパラメトリック推定と関連して，モデルのパラメタを周辺化し，アトリビュート習得パタンを直接サンプリングする崩壊型ギブスサンプリング（collapsed Gibbs sampling）も提案されている（Yamaguchi and Templin, 2022a）．さらに，スライスサンプリングを用いた方法（Xu et al., 2020）や，DINA モデルのパラメタ推定に Pólya-gamma 分布を用いた方法（Zhang et al., 2020）や，Pólya-gamma 分布を用いてモデルパラメタと Q 行列の推定も同時に実行する方法も提案されている（Balamuta and Culpepper, 2022; Jimenez et al., 2023）．

4.3 変分ベイズ推定法

MCMC 法は乱数を生成することにより，パラメタと潜在変数の事後分布を近似する方法であった．MCMC 法の欠点とそれを補う変分ベイズ推定法の利点については以下のようにまとめられる（Bishop, 2006）． まず，MCMC 法はサンプリングの回数を増やせば事後分布の近似精度を高めることができるが，計算量が多くなり推定に時間がかかってしまうことがある．また，MCMC サンプリングが定常分布からのサンプリングになっているかどうかを判断するための手続きが必要であるが，その判断は必ずしも容易ではない．一方，変分ベイズ推定法はあるクラスの確率分布の中から事後分布を最もよく近似する分布を見つける方法であり，計算量を押さえつつアルゴリズムの収束判定も行いやすく，ベイズ的な推定量を得ることができる．また，変分ベイズ推定法は，対数尤度の下界を最適化する方法として定式化されることもある（e.g., Jeon et al., 2017）．上記のように，変分ベイズ推定法は最適化問題を解くことで，MCMC よりも高速な推定を可能にしている．ただし，事後分布を近似する分布（変分事後分布）はデータを得た元での真の事後分布ではないため，変分事後分布における分散が真の事後分布における分散よりも小さくなるなどの問題も有している．

変分ベイズ推定法は，項目反応理論モデルを含む潜在変数モデルにおいても近年適用例が増加してきている方法である．例えば，Jeon et al.（2017）は一般化線形混合モデルに対しての変分ベイズアルゴリズムを提案した．その後，多次元項目反応理論モデルへの適用例（Cho et al., 2021, 2024）もあり，ますますの発展が見られる．変分ベイズ推定法についての理論的な解説は Nakajima et al.（2019）においてなされている．また，入門的レビューについては Blei et al.（2017）や Grimmer（2011）が参考になる．

DCM においては，Yamaguchi and Okada（2020c）において変分ベイズ推定法に基づく DINA モデルの推定アルゴリズムが提案されている．さらに，Yamaguchi（2020）は多枝選択式のデータに対する DINA モデル（multiple choice DINA model）について，Yamaguchi and Okada（2020b）は前述の MCMC 法で用いたモデルと事前分布を仮定した変分ベイズ推定法を提案した．この他，Yamaguchi and Martinez（2024）は隠れマルコフモデルにもとづく縦断的 DCM に対して変分ベイズ推定法を提案し，Yamaguchi（2023a）は外部情報を含めた DCM のうちの一つである二段階 DCM に対する変分ベイズ推定法を提案している．本節では，Yamaguchi and Okada（2020b）にもとづいた DCM における変分ベイズ推定法のアルゴリズムを示す．アルゴ

リズムの具体的な導出方法については，Yamaguchi and Okada (2020b)に詳細な記載がある．

変分ベイズ推定法では，DCM における潜在変数とパラメタの事後分布 $P(\boldsymbol{\mathcal{A}}, \boldsymbol{\Theta}, \boldsymbol{\pi}|\mathrm{X})$ を適当な分布 $q(\boldsymbol{\mathcal{A}}, \boldsymbol{\Theta}, \boldsymbol{\pi})$ によって近似することを考える．$q(\boldsymbol{\mathcal{A}}, \boldsymbol{\Theta}, \boldsymbol{\pi})$ と $P(\boldsymbol{\mathcal{A}}, \boldsymbol{\Theta}, \boldsymbol{\pi}|\mathrm{X})$ の間の Kullback-Leibler divergence を

$$KL(q(\boldsymbol{\mathcal{A}}, \boldsymbol{\Theta}, \boldsymbol{\pi})||P(\boldsymbol{\mathcal{A}}, \boldsymbol{\Theta}, \boldsymbol{\pi}|\mathrm{X})) = \int \sum_{\boldsymbol{\mathcal{A}}} q(\boldsymbol{\mathcal{A}}, \boldsymbol{\Theta}, \boldsymbol{\pi}) \log \frac{q(\boldsymbol{\mathcal{A}}, \boldsymbol{\Theta}, \boldsymbol{\pi})}{P(\boldsymbol{\mathcal{A}}, \boldsymbol{\Theta}, \boldsymbol{\pi}|\mathrm{X})} d\boldsymbol{\Theta} d\boldsymbol{\pi} \tag{4.10}$$

とする．DCM の変分ベイズ推定法においては，$KL(q(\boldsymbol{\mathcal{A}}, \boldsymbol{\Theta}, \boldsymbol{\pi})||P(\boldsymbol{\mathcal{A}}, \boldsymbol{\Theta}, \boldsymbol{\pi}|\mathrm{X}))$ を潜在変数とパラメタの間の独立性の仮定

$$q(\boldsymbol{\mathcal{A}}, \boldsymbol{\Theta}, \boldsymbol{\pi}) = q(\boldsymbol{\mathcal{A}}) q(\boldsymbol{\Theta}, \boldsymbol{\pi}) = \left(\prod_{i=1}^{I} q(\boldsymbol{\alpha}_i) \right) \left(\prod_{j=1}^{J} \prod_{h=1}^{H_j} q(\theta_j h) \right) q(\boldsymbol{\pi}) \tag{4.11}$$

のもと最小化する分布 q^* を変分事後分布とよび，変分事後分布を事後分布の近似分布とする．ここでは，表記の簡略化のため，引数に対応して異なった分布を仮定している．上式の $q(\boldsymbol{\mathcal{A}}) q(\boldsymbol{\Theta}, \boldsymbol{\pi})$ が最小限の独立性の仮定であり，2 つ目の等号はモデルの仮定から導かれる独立性である．これを用いて，変分ベイズ推定法における最適化問題は

$$q^* = \underset{q}{\mathrm{argmin}}\, KL(q(\boldsymbol{\mathcal{A}}, \boldsymbol{\Theta}, \boldsymbol{\pi})||P(\boldsymbol{\mathcal{A}}, \boldsymbol{\Theta}, \boldsymbol{\pi}|\mathrm{X})), \mathrm{s.t.}\ q(\boldsymbol{\mathcal{A}}) q(\boldsymbol{\Theta}, \boldsymbol{\pi}) \tag{4.12}$$

となる．近似する分布に独立性を仮定し，変分分布のクラスを限定して事後分布を近似する方法は平均場近似と呼ばれる．変分ベイズ推定法では，真の事後分布を近似するための変分分布の独立性を仮定した最適化問題を解くため，MCMC 法よりも高速な推定が可能となる．

また，等価な定式化として，周辺対数尤度の下界 $\underline{l}(q)$ を同様の制約条件のもと最適化する分布を変分分布とすることもある．すなわち，

$$\begin{aligned} q^* &= \underset{q}{\mathrm{argmax}} \int \sum_{\boldsymbol{\mathcal{A}}} q(\boldsymbol{\mathcal{A}}, \boldsymbol{\Theta}, \boldsymbol{\pi}) \log \frac{P(\mathrm{X}, \boldsymbol{\mathcal{A}}, \boldsymbol{\Theta}, \boldsymbol{\pi})}{q(\boldsymbol{\mathcal{A}}, \boldsymbol{\Theta}, \boldsymbol{\pi})}, \mathrm{s.t.}\ q(\boldsymbol{\mathcal{A}}) q(\boldsymbol{\Theta}, \boldsymbol{\pi}) d\boldsymbol{\Theta} d\boldsymbol{\pi} \\ &= \underset{q}{\mathrm{argmax}} \{\underline{l}(q)\}, \mathrm{s.t.}\ q(\boldsymbol{\mathcal{A}}) q(\boldsymbol{\Theta}, \boldsymbol{\pi}) \end{aligned} \tag{4.13}$$

により変分事後分布を求める問題を定式することもある．ただし，$P(\mathrm{X}, \boldsymbol{\mathcal{A}}, \boldsymbol{\Theta}, \boldsymbol{\pi})$ はデータとパラメタの同時確率であり，式(3.1)とパラメタの事前分布から，

$$P(\mathrm{X}, \boldsymbol{\mathcal{A}}, \boldsymbol{\Theta}, \boldsymbol{\pi}) = P(\mathrm{X}, \boldsymbol{\mathcal{A}}|\boldsymbol{\Theta}, \boldsymbol{\pi}) P(\boldsymbol{\Theta}) P(\boldsymbol{\pi}) \tag{4.14}$$

と表される．対数周辺尤度の下界 $\underline{l}(q)$ は ELBO(evidence lower bound)とも呼ばれ，本論文の DCM の設定においては解析的な表現が与えられている（Yamaguchi, 2020, Equation 32）．

上記の設定のもと，

$$q(\boldsymbol{\mathcal{A}}) \propto \exp\{\mathbb{E}_{q(\boldsymbol{\Theta}, \boldsymbol{\pi})}[\log P(\mathrm{X}, \boldsymbol{\mathcal{A}}, \boldsymbol{\Theta}, \boldsymbol{\pi})]\}, \tag{4.15}$$

$$q(\boldsymbol{\Theta}, \boldsymbol{\pi}) \propto \exp\{\mathbb{E}_{q(\boldsymbol{\mathcal{A}})}[\log P(\mathrm{X}, \boldsymbol{\mathcal{A}}, \boldsymbol{\Theta}, \boldsymbol{\pi})]\} \tag{4.16}$$

により，最適な変分事後分布を計算できる．しかし，変分事後分布は相互に関係しているので，最適な変分事後分布を得るために反復計算を行う必要がある．期待値の計算 $\mathbb{E}_{q(\boldsymbol{\Theta}, \boldsymbol{\pi})}[\cdot], \mathbb{E}_{q(\boldsymbol{\mathcal{A}})}[\cdot]$ はそれぞれ変分事後分布 $q(\boldsymbol{\Theta}, \boldsymbol{\pi})$ と $q(\boldsymbol{\mathcal{A}})$ について取ることを意味している．前述の事前分布の設定のもと，この変分事後分布はよく知られた分布になることが示されている．まず，$\boldsymbol{\alpha}_i$ の変分事後分布は

$$q^*(\boldsymbol{\alpha}_i) \propto \prod_{l=1}^{L} r_{il}^{\mathcal{I}(\boldsymbol{\alpha}_i=\boldsymbol{\alpha}_l)} \tag{4.17}$$

であり，変分パラメタが $r_{il} = \frac{\rho_{il}}{\sum_{l'=1}^{L}\rho_{il'}}$ であるカテゴリカル分布である．ここで ρ_{il} は

$$\begin{aligned}\log \rho_{il} = {} & \mathbb{E}_{q(\boldsymbol{\pi})}[\log \pi_l] \\ & + \sum_{j=1}^{J}\sum_{h=1}^{H_j} \mathcal{I}(\boldsymbol{\alpha}^*_{jh} = \boldsymbol{\alpha}_l)(x_{ij}\mathbb{E}_{q(\theta_{jh})}[\log \theta_{jh}] + (1-x_{ij})\mathbb{E}_{q(\theta_{jh})}[\log(1-\theta_{jh})])\end{aligned} \tag{4.18}$$

である．上式中に現れる期待値の具体的な値は後述する．つぎに，混合比率パラメタの変分事後分布は，

$$q^*(\boldsymbol{\pi}) \propto \prod_{l=1}^{L} \pi_l^{\delta^*_l} \tag{4.19}$$

と表されることからディリクレ分布であることがわかり，変分パラメタ δ^*_l は

$$\delta^*_l = \sum_{i=1}^{I} \mathbb{E}_{q(\boldsymbol{\alpha}_i)}[\mathcal{I}(\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_l)] + \delta^0_l \tag{4.20}$$

である．最後に，項目パラメタ θ_{jh} の変分事後分布は

$$q^*(\theta_{jh}) \propto \theta_{jh}^{a^*_{jh}-1}(1-\theta_{jh})^{b^*_{jh}-1} \tag{4.21}$$

となるベータ分布であり，変分パラメタ a^*_{jh}, b^*_{jh} は，

$$\begin{cases} a^*_{jh} & = \sum_{i=1}^{I}\sum_{l=1}^{L} \mathcal{I}(\boldsymbol{a}^*_{jh} = \boldsymbol{a}_l)\mathbb{E}_{q(\boldsymbol{\alpha}_i)}[\mathcal{I}(\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_l)]x_{ij} + a^0_{jh} \\ b^*_{jh} & = \sum_{i=1}^{I}\sum_{l=1}^{L} \mathcal{I}(\boldsymbol{a}^*_{jh} = \boldsymbol{a}_l)\mathbb{E}_{q(\boldsymbol{\alpha}_i)}[\mathcal{I}(\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_l)](1-x_{ij}) + b^0_{jh} \end{cases} \tag{4.22}$$

と計算される．上記の変分パラメタの更新に現れる期待値は

$$\mathbb{E}_{q(\boldsymbol{\alpha}_i)}[\mathcal{I}(\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_l)] = r_{il}, \tag{4.23}$$

$$\mathbb{E}_{q(\theta_{jh})}[\log \theta_{jh}] = \psi(a^*_{jh}) - \psi(a^*_{jh} + b^*_{jh}), \tag{4.24}$$

$$\mathbb{E}_{q(\theta_{jh})}[\log(1-\theta_{jh})] = \psi(b^*_{jh}) - \psi(a^*_{jh} + b^*_{jh}), \tag{4.25}$$

$$\mathbb{E}_{q(\boldsymbol{\pi})}[\log \pi_l] = \psi(\delta^*_l) - \psi\left(\sum_{l=1}^{L}\delta^*_l\right) \tag{4.26}$$

と計算される．ここで，$\psi(\cdot)$ はディガンマ関数であり $\frac{d}{dx}\log\Gamma(x) = \int_0^\infty y^{(x-1)}\exp(-y)dy$ と定義される．

DCM に対する変分ベイズアルゴリズムは以下のようになる．まず，変分パラメタ $A^* = \{a^*_{jh}\}^{J,H_j}_{j,h}, B^* = \{B^*_{jh}\}^{J,H_j}_{j,h}, \boldsymbol{\delta}^* = (\delta^*_1, \ldots, \delta^*_L)^\top$ を初期化し，$q(\boldsymbol{\alpha}_i), \forall i,$ の変分パラメタを式(4.18)により更新する(変分 E-step)．混合比率パラメタと項目パラメタの変分事後分布 $q(\boldsymbol{\pi})$ と $q(\theta_j h), \forall j, h,$ の変分パラメタを式(4.20)と(4.22)により更新する(変分 M-step)．変分 E-step および変分 M-step を収束基準(例えば，$\underline{l}(q)$ の変化量が $1 > \epsilon > 0$ 以下になる，など)を満足するまで，変分 E-step と変分 M-step を反復する．上記の更新式により $\underline{l}(q)$ が単調に増加することが保証されている．最後に，得られた変分事後分布をもとにその期待値・中央値や分散などの要約統計量をパラメタの点推定値や事後分散の推定値とすればよい．

ここで示した変分ベイズアルゴリズムは，単調性を完全に満足した事後分布の構成はできて

いない点に注意が必要である．Yamaguchi and Okada (2020c)では，事前分布の平均値が単調性を満たすようにハイパーパラメタを調整しておく方法を提案し，シミュレーションや実データ分析ではフルベイズ推定の場合と類似したの推定値を得ることができることを報告している．しかし，当該の方法の理論的な性質は検討されていない．

5. (一般化)ノンパラメトリック推定法

最尤推定法とベイズ推定法はいずれもパラメトリックな DCM の推定法である．すなわち，これらの方法では DCM の最終的なアウトプットであるアトリビュート習得パタンとモデルのパラメタの両方を仮定していた．これらの方法はサンプルサイズが小さい場合にはアトリビュート習得パタンの推定が安定しないといった問題を持つとされている．これらの推定法に対して，正答確率パラメタや混合比率パラメタといったモデルパラメタを仮定せずにアトリビュート習得パタンを推定する方法として，(一般化)ノンパラメトリック推定法(generalized non-parametric method)も提案されている (Chiu and Douglas, 2013; Chiu et al., 2018)．加えて，Ma et al. (2023a)はノンパラメトリック推定法とパラメトリック法の関連を議論し，これらを統一的に表現できる損失関数に基づく推定法を提案した．本章では，基本的なノンパラメトリック推定法を示し，さらに一般化ノンパラメトリック推定法についても示す．

後述するように，(一般化)ノンパラメトリック推定法で用いられる損失関数では DINA モデルや DINO(deterministic input noisy “OR”-gate)モデル (Templin and Henson, 2006)といった具体的なデータ生成モデルを仮定しているように見える．しかし，(一般化)ノンパラメトリック推定法によるアトリビュート習得パタンの推定値は，データ生成モデルが DINA モデルや DINO モデルでなくとも一致性があることが示されている．

まず，Wang and Douglas (2015)はノンパラメトリック推定法によるアトリビュート習得パタンの推定値の一致性を証明した．より具体的には，DINA モデルを含む結合的(非補償的)な DCM をデータ生成モデルし，既知の Q 行列のもとで，ノンパラメトリック推定法によるアトリビュート習得パタンが一致的であるための項目反応関数について必要な条件は以下の 2 つであることが示されている (Wang and Douglas, 2015, p. 99)．第 1 の条件は，それぞれの項目に対して項目の正答に必要なアトリビュートを習得している場合の正答確率が 0.5 以上であることで，第 2 の条件は項目の正答に必要なアトリビュートを習得していない場合の正答確率が 0.5 未満であることである．データ生成モデルがこれらの条件を満足している場合，ノンパラメトリック推定法によるアトリビュート習得パタンの推定値は，個人ごとに，項目数の増加に伴って一致的であることが示されている．

さらに，Chiu and Köhn (2019)は一般化ノンパラメトリック推定法によるアトリビュート習得パタンの推定値の一致性を証明した．この一致性は，データ生成モデルが LCDM と同等の一般的な DCM である generalized DINA モデル (de la Torre, 2011)で表現できるモデルに対して，一定の仮定のもと成り立つものである (Chiu and Köhn, 2019, p. 841, Theorem 2)．これはすなわち，データ生成モデルがよく利用される DCM の範囲では，一般化ノンパラメトリック推定法のよるアトリビュート習得パタンの推定値が統計的に望ましい性質を有しており，この推定法を用いる根拠があることを意味している．この意味で，一般化ノンパラメトリック推定法は，パラメトリックなモデルを仮定した推定法と同様に一般に利用できる方法であるといえる．

5.1 ノンパラメトリック推定法

ここから，具体的なノンパラメトリック推定法について説明する．ノンパラメトリック推定

法で用いる距離関数を定義するために理想反応(ideal response)を定義する．DINA モデルにおける理想反応は

$$\eta_j^{DINA}(\boldsymbol{\alpha}_l) = \prod_{k=1}^{K} \alpha_{lk}^{q_{jk}} \tag{5.1}$$

で定義される．ここで，0^0 は 1 と定義する．理想反応 $\eta_j(\boldsymbol{\alpha}_l)$ は，l 番目のアトリビュート習得パタンが項目 j で測定されているアトリビュートをすべて習得しているものであれば 1，そうでなければ 0 を取る変数であり，誤差がない場合の l 番目のアトリビュート習得パタンの項目 j への正答・誤答を表している．さて，この理想反応と実際の項目反応とのハミング距離(Hamming distance) $d_h(\cdot)$ を

$$d_h(\boldsymbol{x}_i, \boldsymbol{\eta}^{DINA}(\boldsymbol{\alpha}_l)) = \sum_{j=1}^{J} |x_{ij} - \eta_j^{DINA}(\boldsymbol{\alpha}_l)| = \sum_{j=1}^{J} \left| x_{ij} - \prod_{k=1}^{K} \alpha_{lk}^{q_{jk}} \right| \tag{5.2}$$

とする．ただし，$\boldsymbol{\eta}^{DINA}(\boldsymbol{\alpha}_l) = (\eta_1^{DINA}(\boldsymbol{\alpha}_l), \ldots, \eta_J^{DINA}(\boldsymbol{\alpha}_l))^\top$ である．ノンパラメトリック推定法は，$\boldsymbol{x}_i, \boldsymbol{\eta}_i$ 間の距離を適当に定め，

$$\hat{\boldsymbol{\alpha}}_i = \underset{\boldsymbol{\alpha}_l}{\operatorname{argmin}} \{ d_h(\boldsymbol{x}_i, \boldsymbol{\eta}^{DINA}(\boldsymbol{\alpha}_l)) \} \tag{5.3}$$

と距離関数を最小化するアトリビュート習得パタンを推定値とする方法である．距離関数の定義として，重み付きハミング距離(weighted Hamming distance)を用いることもでき，その場合の距離関数 $d_{wh}(\cdot)$ は

$$d_{wh}(\boldsymbol{x}_i, \boldsymbol{\eta}^{DINA}(\boldsymbol{\alpha}_l)) = \sum_{j=1}^{J} \frac{1}{\bar{p}_j(1-\bar{p}_j)} |x_{ij} - \boldsymbol{\eta}_j(\boldsymbol{\alpha}_l)| = \sum_{j=1}^{J} \frac{1}{\bar{p}_j(1-\bar{p}_j)} \left| x_{ij} - \prod_{k=1}^{K} \alpha_{ik}^{q_{jk}} \right| \tag{5.4}$$

と定義される．ここで $\bar{p}_j = \frac{1}{I}\sum_{i=1}^{I} x_{ij}$ で定義される項目 j の正答確率である．なお，距離関数を最小化するアトリビュート習得パタンが複数ある場合には，それらのなかから一様にランダムで一つのパタンを選択するという操作が必要になる．この手続きは一般化ノンパラメトリック推定法においても必要である．

5.2 一般化ノンパラメトリック推定法

一般化ノンパラメトリック推定法では DINA モデルの理想反応に加えて，DINO モデルの理想反応

$$\eta_j^{DINO}(\boldsymbol{\alpha}_l) = 1 - \prod_{k=1}^{K} (1 - \alpha_{lk})^{q_{jk}} \tag{5.5}$$

を用いる．$\eta_j^{DINO}(\boldsymbol{\alpha}_l)$ は項目 j で必要なアトリビュートのうち少なくとも 1 つ習得しているパタンであれば 1，そうでなければ 0 となる．すなわち，$\eta_j^{DINO}(\boldsymbol{\alpha}_l) = 0$ となるのは項目 j で必要なアトリビュートを全く習得していない場合とも考えられる．さらに，重み w_{lj}(ただし，$0 \leq w_{lj} \leq 1$)を導入し，重み付き理想反応(weighted ideal response) $\eta_j^{(w)}(\boldsymbol{\alpha}_l)$ を

$$\eta_j^{(w)}(\boldsymbol{\alpha}_l) = w_{lj}\eta_j^{DINA}(\boldsymbol{\alpha}_l) + (1 - w_{lj})\eta_j^{DINO}(\boldsymbol{\alpha}_l) \tag{5.6}$$

と定義する．重み w_{lj} は項目 j の反応が DINA 的か DINO 的かどうかをあらわしている．

この重み付き理想反応と項目反応の距離関数を

$$d_g(\boldsymbol{x}_j, \eta_j^{(w)}(\boldsymbol{\alpha}_l)) = \sum_{i=1}^{I} \mathcal{I}(\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_l)(x_{ij} - \eta_j^{(w)}(\boldsymbol{\alpha}_l))^2 \tag{5.7}$$

とする．このとき，上記の距離関数を最小化する重みは

$$\hat{w}_{lj} = \frac{\sum_{i=1}^{I} \mathcal{I}(\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_l)(x_{ij} - \eta_j^{DINO}(\boldsymbol{\alpha}_l))^2}{(\sum_{i=1}^{I} \mathcal{I}(\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_l))(\eta_j^{DINA}(\boldsymbol{\alpha}_l) - \eta_j^{DINO}(\boldsymbol{\alpha}_l))} \tag{5.8}$$

となる．この推定された重みを用いた重み付き理想反応を $\hat{\eta}_j^{(w)}(\boldsymbol{\alpha}_l)$ とし，アトリビュート習得パタンを推定するための損失関数を

$$d_g(\boldsymbol{x}_i, \hat{\boldsymbol{\eta}}^{(w)}(\boldsymbol{\alpha}_l)) = \sum_{j=1}^{J} (x_{ij} - \hat{\eta}_j^{(w)}(\boldsymbol{\alpha}_l))^2 \tag{5.9}$$

とする．ただし，$\hat{\boldsymbol{\eta}}^{(w)}(\boldsymbol{\alpha}_l) = (\hat{\eta}_1^{(w)}(\boldsymbol{\alpha}_l), \ldots, \hat{\eta}_J^{(w)}(\boldsymbol{\alpha}_l))^\top$ である．上記の重みを用いて，一般化ノンパラメトリック推定法によるアトリビュート習得パタンは

$$\hat{\boldsymbol{\alpha}}_i = \underset{\boldsymbol{\alpha}_l}{\operatorname{argmin}} \{d_g(\boldsymbol{x}_i, \hat{\boldsymbol{\eta}}^{(w)}(\boldsymbol{\alpha}_l))\} \tag{5.10}$$

と推定される．

以上で見たように，一般化ノンパラメトリック推定法においては，まずアトリビュート習得パタンの初期値を決め，次に重みパラメタを式(5.8)により更新する．この更新された重みを用いて式(5.10)によりアトリビュート習得パタンを更新し，再度重み w_{lj} の更新を行う．このように，重みとアトリビュート習得パタンの更新を，アトリビュート習得パタンの変化が十分に小さくなるまで反復するのが一般化ノンパラメトリック推定法の手続きである．アトリビュート習得パタンの初期値としては通常のノンパラメトリック推定法による推定値を用いる．

このように一般化ノンパラメトリック推定法においては，補償的な(あるいは disjunctive) DCM の理想反応と非補償的な(あるいは conjunctive な)ものを組み合わせる方法を用いている．パラメトリック法においても，Yamaguchi and Okada (2020a)は DINA モデルと DINO モデルの項目反応関数の混合モデルとしての Hybrid DCM を提案しており，一般化ノンパラメトリック推定法とも類似した考え方をしている．

6. 推定法の選択

ここまで，最尤推定法，ベイズ推定法，ノンパラメトリック推定法の観点から DCM のパラメタ推定についてレビューを行った．本節では，これらのパラメタ推定法について，どのような場合にどのような方法を選択するのが望ましいのか，パラメタ推定法の特徴を踏まえて議論を行う．特に，サンプルサイズの大小とテスト項目数(多い・少ない)の組み合わせの観点から整理を行う．

まず，小サンプル・小項目数の状況は，例えばクラスルームの簡単な確認テストのような状況とみなすことができる．DCM が学習改善のための情報をテストから引き出すための統計モデルであることに鑑みると，小サンプル・小項目数の状況における DCM のパラメタ推定を考察することに意義はあるだろう．ただし，この状況はデータが多くないため，そもそも複雑な統計モデルの推定が容易ではない状況であるという点に留意する必要がある．この状況では，サンプルサイズが小さく，また項目数も多くないため最尤推定ベースの推定法の適用は難しいだろう．この場合には，計算量が多くなったとしても，事前情報を含めたフルベイズ推定を行うのが望ましい可能性がある．ただし，ベイズ推定法の利用のためにはパラメトリックモデル

の特定が必要である点に注意が必要である．また，データが少ないため，事前情報が強すぎるとデータからアトリビュート推定を行う意味も薄くなってしまう点にも注意しなければならない．ノンパラメトリック推定法は，データが少ない場合でもアトリビュート習得パタンの推定がある程度正確であることが数値実験で示されているが，Q 行列の特定が正しい必要がある．項目数が少ない場合には，ノンパラメトリック推定法であっても，アトリビュート習得パタンの推定が安定しにくい可能性もあるので，やはり注意が必要だろう．小サンプル・小項目数の状況でパラメトリックモデルを用いたい場合には，事前に項目パラメタを推定済の項目プールを用意しおくことなど，工夫が必要になると思われる．

次に，小サンプル・大項目数の状況は，例えばクラスルームの定期テストや 1 年の学習成果を振り返るようなまとめテストや定期テストなどの状況が考えられる．多くの項目数が利用できるという点では，先のノンパラメトリック推定法の利用が第 1 の選択肢に挙がるだろう．特に，制約がより少ない一般化ノンパラメトリック推定法が利用しやすいと考えられる．また，小サンプル・小項目数の状況と同様に，ベイズ推定法も候補に挙がる．一つ一つの項目パラメタの推定には不確実性がありながらも，ベイズ推測においてはそれらを考慮して，アトリビュート習得パタンの推定ができると考えられるからである．ただし，MAP 推定は，アトリビュート習得パタンの推定の際に，最尤推定のように項目パラメタの点推定値を使うため，項目パラメタの推定の不確実性は考慮しにくいと考えられる．もし，アトリビュートの数が多い場合や推定時間を短縮したい場合には，変分ベイズ推定法も利用できるだろう．最尤推定法は，やはり項目パラメタの推定の際にサンプルサイズが小さいことが影響し，正確な項目パラメタの推定が難しい可能性がある．また，アトリビュート習得パタンの推定には，項目パラメタ推定値を固定する必要があるが，項目パラメタの不確実性を考慮できないため，最尤推定法の利用には慎重になる必要がある．

さらに，大サンプル・小項目数の状況は，やや想定しにくいが，行政が行う標本調査の一部に，簡単な学力調査が含まれているような状況などが考えられる．この場合，個々人のアトリビュート習得パタンの推定を行うよりも，出題した項目の項目パラメタの特性や混合比率パラメタの推定に興味があると思われる．こうした状況であれば，最尤推定ベースの方法は，項目パラメタの点推定値だけではなく，漸近論に依拠した標準誤差も利用できると考えられるため，周辺最尤推定法を用いた方法が良い可能性もある．また，やや探索的にデータ分析を行いたい場合には，正則化推定法を用いてもよいだろう．MCMC を用いたベイズ的な方法は，サンプルサイズが大きい状況では計算時間が長くなる傾向があるものの，最尤推定法とあまりかわらない推定値が得られる可能性もあるので，積極的に用いる必要はないかもしれない．ただし，事前情報を積極的に活用した場合はその限りではない．この場合には，変分ベイズ推定法や MAP 推定法は，計算速度が MCMC より早いため，利用してもよいだろう．ノンパラメトリック推定法は，その目的に鑑みて，大サンプル・小項目数の推定には適さないのではないかと考えられる．ただし，他の推定法の初期値に用いるアトリビュート習得パタンをノンパラメトリック推定法により推定するという方略は，他の推定法の収束を早めることに繋がると期待できる．

最後に，大サンプル・大項目数の状況は，PISA（Programme for International Student Assessment）調査や国際数学・理科教育動向調査（Trends in International Mathematics and Science Study [TIMSS]）といった国際調査のような状況が思いつく．この場合，先の大サンプル・小項目数の状況と類似して，個々人のアトリビュート習得パタンの推定は必ずしも最終的な目的ではないと考えられる．また，この場合アトリビュート数が多くなるなど，モデルの規模も大きくなると考えられる．こうした場合には，まず許容できる待ち時間で推定が実行できることが望ましいので，変分ベイズ推定法や同時最尤推定法が候補に挙がる．ただし，同時最尤推定法

は点推定値しか与えない方法であるので，標準誤差の算出は別途行う必要がある．周辺最尤推定法や MCMC 法による推定法は，計算コストが高くなりやすいで，大サンプル・大項目数かつモデルが複雑な状況は使いにくいかもしれない．ノンパラメトリック推定法は大サンプル・小項目数の状況と同様に初期値を推定する方法としての活用が考えられる．

以上のように，テストの分析状況や目的によって使いやすい推定法が変わることがわかる．上記の議論から，大雑把に言って，サンプルサイズが小さい場合には，ノンパラメトリック推定法や事前情報を加味できるベイズ推定法が使いやすく，サンプルサイズが大きい場合には計算効率も考慮して，変分ベイズ推定法や同時・周辺最尤推定法が使いやすいと考えられる．

7. 総合考察

本稿では DCM のモデルパラメタ推定法について，現在提案されている方法を概観した．推定法としては，主として最尤推定法，ベイズ推定法，およびノンパラメトリック推定法の 3 つの方法を軸に発展がなされていた．最尤推定法においては，正則化推定法の活用がみられ，ベイズ推定法においては変分ベイズ推定法などの発展がみられた．正則化推定法や変分ベイズ推定法など機械学習でも用いられている方法も積極的に利用されており，活発に研究が進んでいることがうかがえる．本研究でレビューを行った方法により，DCM においては基本的なパラメタ推定についての研究は概ねしつくされていると考えられる．

以下では，推定法についての残された課題について議論を行う．特に，発展的なモデルのパラメタ推定，効率的な推定アルゴリズムの開発，より現実的な状況に即した推定法の開発の必要性といった観点から議論する．

本研究でレビューした方法は，多値反応や多値アトリビュート習得を想定した場合に拡張されている場合や，アトリビュート階層構造を含めた推定法につながる方法も多い．特に，診断情報をより精緻にするために，反応時間やアンケート調査の状況などのテストへの解答パタン以外の外部情報を取り入れたモデルの推定法も必要となる．こうしたモデルは一般にパラメタの数が増え，複雑化されていくと考えられる．そうした場合にも対応できるような効率的な推定法の開発は今後重要な観点である．特に，ある時点までの診断情報を捉えつつ，その後の診断を行ううえでは，最尤推定法よりもベイズ推定によるパラメタ更新を行えるような枠組みが望ましいだろう．特にベイズ推定の中でも，計算量の少ない変分ベイズ推定法などは今後より活用が期待できる推定法と考えられる．

こうした点に関連して，効率的な推定アルゴリズムの開発も必要となる．教育現場への情報機器の普及に伴い，タブレット端末などを用いた学習も一般的になりつつある．こうした現状においては，学習者のテストの反応パタン以外の普段の学習状況などのログデータも大量に取得できると考えられる．また，学習の進展に伴い，アトリビュートの数も次第に増加することが想定できる．こうした複雑なモデルかつ大量のデータを用いて診断情報を即時的に提供するためには，効率的な推定アルゴリズムの開発が不可欠になると考えられる．この意味で，ノンパラメトリック推定法などは高速に機能する方法として，有用であると考えられる．また，パラメトリックな方法においても，変分ベイズ推定法のように，比較的計算量が少ない方法なども提案されているものの，大規模なデータ状況下でも効率的に機能する方法を追求する意義は大きいだろう．

さらに，DCM はその目的から，教室場面など比較的サンプルサイズが小さい状況での利用も未だに一般的である．先の大量のデータにも対応できる推定法に加えて，こうした小サンプルでも，アトリビュート習得パタンを安定して推定できる方法の開発も必要であろう．ノンパラメトリック推定法は，こうした状況でも機能することがシミュレーションにより示されてい

るが，パラメトリックな方法においてもこうした方法が望まれる．

本研究の限界として，Q 行列の推定法についてはレビューを行わなかった点が挙げられる．Q 行列を含めたモデルの推定法も近年活発に研究がなされており，Q 行列推定は DCM の理論・応用の両側面から重要な観点と考えられる．Q 行列推定はモデルパラメタの識別性とも関係し，理論的にも高度な話題であり，DCM の応用研究者がそれらの原典を参照することは必ずしも容易ではない．また，DCM 研究者にとっても，先端的な課題として Q 行列推定の現状を日本語で参照できる包括的なレビューの価値は高いと考えられる．今後は，Q 行列推定の方法の展開の包括的なレビューが望まれる．

加えて，DCM は制約付き潜在クラスモデルとして，識別性など理論的な研究もなされている．後述の Q 行列の推定の話題とも関連して，DCM を現実的な場面で利用するためにも，DCM の理論的な性質の検討は重要である．こうした理論的な発展の現状の解説も必要となるだろう．

さらに，関連して，アトリビュート階層構造の推定法についても重要な研究関心である．本研究では，構造化されていないアトリビュート習得パタンを想定したが，実際にはアトリビュートには習得の順序関係などが想定される．こうしたアトリビュートの習得の順序関係について，理論的な想定のみならずデータから推定を行う方法も，DCM の応用を促す上で必要な観点である．こうしたアトリビュート階層構造の推定法や関連する話題についても包括的なレビューが必要と考えられる．

謝　辞

本研究は JSPS 科研費基盤研究（A）23H00065，基盤研究（B）20H01720，21H00936，23H00985，若手研究 22K13810 の助成を受けたものです．

参 考 文 献

Balamuta, J. J. and Culpepper, S. A. (2022). Exploratory restricted latent class models with monotonicity requirements under POLYA–GAMMA data augmentation, *Psychometrika*, **87**(3), 903–945, https://doi.org/10.1007/s11336-021-09815-9.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*, **4**, Springer, New York.

Blei, D. M., Kucukelbir, A. and McAuliffe, J. D. (2017). Variational inference: A review for statisticians, *Journal of the American statistical Association*, **112**(518), 859–877, https://doi.org/10.1080/01621459.2017.1285773.

Cai, L. (2008). SEM of another flavour: Two new applications of the supplemented EM algorithm, *British Journal of Mathematical and Statistical Psychology*, **61**(2), 309–329, https://doi.org/10.1348/000711007X249603.

Chen, Y., Liu, J., Xu, G. and Ying, Z. (2015). Statistical analysis of Q-matrix based diagnostic classification models, *Journal of the American Statistical Association*, **110**(510), 850–866, https://doi.org/10.1080/01621459.2014.934827.

Chen, Y., Li, X., Liu, J. and Ying, Z. (2017). Regularized latent class analysis with application in cognitive diagnosis, *Psychometrika*, **82**, 660–692, https://doi.org/10.1007/s11336-016-9545-6.

Chen, Y., Culpepper, S. A., Chen, Y. and Douglas, J. (2018). Bayesian estimation of the DINA Q-matrix, *Psychometrika*, **83**, 89–108, https://doi.org/10.1007/s11336-017-9579-4.

Chen, Y., Culpepper, S. A. and Liang, F. (2020). A sparse latent class model for cognitive diagnosis, *Psychometrika*, **85**, 121–153, https://doi.org/10.1007/s11336-019-09693-2.

Chiu, C.-Y. and Douglas, J. (2013). A nonparametric approach to cognitive diagnosis by proxim-

ity to ideal response patterns, *Journal of Classification*, **30**, 225–250, https://doi.org/10.1007/s00357-013-9132-9.

Chiu, C.-Y. and Köhn, H.-F. (2019). Consistency theory for the general nonparametric classification method, *Psychometrika*, **84**, 830–845, https://doi.org/10.1007/s11336-019-09660-x.

Chiu, C.-Y., Köhn, H.-F., Zheng, Y. and Henson, R. (2016). Joint maximum likelihood estimation for diagnostic classification models, *Psychometrika*, **81**, 1069–1092, https://doi.org/10.1007/s11336-016-9534-9.

Chiu, C.-Y., Sun, Y. and Bian, Y. (2018). Cognitive diagnosis for small educational programs: The general nonparametric classification method, *Psychometrika*, **83**, 355–375, https://doi.org/10.1007/s11336-017-9595-4.

Cho, A. E., Wang, C., Zhang, X. and Xu, G. (2021). Gaussian variational estimation for multidimensional item response theory, *British Journal of Mathematical and Statistical Psychology*, **74**, 52–85, https://doi.org/10.1111/bmsp.12219.

Cho, A. E., Xiao, J., Wang, C. and Xu, G. (2024). Regularized variational estimation for exploratory item factor analysis, *Psychometrika*, **89**, 347–375, https://doi.org/10.1007/s11336-022-09874-6.

Culpepper, S. A. (2019). An exploratory diagnostic model for ordinal responses with binary attributes: Identifiability and estimation, *Psychometrika*, **84**(4), 921–940, https://doi.org/10.1007/s11336-019-09683-4.

de la Torre, J. (2009). DINA model and parameter estimation: A didactic, *Journal of Educational and Behavioral Statistics*, **34**(1), 115–130, https://doi.org/10.3102/1076998607309474.

de la Torre, J. (2011). The generalized DINA model framework, *Psychometrika*, **76**, 179–199, https://doi.org/10.1007/s11336-011-9207-7.

Grimmer, J. (2011). An introduction to Bayesian inference via variational approximations, *Political Analysis*, **19**(1), 32–47, https://doi.org/10.1093/pan/mpq027.

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, https://doi.org/10.1007/s11336-008-9089-5.

Hoijtink, H. (1998). Constrained latent class analysis using the Gibbs sampler and posterior predictive p-values: Applications to educational testing, *Statistica Sinica*, 691–711, https://www.jstor.org/stable/24306458.

Hong, C.-Y., Chang, Y.-W. and Tsai, R.-C. (2016). Estimation of generalized DINA model with order restrictions, *Journal of Classification*, **33**, 460–484, https://doi.org/10.1007/s00357-016-9215-5.

Jamshidian, M. and Jennrich, R. I. (2000). Standard errors for EM estimation, *Journal of the Royal Statistical Society*: *Series B* (*Statistical Methodology*), **62**(2), 257–270, https://doi.org/10.1111/1467-9868.00230.

Jeon, M., Rijmen, F. and Rabe-Hesketh, S. (2017). A variational maximization–maximization algorithm for generalized linear mixed models with crossed random effects, *Psychometrika*, **82**, 693–716, https://doi.org/10.1007/s11336-017-9555-z.

Jimenez, A., Balamuta, J. J. and Culpepper, S. A. (2023). A sequential exploratory diagnostic model using a Pólya-gamma data augmentation strategy, *British Journal of Mathematical and Statistical Psychology*, **76**(3), 513–538, https://doi.org/10.1111/bmsp.12307.

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, https://doi.org/10.1177/0146621012203206.

Laudy, O., Boom, J. and Hoijtink, H. (2004). Bayesian computational methods for inequality constrained latent class analysis, *New Developments in Categorical Data Analysis for the Social and Behavioral Sciences*, 63–82, Psychology Press, New York, https://doi.org/10.4324/9781410612021.

Leighton, J. P., Gierl, M. J. and Hunka, S. M. (2004). The attribute hierarchy method for cognitive assessment: A variation on Tatsuoka's rule-space approach, *Journal of Educational Measurement*,

41(3), 205–237, https://doi.org/10.1111/j.1745-3984.2004.tb01163.x.

Liu, X. and Johnson, M. S. (2019). Estimating CDMs using MCMC, *Handbook of Diagnostic Classification Models*: *Models and Model Extensions, Applications, Software Packages* (eds. M. von Davier and Y.-S. Lee), Chapter 31, 629–646, Springer, New York.

Liu, Y., Xin, T., Andersson, B. and Tian, W. (2019). Information matrix estimation procedures for cognitive diagnostic models, *British Journal of Mathematical and Statistical Psychology*, **72**(1), 18–37, https://doi.org/10.1111/bmsp.12134.

Ma, C., de la Torre, J. and Xu, G. (2023a). Bridging parametric and nonparametric methods in cognitive diagnosis, *Psychometrika*, **88**(1), 51–75, https://doi.org/10.1007/s11336-022-09878-2.

Ma, C., Ouyang, J. and Xu, G. (2023b). Learning latent and hierarchical structures in cognitive diagnosis models, *Psychometrika*, **88**(1), 175–207, https://doi.org/10.1007/s11336-022-09867-5.

Ma, W. and Jiang, Z. (2021). Estimating cognitive diagnosis models in small samples: Bayes modal estimation and monotonic constraints, *Applied Psychological Measurement*, **45**(2), 95–111, https://doi.org/10.1177/0146621620977681.

Meng, X.-L. and Rubin, D. B. (1991). Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm, *Journal of the American Statistical Association*, **86**(416), 899–909.

Nakajima, S., Watanabe, K. and Sugiyama, M. (2019). *Variational Bayesian Learning Theory*, Cambridge University Press, Cambridge.

Philipp, M., Strobl, C., de la Torre, J. and Zeileis, A. (2018). On the estimation of standard errors in cognitive diagnosis models, *Journal of Educational and Behavioral Statistics*, **43**(1), 88–115, https://doi.org/10.3102/1076998617719728.

Rupp, A. A. and Templin, J. (2008a). The effects of Q-matrix misspecification on parameter estimates and classification accuracy in the DINA model, *Educational and Psychological Measurement*, **68**(1), 78–96, https://doi.org/10.1177/0013164407301545.

Rupp, A. A. and Templin, J. (2008b). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art, *Measurement*, **6**(4), 219–262, https://doi.org/10.1080/15366360802490866.

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

鈴木雅之, 豊田哲也, 山口一大, 孫 媛 (2015). 認知診断モデルによる学習診断の有用性の検討, 日本テスト学会誌, **11**(1), 81–97, https://doi.org/10.24690/jart.11.1_81.

Tatsuoka, K. K. (1985). A probabilistic model for diagnosing misconceptions by the pattern classification approach, *Journal of Educational Statistics*, **10**(1), 55–73, https://doi.org/10.3102/10769986010001055.

Templin, J. and Hoffman, L. (2013). Obtaining diagnostic classification model estimates using Mplus, *Educational Measurement*: *Issues and Practice*, **32**(2), 37–50, https://doi.org/10.1111/emip.12010.

Templin, J. L. and Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models, *Psychological Methods*, **11**(3), 287–305, https://doi.org/10.1037/1082-989X.11.3.287.

von Davier, M. and Lee, Y.-S. (2019). *Handbook of Diagnostic Classification Models*, Springer, New York.

Wang, C. (2021). Using penalized EM algorithm to infer learning trajectories in latent transition CDM, *Psychometrika*, **86**(1), 167–189, https://doi.org/10.1007/s11336-020-09742-1.

Wang, C. and Lu, J. (2021). Learning attribute hierarchies from data: Two exploratory approaches, *Journal of Educational and Behavioral Statistics*, **46**(1), 58–84, https://doi.org/10.3102/1076998620931094.

Wang, S. and Douglas, J. (2015). Consistency of nonparametric classification in cognitive diagnosis, *Psychometrika*, **80**, 85–100, https://doi.org/10.1007/s11336-013-9372-y.

Xu, G. (2017). Identifiability of restricted latent class models with binary responses, *The Annals of Statistics*, **45**(2), 675–707, https://doi.org/10.1214/16-AOS1464.

Xu, G. and Shang, Z. (2018). Identifying latent structures in restricted latent class models, *Journal of*

the American Statistical Association, **113**(523), 1284–1295, https://doi.org/10.1080/01621459.2017.1340889.

Xu, X., De la Torre, J., Zhang, J., Guo, J. and Shi, N. (2020). Estimating CDMs using the slice-within-gibbs sampler, *Frontiers in Psychology*, **11**, 2260, https://doi.org/10.3389/fpsyg.2020.02260.

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**(1), 21–44, https://doi.org/10.24690/jart.15.1_21.

Yamaguchi, K. (2020). Variational Bayesian inference for the multiple-choice DINA model, *Behaviormetrika*, **47**(1), 159–187, https://doi.org/10.1007/s41237-020-00104-w.

Yamaguchi, K. (2023a). Bayesian analysis methods for two-level diagnosis classification models, *Journal of Educational and Behavioral Statistics*, **48**(6), 773–809, https://doi.org/10.3102/10769986231173594.

Yamaguchi, K. (2023b). On the boundary problems in diagnostic classification models, *Behaviormetrika*, **50**(1), 399–429, https://doi.org/10.1007/s41237-022-00187-7.

Yamaguchi, K. and Martinez, A. J. (2024). Variational Bayes inference for hidden Markov diagnostic classification models, *British Journal of Mathematical and Statistical Psychology*, **77**(1), 55–79, https://doi.org/10.1111/bmsp.12308.

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, https://doi.org/10.2333/jbhmk.44.181.

Yamaguchi, K. and Okada, K. (2020a). Hybrid cognitive diagnostic model, *Behaviormetrika*, **47**(2), 497–518, https://doi.org/10.1007/s41237-020-00111-x.

Yamaguchi, K. and Okada, K. (2020b). Variational Bayes inference algorithm for the saturated diagnostic classification model, *Psychometrika*, **85**(4), 973–995, https://doi.org/10.1007/s11336-020-09739-w.

Yamaguchi, K. and Okada, K. (2020c). Variational Bayes inference for the DINA model, *Journal of Educational and Behavioral Statistics*, **45**(5), 569–597, https://doi.org/10.3102/1076998620911934.

Yamaguchi, K. and Templin, J. (2022a). Direct estimation of diagnostic classification model attribute mastery profiles via a collapsed Gibbs sampling algorithm, *Psychometrika*, **87**(4), 1390–1421, https://doi.org/10.1007/s11336-022-09857-7.

Yamaguchi, K. and Templin, J. (2022b). A Gibbs sampling algorithm with monotonicity constraints for diagnostic classification models, *Journal of Classification*, **39**(1), 24–54, https://doi.org/10.1007/s00357-021-09392-7.

Zhang, Z., Zhang, J., Lu, J. and Tao, J. (2020). Bayesian estimation of the DINA model with Pólya-gamma Gibbs sampling, *Frontiers in Psychology*, **11**, 384, https://doi.org/10.3389/fpsyg.2020.00384.

Recent Development of Parameter Estimation Methods in Diagnostic Classification Models

Kazuhiro Yamaguchi

Division of Psychology, Institute of Human Sciences, University of Tsukuba

Diagnostic classification models (DCMs) or cognitive diagnostic models (CDMs) are a family of models that consider elements of cognitive abilities to solve test items named "attributes" and classify test takers into attribute mastery patterns. DCMs are a useful tool to analyze educational tests and have been applied to various tests. However, estimation methods of DCMs have not been assessed in Japan. This study aimed to review the current state of development of estimation methods of DCMs and to contribute to improving the application and theoretical development of DCMs. As a result, estimation methods were developed according to the maximum likelihood estimation, Bayesian estimation, and non-parametric estimation methods. In particular, regularization methods in the maximum likelihood estimation and variational Bayes, which are relatively new methods, were employed. Finally, the remaining estimation problems and future research topics in this area were discussed.

Key words: Diagnostic classification models, cognitive diagnostic models, parameter estimation methods.

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

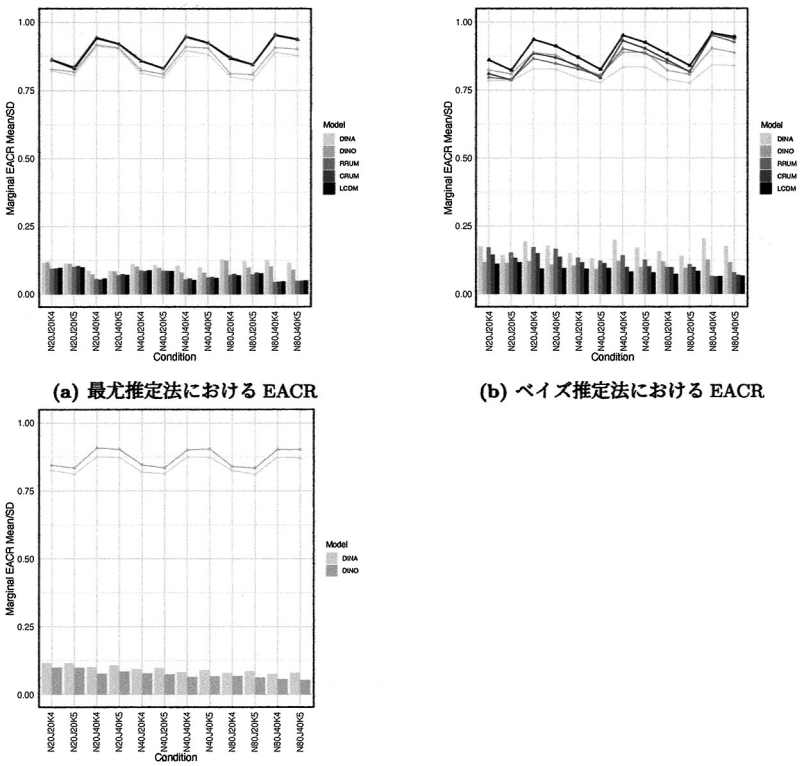
本節では，EACR と PACR の観点からアトリビュート習得パタンの推定精度の結果を示し，Bias と RMSE の観点から項目パラメタ，特に s_j，g_j の推定精度の結果を示す．さらに，各推定法で用いた情報量規準が選択したモデルの割合を示す．

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

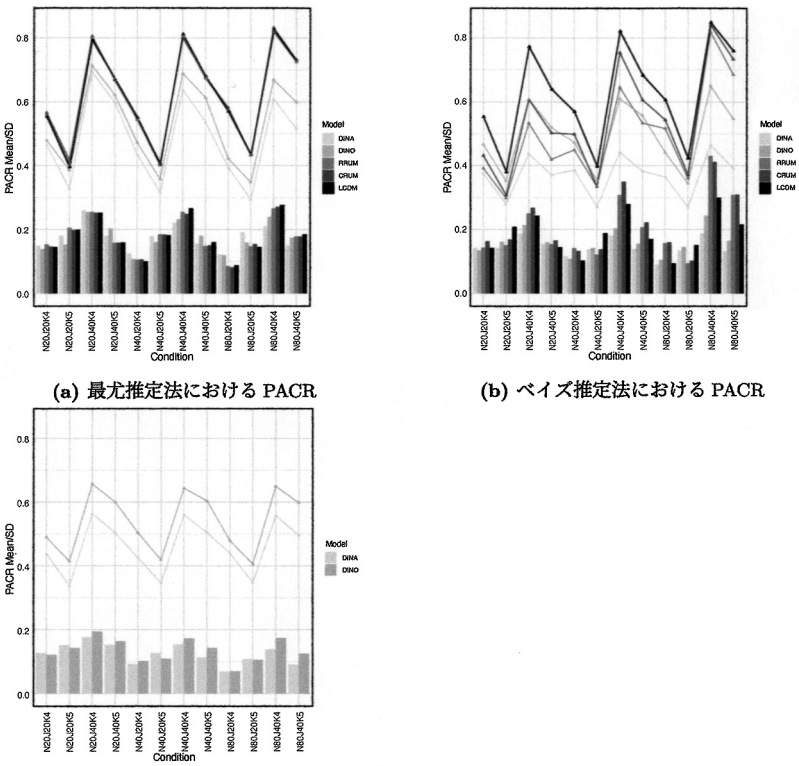
(b) ベイズ推定法における EACR

(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．

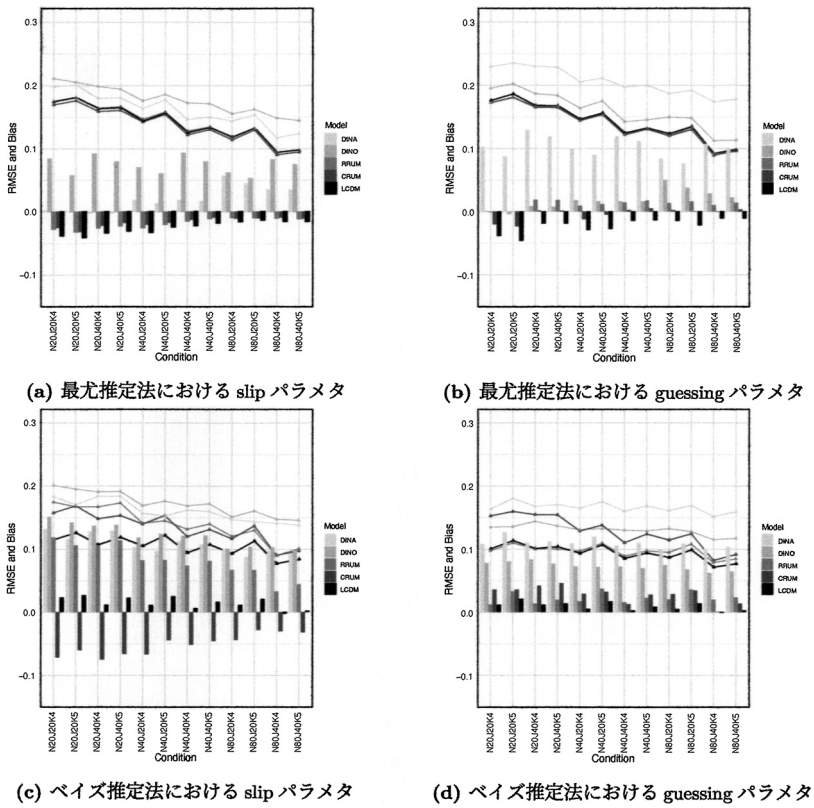
アトリビュート数の要因における η^2 と偏 η^2 は，PACR を従属変数とした場合に EACR を従属変数とした場合と比べて高い値を示した．図 1，2 における解釈も踏まえれば，アトリビュート数を減らすことが，PACR の推定精度を高めることを示唆する．これは，EACR がアトリビュート習得パタンの要素ごとの一致率を表す指標であったのに対して，PACR はアトリビュート習得パタンの一致率を表す指標であり，アトリビュート数が増えるにつれて可能なアトリビュート習得パタンの総数が指数関数的に増えることに起因すると考えられる．

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．

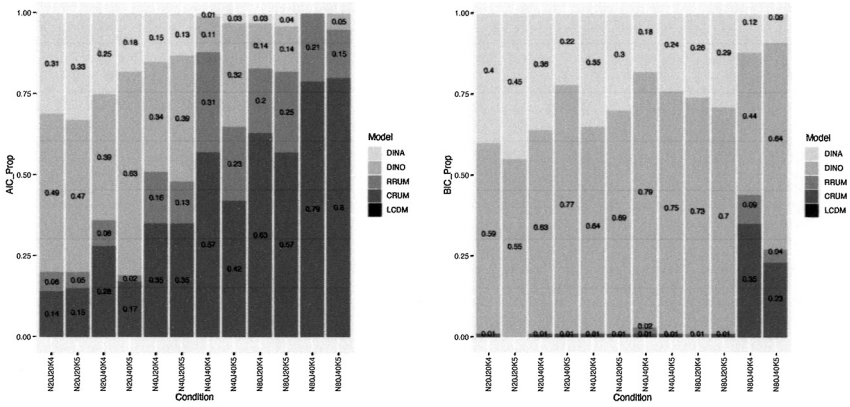
guessing パラメタに関して，DINA モデルが最も Bias の絶対値および RMSE が大きく，RRUM でそれらの値が最も小さかった．DINO モデルの場合と異なり，DINA モデルでは，項目の正答に必要とされるアトリビュートを全て習得している場合のみ理想反応が 1(i.e., 正答)となる．一方 LCDM では，仮に一つのアトリビュートのみを習得している場合でも，その主効果によって正答確率が高くなる．このような項目正答確率とアトリビュートの関係に関する仮定の違いを反映して，DINA モデルにおいて guessing パラメタの Bias が正の方向に大きくなった，つまり guessing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

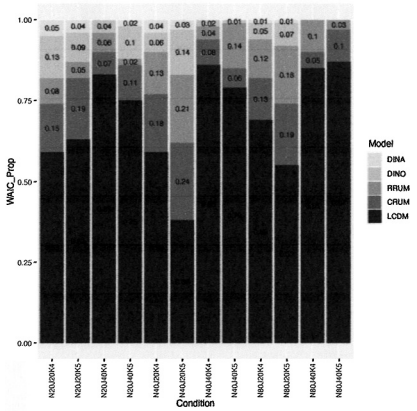


図 5. WAIC が最小となったモデルの割合.

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.

2023 年 11 月改訂

「統計数理」投稿規程

1. 「統計数理」は，統計科学の深化と発展，そして統計科学を通じた社会への貢献を目指すものである．投稿原稿は，統計科学に関連した内容を持つもので，和文の原稿に限る．
2. 投稿原稿は次の 6 種とする．
   a. 原著論文 (Paper)
      統計科学の発展に貢献すると考えられる研究結果.
   b. 総合報告 (Review Article)
      特定の主題に関する一連の研究およびその周辺領域の発展を著者の見解に従って総括的，かつ体系的に報告したもの.
   c. 研究ノート (Letter)
      研究速報，新しい発想，提言，問題提起，事例報告など研究上，記録にとどめておく価値があると認められるものや，既発表の論文等に対するコメントで，研究上，記録にとどめておく価値があると認められるもの.
   d. 研究詳解 (Research Review)
      特定の研究領域における理論的あるいは応用的成果を，最近の結果や知見を加えてわかりやすく説明したもの.
   e. 統計ソフトウェア (Statistical Software)
      有用な計算法や解析法に関する短いプログラムおよびサブルーチンのリスト，利用手引き，実行例など.
   f. 研究資料 (Research Archives)
      歴史的なデータ，入手困難なデータや統計的手法の比較検討のために有用なデータ，あるいは，歴史的文献の翻訳や解説など.
   いずれも原則として，未発表のものに限る.
3. 投稿された原稿は，編集委員会が選定・依頼した査読者の審査を経て，掲載の可否を決定する.
4. 投稿原稿は電子投稿査読システム https://www.editorialmanager.com/toukei/ より投稿するものとする．原稿は pdf ファイルとし，必要なフォントはすべて埋め込み，原稿全体を一つのファイルにまとめることとする．論文が採択になった場合，著者は最終稿のソースファイルとハードコピーを提出するものとする.
5. 著作権
   (1) 掲載される論文等の著作権はその採択をもって統計数理研究所に帰属するものとする．統計数理研究所は，紙媒体の「統計数理」のほか電子媒体などを通じて論文等を公表することができる．特別な事情がある場合は，著者と本編集委員会との間で協議の上措置する.
   (2) 投稿原稿の中で引用する文章や図表の著作権に関する問題は，著者の責任において処理する.
   (3) 著者が自分の論文等を複製，転載，翻訳，翻案等の形で利用するのは自由である．この場合，著者は掲載先に出典を明記する.
6. 原稿は次の執筆要項に従って作成する.

「統計数理」執筆要項

1. 原稿は A4 用紙に 1 行 36 字から 40 字で 1 行おき，1 頁あたり 22 行程度とする．原稿の長さは原則として表・図を含めて 30 頁相当以内とし，各ページにページ番号を付す．図表は別紙にまとめ，本文中には挿入箇所のみを指定する．LaTeX で原稿を作成する場合は，「統計数理」スタイルファイルの使用を推奨する.
   https://www.ism.ac.jp/editsec/toukei/
2. 文体は「である体」とし，句読点は「,」「.」を用いる.
3. 原稿は以下の順に書くものとする.
   ［第 1 頁］ 標題，著者名，所属名，和文要旨 (500 字程度，文献の引用および数式は原則として避ける)，和文キーワード (6 語以内).
   ［第 2 頁］ 英語による標題，著者名，所属名，Abstract (450 ワード程度)，Key words (6 words and phrases 以内)．Abstract は，問題の所在と得られた結果等がそれだけで理解できるようなものとする.

［第 3 頁以降］

① 本文：章，節の番号は，第 1 章にあたるものは，“1.”, 第 1 章 第 1 節にあたるものは，“1.1” というようにつける．また，式の番号は，章ごとに (2.1), (2.2) のようにし，式の左側に配置する．

② 数式：数式は簡明さを心がけ，添字にさらに添字をつけるのはなるべく避ける．

③ 参考文献：書き方は本要項 第 4 項を参照.

④ 表：一枚の用紙に一つの表を書く．表の番号は論文中に現れる順に従って，表 1，表 2,... または，Table 1, Table 2,... のようにする．

⑤ 図：一枚の用紙に一つの図を描く．図はそのまま写真製版できる鮮明なものを用意する．大きさは印刷出来上がりの 1〜2 倍とし，トレースが必要な場合は原則として著者が行うものとする．図の番号は論文中に現れる順に従って，図 1，図 2,... または，Fig. 1, Fig. 2,... のようにする．図は原則としてモノクロ印刷とするが，カラー印刷を必要とする場合は編集委員会に相談すること．

⑥ 注：本文中の注釈は極力避ける．やむを得ず注釈をつける場合は脚注とせず，論文末尾に後注とする．後注は，順番に “1, 2,...” の番号を付け，本文中では上付きで示す.

4. 本文中での参考文献の引用は，著者名 (出版年) とする．たとえば，Efron (1982), 清水・湯浅 (1984), Cox and Snell (1981), 坂元 他 (2004), Nakano et al. (2000).

5. 参考文献の書き方

［雑誌の場合］いずれの場合も雑誌名は省略しないものとする．

① ページ番号がある文献

著者名 (出版年)．標題，雑誌名，巻，ページ〔始–終〕.

【例】Chernoff, H. (1973). The use of faces to represent points in k-dimensional space graphically, *Journal of the American Statistical Association*, **68**, 361–368.

【例】Bligh, E. G. and Dyer, W. J. (1959). A rapid method of total lipid extraction and purification, *Canadian Journal of Biochemistry and Physiology*, **37**, 911–917, https://doi.org/10.1139/o59-099.

② ページ番号がない文献

著者名 (出版年)．標題，雑誌名，巻 に続けて，DOI があるものは DOI を入れ，DOI がない場合は URL と最終アクセス日を記載する.

【例】Lien, G. Y., Kalnay, E. and Miyoshi, T. (2013). Effective assimilation of global precipitation: Simulation experiments, *Tellus A,* **65**, https://doi.org/10.1175/WAF-D-13-00032.1.

【例】Kiraly, F. J. and Oberhauser, H. (2019). Kernels for sequentially ordered data, *Journal of Machine Learning Research*, **20**(31), http://jmlr.org/papers/v20/16-314.html（最終アクセス日 2023 年 10 月 1 日）.

［叢書の中の一巻の場合］

著者名 (出版年)．書名 (編集者名)，叢書名，発行所名，発行地名.

【例】Sakamoto, Y., Ishiguro, M. and Kitagawa, G. (1983). *Akaike Information Criterion Statistics*, Mathematics and Its Applications, Reidel, Dordrecht.

［単行本等の場合］

著者名 (出版年)．書名，発行所名，発行地名.

【例】Cressie, Noel (1993). *Statistics for Spatial Data*, Wiley, New York.

［編集書の中の一部の場合］

著者名 (出版年)．標題，編集書名 (編集者名)，巻，ページ，発行所名，発行地名.

【例】Akaike, H. (1980). Likelihood and the Bayes procedure, *Bayesian Statistics* (eds. J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith), 143–166, University Press, Valencia, Spain.

なお，同じ著者によるものが同一年に複数個現れる場合には，(1980a), (1980b) などとして区別する．文献は，日本人も含め，著者名のアルファベット順に並べる.

6. 著者校正は原則として一回とする．その際，印刷上の誤り以外の字句や図版の訂正，挿入，削除等は原則として認めない.