博士論文

認知診断モデルの理論的および経験的検討

山口一大

目次

図表番号

第1章　序論

本章では，本論文の主題となる認知診断モデルを検討する前に，これまでの教育測定モデルを検討し，認知診断モデルの特徴を述べる。以下，1-1 ではテストを用いた診断情報の必要性について，近年の心理学研究の文脈や日本での形成的評価に言及しながら述べる。1-2 では，これまでに行われてきたテスト理論研究の中で認知能力がどのように統計モデルの中で扱われてきたのか，その歴史的展開を示す。その後，1-3 で認知診断モデルの萌芽的研究を示す。さらに，1-4 で認知能力のモデリングのテスト理論以外での展開を述べ，認知能力のモデリングの潮流を確認する。これを踏まえて，1-5 で CDM の基本的な発想を示す。さらに，1-6 では CDM を応用する際の手続きを示す。さらに 1-7 で CDM の応用研究の例を示し，CDM がどのように利用されているかを確認する。これに加え。1-8 で CDM の比較研究の現状を示す。最後に，1-9 で本論文の目的と全体の構成を示す。

1-1 テストを用いた診断情報の必要性

テストには，定期試験や高校・大学入試，あるいは語学の能力を示すための試験などがある。こうしたテストは，身の回りに数多く存在している。これらのテストは能力の高低を示すためのものとして認識されている。正答・誤答といったテストへの解答者の解答反応データを通じて，人間の能力をモデル化するために，これまでに数多くの研究がなされてきた。こうしたテスト項目への反応を分析するための理論をテスト理論という。

テスト理論は古典的テスト理論（classical test theory, CTT; e.g., Traub, 1997）から項目反応理論（item response theory, IRT; Lord & Novick, 1968）へ発展してきた。CTT では，観測得点は真値と誤差の和で構成されるという極めてシンプルな定式化がなされている。また CTT の枠組みでは，観測得点の分散に占める真値の分散の割合としてテストの信頼性が定義されている。CTT はテストそれ自体が持つ測定の精度の評価を可能にした理論である。しかしながら，CTT は集団に依存する理論であり，項目の特性とテスト解答者の特性を分離できていない。これはつまり，測定する集団が異なった場合には，テスト得点の比較が必ずしも行えないということを意味する。IRT モデルは CTT の欠点を克服し，項目の識別力，困難度といった項目の特性とテスト解答者の能力の分離に成功し，同じテスト項目で構成されたテストを受験していない複数の集団の比較を可能にした。さらに IRT モデルにより，テストの困難度を統制し，等質なテストセットの作成（uniform test assembly; e.g., Belov, 2008）や，コンピュータを用いた適応型テスト（computer adaptive testing, CAT; e.g., Meijer & Nering,

1999）の実用化にも大きく貢献した。こうしたテスト理論の隆盛の中，計算ルールの適用や演算の実行などのテストに解答する際の認知プロセスへの注目が集まっていくこととなる。詳細は次節で述べるが，この流れのもと，認知能力を問題解答の際に正確に適用できない解答者や誤った計算ルールを実行してしまう解答者の特異な反応パタンの分析がなされてきた。こうした分析をバグ分析といい，関連する研究（e.g., Tatsuoka & Tatsuoka, 1987）が盛んに行われた。これらの研究から，解答者を問題正答に必要な複数の認知能力の習得・未習得で分類するという現在の認知診断モデル（cognitive diagnostic models, CDM; Leighton & Gierl, 2007）の体系が整理されてきた。

後述するように，CDM は制約がついた潜在クラス分析（latent class analysis, LCA; Lazarsfeld & Henry, 1968）であり，LCA の特殊なモデルとみなされている（Rupp & Templin, 2008b）。CDM は一般化線形潜在混合モデル（generalized linear latent and mixed models, GLLAMM; Skrondal & Rabe-Hesketh, 2004），あるいは潜在変数モデル（latent variable modeling, LV; e.g., Loehlin, 2004）の下位モデルとしても位置づけられる。IRT モデルと CDM の比較でいえば，IRT モデルは能力の一次元尺度の構成に重点を置いたものであり，CDM は解答者をそれぞれの認知能力の状態に応じて分類するためのモデルである。また，IRT は結果が解答者の人生に大きな影響を持つテスト（ハイステークスなテスト）に適用され，能力の推定精度の保証やテストの公平性を担保するために利用される。一方で，CDM は解答者間の比較・選抜を目的にしたものではなく，クラスルームテストやプレースメントテストなど比較的ローステークスなテストに適用され，テストから次に学習するべき内容を明らかにするために利用される。このように，IRT モデルと CDM はテスト利用の目的が大きく異なっている。統計モデルとして，IRT モデルは解答者の連続的な特性を推定することを目指している。一方，CDM はテスト項目の正答に必要な能力の習得・未習得のパタンを推定することを最終的な目的としている。この能力の習得パタンを用いることにより，テストの結果から次に学習すべき内容を提示することが可能となる。こうした解答者の学習状況を改善するために，その解答者の認知能力の状態を推定し，次に学習するべき内容を提案する一連の判断を「診断」と呼んでいる。これは医者が患者の状況を把握して，病気や症状にあった処置や投薬を行うということと同様である。学習改善のためのアドバイスは能力の習得パタンをもとに行うため，能力の習得パタンを正確に知ることは重要と考えられる。

このような診断を行うことにより，限られた学習時間を効果的に利用した学習計画を立てることができる。また，CDM を利用することにより，テストの設計，データ分析，フィ

ードバックを一貫した枠組みで実行可能となる。つまり，診断に用いる能力に対応する問題項目を作成し，テスト問題に解答するための認知プロセスに合致した CDM を選択し，それぞれの認知能力の習得・未習得のパタンをフィードバックする，というプロセスが一貫する点が CDM を利用する利点となる。

テスト理論のような解答者の能力に注目した理論体系が発展してきている文脈とは別に，発達心理学において個人志向アプローチ（person-oriented approach; Bergman & Lundh, 2015）の考え方が出現した。個人志向アプローチは例えば発達のパタンへの注目，個人への注目，プロセスへの注目，という特徴を持つアプローチである。従来，心理学研究では個人間の変数の比較を行う変数志向アプローチ（variable-oriented approach）であったものの，発達変化への興味の高まりから個人志向アプローチへの注目が高まっていると考えられる。個人志向アプローチでは，変数間の関連を記述するのではなく，個々人の発達の軌跡などに注目し，それらを説明する潜在的なパタンを見出すアプローチである。こうした個人への関心は 1980 年以前からあるものの，全体論的視点（holistic perspective; e.g., Magnusson, 1999）が注目され，それに伴い個人間の相互作用の中での発達が注目されてきた。こうした流れの中で，1990 年前後から徐々に個人志向アプローチと変数志向アプローチを対比させた議論がなされてきた。個人志向アプローチの歴史的なレビューとしては Bergman & Magnusson（1997）が詳しい。具体的にこの方法論を実現する統計的手法には，分類を行うクラスター分析が挙げられている（Bergman & Lundh, 2015）。クラスター分析のみならず，LCA も分類を行う手法であり，個人志向アプローチに即した手法である可能性がある。このように，心理学の特定の分野においても，変数間の関連でなく，個人を分類する視点への注目がなされている。個人志向アプローチからみると，CDM は教育・学習場面での個人の認知能力の習得パタンを詳細に記述するための方法論であり，個人志向アプローチを実現可能にするテストモデルと考えることができる。とくに CDM は事前に設定した認知能力とそれぞれの能力の習得・未習得のパタンに注目するため，クラスター分析や LCA とは異なり出力されてきた結果の解釈が容易である。この点は，他の分類を行うためのモデルとは異なった CDM の特徴と考えられる。CDM はテストへの適用のみならず，こうした発達心理学分野での応用の可能性も秘めている。

また，日本の学校教育での評価として診断的評価・形成的評価への注目が集まっている（植野, 2010）。学習評価の文脈では，学習の前に行われる評価を診断的評価，学習の途中に行われる評価を形成的評価と呼ぶが，近年では形成的評価を「学習のための評価」とみなす

ことがある（二宮, 2014）。形成的評価の目的は学習改善にあり，その目的を達成するためには学習の到達度のみならず，解答者の既有知識の状態を含めた判断が必要であるとされる（二宮, 2014）。CDM はその名に“診断”と含まれているものの，形成的評価のためにも利用可能である。なぜならば，CDM は解答者の認知能力の習得状態を明らかにできるからである。

解答者がいまどのような認知能力を習得しているのか・いないのかという認知状態の診断情報というものは学問的・実用的にも非常に重要な情報である。CDM は解答者の診断に特化した統計モデルであり，CDM により心理学研究のみならず教育・学習場面を含めて，多くの場面で解答者の認知能力の習得状態についての情報を得ることができる。こうした情報によって，解答者それぞれに適した学習方法を提案することが可能となる。心理学研究においても，より解答者にフォーカスし解答者の実態に迫る方法論として，CDM が利用できる。CDM を利用した心理学研究も徐々に増えつつある。例えば，Templin & Henson（2006）では CDM を病的ギャンブリング傾向の診断のために活用し，その診断に CDM が利用できることを示した。また，García, Olea, & de la Torre（2014）や Sorrel, Olea, Abad, de la Torre, Aguado, & Lievens（2016）は状況判断テスト（situational judgement tests）への CDM を適用し，状況判断能力の分類に CDM が有効であることを示した。このように，CDM はテストから解答者を分類するための手法として今後ますます利用されることが期待できる手法である。次節では，CDM の詳細に入るまえに，テスト理論でどのように認知能力をモデルの一部として表現してきたのか，その展開を検討する。

1-2 テスト理論における認知能力のモデリング

本節では，テスト理論の文脈でテスト問題を解答するための能力である認知能力がどのようにモデルに組み込まれてきたのかを検討する。これまで素朴に「認知能力」という用語を使用してきたが，本論文では，「テスト問題解答に必要な複数のきめ細かいスキルやプロセス」を「認知能力」と呼ぶこととする（e.g., de la Torre, 2009a）。さらに，「モデリング」を「問題項目への反応を決定する，認知能力や項目の特性を含む確率的関数を特定すること」として論を進める。

認知能力のモデリングに関して，IRT モデルでは 1970 年代にその萌芽を見ることができる。IRT モデルのうち，認知能力を考慮したモデルとしては Fischer（1973）の線形ロジスティックテストモデル（linear logistic test model, LLTM）が挙げられる。まず，LLTM を説明す

る前に，最も基本的な IRT モデルを示す。IRT モデルにおける 3 パラメタロジスティックモデルの項目反応関数は

$$\Pr\left(x_{ij}=1\middle|\theta_i,a_j,b_j,c_j\right)=c_j+\frac{1-c_j}{1+\exp(-a_j(\theta_i-b_j))} \tag{1.1}$$

と定義される。ここで，$x_{ij}=1$は解答者iの項目jへの正答反応，$x_{ij}=0$は誤答反応でありx_{ij}は 0 か 1 の 2 値しかとらないと仮定している。θ_iは解答者iの潜在特性パラメタ，a_j,b_j,c_jはそれぞれ項目jの識別力パラメタ，困難度パラメタ，当て推量パラメタである。c_jはθ_iが負の極限を取った時の正答確率であり，正答率の下極限である。誤答反応確率は正答反応確率を用いて，$\Pr\left(x_{ij}=0\middle|\theta_i,a_j,b_j,c_j\right)=1-\Pr\left(x_{ij}=1\middle|\theta_i,a_j,b_j,c_j\right)$と定義される。(1.1）式で$c_j=0$とすると，2 パラメタロジスティックモデル，$c_j=0,a_j=1$とすると 1 パラメタロジスティックモデル（Rasch model）となる。

LLTM は Rasch モデルにおける困難度母数をA個の認知操作子（cognitive operator）によって，

$$b_j=\sum_{a=1}^{A}q_{ja}\gamma_a+e \tag{1.2}$$

と分解する。ここで，q_{ja}は項目jにおける第$a\ (=1,\cdots,A<J)$番目の認知操作子の重みを表している。典型的には，項目jにおける第a番目の認知操作子が必要ならば 1，そうでなければ 0 を表すような 2 値をとる。さらに，γ_aは基本母数と呼ばれ認知要素が困難度にどれほど大きく影響するのかを表しており，eは項目間で共通の定数で基準化定数と呼ばれる。LLTM は困難度をその問題項目に必要な認知能力の線形和に分解するモデルである。LLTM は，認知能力の困難度に与える影響の大きさγ_aは項目によらず一定である。つまり，γ_aはその認知要素が項目に含まれることによる難易度の変化の大きさを示している。また，γ_aには解答者の添字iがつかず，認知操作子は解答者の特性としては扱われていない。

その後，多次元項目反応理論（multidimensional IRT, MIRT; e.g., Reckase, 2009）が登場する。MIRT モデルは潜在能力の正答確率の影響の仕方によって，補償モデルと非補償モデルに大別される。補償モデル（e.g., Reckase, 1985; Reckase & McKinley, 1991）の項目反応関数は

$$\Pr\left(x_{ij}=1\middle|\boldsymbol{\theta}_i,\boldsymbol{a}_j,b_j\right)=\frac{1}{1+\exp(-(\boldsymbol{a}_j^{\mathrm{T}}\boldsymbol{\theta}_i-b_j))} \tag{1.3}$$

である。ただし，$\boldsymbol{\theta}_i,\boldsymbol{a}_j$は$D$次元のパラメタベクトルであり，右肩のTは転置記号を表す。補

償モデルは, 適当なD次元の潜在特性の線形結合で表されるモデルである。このモデルでは, 能力パラメタが和の形で正答確率に影響しているため, 1 つの能力パラメタの値が低くとも, 他の能力パラメタの値が高ければ，正答確率が高くなりうることを表現するモデルである。また, (1.3)式において, ロジスティックモデルではなく, 正規累積モデルを用いる場合には, 補償 MIRT は 2 値のカテゴリカル因子分析モデルと等価なモデルであることが示されている（荘島, 2003）。さらに，多次元化することで，能力の間の相関構造も表すことが可能となる。

一方，非補償モデルの項目反応関数は，各次元を 2 パラメタロジスティックモデルとして能力次元で積を取った関数として,

$$\Pr\left(x_{ij}=1\middle|\boldsymbol{\theta}_i,\boldsymbol{a}_j,\boldsymbol{b}_j\right)=\prod_{d=1}^{D}\frac{1}{1+\exp(-a_{jd}(\theta_{id}-b_{jd}))} \tag{1.4}$$

と表される。$\boldsymbol{b}_j$は困難度パラメタをD個まとめたベクトルである。このため，非補償 MIRT は 1 つの次元でも能力パラメタの値が低い場合には，正答確率が極端に低下するモデルとなっている。

MIRT は複数の能力次元を扱えるように 1 次元 IRT モデルを拡張したモデルであった。また，LLTM は 1 次元の IRT モデルの困難度の分解を行ったモデルであった。この MIRT と LLTM を組み合わせたモデルが，Embretson（1984）の General multicomponent latent trait model（GLTM）である。GLTM では，ある項目jに対して，全体項目としてx_{ijT}という課題と，その項目の下位課題x_{ijd}がD個必要である。つまり 1 つの項目に対して$D+1$個の課題を用意することとなる。例えば,「Fist（げんこつ）: Clench（握りしめる） ＝Teeth（歯）: ____，という関係が与えられたとき，下線部に入る単語を，1. Pull（引く），2. Brush（磨く），3. Grit（食いしばる），4. Gnaw（噛み切る），5. Jaw（顎），から選ぶ」という言語アナロジー課題を考える（Embretoson , 1984 の Table 1）。この課題が全体項目である。さらに，下位課題として，ルール構成課題と，ルール評価課題を与える。ルール構成課題は,「Fist : Clench = Teeth : ____」という関係のルールを記述する課題である。ルール評価課題は「ルール : teeth に関係する怒りの反応」，のうち,「Fist : Clench = Teeth : ____」を満足する単語を，1. Pull，2. Brush，3. Grit，4. Gnaw，5. Jaw，から選ぶという課題である。このように，全体項目とそれに解答するための下位課題をあわせて一つの項目として扱う。これらの課題に対して，x_{ijT}の反応確率は

$$\Pr\left(x_{ijT}=1\middle|\boldsymbol{\theta}_i,\boldsymbol{\gamma}_d,\boldsymbol{e},c,g\right)=(c-g)\prod_{d=1}^{D}\frac{1}{1+\exp\left(-\left(\theta_{id}-\left(\sum_{a=1}^{A}q_{jad}\gamma_{ad}+e_d\right)\right)\right)}+g \qquad (1.5)$$

と表現される。$\boldsymbol{\theta}_i$は補償 MIRT と同様にD次元の能力パラメタベクトルであり，$\boldsymbol{\gamma}_d,\boldsymbol{e}$は$D$次元の基本母数ベクトルと基準化定数ベクトルである。パラメタcは，下位課題に全て正答できた場合の全体項目の正答率の上限であり，パラメタgは 1 つでも下位課題に誤答した場合の全体項目の正答確率である。この全体項目と下位課題の同時分布$\Pr(x_{ijT},x_{ij})$を用いて，尤度関数を定義しパラメタの推定を行う。例えば，パラメタcおよびgの最尤推定量は項目反応パタンから容易に計算できる事が知られており，それぞれ，

$$\hat{c}=\frac{\sum_{i=1}^{I}\sum_{j=1}^{J}\left(\prod_{d=1}^{D}x_{ijd}\right)x_{ijT}}{\sum_{i=1}^{I}\sum_{j=1}^{J}\prod_{d=1}^{D}x_{ijd}}, \qquad (1.6)$$

$$\hat{g}=\frac{\sum_{i=1}^{I}\sum_{j=1}^{J}\left(1-\prod_{d=1}^{D}x_{ijd}\right)x_{ijT}}{\sum_{i=1}^{I}\sum_{j=1}^{J}\left(1-\prod_{d=1}^{D}x_{ijd}\right)} \qquad (1.7)$$

と与えられる。$\hat{c}$について，$\prod_{d=1}^{D}x_{ijd}$は下位課題に全て正答していれば 1，1 つでも誤答していれば 0 であり，$\left(\prod_{d=1}^{D}x_{ijd}\right)x_{ijT}$は下位課題に正答しかつ全体項目に正答しているかどうかをそれぞれの解答者と項目について算出している。ここから，$\hat{c}$は下位課題全てと全体項目に正答している解答者・項目の総数を下位課題全てに正答できた解答者・項目の総和で除したものである。言い換えると，$\hat{c}$は下位課題に全て正答できた場合にどれくらい全体項目に正答できるかを表している。$\hat{g}$は$\hat{c}$とは逆に，下位課題に 1 つでも正答できなかった場合（$1-\prod_{d=1}^{D}x_{ijd}$）の総数に対して，下位課題に 1 つ以上誤答した場合に全体正答が正解できる割合を表している。

(1.5)式は，D個の認知プロセスそれぞれに LLTM に相当する項目反応関数を仮定し，それらが非補償関係にあると仮定している。なぜならば，GLTM は人間の情報処理プロセスのモデリングを行ったものであり，1 つでも情報処理が上手く行かないと，それらを全て利用する問題項目には正答できないと考えているからである。さらに，(1.5)式の各次元の項目反応関数$\Pr\left(x_{ijd}=1\middle|\theta_{id},\boldsymbol{\gamma}_d,e_d\right)=\left(1+\exp\left(-\left(\theta_{id}-\left(\sum_{a=1}^{A}q_{jad}\gamma_{ad}+e_d\right)\right)\right)\right)^{-1}$に注目すると，LLTM を応用した項目反応関数になっている。q_{jad}は項目jの各次元dでの認知操作子aの重み，およびγ_{ad}は次元dでのa番目の認知操作子の影響の大きさを表しており，e_dは次元dでの基準化定数である。ここから GLTM は LLTM の次元を拡張した一般化であり，非補償 MIRT モデルの困難度を分解した一般化にもなっている。GLTM において，LLTM の下位課題ごとの困難度の構成要素を仮定しない場合には，単に$A=1$として，(1.5)式を

$$\Pr\left(x_{ij}=1\middle|\boldsymbol{\theta}_i,\boldsymbol{a}_j,\boldsymbol{b}_j,c,g\right)=(c-g)\prod_{d=1}^{D}\frac{1}{1+\exp\left(-\left(\theta_{id}-b_{jd}\right)\right)}+g \tag{1.8}$$

と書き直すことができる。(1.8)式のモデルを Multicomponent latent trait model（MLTM; Whitely, 1980）という。また，MTLM を CDM としさらなる拡張を図ったモデルとして MLTM-D（Multicomponent latent trait model for diagnosis; Embretson & Yang, 2013; Embretson, 2015）も提案されている。

ここまで見てきたように，IRT に基礎をおく一連のモデルは項目反応関数に複数の認知能力を仮定するように発展してきた。具体的には，能力の次元を一次元から多次元へ拡張したモデルや，項目の困難度の源泉としての基礎的認知能力の仮定をするモデルが開発されてきた。このように，認知能力を含むモデルの開発はこれまで多くの研究関心を集めてきたと考えられる。

1-3 認知診断モデルの萌芽的研究

次に CDM の萌芽的研究として重要なのがルールスペース法（Rule space method, RSM; Tatsuoka, 1983; Tatsuoka, 2009; 龍岡・林, 2001; 龍岡・倉元, 2006）である。RSM は符号付き引き算の分析や分数の計算問題の際に誤答してしまう理由の分析（バグ分析）に起源をもつ。幾つかの正しい・誤った一連の計算手続きをルールと呼ぶ。そして，問題解答時の誤った解答ルールの適用を分析するために RSM が開発された。RSM では，ルールスペースと呼ばれる空間での解答パタンを分類する。つまり，潜在能力次元に直交する次元を想定し，それぞれの解答者の項目反応パタンをそれらの二次元空間（ルールスペース）に付置する。その上で，それぞれのルールを適用した場合の項目反応パタンもルールスペース上に付置し，各解答者の解答パタンと各ルールを適用した場合の解答パタン類似度をマハラノビス距離といった距離関数により評価する。これにより，それぞれの解答者がどのようなルールによって解答を行ったのかを推測する。このように RSM は IRT モデルを拡張した診断のためのモデルということができる。

RSM で重要なのがテストによって測定されている能力と項目がどのような関係にあるのかを示した Q 行列（照応行列ともいう; 龍岡・倉元，2006）を明示的にあつかったことである。CDM の文脈では，テストで測定されている能力をアトリビュートと呼ぶ。この Q 行列の導入により，どのようなアトリビュートを習得していると，正答確率がどのように変化するのか定義することができるようなった。現在，CDM と名のつくモデルのほとんどで Q

行列が用いられており，Q 行列は CDM を特徴づける構成要素である。

さらに，アトリビュート階層法（attribute hierarchy method, AHM; Leighton, Gierl, & Hunka, 2004）は RSM を発展させた手法である。具体的には，AHM はアトリビュート間の関係性を積極的にモデルに組み込み，アトリビュートの階層構造を仮定するモデルである。つまり，AHM ではあるアトリビュートの習得には別のアトリビュートが必要であることや，複数のアトリビュートが後続のアトリビュート習得の条件になっていることなど，学習で必要となる習得の順序関係をモデルに取り入れることを目指している。これらの RSM と AHM については第 2 章において再度触れる。

テスト理論においては，解答者をテスト項目に依存しない能力を使って順序づけすることを可能にした IRT モデルの意義は今後もありつづける一方で，解答者によりフォーカスしテストから解答者にとって有用な情報を引き出すためのモデリングへの関心も高まって来ているといえよう。

1-4 テスト理論以外の認知能力のモデリングの潮流

一方で，IRT モデルでの認知能力のモデリングとは別の潮流として，ベイズ統計学を用いたベイジアン認知モデリング（e.g., Lee & Wagenmakers, 2014; Okada & Lee, 2016）の発展が見られる。ベイジアン認知モデリングはテストに限らない認知能力一般に関するモデリングであり，様々な確率分布の組み合わせによってモデルを記述するものである。こうしたモデリングが実行できる背景としては，近年の計算機の性能の伴いマルコフ連鎖モンテカルロ法（Markov chain Monte Carlo method, MCMC method; Geman & Geman, 1984）を実行するための汎用的なフリーソフトの開発がなされてきたことが挙げられる。例えば，BUGS（Bayesian inference using Gibbs sampling）プロジェクトが 1989 年に始まり，WinBUGS（Lunn, Thomas, Best, & Spiegelhalter, 2000; Ntzoufras, 2009）や OpenBUGS（Spiegelhalter, Thomas, Best, & Lunn, 2010）といったソフトが開発されてきている。さらに，JAGS（just another Gibbs sampler; Plummer, 2003）や，近年注目されている No-U-Turn サンプラー（Hoffman & Gelman, 2014）などのハミルトニアンモンテカルロを実行できる Stan（Carpenter, Gelman, Hoffman, Lee, Goodrich, Betancourt, ... & Riddell, 2016）といったフリーソフトの開発も盛んである。このようなソフトウェアによって MCMC によるパラメタ推定がモデル式の記述や若干のプログラミングのみで容易に可能となった。IRT モデルにおいても，Bolt & Lall（2003）や Patz & Junker（1999）により，MCMC によるパラメタの推定法が提案されている。さらに，従来

の線形回帰モデルや共分散構造分析の枠組みにとらわれない柔軟な認知能力のモデリングも可能となっている（e.g., Lee & Wagenmakers, 2014）。こうしたベイジアン認知モデルリングの考え方による実験データ解析や質問紙解答のプロセスの解析なども盛んに行われてきている。

このように，テスト理論以外の文脈においても人間一般の行動の認知プロセスをモデリングすることは心理学一般の興味として存在してきた。また，近年の計算機の発展に伴って，確率分布の複雑な組み合わせによってデータを分析することも可能になってきている。こうした認知能力のモデリングはテスト以外のデータを用いて今後も積極的に展開されると考えられる。次節では，CDM の基本的な考え方について説明し，CDM を適用するデータや最終的な出力の例を示す。

1-5 認知診断モデルの基本的発想

前述のように，IRT がテスト解答者を評価し，その能力特性値を推定することを主要な目的として用いられるのに対し，CDM はテスト解答者の領域ごとの認知能力を診断することを主要な目的として用いられる。こういった目的のテストをとくに認知診断テストなどとも呼ぶ（e.g., Alderson, 2005; Nichols, Chipman, & Brennan, 1995）。CDM ではテストの問題を解くために複数の認知能力が必要であると考え，各認知能力を習得しているかどうかによって問題に正答できる確率が変化すると考える。このような設定において，それぞれのテスト解答者について，認知能力習得の有無のパタンを解答データから推定することが CDM の基本的な枠組みである。

Rupp & Templin（2008b, pp.226）は，より形式的な表現を用いて CDM を以下のように定義している。「CDM は単純または複雑な負荷量構造を持つ確率的，確証的な多次元潜在変数モデルである。CDM はカテゴリカルな観測変数のモデリングに適しており，潜在的でカテゴリカルな説明変数を持つ。説明変数は補償的・非補償的に統合されて潜在クラスを生成する」。このように CDM は，通常「正答・誤答」に対応する従属変数のみならず，「ある能力の習得・非習得」のように説明変数にもカテゴリカル変数を仮定するモデルであって，LCA モデルの特殊な場合とみることができる。すなわち，カテゴリカルで潜在的な説明変数から潜在クラスが定義され，この潜在クラスごとに観測変数への反応確率が異なると考えるのである。

先にあげた，潜在的でカテゴリカルな説明変数はアトリビュート（attribute）と呼ばれる。

アトリビュートは問題項目に正答するために求められる領域ごとの認知能力やスキルの要素に対応する。このアトリビュートに関連した重要な要素として，Q 行列（Tatsuoka, 1983）とアトリビュート習得パタン行列がある。因子分析との対応でいえば，アトリビュート，Q 行列, アトリビュート習得パタンは, それぞれ因子, 因子負荷量行列, 因子得点に相当する。

Q 行列は，それぞれの問題項目に正答するためにどのアトリビュートが求められるのかを表す，{0,1}の 2 値変数を要素にもつ行列である。表 1.1 に分数の計算のテストを想定した，Q 行列の例を示した。この例では，分数の計算の問題を例にしており，アトリビュートとしては，分数の「加減」,「乗除」,「通分」という 3 つを仮定している。こうしたアトリビュートとしては熟達の程度など連続量で表されるものではなく，習得・未習得で表現できる質的なものが扱われることが多い。項目 1 の$1/3 + 4/3$という計算には，単純な分数の足し算・引き算のみが必要であることを示している。さらに項目 2 の$1/2 \times 1/3$には，分数の掛け算・割り算の能力のみが必要であるとしている。最後の項目 3 は$1/2 + 1/3$という項目であるが，この問題に正答するためには各項を通分する必要があり，通分した後に加減の演算が必要となるため，「加減」と「通分」の 2 つのアトリビュートが必要であるということを仮定している。このように Q 行列ではある項目に正答するために，どのアトリビュートが必要であるのかを示しており，ある項目の正答にあるアトリビュートが必要であれば 1 を，そうでなければ 0 を要素として持つものである。これらのアトリビュートや Q 行列の設定はテストの設計者やテスト内容の専門家によって行われる。より具体的には次章で示すが，文献や測定する領域の理論的背景をもとにアトリビュートを設定し，それに合わせた問題を開発したり，既存の問題項目を分析して Q 行列を設定する。

次に，どういったアトリビュートを習得している場合に，各問題に正答できるのかを考える。表 1.2 に，この例における，可能な全てのアトリビュート習得パタンを示した。0 は当該アトリビュートを未習得であること，1 は習得していることを表す。ここではアトリビュート数が 3 なので，$2^3 = 8$個のアトリビュート習得パタンが存在する。例えば，習得パタン 1 は「加減」,「乗除」,「通分」のどのアトリビュートも習得していないパタンであり，習得パタン 8 は逆に 3 つ全てのアトリビュートを習得しているパタンであることが読み取れる。習得パタン 2~4 は各アトリビュートを 1 つだけ習得しているパタンで，習得パタン 5~8 は 2 つのアトリビュートを習得しているパタンである。これらの習得パタンから，例えば項目「1/3＋4/3」に対しては,「加減」のアトリビュートを習得している習得パタン（2，5，6，8）の解答者は正答する可能性が高く,「加減」アトリビュートを習得していないパタン（1,

3，4，7）の解答者は正答する可能性が低いと考えられる。もう一つ例を挙げると，項目「1/2 + 1/3」に対しては，「加減」と「通分」アトリビュートの 2 つを習得している習得パタン 6, 8 の解答者の正答確率は高く，それ以外の習得パタン（1, 2, 3, 4, 5, 7）の解答者が正答する可能性は低いと考えられる。アトリビュート習得パタン行列は，こうした習得パタンをすべての解答者についてまとめたものである。

このように，CDM ではアトリビュート習得パタンを生成する規則と，それらの規則によって表されるクラスによって項目反応がどのように異なるのかを表すモデルである。つまり，CDM の特徴は，カテゴリカルな潜在説明変数（アトリビュート）と Q 行列を用いて項目反応確率を規定し，アトリビュート習得パタンごとに項目への反応確率が異なると考えることである。このように，CDM はテストへの解答データを用いて，解答者のアトリビュート習得パタンを推定し，その習得パタンを学習に活用するためのモデルである。また，CDM は潜在変数が離散的で確証的なモデルであり，統計学的モデルとしては制約付き潜在クラスモデルとみなされる。ただし，CDM の研究論文ではこうした定義に当てはまらないモデルに対しても「認知診断モデル」という命名を行っている場合も散見される。また，項目反応モデルや因子分析モデルもその使い方によってテストからの診断情報を抽出することができる。こうしたことから，厳密に CDM の定義を述べることは難しいものの，CDM の名を関する多くのモデルで Q 行列が用いられている点は CDM の特徴と考えられる。なお，ここで例示したのは，CDM の中でもとくに基本的なモデルである DINA モデル（deterministic inputs noisy “and” gate model; Haertel, 1989; Junker & Sijtsma, 2001; Maris, 1999）に則った CDM の発想であった。

1-6 認知診断モデルの適用手続き

CDM は広く開発が進んでおり，応用研究についても一定の蓄積が見られる。本節では，CDM に適したテスト作成の手続きを示す。これにより，具体的なテスト実施上の問題などを考察する手がかりを得ることを目指す。具体的な手続きについては DiBello, Roussos, & Stout（2006）に詳しく解説がなされている。また，Rupp（2007）は RSM に関した解説を行なっている。Jang（2009）は Q 行列の構成の方法に詳しい。Lee & Sawaki（2009）においても英語テストでの認知診断モデルの適用方法が示されている。図 1.1 に Lee & Sawaki（2009）で示されている既存のテストに対する CDM の適用手続きを示した。図 1.1 では分析に際して，アトリビュートの定義を行う際にもデータ解析などを通じて再度アトリビュートを修

正する手続きが含まれていることを示している。以下の手続きは基本的に DiBelo et al.（2006）に依拠する。

認知診断モデルを適用する対象のテストである認知診断テストの実施では以下の 6 つの段階を踏むことが一般的である。その 6 つの段階は，1.検査目的の記述，2.診断的興味のある潜在的アトリビュート（アトリビュート空間）の記述，3.テスト項目の分析と開発，4.利用するモデルの特定，5.統計モデルの選択と結果の評価，6.解答者・教師・その他のテスト結果を利用する人に対する検査結果の報告，である。以下，それぞれの観点について詳細を示す。

1-6-1　検査目的の記述

はじめに，検査の目的は明確に定義される必要がある。この目的が潜在的アトリビュート空間を定義する際に意味を持つ。例えば，検査の目的が解答者の選抜や順位付けであれば，テストに下位能力を想定する必要は必ずしもなく，一次元的な能力として評価を行えばよい。ただし，これはテストの一次元性が担保されている状況においてである。対象とする潜在変数やアトリビュート空間が一次元以上の変数によってモデリングされるかどうか，分類目的に関して離散的か，スケーリング目的に関して連続的かどうかに関してといった点が，具体的なモデルを決める上で重要となる。この上で，2 つの実践的なスキル診断状況が考えられる。第 1 に，すでにある検査データからより豊富な情報を引き出すポストホック分析（あるいは，retrofitting とよばれる）場合がある（e.g., Jang, 2009; Lee & Sawaki, 2009; Li, 2011）。このアプローチは，適用するテストが多次元でないことや，背後に認知理論が仮定できないことなどから疑問視されることもある（Gierl, Cui, & Zhou, 2009）。第 2 のアプローチでは，アトリビュートによる診断的目的のために，はじめからデザインする（e.g., 孫・島田・谷部, 2015; Tjoe & de la Torre, 2013, 2014）。特に CDM の適用研究としては前者の既存のテストデータの再分析的研究が多く見られる。これらの応用研究については次節で述べる。次に診断的に興味のある潜在的スキルの記述を行う。

1-6-2　アトリビュート空間の記述

次に，テストの目的を果たすために測定されるスキルや他のアトリビュートの詳細な定式化が必要となる。アトリビュート空間の詳細な表現は，認知科学や教育心理学の知見を基盤にして，テストの目的に沿ってなされる。認知理論はよいテストのデザインやその実施に

関して重要であり，アトリビュート空間の設定としてどの定式化が認知的・計量心理学的に正しいかを第一に考える必要がある。アトリビュート空間を記述することは，テストのスコアを診断的に利用するためのテストの内容的エビデンス（e.g., Messick, 1995; 村山, 2012）とみなすことも可能であろう。

ここで，アトリビュート数は過度に多すぎてはならず，統計的な取り扱いやすさの問題から制御される必要がある。アトリビュート数が増えることによって，推定すべきパラメタの数が極端に多くなってしまうことがある。必要なパラメタはモデルによって異なっているが必要なモデルの倹約性の観点からも多すぎるパラメタ数は問題を生じさせる可能性がある。

1-6-3　検査課題の分析と開発

テスト問題の開発段階では，どの程度どの種類のアトリビュートが含まれていて，どの水準の難易度か，どんな形式の交互作用（例えば，アトリビュートの非補償・補償関係）があるかを理解するための課題分析を含むことが望ましい。また，アトリビュートは統計的に不適切ほど多くすべきではないが，診断的な有用性や解釈可能性を損なうほど減らしすぎるのも得策ではない。はじめに提案されたアトリビュートセットと統計的によいアトリビュート間を合致させることが有効な診断テストの作成に不可欠である。

アトリビュートの交互作用のモデルの選択は検査の目的や診断の状況に依存している。この他，考慮するべきアトリビュートの特性は，スキルの難易度の順序性，アトリビュートのサイズ（粒度 ; grainsize）やアトリビュート間の階層性を考慮した Q 行列の構成が必要である（Rupp, Templin, & Henson, 2010）。

さらに，課題分析過程で，特定の課題-スキルコーディングルールを開発する。これを用いて，独立なコーダーがアトリビュートを問題項目に割り当て，Q 行列を作成する。このことから，先に示したように Q 行列はアトリビュートと問題項目の仮定されたリンクとみなされる。はじめに作成された Q 行列は次の統計的な分析によって修正される。

データ分析によって，仮定された Q 行列の内的一貫性のチェックや，課題の困難度，アトリビュートの整合性を検討する。スキルの習得・未習得の 2 値の場合，推定されたアトリビュートの習得の母比率をそのアトリビュートの困難度と呼ぶ。

具体的な研究では，例えば，Tjoe & de la Torre（2013, 2014）は比例推論能力を診断するテスト作成の例を示しており，文献研究によってアトリビュートの選定を行い，さらに中学

校の数学教員との協議を繰り返し，アトリビュートを洗練させている。Jang（2009）は，テストを受けた直後の学生に即時回顧させてどのように問題を解答したのかインタビューを行ったり，複数の言語学を専門とする研究者間でどのようなアトリビュートが妥当であるか協議し，専門家が独立に作成した Q 行列の評定の一致率を算出するなど，Q 行列を作成するための工夫を行なっている。

Lee & Sawaki（2009）の言語検査での CDM の適用では，まず第 1 にテストで測定するアトリビュートを定義し，次に Q 行列を設定した。また，経験的なデータ，当該領域の理論的な側面や，問題解答中の発話思考を元にしたプロトコルを用いて妥当な Q 行列の構成を目指している。

以上のように一般的にはテストでの測定が望まれるアトリビュートを設定し，そのアトリビュートがテスト項目それぞれで必要かどうかを検討していく。これは測定が行われる領域での理論的な合理性や妥当性を検討しつつ，専門家の手によって進められる。

前述のように，一度設定した Q 行列はテストを実施しデータ分析を行なった後に修正を行う。この修正においても前述したような専門家の手によるものと統計的な推定を用いた方法がある。例えば，日本語語彙理解力テストにおいてルールスペース法を適用した，倉元・スコット・笠居（2003）では一度分析を行った後に国語教師（領域の専門家）によって再度検討する方法が取られている。また，Lee & Sawaki（2009）では実データ分析と Q 行列の修正，またはアトリビュートの再定義を行き来しながら Q 行列を決定するプロセスを紹介している。その一方，正しいアトリビュートサイズを決定することは難しいという指摘も同時行っている。

1-6-4　　計量心理学的モデルの特定

検査目的やアトリビュートの仮定から，分析モデルを決めるが，実データへの適合もモデルを選択する材料として挙げられる。また，具体的なモデルに関しては理論的な背景が必要であることはこれまでにも述べたが，第 2 章で示すように CDM とみなされるモデルは非常に数も多く，また，必ずしも統一的なフレームワークに含まれるモデルばかりではないため，モデル選択は容易ではない。認知的な理論との整合性もモデルの選択には非常に重要であるが，それに加えて，既存のテストや得られたデータがどのような認知プロセスやデータ生成プロセスに従って得られたものなのか詳細に検討する必要があるだろう。

1-6-5　　モデル修正と評価

選択されたモデルを用いて項目パラメタとアトリビュート習得パタンを推定するが，合わせてモデルの修正も行う。具体的には，Q 行列の要素を改めて検討したり，不必要なアトリビュートを削除するといった変更が必要である。このとき，診断目的によって課される制約を満足する一方でモデルをできるだけ簡潔に保たなければならない。すべてのモデルにおいて，項目パラメタはどれくらい項目が診断目的に関してうまく機能するかを示している。

さらに，通常のテストと同様に，テストの信頼性を確認したり，妥当性の証拠についての言及も行う。まず，信頼性の推定は二度のテストでのアトリビュート習得の一貫性によってなされる。Templin & Bradshaw（2013）は古典的テスト理論における信頼性係数の概念を CDM 用に拡張し，2 回のテストで，あるアトリビュートの習得・未習得について，クロス集計表を作り，テトラコリック相関係数をアトリビュートkについての信頼性係数の推定値とすることを提案した。この信頼性は IRT モデルにおけるシミュレーションにもとづくリサンプリング法による信頼性（詳細は，Templin & Bradshaw（2013）の pp.261 -.262）より一貫して高いということが報告されている。一方で，Gierl, Ciu, & Zhou（2009）も AHM にもとづいた信頼性の定義を行なっており，CDM の中でも信頼性の定義は必ずしも一貫してない。

妥当性の証拠について，反応パタンに基づく証拠や外的変数との関連に基づく証拠も重要となる。反応パタンに基づく証拠としては，例えば，異なった習得パタンに分類された解答者間の行動に差異があるかを調べることが挙げられる。外的変数に基づく証拠では，通常の心理学研究と同様に，収束・弁別的証拠を収集することが望まれる。また，統制群と学習法の教示といった処置群を含む妥当性のエビデンスを集めることは有益である。このほか，テスト項目やテスト解答者が適切であるかどうかを判断する指標も，Fusion Model に限定的ではあるものの提案されている（Hartz & Roussos, 2008）。また，解答者の真の認知状態がモデルによって表現されているかどうかを尤度比検定によって検討する方法も提案されている（Liu, Douglas, & Henson, 2009）。

認知診断テストはテストの実施上，テストの測定している概念の定義や，アトリビュートと項目との関係の詳細な検討を行なっており，その意味で CDM の適用の中に妥当性検証のプロセスが自然に組み込まれていると考えられる。この意味で，最終的なモデル選択で，一次元 IRT 的モデルが選択されたとしても，アトリビュートの検討や Q 行列の作成は妥当

性を高める意味で意義があろう。Q 行列の作成は，Embretson（1998）の言うところの認知システム・デザインアプローチと呼ばれる認知モデルに依拠したテスト項目作成のアプローチに繋がると言えよう（e.g., Gierl & Haladyna, 2012）。また，想定される認知要素間の関係を共分散構造分析によって検討する方法も提案されている（Dimitrov & Raykov, 2003）。

1-6-6　スコアの報告

スコアの報告では，教師や生徒といったテストの結果を利用する人にとって，読みやすく，解釈しやすいものである必要がある。例えば，それぞれの能力をx軸に，習得確率をy軸に持つような棒グラフを用いてフィードバックを行なうことが考えられる。また，スコア報告は分析の結果に関して学習するべきことの教示など，学習行動に関するアドバイスを含む場合もある（e.g., Roberts & Gierl, 2010）。

1-7 認知診断モデルの適用研究の具体例

本節では，CDM の適用例を具体的にみることで CDM の応用研究の実態を示す。CDM の適用事例のうち，応用事例の多い数学テスト（鈴木・豊田・山口・孫, 2015），英語テスト（Li, Hunter, & Lei, 2016）をより具体的に紹介する。これらの代表的な応用例を検討するとともに，近年注目されているコンピュータ適応型テストと組み合わせた実践例（Liu, You, Wang, Ding, & Chang, 2013）も紹介する。

鈴木他（2015）は，小・中学校で利用される標準学力検査の 1 つである数研式標準学力検査 NRT「中学 1 年生数学」（以下，数研式 NRT）の一部を認知診断に用いた。アトリビュートは，数学的問題解決の認知プロセスに関する知見に基づき，また問題項目に鑑みて「計算力」,「概念理解」,「図形操作力」,「論理力」の 4 つを設定した。各アトリビュートの定義や項目例は鈴木他（2015）の表 5 に示されており，ここではそれらを簡単に紹介する。「計算力」は方程式を解くといった計算を遂行するアトリビュートである。「概念理解」は数学の用語や公式の理解である。「図形操作力」は図形を心的に操作し問題を解くアトリビュートである。そして「論理力」はすじ道を立てて問題の解法を探索するアトリビュートである。

これらのアトリビュートについて，著者 2 名の協議により 49 の問題項目に対する Q 行列を作成した。Q 行列の妥当性を検討するために評定者間信頼性を確認した。具体的には，第 1 著者ではない著者 1 名がアトリビュートの定義および問題例と検定教科書を用いて，はじめに Q 行列を作成した著者とは独立に Q 行列を作成し，アトリビュートごとにはじめ

に作成されたQ行列との一致率が算出された。解答者数は948名であった。CDMの具体的なモデルとしては，G-DINA モデル（generalized DINA model; de la Torre, 2011），A-CDM（additive CDM; de la Torre, 2011）, DINA モデル，および DINO モデル（deterministic imputs noisy “or” gate model; Templin & Henson, 2006）が用いられた。これらの具体的なモデルについては，第2章で詳細に検討を行う。

A-CDM によって推定されたアトリビュート習得パタンの要約（鈴木他（2015）の表 11の一部）を表 1.3 に示した。表 1.3 から，どのアトリビュートも習得していない人数が最も多く，ついで全てのアトリビュートを習得しているパタンの人数が多く，三番目に「計算力」「図形操作力」を習得し，「概念理解」「論理力」が未習得であるパタンの人数が続いていた。この他にも複数のアトリビュート習得パタンが確認された。これにより，素点からはわからない解答者の現状が明らかになった。具体的には，テストの得点が高くとも，実は「図形操作力」が習得できていない解答者や，逆に得点が低くとも「図形操作力」と「論理力」を習得できている解答者が存在することが浮き彫りになった。テスト得点は高いが「図形操作力」を習得できていない解答者には，この能力が今後のつまずきにならないような対応が必要と考えられる。一方，「図形操作力」と「論理力」を習得できている解答者には，計算スキルの訓練や，数学用語を普段から理解しようと試みることで，学習改善につながる可能性が考えられる。このように，CDMを用いることで，単純なテスト得点だけからはわからない解答者の数学能力の習得状況が明らかになり，効果的な学習を促すための介入の指針を得ることができた。

Li et al.（2016）はミシガン英語調査バッテリー（Michigan English Language Assessment Battery, MELAB）の読解力テストにCDMを適用した。MELABは，アメリカの大学における学問・研究にあたっての，ネイティブでない大人の話者の発展的なレベルの英語能力を評価するテストである。アトリビュートは「語彙（Vocabulary）」，「構文（Syntax）」，「明示的情報の抽出（Extracting explicit information）」，「暗黙の情報の理解（Understanding implicit information）」4つであり，Q行列は関連する文献の調査やテスト解答者の発話思考プロトコル，専門家の評定をもとに過去に作成されたものを用いた。解答者は2019人であった。G-DINA モデル，A-CDM, R-RUM（reduced reparametrized unified model; e.g., Hartz & Roussos, 2008）, DINA モデル，DINO モデルを比較の対象とした。

結果として，赤池情報量規準（Akaike information criterion, AIC; Akaike, 1974）の観点からはG-DINA モデルが，ベイズ情報量規準（Bayesian information criterion, BIC; Schwarz, 1978）

の観点からは A-CDM が最も支持された。また，表 1.4 はモデルごとのアトリビュート習得パタンの割合を示したものである。パタンの列の 4 つの数字は，左から，「語彙」，「構文」，「明示的情報の抽出」，「暗黙の情報の理解」の習得（1）・未習得（0）を表す。表 1.4 から，利用するモデルによって，結果として得られるアトリビュート習得パタンの割合にも相違があることがわかる。とくに DINO モデルは 1 つもアトリビュートを習得していないパタン（0000）が他のモデルよりも少なく，語彙のみを習得しているパタン（0100）や，全てのアトリビュートを習得しているパタン（1111）がみられ，他のモデルと挙動が異なっていた。ただし，全体的な傾向をみると，どのモデルにおいても 1 つもアトリビュートを習得していないパタンと全てのアトリビュートを習得しているパタンの割合が多い傾向が見られた。

MELAB リーディングテストでは相対的に A-CDM の適合がよかったことから，読解に必要なアトリビュートはそれぞれが個別的・直接的に正答率に寄与することが考えられる。つまり，Li et al.（2016）で用いられた 4 つのアトリビュートをそれぞれ個々に学習することで，読解能力を向上させられる可能性が示された。

現代では CDM とコンピュータ適応型テスト（computerized adaptive testing, CAT; van der Linden & Glas, 2000）を組み合わせて運用すれば，それぞれの解答者にとって，どの分野にもっと力を入れて学習を進めるのがよいかといった診断情報をすぐにフィードバックし，今後の学習に生かしてもらうようなテストの運用が可能になっている。こうしたアプローチは認知診断的コンピュータ適応型テスト（cognitive diagnostic computer adaptive testing, CD-CAT; Cheng, 2009）と呼ばれる。

Liu et al.（2013）は中国における英語の到達度試験において，CD-CAT を用いて大規模な運用を行った研究を報告している。この試験では，8 つのアトリビュートからなる 400 の問題項目が，11 人の専門家によって作成された。その後，6 人の英語教師がこの Q 行列によるアトリビュートの定義を評価し，不一致が見られた箇所は議論をした上で，やはり意見の一致をみなかった問題項目は削除された。CDM としては DINA モデルが用いられ，これをシャノンのエントロピーを用いた CAT と組み合わせることによって適応型の問題項目の出題が実装された。この CD-CAT システムを用いて，北京の 8 つの学校の 5，6 年生 584 名による実践研究が行われた。CD-CAT システムによる結果と，学校における生徒の英語の達成度や教員による評定との整合性が評価され，システムの妥当性が確認された。

このシステムはその後実際に運用されるようになり，2011 年には大連において，2,000 台の PC を用いて約 30,000 人の生徒がこの CD-CAT システムを利用して，学校現場への導入

に向けた成功事例となった。こうしたCD-CATの取り組みは近年多くなされており（Liu et al., 2013; Luo, Ding,Wang, & Xiong, 2016），CDMの今後ますます発展するであろう応用分野といえる。

1-8 認知診断モデル比較の現状

第2章で詳細を検討するが，CDMでは非常に多くモデルが開発されている。しかしながら，利用できるモデルの選択肢が多くなることにより，モデルの利用者にはどのようなモデルを利用するべきなのか混乱が生じる。CDMを活用していく上では，開発されてきたさまざまなモデルを実データに適用し，各種モデルの当てはまりのよさを比較したり，解釈可能性を吟味したりといった経験的な蓄積が必要といえる。

このようなモデル間の比較は，モデル開発を行う場合と，CDMを利用する場合の両側面に対して寄与がありうる。まず，第1のモデル開発に対する寄与として，理論的に開発されたモデルのテストでの振る舞いを調査することによって，現実のテストでどういったモデルがよく機能するのかが明らかになり，その知見からより実態に即したモデル構築への示唆が得られる。第2に，CDMを利用する場合にはデータに適したモデルを使うことが望ましいが，しかし，いたずらにパラメタ数の多い複雑なモデルを採用するとその解釈や汎化性能に問題が生じうる。そのため，AICやBICといった情報量規準を用いたモデル比較を行うことによって，データへのモデルの過適合を防ぎ，将来のデータの予測や，確率構造の適切さといった観点からモデルを評価することが可能になる。また，一般的なモデル比較の手続きとして，ベイズ統計学を用いた方法として事後予測分布や事後予測p値，あるいはベイズファクターを利用した方法なども利用することができる。

そうしたCDMのモデル比較研究としては，CDMに含まれるモデル間での比較，および，CDMに含まれる1つまたは少数のモデルとその他の関連するモデルとの比較が，部分的に行われてきた。

まず，前者のCDM間でモデル比較を行った先行研究について述べる。Li et al.（2016）は，2019名のテスト解答者を対象としたMELABリーディングテスト（4枝選択，20問）のデータについて，CDMに含まれるモデル間での比較を行った。具体的には，G-DINAモデルとその下位モデルに位置付けられるDINAモデル，DINOモデル，A-CDM，R-RUMを比較検討し，AICの観点からはG-DINAモデル，BICの観点からはA-CDMが優れていることを示した。また，鈴木他（2015）は中学生1年生の数学テスト（数研式NRT）のデータにつ

いて G-DINA, A-CDM, DINA モデル，DINO モデルを比較し，AIC の観点からは G-DINA モデル，BIC および CAIC の観点から A-CDM を採用した。

次に，後者の 1 つまたは少数の CDM と他のモデルとの比較を行った先行研究について述べる。Kunina-Habenicht, Rupp, & Wilhelm（2012）は小学生向けの算数の診断テストを自作し，CDM の 1 つである GDM（general diagnostic model; von Davier, 2008）と探索的・確認的カテゴリカル因子分析モデルとを比較した結果，確認的因子分析モデルの適合がよいことを報告した。また，Lee & Sawaki（2009）は TOEFL iBT データを用いて，GDM と R-RUM という 2 つの CDM と，LCA の比較を行ったものの，統計的なモデル適合の指標は用いず，スキル習得パタンの類似性を考察した。さらに，de la Torre & Karelitz（2009）は，あるアトリビュートを習得した後に次のアトリビュートを習得するという，直線的な構造を仮定した CDM と IRT モデルでのテスト解答者の知識状態の分類の精度を，シミュレーションを用いて比較した。結果として，データ生成モデルが IRT で分析モデルが CDM の場合やその逆の場合に，分類精度が極端に低下することが示された。

1-9 論文の目的・構成

これまで見たように，応用場面ではどの CDM を分析に用いるのか明確な方針は必ずしもあるわけではない。また，CDM の適用に際して，計量心理学的モデル，すなわち項目反応関数を決定することは理論的にも難しいということが述べられている。また，モデルの当てはまりについての議論は限られており，多くの研究で決め打ちされたモデルが利用されている場合もある。さらに，モデル比較研究は一定数行われてきてはいるものの，実データについて，複数の CDM 間の当てはまりを統計学的に評価した先行研究はごく限られたものである。このような，CDM 同士での比較が不十分であるため，どのようなモデルがどのようなデータに適しているのか，未だ十分に明らかになっているとはいえない。これに加えて，同じテスト項目を使って，さまざまなデータ・セットに対して複数のモデルの適合を調査した研究もない。しかし，データに適している CDM が明らかでないことによって，不適切なモデルをデータに適用してしまう可能性もあり，それによって適切な診断情報を提供できないことも考えられる。テストから診断情報を抽出することが CDM の大きな目的であるため，その診断情報が誤っていることは好ましくない。そのため，データに適合しやすいモデルを探索したり，どういったデータにどういったモデルが適合しやすいのか，その傾向を探ることによって，CDM を応用する際の重要な知見を提供しうると考えられる。また，

現実のテストを用いたCDMの比較を行うことによって，実データでどういったモデルが有用であるのか示唆が得られ，モデル開発にも示唆が得られると考えられる。

また，Li et al.（2016）は利用するモデルによって，アトリビュート習得パタンの推定が異なる事を示している。診断結果を利用しようとした場合にはこうした習得パタンの違いは無意味な学習内容や適切ではない学習内容を推奨してしまう懸念もある。その結果として，効果的な学習に結びつかずテストの効果的な利用にはつながらない可能性がある。こうしたことからも，特にモデル比較に焦点を当てた検討が必要であると考えられる。

さらに，モデルが数多く開発されているが，どのモデルがどのような特徴があるのか，その全容を把握するのは必ずしも容易ではない。そのため，これまで開発されてきたモデルの持つ特徴を改めて整理し，理解しやすい観点から再度分類することによりCDMを応用しやすくする必要もあるだろう。

また，応用に際してはモデルのみならず，Q行列も分析者が設定しなければならない。認知診断テストを作成したり，CDMを適用するためには，アトリビュートの定義やQ行列の特定が必要になる。応用場面では，複数の専門家による独立な評定と協議などによって設定されることが一般的な手続きといえる。先に示した応用研究の例においても，専門家の協議などによってアトリビュートの定義がなされていた。しかし，そのように設定されたQ行列の設定はどれほど正しいのかはわからない。もしQ行列の設定が誤っており，その誤りがアトリビュート習得パタンへ悪影響を与えていた場合には，解答者への適切な指導ができずに問題が生じるだろう。この点に関して，Q行列も習得パタンの推定に影響を与えることが知られている（e.g., Rupp & Templin, 2008a）。これは，Q行列の設定を誤ること（誤設定）として研究がなされている。詳細は後述するが，先行研究ではテストに必要なアトリビュートの関係までは考慮していないという点に問題がある。

以上を踏まえて，CDMを応用する際の理論的・経験的に検討が必要な問題を整理する。まず，CDMはこれまで数多く開発されているものの，CDMの現在までの発展の様相や関連する研究知見は必ずしも十分には把握されていない。また，これまでに開発されたモデルのうち，どういったモデルが応用に利用しやすいのか，あるいはどういった特徴をもつモデルが実データに適合するのかということは十分に検討されていない。これに加えて，テストで必要なアトリビュートの関係を考慮したQ行列での誤設定の研究は十分になされていない。上記の問題意識のもと，本研究では，CDMを利用するための基礎研究として，これまでに提案されてきた各種のCDMを理論的に再検討し，Q行列の設定と診断の関連について

詳細に検討し，実データでよく適合するモデルを探索する。より具体的には，1. これまでに提案されている CDM をパラメタリゼーションから再検討し，2. CDM で重要である Q 行列の設定に関して現実のテストに近い状況を想定した場合の誤設定がアトリビュート習得パタンやモデルのパラメタに与える影響をシミュレーション実験により検討し，3. TIMSS（Trends in International Mathematics and Science Study; 国際数学・理科教育動向調査）の 2007 年 4 年生の算数データを用いた CDM のモデル比較を経験的に行うことを主要な目的とする。以下に本論文の構成を示す（図 1.2）。

第 2 章では，近年数多く開発されている CDM の理論的な整理を行う。これまではアトリビュートの関係から補償・非補償などと区別されていたモデルを統合的なモデルのパラメタリゼーションから再度検討する。これに加え，CDM において重要な要素とされる Q 行列に関する研究を整理する。Q 行列に関する研究としては，Q 行列の推定と誤設定の問題を扱う。

第 3 章では，Q 行列の設定問題に焦点をあて，アトリビュートの間の習得の依存関係（階層構造）がある場合特有の誤設定がアトリビュート習得パタンや項目パラメタの推定にどのような問題を生じさせるのかシミュレーション実験により検討する。第 2・第 3 章が理論的な検討を行う役割を果たす。

第 4 章・第 5 章は第 2 章・第 3 章での CDM の理論的な検討を踏まえて，経験的な検討を行う。具体的には，第 2 章で再検討したモデル群を用いてモデル比較研究を行う。第 4 章では，TIMSS の日本人データを用いて，第 2 章で再検討したモデルのうちどのようなモデルが実データで適合するのか，またモデルのパラメタやアトリビュート習得パタンがどのように推定されるのかを明らかにする。このようなモデルの適用を通じて，理論的なモデルが実データでどのように有用な情報を提供しうるのか，またこれまでに提案されてきたモデルのうちどのようなモデルが有効なモデルとして利用できるのか示唆を得ることを目的とする。

第 5 章では，TIMSS データのうち，対象とする国と地域を拡張して，日本人データで得られた知見についての一般化可能性についての検討を行う。これにより，解答者のアトリビュートの状態に適したモデルを探索し，その傾向を明らかにする。

第 6 章では，第 2 章から第 5 章で得られた理論的・経験的な知見を総括し，得られた知見を考察する。また，今後の研究として必要な展望を述べる。

表 1.1 Q 行列の例

項目	アトリビュート		
	加減	乗除	通分
1: 1/3 + 4/3	1	0	0
2: 1/2 × 1/3	0	1	0
3: 1/2 + 1/3	1	0	1

表 1.2 分数の計算におけるアトリビュート習得パタン

習得パタン番号	アトリビュート		
	加減	乗除	通分
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

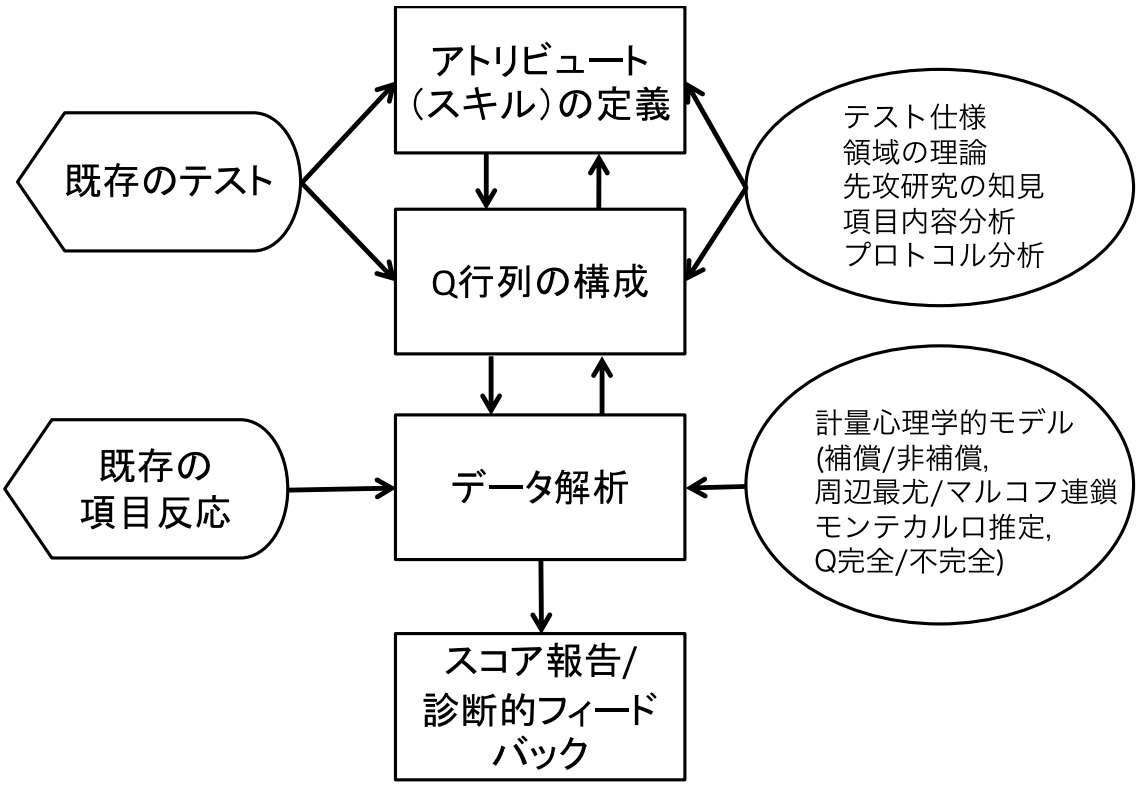


図 1.1 既存のテストに対する認知診断モデルの適用過程（Lee & Sawaki, 2009 を改変）

表 1.3 数研式 NRT でのアトリビュート習得パタンおよびテスト得点

パタン	計算力	概念理解	図形操作力	論理力	人数	テスト得点
1	1	1	1	1	165	40.3
2	1	1	1	0	46	32.2
3	1	1	0	1	24	36.5
4	1	1	0	0	54	30.8
5	1	0	1	1	34	31.1
6	1	0	1	0	122	25.8
7	1	0	0	0	52	21.7
8	0	1	1	1	5	29.0
9	0	1	1	0	2	24.0
10	0	1	0	1	2	28.0
11	0	1	0	0	30	21.4
12	0	0	1	1	24	22.0
13	0	0	1	0	58	18.5
14	0	0	0	1	2	21.5
15	0	0	0	0	328	12.7

Note. 鈴木他 (2015)の表11の一部を抜粋した。計算力と論理力を習得し，概念理解と図形操作力が未習得のパタンに分類された解答者はいなかったため，パタンの総数は15になっている。

表 1.4 MELAB リーディングテストでのアトリビュート習得パタンの比較

パタン	G-DINA	R-RUM	A-CDM	DINA	DINO
0000	.450	.409	.461	.247	.571
1000	.011	.027	.000	.006	.002
0100	.020	.069	.011	.247	.015
1100	.002	.010	.000	.000	.043
0010	.006	.006	.012	.005	.012
1010	.046	.069	.034	.018	.000
0110	.000	.001	.000	.023	.000
1110	.009	.013	.003	.021	.000
0001	.049	.010	.055	.000	.001
1001	.000	.000	.000	.022	.000
0101	.024	.023	.029	.025	.010
1101	.000	.000	.000	.022	.004
0011	.081	.048	.098	.000	.000
1011	.082	.072	.081	.055	.161
0111	.042	.056	.036	.011	.020
1111	.177	.189	.179	.298	.161

Note. Li et al. (2016)のTable 6を一部改変した。4桁の習得パタンは左から,「語彙」,「構文」,「明示的情報の抽出」,「暗黙の情報の理解」であり, 1が習得, 0が未習得に対応する。

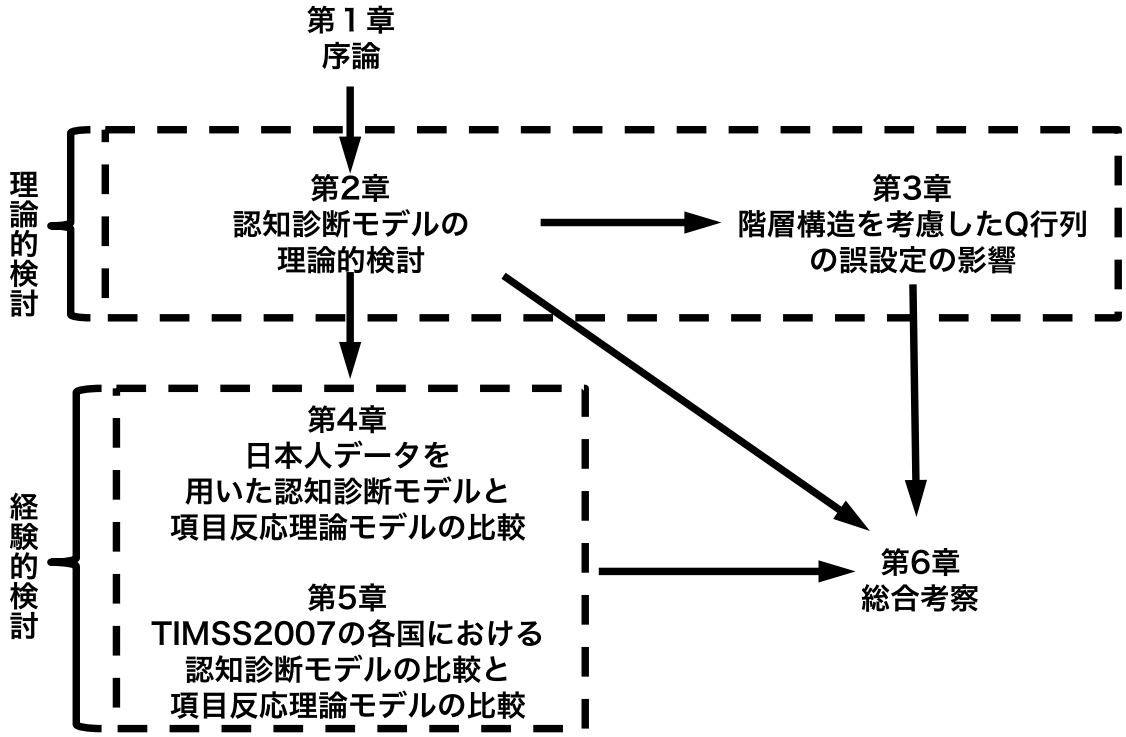

第１章
序論
理論的検討
第2章
認知診断モデルの
理論的検討
第3章
階層構造を考慮したQ行列
の誤設定の影響
経験的検討
第4章
日本人データを
用いた認知診断モデルと
項目反応理論モデルの比較
第5章
TIMSS2007の各国における
認知診断モデルの比較と
項目反応理論モデルの比較
第6章
総合考察

図 1.2 本論文の構成

第2章　認知診断モデルの理論的検討

第 2 章では，現在の認知診断モデル（CDM）研究を広く検討し，CDM 開発の現状を整理するとともに，現在提案されているモデルを新たな観点から再分類することを目的とする。これに加えて，CDM の重要な要素である Q 行列についての実践的な問題点についても検討を行い，CDM 適用研究に向けた理論的整理を目指す。2-1 では 2-2 以降のモデルを記述するための記号を導入する。2-2 では現在開発されている CDM の再分類を行う。とくに，2-2-1 から 2-2-3 でアトリビュートの非補償・補償関係および，それらの統合的モデルを主に検討する。統合的モデルは多くのモデルを包含したモデルであるものの，下位モデルの関係が不明確であり，それぞれのモデルを応用に利用しにくい可能性がある。そこで，2-2-4 で 1 項目あたりのパラメタ数に注目し，モデルの複雑さの観点から下位モデルを再度分類することを試みる。それらと合せて，2-2-5 で多値型反応に適したモデルなどより発展的なモデルについても整理することで，それまでのモデルの特徴を再考する。さらに，2-2-6 で潜在クラスモデルベースではないタイプの認知診断モデルについても検討を行い，潜在クラスモデルベースのモデルとの違いについても注目する。これに加えて，2-2-7 で CDM の中でも基本的なモデルである DINA モデルでのパラメタ推定法について述べる。CDM でのパラメタ推定法の多くは最尤推定法であり，DINA モデルでのパラメタ推定法を理解することでその推定のプロセスが理解できる。2-3 では Q 行列にまつわる諸問題を整理する。特に 2-3-1 ではデータから Q 行列を推定する方法を，2-3-2 では Q 行列の誤設定研究をとりあげる。Q 行列の誤設定は応用場面でもしばしば問題となるため，次の第 3 章でシミュレーションを用いてより詳細に検討する。

2-1 記号の準備

本章で紹介する CDM のモデルを記述するにあたり，以下の記号を定義する。

$i(=1, \cdots, I)$：解答者番号

$j(=1, \cdots, J)$：項目番号

$k(=1, \cdots, K)$：アトリビュート番号であり，$K < J$である。

$x_{ij} = \{0, 1\}$：解答者iの項目jへの反応であり，0 が誤答，1 が正答に対応する。

$q_{jk} = \{0, 1\}$：Q 行列のj行k列要素であり，0 は項目jにアトリビュートkは不要であることを意味し，1 は必要であることを表す。

$\alpha_{ik} = \{0, 1\}$：アトリビュート習得パタン行列のi行k列要素であり，0 は解答者iがアトリビ

ュートkを未習得であることを意味し，1 は習得していることを表す。

$\boldsymbol{x}_i = [x_{i1}, \cdots, x_{ij}, \cdots, x_{iJ}]^{\mathrm{T}}$：解答者$i$の項目反応ベクトルであり，記号Tはベクトルの転置を表す。

$\boldsymbol{\alpha}_i = [\alpha_{i1}, \cdots, \alpha_{ik}, \cdots, \alpha_{iK}]^{\mathrm{T}}$：解答者$i$のアトリビュート習得パタンベクトル

$\boldsymbol{q}_j = [q_{j1}, \cdots, q_{jk}, \cdots, q_{jK}]^{\mathrm{T}}$：項目$j$の Q ベクトル

$\mathrm{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_i, \cdots, \boldsymbol{x}_I]^{\mathrm{T}}$：$I \times J$の項目反応行列

$\mathrm{A} = [\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_i, \cdots, \boldsymbol{\alpha}_I]^{\mathrm{T}}$：$I \times K$のアトリビュート習得パタン行列

$\mathrm{Q} = [\boldsymbol{q}_1, \cdots, \boldsymbol{q}_j, \cdots, \boldsymbol{q}_J]^{\mathrm{T}}$：$J \times K$の Q 行列

その他必要な記号は必要に応じて導入する。

2-2 認知診断モデルの再分類

本節では，これまでに提案されてきた CDM のモデルを理論的に再分類することを目指す。既存の CDM を論じた文献には Rupp et al.（2010）, DiBello et al.（2006）, Lee & Sawaki（2009）などがある。本節では，先行研究でなされてきた分類の観点ではない新しい観点を導入することで CDM の統計モデルとしての特徴を再検討することを目指す。

CDM を分類するうえで重要となるのはアトリビュートがどのように正答確率に影響を与えるのかということである。CDM においても，多次元 IRT モデルと同様に，アトリビュート間の非補償関係・補償関係による区別が存在する。本節ではまず，これまでのアトリビュートの補償・非補償関係からモデルを整理する。さらに，それらの補償関係・非補償関係を越えた複数のアトリビュートの組合せの効果（交互作用効果）を組み込んだ統合的モデルも存在する。統合的モデルは多くのモデルを下位モデルとして含む一般性の高いモデルである。

先行研究では言及されていないモデル分類の観点として，1 項目あたりのパラメタの数が挙げられる。これは，モデルが表現できるデータの構造の複雑さと考えることができる。つまり，1 項目に対するパラメタの数が多いほどモデルが複雑であり，表現力が高いと考えられる。一方，パラメタ数の少ないモデルは制約が厳しく，適用できる場面が限られている。しかし，パラメタ数が少ないモデルであっても，先行研究からアトリビュートを厳密に定義しそのアトリビュートに適合するようにテスト項目を作成し，テストの設計を厳密に行った場合には不必要にパラメタ数が多いモデルよりもデータによく適合する可能性がある。さらに，DINA モデルや DINO モデルは想定されているパラメタの意味が簡潔であり理解し

やすいという利点も有している。一般性の高いモデルを用いて，その下位モデルの特徴を観察すると，モデルの複雑さから，これまでのモデルを整理することができる。すなわち，DINA モデルのようなモデルの制約が厳しく倹約的なモデル，アトリビュートが正答確率に加法的に影響を与える主効果モデル，主効果に加えてアトリビュート間の交互作用を含む飽和モデルが挙げられる。こういった観点からのモデルを検討し直すことにより，それぞれのモデルがもつ特徴を明確にすることができる。これにより，実データに CDM を適用する際にどのようなモデルが有用なのか新しい観点を提供することができる。

このような統合的モデルに含まれるモデル群とは別に，多肢選択に対応するモデルやアトリビュートが 2 値でないモデルなど様々な発展が見られる。こうしたモデルについても検討する。また，現在主流である CDM は潜在クラスモデルに帰着されるものが多い一方で，潜在クラスモデルに帰着されないものの，認知診断モデルと呼ばれるモデルもある。これらモデルを検討し，現状開発されている CDM の全容を把握することを目指す。

本章で扱う CDM の一覧を表 2.1 に示す。ただし，表 2.1 ではノンパラメトリックな方法・LCA にもとづかない方法といった特殊なものは，分類に適さないため除外した。表 2.1 では，反応データとアトリビュートのカテゴリ数，アトリビュート以外の潜在変数の有無，補償・非補償性，応用例，実装されているソフトウェア，1 項目あたりのパラメタ数，提案された主要な論文を示した。

2-2-1　非補償モデル

CDM の中で最もよく知られている DINA モデルを含む，非補償（noncompensatory）モデルを検討する。非補償モデルは，項目に関連する複数のアトリビュートのうち，全てが正答のためには必須であり，1 つでも求められるアトリビュートを習得していない場合には正答確率が大きく低下することが特徴である。

まず，非補償モデルの代表的モデルである DINA モデルについて示す。DINA モデルの項目反応関数を定義するために理想反応と項目パラメタを定義する。DINA モデルでは，誤差がない場合に解答者iが項目jへ正答できるか否かを表す理想反応η_{ij}を

$$\eta_{ij} = \prod_{k=1}^{K} \alpha_{ik}^{q_{jk}} \tag{2.1}$$

によって定義する。ただし，$0^0 = 1$とする。(2.1)式から，解答者iが項目jに正答するためのアトリビュートを全て習得している場合$\eta_{ij} = 1$であり，そうでなければ$\eta_{ij} = 0$となること

がわかる。さらに，このη_{ij}で条件付けたもとでの項目反応確率を

$$\Pr(x_{ij}=0|\,\eta_{ij}=1)=s_j, \tag{2.2}$$

$$\Pr(x_{ij}=1|\,\eta_{ij}=0)=g_j \tag{2.3}$$

とする。(2.2)式のs_jは，項目に正答するためのアトリビュートを全て習得しているにもかかわらず解答者が誤答してしまう確率を表しており slip パラメタと呼ばれる。(2.3)式のg_jは，正答に求められるアトリビュートが全て揃っていない解答者が偶然問題に正答する確率を表しており，guessing パラメタと呼ばれる。slip パラメタと guessing パラメタが DINA モデルの項目パラメタである。slip パラメタと guessing パラメタを用いて，DINA モデルの項目反応関数は

$$\Pr(x_{ij}=1|\eta_{ij},s_j,g_j)=\left(1-s_j\right)^{\eta_{ij}}g_j^{1-\eta_{ij}} \tag{2.4}$$

と定義される。DINA モデルという名称は，このようにアトリビュート習得パタンが決定論的に作用して理想反応が決まり，これに誤差（ノイズ）が加わって項目への正答・誤答が確率的に決まるという考えに由来している。DINA モデルのモデル名にある“and gate”とは，項目に必要とされる全てのアトリビュートを習得していなければ正答の可能性が低いことを意味している。de la Torre（2009b）に示されたモデル図を改変し，DINA モデルの解答反応 s プロセスを図示すると，図 2.1 となる。

DINA モデルとは異なり，NIDA モデル（noisy inputs deterministic “and” gate model; Junker & Sijtsma, 2001）は slip, guessing パラメタがアトリビュートkに特有のものとするモデルである。すなわち，NIDA モデルでは，slip, guessing パラメタが

$$\Pr(\eta_{ijk}=0|\alpha_{ik}=1,q_{jk}=1)=s_k, \tag{2.5}$$

$$\Pr(\eta_{ijk}=1|\alpha_{ik}=0,q_{jk}=1)=g_k \tag{2.6}$$

と定義され，slip と guessing がアトリビュートのパタメタになっている点が特徴である。このとき，項目反応関数は

$$\Pr(x_{ij}=1|\eta_{ijk},s_k,g_k)=\prod_{k=1}^{K}\left[(1-s_k)^{\alpha_{ik}}g_k^{1-\alpha_{ik}}\right]^{q_{jk}} \tag{2.7}$$

と定義される。また，解答反応プロセスは図 2.2 に示した。項目反応関数から明らかなように，NIDA モデルは項目特有のパラメタを仮定していない点が特徴である。モデル名称の“noisy inputs”はアトリビュートを適用できるか否かが確率的に決まることを意味している。

DINA モデルと NIDA モデルでは，項目の正答に求められるアトリビュートが備わっているか否かによって，正答確率が異なった。一方，Fusion model あるいは R-RUM（reduced

reparametrized unified model; Roussos, DiBello, Stout, Hartz, Henson, & Templin, 2007; Hartz & Roussos, 2008）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ときに正答確率を低下させる罰則パラメタをもつモデルである。R-RUM の項目反応関数は

$$\Pr\left(x_{ij}=1|\pi_j,\boldsymbol{r}_j,q_{jk},\alpha_{ik}\right)=\pi_j\prod_{k=1}^{K}r_{jk}^{q_{jk}(1-\alpha_{ik})} \tag{2.8}$$

によって与えられる。ここで$0<\pi_j<1$は項目jに求められるアトリビュートを全て習得している時の正答確率を表し，$0<r_{jk}<1$はアトリビュートkを習得していない場合に正答確率を減少させる罰則パラメタである。$\boldsymbol{r}_j$は項目パラメタr_{jk}をまとめたベクトルである。R-RUM の解答反応プロセスを図 2.3 に示した。なお，R-RUM のもととなった RUM（DiBello, Stout, & Roussos, 1995）は(2.8)式の関数に Rasch モデルの項目反応関数を乗じたモデルであり，

$$\Pr\left(x_{ij}=1|\ \pi_j,\boldsymbol{r}_j,q_{jk},\alpha_{ik},\theta_i,b_j\right)=\left(\pi_j\prod_{k=1}^{K}r_{jk}^{q_{jk}(1-\alpha_{ik})}\right)\left(\frac{1}{1+\exp\left(-(\theta_i-b_j)\right)}\right) \tag{2.9}$$

と定義される。RUM モデルは一次元の能力パラメタθ_iと項目の困難度パラメタb_jによっても正答確率が影響されるモデルである。この能力パラメタθ_iの関数はアトリビュートの関数としての正答確率に掛け算の形で影響を与えている。このため，能力パラメタθ_iは非補償的に項目の正答確率に機能していると解釈できる。RUM の解答反応プロセスを図 2.4 に示した。

2-2-2 補償モデル

補償（compensatory）モデルの特徴は，正答のために求められるアトリビュートのうち未習得のものがあったとしても，ほかに習得しているアトリビュートがあれば，それに応じて正答確率が上がることである。本節では，非補償モデルの節で示した DINA モデル, NIDA モデル, R-RUM に対応する補償モデルである，DINO モデル, NIDO モデル, C-RUM を導入する。

まず，DINO モデル（deterministic inputs noisy “or” gate model; Templin & Henson, 2006）では，DINA モデルとは異なり項目の正答に求められるアトリビュートを 1 つ以上習得している場合に理想反応ω_{ij}が 1 になると考える。理想反応ω_{ij}は

$$\omega_{ij} = 1 - \prod_{k=1}^{K} (1 - \alpha_{ik})^{q_{jk}} \tag{2.10}$$

によって与えられる。DINO モデルにも slip パラメタと guessing パラメタがあり，それぞれ(2.2)式，(2.3)式のη_{ij}をω_{ij}に置き換えることによって得られる。DINO モデルの項目反応関数は，それらのパラメタを用いて(2.7）式と同様に定義される。すなわち，まず DINA モデルと同様にω_{ij}を用いてs_jとg_jを

$$\Pr(x_{ij} = 0 | \omega_{ij} = 1) = s_j, \tag{2.11}$$

$$\Pr(x_{ij} = 1 | \omega_{ij} = 0) = g_j \tag{2.12}$$

と定義する。これらの slip, guessing パラメタを用いて，DINO モデルの項目反応関数は,

$$\Pr(x_{ij} = 1 | \omega_{ij}, s_k, g_k) = (1 - s_j)^{\omega_{ij}} g_j^{1-\omega_{ij}} \tag{2.13}$$

と定義される。DINO モデルにおける解答反応プロセスを図 2.5 に示した。

NIDO モデル（noisy inputs deterministic or gate model; Templin, 2006）は補償的な NIDA モデルであり，アトリビュートごとにτ_k, β_kという 2 つのパラメタを仮定する。τ_kは各アトリビュートを習得していない場合の正答確率を規定するパラメタであり，β_kは項目の正答に必要なアトリビュートを習得しているときに正答確率を変化させるパラメタである。このときの項目反応関数は

$$\Pr(x_{ij} = 1 | \boldsymbol{\tau}, \boldsymbol{\beta}, \boldsymbol{\alpha}_i, \boldsymbol{q}_j) = \frac{1}{1 + \exp\left(-\sum_{k=1}^{K} (\tau_k + \beta_k \alpha_{ik}) q_{jk}\right)} \tag{2.14}$$

と定義される。また，$\boldsymbol{\tau}$, $\boldsymbol{\beta}$はそれぞれアトリビュートパラメタτ_k, β_kを項目jに必要な個数まとめたベクトルである。NIDO モデルは NIDA モデルと同様に，パラメタτ_k, β_kは項目間で同等とみなしている。NIDO モデルの解答反応プロセスを図 2.6 に示した。

C-RUM（Compensatory RUM; Templin, 2006）はアトリビュートが個別に項目の正答確率に寄与することを表現する。その項目反応関数は

$$\Pr(x_{ij} = 1 | \lambda_{j0}, \boldsymbol{\lambda}_j, \boldsymbol{\alpha}_i, \boldsymbol{q}_j) = \frac{1}{1 + \exp\left(-\left(\lambda_{j0} + \sum_{k=1}^{K} \lambda_{jk} \alpha_{ik} q_{jk}\right)\right)} \tag{2.15}$$

である。$\boldsymbol{\lambda}_j$はそれぞれの項目の正答に求められる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パラメタλ_{jk}をまとめたベクトルである。それぞれの項目に対して，これらのパラメタはK個必要であるわけではなく，$\sum_k q_{jk}$個のみ仮定され，q_{jk}が 0 にあたるλ_{jk}は推定されない。λ_{j0}は切片に対応する当て推量パラメタである。C-RUM の解答反応プロセスを図 2.7 に示した。C-RUM と類似したモデルとしては，カーネルとしてアトリビュート

の線形関数を仮定するのではなく，直接正答確率を変化させるように項目パラメタを仮定する A-CDM（Additive CDM; de la Torre, 2011）や，LLM（linear logistic model; Maris, 1999）がある。

2-2-3　統合的モデル

統合的モデルは，複数のアトリビュートを習得している場合にアトリビュート間の交互作用や，アトリビュートと Q 行列の要素を引数とする関数を仮定することで，補償・非補償モデルを統合する一般的なモデルの枠組みを提供するモデルであり，飽和モデルとも呼ばれる（Li et al., 2016）。具体的には対数線形認知診断モデル（log linear cognitive diagnostic model, LCDM; Henson, Templin, & Willse, 2009），一般化 DINA（generalized DINA, G-DINA; de la Torre, 2011）モデル，一般化認知診断モデル（generalized cognitive diagnostic model, GDM; von Davier, 2007a, 2014a）の 3 つのモデルが挙げられる。これらのモデルはパラメタ変換することで，同等の推定値を得られる場合があることが示されている（de la Torre, 2011; von Davier, 2014b）。まず，IRT モデルとの類似性の高い LCDM について述べる。

LCDM の項目反応関数は

$$\Pr(x_{ij}=1|\lambda_{j0},\boldsymbol{\lambda}_j,\boldsymbol{\alpha}_i,\boldsymbol{q}_j)=\frac{1}{1+\exp\left(-\left(\lambda_{j0}+\boldsymbol{\lambda}_j^{\mathrm{T}}f(\boldsymbol{\alpha}_i,\boldsymbol{q}_j)\right)\right)} \tag{2.16}$$

である。ここで，$\boldsymbol{\lambda}_j$は項目jのパラメタベクトルである。$f(\boldsymbol{\alpha}_i,\boldsymbol{q}_j)$は$\boldsymbol{\alpha}_i$と$\boldsymbol{q}_j$を引数として，それらの組合せの項を出力する関数であり

$$\boldsymbol{\lambda}_j^{\mathrm{T}}f(\boldsymbol{\alpha}_i,\boldsymbol{q}_j)=\sum_{k=1}^{K}\lambda_{jk}\alpha_{ik}q_{jk}+\sum_{k=1}^{K-1}\sum_{k'>k}\lambda_{jkk'}\alpha_{ik}\,\alpha_{ik'}q_{jk}q_{jk'}+\cdots \tag{2.17}$$

と表される。(2.17)式は，(2.15)式で示されているアトリビュートの主効果だけではなく，2 次の交互作用やより高次の交互作用も含む。ここでは表記が煩雑になるため，最高次の交互作用の記述は省略した。LCDM の解答反応プロセスを図 2.8 に示した。

このほか，LCDM, G-DINA, GDM においては，パラメタ$\boldsymbol{\lambda}_j$に適切な制約をおくことで，2-2-1，2-2-2 で述べた各種補償・非補償モデルをその下位モデルとして表現できる。具体的には，LCDM の切片と全ての主効果パラメタに等値制約をおき，最高次の交互作用のみを推定すれば(2.4)式の DINA モデルが，主効果・交互作用のパラメタの絶対値に等値制約をおき，n次の交互作用に重み$(-1)^{n+1}$をかければ(2.13)式の DINO モデルが（例えば，主効果λ_{j1}には1,1 次の交互作用λ_{j12}には，-1の重みをかける），さらに交互作用項を推定しなければ

(2.15)式の C-RUM を表現できる。このほか, R-RUM（Chiu & Köhn, 2016）や NIDA, Generalized NIDA（de la Torre, 2011）といったモデルも，LCDM や G-DINA モデルは包摂している。より詳細な関係に関して，詳しくは Henson et al.（2009）や de la Torre（2011）を参照のこと。

G-DINAモデルは項目反応関数が(2.16)式のLCDMとほぼ同様であり, パラメタの値をそのまま正答確率の変化量と解釈することができる。G-DINAモデルは一般化線形モデルのようにlogリンク（対数リンク），ロジットリンク（logitリンク），恒等リンクといったリンク関数を用いることが可能である。例えば，ロジットリンクを用いると，G-DINAモデルはLCDMと等価なモデルとなる（de la Torre, 2011），項目jの正答に関わるアトリビュート数は$K_j^* = \sum_{k=1}^{K} q_{jk}$であり，アトリビュートの習得パタンの数は，$2^{K_j^*}$となる。また，項目$j$の正答に関わるアトリビュートを要素に持つベクトルを$\boldsymbol{\alpha}_{lj}^*$で表す。ここで，添え字$l$は，特定のアトリビュートの習得パタンを示す。このとき，G-DINAモデルのアトリビュート習得パタン$\boldsymbol{\alpha}_{lj}^*$をもつ解答者が項目$j$に正答する項目反応関数は

$$\Pr\left(x_j = 1|\boldsymbol{\alpha}_{lj}^*\right) = \delta_{j0} + \sum_{k=1}^{K_j^*} \delta_{jk}\alpha_{lk} + \sum_{k'=k+1}^{K_j^*} \sum_{k=1}^{K_j^*-1} \delta_{jkk'}\alpha_{lk}\alpha_{lk'} + \cdots + \delta_{j12\cdots K_j^*} \prod_{k=1}^{K_j^*} \alpha_{lk} \tag{2.18}$$

と表される。ここでは1からK_j^*番目のアトリビュートについてのみ考えているとしても一般性を失わない。モデル識別の条件や，パラメタの制約の詳細については，de la Torre（2011）に記載されている。$\Pr\left(x_j = 1|\boldsymbol{\alpha}_{lj}^*\right)$は項目の正答に関係するアトリビュートの主効果項と，複数のアトリビュート間の交互作用効果項から構成される。例えば，δ_{j0}は必要なアトリビュートを全く習得していない場合の正答確率であり切片項である。δ_{jk}は1つのアトリビュート（α_k）を習得している場合の正答確率の変化量を示しており，α_kの主効果項と解釈することができる。交互作用項$\delta_{jkk'}$（$k \neq k'$）は，α_kと$\alpha_{k'}$の異なる2つのアトリビュートを習得している場合の正答確率の変化量であり，一次の交互作用効果といえる。同様に，$\delta_{j12\cdots K_j^*}$は項目jの中での最高次の交互作用効果である。G-DINAモデルの解答反応プロセスを図 2.9に示した。

(2.18)式のG-DINAモデルでDINAモデルを表現するには，最高次の交互作用$\delta_{j12\cdots K_j^*}$と切片を推定し$\delta_{jk}, \delta_{jkk'}$といった主効果や低次の交互作用効果を0と制約すればよい。交互作用項は複数のアトリビュートを習得している場合の正答確率の変化量であり，最高次の交互作用はその項目の正答に必要な全てのアトリビュートがある場合にのみ正答率が上昇することから非補償的な影響を表す項と考えられる。DINAモデルにおいて，アトリビュートの交互作用効果とは，例えば分数の計算問題において，足し算と掛け算が必要な問題のときにど

ちらの演算を優先するのか，あるいは，片方の演算の結果を次にどのように利用するのか，といった2つ以上のアトリビュートが揃った場合に生じる新しい能力やスキルとしても解釈できる。DINAモデルのパラメタにG-DINAモデルのパラメタを対応させるとguessingパラメタは$g_j = \delta_{j0}$，slipパラメタは$s_j = 1 - \delta_{j0} - \delta_{j12\cdots K_j^*}$と表現される。

LCDMの場合と同様に，G-DINAモデルでDINOモデルを表現する際には$\delta_{jk} = -\delta_{jk'k''} = \cdots = (-1)^{K_j^*+1}\delta_{j12\cdots K_j^*}$と制約をかける。これにより，DINOモデルと同様に，一つでもアトリビュートを習得している場合に一様に項目に正答できる確率が上昇する項目反応関数を表すことができる。A-CDMを表現するためには，(2.18)式のG-DINAモデルのパラメタのうち，交互作用効果をすべて0に固定し（$\delta_{jkk'} = \cdots = \delta_{j12\cdots K_j^*} = 0$），主効果のみを推定すればよい。

また，(2.18)式のG-DINAモデルにおいて$\Pr(x_j = 1|\boldsymbol{\alpha}_{lj}^*)$に対してロジットリンク関数および対数リンク関数を仮定することで，それぞれLogit CDM, Log CDMというモデル群を生成できる。これまで示してきたDINA, DINO, A-CDMは恒等リンク関数を仮定したモデルである。Logit CDMは$\mathrm{logit}\left(\Pr(x_j = 1|\boldsymbol{\alpha}_{lj}^*)\right)$が$\alpha_{lk}$の線形関数になるモデルであり，LCDMと同値のモデルである。ここで$\mathrm{logit}(\cdot)$は$\mathrm{logit}(\Pr(\cdot)) = \log(\Pr(\cdot)/(1 - \Pr(\cdot))$という変換である。LLMは項目反応のカーネルを

$$\mathrm{logit}\left(\Pr(x_j = 1|\boldsymbol{\alpha}_{lj}^*)\right) = \delta_{j0} + \sum_{k=1}^{K_j^*}\delta_{jk}\alpha_{lk} \tag{2.19}$$

とするモデルであり，すなわち交互作用のないLogit CDMである。

Logit CDMと同様に，Log CDMは$\log\left(\Pr(X_j = 1|\boldsymbol{\alpha}_{lj}^*)\right)$が$\alpha_{lk}$の線形関数になるモデルである。具体的には

$$\log\left(\Pr(x_j = 1|\boldsymbol{\alpha}_{lj}^*)\right) = \nu_{j0} + \sum_{k=1}^{K_j^*}\nu_{jk}\alpha_{lk} \tag{2.20}$$

とすれば，Log CDMの交互作用がないモデルを表現できる。さて，R-RUMはこのLog CDMに含まれるモデルとして定式化される。まず，R-RUMの項目反応関数は(2.8)式を

$$Pr(x_j = 1|\boldsymbol{\alpha}_{lj}^*) = \pi_j\prod_{k=1}^{K_j^*}r_{jk} \times \prod_{k=1}^{K_j^*}\left(\frac{1}{r_{jk}}\right)^{\alpha_{lk}} \tag{2.21}$$

と変形する。ここで，項目jに必要なアトリビュートのみを考慮すればよいので，KをK_j^*とすることができる。(2.21)式の対数をとると

$$\log\left(\Pr\left(x_j = 1|\boldsymbol{\alpha}_{lj}^*\right)\right) = \log\left(\pi_j\right) + \sum_{k=1}^{K_j^*} \log\left(r_{jk}\right) - \sum_{k=1}^{K_j^*} \alpha_{lk} \log\left(r_{jk}\right) \quad (2.22)$$

となり，$\nu_{j0} = \log\left(\pi_j\right) + \sum_{k=1}^{K_j^*} \log\left(r_{jk}\right)$および$\nu_{jk} = -\log\left(r_{jk}\right)$とおけば，R-RUMがLog CDMの下位モデルであることがわかる。以上のように，LCDMと同様に多くのモデルがG-DINAモデルの下位モデルに位置づけられる。

G-DINAモデルとLCDMはパラメタリゼーションとして直接確率を表すパラメタδを仮定するか，対数線形モデルのように逆ロジット変換をして線形モデルになるパラメタλを仮定しているのかのみが異なり，本質的なコンセプトは同等と考えられる。ただし，G-DINAモデルはリンク関数を明示的に導入している点，LCDMはアトリビュートとQ行列の要素を引数とする関数を任意に指定できる点がそれぞれのモデルの違いである。LCDMの関数$\boldsymbol{\lambda}_j^{\mathrm{T}} f(\boldsymbol{\alpha}_i, \boldsymbol{q}_j)$の設定次第では，G-DINAモデルが表現できないモデル設定も可能と考えられる。

G-DINA モデル・LCDM に並び一般性の高い GDM は 2 値に限らない離散的な潜在特性や項目反応を統合的に扱うことが可能であり，後述する多母集団解析，LCA モデルなど不均質な集団の解析に適した方法も扱えるよう拡張されている。また，LCDM における$f(\boldsymbol{\alpha}_i, \boldsymbol{q}_j)$関数に相当する関数を柔軟に設定することで，離散的な潜在特性を用いて通常の IRT モデルとほぼおなじモデルを構成する方法も提案されている（von Davier, 2007b）。

上記のように G-DINA モデル・LCDM および GDM は，2 値の観測変数の CDM を包含するモデルといえる。特に LCDM，G-DINA モデルが包含するモデル群は，統合モデルに制約をかけて表現が可能であるため，現在主流な CDM として位置づけることができよう。

2-2-4　パラメタ数からみたモデルの再分類

統合的モデルから，その下位モデルが含んでいるモデルの性質を考えることができ，新しいモデルの分類を行うことができる。de la Torre（2011）や Henson et al.（2009）では，おもに CDM でしばしば注目されていたアトリビュートの補償・非補償関係に注目していた。しかし，下位モデルのパラメタの複雑さに注目したモデルの分類はこれまで行われていない。とくに，1 つの項目あたりのパラメタ数に注目したモデルの分類は行われていない。こうした点に注目することにより，モデルの制約の強さについてより直感的に理解することができる。さらに，モデルを利用する際の観点の 1 つとして，パラメタ数に注目するというという新しい視点を与えることに繋がる。こうしたことにより，どのモデルがどのように強い制約を持ったモデルなのかがわかり，モデルの複雑さからデータに適合しやすいモデルを選

択することもできるようになる。以下では，統合的モデルで表現できるモデルを，倹約モデル，主効果モデル，飽和モデルの 3 つに大別する。

まず，倹約モデルには DINA モデルと DINO モデルが含まれる。DINA・DINO モデルは 1 項目あたりの項目パラメタが 2 つのみであり，これは数多く提案されている CDM の中でも最も少ないモデルである。つまり，DINA・DINO モデルは最も制約が厳しいモデルであり，これらのモデルを実データに適用したとしてもデータの複雑な構造に適合しない可能性がある。その一方，2 つの項目パラメタは slip パラメタと guessing パラメタであり，パラメタの意味は理解しやすいという利点もある。

主効果モデルは，それぞれのアトリビュートが加法的に問題の正答確率に影響を与えるモデルであり，アトリビュートの間の交互作用効果を仮定しないものである。主効果モデルとしては，R-RUM，A-CDM，LLM が含まれる。これらのモデルは制約が厳しすぎることもなく，また G-DINA モデルや LCDM ほど複雑でもない中庸なモデル群である。主効果モデルは倹約モデルよりも 1 項目あたりのパラメタ数が多く，データに適合しやすいことが期待できる。また，項目パラメタがそれぞれのアトリビュートが正答確率に与える影響の大きさと考えることができ，解釈も比較的単純である点も特徴的である。しかし，アトリビュートの交互作用が表現できないため，複数のアトリビュートがあって初めて得られる効果の検証はできない。

最後の飽和（saturated）モデルは統合的モデルを指す。つまり，ありうるアトリビュートの組み合わせの効果をモデル化することで，より複雑な項目反応関数を表現できるものである。これらのモデルは理論的にすべてのアトリビュートの組み合わせをモデルに含んでいるため，他の下位モデルよりも尤度の点で高い適合を示すことが特徴として挙げられる。また，主効果，交互作用の両方を含んでおりアトリビュートの複雑な影響の関係も表現することができる。

これらの倹約モデル，主効果モデル，飽和モデルが適している場面について，パラメタの制約の強さから考えてみる。これらのモデルは制約の強さが異なっていることから，制約が厳しいモデルほどそれに対応するようにアトリビュートに関する事前知識が求められる可能性がある。

まず，倹約モデルは最も制約の強いモデルであるため，DINA モデルや DINO モデルの使用を前提として，テストを作成する必要がある。診断を目的として新しく作成されたテストのようにアトリビュートを明確に定義した場合など，アトリビュートに対して詳細な想定

がある場合には，DINA モデルや DINO モデルでもデータに適合する可能性は高くなる。こうした制約の強い DINA モデルや DINO モデルが適合することで，他のモデルが適合した場合よりもより診断情報が信用できる可能性もある。つまり，倹約モデルが適合すれば診断の結果も理論的背景を踏まえたものであり，信用できるということである。

飽和モデルはアトリビュートの交互作用をすべて考慮している。そのため，アトリビュートの組み合わせの効果について事前に予測ができない場合には，飽和モデルを用いることでアトリビュートの組み合わせの効果を探索することができる可能性がある。何故ならば，アトリビュートがどのように正答確率に影響するのかわからないとしても，飽和モデルは主効果・交互作用を全て含んでおり，アトリビュートの不要な組合せの効果はないのであれば，その組み合わせの効果に対応するパラメタが 0 と推定されることが期待されるからである。また，飽和モデルは複雑なアトリビュートの効果を表現するためにも利用できるため，積極的な仮説がある場合にも有効となる可能性はある。もちろん，複雑なアトリビュートの関係を事前に想定した上で飽和モデルを利用することもできるが，探索的な使用方法も可能であり，飽和モデルの柔軟性が役に立つと考えられる。ただし，飽和モデルはパラメタ数が多い分，他のモデルよりも，パラメタの推定の標準誤差が大きくなりやすいため，他のモデルよりも大きめのサンプルサイズが必要と考えられる。

主効果モデルは，制約の強さが倹約モデルと飽和モデルの中間に当たるモデルであり，想定したアトリビュートのうち幾つかが正答確率に影響を与えているような場合に有用である可能性がある。主効果モデルは積極的に複数のアトリビュートの関係を想定できず，個別に影響を与える場合に有用である可能性がある。言い換えると，厳密なアトリビュートの交互作用が定義出来なくても主効果モデルはデータに適用できる可能性がある。また，主効果モデルは，その項目パラメタを参照することで，Q 行列を修正するための簡便な方法として利用できることもある。例えば，ある問題の正答に必要なアトリビュートが複数想定される場合，それらのアトリビュートが負荷すると仮定した Q 行列を作成し，項目パラメタを推定することで，不要なアトリビュートの項目パラメタは正答に影響しない値（0 や 1 など）と推定されることが期待される。つまり，不要なアトリビュートは正答確率に対して影響を及ぼさないため，主効果モデルを用いることにより推定された項目パラメタをもとに，Q 行列を修正するという手続きを踏むことができる。

2-2-5　発展モデル

ここまで述べたモデルは，LCA をベースにした項目反応が 2 値，潜在変数がカテゴリカルである場合についてであった。本項ではこれを拡張した，多枝選択解答に適したモデル，多値アトリビュートを仮定したモデル，複数の Q 行列を仮定するモデル，アトリビュートに対して高次の連続的潜在変数を仮定したモデル，変量効果モデル，潜在移行分析を応用したモデルについて検討する。

まず，名義尺度の形式である多枝選択式解答に適したモデルとして Multiple choices DINA（MC-DINA; de la Torre, 2009a; DiBello, Henson, & Stout, 2015; Ozaki, 2015）がある。MC-DINA では，多枝選択枝のそれぞれに対してアトリビュートを付与し，アトリビュートパタンごとに，選択枝の選択確率をモデリングする。例えば，各項目のアトリビュートが付与されている選択枝数をHとし，選択枝の番号をhとする。さらに$\boldsymbol{q}_{jh} = [q_{ih1}, \cdots, q_{ihk}, \cdots, q_{ihK}]^{\mathrm{T}}$という長さ$K$の列ベクトルを仮定する。要素$q_{ihk}$は，他のモデルの Q 行列の要素のように，選択枝$h$にアトリビュート$k$が必要であれば 1，そうでなければ 0 をとる。すべての選択枝にアトリビュートが付加されているとすると，アトリビュートパタン$\boldsymbol{\alpha}_i$は$H+1$種類のグループに分類できる。また，ここで$q_{j0} = 0$とする。このとき，解答者iの項目jへの異なる反応を示すグループを$c_{ij}\ (= 0, \cdots, H)$として番号をつける。さらに，このc_{ij}を決定する関数を

$$c_{ij} = \arg\max_{h'} \{\boldsymbol{\alpha}_i^{\mathrm{T}} \boldsymbol{q}_{jh'} | \boldsymbol{\alpha}_i^{\mathrm{T}} \boldsymbol{q}_{jh'} = \boldsymbol{q}_{jh'}^{\mathrm{T}} \boldsymbol{q}_{jh'}\} \tag{2.23}$$

とする。ただし，$h' = 0, \cdots, H$であり，$\boldsymbol{\alpha}_i$がどの選択枝の$\boldsymbol{q}_{jh'}$のパタンにも合致しない場合は$c_{ij} = 0$とする。$\boldsymbol{q}_{jh}$の具体例を使って，(2.22）式の決定方式を考える。表 2.2 では，項目jに対して，3 つの選択枝が存在し，テスト全体で仮定されている 4 つのアトリビュートのうちはじめの 3 つが正答に必要と仮定されている。このとき，例えば，はじめの 3 つのアトリビュートを全て習得しているパタンでは，(2.23)式の$\boldsymbol{\alpha}_i^{\mathrm{T}} \boldsymbol{q}_{jh'} = \boldsymbol{q}_{jh'}^{\mathrm{T}} \boldsymbol{q}_{jh'}$を満たす$h'$は1,2,3である。この中で$\boldsymbol{\alpha}_i^{\mathrm{T}} \boldsymbol{q}_{jh'}$が最大になるのは$h' = 3$のときであり，このアトリビュート習得パタンを持つ解答者iの項目jでのグループは$c_{ij} = 3$となる。また，4 つ全てアトリビュートを習得しているパタンも$c_{i'j} = 3$となる。2 番目，3 番目のアトリビュートを習得しているパタンでは，同様に，$\boldsymbol{\alpha}_i^{\mathrm{T}} \boldsymbol{q}_{jh'} = \boldsymbol{q}_{jh'}^{\mathrm{T}} \boldsymbol{q}_{jh'}$を満たす$h'$は 2, 3 となり，グループ番号は3となる。どのアトリビュートも習得していない場合は$h' = 0$であり，グループ番号も0となる。

MC-DINA モデルは，このグループ番号を用いて選択枝hへの反応確率$\Pr(x_{ij} = h | c_{ij} = c)$を考えるモデルである。これは解答者$i$が項目$j$に対して，グループが$c\ (= 0, \cdots, H)$であるときに選択枝$h$を選ぶ確率であり，$c$で条件付けたとき$\sum_h \Pr(x_{ij} = h | c_{ij} = c) = 1$を満たす。つ

まり，解答者は選択枝のうちどれかには反応すると仮定する。このように選択枝ごとにアトリビュートを導入することで，正答誤答の 2 値反応では扱えていない錯乱枝（distractor; 一見すると正答にもみえる誤答選択枝）も含めた，より豊かな情報を解答確率のモデリングに利用できることが MC-DINA モデルの利点である。

Ozaki（2015）は複数選択枝項目をモデリングする際にパラメタを de la Torre（2009a）のモデルよりも少なく表現できるモデルとともに，de la Torre（2009a）では表現できないモデルを提案した。DiBello et al.（2015）は複数選択枝にもとづくスコアリングに対する一般化診断分類モデル（generalized diagnostic classification models for multiple choice option-based scoring, GDCM-MC）を提案した。GDCM-MC で特徴的なのは，アトリビュート空間の拡張，Q 行列の拡張，当て推量のモデリング，モデリングフレームワークの拡張にある。まず，1 つ目の特徴としてはアトリビュートを習得すべきスキルのみならず，問題のある誤概念も含むものとして扱うようにした点が挙げられる。次に，Q 行列の要素として，0, 1 に加えて“*N*”と呼ばれる状態を追加した。これは，特定のアトリビュートが特定の選択枝の選択に影響を与えないことを明示する状態である。また複数選択枝のモデリングでは当て推量のモデル化が問題となるが，GDCM-MC では当て推量もアトリビュートの状態によって決まるパラメタとして推定する。最後に，各選択枝を選択する確率を表現するために IRT モデルの名義尺度モデルのようなモデル表現を行うことで，カーネルの関数を柔軟に表現することが可能となっている。GDCM-MC は多肢選択枝を分析する CDM の中では非常に一般性が高い表現を可能としている。

多値アトリビュート G-DINA（polytomous attributes G-DINA, pG-DINA）モデルでは，アトリビュートは習得水準を示す順序カテゴリカル変数とみなされる（Chen & de la Torre, 2013）。pG-DINA モデルでは，Q 行列の要素も多値となり，特定アトリビュート水準習得項目（specific attribute level mastery items）という，同じアトリビュートが求められる項目であっても，異なった習得段階を表現する項目を導入する。モデルでは，多値アトリビュートを当該の項目に求められるアトリビュートのみを抽出した退化アトリビュートベクトル（reduced attribute vector）と，これを 2 値に変換した崩壊アトリビュートベクトル（collapsed attribute vector）を用いる。この崩壊アトリビュートベクトルを用いて項目反応関数を構成する。

de la Torre & Douglas（2008）や Huo & de la Torre（2014）の Multiple-strategy DINA（MS-DINA）モデルでは，通常の DINA モデルと異なり，複数の Q 行列を用いて多様な解答方略

を表現する。ここで，$m(=1,\cdots,M)$を方略を表す番号とし，$\mathbf{Q}_{\mathrm{m}}$をm番目の方略の Q 行列とする。またq_{jkm}を$\mathbf{Q}_{\mathrm{m}}$における項目jのk番目の要素とし，方略ごとに$\eta_{ijm}=\prod_{k=1}^{K}\alpha_{ik}^{q_{jkm}}$とする。このとき，解答者$i$の項目$j$への理想反応を$\eta_{ij}=\max\{\eta_{ij1},\cdots,\eta_{ijm},\cdots,\eta_{ijM}\}$によって与える。このように$\eta_{ij}$を定めたあとは，DINA モデルと同様に項目反応関数を定義できる。

さらに，解答者が項目ごとにM個の方略を使い分けていることを有限混合モデル（McLaclen & Peel, 2000）として拡張することを考える。$\delta_{ij}(=1,\cdots,M)$を解答者iが項目jで採用する方略とするとき，項目反応関数を

$$\Pr(x_{ij}=1)=\sum_{m=1}^{M}\Pr(\delta_{ij}=m|\boldsymbol{\alpha}_i)\left(1-s_{jm}\right)^{\eta_{ijm}}g_{jm}^{1-\eta_{ijm}} \tag{2.24}$$

によって与える。ただし$\sum_m \Pr(\delta_{ij}=m|\boldsymbol{\alpha}_i)=1$である。このような混合パラメタ$\delta_{ij}$を導入することで，各解答者がどのような方略を用いているかをより詳細に検討することが可能となる。ここでのδ_{ij}パラメタは方略の混合割合を示すもので G-DINA モデルの項目パラメタとは異なっている点に注意が必要である。

高次 DINA（higher order DINA, HO-DINA; de la Torre & Douglas, 2004）モデルはαを従属変数とし，θの適当な単調増加関数としたモデルである。α_{ik}は 2 値変数であるため，2 パラメタロジスティックモデルを仮定すると，

$$\Pr(\alpha_{ik}=1|\theta_i,a_k,b_k)=\frac{1}{1+\exp(-a_k(\theta_i-b_k))} \tag{2.25}$$

とモデル化される。a_kはアトリビュートkの習得に関わる識別力パラメタ，b_kは困難度パラメタである。なお，本稿ではθは一次元と仮定しているが，多次元を仮定することも可能である。また，このような階層モデルはレベル 2（例えば，学校）のグループレベルのパラメタを仮定した Hierarchical GDM（H-GDM）としてより一般的な形式で与えられている（von Davier, 2007a）。このモデルは，これまで陽に仮定されてこなかったアトリビュートの相関を明示したものと考えられる。関連したモデルとしては，Templin & Henson（2006）のテトラコリック相関モデルが挙げられるが，これはアトリビュート間の相関構造をモデル化したものである。このような構造を仮定する事によって，アトリビュート習得パタンと項目パラメタの推定を頑健に行うことができる（Templin, Henson, Templin, & Roussos, 2008）。

高次変量効果 DINA（higher order random effect DINA, HO-R-DINA; Ayers, Rabe-Hesketh, & Nugent, 2013）モデルは，複数回のテストを実施した場合に HO-DINA の高次能力パラメタに対して，$\gamma_i^{(0)}$や$\gamma_i^{(1)}$という個人差を仮定したモデルであり，

$$\theta_{it} = \gamma_i^{(0)} + (t-1)\gamma_i^{(1)} \tag{2.26}$$

である。ただしtは時点を表す。$\gamma_i^{(0)}$や$\gamma_i^{(1)}$は適当な平均と分散を持つ正規分布にしたがうと仮定すれば，先に言及した階層モデルと同様になる。また，これはθ_{it}を従属変数とした潜在成長曲線モデル（Latent growth curve mode; e.g., Preacher, Wichman, MacCallum, & Briggs, 2008）とみなすことができる。つまり HO-R-DINA は反復測定に適した CDM である。

潜在移行分析（Collins & Lanza, 2010）を応用したモデル（Latent transition CDM, LTA-CDM）も提案されている。具体的には，Li, Cohen, Bottge, & Templin（2016）は$\Pr(\alpha_{it}|\alpha_{i(t-1)})$という，ある時点でのアトリビュートパタンで条件付けた次の時点での各アトリビュートパタンの確率を考えることで，アトリビュートパタンの変遷をモデル化している。

LCA モデルを基調としたモデルとしては，多母集団同時解析を組み合わせた Multi-group CDM（George & Robitzsch, 2014; Xu & von Davier, 2008）や，slip や guessing に対する変量効果を仮定した Random effect DINA（Huang & Wang, 2014）といったモデルも開発されている。Random effect DINA モデルでは slip，guessing パラメタに Rasch モデルなどを適用し，解答者ごとに slip, guessing パラメタが変化することを許容している。先に登場した HO-DINA モデルでは，アトリビュートの習得確率を規定する能力値パラメタを仮定しており，slip, guessing パラメタに個人差はなく，Random effect DINA モデルではアトリビュートの習得確率は解答者によって変化しないものの slip, guessing パラメタが解答者によって異なるという点で異なっている。

さらに，Templin & Bradshaw（2014）は LCDM を基調にした，Hierarchical Diagnostic Classification Model（H-DCM）を提案した。H-DCM では，アトリビュート間の従属関係（階層関係）を仮定したものであり，後述のアトリビュート階層法に似た発想を取り入れている。アトリビュートの階層関係を仮定することによって，不要な習得パタンを削減したり，結果の解釈を容易にすることが可能になると期待できる。

発展的モデルでは，アトリビュートの補償・非補償という関係をモデル化するよりも，回答形式や反復的なデータ収集に適したモデル化が行われていると考えられる。また，発展的なモデルのベースとなっているのは，倹約的なモデルである DINA モデルである傾向がみられる。このように，発展的なモデルも基本的には LCA に基づくモデルに基本をおき，様々なパラメタを追加することで，複雑な状況を表現するモデルを開発する方向性がみられる。

2-2-6　非潜在クラス型モデル

ここまでは，LCA モデルを基本としたモデルを検討してきた。しかし，CDM とみなされる分析モデルはこれにとどまらず，ノンパラメトリック分類法，第 1 章でも紹介したルールスペース法およびアトリビュート階層法，IRT モデルの拡張なども包含しうる。さらに，ベイジアンネットワーク（e.g., 繁桝・植野・本村, 2007）も診断に利用されるという意味では CDM といえる。以下ではそれらのモデルの特徴を示す。

ノンパラメトリック分類法（nonparametric classification method; Chiu & Douglas, 2013）は DINA モデルで定義した理想反応パタンと観測されたパタンの乖離度$d(\boldsymbol{x}_i, \boldsymbol{\eta}^{(l)})$を用いる。ただし$\boldsymbol{\eta}^{(l)}$は$l\ (=1,\cdots,2^K)$番目のアトリビュート習得パタンにおける理想反応を表す。解答者の反応と理想反応はいずれも要素が 2 値のベクトルであるため，乖離度としてはハミング距離を用いる。ハミング距離は 2 値のデータの不一致度を示す距離であり，2 つの 2 値ベクトルのそれぞれの要素が一致していない個数を数えたものである。アトリビュートパタンの推定値$\hat{\boldsymbol{\alpha}}_i$はハミング距離を最小化するパタンである。さらに，項目jに正答した解答者の割合を$\bar{p}_j$として，$\bar{p}_j(1-\bar{p}_j)$という分散の逆数を重みとした，重み付きハミング距離も利用される。Chiu, Douglas, & Li（2009）は K-means 法を CDM に応用した手法を提案しており，分類の一致性も示されている（Chiu & Köhn, 2015）。

第 1 章でも紹介した RSM（Rule space method; Tatsuoka, 2009; 龍岡・倉元, 2006）は Q 行列からアトリビュート習得パタンと習得パタンに対応する理想反応パタンを導出する。さらに，1 次元の IRT モデルの潜在特性値に直交する軸を導入し，その空間へ知識状態ごとの理想反応パタンと，解答者の反応パタンを布置する。そして，マハラノビス距離等を用いて解答者の反応パタンに最も近い理想反応の知識状態に分類を行う。

AHM（Attribute hierarchy method; Leighton et al., 2004; Gierl, 2007）は RSM を発展させ，アトリビュートの間の階層的な関係を積極的に取り入れたモデルである。アトリビュートの階層関係とは，あるアトリビュートを習得するためには他のアトリビュートを予め習得していなければならない，といった習得の順序関係や認知モデル上での処理の順序関係である。具体的には，「直線型」，「分岐型」，「収束型」，「非構造」などの構造が仮定されており，この認知能力の積極的なモデル化が RSM との相違点であるとされる（Leighton et al., 2004）。

IRT モデルベースの CDM としては, MLTM-D（Multicomponent latent trait model for diagnosis; Embretson & Yang, 2013）が挙げられる。MLTM-D は非補償多次元 IRT モデルを拡張して診

断を可能にしたものである。MLTM-D ではこれまで示した分類を志向したモデルと同様に Q 行列を用いてどのアトリビュートが項目に対して求められるのかを項目パラメタとして反映しながら，解答者のパラメタは従来の多次元 IRT モデルのように各アトリビュートに連続的な潜在特性を仮定するモデルである。

この他に，IRT モデルをテスト全体に適用して得られる解答者の潜在特性θ_iを用いて，$\Pr(\alpha_{ik} = 1|\theta_i)$を得ることを目指した最小二乗距離モデル（least squares distance model, LSDM; Dimitrov, 2007; Dimitrov & Atanasov, 2012）も提案されている。このモデルは，HO-DINA とは異なり，一般的な能力とその下位能力としてのアトリビュートを仮定するわけではない。他の CDM と同様に補償モデル（Dimitrov, 2007）と，非補償モデル（Dimitrov & Atanasov, 2012）が提案されている。

ベイジアンネットワーク（Almond, DiBello, & Zapata-Rivera, 2007）は有向非巡回グラフにもとづいて，同時確率分布を条件付き確率に分解する。診断的検査のモデリングにおいては，解答者の能力ノードと能力間のリンク関係からなる能力モデルと，観測変数と親のノードである能力からなる証拠モデルを構成し，ノードと親ノードの関係としてこれまでも述べられてきた補償関係，非補償関係といった関係性を仮定することが可能である（Almond et al., 2007）。ベイジアンネットワークは，潜在変数モデルや LCA モデルも表現できる非常に一般性の高い枠組みである。教育測定の文脈では，解答者の問題解決過程のモデリングや，測定における証拠中心アプローチでの中核的な役割を果たしている（Mislevy, 1994; Mislevy & Haertel, 2006）。

非潜在クラス型の CDM においては，ノンパラメトリック分類法が DINA モデルを基本としており，MLTM-D が MLTM を元にしているものの，多くのモデルは LCA に基礎をおく CDM とは大きく異なっている。これらのモデルではそれぞれに特徴が大きく異なっており，モデル間の関係について系統的な類似性を見出すことは難しい。こうしたモデルも認知診断モデルの名前がついていたり，診断のために利用されていることから，認知診断が多様なモデルを包含している事がわかる。

2-2-7　パラメタ推定法

CDM の推定方法は，EM アルゴリズムを用いた周辺最尤推定法（marginal maximum likelihood estimation vie the EM algorithm, MMLE-EM; de la Torre, 2009b）やマルコフ連鎖モンテカルロ（Markov chain Monte Carlo; e.g., Bolt & Lall, 2003）法など通常の潜在変数モデルで

利用される推定方法が利用可能である。ここではDINAモデルに関してのde la Torre（2009b）が示したEMアルゴリズムについて概説する。これまでに提案されているCDMの多くはMMLE-EMにもとづくものである。そのため，CDMのEMアルゴリズムベースでの項目パラメタ推定法の手続きは，DINAモデルでの推定方法を発展させた形式になることが多い。DINAモデルのEMアルゴリズムを理解することで，他のモデルでの推定方法を理解するための基礎的な考え方を理解することができると考えられる。

まず，$\Pr(\boldsymbol{\alpha}_l)$は$l(=1,\cdots,2^K)$番目のアトリビュート習得パタンを得る事前確率とする。さらに，アトリビュート習得パタンで条件付けたもとでの項目反応パタン$\boldsymbol{x}_i$を得る確率である尤度を$\Pr(\boldsymbol{x}_i|\boldsymbol{\alpha}_l)$とする。このとき，解答パタン$\boldsymbol{x}_i$を得た元でのアトリビュートパタン$\boldsymbol{\alpha}_l$を得る事後確率$\Pr(\boldsymbol{\alpha}_l|\boldsymbol{x}_i)$は

$$\Pr(\boldsymbol{\alpha}_l|\boldsymbol{x}_i)=\frac{\Pr(\boldsymbol{x}_i|\boldsymbol{\alpha}_l)\Pr(\boldsymbol{\alpha}_l)}{\sum_{l=1}^{L}\Pr(\boldsymbol{x}_i|\boldsymbol{\alpha}_l)\Pr(\boldsymbol{\alpha}_l)} \tag{2.27}$$

となる。この時，アトリビュート習得パタン$\boldsymbol{\alpha}_l$をもつ解答者の期待度数を$I_l=\sum_{i=1}^{I}\Pr(\boldsymbol{\alpha}_l|\boldsymbol{x}_i)$，項目$j$に正答する期待度数を$R_{jl}=\sum_{i=1}^{I}x_{ij}\Pr(\boldsymbol{\alpha}_l|\boldsymbol{x}_i)$とする。さらに，$I_{jl}^{(0)}$を項目$j$の正答に必要なアトリビュートを1つ以上欠いている解答者の期待度数，$R_{jl}^{(0)}$を$I_{jl}^{(0)}$の中で項目jに正答する期待度数をあらわすとする。これらに加え，$I_{jl}^{(1)}$を項目jに必要なアトリビュートを全て持っている解答者の期待度数，$R_{jl}^{(1)}$を$I_{jl}^{(1)}$の中で正答する解答者の期待度数とする。全てのjに対して$I_{jl}^{(0)}+I_{jl}^{(1)}=I_l$，$R_{jl}^{(0)}+R_{jl}^{(1)}=R_l$で一定となる。これらを用いて，$s_j,g_j$の推定量を，

$$\hat{s}_j=\frac{I_{jl}^{(1)}-R_{jl}^{(1)}}{I_{jl}^{(1)}} \tag{2.28}$$

$$\hat{g}_j=\frac{R_{jl}^{(0)}}{I_{jl}^{(0)}} \tag{2.29}$$

とすることによって，対数尤度を項目パラメタで偏微分したスコア関数を最大化することができる。これを用いてDINAモデルの項目パラメタの推定のためのEMアルゴリズムをまとめると，

Step1　$\boldsymbol{s},\boldsymbol{g}$に適当な初期値を与えて初期化する

Step2　$I_{jl}^{(0)},I_{jl}^{(1)},R_{jl}^{(0)},R_{jl}^{(1)}$を現在の$\boldsymbol{s},\boldsymbol{g}$にもとづいて計算する

Step3　$\hat{\boldsymbol{s}},\hat{\boldsymbol{g}}$を(2.28)式および(2.29)式によって計算し，$\boldsymbol{s}=\hat{\boldsymbol{s}},\boldsymbol{g}=\hat{\boldsymbol{g}}$とする

Step4　Step2, Step3を収束するまで実行する

という手続きをとる。これがDINAモデルの項目パラメタの最尤推定値を求める手続きと

なる。

アトリビュート習得パタン$\boldsymbol{\alpha}_i$は，MMLE，事後確率最大化法（MAP 推定法）や期待事後推定法（EAP 推定法）によって推定される。MMLE と MAP 推定法では，データを得た元で最も蓋然性が高いアトリビュート習得パタンを推定値とするが，EAP 推定法ではアトリビュート習得パタンを周辺化するため，各アトリビュートの習得確率を得て，習得確率が.5 を超える場合にアトリビュートを習得しているとみなす。ただし，EAP 推定法を用いる場合には，情報を捨象せずに習得確率を用いることもある。どのような推定法を用いるかは，テスト結果のフィードバックの形態などによって決定される。de la Torre & Douglas（2004）は DINA モデルや HO-DINA での MCMC のサンプリングアルゴリズムを提案した。WinBUGS（Spiegelhalter, Thomas, & Best, 1999）などの MCMC サンプラーを用いた DINA モデルのベイズ推定は DeCarlo（2012）のパラメタリゼーションによって可能である。

LCDM やその下位モデルは Mplus（Muthén & Muthén, 1998-2017）などの汎用的な潜在変数モデルを推定するためのソフトウェアを用いて推定可能である（Templin & Hoffman, 2013）。G-DINA や下位モデルに分類されるモデル，および HO-DINA は統計解析環境 R で利用できる CDM パッケージ（George, Robitzsch, Kiefer, Gross, & Uenlue, 2016）や NPCD パッケージ（Zheng & Chiu, 2016）, GDINA パッケージ（Ma & de la Torre, 2017）により推定可能である。この他のモデルは，各研究者が開発しているソフトウェアなどで推定可能である。例えば，GDM の推定は mdltm（von Davier, 2006）という独自開発されたソフトウェアにより実行できる。CDM の推定に利用できるソフトウェアを，モデルごとに表 2.1 に記載した。かなり多くのモデルは何らかのパッケージを用いて推定できるものの，複雑なモデルについてはこの限りではなく，MCMC 法を用いるなどして推定する必要がある。なお，Rupp（2009）や Lee & Sawaki（2009），Li et al.（2016）も同様にソフトウェアの情報をまとめている。

2-3 Q 行列に関する研究

Q 行列の作成に関しては前章で示したように，まず当該のテスト領域ではどのようなアトリビュートが問題解決に必要であるかを同定する必要がある。続いて，項目の分析や，教科の専門家の意見なども参考にしながら項目とアトリビュートの関係を定めていく。場合によっては，解答者の発話プロトコルを用いて，解答者がどのような解答プロセスを経ているのかを検討することも行われる。このように 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ものであり，時間と労力のかかるプロセスである。

先にCDMの応用研究を検討した際に明らかとなったように,CDMの応用研究の多くは,既存の大規模調査やテストにCDMを適用しており,特定の診断のためにテストを作成しているケースは必ずしも多くないといえる。既存のテストにCDMを適用する場合には，そのテストが目的としたことや，テストの仕様書に依拠して，アトリビュートの設定やQ行列の作成がなされる。しかしながら，その仕様書などはあくまでテスト作成者の意図にすぎない場合もあり，解答者が行っている問題解答時に使用されるアトリビュートと必ずしも一致するとは限らない。先に述べたように，Q行列は先行研究や関連分野の知見を用いて決定されるものの，Q行列の作成は非常に手間の掛かる作業である。しかし，その一方で実データに即したQ行列を作成するための方法論も重要である。

そこで，近年データからQ行列を推定する手法の開発や，Q行列の設定を誤った場合の影響についての研究がなされている。また，本稿ではあまり触れないがこの他にQ行列に関する話題として，モデルの識別性を規定するQ行列の設定が挙げられる（e.g., Xu, 2017; Xu & Zhang, 2016）。モデルの識別の問題は，例えば探索的因子分析モデルや共分散構造分析でしばしば取り上げられる問題であるが，CDMにおいても最近議論され始めている話題である。モデルが識別されるとは，尤度が異なれば項目パラメタ・アトリビュート習得パタンが異なるということである。この識別性を保証するためには，Q 行列の構成が重要である。項目パラメタ・アトリビュート習得パタンの識別性を保証するためのQ行列が満たすべき条件についても研究が進んでいる（e.g., Xu & Zhang, 2016）。とくに，Q 行列が完備（complete）であることが識別性の必要条件となる（Xu & Zhang, 2016）。Q行列が完備とはQ 行列がそれぞれのアトリビュートを単独で測定している項目を含んでいる場合のことを指し，そうでなければQ行列は完備ではないという（Xu & Zhang, 2016）。このように識別性の問題はCDMにおいても重要な話題であるものの，まだ研究が始まったばかりであるため，ここでは詳細な議論は省略する。しかし，アトリビュート習得パタンの推定を利用する応用場面では識別性の問題は重要な点であるため，第 4 章・第 5 章の経験的検討の際には再度この識別性について触れる。そのため本稿では，Q行列の推定方法と，誤設定研究の2点に焦点を絞り先行研究の検討を行う。

2-3-1　Q行列の推定研究

先に述べたように，Q行列の設定は分析に先立って行われる。しかしながら，近年Q行

列をデータから推定するという研究も盛んに行われている。本節では，Q 行列の推定に関しての最近の展開を示すこととする。

Q 行列は CDM に特有の要素であり，テスト作成者によって設計されるものであることは，すでに述べた。しかしながら，後述するように Q 行列の特定を誤ることによって，アトリビュート習得パタンや項目パラメタの推定にバイアスが生じる（e.g., Baker, 1993; Liu, 2017; Im & Corter, 2011; Kunina‐Habenicht et al., 2012; Liu & Huggins-Manley, 2016; Rupp & Templin, 2008a）。通常は真の Q 行列を知り得ないため，テストでどの程度 Q 行列の誤設定の影響があるのかは未知数である。この Q 行列の誤設定は解析者が Q 行列を設定する以上はいかなる場合に生じても不自然ではない。このような誤設定の問題を回避するために，最近 Q 行列の推定方法が盛んに研究されている。本節ではこの問題に関して，現在まで得られている方法論を検討する。CDM は確証的な手法であるが，前述のように Q 行列の設定は必ずしも容易ではない。そのため，Q 行列の修正・推定方法は CDM の利用者にとって Q 行列の妥当な設定を行うことを援助し，診断結果が正答である可能性を増加させるため有益な手法であるといえる。Q 行列の推定方法は，モデルを仮定したもとで，ベイズ推定や最尤推定，残差平方和最小を目指す手法，行列因子分解に基づく手法と様々なバリエーションがある。以下，それぞれについて検討する。

DeCarlo（2012）は，Q 行列の特定の要素のみを確率変数として推定を行うベイズ的アプローチを示した。さらに，全ての Q 行列の要素を確率変数として，MCMC によって推定する手法も提案されている（Chung, 2014）。ただし，これらの手法を利用するためには，モデルの設定が正しいことが前提となっている点に注意が必要である。

Sun, Ye, Inoue, & Sun（2014）, Sun, Ye, Shi, Wang, & Sun（2014）, Sun, Ye, Sun, & Kameda（2015）は Liu, Xu, & Ying（2012）, Liu, Xu, & Ying（2013）, Xiang（2013）などで提案されたデータ駆動の Q 行列推定法を改良し，DINA モデルにもとづく Q 行列の最尤推定法を提案している。特に，特徴的なのが，Boolean Matrix Decomposition というブール代数演算を伴う行列積を分解する手法を提案している点である。DINA モデルの理想反応は Q 行列とアトリビュート習得パタンのブール代数演算を用いて表現することができる。このことを利用し，項目反応から Q 行列とアトリビュート習得パタンを分解する手法として Boolean Matrix Decomposition が開発された。

de la Torre（2008）は DINA モデルをベースにし，$1-s_j-g_j$を最大化するアトリビュートパタンを Q 行列の行ベクトルとする推定を提案している。さらに，この方法を発展させ，

de la Torre & Chiu（2016）は項目ごとに誤った要素を特定するための指標を開発し，シミュレーション実験によって，多くのモデルで誤設定された要素の修正がなされることを確認している。

Chiu（2013）は観測反応と理想反応の間の残差 2 乗和（residual sum of squares, RSS）の最小化を目指す方法を Qmatrix refinement method としている。具体的な項目jの RSS は

$$\mathrm{RSS}_j = \sum_{i=1}^{I}\left(x_{ij} - \eta_{ij}\right)^2 = \sum_{l=1}^{2^K}\sum_{i\in C_l}^{I}\left(x_{ij} - \eta_{jl}\right)^2 \tag{2.30}$$

と定義される。$i \in C_l$はl番目のクラスに属する解答者を意味する。Q 行列の行ベクトル$\boldsymbol{q}$の誤設定の損失関数としての RSS は解答者の分類が正確であれば，$\boldsymbol{q}$ベクトルの設定が正しいときに最小になることが期待される。

Q 行列の推定に行列因子分解を利用する方法は，Q-matrix method（Barnes, 2005）によって提案されたものが最初期の研究である。その後，非負行列因子分解（Non-negative matrix factorization; NMF; Lee & Seung, 1999; Lee & Seung, 2001; Cichocki, Zdunek, Phan, & Amari, 2009）を用いた方法の提案がなされてきた。NMF は要素が非負である行列を 2 つ以上の要素が非負である行列に分解する手法である。NMF は要素が正の行列のみを用いて観測データの近似を行う手法であり，通常の重み付き加重和とは異なって四則演算のうち加法演算のみが許容されている。この制約によって，分解した行列の要素に 0 が多くスパースになりやすく，解釈が容易になる傾向がある。このような傾向があり，Lee & Sueng（1999）は顔画像に NMF を適用することによって，眉や目や口，鼻などの顔のパーツのみを抽出することが可能になったことを報告している。

さて，CDM のうち DINA モデルでは項目反応パタンは Q 行列とアトリビュート習得パタン行列の積として，

$$\neg(X^T) = Q(\neg(A)) \tag{2.31}$$

と表現できる（Barnes, 2005; Desmarais, 2011; Desmarais, Beheshti, & Naceur, 2012）。ここで，¬は否定演算子で行列の 0 以外の要素を 0 に，0 を 1 に変換する演算である。Xは項目反応行列であり，Q 行列，アトリビュート習得パタン行列はすべて非負であるため，NMF を適用することが可能となる。ただし，推定された Q 行列はそのままでは正の実数をとるものであり，何らかの基準にもとづいて 0，1 に値を丸め込む処理が必要である。

Desmarais, Beheshti, & Xu（2014）では NMF が de la Torre（2008）や Chiu（2013）の手法

よりも，精度よく推定できることを示している。NMF による Q 行列推定は理解しやすく，実用的である一方，推定された Q 行列の統計的性質は十分に明らかになっていない。

このように，Q 行列の推定の研究は近年極めて活発であり，CDM のユーザーにも使いやすいようにソフトウェアの整備も徐々に進んできており，CDM を利用しやすくするための基盤が整いつつあるといえる。例えば，これまでに紹介した手法の中で，de la Torre & Chiu（2016）の方法や Q-matrix refinement method は，R のパッケージを用いて実行可能であり，利便性は高いといえる。

Q 行列の推定は有用である。しかし，Q 行列の推定では因子分析における因子数の決定と同様にアトリビュート数を決定する問題がある。この問題に対して Beheshti, Desmarais, & Naceur（2012）は, 特異値分解を用いたアトリビュート数決定の方法を提案しているものの，CDM の文脈においてはアトリビュート数を決めるための統計的手法についての検討はあまりなされていない。通常の CDM の適用場面では CDM が確証的モデルでありアトリビュート数は理論的に既知であると考えことによって，アトリビュート数決定の問題を回避している。しかし，アトリビュート数を統計的に決定する方法は，実データに適したアトリビュートを保証するために必要な方法論であろう。アトリビュート数決定に関して，応用上行われる実践としては，例えば，Jang（2009）では，英語テストに CDM を適用する際に，そのアトリビュートが負荷している項目数が 3 以下のアトリビュートを分析から除外するといった手続きをとっている。Q 行列の推定手法と合わせて，テストの解答に関連した理論的な正当性を合わせて議論がなされる必要があるだろう。

2-3-2　Q 行列の誤設定研究

Q 行列をデータから推定する研究が活発であり，十分なデータがあれば Q 行列の推定が可能であるということも示されつつある。しかしながら，現状の CDM の応用ではそういった Q 行列の推定法を利用できるほどのサンプルサイズを確保できる状況は少ない。診断的目的にテストを作成した場合には分析に利用できるサンプルサイズは数百から多くとも 2000 程度である。例えば，学校場面で CDM を適用しようとする場合には，1 学学年のサンプルサイズが限度であろう。

そのような状況が通常の状況であり，やはり現状の Q 行列の設定は先に示した当該テスト領域の専門家や教科教育の専門家，その他文献研究，テスト項目の内容分析，発話プロトコルの分析などを駆使した比較的定性的情報に依存せざるを得ない。そうであるならば，Q

行列の設定がどれほど正しいのか，量的に評価することは難しく，正しくない Q 行列の設定を行ってしまうこともあると考えられる。

これまでの Q 行列の設定を誤ることについての研究（Q 行列の誤設定）から，Q 行列の誤設定はアトリビュート習得パタンや，項目パラメタの推定値にバイアスをもたらすことが明らかになっている。例えば，Rupp & Templin（2008a）は DINA モデルを用いて Q 行列の誤設定の影響を系統的に検討し，誤設定によって解答者のアトリビュート習得パタンの推定にバイアスが生じること，さらに項目パラメタの推定値にも影響があることを示した。また，Im & Corter（2011）は RSM を用いた場合でも，Q 行列の誤設定によって，同様にアトリビュート習得パタンの推定精度が低下することを示した。この他, Baker（1993）は LLTM の計画行列（CDM の Q 行列に相当）の誤りについて検討した。具体的には，Q 行列の誤設定の量が同じ場合，1 つの項目の解答に必要なアトリビュートが多い場合のほうが少ない場合よりも項目パラメタの推定値への影響が少ないことを示した。さらに，Baker（1993）は一箇所でも Q 行列に誤りがある場合にパラメタの推定に与える影響は少なからずあることを指摘した。このように，誤設定は CDM において深刻な問題であることが示されてきた。

先に示した Q 行列作成手続きを踏まえて，Q 行列の誤設定がどのように生じうるのか，その作成の手続きから整理を行う。先に示したように，Q 行列の作成には複数の段階があるため（e.g., Lee & Sawaki, 2009; Tjoe & de la Torre, 2014），その段階に改めて触れる。まず，文献調査によって，当該テスト領域の認知理論や学習理論にもとづいてアトリビュートを定義する。次に，明らかになったアトリビュートが妥当であるかを複数の専門家の合議によって検討する。ここで不適切なアトリビュートは再度定義を行う。既存のテストに CDM を当てはめる場合には，テスト項目を分析し，問題文や出題意図，テストの仕様書を参照してアトリビュートを決定する。また，新規に診断テストを作成する場合には，決定したアトリビュートを測定している項目の作成を行う。さらに，このようにアトリビュートを定義した後にどの項目にどのアトリビュートが必要であるかを検討する。この際には，複数人の専門家が Q 行列を作成した後に，一致率を算出し，不一致である部分を修正する。また，問題を解く際の解答者の発話思考のプロトコルデータを参考にして Q 行列を作成・修正を行なうこともある。これらの手続きは繰り返し行われ，アトリビュートの統合・追加を経て Q 行列が作成される。

このように，入念な調査にもとづいて Q 行列が作成されるものの，必ずしも Q 行列が一意に定まるわけではない。例えば，Li（2011）では MELAB リーディングテストに CDM を

適用し，Q 行列を作成する際に複数の専門家が項目の分析を行った。この際に，専門家同士の一致率にスピアマンのρを用いており，.3 以上という一致率を得た。この値は統計的に有意であったものの，専門間の見解が必ずしも一致していないことを表している。この結果から，項目に付与したアトリビュートがどれほど正しいのか，その程度はわからず，誤設定の可能性は拭えない。以上の議論から，誤設定は CDM を適用する際に普遍的に生じる問題といえる。

これらの誤設定のなかで，特に問題となる誤設定の特定を行う。誤設定は「アトリビュート数の誤設定」と「要素の誤設定」に大別できる。まず，「アトリビュート数の誤設定」は，テストで測定しているアトリビュートセットの定義が不適切であることによって生じる。「要素の誤設定」は項目に付与されるべきアトリビュートが付与されない，あるいは逆に項目に不要なアトリビュートが付与される誤設定である。これら 2 つの誤設定のうち，「アトリビュート数が変化する誤設定」は，根本的な理論の構築などを行い直すことで対応可能である。また，文献のレビューによって認知モデル・情報処理プロセスが仮定されているのであれば，アトリビュート構造が既知の状態と考えられるため，このような誤設定は生じにくい。一方で，「要素の誤設定」は出題意図と解答者の解答方法の食い違いが生じている場合には看過できない可能性がある。この意味で，後者の誤設定はより実際的に生じうる問題であり，その詳細についての検討が行われなければ，診断の有用性を損ないかねない問題であるといえよう。

Q 行列の誤設定の悪影響が指摘されている一方で，先行研究では現実的なテストにみられるアトリビュートの階層構造（attribute hierarchy structure, AHS）を十分に考慮した検討は行われていない。AHS とは，あるアトリビュートが別のアトリビュートに従属する構造であり，CDM が着目する側面の 1 つである構造化された手続き的スキル・知識である（Yang & Embretson, 2007）。具体的なテストでは，Gierl et al.（2009）は AHM を用いて，代数の問題の解答に必要なアトリビュートの関係を記述しており，“演算記号や示された数字の理解能力”の習得を前提として，そのあとに，“線形関数を解くスキル”や“簡単な代入問題（1 文字に 1 変数を代入）に必要なスキル”を習得することを示した。この他の AHS の例は，SAT®の代数の分析（Gierl, Wang, & Zhou, 2008; Gierl, Zheng, & Cui, 2008; Gierl et al., 2009），三段論法のメンタルモデルの分析（Leighton et al., 2004），批判的読解スキルへの適用（Wang & Gierl, 2011）にみられる。このように，AHS は特別なものではなく，龍岡・林（2001）は構造を持たず独立なアトリビュートを仮定して実用的なテストを作成することは不可能で

あると述べている。

AHS は診断モデルを利用するために必要な構造であるといえるが，このような階層構造以外に，仮にアトリビュート間に相関が仮定される場合，さらに高次因子を仮定したモデルも提案されている（e.g., Rupp at al., 2010）。これらは，アトリビュート間の相関構造をモデル化したもので，学習の順序についての理論的な想定はとくにない。つまり，このモデルの仮定する相関構造は AHS を意味しておらず，また，認知モデルに必ずしも則っておらず適切とはいえない。AHS を仮定することはテストの妥当性の証拠を集める観点からも推奨されることであり（Yang & Embretson, 2007），そのような認知的な理論がない中でのフィードバック情報は利用価値が十分には発揮できないと考えられる。

このように，現実的なテストの状況を想定した場合には AHS を想定する必要があり，その場合の Q 行列の誤設定の影響について検討する必要があるといえる。これについては，次章でシミュレーション研究を用いてより詳細に検討する。

表 2.1 近年提案されているモデルの特徴

モデル	反応データ	アトリビュートのカテゴリ水準	アトリビュート以外の潜在変数	補償・非補償	応用研究例	推定ソフト	1項目あたりの項目パラメータ数	文献	備考
DINA	2値	2値	無	非補償	Lee, Park, & Taylan (2011)	CDM・GDINA	2	Haertel (1989)	
DINO	2値	2値	無	補償	Templin & Henson (2006)	CDM・GDINA	2	Templin & Henson (2006)	
NIDA	2値	2値	無	非補償	なし	CDM・GDINA	$2K_j^*$	Junker & Sitjma (2001)	K_j^* は項目jに必要なアトリビュート数を意味する。
NIDO	2値	2値	無	補償	なし	CDM・GDINA	$2K_j^*$	Templin (2006)	
LLM	2値	2値	無	補償	Chen & de la Torre (2014)	CDM・GDINA	$1+K_j^*$	Maris (1999)	
A-CDM	2値	2値	無	補償	Chen & de la Torre (2014)	CDM・GDINA	$1+K_j^*$	de la Torre (2011)	
R-RUM	2値	2値	無	非補償	Jang (2009)	CDM･GDINA･Mplus	$1+K_j^*$	Roussos, DiBello, Stout, Hartz, Henson, & Templin (2007)	
C-RUM	2値	2値	無	補償	山口 (2016)	Mplus	$1+K_j^*$	Templin (2006)	
G-DINA	2値	2値	無	飽和	鈴木他(2015)	CDM・GDINA	$2^{K_j^*}$	de la Torre (2011)	
LCDM	2値	2値	無	飽和	Templin & Hoffman (2013)	Mplus	$2^{K_j^*}$	Henson, Templin, & Willse (2009)	
GDM	多値	多値	無	飽和	von Davier (2008)	mdltm・CDM	$1+K_j^*$	von Davier (2007a)	パラメータ数は関数 $f(\boldsymbol{\alpha}_i, \boldsymbol{q}_j)$に依存する。
MC-DINA	多値	2値	無	非補償	なし	CDM・GDINA	$(1+H_j^*)(H_j-1)$	de la Torre (2009a)	H_j は項目jの選択枝の数,H_j^*はアトリビュートが付加された選択枝の数を表す。
HO-DINA	2値	2値	有	非補償	なし	CDM・GDINA	2	de la Torre & Douglas (2004)	
MS-DINA	2値	2値	無	非補償	なし	WinBUGS	$2M$	de la Torre & Douglas (2008)	M は方略数を表す。
pG-DINA	2値	多値	無	飽和	Chen & de la Torre (2013)	CDM・GDINA	$2^{K_j^*}$	Chen & de la Torre (2013)	
H-GDM	多値	多値	有	飽和	von Davier (2007a)	mdltm	$SG(K_j^*+1)$	von Davier (2007a)	Sはクラスター数，Gはグループ数であり，各クラスター内に同じ数のグループを仮定した。
H-DCM	2値	2値	無	飽和	Templin & Bradshaw (2014)	Mplus	$2^{K_j^*}$	Templin & Bradshaw (2014)	
Multi-group CDM	2値	2値	無	非補償	Xu & von Davier (2008a)	CDM･GDINA･mdltm･Mplus	$G(K_j^*+1)$	Xu & von Davier (2008a)	モデルによってはこの限りではない。
Random effect DINA	2値	2値	有	非補償	Huang & Wang (2014)	WinBUGS	2	Huang & Wang (2014)	
LTA-DINA	2値	2値	有	非補償	Li, Cohen, Bottge, & Templin (2015)	WinBUGS	2	Li, Cohen, Bottge, & Templin (2015)	
HO-R-DINA	2値	2値	有	非補償	Ayers, Rabe-Hesketh, & Nugent (2013)	WinBUGS	2	Ayers, Rabe-Hesketh, & Nugent (2013)	

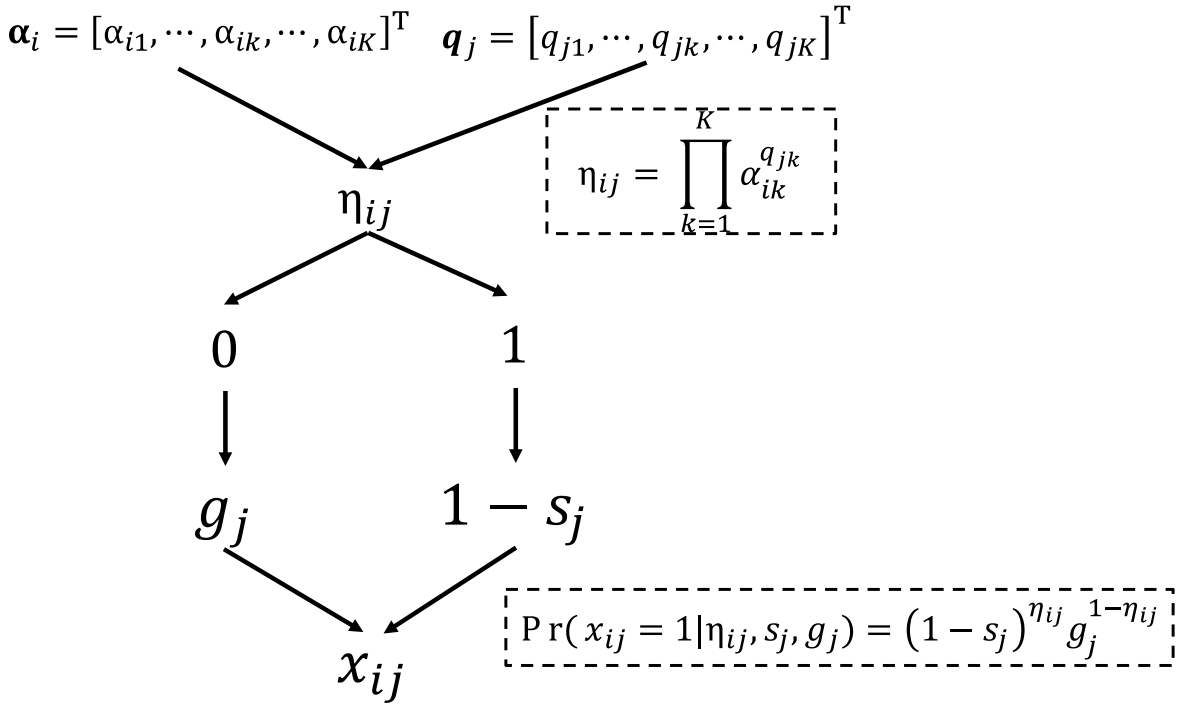


図 2.1 DINA モデルにおける解答反応プロセス（de la Torre, 2009b を改変）

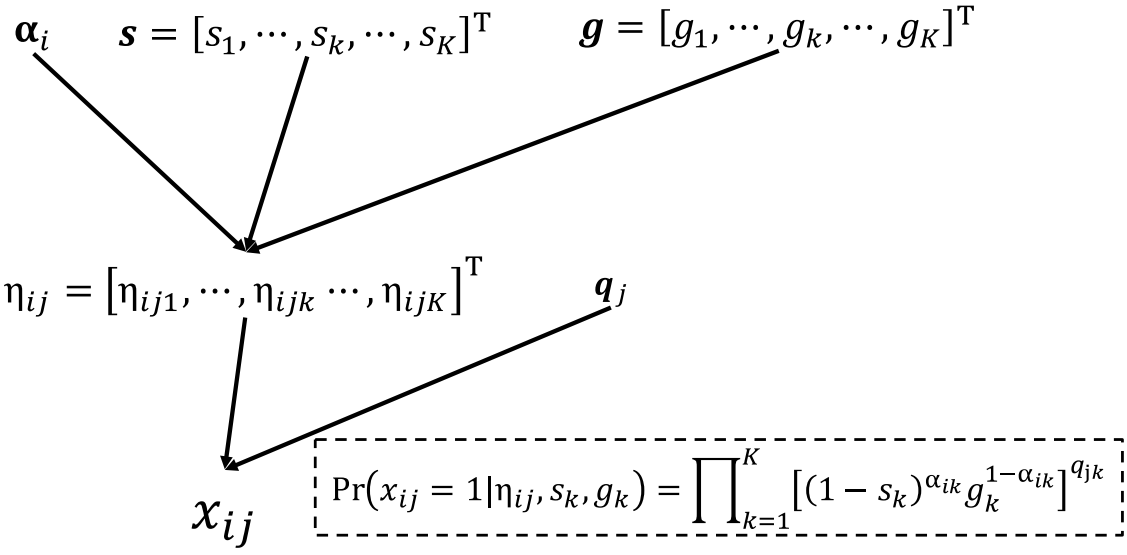


図 2.2 NIDA モデルにおける解答反応プロセス

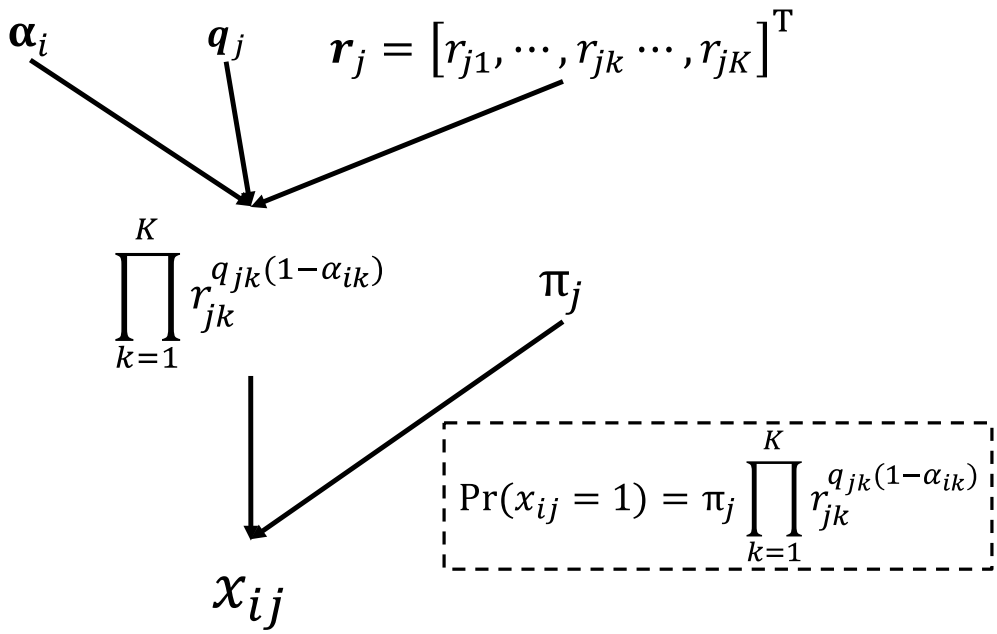


図 2.3 R-RUM モデルにおける解答反応プロセス

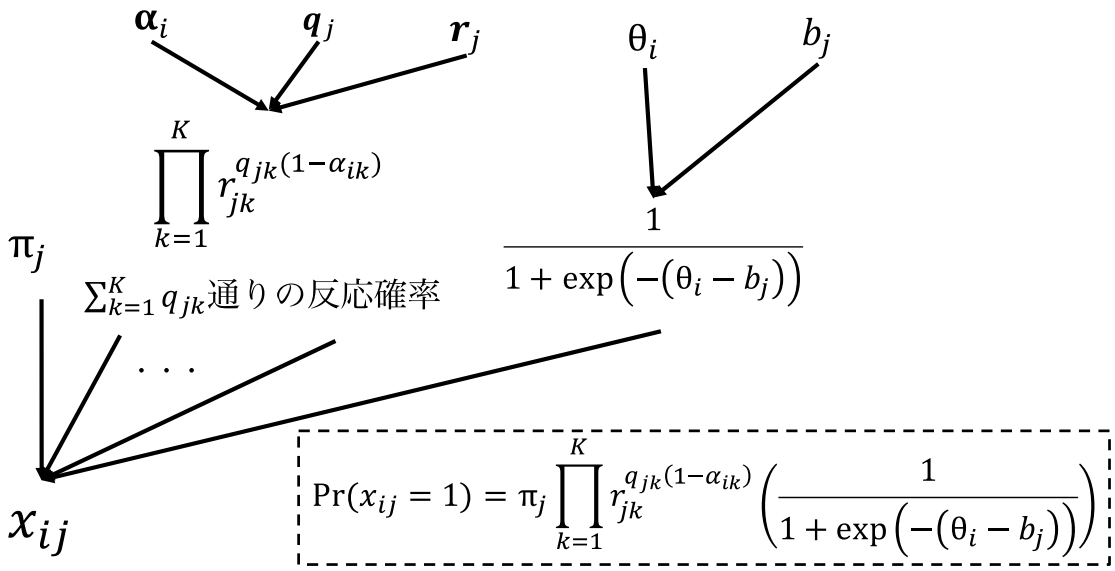


図 2.4 RUM モデル（オリジナル）における解答反応プロセス

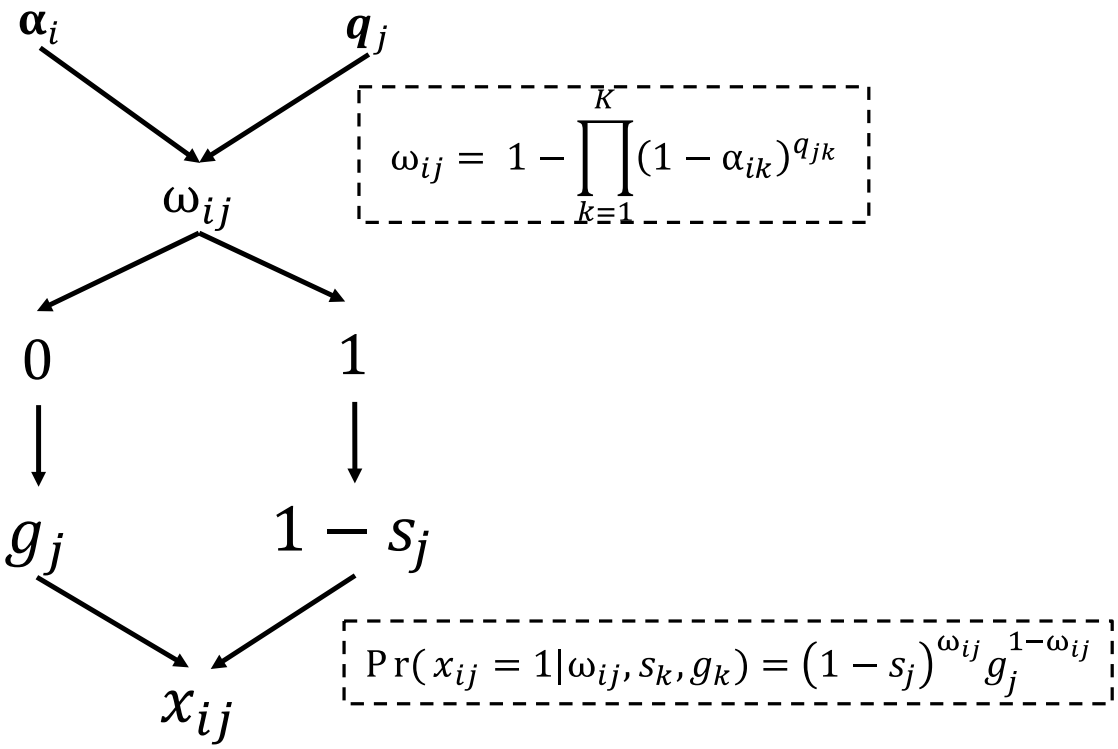


図 2.5 DINO モデルにおける解答反応プロセス

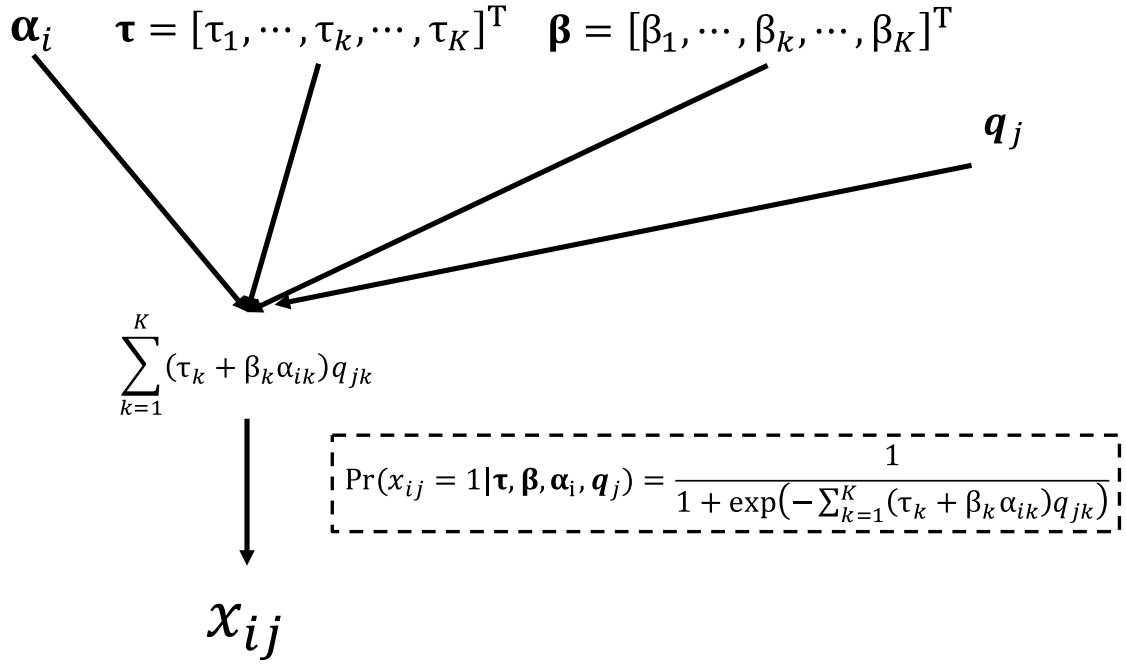


図 2.6 NIDO モデルにおける解答反応プロセス

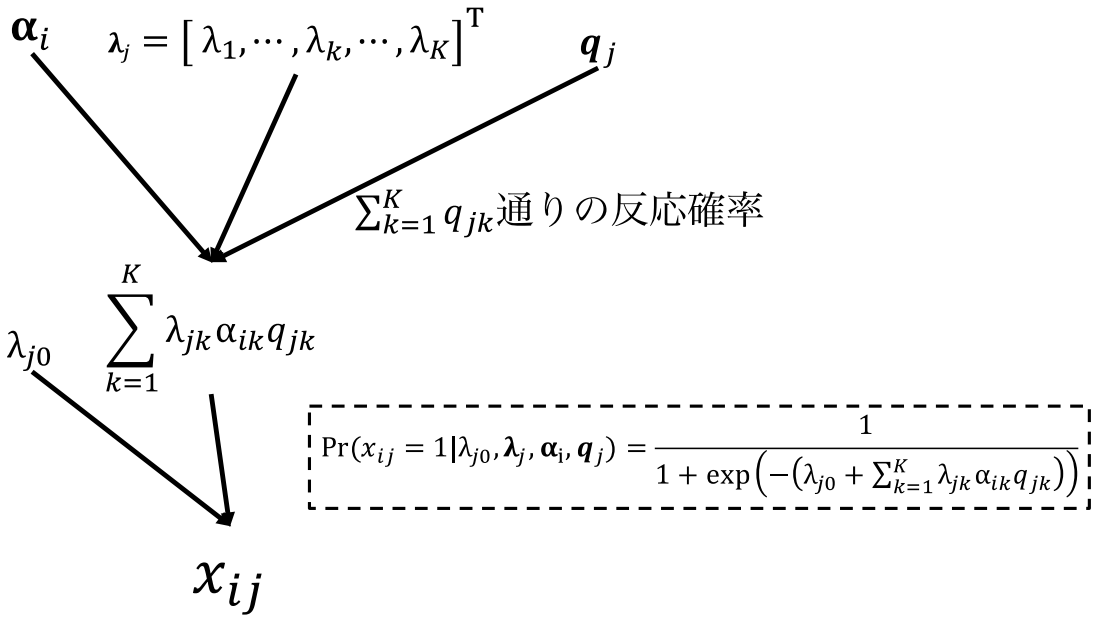


図 2.7 C-RUM における解答反応プロセス

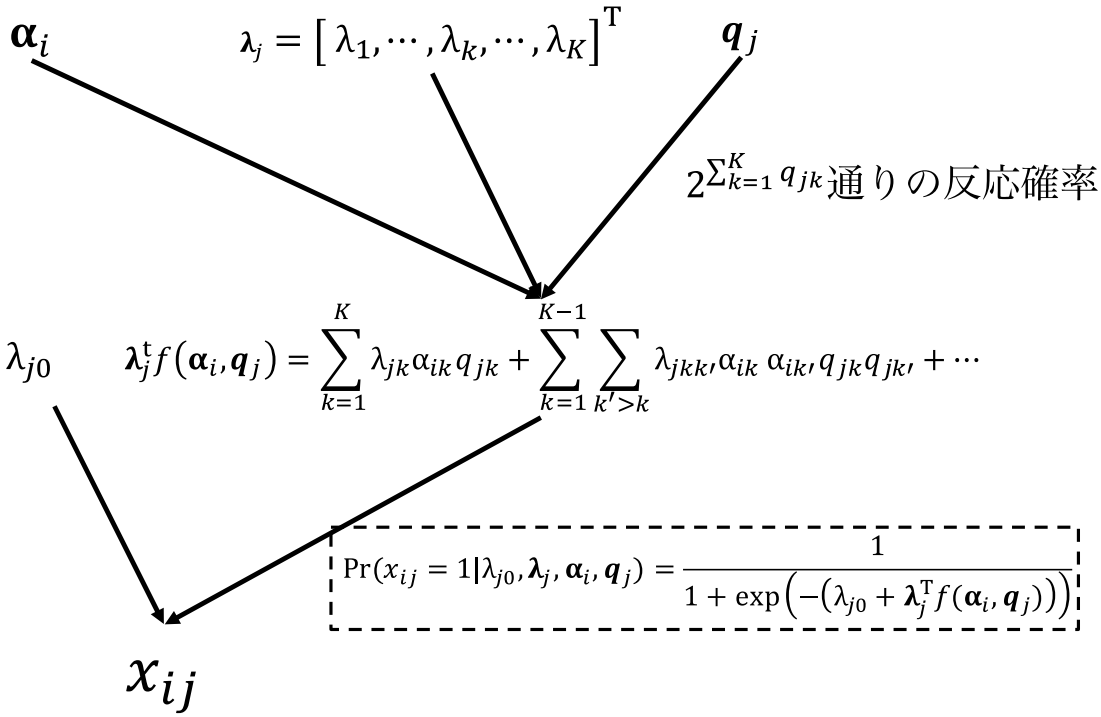


図 2.8 LCDM における解答反応プロセス

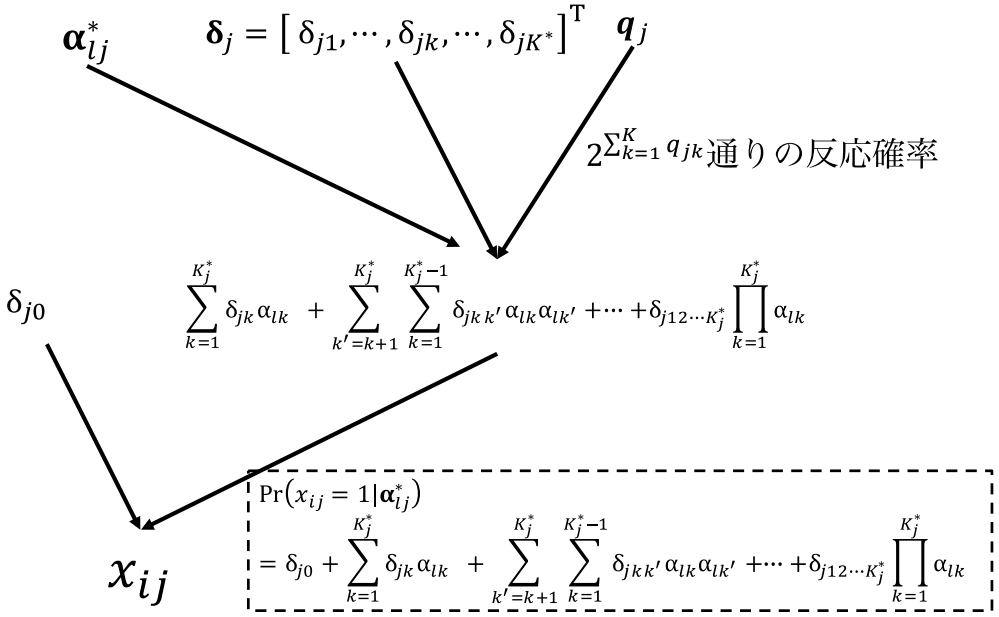


図 2.9 G-DINA モデルにおける解答反応プロセス

表 2.2 アトリビュート数 4 で，3 つの選択枝がある項目jでの$\boldsymbol{q}_{jh}$の例

	アトリビュート			
選択枝	1	2	3	4
1	0	1	0	0
2	0	1	1	0
3	1	1	1	0

第3章　階層構造を考慮した Q 行列の誤設定の影響

第 2 章で述べたように，現実的なテストにみられるアトリビュートの間の従属関係である階層構造（attribute hierarchy structure, AHS）を考慮した場合の Q 行列の誤設定の影響の検討は十分に行われていない。現実のテストではこうした AHS を仮定する状況がしばしば生じるため，この場合の誤設定の影響を検討することは CDM の応用のために必要と考えられる。本章では，より具体的な AHS を考慮した場合の誤設定を同定し，それらの誤設定と AHS の組合せが，アトリビュート習得パタンに与える影響について，シミュレーション実験を通じて検討する。

3-1 背景

3-1-1　　問題と目的

まず，AHS の設定により Q 行列には特定の構造が生じ，誤設定の種類が限定される。例えば，アトリビュート A のあとに B を習得することを仮定した場合に，アトリビュート A，アトリビュート B が必要な項目を，アトリビュート B のみ必要とする誤りは論理的に生じ得ない。Rupp & Templin（2008a）は誤設定の仕方によって解答者のアトリビュート習得パタンの推定の精度が異なることを示している。このことから，AHS に特有の誤設定によって，その影響が異なると予想されるが，その影響の程度は明確ではない。つまり，誤設定の種類と AHS の種類の間にどのような関係があるのかを検討し，どの AHS がどのような誤設定に頑健であるのかを検証する必要があるといえる。これに加えて，誤設定は項目パラメタの推定値に影響することが示されており（Rupp & Templin, 2008a），項目パラメタの推定精度がアトリビュート習得パタンに影響を与える可能性がある。このため，項目パラメタについても，合わせて検証する必要があるといえる。

本章では，この点に関して，Q 行列の誤設定がアトリビュート習得パタンの推定の正確さ（以下，診断精度）に与える影響と項目パラメタの推定精度に与える影響を，複数の AHS の比較実験から明らかにする。Q 行列はテストによって様々な形をとるため，シミュレーションを用いて様々な条件を検討することが適切であると考えられる。また，これまでの誤設定の研究では項目数，サンプルサイズを要因として操作していないものの，これらの要因はテストの構成やテストの実施規模に関わると考えられる。そのため，このような要因を操作し誤設定の影響を緩和できる条件を探索することも目指す。

以下では，アトリビュート階層構造での誤設定について精査する。その後，シミュレー

ションの枠組みを示し，AHS による診断精度の違いと項目パラメタの推定精度についての結果を示す。最後に，シミュレーションの結果に関して考察を行いテスト作成へ示唆を述べる。

3-1-2　アトリビュート階層構造と誤設定

Leighton et al.（2004）はアトリビュートの構造として，“直線（Linear）”，“収束（Convergent）”，“分岐（Divergent）”，“非構造（Unstructured）”を示した。“直線”は，アトリビュートを順番に修得する構造である。“収束”はあるアトリビュートを習得した後に，複数のアトリビュートを修得し，さらにそれらのアトリビュートの習得を前提として最終的には 1 つのアトリビュートを習得する構造である。“分岐”はあるアトリビュートを習得した後に，習得したアトリビュートからさらに複数のアトリビュートを習得する構造である。最後の“非構造”は，1 つのアトリビュートを習得した後に，複数のアトリビュートを並列に複数習得する構造である。しかし，それらの構造は次に示す 3 種類のより基本的な関係から成り立っている。それらの構造を，改めて，直線型，分岐型，収束型と呼ぶこととする（図 3.1）。表 3.1 に 3 種類の AHS に対応した Q 行列を示した。さらに，アトリビュートの階層構造を明確にするために，それぞれのアトリビュートがどのような習得の段階のものであるのかを“階層 1”，“階層 2”などというように呼ぶこととする。以降では，アトリビュートをダブルクオーテーションでくくり，“アトリビュート A”を単に“A”と略記する。直線型は前述の“直線”と同じ構造である（図 3.1 左）。図 3.1 の直線型は，3 つの階層を持ち，階層 1 に“A”，階層 2 に“B”，階層 3 に“C”が配置されている。直線型は表 3.1 の左列に示す Q 行列で表される。つまり，直線型では項目に“B”のみや"C”のみが付与されることはなく，より基本的な“A”も同時に付与される。

図 3.1 中央の分岐型は表 3.1 中央列の Q 行列で表される。これは，“A”を習得した後に，“B”または“C”を習得すると仮定し，1 つのアトリビュートの習得を前提にその後複数のアトリビュートを習得する構造である。ここでの分岐型は，2 つの階層を持ち，階層 1 に“A”，階層 2 に“C”および“B”を持つ構造である。最後に，図 3.1 の右に示した収束型の Q 行列は表 3.1 右列である。収束型では“C”を習得する経路が複数あることを意味しており，“A”を習得したあとに“C”を修得する場合と，“B”を習得したあとに“C”を修得する場合を想定する。収束型は，分岐型と同様に 2 つの階層を持ち，階層 1 に“A”および“B”，階層 2 に“C”を仮定する。分岐型は，例えば，“代数の演算能力”または“グラフを用いた解法”のうちどちらかを習得した後に“一次方程式の解法”を学ぶ状況を表してい

る。

これらの AHS での誤設定を整理する。まず，表 3.3 上段左側の項目 1 や下段左側の項目 2，3 のように 0 である要素を 1 と仮定する条件（過大条件）と，逆に表 3.3 上段右側の項目 2，3 や下段右側の項目 4 のように 1 である要素を 0 と仮定する条件（過小条件）が想定できる。

つぎに AHS を想定した場合に特有の誤設定がある。AHS によりあるアトリビュートを習得するために前提として必要となるアトリビュート数が定まる。同じ階層に位置するアトリビュートを同じ能力水準にあるアトリビュートと呼び，互いに異なった階層にあるアトリビュートを異なった能力水準にあるアトリビュートと呼ぶこととする。例えば，分岐型の構造を考えた時，“B”，“C”を習得するためには最も基本的な“A”のみを習得している必要があり，“B”と“C”は同じ能力水準にあるアトリビュートと考える事ができる。一方で，分岐型での“B”と“A”や“C”と“A”は異なった能力水準のアトリビュートである。アトリビュートについて定義した階層は，問題項目についても同様に考えることができ，ある項目がどの階層に属しているのかを問題に負荷しているアトリビュートをもとに規定できる。例えば，表 3.1 の直線型を仮定した場合，項目 1 は“A”のみを測定するものであり，階層 1 に属する項目ととらえることができる。AHS を仮定した場合に特有の 1 つ目の誤設定は，特定の能力水準全体（今の例では，“B”，“C”の水準全体）の有無に関わる誤設定であり，ある項目において特定の水準がまったく無くなるかどうかを問題としている。つまり，ある項目において，アトリビュートを追加・削除することによって，項目の所属する階層が変化する誤設定が存在する。この誤設定は異なった習得水準のアトリビュートに関わる誤りであるため，これを異水準間誤設定とよぶ。異水準間誤設定は習得順序に関係した誤設定であり，項目の難しさを誤る誤設定とみなすことができる。ここで，特定の項目に必要なアトリビュートの組を{“A”，“B”，“C”}と表すと，例えば，直線型での誤設定をまとめた表 3.2 の左側の項目 2 のように{1, 1, 0}という“A”，“B”の 2 つのアトリビュートが必要な項目に対して，{0, 1, 0}という先に習得する“A”は仮定せず，“B”のみが解答に必要ということは論理的には考え難い。そのため，この項目の誤設定としてありうるのは，{1, 0, 0}と“A”のみが必要と仮定する場合に限定される。具体的な場合では，英語の読解問題などで，“語彙能力”の後に“読解能力”としたとき，作問者が“語彙能力”だけで解けると仮定したものの，実際には文脈に則した意味を答えなければならず，“読解能力”も必要である項目があった場合は異水準間誤設定が生じる。

AHS に特有の 2 つ目の誤設定は，項目の属する階層を変化させない誤設定で，分岐型と収束型の構造において生じるものである。これは，例えば，表 3.3 の下段の項目 2, 3, 4 のように，“B”と“C”に関係した誤設定であり，表 3.3 下段左側の項目 2 のように{1,1,0}という“A”，“B”のみ必要な項目を{1, 1, 1}と“A”，“B”，“C”全てが必要であるとする，あるいは，表 3.3 下段右側の項目 4 の{1, 1, 1}という“A”，“B”，“C”の全てが必要である項目を“A”，“B”のみあるいは“A”，“C”のみが必要であるとする誤設定である。ただし，表中のスペース省略するため，1 つの項目に複数の誤設定がありうる場合は，誤設定をする要素のうちどれか 1 つを選択して，誤設定項目とすることを意味している。このような誤設定があったとしても，項目の階層段階は誤設定がない場合と変わらない。つまり，異水準間誤設定とは異なり，誤設定によってある項目が測定しているアトリビュートの水準は変化せず同じ能力水準の中で能力の質をとらえ違う誤設定である。そのため，この誤設定を同水準内誤設定と呼ぶことする。このような誤設定は分岐型，収束型に特有の誤設定であり，直線型では定義ができない。同水準内誤設定は，数学の“演算記号”の理解を前提とした場合に，本来は“一次方程式の解法”アトリビュートのみが必要である項目に“数の代入の能力”も仮定してしまう場合の誤設定である。表 3.4 に収束型の誤設定状況を示した。

異水準間・同水準内誤設定についてまとめると，どちらも項目の階層段階に注目した誤設定であるものの，誤設定によって項目の階層段階がどのように変化するのかに違いがある。異水準間誤設定は誤設定によって項目の階層が変化する誤設定であり，同水準内誤設定は項目の階層が変化しない誤設定であるといえる。このように AHS により過大・過小条件と異水準間・同水準内誤設定の組み合わせの誤設定を検討可能となった。

3-2 方法

3-2-1 検討要因

シミュレーションで操作する要因をアトリビュート階層構造（AHS），誤設定の種類，項目数，サンプルサイズとした。まず，AHS は直線型，分岐型，収束型の 3 種類を用いた。アトリビュート数はそれぞれの構造に対して 3，4 の 2 つを仮定した。

誤設定は前述したアトリビュートを過大・過小に見積もる場合と，同・異水準誤設定の組み合わせを用いた。誤設定は分析に用いる Q 行列 1 つにつき 1 箇所のみに限定し，最小の誤設定がどの程度大きな影響を持つのかを検討した。

項目数を要因として変化させる際には，追加する項目にはどのようなアトリビュートが必

要であるかによって分析の結果が変化すると予測された。そのため，本章では AHS を表すことができる最小限項目数を項目数操作の 1 単位（基本項目数）とした。具体的には，この基本項目数の 1 倍，2 倍，3 倍の 3 条件を設定した。例えば，直線型の構造の 2 倍の項目条件は表 3.5 に示したようになる。本シミュレーションでは誤設定数は 1 に固定したため，項目数を変化させることで，Q 行列の要素数に対して，誤設定の割合が変化することとなった。つまり，項目数の増加に伴って，誤設定の割合が減少するように設定した。サンプルサイズは Rupp & Templin（2008a）や de la Torre（2008）らの研究を参考に，より小さいサンプルサイズとより大きいサンプルサイズも検討の対象とし, 500, 1000, 3000, 10000 の 4 条件であった。

3-2-2　シミュレーションの手続き

項目反応データを生成するために，DINA モデルの項目パラメタは全ての項目で$s_j = g_j = .2$と一定にした。Rupp & Templin（2008a）は項目パラメタを一様乱数から生成したが，本章では項目パラメタの生成に伴う変動を除去するために，項目によらず真の項目パラメタを一定とした。次に，Q 行列をアトリビュート数と AHS，項目数条件に応じて決定した。最後に，解答者のアトリビュート習得パタンは AHS を固定した場合に，習得パタンが一様分布するように乱数で生成した。

分析に利用する誤設定Q行列は, 前述のようにAHSと誤設定の種類によって決定された。データ発生のために仮定した AHS から真の Q 行列を作成し，真の Q 行列の要素を誤設定の条件に合わせて変更した。このとき，1 つの条件に対して複数の誤設定 Q 行列が仮定される場合があった。例えば，表 4 で示したように，アトリビュートが 3 つで分岐型の構造がある場合，異水準間誤設定・同水準内誤設定のそれぞれについて，過大条件・過小条件のどちらでも 2 つの誤設定 Q 行列が想定された。

作成された項目反応データと誤設定 Q 行列・真の Q 行列を用いて，各解答者のアトリビュート習得パタンを MAP 推定した。さらに，AHS を仮定する場合，論理的に存在しないアトリビュート習得パタンの事前確率を 0 とする制約をかけた推定を行なった。

上記の要因の組み合わせごとに 300 回シミュレーションを行なった。シミュレーションには，統計解析環境 R を用い，パラメタの推定には CDM パッケージ（Gerge et al., 2016）を用いた。Rupp & Templin（2008a）では CDM の推定には時間がかかることが述べられており，本章においても計算時間がかかることが予測された。また，シミュレーション回数を増やし

た場合，シミュレーションに用いたコンピュータがハングアップしてしまったため，プログラムが実行できる範囲でのシミュレーション回数を選択した。しかし，後に分散分析を行う際には十分な反復回数であると考えられた。

3-2-3　評価指標

誤設定が生じた場合の解答者のアトリビュート習得パタンの正確性を評価する指標として，アトリビュート習得パタンの平均一致率を算出するために，解答者のアトリビュート習得パタンの推定値と真値が一致した場合に 1 をとり，それ以外の場合は 0 を取る変数M_{il}を

$$M_{il} = \begin{cases} 1, & \text{if } \hat{\boldsymbol{\alpha}}_{il} = \boldsymbol{\alpha}_{il}, \\ 0, & \text{otherwise} \end{cases} \tag{3.1}$$

と定義した。ここで$l\ (= 1, \cdots, 300)$はシミュレーションの回数である。$\hat{\boldsymbol{\alpha}}_{il}$は推定された解答者$i$のアトリビュート習得パタンを表し，$\boldsymbol{\alpha}_{il}$は解答者$i$の真のアトリビュート習得パタンである。これをそれぞれの誤設定を含む Q 行列ごとに算出した。例えば，アトリビュート数が 3 つの時の分岐型の AHS 図 3.1 中央）について考える。このとき，正しい Q 行列は 1 つしか存在し得ないが，異水準間誤設定の過大条件は表 3.3 アトリビュート数が 3 の分岐型の Q 行列での誤設定の例で示されるように 2 つの Q 行列が想定された。同様に，表 3.3 アトリビュート数が 3 の分岐型の Q 行列での誤設定の例より，異水準間誤設定で過小条件，同水準内誤設定の過大条件，および同水準内誤設定で過小条件も 2 つの場合が考えられた。このように，それぞれの構造のそれぞれの誤設定に対して複数の Q 行列を想定してそれぞれの場合に対して，推定された解答者のアトリビュート習得パタンと真のアトリビュート習得パタンの一致を評価した。1 つの誤設定条件に対してこれらの複数の Q 行列が存在するので，誤設定条件の効果を検討するためにそれぞれの誤設定条件に関して Q 行列の数に関するMの平均値を$\bar{M}_{il}$とした。これは，例えば分岐型の異水準間誤設定の過大条件の中では“B”，“C”のどちらを誤っても同様であるため，平均をとった。そのうえで，シミュレーションの 1 サイクルでの一致割合を

$$M_{\cdot l} = \frac{1}{I}\sum_{i=1}^{I} \bar{M}_{il} \tag{3.2}$$

と計算した。ただし，$i\ (= 1, \cdots, I)$は解答者を意味する。さらに，L回のシミュレーションでの平均一致率とその標準偏差を

$$M_{\cdot\cdot} = \frac{1}{L}\sum_{l=1}^{L} M_{\cdot l}, \tag{3.3}$$

$$SD(M_{\cdot l}) = \sqrt{\frac{1}{L-1}\sum_{l=1}^{L}(M_{\cdot l} - M_{\cdot\cdot})^2} \tag{3.4}$$

とした。

また，1 回のシミュレーションの結果を 1 ケースとみなして，$M_{\cdot l}$を従属変数として分散分析を行なった。要因は AHS が直線型，分岐型，収束型の 3 水準，推定に利用した Q 行列の条件 5 水準（正確な Q 行列，異なった習得水準間誤設定で過大・過小にアトリビュートを仮定，同じ水準間誤設定で過大・過小にアトリビュートを仮定），アトリビュート数条件が 3，4 の 2 水準，項目数が 1 倍，2 倍，3 倍条件の 3 水準，サンプルサイズが 500，1000，3000，10000 の 4 水準の 5 要因とし，結果の解釈のしやすさから，交互作用は 1 次のみを検討した。効果量は要因の平方和であるSS_{f}，当該の要因以外の要因を統制した残差平方和をSS_{e}としたときに，$\eta_p^2 = SS_{\mathrm{f}}/(SS_{\mathrm{f}} + SS_{\mathrm{e}})$ で定義されるη_p^2を用いた。

項目パラメタの評価には，シミュレーションごとに，設定が正しい項目と誤設定がある項目のそれぞれに対して真値と推定値の平均自乗誤差平方根（root mean square error, RMSE）を計算した。例えば，slip パラメタについて，$\hat{s}_l$をシミュレーションlでの推定値，とした場合，

$$RMSE = \sqrt{\frac{1}{L-1}\sum_{l=1}^{L}(\hat{s}_l - .2)^2} \tag{3.5}$$

によって算出した。誤設定が複数ある条件では，$\hat{s}_l$に推定値の平均を用いた。ここで，(3.5) 式の.2 は真値である。guessing パラメタについても同様に算出した。なお，RMSE 誤設定条件ごとに算出し，複数の条件がある場合にはその平均値を評価対象とした。設定が正しい項目は 1 つの条件で複数あるためその平均を用いた。

3-3 結果

3-3-1 誤設定が診断精度に与える影響

アトリビュート数 4 の結果はアトリビュート数 3 の結果と類似していたため，ここではアトリビュート数 3 の結果のみを示す。各条件の診断精度の平均値を図 3.2 から図 3.4 に示

した。これらの図では項目数条件ごとに 3 つのグラフを並列させた。さらに，項目数条件ごとに分析に用いた Q 行列の条件を示した。“正 Q”は正しい Q 行列条件，“異大”は異水準間誤設定の過大条件，“異小”は異水準間誤設定の過小条件，“同大”は同水準内誤設定で過大条件，“同小”は同水準内誤設定で過小条件であった。これらの条件ごとにサンプルサイズ条件での診断精度の値（$M_{..}$）を図示した。また，これらの図のエラーバーは標準偏差（$SD(M_{.l})$）である。

分析の結果，全ての主効果，一次の交互作用項が有意であった。その中でも，特にη_p^2の大きい要因は項目数であった（$F(2,93534) = 1403000, p < .001, \eta_p^2 = .968$）。ついで，$\eta_p^2$ が大きかったのは，Q 行列条件（$F(4, 93534) = 1403000,\ p < .001, \eta_p^2 = .502$），Q 行列条件と項目数の交互作用（$F(8, 93534) = 5033, p < .001,\ \eta_p^2 = .301$），AHS と項目数の交互作用（$F(4, 93534) = 9172, p < .001, \eta_p^2 = .282$）,であった。これら以外の要因の$\eta_p^2$は.1 未満であり，影響の大きさは小さかった。たとえば，サンプルサイズの条件は（$F(3, 93534) = 566, p < .001, \eta_p^2 = .018$）であり，有意ではあるものの，相対的にみて影響の小さい要因であった。

このことから，項目数条件は正答率の変化を最も説明する要因であり，図 3.2 から図 3.4 より，項目数の増加にともなって顕著に診断精度が向上している様子が観察された。加えて，Q 行列条件が有意であり，Q 行列に 1 つでも誤設定があることによって診断精度が低下することが示された。

さらに，AHS と項目数の交互作用が有意であり，AHS の種類によって項目数が増加したときの診断精度の向上のしかたが異なっていることが示唆された。また，Q 行列条件と項目数の交互作用がみられ，項目数によって誤設定の影響が異なっていた。図 3.2 から図 3.4 から，収束型では項目数増加による診断精度の回復しやすく，直線型，分岐型は相対的に回復が遅い様子がみられた。

これらの交互作用の効果を検討するため AHS ごとに同様の分散分析を行なった。直線型においても，全体の分析と同様に項目数の影響が大きく（$F(2, 21568) = 177973.44, p < .001, \eta_p^2 = .943$），ついで，分析に用いた Q 行列の条件（$F(2, 21568) = 26944.97, p < .001, \eta_p^2 = .714$），項目数と Q 行列の条件の交互作（$F(4, 21568) = 2269.76, p < .001, \eta_p^2 = .296$）の 3 条件において，$\eta_p^2$が.2 以上の要因であった。Tukey 法による多重比較の結果，項目数が多くなるほど，診断精度が回復する傾向がみられ，誤設定 Q 行列は正しい Q 行列よりも有意に診断精度が低いことが示された。この傾向は，項目数が増加してもみられた。誤設定の間では，過大条件の方が過小条件よりも診断精度が良好であった（得点差 = .015,

95%CI[.014, .016])。

図 3.3 より，分岐型の異水準間誤設定の過小条件で極端に診断精度が低下した様子がみられた。この条件においてアトリビュート習得パタンの推定結果がどのようになっているかを具体的に調査し，診断精度が低下した一因を明らかにすることを試みた。このために，当該誤設定の過大条件と過小条件それぞれで，真のアトリビュート習得パタンと推定された習得状況の一致率を百分率で算出した（表 3.6, 表 3.7)。表 3.6 はアトリビュートの設定が{1, 0, 0}である項目を{1, 0, 1}と過大に誤設定した場合の結果であった。真のアトリビュート習得パタンが{1,0,0}である解答者が{0, 0, 0}と推定された解答者が 53.13%であり，{0, 0, 0}である解答者が{1, 0, 0}と推定された率も 42.31%になり，誤設定した項目に関連するアトリビュート習得パタンの推定を誤っていた。

表 3.7 は{1,0,1}の項目を{1,0,0}であると過小に誤設定した場合の結果であった。この場合，{1,0,0}，{1,0,1}の習得状況が推定されていない。そのため，表 3.7 の 2 行目，3 行目に見られるように, 真のアトリビュート習得パタンが{1,0,0}の解答者と{1,0,1}の解答者のうち, それぞれ 80.48%と 71.61%が{0, 0, 0}と推定されており，これにより顕著な影響が見られたと考えられる。

収束型の構造では，項目数，Q 行列の条件，それらの交互作用に加えて，アトリビュート数条件でη_p^2が.2 を超え（$F(1,35954) = 11760, p < .001, \eta_p^2 = .247$)，アトリビュート数が 4 よりも 3 で高い推定精度を示した。また，項目数が 1 倍での同水準間の誤設定で過大・過小条件の差は見られなかった。

3-3-2　誤設定が項目パラメタに与える影響

アトリビュート数が 3 つの場合の，直線型，分岐型収束型の項目パラメタの RMSE を図 3.5 から図 3.7 に示した。図の横軸には，診断精度と同様に“正 Q”，“異大”，“異小”，“同大”，“同小”の 5 条件を示し，1 倍から 3 倍の条件を併記した。グラフの各行は上から，誤設定項目の slip パラメタ，guessing パラメタについて，設定が正確な項目な slip パラメタ，guessing パラメタについての結果を意味している。ただし，設定が“正 Q”条件では，誤設定項目が存在しないため，グラフの上二行の“正 Q”条件の結果は空欄とした。

図 3.5 より，直線型では，誤設定がある項目について，過小にアトリビュートを仮定する場合に slip パラメタの RMSE が“正 Q”条件よりも相対的に大きく，過大にアトリビュートを仮定する場合に，guessing パラメタの RMSE が増加する傾向がみられた。これは，

Rupp & Templin（2008a）と同様の傾向であった。この傾向は分岐型，収束型でもみられた。直線型の設定が正確な項目においては，“異小”条件において，slip パラメタの RMSE が，“異大”条件において slip パラメタの RMSE が増加している様子がみられた。

分岐型においては，図 3.6 から，設定が正確な項目においては，ほとんどの場合に RMSE が非常に小さいものであったが，診断精度が顕著に低下していた項目数 1 倍の“異小”条件の guessing パラメタの RMSE が高い傾向がみられた。slip パラメタにおいては，このような結果はみられなかった。誤設定がある項目の slip パラメタでは，“異大”条件の方が“異小”条件よりも RMSE が高かった。誤設定項目 slip パラメタの RMSE は，“異小”，“異大”条件のどちらにおいても同程度であった。

収束型では，図 3.7 に示したように，設定が正確な項目の RMSE は 0 に近いものであった。また，誤設定がある項目で，過小にアトリビュートを仮定する条件においては slip パラメタが，過大にアトリビュートを仮定する条件では，guessing パラメタの RMSE が大きい結果となった。

どの AHS においても，項目パラメタの RMSE については，項目数やサンプルサイズの条件は一貫した傾向があまりみられなかった。つまり，項目数条件や，サンプルサイズ条件よりも，どのような誤設定条件であるかが，RMSE の大きさを左右する要因と考えられる。

3-4 考察

3-4-1 得られた知見

本章では Q 行列の誤設定が診断精度に与える影響を複数の AHS の比較から検討することが第 1 の目的であった。結果から，想定する AHS によって影響を受けやすい誤設定や誤設定による影響の程度が異なることが示された。例えば，分岐型の構造では異水準間誤設定条件で過小に項目を想定する場合に診断精度が最も影響を受け，収束型の構造では同水準間誤設定の過大・過小どちらでも一定程度の診断精度の低下を示した。また，診断精度の低下はみられるものの，過小条件よりも過大条件のほうが高い診断精度を維持していた。第 2 の目的は，項目数，サンプルサイズを要因として操作し誤設定の影響を緩和できる条件を探索することであった。分散分析の結果から，分析する Q 行列に誤設定があったとしても，項目数の増加にともなって診断精度が回復することが示された。一方で，サンプルサイズは診断精度に影響を与えなかった。

項目パラメタについては，アトリビュートを過大に誤設定した場合には誤設定項目の

guessing パラメタの RMSE が増加し，過小に見積もる場合には slip パラメタの RMSE が増加するという先行研究と類似した結果が得られた。さらに，分岐型構造では，異なった水準間でアトリビュートを過小に誤設定した場合に設定が正しい項目の guessing パラメタにも誤設定の影響がみられた。直線型の構造においても，過小にアトリビュートを仮定した場合には，guessing パラメタに，過大にアトリビュートを仮定した場合には，slip パラメタに影響がみられた。

本章で得られた知見は，1. 分岐型で誤設定の影響が最も強くみられ，直線型や収束型では誤設定の影響はあるものの相対的に見て小さい，2. 全体の項目数に対して誤設定の割合が小さくできれば誤設定の影響が緩和できる可能性がある，3. サンプルサイズは誤設定の影響とは関係しない，4. 特定の構造において，誤設定の影響は設定が正しい項目にも影響しうる，という 4 点にまとめられる。

また，AHS はアトリビュートの粒度によっても変化しうるものである。さらに，AHS の設定はテスト項目の性質や背景理論に依存して決まるものである。そのため，当該のテストがどのような AHS を仮定しているのか意識的になる必要がある。場合によっては，本研究で示した誤設定に頑健ではない階層構造であることもありうる。誤設定の影響は AHS によって異なることが示されたため，診断テストを作成する際には，当該の階層構造が誤設定に頑健であるかどうかという観点にも留意しつつ，誤設定を起こさないように注意を払って，Q 行列の設定やテストの内容を決定する必要がある。

3-4-2　誤設定のアトリビュート習得パタンへの影響

誤設定で診断精度が低下した一因は，表 3.7 に見られるように，項目数が 1 倍で誤設定がある場合に特定のアトリビュートが単独で測定されておらず，その結果として，特定のアトリビュート習得パタンの推定ができないことである可能性がある。そのため，設定が正確な項目が追加されたことにより，誤設定の影響を相殺する形で診断精度が回復したと考えられる。

また，誤設定がある場合に異誤設定と同誤設定では異誤設定の方が診断精度に与える影響が大きかった。これは，異誤設定は誤った項目で測定しているアトリビュートの情報を補う項目が十分に無いためと考えられる。一方で，同誤設定では誤設定がある項目以外の項目に誤設定を補う項目があるため，相対的に影響が小さかった可能性がある。

項目に誤設定がある場合には，項目パラメタの推定が正確にならず，アトリビュート習

得パタンの推定に影響がみられた可能性が示唆された。より具体的には，分岐型の構造において，異水準間誤設定を仮定した場合にはアトリビュート習得パタンの推定精度が極端に低下しており，その条件では誤設定項目のみならず，設定が正しい項目の項目パラメタのRMSE も増大していた。また，DINA モデルの推定には EM アルゴリズムが用いられており（de la Torre, 2009b），アトリビュート習得パタンと項目パラメタの 2 種類のパラメタについての計算を交互に行う。そのため，誤設定項目によって，アトリビュート習得パタンの期待値の計算やその期待値にもとづいた項目パラメタの最大化が誤ってしまったと考えられる。正しい設定の項目パラメタにも誤設定の影響がみられた背景としては，このような交互にパラメタを最適化するプロセスによるものである可能性もある。ただし，これについては今後より詳細な検証が必要であろう。

3-4-3　誤設定の項目パラメタへの影響

項目パラメタに関して，関連した話題として重要なのは推定量のバイアスについてであろう。今回は誤設定による項目パラメタのバイアスの程度は検討を行わなかった。しかしながら，誤設定の種類と誤設定によるバイアスの程度を調査することによって，診断精度がどれほど影響を受けるのかということも，今後検討する価値のある事項である。今回の結果からは，過大にアトリビュートを誤設定する場合には guessing パラメタに正のバイアスが，逆に過小にアトリビュートを誤設定する場合には slip パラメタが正のバイアスをうける可能性が予測できる。こういった可能性を検証するとともに，アトリビュートの構造のうち，バイアスを受けやすい構造を明らかにする研究も必要である。

AHS が診断精度に与える影響には項目数との交互作用があり，AHS ごとに項目数による診断精度の回復のしかたは異なっていた。これは，それぞれの AHS ごとに基本項目数やその設定が異なっており，追加される項目が異なっていたことが原因と考えられる。例えば，収束型は追加される項目数が多く，他の設定より相対的に誤設定の割合が小さくなったため，誤設定の影響が希釈されたと考えることができる。この意味で，ありうるアトリビュート習得パタンを限定するほど誤設定の影響が大きくなる可能性が示唆された。

3-4-4　限界と展望

AHS を仮定することによって，診断モデルを利用する有用性が高まることはすでに述べたが，用いた理論的背景が正しくない場合は AHS の指定自体を誤ってしまうことも考えら

れる。これは，例えば，分岐型として“A”，“B”，“C”を仮定したとしても，テストでは“B”，“C”を測定するための項目が不足しており，“B”，“C”を統合した“B’”というアトリビュートにせざるをえないこともありうる。この場合，理論的な構造とテストで測定できているアトリビュートの間には不一致が生じており，テストで測定している構造が“A”と“B’”の直線型となり，構造自体の誤設定が生じたといえる。

また，誤設定があっても新たに追加された項目が正しく，誤設定によるクラス間の違いを区別できるようになれば，診断精度が十分に回復することが示された。しかしながら，本章で追加された項目は付与されるアトリビュートが全て正しいという状況であるため，結果の一般化には注意が必要である。より現実的には，誤設定がある項目に類似した項目にも誤設定が混入する可能性が高い。この場合，追加する項目は結局アトリビュートの付与を誤っており，誤設定の影響を緩和するよりもむしろ影響を増大させることも想定される。たしかに，正しい項目を追加すれば誤設定の影響は小さくできるが，追加する項目が正しい設定であるかを十分吟味する必要がある。

さらに，本章で検証したアトリビュート数は 3 と 4 に限られている点にも注意が必要である。現実的なテストの設定で診断の意義が生じるのはより多くのアトリビュートが必要となる場合であろう。その場合にはより多くの項目が必要となる。多くのアトリビュートが必要となると，アトリビュートの構造の特定に時間を要したり，項目パラメタやアトリビュート習得パタンの推定の計算時間が増大する可能性が指摘できる。加えて，あまりに多くのアトリビュートを想定することによって，計算機の処理の限界を超える可能性も考えられる。このような危険性に加え，現実のテストでは実施時間が限られているため，無制限にテスト項目を増やすことはできない。このような場合には，本章で行ったように，項目数を増やすことによって誤設定の影響を緩和することが難しいと考えられる。このことから，より慎重にアトリビュートと項目の関係を吟味し，予備テストを行うといった入念な準備によって，誤設定が生じる可能性を減少させる必要性が指摘できよう。

サンプルサイズは診断精度に影響を与えなかったが，これは今回仮定した AHS のもとで，解答者がどのアトリビュート習得パタンにも一様に存在することを仮定したためであると考えられる。また，今回はアトリビュート習得パタンの数に比して十分大きなサンプルサイズを仮定したため，サンプリングによるアトリビュート習得パタンの偏りの影響は小さかったと考えられる。

本章の結果から，誤設定に対処するための方法の 1 つは項目数を増加させることである

可能性が示唆された。とくに誤設定が疑われる項目がある場合には，その項目の正答に必要なアトリビュートを測定している項目を増やす努力が必要といえる。しかし，本章では誤設定は 1 箇所に限定されているため，誤設定が多い場合に項目を増やすことによってどれほど診断精度を改善できるのかは未だ不明である。また，アトリビュートが必要でないと仮定するよりも，必要であると仮定したほうが相対的に見て，影響が小さいといえる。さらに，AHS の構造によって，必要な項目数が異なっている可能性がある。このことから，逆に，項目数が一定である場合に，どのような AHS で十分な診断精度が得られるのかを検討可能といえる。

ただし, 本章の結果は全ての項目で slip・guessing パラメタの真値が.2 である。slip・guessing パラメタが大きい場合に本章の結果と同程度の診断精度を得るためには，よりも多くの項目数が必要と考えられる。さまざまな場合が想定されるため，テストを設計する際には，本章のようなシミュレーションを行なうことが望ましいであろう。

今後の検討課題として，そもそも構造の仮定を誤る場合，アトリビュート数の設定を誤る場合など，理論的想定とテストで測定していることの不一致を検討する必要があると考えられる。さらに，より具体的な誤設定への対応手法の開発が挙げられる。例えば，本章の条件でのデータから誤設定を検出する既存の手法（e.g., de la Torre, 2008）の有用性の検討や，データから正しい Q 行列を探索する手法の開発，誤設定に頑健な CDM の開発といった誤診断を低減する手法が望まれる。また，CDM 以外のモデルを用いた場合なども併用することで診断情報が適切であることを保証する方法についても議論が必要である。

本章で示したように，Q 行列の誤設定はアトリビュート習得パタンの推定のみならず項目パラメタの推定にも悪影響を与えるということがわかる。CDM のアプリケーションに際しては誤設定の可能性を考慮しつつ，結果の解釈に留意する必要がある。

また，本章で検討した階層構造がある場合のうち幾つかの条件において Q 行列は識別性を持たないことから注意が必要となる。誤設定によって識別性を持たない Q 行列のなかでも問題の大きい Q 行列を用いてしまったため，アトリビュート習得パタンの推定が極端に悪くなった可能性もある。

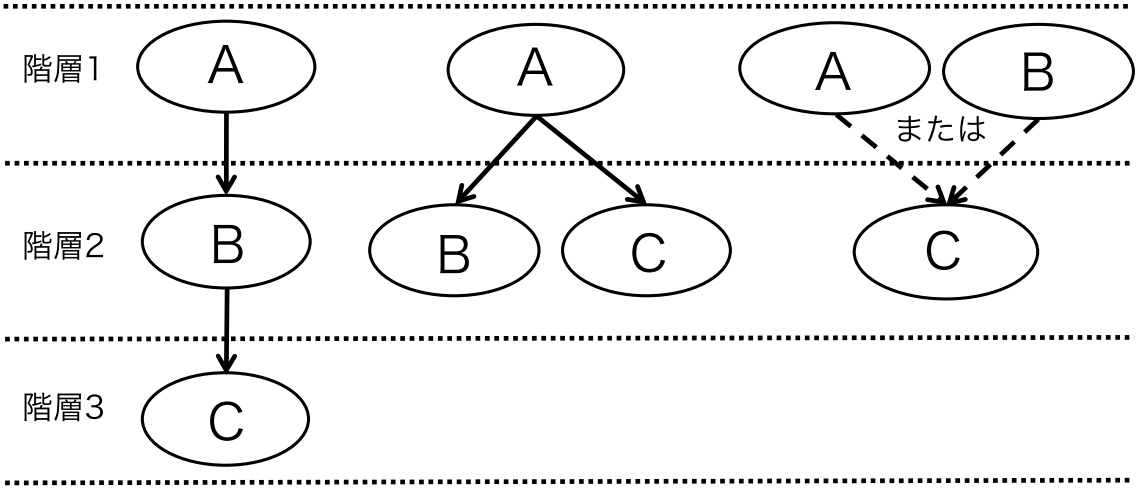


図 3.1 直線型（左），分岐型（中央），収束型（右）のアトリビュート構造例と階層の関係

表 3.1 アトリビュート数が 3 の場合の 3 種類のアトリビュート階層構造

	直線型			分岐型			収束型		
	アトリビュート			アトリビュート			アトリビュート		
項目	A	B	C	A	B	C	A	B	C
1	1	0	0	1	0	0	1	0	0
2	1	1	0	1	1	0	0	1	0
3	1	1	1	1	0	1	1	1	0
4				1	1	1	1	0	1
5							0	1	1
6							1	1	1

表 3.2 アトリビュート数が 3 の直線型の Q 行列での誤設定の例

		アトリビュート					
	項目	A	B	C	A	B	C
異	1	1	**0**→1	**0**→1	1	0	0
水	2	1	1	0	1	**1→0**	0
準							
間	3	1	1	1	1	1	1
		要素を過大に仮定			要素を過小に仮定		

“異水準間”は“異水準間誤設定”を意味する。また，0→1は要素を0から1,1→0は要素を1から0へ変更する誤設定を意味する。ただし,変更は分析対象のQ行列1つにつき，1箇所とした。

表 3.3 アトリビュート数が 3 の分岐型の Q 行列での誤設定の例

		アトリビュート					
	項目	A	B	C	A	B	C
異	1	1	**0**→1	**0**→1	1	0	0
水	2	1	1	0	1	**1→0**	0
準	3	1	0	1	1	0	**1→0**
間	4	1	1	1	1	1	1
同	1	1	0	0	1	0	0
水	2	1	1	**0**→1	1	1	0
準	3	1	**0**→1	1	1	0	1
内	4	1	1	1	1	**1→0**	**1→0**
		要素を過大に仮定			要素を過小に仮定		

“異水準間”は“異水準間誤設定”を意味し，“同水準内”は“同水準内誤設定”を意味する。また，0→1は要素を0から1,1→0は要素を1から0へ変更する誤設定を意味する。ただし，変更は分析対象のQ行列1つにつき，1箇所とした。

表 3.4 アトリビュート数が 3 の収束型の Q 行列での誤設定の例

		アトリビュート					
	項目	A	B	C	A	B	C
異水準間	1	1	0	**0→1**	1	0	0
	2	0	1	**0→1**	0	1	0
	3	1	1	**0→1**	1	1	0
	4	1	0	1	1	0	**1→0**
	5	0	1	1	0	1	**1→0**
	6	1	1	1	1	1	**1→0**
同水準間	1	1	**0→1**	0	1	0	0
	2	**0→1**	1	0	0	1	0
	3	1	1	0	**1→0**	**0→1**	0
	4	1	**0→1**	1	1	0	1
	5	**0→1**	1	1	0	1	1
	6	1	1	1	**1→0**	**1→0**	1
		要素を過大に仮定			要素を過小に仮定		

“異水準間”は“異水準間誤設定”を意味し，“同水準内”は“同水準内誤設定”を意味する。また，0→1は要素を0から1,1→0は要素を1から0へ変更する誤設定を意味する。ただし，変更は分析対象のQ行列1つにつき，1箇所とした。

表 3.5 アトリビュート数が 3 で項目数が 2 倍の直線型の Q 行列

	アトリビュート		
項目	A	B	C
1	1	0	0
2	1	1	0
3	1	1	1
4	1	0	0
5	1	1	0
6	1	1	1

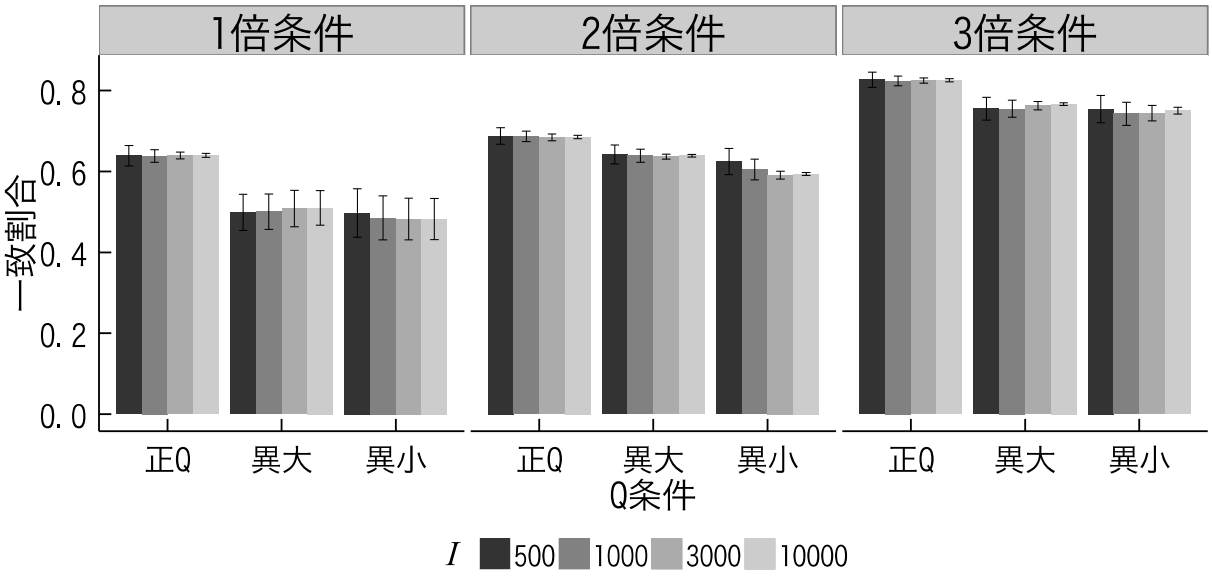


図 3.2 アトリビュート数 3 の直線型の診断精度

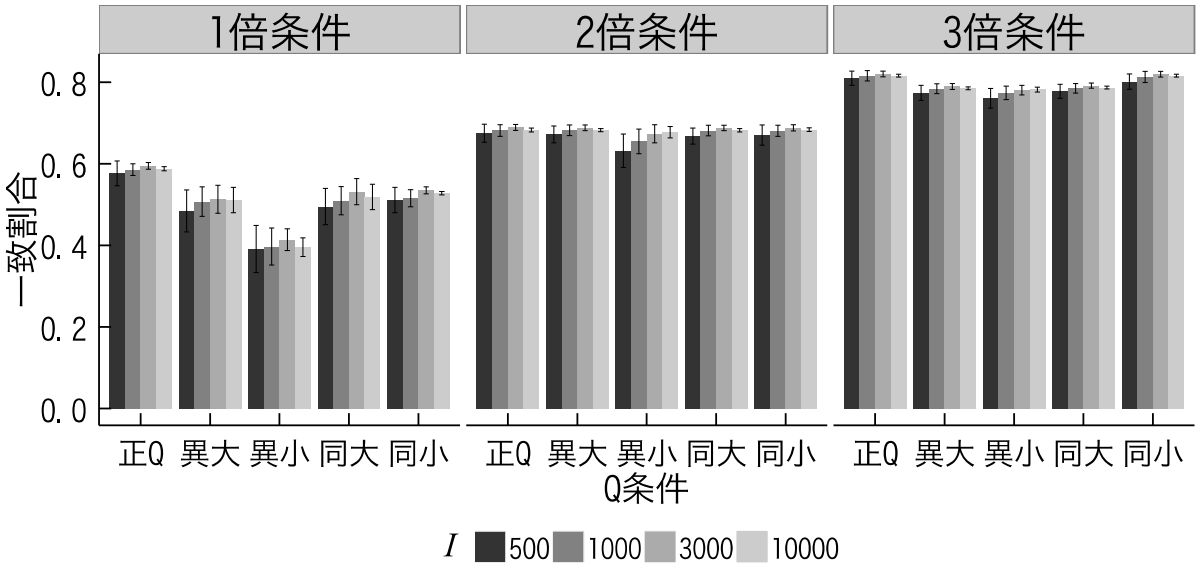


図 3.3 アトリビュート数 3 の分岐型の診断精度

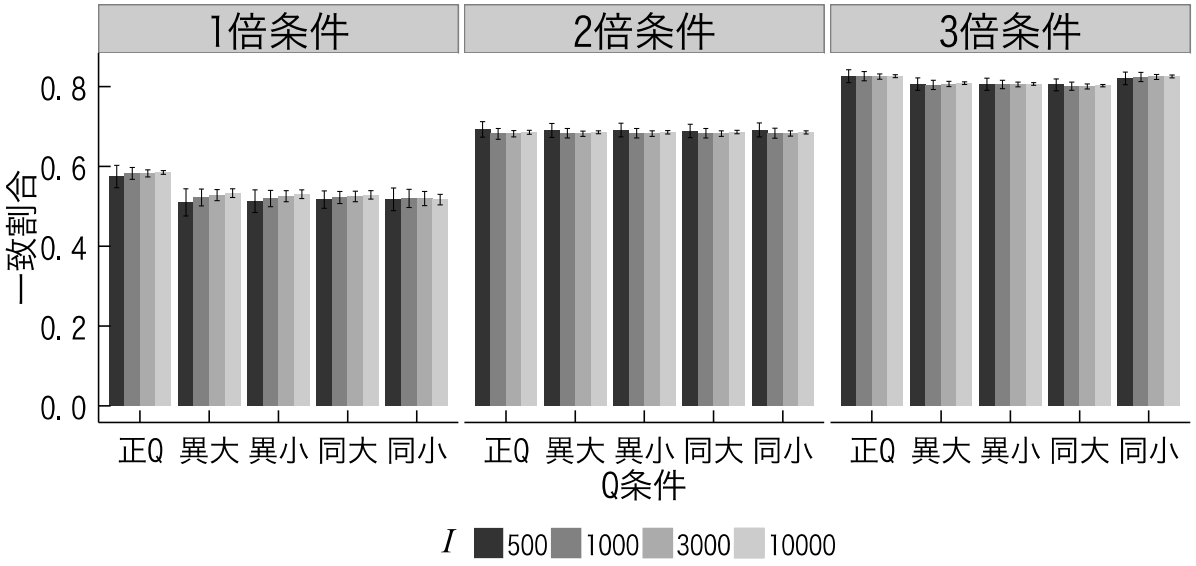


図 3.4 アトリビュート数 3 の収束型の診断精度

表 3.6 {1,0,0}である項目を{1,0,1}に誤設定した場合の真の能力状態と推定された能力状態の一致率（%）

		推定された能力状態					
		{0,0,0}	{1,0,0}	{1,0,1}	{1,1,0}	{1,1,1}	合計
真の能力状態	{0,0,0}	**22.67**	42.31	17.36	12.40	5.26	100
	{1,0,0}	53.18	**10.25**	14.59	13.11	8.87	100
	{1,0,1}	12.83	2.63	**55.45**	3.12	25.97	100
	{1,1,0}	12.92	2.49	6.10	**51.88**	26.60	100
	{1,1,1}	3.65	0.20	7.66	3.09	**85.39**	100

能力状態は{A, B, C}であり，1が習得，0が未習得を意味する。ただし，I = 10000で，項目数が1倍の条件場合の結果である。

表 3.7 {1,0,1}である項目を{1,0,0}に誤設定した場合の真の能力状態と推定された能力状態の一致率（%）

		推定された能力状態					
		{0,0,0}	{1,0,0}	{1,0,1}	{1,1,0}	{1,1,1}	合計
真の能力状態	{0,0,0}	**92.11**	0	0	3.08	4.81	100
	{1,0,0}	80.48	**0**	0	12.96	6.55	100
	{1,0,1}	71.61	0	**0**	13.73	14.67	100
	{1,1,0}	31.64	0	0	**50.81**	17.55	100
	{1,1,1}	12.33	0	0	11.82	**75.85**	100

能力状態は{A, B, C}であり，1が習得，0が未習得を意味する。ただし，I = 10000で，項目数が1倍の条件場合の結果である。

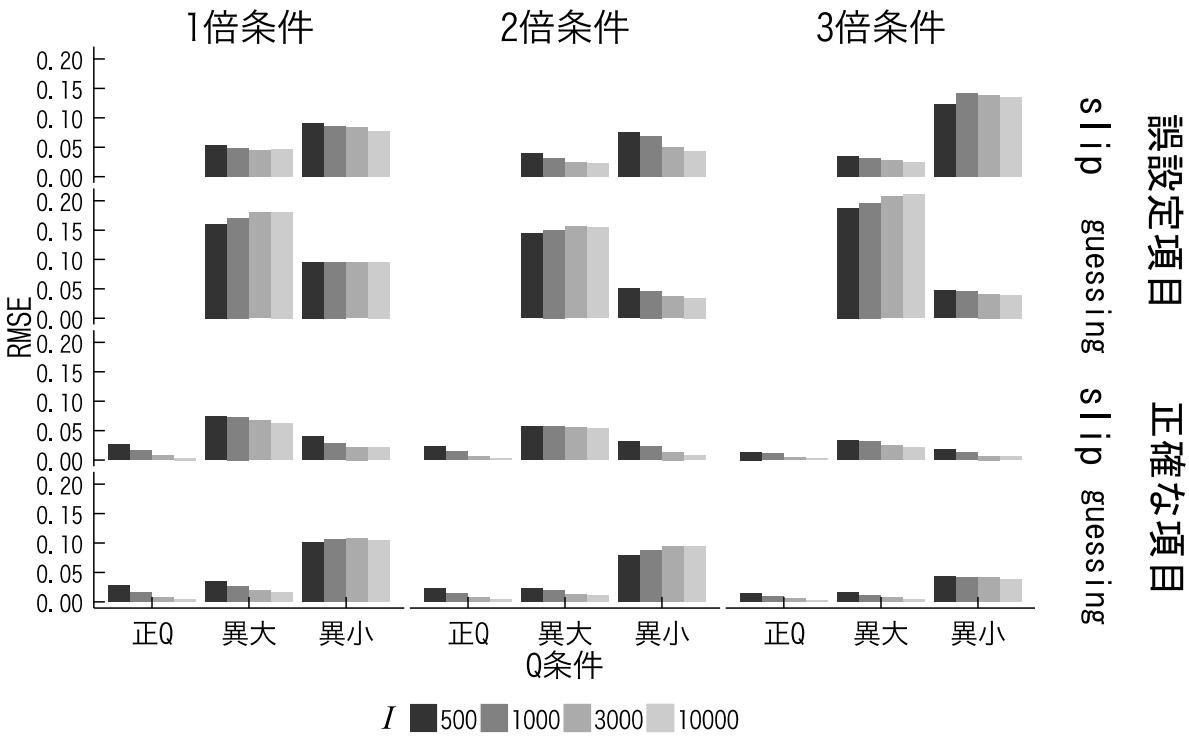


図 3.5 アトリビュート数 3 の直線型の項目パラメタの RMSE

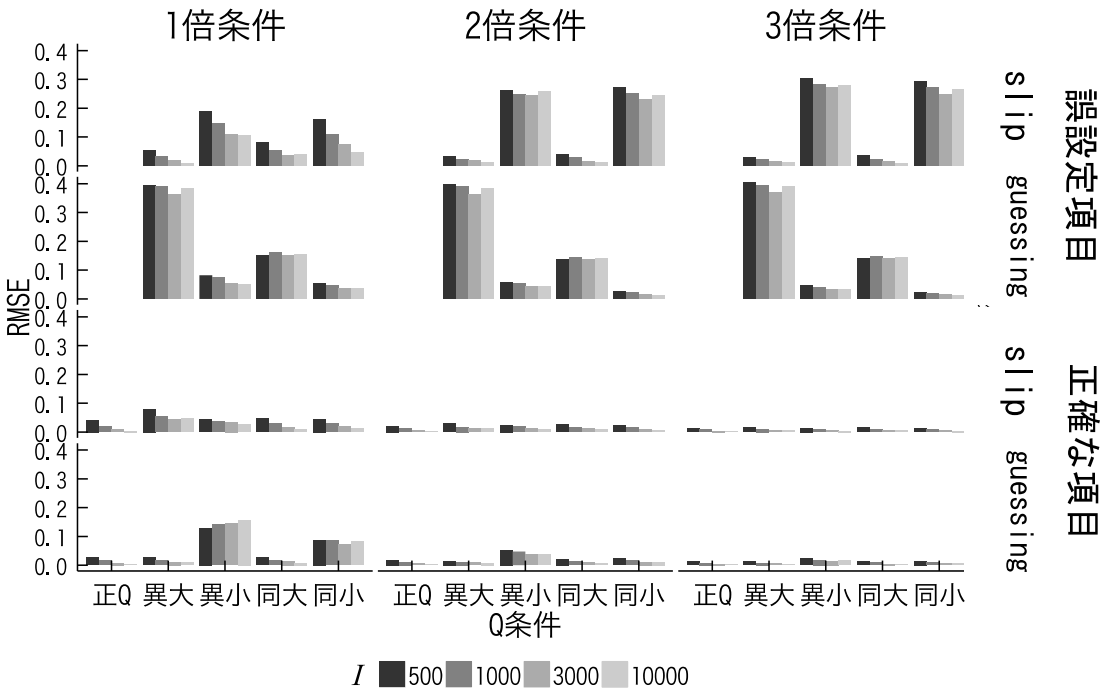


図 3.6 アトリビュート数 3 の分岐型の項目パラメタの RMSE

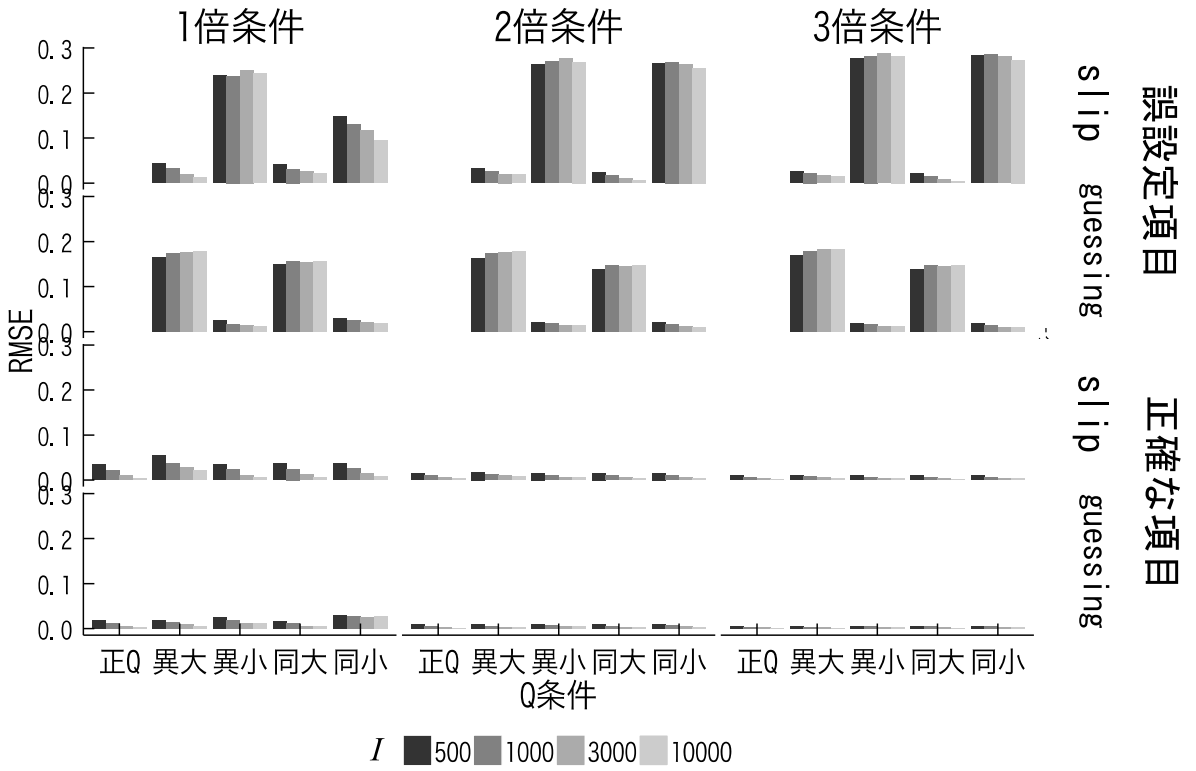


図 3.7 アトリビュート数 3 の収束型の項目パラメタの RMSE

第4章　TIMSS2007 日本人データを用いた認知診断モデルと項目反応理論モデルの比較

第 4 章では，第 2 章でのモデル比較の必要性の議論から，実データを用いた CDM の経験的なモデル比較を行う。実データとしては，国際数学・理科教育動向調査の算数データを用いることで，目的を達成する。

4-1 問題と目的

TIMSS（Trends in International Mathematics and Science Study; 国際数学・理科教育動向調査）は，国際教育到達度評価学会（International Association for the Evaluation of Educational Achievement; IEA）が実施する，小・中学生を対象とした算数・数学と理科の到達度に関する国際調査である。TIMSS は，直近の回では 60 か国，延べ 60 万人以上が受験する非常に大規模な調査である。わが国も 1964 年の第 1 回国際数学教育調査以来，幾多の国際研究調査に参加し, その結果は教育政策や学習指導で活用されている（国立教育政策研究所, 2009）。しかし, このように影響力の大きな調査であるにも関わらず, 日本においてこれまで TIMSS の調査から, 解答者や学校単位で定量的な診断的情報を得て, その後の個々の生徒の学習に活用していくような試みはなされていない。TIMSS で実施される問題項目は，その内容領域と認知的領域の両側面について, 国際的な代表性が担保された質の高い調査である。したがって, その実施結果から生徒や学校を単位とした診断的な情報を得て, 生徒自身や教師にフィードバックすることができれば, 単に調査をして終わりというだけでなく, 調査に参加する生徒や学校にとっても有益であると考えられる。特に算数・数学テストは計算手続きや解答スキルをアトリビュートとして定義しやすく, CDM の適用に適している。例えば, CDM の適用例として最も有名なのは Tatsuoka（1983）の分数の計算テストである。一方，理科テストは知識を問う問題が多く，項目間で共通のアトリビュートを想定しにくいと考えられる。これは社会のテストでも同様で，知識を中心に問うようなテストに CDM を適用した例はおそらく存在しない。また，CDM の PISA のリーディング英語テストへの適用は Chen & de la Torre（2014）にみられるものの，TIMSS に比しての適用例は多くない。こうしたことから，TIMSS への CDM の適用は教育実践への示唆も踏まえて適切であると考えられる。

生徒や教師への診断情報のフィードバックを可能にするためには, まず TIMSS の調査における項目反応がどのような CDM によってよく表現できるかを調べなければならない。このために，既存のさまざまな CDM のモデル間で当てはまりのよさを比較し，その知見の蓄

積が必要である。

TIMSS のデータを CDM で扱う先行研究としては，たとえば，Tatsuoka, Corter, & Tatsuoka（2004）は TIMSS の 1999 年版のデータを用いて，RSM によって 20 の国でアトリビュート習得パタンを比較し，国によって習得パタンが異なっていることを示した。また，Choi, Lee, & Park（2015）が 2003 年版の中学校 2 年生のデータを用いて DINA モデルによってアメリカと韓国の比較を行っており，やはり国によって習得パタンの違いがあることを示した。しかしながら，これらの研究の主眼は国際比較であり，CDM 間，もしくは CDM と他のモデルの間での，モデル適合についての経験的な比較や議論はされていない。

TIMSS のデータを分析した先行研究の中で，CDM と他のモデルとの比較を行っているのは，Lee, Park, & Taylan（2011）の研究のみである。Lee らは 2007 年版の TIMSS（TIMSS 2007）の小学校 4 年生の算数の問題の一部を使って，CDM の 1 つである DINA モデルと，1,2,3 パラメタロジスティックモデル（以下，1-3PL モデル）との比較を行った。その結果，アメリカ（全体），ミネソタ州，マサチューセッツ州のデータでは情報量規準の観点から DINA モデルの適合がよいことを示した。しかしこの研究でも，現在提案されている複数の CDM 間でのデータに基づく比較検討が行われているわけではない。

以上のような現状を踏まえ，第 4 章では TIMSS の実データを用いて，経験的に複数の CDM 間の当てはまりのよさを比較する。本章の目的は以下の 3 点である:（1）日本の TIMSS2007 のデータを用いて，アメリカのデータでは IRT モデルより CDM の当てはまりがよいという，Lee et al.（2011）の知見の再現性を検証する（2）先行研究で提案されてきた複数の CDM を同データに当てはめ，当てはまりのよさを経験的に検討するとともに，相対的に適合したモデルにおけるアトリビュート習得パタンを用いてアメリカのデータとの相違を明らかにする（3）なぜ同データにおいて，そのモデルが適合したのか，その理由を他のモデルと比較しながら議論・考察する。

本章の構成は次の通りである。4-2 では，先行研究を踏まえてこれらのモデルを比較するための方法について述べる。4-3 で結果を示し，4-4 でこの結果を踏まえてモデル適合についての考察と議論を行う。

G-DINA モデルとその各種下位モデルはフリーの統計解析環境 R の G-DINA パッケージ（Ma & de la Torre, 2017）によって解析可能であり，本研究では G-DINA パッケージを用いて各種 CDM の項目パラメタやアトリビュート習得パタンを推定した。

4-2 方法

本節では，分析に用いたデータについて設定，アトリビュートおよびQ行列の詳細，問題項目の具体的な設定，解析の設定について述べる。

4-2-1 データ

TIMSS2007の4年生の算数データの中から，Lee et al.（2011）で用いられたものと同様に，問題冊子4，5のいずれかに解答した解答者のデータを用いた。調査対象の詳細や標本の抽出方法については国立教育政策研究所（2009）に記載されている。

データはTIMSS2007のインターナショナルデータベースから取得し，国を示す識別コードに従って日本人サンプルデータを選択した。さらに，問題冊子を示すコードに従って当該のデータを抽出した。分析対象とした項目に1問も解答していない解答者1人を除いた最終的な日本人サンプルは639人（女性 =323人，男性 =316人）であった。各問題冊子に含まれる項目が異なっており，デザインによる欠測が生じているため，Missing at random（MAR）が成立しているとみなして完全情報最尤推定で分析を行った。

4-2-2 アトリビュートおよびQ行列

アトリビュートはLee et al.（2011）のTable 2に示されている15個を利用した。アトリビュート数について，TIMSSにCDMを適用した研究であるTatsuoka et al.（2004）では21個，Choi et al.（2015）では12個のアトリビュートを用いており，Lee et al.（2011）で用いた15個という数は中間的な数であった。TIMSS 2007の算数テストは複数のトピックを持つ3つの内容領域と38の目的がある。Lee et al.（2011）は，3人の数学教育の学位を持つ研究者と 2 人の内容領域の専門家と共に，それぞれのトピックに対して当該テストで必要な初期のアトリビュートを決定するための協議を行っている。その後，Lee et al.（2011）では調査目的に合致するようにアトリビュートを修正，最終的に15個のアトリビュートを得た。こうした手続を経ているため，本研究ではアトリビュートの内容について一定の妥当性があると判断して，Lee et al.（2011）で定義されたアトリビュートとQ行列を利用した。

15個のアトリビュートのうち，「数」の内容領域についてのものが8つ（「数1」～「数8」），「図形と測定」の内容領域についてのものが4つ（「図形と測定9」～「図形と測定12」），「資料の表現」の内容領域についてのものが3つ（「資料の表現 13」～「資料の表現 15」）であった。以下の説明はLee et al.（2011）のTable 2の記述に準拠したものである。

3 つの内容領域について，各々さらに細分化してアトリビュートが付与された。「数」は「整数」（Whole Numbers）に 4 つ，「分数と小数」（Fractions and Decimals）に 2 つ，「整数の数表現」（Number Sentences with Whole Numbers）に 1 つ，「パタンと関係」（Patterns and Relationships）に 1 つのアトリビュートが与えられた。

「整数」のアトリビュートは「1. 位置の値の知識を実行するだけでなく，整数の表現，比較，順序づけすること」，「2. 四つの演算子を用いた整数の計算と倍数の認識および計算の推論」，「3. 実生活の文脈での問題解決（例 : 測定，お金の問題）を含む問題解決」，「4. 比率を含む問題解決」，であった。

「分数と小数」のアトリビュートは，「5. 全体の一部としての小数と分数の同値性の理解・表現・認識」，「6. 加法と減法を含む単純な分数と小数の問題解決」，であった。「整数の数表現」のアトリビュートは「7. 数の文章や表現での未知を含む欠損した数や単純な状況のモデルと演算を見つけること」で，「パタンと関係」のアトリビュートは「8. パタンとその拡張の関係を記述すること（例 : 所与の規則のもとで整数の組を生成する，所与の整数の組からルールを同定する)」であった。

「図形」の内容領域では，「直線と角」（Lines and Angles）に 1 つ，「2 次元・3 次元図形」（Two- and Three-dimensional shapes）に 2 つ，「位置と運動」（Location and Movement）に 1 つのアトリビュートが与えられた。「直線と角」のアトリビュートは「9. 数と直線を描くためのそれらの性質の理解と測定および推測」であった。「2 次元・3 次元図形」のアトリビュートは「10. 幾何図形と形状とその関係と初等的な性質の認識・比較・分類」，「11. 周囲の長さと領域と体積の計算と推論」であった。「位置と運動」のアトリビュートは「12. 図形の描画と認識およびその動きを認識するためのインフォーマルな座標での場所の点の認識」であった。

最後の「資料の表現」では，「読み取りと解釈」（Reading and Interpreting）に 2 つ，「表現」（Organizing and Representing）に 1 つのアトリビュートが与えられた。「読み取りと解釈」のアトリビュートは「13. 表・統計図表・棒グラフ・円グラフからのデータの読み取り」，「14. データからの情報の使い方の理解と計算」で，「表現」のアトリビュートは「15. 表・統計図表・棒グラフをも用いたデータの整理と異なった表現の理解」であった。

Q 行列も Lee et al.（2011）の Table 3 に示されているものを利用した（表 4.1)。それぞれのアトリビュートが必要な項目数は最小で 2 つ（例，「数 7」)，最大で 16 個（「数 2」）であった。また，1 つの項目に必要なアトリビュート数は最小で 1 個（例，M041056)，最大 6 個

（M041336）であった。

4-2-3　問題項目

問題項目は Lee et al.（2011）に従って抽出した。具体的には，「数」（Number），「図形と測定」（Geometric Shapes & Measurement），「資料の表現」（Data & Display）という 3 つの内容領域から問題を抽出した。この内容領域名は国立教育政策研究所（2009）の調査報告書に従ったものである。採点について，多枝選択の問題 15 問はコードブックに従い，正答選択枝を選んでいれば 1，そうでなければ 0 とした。また，部分点を含む項目については，満点を 1 とし，部分点は誤答（= 0）とした。

付録に実際の問題項目を全て示した。ここではそれらの項目を具体的に示しながら，付与されたアトリビュートとの関係を簡単に記述する。ここでは出題された問題順ではなく数領域，図形と測定領域，資料の表現領域のそれぞれを中心に測定している項目をまとめて記述する。

まず，主に数領域のアトリビュートのみが必要な項目を記述する。これらの問題としては分数の計算や整数の性質を理解しているか，また実生活の中での立式や状況を式で表現することなどが含まれている。数量域を測定する問題は他の 2 つの領域に比べて，問題数が比較的多いという特徴がある。次に具体的な問題と解答に必要なアトリビュートを示す。項目 1（M041052）は「3 つの一 ＋2 つの十 ＋4 つの百に等しい数はどれか」という問題である。アトリビュートとしては「数 1」と「数 2」のアトリビュートが問題の正答に必要と仮定されており，整数の知識や四則演算の計算ができることが問題正答に必要とされた。項目 2 は 12 個のクッキーの絵が与えられ，1/3 のクッキーを含むように丸を描くことが求められる問題であった。この問題は「数 5」のアトリビュートが必要であり，全体の一部としての小数が理解できれば正答できると仮定される問題である。項目 3（M041069）は 2/3 に等しい分数を選択する問題で，「数 2」，「数 4」および「数 5」のアトリビュートが必要と仮定されている。これらのアトリビュートは基本的な計算と比率を含む問題解決および，単純な分数の問題解決を表している。項目 4（M041076）は「ジョーは 3/10 のお金をペンに，5/10 のお金を本に使った。彼が使ったお金を分数で表すとどうなるか」という問題である。アトリビュートとしては，「数 3」と「数 6」が必要であり，実生活上の文脈での問題解決を含む単純な分数の問題と考えることができる。項目 5（M041281）は「レインは 32 本の鉛筆と 4 つの鉛筆用の箱を持っている。彼は同じ本数の鉛筆をそれぞれの箱に入れた。彼がそれぞれの

箱に何本の鉛筆を入れたかを意味する式を選択しなさい。」という問題であり，「数 2」，「数 3」および「数 8」の 3 つのアトリビュートが必要な問題である。これらのアトリビュートが付与されていることから，この問題には実生活の文脈で倍数を認識し，パタンと関係を記述する能力が必要という解釈を行うことができる。項目 15（M031303）と項目 16（M031309）は計算中心の問題であり，文章から計算式を立式したり，文章の条件から計算をする問題であった。この問題に必要なアトリビュートとしては，「数 2」，「数 3」が必要であり実生活での計算ができるかどうかを問う問題と考えることができる。項目 17（M031245）は割り算の等式の一部が空欄になっており，その空欄を解答する問題であり，「数 1」，「数 7」のアトリビュートが必要な問題であった。これはほぼアトリビュート「数 7」の定義に沿った内容を問う問題項目であったといえる。項目 23（M031173）は項目 15（M031303）と項目 16（M031309）と同様のアトリビュートが必要な計算問題であり，実生活場面で必要な計算問題であった。

次に図形と測定のアトリビュートを中心に問う問題について記述する。図形と測定のアトリビュートには単に図形の性質を知っていれば解答できる問題と多少の計算が必要な問題が存在し，数領域と同様に比較的問題数が多い。項目 6（M041164）はいくつかの図形（絵）が提示され，点線の部分で対称（symmetry）となっているものを選択する問題であった。この問題では「図形と測定 10」，「図形と測定 12」のアトリビュートが必要という問題で，対称という初等的な図形の性質を理解しているか，またグラフではない図形を認識できるかどうかが問われている問題と考えられる。項目 7（M041146）は座標に長方形の 2 辺のみが描かれており残りの 2 辺を記入する問題であった。アトリビュートとしては，「図形と測定 9，10，12」のアトリビュートが必要とされた。ここから，直線を描くための知識や長方形という図形の性質の理解，座標の認識などが求められると考えられる。項目 8（M041152）は「パトリックはフェンスの片面にペンキを塗っている。フェンスは 4 メートルの長さと 3 メートルの高さである。パトリックはどの広さを塗らなければならないか。」という問題でフェンスの図も提示されている。「数 1，2，3」および「図形と測定 10，11」の合計 5 のアトリビュートが必要な問題である。実生活でありうる状況設定で，必要な面積を求める問題であり，比較的多くのアトリビュートが必要であると想定されている。項目 9（M041258A），項目 10（M041258B）は 2 つの三角形が提示され，それらが同じであることを示す方法と異なっていることを示す方法を 1 つずつ記述することが求められる問題であった。項目 9（M041258A）はアトリビュートとして「図形と測定 10」のみが必要である項目であり，図

形の性質や比較の方法を理解していれば解答できる問題と仮定されている。項目 10（M041258B）は必要なアトリビュートとして,「図形と測定 9，10」の 2 つが必要である問題であり，項目 10（M041258A）とは異なった側面を測定していると考えられる。項目 11（M041131）は「数 2，3，4」および「図形と測定 9」が必要な項目であり，人間と木の図および人間の身長が与えられたもとで, 木の高さを推定する問題である。項目 22（M031219）は長方形の性質を問う問題であり,「図形と測定 10，11，12」のアトリビュートのみが必要とされる珍しい問題であった。この問題では長方形の紙を細長い 2 つの紙片に分割し L 字に形を変えた図形の面積が元の長方形と同じか否かを尋ねる問題であり，四則演算などの計算は必要でない問題であった。項目 24（M031085）は「図形と測定 10」のみが必要の項目であり，正三角形の性質が理解できていれば正答できる問題であった。

最後の資料の表現に関連した項目は，図形と測定のアトリビュートは全く含まないが全ての項目で数領域のアトリビュート項目が必要であった。また, 他の 2 つの領域に比して資料の表現領域の問題数は少ない傾向が見られる。項目 12（M041275）は表の形式で与えられたデータの頻度を棒グラフに変換する問題であった。この問題では「数 1」および「資料の表現 13，15」が想定されているアトリビュートである。資料の表現のアトリビュートが求められており，表のデータを読みとり，別表現を行うのみならず，数字体の知識も求められる問題となっている。項目 13（M041186）はジョンが 4 日間で収穫したりんごの個数が表にまとめられており, 特定の個数を収穫した曜日を答える問題であった。アトリビュートとしては,「数 1，2，4」および「資料の表現 13」が必要とされている。問題の正答には表に表されているりんごのシンボルの個数を数値に変換し，半分のりんごのシンボルの個数を計算する必要があり，単純な表の読み取りだけでなく割り算の実行が必要な問題となっている。項目 14（M041336）はクラスの人数とともに，クラス A の男女の比率が円グラフで，クラス B の男女の人数棒グラフで与えられており，クラス A に所属する女子がクラス B よりも何人多いのかを問う問題であった。アトリビュートとしては,「数 1，2，5，6」および「資料の活用 13，14」で合計 6 つ必要とされる問題であった。小数や分数の理解のみならず, 表現が異なった 2 つのグラフから人数を算出することも求められており, 数多くのアトリビュートが必要な問題となっている。項目 18（M031242A），項目 19（M031242B），項目 20（M031242C）は二種類の自転車のレンタルの初期料金と時間あたりの料金が与えられており, 与えられた情報から表の情報を埋めたり, 何時間レンタルすると 2 つの料金が一致するのかを解答したり，特定の時間レンタルした場合にどちらのレンタル条件のほうが

費用を押さえることができるのかを解答する問題であった。これらの問題では「数 2，3，8」や「資料の表現 14」のアトリビュートが必要であり，与えられた資料の情報を読み取り，実際に自転車を借りた時間に応じた料金を計算したり，その料金の変化パタンを理解することが求められる問題と考えられる。項目 21（M031247）は文章題であり，子どもと大人で料金が異なるチケットの合計料金を求めるものである。この問題でもこれまでに見てきた問題の多くと同様に，「数 2，3」のアトリビュートが必要であり，さらに「数 7」のアトリビュートも必要とされている。「数 7」のアトリビュートは日常の状況をモデル化することを表したものであり，項目 21 の子供の人数やチケットの値段を適切に反映した立式が求められた。項目 25（M031172）は家のシンボルと地域が書かれた表を完成させる問題であり，項目 13（M041186）と類似した問題であった。アトリビュートとしては，「数 1，2」および「資料の活用 13，15」が必要な問題であり，提示された条件に合致する個数の家のシンボルを解答するものであった。

4-2-4　解析の設定

項目の基礎統計量として，正答率（および，95%CI），標準偏差，IT 相関（当該の項目を含まない合計点を利用し，95%CI も算出）を psych パッケージ ver.1.7.5（Revelle, 2017）の alpha 関数を用いて算出した。IT 相関の 95%信頼区間の算出には非心 *t* 分布にもとづく点双列相関係数の信頼区間の構成方法（南風原, 2014）を用いた。非心度の算出には R の MBESS パッケージ ver.4.3.0（Kelley, 2017）を利用した。さらに，全てのテスト項目を用いてテストとしての信頼性係数αとその 95%信頼区間も算出した。モデル比較では，G-DINA，DINA，DINO，A-CDM，LLM，R-RUM，1-3PL の各種モデルで推定を行った。CDM の推定には R の G-DINA パッケージ（Ma & de la Torre, 2016; ver. 0.13.0）を用い，1-3PL モデルの項目パラメタの推定には ltm パッケージ（Rizopoulos, 2006）の tpm 関数を用いた。1PL モデルを推定する際には，全ての項目の識別力パラメタを 1，当て推量パラメタを 0 とした。CDM を推定する際には G-DINA 関数を利用した。推定の際には，初期値に関数引数として nstart を 500 に設定し，最大イテレーション回数は 3000 回，収束基準を.0005 に設定した。項目パラメタの推定には周辺最尤推定を用い，アトリビュートパタンの分布には関数の既定である一様分布を仮定した。アトリビュートの習得パタンの算出には EAP 推定値（期待事後推定値，Expected A Posteriori estimate）を用いて，それぞれのアトリビュートの習得確率が.5 を超えるものを習得（=1），そうでなければ未習得（=0）とした。IRT モデルの項目パラメタ

の推定には，周辺最尤推定を用いた。また，IRT モデルの潜在特性の推定には，周辺最尤推定された項目パラメタを用いて，事前分布を標準正規分布とした経験ベイズ推定による MAP 推定値（事後確率最大化推定値, Maximum A Posteriori estimate）を用いた。

モデル比較の指標として, −2 対数尤度（デビアンス）, AIC, BIC, 平均絶対相関偏差（mean absolute deviation correlation, MADcor; DiBello, et al. 2006），および標準化平均平方残差平方根（standardized root mean square residual, SRMSR; Maydeu-Olivares, 2013）を用いた。AIC は−2 対数尤度+2 項目パラメタ数，BIC は−2 対数尤度+サンプルサイズ log(項目パラメタ数)によって算出した。情報量規準である AIC，BIC はモデル適合の相対指標であり，それぞれのモデルの間で値を比較し，より小さい値であるモデルが適していると判断する。

一方，MADcor と SRMSR は絶対適合指標であり，そのモデルが当該のデータにどれだけ適合しているかを示すものであり，値が 0 に近いほど適合していることを意味する。ただし，これらの絶対指標については，CDM の各モデルよりも IRT モデルにおいて良い値を示すことが期待される。なぜならば，IRT モデルは単一の連続的な潜在変数のみを仮定している一方，CDM はアトリビュートの習得・未習得を表す離散的な潜在変数であるアトリビュートを想定しているからである。また，CDM において絶対指標はパラメタ数が少ないモデルよりも多いモデルで良い適合を示すことが期待される。そのため，本研究においては AIC と BIC を主要なモデル比較の指標とし，MADcor と SRMSR は参考程度に記述するに留める。

MADcor は

$$\text{MADcor} = \frac{2}{J(J-1)} \sum_{j'<j} |r_{j'j} - \hat{r}_{j'j}| \tag{4.1}$$

と定義される。ここで，$r_{j'j}$は項目j'とjのサンプルでの観測相関係数であり，$\hat{r}_{j'j}$は期待相関係数である。それゆえ，MADcor は$\hat{r}_{j'j}$と$r_{j'j}$の差の絶対値をすべての項目の組み合わせについて平均した指標である。同様に，SRMSR は

$$\text{SRMSR} = \sqrt{\frac{2}{J(J-1)} \sum_{j'<j} \left(r_{j'j} - \hat{r}_{j'j}\right)^2} \tag{4.2}$$

と定義される。SRMSR は，期待相関と観測相関係数の差の二乗をすべての項目の組み合わせについて平均したものの平方根を取った指標である。これらの定義式は，CDM パッケージ（George et al., 2016）のヘルプファイルに記述されている。IRT モデルにおける絶対適合指標の算出には, tpm 関数で推定した項目パラメタをもとにして, TAM パッケージ（Robitzsch,

Kiefer, & Wu, 2017）の tam.modelfit 関数を用いた。

また，アトリビュートの習得確率を従属変数，2PL モデルで推定した潜在特性を独立変数にしたロジスティック回帰分析を行った。ロジスティック回帰分析は R の glm 関数を利用した。

4-3 結果

本節では，基礎的な項目の統計量を示し，そもそものテスト項目の性質を確認し，その後にモデル比較を実行する。さらに，アトリビュート習得パタンについて，複数のモデルを用いて推定した結果を示し，1 次元潜在特性とアトリビュート習得確率の関連を分析する。最後に，項目パラメタの推定値についての結果を提示し，CDM からみた TIMSS2007 の項目の特性を示す。

4-3-1 記述統計量

モデル比較の結果に入る前に，項目の特性を理解するために項目レベルの統計量を示す（表 4.2）。表 4.2 では，項目に解答した人数（*N*），正答率（*Mean*），標準偏差（*SD*），正答率の 95%信頼区間の下限・上限，当該項目を除いた項目場合の平均得点と当該項目の IT 相関とその 95%信頼区間の下限・上限の推定値を示した。正答率と IT 相関は古典的テスト理論における困難度と識別力に対応し，項目の性質を理解するために役立つ指標である。項目 1~14 は計画欠損によりおよそ半数の解答者が解答していないため，解答人数が少なくなっている。なお，信頼性係数としてα係数と 95%信頼区間を算出したところ，α＝.83995（%CI[.855,.871]）となり，通常のテストの信頼性は担保されていると考えることができる。

個々の項目の正答率に関しては，全体的に.6 を超える項目が多く，比較的易しい項目が多かったといえる。その一方で，項目 3（M041056）や項目 21（M031247）などは正答率が.200 を下回っており困難度が高い項目であった可能性がある。IT 相関に関しては.2～.5 程度の項目が多く，最も高い値を示した項目 23（M031085）であっても.560（95%CI[.507, .606]）という値であった。最も IT 相関が低い項目は項目 6（M041164）であり，その値は.236（95%CI[.128, .334]）となっていた。95%信頼区間を考慮してもあまり高い値は示していないと考えられる。項目の性質についてまとめると，全体的に困難度も識別力も低めの項目であるものの，一定の項目数によってテストとしての信頼性を担保できていると考えられる。

ただし，識別力の高さはさほど高くない点と，問題の内容領域が複数想定されているテストである点を考慮すると，テスト全体としては 1 次元的な能力よりも多次元的な能力を考慮する必要性があると考えられる。

4-3-2　モデル比較

表 4.3 に，各モデルの−2 対数尤度，AIC, BIC，MADcor, SRMSR，項目パラメタ数を示した。表 4.3 の各列で値が最小であるセルの色を灰色にして区別した。この結果から，AIC, BIC の観点からは R-RUM が支持された。また，G-DINA モデルは飽和モデルであることから理論的に予測されるとおり対数尤度では最もよく，また AIC の観点から 2 番目によいモデルであった。DINA モデルは BIC の観点では R-RUM についでよいモデルであった。CDM の中では，DINO モデル，A-CDM，LLM の当てはまりは相対的に低かった。本データにおいては，IRT モデルはすべて，どの認知診断モデルよりも当てはまりがよくなかった。

絶対適合指標に関しては期待されたように MADcor および SRMSR のいずれも 3PL の適合が良いことが示された。また，CDM では G-DINA モデルではなく A-CDM で MACcor と SRMSR が最小となった。

4-3-3　アトリビュート習得パタンの比較

表 4.4，表 4.5 および表 4.6 に CDM で当てはまりのよかった R-RUM, G-DINA, DINA モデルの各内容領域（「数」，「図形と測定」，「資料の活用」）での人数の多い上位 10 位のアトリビュート習得パタンを示した。ただし，「資料の活用」は全部で 8 つのパタンしかないため，全てのパタンについての結果を示した。なお，アトリビュート習得パタンの推定は全てのアトリビュートを考慮して推定した後に，当該の領域について集計を行ったものである。

「数」の内容領域では，G-DINA モデルと DINA モデルではすべてのアトリビュートを習得しているパタンに属する人数が最も多かったが，R-RUM では第 4 番目と第 7 番目のアトリビュートが未習得であるパタンが最も多かった。また，G-DINA モデルと DINA モデルは，比較的類似した習得パタンを示したものの，R-RUM では 4 番目のアトリビュートが 3 位まで未習得である点が異なっていた。また，上位 10 位のアトリビュート習得パタンに含まれる解答者の割合は, R-RUM で.676, G-DINA モデルで.485, DINA モデルで.773 であり，DINA モデルを用いた場合には特定のアトリビュートパタンに偏りがみられる可能性が高いことを示唆している。一方で，G-DINA モデルでは上位 10 のアトリビュート習得パタン

である解答者は 50%に満たない割合であり，他のモデルに比べて推定される習得パタンの種類が多いことを反映している可能性がある。また，先行研究と同様，どのモデルを利用するかによって，アトリビュート習得パタンの推定値がかなり異なっていることが示された。

「図形と測定」の内容領域では R-RUM と G-DINA の 1 位から 3 位までの習得パタンは同じであり，習得の人数も比較的似通っていた。一方，DINA モデルでは，すべてのアトリビュートを習得しているパタンが 3 位に入っていたが，R-RUM，G-DINA モデルと異なり，1 つ目のアトリビュートを習得しているパタンが多かった。1〜3 位の習得パタンに注目すると，R-RUM と G-DINA モデルは一致しており，比較的類似した結果を示したものの，DINA モデルは R-RUM，G-DINA モデルとは異なった習得パタンの割合が多く，異なった傾向を示したといえよう。

「資料の表現」の内容領域では R-RUM，G-DINA モデルと DINA モデルのパタンが類似しており，どのモデルにおいても，「111」と「101」のパタンが 1 位と 2 位に見られた。G-DINA モデルと DINA モデルの 1 位はどちらも「111」というパタンであったが，その割合は異なっており，G-DINA モデルは.482，DINA モデルは.729 であった。3 位のパタンは R-RUM では「000」,G-DINA モデルでは「110」，DINA モデルでは「010」とすべて異なっていた。

Lee et al.（2011）のアメリカのデータの結果では「数」の内容領域ではすべてのアトリビュートを習得しているパタンの割合が最も多く，ついで，2 番目のアトリビュートのみを習得しているパタンが 2 位，3 位はひとつも習得していないパタンであり，日本のデータと 1 位の傾向は類似していたものの，2 位・3 位の習得パタンは異なっていた。「図形と測定」の内容領域や，「資料の表現」の内容領域でも同様に，1 位が完全習得パタンであるが，アメリカのデータでは 2 位・3 位は一つもアトリビュートを習得していないパタンや一つないし二つしかアトリビュートを習得していないパタンが占められていた。日本では習得されているアトリビュート数が比較的多いパタンが上位を占める傾向が見られた。

4-3-4　潜在特性とアトリビュート習得確率の関係

本研究で与えられたアトリビュートが解答者の全般的な算数能力を反映しているかどうかを確認し，アトリビュートの性質をより詳細に解釈するために，2PL モデルで推定した潜在特性と R-RUM で推定したアトリビュートの習得確率の関係を分析した。アトリビュートが算数能力の一部を構成するものであれば，全般的な算数能力が高くなるほどアトリビュ

ート習得確率が高くなることが期待される。表 4.7 に，R-RUM での各アトリビュート習得確率を従属変数とし，2PL モデルの潜在特性値を独立変数としたロジスティック回帰分析の結果を示した。アトリビュートごとに切片と回帰係数の推定値および 95%信頼区間（標準誤差の 1.96 倍を用いて算出），オッズ比，オッズ比の 95%信頼区間，p値，R^2値を示した。R^2値はデビアンスを用いて定義される。まず，2 項分布の場合，デビアンスはデータ自身を用いた対数尤度l_fとモデルの対数尤度l_pにより，デビアンス$= -2(l_p - l_f)$と定義される。さらに，Residual deviance と Null deviance を算出する。Residual deviance は回帰係数と切片項を推定したモデルのデビアンス，Null deviance は切片項のみを推定したモデルのデビアンスである。これらのデビアンスを用いて，R^2値は，$R^2 = 1 - \frac{\text{Residual deviance}}{\text{Null deviance}}$と算出される（服部, 2011）。ロジスティック回帰分析におけるR^2は重回帰分析におけるそれと同等の解釈をすることはできない。しかし，ここでは，分散説明率と同様の解釈を行ったとして，独立変数によって従属変数の分散のうち 10%程度を説明できることを一つの基準として結果を解釈する。この結果は，潜在特性の高低によって，習得を強く予測されるアトリビュートと，そうでないアトリビュートがあることを示すこととなる。例えば，「数 2」や「数 7」，「図形と測定 10」，「資料の表現 15」のアトリビュートは高い回帰係数の値を示した。一方で，「数 1」や「図形と測定 12」，「資料の表現 14」のアトリビュートは相対的に見て小さい回帰係数であった。これは，「数 1」ではアトリビュートを習得している確率が高い解答者が相対的に多く，習得確率の分布に偏りが見られたことに起因するものと考えられる。「図形と測定 12」では，習得確率が 0 や 1 付近だけでなく，.7 から.8 周辺の習得確率に多くの解答者が集中しており，やはり，習得確率の偏りがみられ，これによって，回帰係数が小さく推定された可能性がある。さらに，「資料の表現 14」では，潜在特性が低い場合でも習得確率が高い解答者も多く，全般的な算数能力を用いた習得確率の予測が難しい分布になっていたと考えられる。

さらに，図 4.1 にそれぞれのアトリビュートの習得確率と 2PL モデルの潜在特性との間の散布図を示した。図中の曲線は推定されたパラメタ値によるロジスティック曲線である。このような図によって，アトリビュート習得確率がどのように 2PL モデルで推定される潜在特性と関連があるのかをひと目で把握することが可能になる。この図からも，潜在特性値が習得確率をよく予測するアトリビュートと，そうでないアトリビュートがあることが見てとれる。例えば，「数 1」は比較的習得が容易なアトリビュートであることが読み取れるが，一方，「数 4」は比較的習得が困難なものであることがわかる。さらに，「図形と測定 9」，

「図形と測定 12」や「資料の表現 14」などは表 4.7 に示したR^2値も.151 や.055，.197 であり，相対的にロジスティック曲線の当てはまりがよくなかった。

4-3-5　　項目パラメタの推定値

アトリビュート習得確率と一次元的な潜在特性との関連に加えて，最もデータに適合していた R-RUM の項目パラメタの推定値を検討することで，モデルがどのようにデータに適合しているのか検討を行った。表 4.8 に R-RUM の項目パラメタの推定値を示した。

全てのアトリビュートを習得している場合の基本正答確率である$\hat{\pi}$がいずれも高いことから，これらのアトリビュートによって，項目に正答する確率を適切に捉えることができていると考えられる。ただし，項目 11 や 21 は比較的$\hat{\pi}$の値が小さく，これらのアトリビュートだけでは正答確率をとらえられない項目である可能性も示唆された。

項目パラメタの推定値が 1 に近い値を示しているものもみられ，実質的に罰則としての意味をなさない項目パラメタもみられた。例えば，項目 4 や項目 6 では，単一のアトリビュートのみが影響を与えている可能性が示唆された。

しかし，そのような中でも，項目 1 や項目 3，項目 11 などでは，複数の項目パラメタが罰則項として機能している様子も観察された。このように，項目の正答に必要な 1 つでもアトリビュートが欠けていることによって，正答率に影響を与えることからπパラメタには各アトリビュートが正答に寄与する部分だけでなく，複数のアトリビュートの組合せの効果も含まれている可能性がある。ただし，R-RUM のモデル式の定義からは DINA モデルのように明示的な交互作用を含まないため，解釈には注意が必要といえよう。

4-4 考察

本章の目的は，TIMSS2007 の 4 年生日本人サンプルデータを用いて，(1) CDM の方が IRT モデルよりもデータへの適合がよいという Lee et al.（2011）のアメリカでの結果が再現できるかを確認し，(2) 複数の CDM を同一のデータに当てはめて，当てはまりのよさを経験的に検討し，相対的に適合したモデルでのアトリビュート習得パタンを用いてアメリカのデータとの相違を明らかにし，(3) なぜそのモデルが適合したのかその理由を他のモデルと比較しながら議論・考察することであった。

4-4-1　　得られた知見

結果として，1 点目については，情報量規準から IRT モデルよりも CDM の適合がよいことが示され，この点で Lee et al.（2011)の結果は日本人サンプルにおいても再現された。さらに，2 点目の CDM 内の比較では，先行研究では検討されなかった各種モデルの比較を行った結果，R-RUM の適合がよいことが示された。アトリビュート習得パタンに関しては，日本人データとアメリカデータでの傾向は，利用しているモデルの違いはあるにしても，異なっている可能性が示された。2PL から推定された潜在特性値とアトリビュート習得パタンの関係の分析では，アトリビュート自体がどの程度の困難度なのか，当該のテストで測られる能力の高さによってアトリビュートの習得を予測できるのかを検討し，潜在特性の高さと関連の高いアトリビュートと関連の低いアトリビュートが混在していることが明らかになった。

上記の結果を踏まえ，3 点目の目的であったモデル適合についての考察を行う。IRT よりも CDM の方が当てはまりがよいという本章と先行研究で共通する結果から，TIMSS2007 の 4 年生の算数データでは，1 次元的な算数能力を測定しているというよりも，むしろ，複数の能力を総合した解答が求められるテスト項目になっていると推測できる。絶対適合度指標としては 3PL ロジスティックモデルの適合が良かったものの，これは潜在変数が連続 1 次元であったことに起因すると考えられる。また，TIMSS は CDM を適用する前提で作成されていないため，SRMSR などの絶対適合指標では良い適合の値を示さなかった可能性はある。本章において，情報量規準の観点からは CDM の中でも特に R-RUM の適合がよかったことから，TIMSS2007 のアトリビュートには非補償的関係があることが示唆される。つまり，複数のアトリビュートが存在したときに，アトリビュートを習得していないことによって，項目に正答する可能性が極めて小さくなりうるといえる。項目パラメタの推定値からも，この様子が観察された。これは，やはり非補償モデルである DINA モデルの適合が，BIC の観点で二番目によかったこととも整合的である。合わせて，G-DINA モデルも AIC の観点では適合していたことについて，G-DINA モデルの特色である複数のアトリビュートを習得していることによる交互作用効果ということに注目すれば，R-RUM，DINA モデルのように相対的に適合していたことにも整合する。R-RUM を G-DINA モデル式から構成した場合，一見すると加法的なモデルに近いが，項目パラメタが加算的にではなく，掛け算的に影響を与えている点に注意が必要である。これは，各アトリビュートの影響が線形ではないことを意味しており，基本正答率にはアトリビュートの単純加算以上の交互作用効果といっ

た要素が含まれている場合もある。ただし，R-RUM は G-DINA モデルの下位モデルとしては log リンク関数をとった場合の主効果モデルと解釈される。これは，R-RUM では，アトリビュートが個別の影響力を持っているということを意味している。

アトリビュート習得パタンはモデルによって異なっていた。DINA モデルは全てのアトリビュートを習得しているパタンの人数が多い傾向を示し，R-RUM や G-DINA モデルは必ずしもそのような傾向ではなく，他の習得パタンである解答者も一定数いるということが推定された。情報量規準の意味で最適な基準のモデル以外は参考程度の結果ではあるものの，この結果はデータに適用するモデルによってアトリビュート習得パタンはかなり異なっている可能性があることを示唆している。

4-4-2　　アトリビュートの影響

具体的なアトリビュートから TIMSS2007 の交互作用についての考察を行う。TIMSS2007 の問題項目では一つしかアトリビュートが必要ではない項目が 3 項目にとどまっており，残り 22 項目には複数のアトリビュートが必要であった。また，複数の内容領域のアトリビュートが必要な項目も散見された。「資料の表現」のアトリビュートに至ってはすべてが「数」,「図形の測定」のアトリビュートと同時に測定されていた。アトリビュートのうち，4 項目以上で測定されていたものは,「数 1. 位置の値の知識を実行するだけでなく，整数の表現，比較，順序づけすること」(6 項目),「数 2. 四つの演算子を用いた整数の計算と倍数の認識および計算の推論」,（16 項目),「数 3. 実生活の文脈での問題解決（例：測定，お金の問題）を含む問題解決」（11 項目),「図形と測定 9. 数と直線を描くためのそれらの性質の理解と測定および推測」（7 項目),「資料の表現 13. 表・統計図表・棒グラフ・円グラフからのデータの読み取り」（4 項目）の合計 5 つであった。これ以外のアトリビュートは 3 項目以下の項目で測定されていた。

さて，ここではとくに多くの項目で測定されているアトリビュートに注目する。例えば，項目 5 は「数 2」,「数 3」,「数 8」のアトリビュートが必要であり，実生活上での算数の活用ができるかどうかを測定している項目といえよう。項目パラメタの推定値からは,「数 2」,「数 8」のアトリビュートが項目の正答に重要であることが示された。このような項目では純粋な算数の知識や計算のみならず，実生活での文脈や具体的な状況に合わせて既存の知識を適用する必要があると考えられる。こういった具体的な文脈での効果が一種の交互作用効果と解釈できる可能性がある。このほか,「数 1」,「数 2」,「数 3」,「図形と測定 10」,

「図形と測定 11」を含む項目 8 では，項目パラメタの推定値から実質的には，「数 2」と「図形と測定 10」が主要なアトリビュートであった。このため，この問題には図形を正しく理解・分類し，その上で四則演算を用いて解答することが求められるのであろう。ここでは，図形のどのような値に注目するのかといった，その値をどのように活用するのかということも問われていると考えられ，単純なアトリビュートの効果の和ではない可能性が示唆される。「数 1」，「数 2」，「資料の表現 15」を測定している項目 25 では，「資料の表現 15」が最も正答に寄与しており，ついで「数 2」が影響している。これは，図表を変換して様々な表現形式として理解する必要があり，変換の際に数値的な計算が必要である可能性を示唆しており，資料を活用する際に必要になる数値の読み取りと具体的な計算を適用することが複合的に必要と考えられる。この他にも多数の項目で複数のアトリビュートが必要とされており，日常生活の文脈に合わせて知識を総合的に活用したり，知識の組み合わせが必要な問題が多く含まれていることから，R-RUM や G-DINA モデル，DINA モデルといった交互作用を含むモデルがデータに適合していたことも理解できよう。R-RUM は交互作用項を明示的に含まないが，全てのアトリビュートを習得している場合が基本的な正答率であり，その正答率に交互作用が含意されていると解釈できる。

4-4-3　限界と展望

本章の結果は，IRT との比較について Lee et al.（2011）の結果は再現されたものの，CDM 間の比較では異なるデータを用いた Li et al.（2016），鈴木他（2015）の知見とは必ずしも一致していない。Li et al.（2016）の結果では AIC の観点から G-DINA モデルが採用されており，この点は本章の結果に近いものであったものの，BIC の観点や鈴木ら（2015）の結果では A-CDM が採用されている。テスト領域の相違やそれに伴うアトリビュートの前提の相違によって，本章の結果と先行研究の結果が異なったと考えられる。

また，アトリビュート習得パタンは同じテスト項目やアトリビュートを用いても，分析するデータ・セットが異なれば一致しない様子が観察された。さらに，用いるモデルによっても，習得パタンの推定値はかなり変化する様子も観察された。これらの結果から，CDM の適用においては，当該のテスト領域でどのように問題解決が図られるのかというモデル化を慎重に行い，データごとに複数のモデルを比較する必要があるといえる。さらにいえば，テスト全体に同じ項目反応関数を仮定するのではなく，項目単位でのモデリングも必要になる可能性もある。

潜在特性値とアトリビュート習得パタンの関係の分析からは，アトリビュートの統計的な性質を考慮する必要性が示唆された。なぜならば，ほとんどの解答者が習得しているアトリビュートや，逆に習得率の低いアトリビュートなどが存在することや，テスト全体で必要なアトリビュートと関係がないアトリビュートが存在することが示されたからである。このようなアトリビュートは診断の目的に照らして必要な情報を提供しないばかりか，推定するパラメタが多くなることや，極端な習得パタンに解答者が集中することによって，不安定な推定をもたらす危険性を孕んでいる。このような潜在的な問題に対処するために，例えば，IRT モデルの項目パラメタに相当するような困難度や識別力といったものを導入し，パラメタからアトリビュートの性質を調べることなども必要であるかもしれない。

本章の限界としては，用いたデータが日本人サンプルのみであった点が挙げられる。そのため，R-RUM が TIMSS2007 に適合していたとしても，その結果が日本人サンプルに限定的である可能性を否定できない。さらに，Lee et al.（2011）や Choi et al.（2015）のように，複数の国や地域のデータを用いて，モデル比較を行うことで，TIMSS データに適した CDM が何であるか，より詳細な検討が必要であろう。また，国際比較に加えて，2011 年に行われた TIMSS2011 の日本人データにおいても同様の適合がみられるのかといった縦断的な観点からも，モデル適合に関する知見の蓄積をすすめることが望ましいだろう。さらに，日本と他国での適合のよいモデルの違いを検討する際に，国ごとの教育制度や教育課程，家庭の学習環境などを交えて，検討を行うことによって診断情報を有効に利用することができよう。TIMSS データでは，教員や家庭に対する質問紙調査の結果も利用できる。今後はこういった観点からの検討も必要である。

本章では，Lee et al.（2011）との比較可能性を担保するために，先行研究と同様の設定を用いて検討を行った。しかし，アトリビュートや Q 行列についての再考も必要である可能性が示唆された。項目パラメタの推定結果から，項目に不必要に付加されているアトリビュートも存在しており，また，全体として負荷している数が少ないアトリビュートもみられた。テスト全体としては必要なアトリビュートであっても，どの項目に必要であるかということは，推定された項目パラメタなども勘案して修正が必要であろう。そのうえで，アトリビュート数を減らし，よりデータを表現できるアトリビュートのセットを再度検討し，本章結果の一般化可能性を検討する必要もあるだろう。

これらの限界のうち，特に一般化可能性に着目し，次章ではさらに多くの国のデータを使ってモデル比較を実行する。詳細は次章で示すが，TIMSS に参加している様々な学力層

の国において，選択されるモデルを検討することによって，本章の結果がどれほど一般化可能な知見であるのかを検討する。

表 4.1 TIMSS2007 の 4 年生の算数の Q 行列（Lee, Park, & Taylan（2011）の Table 3 を一部改変）

		数								図形と測定				資料の表現			
項目番号	項目名	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	合計
1	M041052	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	2
2	M041056	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
3	M041069	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0	3
4	M041076	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	2
5	M041281	0	1	1	0	0	0	0	1	0	0	0	0	0	0	0	3
6	M041164	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	2
7	M041146	0	0	0	0	0	0	0	0	1	1	0	1	0	0	0	3
8	M041152	1	1	1	0	0	0	0	0	0	1	1	0	0	0	0	5
9	M041258A	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1
10	M041258B	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	2
11	M041131	0	1	1	1	0	0	0	0	1	0	0	0	0	0	0	4
12	M041275	1	0	0	0	0	0	0	0	0	0	0	0	1	0	1	3
13	M041186	1	1	0	1	0	0	0	0	0	0	0	0	1	0	0	4
14	M041336	1	1	0	0	1	1	0	0	0	0	0	0	1	1	0	6
15	M031303	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	2
16	M031309	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	2
17	M031245	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	2
18	M031242A	0	1	1	0	0	0	0	1	0	0	0	0	0	0	0	3
19	M031242B	0	1	1	0	0	0	0	0	0	0	0	0	0	1	0	3
20	M031242C	0	1	1	0	0	0	0	1	0	0	0	0	0	1	0	4
21	M031247	0	1	1	0	0	0	1	0	0	0	0	0	0	0	0	3
22	M031219	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	3
23	M031173	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	2
24	M031085	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1
25	M031172	1	1	0	0	0	0	0	0	0	0	0	0	1	0	1	4
	合計	6	16	11	3	3	2	2	3	3	7	2	3	4	3	2	70
	Mean	.240	.640	.440	.120	.120	.080	.080	.120	.120	.280	.080	.120	.160	.120	.080	2.800
	SD	.427	.480	.496	.325	.325	.271	.271	.325	.325	.449	.271	.325	.367	.325	.271	1.200

表 4.2 各項目の平均正答率および IT 相関

項目番号	項目名	*N*	*Mean*	*SD*	*Mean* 95%CI[下限, 上限]	IT相関	IT相関 95%CI[下限, 上限]
1	M041052	317	.861	.346	[.796 , .926]	.399	[.303 , .481]
2	M041056	308	.568	.496	[.490 , .647]	.361	[.261 , .448]
3	M041069	308	.188	.392	[.118 , .258]	.349	[.247 , .437]
4	M041076	306	.725	.447	[.651 , .800]	.503	[.417 , .574]
5	M041281	309	.790	.408	[.718 , .861]	.457	[.366 , .534]
6	M041164	313	.690	.463	[.615 , .765]	.236	[.128 , .334]
7	M041146	316	.813	.390	[.744 , .882]	.534	[.454 , .600]
8	M041152	314	.780	.415	[.709 , .851]	.371	[.273 , .457]
9	M041258A	301	.568	.496	[.489 , .648]	.393	[.294 , .477]
10	M041258B	299	.274	.447	[.198 , .350]	.328	[.223 , .419]
11	M041131	314	.433	.496	[.355 , .511]	.289	[.184 , .382]
12	M041275	316	.930	.255	[.875 , .986]	.279	[.174 , .373]
13	M041186	316	.665	.473	[.589 , .740]	.432	[.339 , .510]
14	M041336	310	.455	.499	[.376 , .533]	.362	[.262 , .449]
15	M031303	630	.743	.437	[.691 , .795]	.460	[.398 , .515]
16	M031309	626	.816	.388	[.768 , .865]	.478	[.418 , .532]
17	M031245	628	.339	.474	[.285 , .393]	.480	[.419 , .533]
18	M031242A	626	.879	.327	[.834 , .923]	.404	[.338 , .464]
19	M031242B	616	.856	.352	[.809 , .902]	.486	[.426 , .539]
20	M031242C	629	.808	.394	[.759 , .857]	.353	[.284 , .416]
21	M031247	609	.151	.358	[.104 , .199]	.300	[.227 , .368]
22	M031219	627	.659	.475	[.605 , .713]	.387	[.320 , .448]
23	M031173	627	.702	.458	[.649 , .755]	.560	[.507 , .606]
24	M031085	623	.722	.448	[.670 , .775]	.478	[.417 , .531]
25	M031172	606	.814	.390	[.764 , .863]	.498	[.438 , .551]

表 4.3 日本人データにおける IRT モデルと CDM のモデル比較

モデル	−2対数尤度	AIC	BIC	MADcor	SRMSR	項目パラメタ数
3PL	10672.95	10722.95	10834.45	.042	.059	75
2PL	10527.85	10627.85	10850.85	.044	.061	50
1PL	10499.14	10649.14	10983.63	.066	.087	25
G-DINA	8902.79	9426.79	10595.28	.207	.182	262
DINA	9316.41	9506.41	9930.10	.184	.165	95
DINO	10387.21	10487.21	10710.21	.190	.171	50
A-CDM	10367.09	10467.09	10690.09	.182	.164	50
LLM	9398.40	9588.40	10012.09	.197	.175	95
R-RUM	9085.58	9275.58	9699.27	.188	.169	95

Note. N=639,1-3PL=１-3パラメタロジスティックモデル

表 4.4 日本人データにおける R-RUM，G-DINA，DINA モデルでの数量域のアトリビュート習得パタンの比較

R-RUM			G-DINA			DINA		
パタン	人数	割合	パタン	人数	割合	パタン	人数	割合
11101101	122	.191	11111111	84	.131	11111111	224	.351
11101111	116	.182	11101101	56	.088	11110101	86	.135
11100101	55	.086	10110101	50	.078	11111101	80	.125
10100101	31	.049	10101101	29	.045	11101101	30	.047
11111111	24	.038	00000000	19	.030	11110001	14	.022
11100100	23	.036	11011111	17	.027	11100101	14	.022
10000000	19	.030	10100101	15	.023	11110111	13	.020
11101110	15	.023	01111101	14	.022	00000101	12	.019
11100111	13	.020	01111110	13	.020	00000001	10	.016
10101101	14	.022	11101111	13	.020	11111001	11	.017
合計	432	.676		310	.485		494	.773

表 4.5 日本人データにおける R-RUM，G-DINA，DINA モデルでの図形と測定領域のアトリビュート習得パタンの比較

R-RUM			G-DINA			DINA		
パタン	人数	割合	パタン	人数	割合	パタン	人数	割合
0111	239	.374	0111	219	.343	1111	340	.532
0101	82	.128	0101	69	.108	1011	92	.144
1111	61	.095	1111	67	.105	1110	52	.081
0001	59	.092	0011	34	.053	0001	39	.061
0011	54	.085	0001	33	.052	0111	27	.042
0000	40	.063	0010	32	.050	0101	24	.038
1101	16	.025	0000	26	.041	0100	11	.017
0110	23	.036	0110	26	.041	0110	18	.028
0010	30	.047	1101	26	.041	0011	19	.030
0100	12	.019	1011	24	.038	1101	5	.008
合計	616	.964		556	.870		627	.981

表 4.6 日本人データにおける R-RUM，G-DINA，DINA モデルでの資料の活用領域のアトリビュート習得パタンの比較

R-RUM			G-DINA			DINA		
パタン	人数	割合	パタン	人数	割合	パタン	人数	割合
101	231	.362	111	308	.482	111	466	.729
111	210	.329	101	98	.153	101	77	.121
000	77	.121	110	54	.085	010	33	.052
001	53	.083	001	42	.066	000	23	.036
011	28	.044	011	41	.064	011	20	.031
100	16	.025	010	36	.056	100	8	.013
110	13	.020	100	32	.050	110	7	.011
010	11	.017	000	28	.044	001	5	.008
合計	639	1.000		639	1.000		639	1.000

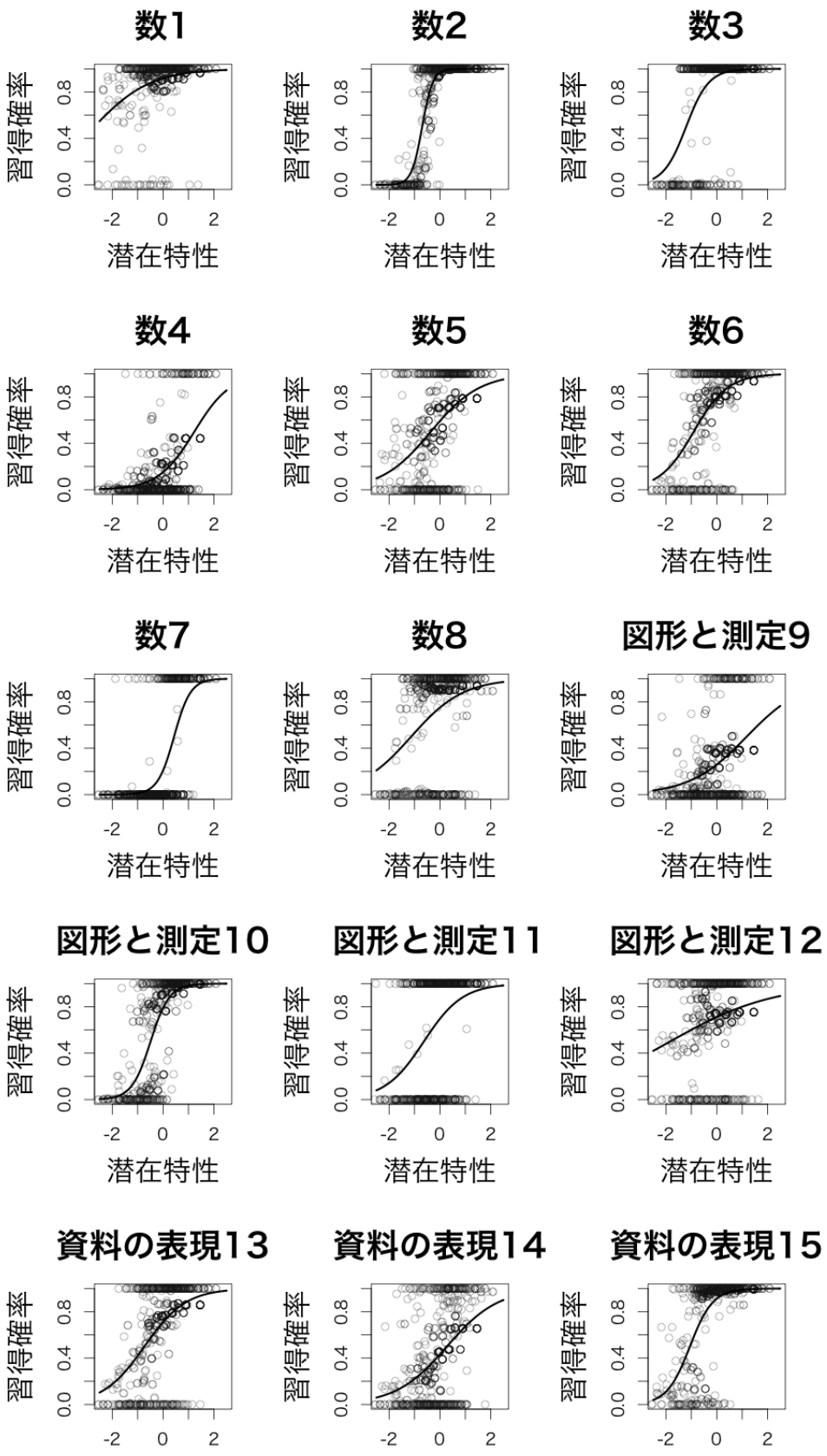


図 4.1 各アトリビュート習得確率と 2PL モデルで推定した潜在特性の関係

表 4.7 2PL モデルを用いた潜在特性値を独立変数，R-RUM で推定した各アトリビュートの習得確率を従属変数としたときのロジスティック回帰分析の結果

従属変数 (アトリビュート)	推定値 上段：切片 下段：回帰係数	95%CI [下限，上限]	オッズ比	オッズ比95%CI [下限，上限]	p	R^2
数1	2.423	[2.114 - 2.732]			<.001	.153
	0.891	[0.585 - 1.198]	2.438	[1.794 - 3.312]	<.001	
数2	3.342	[2.733 - 3.951]			<.001	.746
	4.791	[3.891 - 5.691]	120.428	[48.972 - 296.150]	<.001	
数3	2.702	[2.304 - 3.099]			<.001	.355
	2.241	[1.818 - 2.663]	9.399	[6.162 - 14.337]	<.001	
数4	-1.726	[-1.979 - -1.472]			<.001	.280
	1.407	[1.089 - 1.724]	4.082	[2.971 - 5.609]	<.001	
数5	0.373	[0.202 - 0.544]			<.001	.201
	1.037	[0.811 - 1.263]	2.820	[2.250 - 3.535]	<.001	
数6	1.350	[1.126 - 1.575]			<.001	.336
	1.490	[1.207 - 1.772]	4.436	[3.344 - 5.885]	<.001	
数7	-1.327	[-1.603 - -1.050]			<.001	.423
	3.082	[2.581 - 3.582]	21.795	[13.213 - 35.951]	<.001	
数8	1.102	[0.907 - 1.297]			<.001	.142
	0.973	[0.740 - 1.206]	2.647	[2.097 - 3.341]	<.001	
図形と測定9	-1.049	[-1.238 - -0.859]			<.001	.151
	0.892	[0.650 - 1.134]	2.439	[1.915 - 3.108]	<.001	
図形と測定10	1.332	[1.071 - 1.593]			<.001	.546
	2.699	[2.266 - 3.132]	14.861	[9.638 - 22.915]	<.001	
図形と測定11	0.831	[0.641 - 1.022]			<.001	.170
	1.311	[1.057 - 1.566]	3.712	[2.878 - 4.787]	<.001	
図形と測定12	0.900	[0.725 - 1.076]			<.001	.055
	0.489	[0.286 - 0.692]	1.631	[1.331 - 1.998]	<.001	
資料の表現13	0.902	[0.710 - 1.094]			<.001	.256
	1.225	[0.977 - 1.473]	3.404	[2.657 - 4.362]	<.001	
資料の表現14	-0.274	[-0.442 - -0.105]			.001	.197
	0.993	[0.767 - 1.219]	2.700	[2.153 - 3.385]	<.001	
資料の表現15	2.340	[1.994 - 2.687]			<.001	.450
	2.215	[1.814 - 2.616]	9.165	[6.137 - 13.686]	<.001	

表 4.8 R-RUM の項目パラメタの推定値

			数								図形と測定				資料の表現		
項目番号	項目名	π	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	M041052	.986	.611	.618													
2	M041056	1.000					.000										
3	M041069	.998		1.000		.000	.307										
4	M041076	1.000			1.000			.000									
5	M041281	1.000		.519	1.000					.581							
6	M041164	1.000										1.000		.000			
7	M041146	1.000									.990	.480		1.000			
8	M041152	.920	1.000	.780	.929							.622	1.000				
9	M041258A	.737										.315					
10	M041258B	1.000									.000	.849					
11	M041131	.641		.345	.766	.885					.870						
12	M041275	1.000	.312												1.000		1.000
13	M041186	1.000	1.000	1.000		1.000									.000		
14	M041336	1.000	1.000	1.000			1.000	1.000							1.000	.145	
15	M031303	.875		.383	.884												
16	M031309	.925		.466	1.000												
17	M031245	1.000		1.000					.000								
18	M031242A	.985		1.000	.266					1.000							
19	M031242B	1.000		1.000	.000											.999	
20	M031242C	1.000		.992	.999					.320						1.000	
21	M031247	.307		.001	.907				.404								
22	M031219	1.000										1.000	.000	1.000			
23	M031173	.848		.283	.896												
24	M031085	.936										.295					
25	M031172	1.000	.998	.720											1.000		.167

Note. GDINA関数ではR-RUMモデルの標準誤差が算出されないためここでは項目パラメタの推定値のみを示した。

第5章　TIMSS2007 の各国における認知診断モデルと項目反応理論モデルの比較

第 5 章では，第 4 章で得られた知見の一般化を目指して TIMSS 2007 に参加した様々な国を選択して 4 年生の算数データを用いて CDM のモデル比較を行う。さらに，第 4 章と同様に，各国での問題項目の基礎統計量やアトリビュート習得パタン，1 次元潜在特性とアトリビュート習得確率との関連，および項目パラメタの推定値についての結果も合わせて検討する。

5-1 問題と目的

第 4 章では，TIMSS 2007 の日本人データを用いて倹約モデル（DINA モデル，DINO モデル），主効果モデル（A-CDM，R-RUM，LLM），飽和モデル（G-DINA モデル），1-3 パラメタロジスティック IRT モデルを比較し，主効果モデルである R-RUM の適合が相対的によいことを示した。

しかしながら，この結果は日本人データに限定された結果である可能性も拭いきれない。TIMSS においてどの CDM が適合するのかを検討するためには，他の国や地域においても同様にモデル比較を行う必要がある。第 1 章で述べたように，CDM の比較研究では同じ問題項目による国際データの比較研究は行われておらず，特定の地域データで得られた結果の一般化可能性についてはこれまでには検討されてこなかった。TIMSS データを用いることで，同一問題に解答した各国のデータの比較が可能となる。国際データを比較することにより，解答者の思考を理解するだけでなく，教室場面での教育方法を変えることや国際的なデータから各国への教育政策への示唆を与えるといったことが可能といわれている（Cai, Mok, Reddy, & Stacey, 2016）。

そこで，第 5 章では第 4 章で得られた結果を一般化することを目的として，TIMSS スコアを基準に，広範に国を選択し，様々な国での基本的な IRT モデルと種々の CDM の比較を行うことを目的とした。あるモデルが広い範囲に渡って，最も適合が良いと言えるかどうかを調べるには，できるだけ異質性の高いサンプルを選択する必要がある。この異質性について，本研究ではその異質性を，後述する TIMSS の到達度スコアの差異で評価することとした。個々の国のデータを個別に検討することにより，モデル比較の詳細な検討を試みた。もし，日本人データで得られた結果が他の国でも同様に得られたならば，TIMSS の解答プロセスとして，R-RUM が有力である可能性が高くなるといえよう。また，もし R-RUM 以外

のモデルが適合したとして，第 2 章で検討した倹約モデル，主効果モデル，統合モデルのうち，どのようなタイプの CDM が適合するのかを明らかにし，その傾向を調査することによって，TIMSS における項目反応関数としてどのようなものが適切であるのか，検討することが可能である。この目的に加えて，第 4 章と同様に，各国での項目の記述統計量を比較したり，各国での最も適合のよいモデルを用いたアトリビュート習得パタンの比較や，モデルの項目パラメタの精査を通して TIMSS データでの CDM の振る舞いを精緻に調査することも本章の目的とした。さらにアトリビュートの妥当性の検討として，TIMSS の算数スコアや 2PL ロジスティックモデルで推定した一次元潜在特性とアトリビュートの習得確率や各国のアトリビュート習得数についての関係を調査した。

5-2 方法

第 4 章および，Lee et al.（2011）と同様に本章では TIMSS 2007 の 4 年生の算数データのうち，ブックレット 4，5 に解答した解答者データを分析した。データには Lee et al.（2011）で用いられたアメリカのデータのみならず，他に 6 カ国のデータを用いた。具体的には，Martin, Mullis, & Foy（2008）の Exhibit 1.1 に示されている TIMSS の到達度スコアをもとに，高到達度，平均到達度，低到達度に当たる国を 2 カ国ずつ選定した。高到達度国は TIMSS 到達度スコアの第一位，第二位に該当する香港 SAR[1]（以下，香港とする）とシンガポールを，平均到達度国としては TIMSS 平均スコアの上下に隣接するスロベニアとアルメニアを選択し，低到達度国としては最もランキングが低いカタールとイエメンを選択した。

日本人データでの解析と同様に，項目への反応が 1 つもないかあるいは 1 つしかないデータを除外した。その結果として，最もサンプルサイズが大きい国はアメリカ（$N = 1{,}130$）で，最もサンプルサイズが小さい国は香港であった（$N = 543$）。表 5.1 に全ての国のサンプルサイズを示した。また，日本人データの場合と同様に，各国における各項目の正答率（および，95%CI），標準偏差，IT 相関（当該の項目を含まない合計点を利用し，95%CI も算出）を psych パッケージ ver.1.7.5（Revelle, 2017）の alpha 関数を用いて算出した。また，各国での 25 項目全体のテストとしての信頼性係数αとその 95%信頼区間も算出した。アトリビュートおよび Q 行列は Lee et al.（2011）と同様のものを利用した。利用したモデルおよびモデル比較に利用した指標も第 4 章と同様であり，モデルとしては 1-3PL IRT モデル，

[1]本論文では「国と地域」を一括して「国」と表現する。

G-DINA モデル，DINA モデル，DINO モデル，R-RUM，A-CDM，LLM を対象とし，指標としては−2 対数尤度（デビアンス），情報量規準である AIC と BIC，絶対指標である MADcor と SRMSR を利用した。1-3PL IRT モデルの推定には R の ltm パッケージ（Rizopoulos, 2006）の tpm 関数を利用し，項目パラメタの推定には周辺最尤推定を用いた。また，G-DINA モデルとその下位モデルの推定には GDINA パッケージ（Ma & de la Torre, 2017）に含まれる GDINA 関数を利用した。GDINA 関数の設定はデフォルトとし，項目パラメタの推定には EM アルゴリズムによる周辺最尤推定法を用いた。項目パラメタの標準誤差の推定は，パッケージのデフォルトである完全情報行列の逆行列を，外積近似した行列の対角要素の正の平方根を用いた。ただし，現状の G-DINA 関数では R-RUM の項目パラメタの標準誤差は算出されないため，R-RUM モデルが選択された場合には項目パラメタのみを提示することとした。適合度指標の計算も第 4 章と同様に行った。IRT モデルでの絶対適合指標の算出には，tpm 関数で推定した項目パラメタを用いて，TAM パッケージ（Robitzsch et al., 2017）の tam.modelfit 関数を用いた。

妥当性の検討のために，2PL ロジスティックモデルで推定した一次元潜在特性を独立変数，アトリビュートの習得確率を従属変数としたロジスティック回帰分析を行った。潜在特性の推定には，前述の ltm パッケージの tpm 関数を利用し，推定値としては MAP 推定値を用いた。アトリビュート習得確率はアトリビュート習得パタンを周辺化した周辺確率を用いた。また，アトリビュート習得パタンは EAP 推定値，ロジスティック回帰分析は R の glm 関数を利用した。加えて，アトリビュートの領域ごとに，一次元潜在特性と各国のアトリビュート習得数とのピアソンの積率相関を算出した。また，TIMSS の公式の到達度スコアは Martin et al.（2008）の当該の国の値を利用した。

5-3 結果

結果は第 4 章と同様に問題項目の基礎統計量，モデル比較，アトリビュート習得パタン，CDM の結果と IRT モデルで推定した 1 次元潜在変数の関連，さらに項目パラメタの推定値についての分析結果を提示する。

5-3-1 各国の項目の基礎統計量

表 5.2〜表 5.8 に各国のデータを用いた項目の正答率，標準偏差，標準誤差，IT 相関，等の基礎統計量を示した。イエメンにおいて，項目 20（M031247）は，欠測により他の変数

と 2 つのデータが揃っている解答者の解答が全て 0 となり，相関係数が算出できないため IT 相関の分析から除外した。また，正答率の 95%信頼区間の算出には，正規近似を用いたため, 信頼上限が 1.000 を超過した場合がみられた。この場合は, 値を 1.000 として扱った。IT 相関の 95%信頼区間は第 4 章と同様に，非心 *t* 分布にもとづく点双列相関係数の信頼区間の構成方法（南風原, 2014）を用い，非心度の算出には R の MBESS パッケージ ver.4.3.0（Kelley, 2017）を利用した。

全体的な傾向として，TIMSS の公式スコアの順位が高い国において正答率が高い場合が多く，順位が低くなるに連れて正答率が低下していく傾向がみられた。IT 相関は項目の困難度が.4 など適度な値であっても必ずしも高いわけではなく，かなり低い項目も含まれていた。さらに，カタールやイエメンにおいては，IT 相関がきわめて低く，場合によっては負となる項目も散見された。以下，各国の結果を簡単に概観する。

まず，アメリカデータ（表 5.2）の正答率が.900 を超える比較的困難度の高い項目としては，項目 6（M041164）が挙げられ，その正答率は.914（95%CI[.870, .958]）であった。また，正答率が.100 を下回る低い項目としては，項目 10（M041258B）が.084（95%CI[.040, .128]）のみであった。その他項目に関しては，.200 から.800 程度の正答率に収まっており，極端な困難度の偏りはなかったと考えられる。また，アメリカデータはサンプルサイズが 500 以上であり平均値の標準誤差が比較的小さく，また 95%信頼区間の下限・上限の差も.100 程度に押さえられており，安定した推定値が得られている。IT 相関に注目すると，推定値として最も低い値は項目 22（M031249）の.102（95%CI[.043, .160]）であった。一方，項目 10（M041258B）や項目 11（M041131）の IT 相関の推定値は.168（95%CI[.085, .246]）と.163（95%CI[.081, .242]）であり，95%信頼下限が.100 を下回る程度には低い値であった。一方，最も IT 相関が高かった項目は項目 16（M031303）であり，値は.572（95%CI[.534, .607]）であり信頼下限が.500 を上回っており，比較的高い値であったと判断することができる。また，全体的には.200 から.500 程度の IT 相関の値を示しており，アメリカデータにおいては大きな問題のある項目は少なかったと考えることができる。

香港データ（表 5.3）においては項目 21（M031247）の正答率が.144（95%CI[.092, .196]）であり，若干低い正答率であったものの，それ以外の項目では.6 以上の正答率を示した項目が多かった。このことから，香港の解答者にとっては比較的容易な問題が揃っていたと考えられる。IT 相関については，項目 11（M041131）が. 096（95%CI[−.024, .211]）であり，95%信頼区間が 0 を含むほどに低く，識別力が低い項目であったと考えられる。他の項目とし

て，項目 6（M041164），項目 7（M041146），項目 8（M041152），項目 9（M041258A），項目 10（M041258B）および項目 12（M041275）なども，IT 相関の 95%信頼下限が.1 を下回っており，識別力が低い項目であったといえる。その一方で，IT 相関は.3 から.6 弱程度の項目が多く見られた。

シンガポールデータ（表 5.4）においても，香港データと同様に正答率が.5 を上回る様な項目が多く，全体として正答率が高い傾向がみられた。また，香港データとはことなり，極端に正答率が低い項目は項目 11（M041131）であり，*Mean* = .442（95%CI[.369, .516]）という値を示した。また，IT 相関に関しては，.4～.6 程度の項目が多く，シンガポールデータにおいてはこれらの項目の識別力が高い傾向がみられた。ただし，項目 22（M03121）の IT 相関は.218（95%CI[.147, .284]）であり，やや低い値を示したものの，95%信頼下限が.1 を下回る項目は存在しなかった。

スロベニアデータ（表 5.5）の困難度は.5~.7 程度の項目が散見された。ただし，項目 3（M041069）や項目 10（M041258B）などは，正答率が.05 を下回る項目であり，かなり正答率が低い項目であった。項目 3（M041069）は困難度が高い項目であったものの，IT 相関としては.207（95%CI[.094, .310]）という値を示した。IT 相関については，.4 以上の項目も多く見られた。項目 22（M031219）などは，.111（95%CI[.031, .189]）であり，低い識別力の項目であったと考えられる。

アルメニアデータ（表 5.6）において，項目の正答率の最小値を示した項目は，項目 21（M031247）であり，.122（95%CI[.071, .172]）という値を示した。他の項目の正答率は.4 以上の正答率を示しており，スロベニアなどと同程度の正答率を示したと考えられる。IT 相関に関しては，項目 2（M041056）や項目 10（M041258B）は.066（95%CI[−.053, .181]）と.026（95%CI[−.119, .169]）であり，95%信頼区間が 0 を含み，全体の得点と項目に正答するか否かについて相関があるとはいえないということを示していた。他の項目の IT 相関は.2~.4 程度であり，若干低い値を示した項目が多くを占めていたと考えられる。

カタールデータ（表 5.7）はこれまでにみてきた国とは異なり，.1 を下回る正答率を示した項目が散見された。また，.1~.3 程度の正答率の項目も多くみられた。最も正答率が高かったのは項目 1（M041052）であり，値は.633（95%CI[.571, .695]）であった。以上のことから，カタールの解答者においては，このテストは困難度が高い項目が多いテストであったと考えられる。また，IT 相関においては推定値が負となる項目（項目 13（M041186），項目 14（M041336），項目 17（M031245））もみられた。さらに，.2 を下回る IT 相関を示した項目

も散見され，95%信頼区間が 0 を含む項目も多く見られた。最も高い IT 相関を示した項目としては，項目 16（M031309）であり，.344（95%CI[.282, .402]）という値を示した。これらの項目分析から，カタールの解答者に対しては TIMSS の問題項目が適切でなはなかった可能性が示唆された。

イエメンデータ（表 5.8）は，カタールデータと類似しており，比較的低い正答確率を示す項目が多い傾向が見られた。具体的には，項目 7（M041146），項目 10（M041258B），項目 12（M041275），項目 16（M031309），項目 21（M031247）などが，正答率を.1 を下回る項目であった。また，IT 相関は項目 10（M041258B）で.598（95%CI[.517,.661]）という値が最大であったものの，95%CI が 0 を含む項目も一定数みられ，また，項目 17（M031245）では−.111（95%CI[−.179, −.040]）というように IT 相関が負に有意となるという極端な項目もみられた。

表 5.9 に各国のα係数とその信頼区間を示した。表 5.9 から，アメリカ，香港，シンガポール，スロベニア，アルメニアにおいては，α係数が.8 程度あり，95%信頼区間を考慮しても.75 を上回っている可能性が高いと判断でき，TIMSS 2007 のテストは算数能力を測定しているテストとしての一定の信頼性があると考えられる。一方，カタールとイエメンデータにおけるα係数の推定値は .415 と .415（95%CI[.364,.467]）と .564（95%CI[.523,.606]）であり，かなり低い値であった。これらのことから，今回使用したカタールとイエメンのデータにおいては，TIMSS 2007 のテストは通常のテストとしての信頼性は必ずしも高いものとはいえない可能性が示唆された。

5-3-2　　モデル比較

表 5.10〜表 5.16 に各モデルの対数尤度のマイナス 2 倍（デビアンス），AIC, BIC, MADcor, SRMSR，項目パラメタ数を示した。各表において，各指標が示す最良のモデルのセルに灰色をつけた。注意点として，スロベニア（表 5.13）とカタール（表 5.15）において，3PL の推定で不適解が得られたが，参考値として表には各種指標を示した。また，MADcor と SRMSR を算出するための相関係数を計算するために完全にペアができる項目のみを用いた。カタールにおいては，M031247 の項目での項目反応がすべて 0 であったため，この項目は絶対指標の計算から除外した。G-DINA モデルは飽和モデルであるため，全ての国において最小のデビアンスを示すことが確認された。

AIC と BIC はシンガポール（表 5.12）を除くほとんどの国で同様の傾向を示した。さら

に加えて，全ての国において CDM は 1-3PL IRT モデルよりもよい適合を示した。これは，IRT モデルよりも，CDM の適合がよいという Lee et al.（2011）での結果を再現していることを示している。

アメリカ（表 5.10）とスロベニア（表 5.13）においては非補償モデルかつ主効果モデルである R-RUM が最もよい適合を示した。他の 5 つの国では補償モデルかつ主効果モデルが最も適合が良いモデルとして選択された。LLM は香港（表 5.11），カタール（表 5.15），そしてイエメン（表 5.16）において最も適合したモデルであり，A-CDM はシンガポール（表 5.12），アルメニア（表 5.14）において最も適合したモデルであった。シンガポールは AIC の観点からは G-DINA モデルが支持されていたものの，2 番目に良い当てはまりを示した A-CDM との AIC の差は 1 未満であり，情報量規準の間に大きな違いはみられなかった。

絶対適合度指標においては，日本データと同様にすべての国で 3PL モデルの適合が最も良いことが示された。また，CDM では，ほとんどの場合 G-DINA モデルが最適なモデルとして選択された。IRT モデルが絶対適合指標においてよい値を示したのは，日本人データと同様の傾向であった。ただし，CDM 内での比較では日本人データで A-CDM の適合が良かったということと異なった結果が得られた。

さらに，BIC の観点から最も適合が良かったモデルを用いて，各国での習得されたアトリビュート数の推定値の要約統計量を表 5.17 に示した。表 5.17 の結果は EAP 推定にもとづいたものであり，各解答者の各アトリビュートについて事後平均習得確率が.5 を上回ったアトリビュートを習得とみなした。香港とシンガポールの解答者は平均的にみて数アトリビュートのうち 6 つ，図形と測定アトリビュートのうち 3 つ，資料の表現アトリビュートのうち 2 つあるいは 3 つのアトリビュートを習得していた。このように，香港およびシンガポールの解答者は数，図形と測定，資料の表現の 3 つのそれぞれの領域においてわずか 1 つか 2 つのアトリビュートのみ未習得であるということができる。スロベニアとアルメニアは，香港とシンガポールよりもアトリビュートの習得数が少ないものの，カタールとイエメンよりも多いアトリビュートを習得していた。スロベニアとアルメニアの結果は，アメリカの結果に類似していた。カタールとイエメンはアトリビュートの各領域において平均的にみてわずか 1 つか 2 つのアトリビュートのみ習得していた。さらに，推定されたアトリビュート習得数の中央値から，カタールとアルメニアのほとんどの解答者はアトリビュートの各領域のうち 1 つのみ習得あるいは 1 つもアトリビュートを習得していないとわかる。以上より，これらの結果はアトリビュート習得パタンが国によって顕著に異なっているこ

とと，推定された習得アトリビュートは算数能力に関係していることを示唆していると考えられる。さらに，次節ではより具体的なアトリビュート習得パタンを検討し，各国での習得パタンの異同の結果を示す。

5-3-3　アトリビュート習得パタンの比較

本節では，前節で示された各国の各アトリビュート領域でのアトリビュート習得パタンの相違をより具体的に比較することによって，国ごとの解答者の認知状態について考察するための一助を得ることを目指す。表 5.18〜表 5.24 に各国の各学習領域におけるアトリビュート習得パタンのうち，上位 10 位の習得パタンを示した。ただし，資料の活用領域は合計で 8 つの習得パタンしか存在しないため，全てのパタンを示した。表 5.18〜表 5.24 では，列に 3 つのアトリビュートの領域（数，図形と測定，資料の表現）を示した。各アトリビュート領域では，アトリビュート習得パタン，その習得パタンを示した人数，および割合を示した。第 4 章と同様にこれらの習得パタンを算出する際には，15 個すべてのアトリビュートのパタンを同時に推定したのち，それぞれのアトリビュートの領域ごとに人数を集計した。

アメリカデータにおけるアトリビュート習得パタンの結果を表 5.18 に示した。アメリカデータにおいては，数領域のアトリビュートパタンとして，全てのアトリビュートを習得しているパタン，「数 6」アトリビュートのみを習得しているパタン，数 1〜3，6，8 のアトリビュートを習得しているパタンの 3 つが上位を占めていた。また，10 位までの習得パタンのうち半数以上は数領域のアトリビュートの半分以下のアトリビュートしか習得していないパタンが見られた。また，数領域のアトリビュートの多くを習得している解答者とそうでない解答者の 2 つの極端なパタンが比較的多いことが示された。図形と測定領域においては，上位 4 位まではアトリビュート習得数が 2 つ以下の習得パタンしかなく，アメリカの解答者は図形領域が得意ではない可能性が示唆された。また，図形と測定領域のうち「図形と測定 11，12」のアトリビュートを習得しているパタンが上位に位置しており，「図形と測定 9，10」の 2 つよりも相対的にみて習得されやすい可能性が示唆された。資料の活用領域に関しては，3 つのアトリビュートを習得しているパタンが 1 位であり，資料を活用することについては比較的得意とする解答者が多い可能性が示された。また，「資料の活用 15」のアトリビュートを習得しているパタンが上位 4 位を占めており，このアトリビュートの習得が容易である可能性を示唆している。また，全体的としては，数領域は第 10 位までのア

トリビュート習得パタンに含まれる解答者の割合は.430 であり，他の習得パタンに分類される解答者が他にも多い傾向がみられた。図形と測定領域では，第 10 位までの習得パタンに含まれる解答者の割合は.865 であった。資料と活用に関しては上位 3 つのパタンに 8 割以上の解答者が含まれることが示された。

香港データ（表 5.19）においては，全体的に多くのアトリビュートを習得しているパタンが多く見られた。数領域においては，上位 10 のパタンは全て 4 つ以上のアトリビュートを習得しているパタンであった。特に，数領域の上位 3 位は全てのアトリビュートを習得しているか，1 つのアトリビュートのみ未習得であるパタンであり，これらのパタンには 35%以上の解答者が含まれることが示された。これは香港データの学力の高さを反映していることとも整合する結果と考えることができる。図形と測定領域においては，上位 5 位までのパタンは 2 つ以上のアトリビュートを習得しているパタンであった。さらに，1 位から 3 位までのパタンのみに半数以上の解答者が含まれるという事が示された。香港データにおいては，上位 3 つの習得パタンから「図形と測定 10，11」のアトリビュートが比較的学習されにいくい傾向があると考えられる。資料の活用領域においても，上位 3 位のパタンに 8 割の解答者が含まれ，それらのパタンは 2 つ以上のアトリビュートを習得しているものであった。ただし，「資料の活用 13」を未習得であるパタンが 3 位から 5 位を占めており，四分の一の解答者がこのアトリビュートを習得していない傾向がみられた。

シンガポールデータ（表 5.20）の結果は，香港の結果に類似しており，数領域，図形と測定領域，資料の活用領域の全ての領域において第 1 位の習得パタンは全てのアトリビュートを習得しているパタンであった。数領域については，第 10 位が 1 つもアトリビュートを習得していないパタンであった点を除けば，どの習得パタンも少なくとも 5 つ以上のアトリビュートを習得しているパタンであった。また，全ての数領域のアトリビュートを習得している解答者が 3 割以上存在している点も特徴的と考えられる。図形と測定領域においては，第 2〜4 位の習得パタンが「図形と測定 11」を未習得であるパタンであり，シンガポールデータではこのアトリビュートの習得が不十分であるという点が明らかとなった。これに加えて，図形と測定領域の第 3〜9 位のパタンは「図形と測定 9」のアトリビュートが未習得であるパタンであり，このアトリビュートも学習が必要である可能性が示唆された。資料と表現領域は香港データにおけるものとほぼ同様であったものの，シンガポールデータにおいては，第 1 位のパタンであると推定された解答者の割合が 6 割近く，かなり多くの解答者が資料の読み取りやグラフの活用ができているということが示唆された。

スロベニアデータ（表 5.21）では，アトリビュート習得パタンの傾向はややアメリカに類似していた。具体的には，数領域において多くのアトリビュートを習得しているパタンとわずか数個のアトリビュートのみ習得しているという極端なパタンが上位 10 位までのなかに混在している点が類似している。ただし，アメリカとは習得パタンの種類やランキングに違いがみられる。数領域では，スロベニアは 1，2，3，5，8 位のパタンにおいてアトリビュート習得数が多いことを示し，アメリカは 1，3，4，8 位のアトリビュート習得パタンにおいて習得アトリビュート数の多いことを示した。図形と測定領域の上位 3 位のアトリビュート習得数はアメリカよりもスロベニアの方が多い結果が示された。スロベニアの図形と測定領域の特徴としては，「図形と測定 9」のアトリビュートを未習得であるパタンが 1 位と 2 位である点が挙げられる。資料の活用では，スロベニアの 1 位は全てのアトリビュートを習得しているパタンであったものの，2 位は「資料の活用 15」のみを習得しているパタン，3 位は「資料の表現 13」のみ習得していないパタンであり，「資料の表現 13」を未習得である解答者が多い傾向が示されたと考えられる。

アルメニアデータ（表 5.22）では，数領域のパタンでは第 2 位と第 6 位のパタンはアトリビュート習得数が少ないアトリビュートであったものの，それ以外のパタンは未習得のアトリビュートが 3 つ以下のパタンであった。また，数領機では「数 4，6，7」のアトリビュートが未習得であるパタンが上位 10 位に含まれており，この点がアルメニアデータでのアトリビュート習得パタンの特徴であると考えられる。図形と測定領域と資料の活用領域の上位 3 位の習得パタンはスロベニアデータのものと類似していた。

カタールデータ（表 5.23）のアトリビュート習得パタンはこれまでにみてきた国とは大幅に異なっており，各アトリビュート領域においてアトリビュート習得数が非常に少ない傾向がみられた。それぞれのアトリビュート領域のパタンを概観する。数領域においては，3 つ以下のアトリビュート数しかないパタンで構成されていた。さらに，上位 10 位までのアトリビュート習得パタンには「数 5, 6, 7」を習得しているパタンは含まれていなかった。ここから，「数 5，6，7」を未習得である解答者が 6 割以上いるということが示唆された。図形と測定領域においても 1 位から 4 位までの習得パタンは 1 つもアトリビュートを習得していないパタンと，「図形と測定 10，11，12」のうちどれか 1 つのみアトリビュートを習得している 3 つのパタンのみで構成されていた。さらに，「図形と測定 9」を習得しているパタンは 9 位まで出現せず，8 割以上の解答者がこのアトリビュートを習得していないという実態が明らかとなった。資料の活用領域では，3 割以上の解答者がどのアトリビュートも

習得していないという状況も示唆された。以上のことから，カタールにおいては，TIMSS の公式スコアと関連してアトリビュート習得数が少ない解答者が多く，かなり多くの解答者の算数能力の習得が不十分であることが示された。

イエメンデータ（表 5.24）では，カタールと類似して非常に少ないアトリビュート数のみが習得されていることが示唆された。数領域については，「数 1」のアトリビュートは比較的習得されている傾向がみられたが，「数 8」のアトリビュートを習得しているパタンは 8 位であり，かなり少ない傾向が見られた。イエメンにおける図形と測定領域や資料の活用領域のアトリビュート習得パタンは，1〜3 位ではかなり類似したものと考えられる。

ここまでみてきたように，国によってどのような習得パタンが多いのか，また特定のアトリビュートのうちどのようなアトリビュートを習得済みの解答者が多いのかといったことは国ごとに特徴があり，類似した国もあるということが示されたと考えられる。

本節では，アトリビュートの習得パタンの違いを具体的に検討してきた。TIMSS の公式到達度スコアで学力の水準が比較的近い国であっても，アトリビュートの習得パタンを精査することによって単純な学力の高低では明らかにならなかった側面が示されたといえる。しかし，全体的なアトリビュートの習得数は TIMSS 公式の到達度スコアが高い国ほど多い傾向が見られた。次節では，全体的なアトリビュート習得数に関しての分析のみならず，個別のアトリビュートを習得している確率を，1 次元の潜在特性がどれほど予測するのかについて詳細に検討を行う。さらに，1 次元の潜在特性との関係や TIMSS の公式のスコアとの関係に関しても次節でより具体的に検討を行う。

5-3-4　CDM の結果と IRT モデルの相関

本項では，アトリビュートの相関を確認するために，3 つの分析を行った。第 1 に推定されたアトリビュート習得確率を従属変数とし，2PL モデルで推定した潜在特性を独立変数としたロジスティック回帰分析を各国のデータについて行った。第 2 に，各国の各アトリビュート領域と全てのアトリビュートの合計アトリビュート習得数と IRT モデルで推定された特性値との相関係数を算出した。第 3 に，アトリビュート領域の合計アトリビュート習得数が TIMSS の公式スコアとどれほどの相関を持つのかを検討した。これらの分析を行うことにより，アトリビュートの習得確率のみならずアトリビュートを習得している個数がどれほど 1 次元的な能力と対応関係を持つのかを検討した。

もしも，IRT モデルで測定された算数能力が CDM のアトリビュートに関する結果と相

関しないならば，CDM で仮定したアトリビュートが算数能力を反映しているとはいえない可能性が考えられる。この点を確認するために，IRT モデルで推定した潜在特性とアトリビュートとの関連を調査する。アトリビュート習得パタンの推定には，AIC よりも強いペナルティを課す BIC の観点で各国の最も適合が良いモデルを用いた。

結果を表 5.25～表 5.31 に示した。また，各国で，アトリビュートごとに横軸に 1 次元潜在特性をとり，縦軸にアトリビュート習得確率をとった散布図に推定されたロジスティック曲線を同時に描画した図を作成した（図 5.1～図 5.7）。例えば, von Davier（2008）でも同様の図を作成している。この分析により，アトリビュート習得確率が全体的な潜在特性をどれほど反映しているのかを検討することが可能である。表 5.25～表 5.31 では，第 4 章と同様に，従属変数のアトリビュートごとに切片と回帰係数の推定値（$\hat{\beta}$と表記）および 95%信頼区間（標準誤差の 1.96 倍を用いて算出），オッズ比，オッズ比の 95%信頼区間，p値，R^2値を示した。R^2値の算出方法なども第 4 章にと同様の算出方法を採用した。

アメリカデータ（表 5.25，図 5.1）の数領域では.2 以上のR^2値をとっているアトリビュートが多く，とくに「数 2」は$\hat{\beta} = 4.128\ (95\%\mathrm{CI}[3.623, 4.632],\ p < .001), R^2 = .659$という値を示していた。このことは，図 5.1 でもロジスティック回帰曲線がアトリビュート習得確率によく適合している様子が観察できる。しかし，「数 6」は$\hat{\beta} = 1.073(95\%\mathrm{CI}[0.884, 1.262],\ p < .001), R^2 = .193$や「数 8」は$\hat{\beta} = 0.990(95\%\mathrm{CI}[0.815, 1.164],\ p < .001), R^2 = .104$と相対的に低い$R^2$値であり，図 5.1 からもロジスティック回帰曲線が推定されたアトリビュート習得確率にあまり適合していない様子が観察された。アメリカデータにおける図形と測定領域のアトリビュートは数領域のアトリビュートよりも全般的に低いR^2値を示しており，「図形と測定 12」において$\hat{\beta} = 1.005(95\%\mathrm{CI}[0.836, 1.175],\ p < .001), R^2 = .188$が最大の値であった。一方，「図形と測定 9」は$\hat{\beta} = 0.463(95\%\mathrm{CI}[0.313,0.613],\ p < .001), R^2 = .048$，「図形と測定 11」は$R^2 = .011$というかなり低い値を示した。図 5.1 でもその様子が観察され，とくに，「図形と測定 11」が一次元の潜在特性を反映していない可能性を示唆した。資料の表現領域は「資料の表現 13, 14」において.227 以上のR^2値を示し，比較的高い適合を示したと考えられる。ただし，「資料の表現 15」は$\hat{\beta} = 0.756(95\%\mathrm{CI}[0.571, 0.940],\ p < .001), R^2 = .106$であり，資料の表現領域の中では最も低い値を示した。

香港データにおいて（表 5.26，図 5.2），数領域のアトリビュートは，アメリカデータと同様に.100 以上という比較的高いR^2値を示していた。しかしながら，「数 4」，「数 6」および「数 8」アトリビュートのR^2値はそれぞれ.020，.085，.052 であり，$\hat{\beta} =$

$0.311(95\%\mathrm{CI}[0.101, 0.522], p = .004)$と$\hat{\beta} = 0.666(95\%\mathrm{CI}[0.426, 0.906], p < .001)$および$\hat{\beta} = .558(95\%\mathrm{CI}[0.326, 0.791], p < .001)$であり有意ではあるものの係数自体は信頼区間を考慮しても小さい値であった。図形と測定領域においては，「図形と測定 9」の$\hat{\beta} = 0.326(95\%\mathrm{CI}[0.097,0.555], p = .005), R^2 = .021$で香港データの中では最小の$R^2$値を示した。資料の表現領域では，どのアトリビュートにおいても，$.200 > R^2 > .100$であり，若干低めながらも一定程度の適合を示したと考えられる。また，図 5.2 から，適合の悪いアトリビュート（例えば「図形と測定 12」など）は潜在特性の値が−2 など低い値であっても習得確率が.30 を超えるようなロジスティック曲線を描いており，全般的に習得が容易なアトリビュートであった可能性が示唆される。

シンガポールデータ（表 5.27，図 5.3）においては，数領域において，どのアトリビュートでも$R^2 > .100$であった。最も低いR^2で値を示したのは「数 4」アトリビュートであるが，$\hat{\beta} = 0.728(95\%\mathrm{CI}[0.531,0.924], p < .001), R^2 = .111$であった。「数 2」アトリビュートは逆に非常に高い$R^2 = .847$という値を示しており，$\hat{\beta} = 7.412(95\%\mathrm{CI}[5.565, 9.259], p < .001)$という値を示した。この「数 2」アトリビュートの適合の良さは，図 5.3 からも明らかにわかる。図形と測定領域は，「図形と測定 11」のみが$\hat{\beta} = 0.685(95\%\mathrm{CI}[0.495, 0.875], p < .001), R^2 = .056$という.100 を下回った低い適合であった。一方，資料の表現領域はすべてのアトリビュートにおいて，$R^2 > .300$あり，潜在特性がアトリビュートの習得確率をよく予測していることが示された。

スロベニアデータについて（表 5.28，図 5.4），数領域においては全てのアトリビュートにおいて，$R^2 > .100$であり，最も低いR^2は「数 8」アトリビュートであり，その値は.198，$\hat{\beta} = 1.450(95\%\mathrm{CI}[1.176, 1.725], p < .001)$であった。また，数領域において最も高い適合を示したのは「数 2」であり，$\hat{\beta} = 3.357(95\%\mathrm{CI}[2.812, 3.903], p < .001), R^2 = .579$であった。図形と測定領域においては，「図形と測定 12」のみ$R^2 > .200$である一方，他の 3 つのアトリビュートでのR^2値はそれぞれ，.023（「図形と測定 11」）～.058（「図形と測定 10」）と低い値を示した。最後に資料の表現領域に関しては，「資料の表現 13」において$\hat{\beta} = 0.799(95\%\mathrm{CI}[0.585, 1.012], p < .001), R^2 = .118$であり，資料の表現領域の中では相対的に一次元潜在特性の説明力が高くないことが示された。一方で「資料の表現 14，15」では，$R^2 = .266$~$.261$という値を示した。図 5.4 からも，「図形と測定 13」などでは，推定されたアトリビュート習得確率と潜在特性の間の関数関係は不明瞭であることが読み取れる。

アルメニアデータでは（表 5.29，図 5.5），アメリカ・香港・シンガポール・スロベニア

と比較して$R^2 < .100$であるアトリビュートの個数が多い傾向が見られた。数領域において，「数 4」では$\hat{\beta} = 0.538\ (95\%\mathrm{CI}[0.322, 0.755], p < .001), R^2 = .059$，「数 5」では$\hat{\beta} = 0.012\ (95\%\mathrm{CI}[-0.266, 0.290], p = .933), R^2 = .000$，「数 6」では$\hat{\beta} = 0.514\ (95\%\mathrm{CI}[0.305, 0.723], p < .001), R^2 = .059$であり，かなり低い値もみられ，ロジスティック回帰曲線の適合の悪さが顕著にみられるアトリビュートも含まれている事が示された。図形と測定領域では，「図形と測定 9」において$\hat{\beta} = 0.152\ (95\%\mathrm{CI}[-0.054, 0.359], p = .147), R^2 = .005$であり，「数 5」，「数 6」と同様にモデルの適合が極めて悪いといえる。図 5.5 からも，これら極めて適合が悪いアトリビュートのロジスティック曲線はほぼx軸に並行の直線であり，アルメニアデータにおいては 1 次元の潜在特性の高低がアトリビュート習得確率を予測できないことがあることが明らかに見て取れる。「図形と測定 12」では$\hat{\beta} = 0.669\ (95\%\mathrm{CI}[0.421,\ 0.916], p < .001), R^2 = .085$であり，若干の適合の悪さが見られた。資料の表現領域では，「資料の表現 14」において$\hat{\beta} = 0.840\ (95\%\mathrm{CI}[0.615, 1.065], p < .001), R^2 = .090$,「資料の表現 15」において$\hat{\beta} = 0.850\ (95\%\mathrm{CI}[0.624, 1.077], p < .001), R^2 = .141$であり，やや低めではあるものの一定程度の適合を示したと考えられる。

カタールデータ（表 5.30，図 5.6）およびイエメンデータ（表 5.31，図 5.7）においては他の国と異なって，回帰係数が負に推定されるアトリビュートが見られたことが特徴的であった。また，いずれの国においてもR^2値は全般的に低めの値を示していた。カタールの数領域では，「数 1」において$\hat{\beta} = 1.662\ (95\%\mathrm{CI}[1.365,1.958], p < .001), R^2 = .235$であり，という最大の$R^2$であった。この他，「数 4」において$\hat{\beta} = -0.320\ (95\%\mathrm{CI}[-0.527, -0.114], p = .002), R^2 = .015$，負の値で有意に推定され，「数 7」において$\hat{\beta} = -0.338\ (95\%\mathrm{CI}[-0.683,0.007], p = .057), R^2 = .007$であり，回帰係数が負に推定されたものの有意ではなかった。また，「数 5」では$\hat{\beta} = 0.073\ (95\%\mathrm{CI}[-0.175,\ 0.320], p = .565), R^2 = .001$,「数 2」は$\hat{\beta} = 0.611\ (95\%\mathrm{CI}[0.405,\ 0.817], p < .001), R^2 = .033$という非常に低い値であった。図形と測定領域では，「図形と測定 9」は$R^2 > .100$であったものの，「図形と測定 11」と「図形と測定 12」はそれぞれ$\hat{\beta} = 0.334(95\%\mathrm{CI}[0.138,\ 0.531], p < .001), R^2 = .009$と$\hat{\beta} = 0.327(95\%\mathrm{CI}[0.134, 0.520], p < .001), R^2 = .015$という値を示した。資料の表現領域はすべてのアトリビュートにおいて，$R^2 < .100$であった。また，「資料の表現 13」は$\hat{\beta} = -0.203(95\%\mathrm{CI}[-0.398, -0.008],\ p = .041), R^2 = .006$と有意な負の値であった点も特徴的であった。図 5.6 では，潜在特性とアトリビュート習得確率の関係が不明瞭であるアトリビュートが多く，また，アトリビュート習得確率が特定の習得確率の周辺で固まっているもの

も視覚的に確認することができる。例えば「数 5」アトリビュートでは，.2～.3 周辺の習得確率と推定された解答者が多いという傾向が見られた。

最後のイエメンデータにおいては（表 5.31，図 5.7），数領域の「数 1」，「数 2」，「数 4」，「数 5」の 4 つのアトリビュートにおいて$R^2 < .100$であった。「数 2」アトリビュートはシンガポールなど他の国ではかなり高いR^2値を示しており，回帰係数も正の値であったが，イエメンにおいては$\hat{\beta} = -0.834\ (95\%\text{CI}[-1.114, -0.553],\ p < .001), R^2 = .048$であり，有意に負の値を示した。また，「数 7」は$R^2 = .290$という高い値を示したものの，やはり$\hat{\beta} = -2.858$（$95\%\text{CI}[-3.339, -2.377],\ p < .001$）で有意に負の値を示した。図形と測定領域については，「図形と測定 10」のみが$\hat{\beta} = 1.298\ (95\%\text{CI}[1.038, 1.559],\ p < .001), R^2 = .137$であったものの，他の「図形と測定 9」，「図形と測定 11」および「図形と測定 12」は$R^2 < .100$であり，特に「図形と測定 9」および「図形と測定 12」は非常に低いR^2値を示した。そして，資料の表現領域では，「資料の表現 15」のみ$\hat{\beta} = 1.215(95\%\text{CI}[0.919, 1.511],\ p < .001), R^2 = .122$であり$R^2 > .100$であったが，「資料の表現 13」および「資料の表現 14」は$R^2 < .100$であった。また，「資料の表現 13」は回帰係数の推定値が$\hat{\beta} = -0.070\ (95\%\text{CI}[-0.297, 0.157], p = .547), R^2 = .001$であり，負の値であったものの 5%水準で有意ではなかった。

アトリビュート習得確率と 2PL ロジスティックモデルで推定された 1 次元潜在特性のロジスティック分析の結果から，ほとんどの国において，アトリビュート習得確率は全般的な算数能力を反映していると考えるられる。しかし，中には 1 次元潜在特性によって習得確率を予測できないアトリビュートが存在したり，場合によっては回帰係数が負に推定されるアトリビュートも存在した。ただし，回帰係数が負に推定されたものはカタールデータかイエメンデータに限定されていた。R^2の値が.100 未満または回帰係数が負に推定されたアトリビュートは，アメリカでの「図形と測定 9」，「図形と測定 11」，香港の「数 4」，「数 6」，「数 8」，「図形と測定 9」，シンガポールの「図形と測定 11」，スロベニアの「図形と測定 9」，「図形と測定 10」，「図形と測定 11」，アルメニアの「数 4」，「数 5」，「数 6」，「図形と測定 9」，「図形と測定 12」，「資料の表現 14」，「資料の表現 15」，カタールの「数 2」，「数 3」，「数 4」，「数 5」，「数 6」，「数 7」，「図形と測定 10」，「図形と測定 11」，「図形と測定 12」，「資料の表現 13」，「資料の表現 14」，「資料の表現 15」，およびイエメンの「数 1」，「数 2」，「数 4」，「数 5」，「数 7」，「図形と測定 9」，「図形と測定 11」，「図形と測定 12」，「資料の表現 13」，「資料の表現 14」であった。アトリビュート単位で考えると，「数 1」や「数 2」はどの国

でも比較的一貫して高いR^2値を示しており，全般的な算数能力に対応するアトリビュートと結論できる。その一方で，「図形の測定 9」，「図形の測定 11」などのアトリビュートはどの国においても比較的低いR^2値を示しており，全般的な算数能力を部分的にしか反映していない可能性もある。また，国によって，全体的な算数能力がそれぞれのアトリビュート習得確率を説明できる程度は異なっていた。ただし，これらの結果を踏まえても，全体的にはアトリビュートの習得確率に対して，全般的な算数能力の回帰係数の推定値は正であるものが多く，個々のアトリビュートは算数能力を一定程度反映している一つのエビデンスが得られたと考えることができる。

アトリビュートと IRT モデルで推定された 1 次元潜在特性との関連の分析の第 2 番目として，それぞれの国において，最良の適合を示した CDM を用いてアトリビュート習得数と 2PL モデルによって推定された能力の間の相関係数を算出した。結果は表 5.32 に示した。全ての国において，全てのアトリビュート領域において中程度から強い相関係数を示した。カタールとイエメンにおいては，全ての領域において中程度の相関を示したものの資料の表現においてはやや低い相関係数であった。アメリカ，香港，シンガポール，スロベニア，アルメニアは全て数領域において強い相関を示したものの，図形と測定領域と資料の表現領域においては，.5 以上.8 以下の相関に留まった。

第 3 に各国の平均的なアトリビュート習得数と各国の TIMSS 2007 の公式到達度スコア（詳細な値は Martin et al., 2008 の Exhibit 1.1 をみよ）との相関係数を算出した。この結果は表 5.33 に示した。ここでのサンプルサイズは研究に用いた国の数であるので 7 となる。そのため，95%信頼区間も比較的広い。しかし，得られた相関係数は顕著に強いものであり，全ての推定値は.9 を越えていた。これは，推定されたアトリビュート数は一般的な算数能力を反映したよい指標になっていることを考えることができる。これら 2 つの分析結果は，本章で用いたアトリビュートが 1 次元的な能力を反映しているという証拠として解釈できる。

5-3-5　　項目パラメタの推定値

本章ではこれまで，モデル比較，各国のアトリビュート習得パタンの比較，アトリビュートの妥当性の検討を行った。最後にそれぞれの国で最も適合が良かったモデルの項目パラメタの推定値と標準誤差を示し解答者の解答の方略について考察する一助を得ることとする。項目パラメタの推定結果は表 5.34～表 5.40 に示した。R-RUM においては，GDINA パッケージが本来のモデルでのパラメリゼーションでの標準誤差の算出に未対応であるた

め，項目パラメタの推定値のみを記載した。また，数値計算の問題で R-RUM の項目パラメタの推定値で 1.00 を上回って推定されたものは 1.00 と表記した。LLM と A-CDM については G-DINA モデルのデルタパラメタリゼーションによる項目パラメタの推定値（R-RUM では$\hat{\pi}$, $\hat{r}$, その他モデルでは$\hat{\delta}$などで表現）と標準誤差（SE）を記した。R-RUM，LLM，A-CDM の何れのモデルも切片パラメタと各アトリビュートの影響の大きさを反映するパラメタからからなりたっており，Q 行列の要素が 1 である箇所に，推定値を記述した。また，表中の空欄は該当するパラメタが存在しないことを表している。

アメリカデータにおける R-RUM の項目パラメタの推定値（表 5.34）では，まず全体的に$\hat{\pi}$の推定値が 1.000 に近いものが多いことが挙げられる。つまり，仮定したアトリビュートを全て習得している場合には当該項目への正答確率が極めて高くなることを意味している。しかし，項目 10（M041258B）は$\hat{\pi}_{10} = .108$ときわめて低く，図形と測定領域の「図形と測定 9」,「図形と測定 10」のアトリビュートを習得していたとしても正答率が高くなく，そもそもの正答確率が低い可能性がある。また，$\hat{r}_{10,9} = .881$，$\hat{r}_{10,10} = .713$であり，いずれも比較的高い値である。R-RUM の項目パラメタ$\hat{r}$は当該アトリビュート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として解釈されるがその値が高いということはそのアトリビュートを習得していなくても，項目への正答確率の低下が低いといえる。つまり，そのアトリビュートは当該の項目に必要でなかった可能性を示唆している。また，項目 21（M031247）は$\hat{\pi}_{21} = .358$であり，項目 10（M041258B）についで低い値を示した。ただし，項目 21（M031247）のパラメタの推定値は$\hat{r}_{21,2} = .201$, $\hat{r}_{21,3} = .518, \hat{r}_{21,7} = 1.000$であり，「数 2」,「数 3」アトリビュートは項目の正答に必要なアトリビュートである一方，「数 7」アトリビュートは未習得であっても正答率に影響しない事が示唆された。これに加えて，Q 行列の設定としては複数のアトリビュートが必要であることが想定されている項目ではあるものの，$\hat{r}$は 1.00 に極めて近い値をとっており，実質的に一つのアトリビュートが当該項目の正答に重要である可能性がある項目が散見された。例えば，項目 1（M041052）では「数 1」と「数 2」アトリビュートが必要であると仮定されているものの，$\hat{r}_{11} = 1.000, \hat{r}_{12} = .670$であり，「数 2」アトリビュートのみが正答確率の変動に寄与している可能性が考えられる。他の項目においても，一つのアトリビュートのみが正答確率に関連している様子が散見された。一方で，複数の$\hat{r}$が 1.000 ではない項目も存在した。例えば，項目 5（M041281）において，$\hat{r}_{52} = .410, \hat{r}_{53} = .678$，$\hat{r}_{58} = 1.000$であり，「数 8」アトリビュートは正答に関連がないものの，「数 2」および「数 3」アトリビュートはいずれも罰則項として機能してい

たと考えられる。また，項目 5（M041281）の$\hat{\pi}_5 = .960$であることから，「数 2」のみ習得していない解答者の項目 5 の正答確率は.394，「数 3」のみ習得していない解答者の正答確率は.651，両方のアトリビュートを習得していない解答者の正答確率は，.267 となる。ただし，これらの結果は点推定値にのみ依拠した結果の解釈であり，推定値の標準誤差を考慮できていない点については注意する必要がある。

香港データにおける LLM の項目パラメタの推定値とともに標準誤差を表 5.35 に示した。項目パラメタの標準誤差に着目すると，非常に大きな値をとるパラメタが散見され，推定値が安定していない可能性がある。例えば項目 1（M041052）の「数 1」アトリビュートの$\hat{\delta}_{11} = 16.387$であり，$SE = 261.102$であった。これは，LLM がリンク関数として，logit 関数を選択しており，パラメタの取りうる範囲が$-\infty$から$+\infty$であることにも関連している可能性がある。結論として，香港データの項目パラメタの推定値は必ずしも安定していないと考えられるため，項目パラメタの具体的な値についての解釈には慎重になる必要があり，積極的な解釈は控えたほうが良い可能性が高い。後述のように，カタールのデータでも LLM が支持されたものの，香港データと同様にSEが非常に大きい項目が散見された。ただし，例えば項目 5（M041281）では$\hat{\delta}_{52} = 3.041$，$SE = 0.641$で他の項目よりも標準誤差の値が小さいものもみられた。

シンガポールデータに A-CDM を適用した場合の項目パラメタの推定値と標準誤差を表 5.36 に示した。シンガポールデータは，香港データとは異なり，標準誤差が極端な値を取ることはなかった。しかし，アメリカデータでみられたように項目パラメタが 1.000 や.000 といった極端な値を取る場合もみられた。例えば，項目 4（M041076）は$\hat{\delta}_{40} = \hat{\delta}_{43} = .000$であり，$\hat{\delta}_{46} = 1.000$，推定値が極端なパラメタであったといえる。また，標準誤差に注目すると，低いもので項目 24（M031085）の切片パラメタの *SE* が.023，高いもので項目 12 の「資料の表現 15」のパラメタの *SE* が.338 であった。近似的な 95%信頼区間を考える場合には，この値を 2 倍したものが信頼区間の半幅に対応し，$.023 \times 2 = .026$と$.338 \times 2 = .676$といった値が得られる。これらのことから，シンガポールデータは香港データの推定値と比較して安定した推定値を得られているように見える反面，具体的な変動の幅は直観に反して小さいとはいえないものと考えられる。こういった標準誤差や信頼区間を具体的に考慮することは，結果の解釈を行ううえでも必要な観点と考えられる。

スロベニアはアメリカデータと同様に R-RUM の適合がよい国であった。スロベニアデータに R-RUM を適用したときの項目パラメタの推定値を表 5.37 に示した。スロベニアデ

ータでの項目パラメタの推定値の高低のパタンはアメリカの推定値に類似した項目が多い傾向がみられた。例えば，項目 22（M031219）の項目パラメタは「図形と測定 11」のパラメタが非常に低く，その他の項目パラメタの推定値は 1.000 に非常に近いものであった。その一方，局所的に見ると，スロベニアデータとアメリカデータでかなり異なった推定値が得られた項目もみられた。例えば，項目 1（M041052）について，スロベニアでは$\hat{\pi}_1 = 1.000, \hat{r}_{11} = .178, \hat{r}_{12} = 1.000$，アメリカでは，$\hat{\pi}_1 = .919, \hat{r}_{11} = 1.000, \hat{r}_{12} = .670$という推定値が得られた。これは，スロベニアデータでは項目 1 について，「数 1」のアトリビュートが問題の正答に重要であることを示唆する一方，アメリカデータでは「数 2」のアトリビュートが重要であることを示唆しており，問題の正答に必要とされるアトリビュートが国によって異なっている可能性を示唆する結果と考えることができる。このような二カ国の間の相違は項目 3（M041069）にもみられ，そもそも切片パラメタの推定値からして大きな相違がみられた。このような細かい点での国による推定値の違いは一考の余地があるだろう。

アルメニアは A-CDM が支持された国であった（表 5.38）。アルメニアにおいても他の国と同様に，項目の正答に必要なアトリビュートのうち 1 つのアトリビュートのみの推定値が大きい項目がみられた。例えば，項目 12（M041275）の「資料の表現 15」は$\hat{\delta}_{12,14} = 1.000$であった。一方，複数のアトリビュートが項目の正答確率に個別に高い影響を持つ項目がみられた。例えば，項目 2（M041056）は「数 1」か「数 5」の 2 つのアトリビュートのうちどちらかが必要であれば正答できることを示唆しており，1 つの項目に正答に特定のアトリビュートを習得していれば正答できることが示唆されている。これは，複数の解答ストラテジがある現象が存在することを示唆している。また，複数のアトリビュートが項目 15（M031303）は「数 1，2，3」のアトリビュートが必要な項目であり，パラメタの推定値はそれぞれ$\hat{\delta}_{15,0} = .395, \hat{\delta}_{15,1} = .266, \hat{\delta}_{15,2} = .309, \hat{\delta}_{15,3} = .266$である。これは，項目 15（M031303）が必要なアトリビュートを習得しているほど，段階的に習得確率が高くなる項目であると解釈できる。また，項目パラメタの推定値が負に推定されたものもみられた（例えば，項目 3（M041069），「数 4」アトリビュートの$\hat{\delta}_{34} = -.009$など）。しかし，標準誤差を考慮すればこれは推定の誤差と考えられる。

カタールデータでは LLM が支持された（表 5.39）。前述の香港と同様に標準誤差の推定値が非常に大きい値である場合が散見された。項目パラメタの推定値の標準誤差が大きい中でも比較的安定した標準誤差を示したものとしては，項目 8（M041152）がある。項目 8（M041152）では，「数 3」が最も高い推定値を示した（$\hat{\delta}_{83} = 2.215$）ものの，「図形と測定

10」は負の推定値を示した（$\hat{\delta}_{8,10} = -.994$）。このようにカタールデータでは項目パラメタが負に推定される，つまりアトリビュートを習得している場合にはむしろ正答率が下がるという場合が散見された点も特徴と考えられる。ただし，これらの推定値は標準誤差を 1.96 倍した近似的な信頼区間が 0 を含む場合も多く，推定の誤差である可能性も考えられる。

イエメンデータにおいても香港，カタールと同様に LLM が支持された（表 5.40）。イエメンデータにおいても，項目パラメタの標準誤差は非常に大きい場合が散見された。3 カ国に渡って LLM の項目パラメタの推定値の標準誤差が大きいことから，少なくとも今回の TIMSS を分析するためのモデルとして LLM は項目パラメタの推定値が安定しないモデルである可能性がある。また，LLM がテスト分析では推定が安定しにくいモデルである可能性が示唆される。このようなことから，実データで LLM の項目パラメタの推定を安定させる方法論が必要であるといえる。

項目パラメタの推定値についての結果をまとめると，次のようになる。まず，TIMSS データは Lee et al.（2011）が想定していたように複数のアトリビュートが項目の正答に必要であるという仮定は必ずしも正しくない可能性がある。これは，特定のアトリビュートの項目パラメタのみが大きく項目の正答確率に寄与している点から示唆された。その一方で，項目によっては，複数のアトリビュートを習得していれば習得しているほど，正答率が高くなる項目が存在することも明らかになり，項目によっては複数のアトリビュートが必要となることが示唆された。また，複数のアトリビュートが別々に項目の正答に寄与している場合もみられ，いわゆる別解や複数のアトリビュートのどちらを使っても正答できる場合があることが示された。上記の結果から，Lee et al.（2011）の用いたアトリビュートについては，部分的には正しい可能性があるものの，一方で過剰に不要なアトリビュートを付与している可能性が高い可能性が考えられる。アトリビュートの設定の妥当性については，第 6 章で再度考察を行う。

国別の結果の違いに関して，適用しているモデルが必ずしも同じではないため直接的な比較が難しいと考えられる。しかし，TIMSS データに適合したモデルは全て主効果モデルであるため，どのアトリビュートが項目の正答に寄与しているのかという観点においては比較が可能であると考えられる。この点に注目してみると，国によって項目パラメタの推定値が類似している部分と，異なっている部分がみられることがわかる。こういった国の間での共通する点と異なっている点については，より詳細な検討が必要である。

モデルの設定について，A-CDM と LLM は類似したパラメタリゼーションであるものの，

項目パラメタの安定性の観点からいえば，A-CDM の方が項目パラメタの推定値が安定している可能性があることが示された。ただし，これはこのデータに特有の現象である可能性もあり，必ずしも一般化できることではないものの，応用では注意が必要となるだろう。また，標準誤差に注目すると一見小さい値を取っているものでも，95%信頼区間としては比較的広いものも多くみられた。推定値が安定しないということは，サンプルによって結果が大きく変動しうることを示唆しており，結果の解釈に注意が必要な点のうちの一つであると考えられる。

項目パラメタの詳細を検討することによって，実データでのモデルの挙動の一側面を検討した。この結果から，Lee et al.（2011）の設定では項目パラメタの推定が安定してない可能性が示唆された。具体的には，アトリビュート数が 15 あるにも関わらず，サンプルサイズが数百から多くて 1000 程度という設定であったため，サンプルの変動による影響がかなりある可能性が示唆された。つまりどの国においても，想定される習得パタンのほうがサンプルサイズよりも大きく，必然的に特定のアトリビュート習得パタンに属する解答者がいないという状況が生じうる。また，一般にアトリビュート間には相関が生じるため，習得パタンは一様に分布しているわけではなく，偏りが生じていると考えるのが自然である。こうしたことから，特定のアトリビュート習得パタンに解答者がほとんどいないことによって，項目パラメタの推定値の標準誤差が大きいものも生じていた可能性も考えられる。項目パラメタの標準誤差が極めて大きいものについては，注意が必要であり，点推定値と合わせて考察の対象としていくことが望ましい。これらの結果から得られる重要な示唆としては，アトリビュート数の設定は慎重に行うべきであり，また，アトリビュート数が多い場合にはかなりのサンプルサイズが必要になることに留意しなければならない。また，第 2 章で述べたように，必要なアトリビュートをどのように設定するか（Q 行列の設定）も重要な問題として解決が望まれる。

5-4 考察

本章の考察においては，主要な目的であるモデル比較の結果をまず取り上げ，つぎにアトリビュート習得パタン，アトリビュートと IRT モデルの潜在特性の関係について考察を行い，最後に項目パラメタの推定値についての考察を行う。

5-4-1　　モデル比較について

TIMSS 2007 の算数調査の本章の知見は本章の 3 つの目的に照らして，次のようにまとめられる。第 1 に，CDM は本章で用いた全ての国において IRT モデルよりもよい適合を示した。第 2 に，全ての国でアトリビュートが個別に正答確率に寄与する主効果モデルの適合がよい事が示唆された。第 3 に，最も適合がよい CDM は国によって若干のばらつきがみられたものの，DINA モデルや DINO モデルといった古典的で比較的パラメタ数が少ない倹約モデルや，検討に用いたモデルの中では最も複雑な飽和モデルである G-DINA モデルは最も適合が良いモデルとしては選択されなかった。

まず，はじめの主要な知見はアメリカデータで実証された Lee et al.（2011）の知見に一致しており，Lee et al.（2011）の主張を一般化する証拠が得られたといえる。TIMSS において，IRT モデルよりも CDM がよく適合したという事実は，解答者が問題の解答に 1 つ以上のアトリビュートを必要としているということであり，TIMSS が算数の細やかな能力をアセスメントするテストとして有用であるということを示唆している。一次元 IRT モデルは，ハイステークスなテストにおいて解答者の序列を決定するのに有用かもしれないが，それらは解答者の問題解決プロセスを説明するには単純すぎると考えられる。IRT モデルと異なって，CDM は TIMSS の目的にもとづいた解答者の項目反応行動を反映しているモデルである。

もちろん，CDM と IRT モデルのどちらを選択するのかはテストの目的やデータ解析の目的に依存している。例えば，一次元 IRT モデルは，大規模コンピュータベーストアダプティブテストの開発時など，診断ではなく特定の 1 次元の特性にもとづいて解答者の序列を決定する目的にはより適しているかもしれない。しかしながら，本章では CDM が TIMSS データでは IRT モデルよりもよい適合を示した。CDM の適用が目的に適っているのであれば，CDM は解答者についての診断情報を含めてかなり多くの情報を抽出するのに役立つといえる。

本章の第 2 の知見は，主効果モデルがどの国においても，選択されたという点である。国際的なデータを用いたモデル比較からは，TIMSS 2007 の 4 年生の算数データにおいては，それぞれの項目の正答に必要なアトリビュートのうち，問題の解答に際して全てのアトリビュートを利用することを求めているわけではない可能性が明らかとなった。言い換えれば，ほとんどの国の解答者は問題に解答するために知識を同時に利用しているというよりは選択的に利用していると考えられる。あるいは，TIMSS の解答者はそれぞれのアトリビ

ュートを個別に適用することで問題に解答している可能性が考えられる。本章においては，TIMSS の到達度スコアの観点から幅広く国を選択している。このため，主効果モデルが他の倹約モデルや飽和モデルよりもよいという本章の知見は他の国においても一般化できる可能性を有している。さらなる研究によって，この点に関して明らかになると考えられる。

主効果モデルが相対的に適合したという結果が得られた理由のうちの一つに Q 行列の設定が考えられる。それぞれのモデルについて，どのようなアトリビュートがどのようにそれぞれの項目に影響を持ちうるのかを規定しているのは Q 行列である。本研究では Lee et al.（2011）で用いられた設定を利用して分析を行い，Q 行列も先行研究によって定義されたものを利用した。第 4 章において本研究で利用した Q 行列を示したが，殆どの項目でアトリビュートが複数必要であり，そういった項目のアトリビュートの個数は 2〜6 であった。DINA モデルは項目の正答に必要なアトリビュートをすべて習得している場合に正答率が上昇するモデルであり，必要なアトリビュートが多くなると適合が悪くなる可能性がある。また，DINO モデルも必要なアトリビュートを一つ以上習得している場合に習得パタンが上昇するモデルであり，複数必要なアトリビュートのうちどれか一つだけ習得していればよいという設定には無理があったと可能性が考えられる。

本章の 3 番目の知見は DINA モデルや DINO モデルと呼ばれる比較的古い倹約的な CDM や，飽和モデルは TIMSS 2007 のテストではどの国においても最適な適合を示さなかったということである。この結果が得られた理由としては，DINA モデルと DINO モデルは今回検討した CDM の中で最も制約の強いモデルであったことが原因である可能性がある。つまり，DINA モデルや DINO モデルは応用するためには制限が強すぎ，解答者の仮定されたアトリビュートや項目解答プロセスとは乖離があることが原因と考えられる。また，解答者は TIMSS の問題の正しい答えに到達するために，ある種の認知スキルの一部を部分的に適用することのみで十分であった可能性がある。しかし，DINA モデルではこうしたアトリビュートの部分的適用を仮定していない。このような DINA モデルや DINO モデルが他のモデルよりも悪い適合であったという知見は，より制約の少ない CDM がモデル開発者にとって重要であるということを示唆している。モデル開発研究の多くは DINA モデルを基本として発展してきた。なぜならば，DINA モデルは最初期に開発された確率的 CDM のうちの 1 つであり，問題に正答するためには複数のアトリビュートが全て必要であるという比較的わかりやすい設定であったためであると考えられる。しかし，本研究の結果からは，DINA モデルが解答者の知識状態を反映するには制約が厳しすぎる可能性がある。このことに鑑

みて，将来のモデル開発研究は実データへの経験的なモデル適合をより重視して，解答者の実情に即したモデルの開発が望まれる。第 5 章で示された項目パラメタからは，各アトリビュートが正答率に寄与する程度がかなり異なっているということが示唆された。このことからも，DINA モデルや DINO モデルの制約が非常に強く，モデルがデータに適合しにくかった可能性が考えられる。

G-DINA モデルの適合も良くなかったことから，TIMSS 2007 データにおいては，G-DINA モデルの特徴である交互作用効果が過剰なパラメタであり不要であった可能性が考えられる。Lee et al.（2011）の設定では，当初は DINA モデルを適用することを念頭にアトリビュートの設定や Q 行列の設定を行ったものの，本研究の結果から TIMSS の算数データではアトリビュートを問題解決に同時に適用するというメカニズムにはなっていないことが示唆される。また，Lee et al.（2011）の用いた Q 行列では 1 つの項目にかなり多くのアトリビュートを付与していることもあり，1 つの項目に含まれる交互作用パラメタの数が多くなり過ぎた可能性もある。今回定義されたアトリビュートは単純な計算ルールというより複雑な認知プロセスを想定したものであり，アトリビュート 1 つで交互作用も含めた効果を表現してしまっている可能性も考えられる。何れにせよ，今回の TIMSS データにおいては，複雑な交互作用効果が問題の正答に必要というより，主効果モデルのように個別にアトリビュートを適用することで十分正答が可能であったと考えられる。

ただし，本章のモデル適合の結果を一般化する際には留意が必要である。なぜならば，TIMSS は解答者の到達度を測定するためのテストであり，そもそも診断目的に作られていないからである。このため，本章で用いた Q 行列の設定では DINA モデルや DINO モデルに対して不利に働いた可能性がある。もし，テストの設計自体を DINA モデルや DINO モデルでの分析を前提に作成していたとすれば，モデル適合の結果は異なっていた可能性もある。テストデータの分析はテストの目的に準拠したものであることが望ましく，TIMSS では IRT モデルを用いることが TIMSS の目的として適しているといえる。これは，絶対適合指標が IRT モデルにおいて最小となった結果からも支持されると考えられる。

しかし，DINA モデルや DINO モデルが適合しない事自体はそれらのモデルが有用でないことを意味しない。本章のようにアドホックに CDM を適用する場合には他の制約の緩やかなモデルが適していたということを示していたと考えられる。また，後述のように今回利用した実データは TIMSS に限定されているため，必ずしも全てのテストにおいて常に主効果モデルが適しているということを意味しているわけではない点に注意が必要である。例

えば，診断目的を事前に立てて，DINA モデルの項目反応関数が適合するように項目を作成するという状況もありうる。その場合にはもちろん DINA モデルが最も適合するモデルとして選択されることが期待される。しかし，実際には本研究のように既存のテストに CDM を適用するという状況も多いため，そのような状況でどのようなモデルが有用であるのかを知ることには一定の意味があると考えられる。

また，TIMSS が診断を志向して作成されていないということは本章の結果の意義を損なうものではない。本章からは，CDM の適合も相対的に良いことが示唆されており，また通常のテストデータ分析では得ることのできないアトリビュート習得パタンを得ることもできている。また，アトリビュートの習得数は IRT モデルで推定された潜在特性と非常に高い相関を示すものであり，CDM で得られた結果は IRT で明らかにできる側面も含んでいると考えられる。もちろん，IRT モデルで得られる潜在特性と CDM で得られるアトリビュート習得数は性質が異なった統計量であり同一視はできない点は注意が必要である。しかし，CDM を 1 次元性の高いテストに適用したとしても，フィードバックに有用な情報を抽出できることを示すことができたという点には意義があると考えられる。

これら 3 つの知見は全て TIMSS 2007 の算数テストの結果に依拠しているものの，これらの結果は TIMSS 以外のデータ以外にも当てはまると考えられる。教育的な診断アセスメント（e.g., Lee et al., 2011; Choi et al., 2015; Sun, Suzuki, & Kakinuma, 2012）においては，DINA モデルが適用された例は必ずしも多くなく，このことは DINA モデルの適用の難しさに起因している可能性もある。また，例えば，Sun et al.（2012）は小学校 6 年生の分数のテストを DINA モデルにより分析し，DINA モデルの結果にもとづく診断をフィードバックしている。しかし，本章の結果によれば，こうしたテストの項目反応関数として DINA モデルを選択するには制約が強すぎる可能性がある。そのため，LLM や A-CDM といったより制約の緩やかなモデルが Sun et al.（2012）の分数テストにもより適合するかもしれない。もちろん，この主張は検討される必要があるが，もしこの主張が正しければ，より制約の緩やかな CDM を用いることにより，解答者の知識状態を DINA モデルより適切に推定することができる可能性がある。

また，前述のように，テスト設計を行う段階で CDM の適用を前提とし，DINA・DINO モデルや G-DINA モデルに適した項目作成ができれば，本研究の結果とは異なった結果が得られる可能性も多分にある。しかし，第 1 章で述べたように，CDM に則ったテストの作成には膨大な労力が必要となる。クラスルーム単位のテストではハイステークスなテスト

のような細かい設定を行うことは通常想定していない。このようなことから，既存のテストに対してCDMを適用する場面は多く見られるだろう。アプリケーション研究の例でもCDMの適用を前提としてテストを作成しているものは少なく，その多くは既存のテストへのモデル適用が中心的な関心である。本研究の結果から，こういったテストに対してもCDMを適用することによって，CDM特有のフィードバックに役立つ情報が得られ，テストを改良するための方法としての示唆が得られると考えられる。ただし，DINAモデルの適用したい場合には，厳密に複数のアトリビュートを適用できる必要があり，今回想定したアトリビュートの定義や問題項目の設定を再考する必要があるということがわかる。

本章の限界点の1つは，TIMSSの調査では非常に多くの国が参加しているにも関わらず，高々7つの国の分析しか行っていない点が挙げられる。今後の研究としては，本章の知見が他の国や，あるいは異なった時点のデータに一般化できるのかということを評価するということが考えられる。本章においては，アトリビュートとQ行列はLee et al.（2011）によって作成されたものを用いた。この選択はLee et al.（2011）の研究知見の再現性を検討するために必要であった。しかし，安定した項目パラメタやアトリビュート習得パタンの推定を可能にするためには，診断情報として重要なアトリビュートを残してアトリビュート数をより少なくするべきかもしれない。アトリビュートの構造を明確化するという先行研究（e.g., Leighton et al., 2004）は，こうした重要なアトリビュートを同定するのに役立つ可能性がある。また，第2章で述べたように，近年データにもとづいてQ行列を推定する新規な手法も幾つか開発されてきている（e.g., Xiang, 2013）。これらのQ行列の推定手法は未だ発展途上であり，経験的に十分に検討されてきてはいないが，こういった手法の利点を活かして，CDMの妥当な適用とよりよい構成を行うことも可能となる可能性がある。

また，TIMSSの様々な国の解答データにCDMを適用した場合には，主効果モデルが選択されやすかったものの，絶対適合指標の観点からはIRTモデルの適合が良く，あくまで情報量規準という相対指標を利用したためにCDMの適合が良かったという点には注意が必要である。CDMの適合が良いという結果はモデルを相対的に比較した中での結果であり，利用するモデルが変わることによって最適なモデルも変化する可能性もある。こうしたことから，複数の指標を組合せて適したモデルを判断したり，テスト内容や測定の目的に合わせて適しているモデルを選択する必要もあると考えられる。また，今回用いた絶対適合指標は観測された相関係数とモデルで期待された相関係数の差異に注目した指標である。絶対適合度指標は共分散構造分析の文脈でも数多く提案されており（e.g., 星野・岡田・前田，

2005），別の観点からの適合度指標を用いることも考えられる。例えば，0 から 1 までに標準化された指標を用いることで，分散説明率のようにモデルで説明できたデータの分散の大きさなどを解釈することもできる。また，TIMSS が IRT を適用する前提で作成されたテストであったため，絶対適合指標において CDM は IRT モデルよりも適合が悪いことが示された可能性はある。ただし，連続的な潜在変数を用いている IRT はこうした絶対的な適合指での適合がよいことが予測されていたため，そもそも CDM モデルを評価する指標として MADcor や SRMSR はあまり適していなかった可能性もある。

この他，CDM の中では主効果モデルが相対的に選択されたものの，国によって選択されたモデルが異なっていたことから，IRT モデルにおける特異項目機能（differential item functioning, DIF; e.g., 孫・井上, 1995）の問題のように国の間でそもそも項目反応関数が異なっている可能性が考えられる。このような，違いが生じた理由については，後述するような国によっての教育環境の違いの可能性も考えられる。あるいは，今回定義されたアトリビュートの性質が国によって異なっており，それぞれの国で異なった解答方略がとられていたという可能性も考えられる。認知メカニズム自体が国によって異なっているということは直感的には考えにくいものの，アトリビュートをどのように問題に解答に利用するのかその方略は授業実態などによって変わりうるだろう。さらに今回利用したアトリビュートの定義は問題の性質としてどのような能力が必要であるかを述べているに留まっており，それらの能力をどのように適用するのかといった手続き的な観点は含まれていない。そのため，アトリビュートがテスト解答時にどのように適用されるのかが国によって異なっていた可能性があり，今回得られたモデル比較の結果が国によって多少異なっていたという結果に繋がった可能性がある。

5-4-2　アトリビュート習得パタンについて

アメリカを含む 7 つの国でのアトリビュートの習得パタンは，各国で最も適合が良かったモデルを用いた結果を示した。この結果から，同じアトリビュートであっても，国ごとに多くの解答者が習得しているアトリビュートとそうではないアトリビュートが存在することが示された。また，国ごとにアトリビュートの習得パタンはかなり異なっている様子が示された。例えば，アメリカは図形と測定の領域のアトリビュートを 1 つしか習得していない解答者が多く，また数領域のアトリビュートもほとんど習得していないパタンが 2 位にあるなど，特徴が見られた。また，TIMSS の公式のスコアに対応して，スコアが高い国の方

が多くのアトリビュートを習得している傾向がみられ，スコアが低い国ではアトリビュートの習得数も少ないという傾向にも一致していた。

より具体的にどのようなアトリビュートが習得されているのかを見てみると，基本的な計算能力を想定している「数 1」といったアトリビュートは習得されている場合が多い傾向が見られた。その一方，「数 4」や「数 5」など小数や分数の計算に関連したアトリビュートは若干習得がなされにくい傾向がみられた。また，図形と測定領域においては，「図形と測定 11」などがやや習得されにくく，図形の周囲の長さや体積を求めることの習得は難しい傾向がある可能性が示された。また，資料の表現領域では「資料の表現 14」が習得されていないパタンがしばしばみられた。「資料の表現 14」はデータの使い方と計算に関するアトリビュートであった。このようにそれぞれの領域の中でも，特に計算に関するアトリビュートは相対的に習得が難しい可能性が示唆された。算数においては，代数的な計算ができるかが，算数能力の重要な側面を示しており，それぞれの領域でも計算が関係してくるアトリビュートは多くの解答者には難しい可能性があるということを反映している可能性がある。

CDM の利用する主目的として，アトリビュート習得パタンを用いた学習改善がある。こうしたアトリビュート習得パタンから，各国の学習の状況が明らかになり，多くの解答者が未習得であるアトリビュートを選択して学習に盛り込むことで効率的な学習を行うことが可能となりうる。

5-4-3　アトリビュートと IRT モデルの潜在特性の関係について

アトリビュート習得パタンに関連して，アトリビュートの性質を考慮する上で有用と考えられる。アトリビュート習得確率と IRT モデルの潜在特性との関係や，アトリビュート習得数と TIMSS の公式スコアとの関係から，Lee et al.（2011）の設定したアトリビュートは一次元的な潜在特性を反映しており，少なくとも算数能力の特定の側面をとらえている可能性があることが示唆された。もちろん，CDM で推定されたアトリビュートと IRT モデルでの 1 次元潜在特性は仮定されているものが大きく異なっているため，互換性のあるものではないものの，共通する成分を含んでいるという意味では類似したものと考えられる。この意味で，習得アトリビュート数はテストの合計得点などのように，1 次元の能力を荒く反映する指標としても利用できる可能性がある。

また，各アトリビュートの習得パタンを従属変数にし，1 次元潜在特性を独立変数にしたロジスティック回帰分析の結果から，国によって，またアトリビュートによってどの程度

1 次元的な能力を反映しているのかが異なっていた。ただし，多くのアトリビュートは 1 次元の潜在特性と正の関係を示した。このことから，個別のアトリビュートを習得している確率も多くの場合に 1 次元の潜在特性によって予測可能であると考えることができる。さらに，例えばシンガポールデータにおける「数 2」アトリビュートの習得確率は，1 次元潜在特性によって極めてよく説明することができることが示された。

その一方で，潜在特性によって習得確率を予測できないアトリビュートも存在した。さらに，カタールやイエメンデータでは回帰係数の推定値が負に推定されたアトリビュートも存在し，算数能力が高いほどアトリビュート習得確率が低いということを示したアトリビュートも存在した。ただし，このようなアトリビュートはごく少数であった。

第 1 章で述べたように，アトリビュートの設定は基本的に，先行研究の知見や教科の専門家の意見などテストの項目反応以外の情報を使った設定が多く，統計的な観点からの評価はあまり行われてきていない。今回の研究で行った 1 次元潜在特性との関連の検討によって，少なくとも特性の高低によってアトリビュート習得確率をどれほど予測するのかという観点から評価することができる。この分析で，潜在特性がアトリビュート習得パタンを予測しない場合には注意が必要と考えられる。つまり，この場合にはアトリビュートが能力の高低を反映していないということになるので，アトリビュートが学力の高低以外の要素を測定している可能性がある。これは，単純な能力の高低には反映されない重要な要素である可能性があるため，直ちに当該アトリビュートを除外するべき理由にはならないが，アトリビュートの定義を再考する必要があるかもしれない。例えば，香港データの「数 4」（比率を含む問題解決）においては，$\hat{\beta} = 0.311$（$95\%\mathrm{CI}[0.101,\ 0.522], p = .004$）であり，5%水準で有意であったものの，比較的低い値を示した。このような現象が生じた理由として，香港データでは「比率を含む問題解決」というアトリビュートは他のアトリビュートの一部として理解されるなど，必ずしも独立したアトリビュートではない可能性もある。これは香港データにおいてのみの結果だった可能性もあるが，結果の解釈には注意が必要である。また，国によってアトリビュートの定義が変わっている可能性や，解答者の問題解答行動が異なっている可能性もある。さらに，学力水準や，教育内容に合わせて適切な困難度のアトリビュートを設定する必要などが示唆される。

こうした点は後述のアトリビュートの妥当性とも関係する問題であるが，アトリビュートがどの国でも同様であるという仮定が難しい可能性を示していると考えることもできる。CDM で仮定するアトリビュートは純粋な認知能力のようなもののみを指すわけではなく，

教科学習に関連した知識的側面としての性質もある。例えば，本研究で想定した「14. データからの情報の使い方の理解と計算」（資料の活用 14）というアトリビュートは，どのように提示された図や表を読み取るのかといったことを教えられていなければならず，知識的な側面を反映しているアトリビュートと考えられる。こういったアトリビュートは「情報のエンコーディング」や「推論の実行」というよりも，一種の学習目標と見なすことも可能である。さらにいえば，「資料の活用 14」には計算の要素も含まれているが，計算に関しては他のアトリビュートである「2. 四つの演算子を用いた整数の計算と倍数の認識および計算の推論」によって定義されている。このような要素の重複がある場合には，それらのアトリビュートがどのように類似していて，どのように異なっているのかの明確な言及が必要である。

5-4-4　　項目パラメタの推定値について

項目パラメタの推定については，各国で類似した推定値を示した部分と，国によって異なるアトリビュートが正答確率に寄与する部分があることが示された。また，TIMSS データにおいては Lee et al.（2011）が想定したような複数のアトリビュートが必ずしも必要ではない項目が散見された。その一方で，複数のアトリビュートが正答確率に影響を与える場合もあり，部分的には複数のアトリビュートが必要であることが示された。第 6 章において，アトリビュートの定義と問題項目についての考察を深めるが，確かに特定の項目においては過剰にアトリビュートが付与されていた可能性は高いだろう。例えば，項目 4（M041076）は，ジョーといった人名やお金などが題材であり，実生活に関連した問題となっているが，問題文では分数が明示的に与えられており，必ずしも実生活の文脈を理解しないと解けない問題であるのかは疑問が残る。

また，比較的多くの項目において，複数のアトリビュートではなく単一のアトリビュートが問題の正答に寄与している可能性が示唆された点について，この理由としてアトリビュートの定義が抽象的であり，TIMSS の問題項目では今回のアトリビュートを測定するのには不十分であった可能性がある。その結果として，複数のアトリビュートではなく，単一のアトリビュートの影響のみがみられてしまった可能性がある。

さらに，項目パラメタの推定値については，モデルの複雑さに対してサンプルサイズが不十分であり，標準誤差が大きい場合もみられた。主効果モデルを適用することによって，どのアトリビュートがどの項目にどれだけ寄与していたのかを精緻に検討することが可能

となったものの，標準誤差の多い場合もあり，項目パラメタの点推定値については，断定的な判断を行うことには注意が必要である可能性がある。

そのうえで，今回適用したモデルは全て主効果モデルであった点は国の間で共通であり，どういったアトリビュートが問題の正答に寄与しているのかという観点からの比較はできる。また，R-RUM，A-CDM，LLM が支持された国は複数あり，そういった国の間での項目パラメタの比較も可能であると考えられる。例えば，項目 1 について，アメリカでは「数 2」がないと正答率の低下がみられるものの，「数 1」は問題の正答確率に影響を及ぼさない事が示唆される一方，シンガポールでは「数 1」の習得が項目 1 の正答に強い意味を持っており，「数 2」は項目 1 の正答には無意味なアトリビュートと考えることができる。このような国による違いがあることから，国ごとに使用されている認知要素が異なっている可能性もあることが示唆される。

項目パラメタを解釈する上でも，やはりアトリビュートの定義はやはり重要になっており，アトリビュートの設定の妥当性を抜きにして様々な解釈を行うことは難しい。さらに CDM の分析がどれだけ信頼できるものなのかは，どれほどアトリビュートの設定が妥当であるかに強く依存してしまうため，アトリビュートの設定に十分に注意する必要がある。

また，Lee et al.（2011）の Q 行列には誤設定があった可能性も考えられる。なぜならば，今回選択された主効果モデルの項目パラメタの推定値からは，R-RUM であれば$\hat{r}$が 1.000 に近く，A-CDM や LLM であれば$\hat{\delta}$が 0 に近く，特定のアトリビュートが項目の正答に寄与していない場合がみられたからである。このことから，不要なアトリビュートが項目の正答に必要と仮定されている場合も多いと推察できる。これらのことから，問題項目を分析して正答に必要であると想定されたアトリビュートであっても，解答者は定義されたアトリビュートを細かく使い分けて問題に解答していない可能性がある。また，もちろん問題自体にそれらのアトリビュートを細かく区別できるだけの情報がないということもありうる。CDM では，アトリビュート習得パタンごとに問題正答確率が異なっていることを想定しているが，現実的な項目でこのような事が達成されるわけでは必ずしも無く，項目の品質が結果に影響するといえる。

アトリビュートの妥当性とも関係するが，国間の項目反応関数の相違に加えて，そもそも各国で問題解答に必要なアトリビュートが異なっている可能性がある。つまり，通常の項目のようにアトリビュートにも習得が容易なものや，困難度が高いものが存在すると考えられる。現状の CDM においては，アトリビュートは習得パタンを生成するための構成

概念にすぎず，アトリビュートの内容的な側面しか鑑みられてこなかったと考えられる。しかしながら，アトリビュートがどのように問題解答プロセスに寄与するのかというような精緻な議論も必要であろう。現状の CDM はそれぞれのアトリビュートを習得しているか否かの組み合わせによってパタンを区別し，それぞれの習得パタンでの項目反応確率を推定しているにとどまっており，解答のプロセスの詳細を説明するモデルとしては機能していない点が指摘できる。これは，CDM 自体が LCA ベースであることにも起因しているものであり，プロセスを記述するためのモデルとしては未だ不十分な点があるといえよう。AHM などは問題解答中のミクロな解答プロセスの記述ではないものの，学習要素全体として学習の順序関係やアトリビュート間の関係を積極的にモデル化することにより，学習の段階を表現することが可能である。項目パラメタの解釈にはこうしたアトリビュート間の関係について考慮することが望ましいだろう。

表 5.1 各国の男女別の解答者数およびサンプルサイズ

国	女子	男子	サンプルサイズ
アメリカ	587	543	1130
香港	252	291	543
シンガポール	345	372	717
スロベニア	319	301	620
アルメニア	287	299	586
カタール	519	480	999
イエメン	375	461	836

表 5.2 アメリカデータにおける各項目の平均正答率および IT 相関

項目番号	項目名	*N*	*Mean*	*SD*	*SE*	*Mean* 95%CI[下限, 上限]	IT相関	IT相関 95%CI[下限, 上限]
1	M041052	558	.808	.394	.027	[.756 , .860]	.329	[.254 , .398]
2	M041056	547	.419	.494	.030	[.360 , .478]	.437	[.368 , .497]
3	M041069	557	.456	.499	.030	[.397 , .515]	.417	[.347 , .478]
4	M041076	557	.722	.449	.028	[.666 , .777]	.338	[.263 , .406]
5	M041281	552	.667	.472	.029	[.609 , .724]	.518	[.457 , .571]
6	M041164	560	.914	.280	.022	[.870 , .958]	.281	[.203 , .353]
7	M041146	557	.592	.492	.030	[.534 , .651]	.376	[.303 , .441]
8	M041152	560	.484	.500	.030	[.425 , .543]	.384	[.312 , .448]
9	M041258A	535	.809	.393	.027	[.756 , .862]	.286	[.206 , .359]
10	M041258B	550	.084	.277	.022	[.040 , .128]	.168	[.085 , .246]
11	M041131	552	.357	.480	.029	[.299 , .415]	.163	[.081 , .242]
12	M041275	545	.787	.410	.027	[.733 , .841]	.213	[.131 , .289]
13	M041186	559	.750	.434	.028	[.695 , .804]	.369	[.296 , .434]
14	M041336	553	.403	.491	.030	[.345 , .462]	.399	[.327 , .462]
15	M031303	1120	.743	.437	.020	[.704 , .782]	.391	[.342 , .437]
16	M031309	1100	.528	.499	.021	[.486 , .570]	.572	[.534 , .607]
17	M031245	1117	.279	.449	.020	[.240 , .319]	.390	[.340 , .436]
18	M031242A	909	.767	.423	.022	[.724 , .809]	.325	[.267 , .380]
19	M031242B	1065	.422	.494	.022	[.379 , .464]	.534	[.492 , .572]
20	M031242C	1110	.658	.475	.021	[.617 , .698]	.298	[.244 , .349]
21	M031247	1089	.211	.408	.019	[.173 , .249]	.279	[.223 , .331]
22	M031219	1105	.488	.500	.021	[.446 , .529]	.102	[.043 , .160]
23	M031173	1107	.659	.474	.021	[.618 , .699]	.533	[.492 , .570]
24	M031085	1103	.429	.495	.021	[.387 , .470]	.325	[.272 , .375]
25	M031172	1107	.804	.397	.019	[.767 , .841]	.414	[.365 , .458]

表 5.3 香港データにおける各項目の平均正答率および IT 相関

項目番号	項目名	*N*	*Mean*	*SD*	*SE*	*Mean* 95%CI[下限, 上限]	IT相関	IT相関 95%CI[下限, 上限]
1	M041052	269	.788	.409	.039	[.712 , .865]	.396	[.291 , .485]
2	M041056	270	.870	.337	.035	[.801 , .940]	.437	[.336 , .520]
3	M041069	269	.751	.433	.040	[.672 , .830]	.424	[.322 , .510]
4	M041076	268	.888	.316	.034	[.821 , .955]	.329	[.218 , .425]
5	M041281	265	.834	.373	.038	[.760 , .907]	.376	[.268 , .468]
6	M041164	267	.996	.061	.015	[.967 , 1.000]	.159	[.040 , .271]
7	M041146	269	.888	.315	.034	[.821 , .956]	.337	[.227 , .432]
8	M041152	269	.914	.280	.032	[.851 , .978]	.360	[.252 , .453]
9	M041258A	255	.878	.327	.036	[.808 , .949]	.147	[.024 , .262]
10	M041258B	257	.724	.448	.042	[.642 , .806]	.182	[.061 , .294]
11	M041131	270	.437	.497	.043	[.353 , .521]	.096	[-.024 , .211]
12	M041275	269	.896	.306	.034	[.830 , .962]	.203	[.085 , .311]
13	M041186	270	.615	.488	.042	[.532 , .698]	.279	[.165 , .380]
14	M041336	268	.642	.480	.042	[.559 , .725]	.470	[.373 , .550]
15	M031303	542	.928	.259	.022	[.885 , .971]	.360	[.286 , .427]
16	M031309	539	.924	.265	.022	[.880 , .967]	.402	[.330 , .465]
17	M031245	542	.696	.461	.029	[.638 , .753]	.506	[.443 , .560]
18	M031242A	531	.876	.330	.025	[.827 , .925]	.321	[.243 , .392]
19	M031242B	530	.664	.473	.030	[.606 , .723]	.319	[.241 , .390]
20	M031242C	542	.803	.398	.027	[.749 , .856]	.304	[.226 , .375]
21	M031247	499	.144	.352	.027	[.092 , .196]	.251	[.167 , .328]
22	M031219	541	.591	.492	.030	[.532 , .651]	.298	[.220 , .370]
23	M031173	541	.898	.302	.024	[.852 , .945]	.598	[.545 , .643]
24	M031085	540	.650	.477	.030	[.592 , .708]	.339	[.263 , .408]
25	M031172	526	.914	.280	.023	[.869 , .960]	.503	[.439 , .559]

Note. 信頼区間が1を越えたものは1.000と表記した。

表 5.4 シンガポールデータにおける各項目の平均正答率および IT 相関

項目番号	項目名	*N*	*Mean*	*SD*	*SE*	*Mean* 95%CI[下限, 上限]	IT相関	IT相関 95%CI[下限, 上限]
1	M041052	356	.857	.351	.031	[.795 , .918]	.490	[.409 , .558]
2	M041056	353	.762	.426	.035	[.694 , .830]	.588	[.519 , .644]
3	M041069	356	.848	.359	.032	[.786 , .911]	.600	[.533 , .654]
4	M041076	356	.817	.387	.033	[.753 , .882]	.460	[.376 , .531]
5	M041281	351	.866	.341	.031	[.805 , .927]	.481	[.399 , .550]
6	M041164	353	.977	.149	.021	[.937 , 1.000]	.250	[.150 , .342]
7	M041146	355	.670	.471	.036	[.599 , .742]	.468	[.385 , .539]
8	M041152	354	.836	.371	.032	[.773 , .900]	.469	[.386 , .539]
9	M041258A	350	.840	.367	.032	[.777 , .903]	.424	[.335 , .499]
10	M041258B	350	.491	.501	.038	[.417 , .566]	.420	[.331 , .496]
11	M041131	355	.442	.497	.037	[.369 , .516]	.271	[.172 , .361]
12	M041275	356	.871	.336	.031	[.811 , .931]	.444	[.358 , .517]
13	M041186	356	.747	.435	.035	[.679 , .816]	.610	[.545 , .663]
14	M041336	355	.639	.481	.037	[.567 , .712]	.523	[.446 , .587]
15	M031303	715	.906	.292	.020	[.867 , .946]	.562	[.512 , .605]
16	M031309	716	.888	.315	.021	[.847 , .929]	.565	[.516 , .607]
17	M031245	716	.703	.457	.025	[.653 , .752]	.603	[.558 , .642]
18	M031242A	677	.861	.346	.023	[.817 , .905]	.323	[.254 , .386]
19	M031242B	699	.629	.483	.026	[.578 , .681]	.465	[.406 , .516]
20	M031242C	711	.723	.448	.025	[.674 , .772]	.341	[.275 , .401]
21	M031247	708	.500	.500	.027	[.448 , .552]	.578	[.530 , .620]
22	M031219	715	.557	.497	.026	[.505 , .608]	.218	[.147 , .284]
23	M031173	714	.891	.312	.021	[.850 , .932]	.673	[.634 , .705]
24	M031085	716	.729	.445	.025	[.680 , .778]	.538	[.487 , .583]
25	M031172	715	.863	.344	.022	[.820 , .906]	.590	[.543 , .630]

Note. 信頼区間が1を越えたものは1.000と表記した。

表 5.5 スロベニアデータにおける各項目の平均正答率および IT 相関

項目番号	項目名	*N*	*Mean*	*SD*	*SE*	*Mean* 95%CI[下限, 上限]	IT相関	IT相関 95%CI[下限, 上限]
1	M041052	308	.721	.449	.038	[.646 , .796]	.487	[.399 , .560]
2	M041056	301	.472	.500	.041	[.392 , .552]	.432	[.336 , .512]
3	M041069	294	.037	.190	.025	[.000 , .087]	.207	[.094 , .310]
4	M041076	269	.353	.479	.042	[.270 , .436]	.438	[.338 , .522]
5	M041281	292	.726	.447	.039	[.649 , .803]	.391	[.291 , .477]
6	M041164	304	.980	.139	.021	[.938 , 1.000]	.146	[.034 , .252]
7	M041146	306	.621	.486	.040	[.543 , .699]	.428	[.333 , .508]
8	M041152	298	.248	.433	.038	[.174 , .323]	.285	[.178 , .381]
9	M041258A	294	.738	.440	.039	[.662 , .814]	.345	[.241 , .436]
10	M041258B	299	.007	.082	.017	[.000 , .039]	.062	[-.051 , .173]
11	M041131	307	.375	.485	.040	[.297 , .452]	.202	[.091 , .303]
12	M041275	299	.886	.318	.033	[.822 , .950]	.322	[.217 , .415]
13	M041186	306	.745	.437	.038	[.671 , .819]	.490	[.402 , .563]
14	M041336	293	.461	.499	.041	[.380 , .542]	.280	[.171 , .378]
15	M031303	610	.708	.455	.027	[.655 , .762]	.508	[.449 , .560]
16	M031309	586	.633	.482	.029	[.577 , .689]	.470	[.406 , .526]
17	M031245	611	.200	.400	.026	[.150 , .250]	.381	[.313 , .443]
18	M031242A	553	.682	.466	.029	[.625 , .739]	.421	[.351 , .482]
19	M031242B	532	.564	.496	.031	[.504 , .624]	.473	[.406 , .531]
20	M031242C	569	.684	.465	.029	[.628 , .740]	.406	[.337 , .468]
21	M031247	563	.160	.367	.026	[.110 , .210]	.262	[.184 , .335]
22	M031219	596	.539	.499	.029	[.482 , .595]	.111	[.031 , .189]
23	M031173	603	.701	.458	.028	[.647 , .756]	.535	[.478 , .584]
24	M031085	589	.581	.494	.029	[.524 , .637]	.215	[.137 , .289]
25	M031172	595	.739	.439	.027	[.686 , .793]	.446	[.381 , .503]

Note. 信頼区間が1を越えたものは1.000，平均正答率の信頼下限で0を下回ったものは.000と表記した。

表 5.6 アルメニアデータにおける各項目の平均正答率および IT 相関

項目番号	項目名	*N*	*Mean*	*SD*	*SE*	*Mean* 95%CI[下限, 上限]	IT相関	IT相関 95%CI[下限, 上限]
1	M041052	265	.589	.493	.043	[.504 , .673]	.393	[.287 , .483]
2	M041056	274	.850	.357	.036	[.780 , .921]	.066	[-.053 , .181]
3	M041069	261	.613	.488	.043	[.528 , .698]	.383	[.275 , .475]
4	M041076	266	.883	.321	.035	[.815 , .952]	.301	[.188 , .401]
5	M041281	234	.731	.445	.044	[.645 , .816]	.476	[.372 , .560]
6	M041164	255	.769	.423	.041	[.689 , .848]	.238	[.119 , .346]
7	M041146	266	.774	.419	.040	[.697 , .852]	.306	[.193 , .406]
8	M041152	270	.689	.464	.041	[.608 , .770]	.336	[.226 , .431]
9	M041258A	192	.797	.403	.046	[.707 , .887]	.235	[.096 , .358]
10	M041258B	181	.326	.470	.051	[.226 , .426]	.026	[-.119 , .169]
11	M041131	273	.385	.487	.042	[.302 , .467]	.192	[.075 , .300]
12	M041275	132	.462	.500	.062	[.341 , .583]	.282	[.116 , .422]
13	M041186	263	.483	.501	.044	[.397 , .568]	.441	[.340 , .525]
14	M041336	236	.445	.498	.046	[.355 , .535]	.219	[.094 , .333]
15	M031303	528	.742	.438	.029	[.686 , .799]	.420	[.349 , .483]
16	M031309	558	.867	.339	.025	[.819 , .916]	.308	[.232 , .378]
17	M031245	537	.477	.500	.031	[.417 , .537]	.456	[.388 , .515]
18	M031242A	349	.739	.440	.035	[.670 , .809]	.416	[.326 , .492]
19	M031242B	341	.669	.471	.037	[.596 , .741]	.289	[.188 , .378]
20	M031242C	439	.499	.501	.034	[.433 , .565]	.302	[.215 , .381]
21	M031247	493	.122	.327	.026	[.071 , .172]	.224	[.139 , .304]
22	M031219	492	.569	.496	.032	[.507 , .631]	.351	[.272 , .422]
23	M031173	525	.590	.492	.031	[.530 , .650]	.535	[.474 , .588]
24	M031085	506	.575	.495	.031	[.514 , .636]	.453	[.383 , .514]
25	M031172	487	.575	.495	.032	[.512 , .637]	.529	[.464 , .584]

表 5.7 カタールデータにおける各項目の平均正答率および IT 相関

項目番号	項目名	*N*	*Mean*	*SD*	*SE*	*Mean* 95%CI[下限, 上限]	IT相関	IT相関 95%CI[下限, 上限]
1	M041052	482	.633	.483	.032	[.571 , .695]	.164	[.076 , .248]
2	M041056	331	.118	.323	.031	[.057 , .179]	.076	[-.032 , .180]
3	M041069	471	.089	.285	.025	[.041 , .137]	.004	[-.086 , .094]
4	M041076	356	.124	.330	.030	[.064 , .183]	.148	[.044 , .246]
5	M041281	389	.321	.468	.035	[.253 , .389]	.244	[.148 , .331]
6	M041164	426	.444	.497	.034	[.377 , .511]	.074	[-.021 , .167]
7	M041146	400	.245	.431	.033	[.181 , .309]	.204	[.108 , .294]
8	M041152	472	.133	.340	.027	[.081 , .186]	-.048	[-.137 , .042]
9	M041258A	316	.196	.398	.035	[.127 , .266]	.277	[.171 , .371]
10	M041258B	339	.021	.142	.020	[.000 , .061]	.259	[.157 , .352]
11	M041131	494	.134	.341	.026	[.082 , .185]	.017	[-.071 , .104]
12	M041275	350	.180	.385	.033	[.115 , .245]	.110	[.005 , .211]
13	M041186	461	.165	.371	.028	[.109 , .220]	-.024	[-.115 , .067]
14	M041336	461	.132	.339	.027	[.079 , .185]	-.084	[-.172 , .008]
15	M031303	919	.312	.464	.022	[.268 , .356]	.009	[-.055 , .074]
16	M031309	778	.046	.210	.016	[.014 , .078]	.344	[.282 , .402]
17	M031245	935	.096	.295	.018	[.061 , .131]	-.031	[-.095 , .033]
18	M031242A	614	.202	.402	.026	[.152 , .252]	.312	[.239 , .378]
19	M031242B	556	.049	.215	.020	[.010 , .087]	.218	[.137 , .293]
20	M031242C	873	.475	.500	.024	[.428 , .522]	.153	[.088 , .216]
21	M031247	741	.008	.090	.011	[.000 , .030]	.006	[-.066 , .077]
22	M031219	885	.362	.481	.023	[.316 , .407]	.086	[.020 , .150]
23	M031173	930	.178	.383	.020	[.139 , .218]	.190	[.128 , .250]
24	M031085	911	.374	.484	.023	[.329 , .420]	.057	[-.008 , .121]
25	M031172	904	.188	.391	.021	[.147 , .229]	.074	[.008 , .138]

Note. 平均正答率の信頼区間で0を下回ったものは.000と表記した。

表 5.8 イエメンデータにおける各項目の平均正答率および IT 相関

項目番号	項目名	*N*	*Mean*	*SD*	*SE*	*Mean* 95%CI[下限, 上限]	IT相関	IT相関 95%CI[下限, 上限]
1	M041052	383	.608	.489	.036	[.538 , .678]	.056	[-.045 , .154]
2	M041056	171	.199	.400	.048	[.104 , .294]	.551	[.441 , .635]
3	M041069	384	.198	.399	.032	[.135 , .261]	.225	[.128 , .315]
4	M041076	285	.221	.416	.038	[.146 , .296]	.356	[.251 , .447]
5	M041281	270	.267	.443	.041	[.187 , .346]	.380	[.274 , .470]
6	M041164	338	.284	.452	.037	[.212 , .356]	.170	[.065 , .269]
7	M041146	259	.077	.267	.032	[.014 , .140]	.260	[.142 , .365]
8	M041152	388	.237	.426	.033	[.172 , .302]	.028	[-.072 , .126]
9	M041258A	223	.188	.392	.042	[.106 , .271]	.431	[.319 , .523]
10	M041258B	256	.055	.228	.030	[-.004 , .113]	.598	[.517 , .661]
11	M041131	400	.235	.425	.033	[.171 , .299]	-.004	[-.101 , .094]
12	M041275	291	.034	.182	.025	[.000 , .083]	.411	[.312 , .495]
13	M041186	376	.250	.434	.034	[.183 , .317]	-.022	[-.122 , .079]
14	M041336	362	.116	.321	.030	[.058 , .174]	.065	[-.038 , .166]
15	M031303	695	.370	.483	.026	[.318 , .421]	.033	[-.042 , .106]
16	M031309	560	.080	.272	.022	[.037 , .124]	.583	[.529 , .629]
17	M031245	767	.207	.406	.023	[.162 , .252]	-.111	[-.179 , -.040]
18	M031242A	541	.133	.340	.025	[.084 , .182]	.229	[.148 , .305]
19	M031242B	326	.031	.173	.023	[-.014 , .076]	.002	[-.106 , .110]
20	M031242C	701	.372	.484	.026	[.321 , .424]	.078	[.004 , .150]
21	M031247	532	.009	.097	.004	[.001 , .018]	-	- -
22	M031219	710	.296	.457	.025	[.246 , .345]	.109	[.035 , .180]
23	M031173	724	.222	.416	.024	[.175 , .269]	.063	[-.010 , .134]
24	M031085	730	.319	.466	.025	[.270 , .369]	.072	[.000 , .144]
25	M031172	718	.121	.327	.021	[.079 , .163]	.174	[.102 , .243]

Note. M031247はIT相関の算出には利用できなかった。また平均正答率の信頼区間で0を下回ったものは.000と表記した。

表 5.9 各国のα係数の推定値と 95%信頼区間

国	α	95%CI[下限, 上限]
アメリカ	.821	[.806 , .835]
香港	.792	[.768 , .817]
シンガポール	.890	[.879 , .901]
スロベニア	.820	[.801 , .840]
アルメニア	.798	[.775 , .821]
カタール	.415	[.364 , .467]
イエメン	.564	[.523 , .606]

Note. イエメンにおいてはM031247はα係数の推定に利用できなかった。

表 5.10 アメリカデータにおける IRT モデルと CDM のモデル比較

IRT/CDM	モデル	Deviance	AIC	BIC	MADcor	SRMSR	項目パラメタ数
IRT	3PL	21197.05	21347.05	21724.29	.030	.041	75
	2PL	21245.88	21345.88	21597.38	.031	.042	50
	1PL	21605.52	21655.52	21781.27	.061	.078	25
CDM	G-DINA	18383.19	18907.19	20225.04	.072	.072	262
	DINA	21173.55	21273.55	21525.05	.084	.077	50
	DINO	21244.69	21344.69	21596.19	.105	.102	50
	A-CDM	18649.78	18839.78	19317.62	.081	.082	95
	LLM	19088.42	19278.42	19756.27	.070	.074	95
	R-RUM	18584.39	18774.39	19252.23	.080	.081	95

Note. 1-3PLは1-3パラメタロジスティックモデルを意味する。網掛けのセルは各観点での最適な値が得られたものを意味する。

表 5.11 香港データにおける IRT モデルと CDM のモデル比較

IRT/CDM	モデル	Deviance	AIC	BIC	MADcor	SRMSR	項目パラメタ数
IRT	3PL	7812.17	7962.17	8284.45	.045	.059	75
	2PL	7841.21	7941.21	8156.06	.046	.059	50
	1PL	7996.71	8046.71	8154.14	.076	.094	25
CDM	G-DINA	6437.60	6961.60	8087.44	.091	.089	262
	DINA	7707.45	7807.45	8022.31	.135	.121	50
	DINO	7712.32	7812.32	8027.18	.130	.115	50
	A-CDM	6653.40	6843.40	7251.63	.101	.096	95
	LLM	6582.56	6772.56	7180.79	.095	.093	95
	R-RUM	6783.54	6973.54	7381.77	.115	.107	95

Note. 1-3PLは1-3パラメタロジスティックモデルを意味する。網掛けのセルは各観点での最適な値が得られたものを意味する。

表 5.12 シンガポールデータにおける IRT モデルと CDM のモデル比較

IRT/CDM	モデル	Deviance	AIC	BIC	MADcor	SRMSR	項目パラメタ数
IRT	3PL	10554.09	10704.09	11047.22	.035	.047	75
	2PL	10600.62	10700.62	10929.37	.038	.050	50
	1PL	10976.23	11026.23	11140.60	.092	.114	25
CDM	G-DINA	8926.62	9450.62	10649.29	.103	.101	262
	DINA	10561.27	10661.27	10890.03	.155	.136	50
	DINO	10645.89	10745.89	10974.64	.169	.155	50
	A-CDM	9261.37	9451.37	9886.00	.108	.105	95
	LLM	9406.80	9596.80	10031.44	.115	.111	95
	R-RUM	9308.16	9498.16	9932.79	.111	.106	95

Note. 1-3PLは1-3パラメタロジスティックモデルを意味する。網掛けのセルは各観点での最適な値が得られたものを意味する。

表 5.13 スロベニアデータにおける IRT モデルと CDM のモデル比較

IRT/CDM	モデル	Deviance	AIC	BIC	MADcor	SRMSR	項目パラメタ数
IRT	3PL	10697.15	10847.15	11179.38	.038	.049	75
	2PL	10725.67	10825.67	11047.15	.040	.052	50
	1PL	10908.81	10958.81	11069.55	.065	.085	25
CDM	G-DINA	9178.72	9702.72	10863.31	.080	.079	262
	DINA	10552.44	10652.44	10873.92	.123	.113	50
	DINO	10622.81	10722.81	10944.29	.124	.114	50
	A-CDM	9436.26	9626.26	10047.08	.088	.083	95
	LLM	9388.83	9578.83	9999.65	.091	.091	95
	R-RUM	9234.20	9424.20	9845.02	.095	.097	95

Note. 1-3PLは1-3パラメタロジスティックモデルを意味する。網掛けのセルは各観点での最適な値が得られたものを意味する。

表 5.14 アルメニアデータにおける IRT モデルと CDM のモデル比較

IRT/CDM	モデル	Deviance	AIC	BIC	MADcor	SRMSR	項目パラメタ数
IRT	3PL	9420.97	9570.97	9899.10	.060	.082	75
	2PL	9439.23	9539.23	9757.98	.060	.083	50
	1PL	9564.28	9614.28	9723.66	.082	.106	25
CDM	G-DINA	7795.31	8319.31	9465.12	.116	.111	262
	DINA	9256.33	9356.33	9575.00	.146	.136	50
	DINO	9217.19	9317.19	9535.86	.153	.140	50
	A-CDM	8063.70	8253.70	8669.16	.121	.120	95
	LLM	8069.33	8259.33	8674.80	.120	.119	95
	R-RUM	8173.22	8363.22	8778.68	.123	.119	95

Note. 1-3PLは1-3パラメタロジスティックモデルを意味する。網掛けのセルは各観点での最適な値が得られたものを意味する。

表 5.15 カタールデータにおける IRT モデルと CDM のモデル比較

IRT/CDM	モデル	Deviance	AIC	BIC	MADcor	SRMSR	項目パラメタ数
IRT	3PL	13112.33	13262.33	13630.41	.044	.057	75
	2PL	13188.73	13288.73	13534.12	.047	.059	50
	1PL	13564.05	13614.05	13736.74	.075	.091	25
CDM	G-DINA	11479.83	12003.83	13289.40	.093	.094	262
	DINA	13045.53	13145.53	13390.87	.098	.089	50
	DINO	13008.34	13108.34	13353.68	.094	.087	50
	A-CDM	11820.48	12010.48	12476.63	.092	.097	95
	LLM	11720.40	11910.40	12376.54	.101	.104	95
	R-RUM	11949.82	12139.82	12605.96	.100	.105	95

Note. 1-3PLは1-3パラメタロジスティックモデルを意味する。網掛けのセルは各観点での最適な値が得られたものを意味する。

表 5.16 イエメンデータにおける IRT モデルと CDM のモデル比較

IRT/CDM	モデル	Deviance	AIC	BIC	MADcor	SRMSR	項目パラメタ数
IRT	3PL	10402.46	10552.46	10907.20	.056	.073	75
	2PL	10448.07	10548.07	10784.56	.058	.077	50
	1PL	10886.10	10936.10	11054.34	.098	.125	25
CDM	G-DINA	8978.58	9502.58	10741.48	.125	.137	262
	DINA	10313.60	10413.60	10650.03	.129	.124	50
	DINO	10295.34	10395.34	10631.77	.148	.143	50
	A-CDM	9291.09	9481.09	9930.31	.141	.153	95
	LLM	9247.31	9437.31	9886.53	.140	.151	95
	R-RUM	9314.84	9504.84	9954.06	.134	.145	95

Note. 1-3PLは1-3パラメタロジスティックモデルを意味する。網掛けのセルは各観点での最適な値が得られたものを意味する。

表 5.17 アトリビュート習得数の記述統計量

国	アトリビュート	*Mean*	(*SD*)	Median	Skewness	Kurtosis
アメリカ	数	4.263	(2.270)	4	0.082	-1.009
	図形と測定	1.826	(1.115)	2	0.251	-0.687
	資料の表現	2.174	(0.865)	2	-0.737	-0.351
	すべての領域	8.263	(3.498)	8	-0.027	-0.939
香港	数	5.722	(1.846)	6	-0.652	-0.160
	図形と測定	2.972	(1.002)	3	-0.702	-0.315
	資料の表現	2.230	(0.868)	2	-0.951	0.131
	すべての領域	10.924	(2.999)	11	-0.539	-0.289
シンガポール	数	5.922	(2.154)	7	-0.948	-0.021
	図形と測定	2.580	(1.235)	3	-0.444	-0.891
	資料の表現	2.379	(0.867)	3	-1.197	0.379
	すべての領域	10.881	(3.812)	12	-0.922	-0.114
スロベニア	数	4.590	(2.222)	5	-0.278	-0.926
	図形と測定	2.174	(1.075)	2	-0.178	-0.532
	資料の表現	1.773	(1.010)	2	-0.203	-1.137
	すべての領域	8.537	(3.527)	9	-0.256	-0.643
アルメニア	数	4.790	(2.039)	5	-0.317	-0.948
	図形と測定	2.348	(0.983)	2	-0.124	-0.475
	資料の表現	1.539	(1.073)	1	-0.002	-1.262
	すべての領域	8.677	(3.404)	9	-0.236	-0.862
カタール	数	2.004	(1.208)	2	0.729	0.271
	図形と測定	1.209	(1.007)	1	0.515	-0.391
	資料の表現	1.019	(0.912)	1	0.532	-0.598
	すべての領域	4.232	(2.132)	4	0.596	0.137
イエメン	数	2.300	(1.320)	2	1.159	1.753
	図形と測定	0.962	(0.919)	1	0.686	-0.300
	資料の表現	0.610	(0.712)	0	0.884	0.060
	すべての領域	3.872	(2.142)	4	1.040	1.772

Note. アトリビュート習得数は期待事後推定値(EAP)により算出した。

表 5.18 アメリカデータにおける各アトリビュート領域の習得パタン

順位	数			図形と測定			資料の活用		
	パタン	人数	割合	パタン	人数	割合	パタン	人数	割合
1	11111111	135	.119	0001	175	.155	111	490	.434
2	00000100	56	.050	0011	121	.107	101	290	.257
3	11100101	47	.042	0000	120	.106	001	137	.121
4	11111101	39	.035	0010	112	.099	011	64	.057
5	01000100	38	.034	0111	110	.097	100	49	.043
6	00100101	37	.033	0101	101	.089	000	48	.042
7	00000101	36	.032	1111	95	.084	110	41	.036
8	11101101	34	.030	0100	56	.050	010	11	.010
9	00000000	32	.028	1101	48	.042	-	-	-
10	00100001	30	.027	1001	41	.036	-	-	-
	合計	484	.430		979	.865		1130	1.000

表 5.19 香港データにおける各アトリビュート領域の習得パタン

順位	数			図形と測定			資料の活用		
	パタン	人数	割合	パタン	人数	割合	パタン	人数	割合
1	11111111	109	.201	1111	202	.372	111	252	.464
2	11101111	56	.103	1101	93	.171	101	121	.223
3	11011111	34	.063	1001	63	.116	011	58	.107
4	11001101	26	.048	1011	58	.107	001	51	.094
5	11101110	24	.044	0111	29	.053	000	29	.053
6	11001111	23	.042	0001	21	.039	110	14	.026
7	11000101	10	.018	0101	17	.031	010	13	.024
8	10111011	8	.015	0011	15	.028	100	5	.009
9	11001001	8	.015	1000	12	.022	-	-	-
10	11001110	8	.015	1010	9	.017	-	-	-
	合計	306	.564		519	.956		543	1.000

表 5.20 シンガポールデータおける各アトリビュート領域の習得パタン

順位	数			図形と測定			資料の活用		
	パタン	人数	割合	パタン	人数	割合	パタン	人数	割合
1	11111111	230	.321	1111	214	.298	111	427	.596
2	11101111	63	.088	1101	110	.153	101	101	.141
3	11101101	32	.045	0101	69	.096	001	69	.096
4	11111101	24	.033	0001	52	.073	011	51	.071
5	11001111	21	.029	0111	46	.064	000	30	.042
6	11011111	20	.028	0000	43	.060	100	14	.020
7	11001110	19	.026	0010	38	.053	110	13	.018
8	11111110	18	.025	0110	31	.043	010	12	.017
9	11100101	16	.022	0011	25	.035	-	-	-
10	00000000	15	.021	1110	21	.029	-	-	-
	合計	458	.639		649	.905		717	1.000

表 5.21 スロベニアデータにおける各アトリビュート領域の習得パタン

順位	数			図形と測定			資料の活用		
	パタン	人数	割合	パタン	人数	割合	パタン	人数	割合
1	11111101	66	.106	0111	94	.152	111	190	.306
2	11111001	58	.094	0101	88	.142	001	135	.218
3	11111111	53	.085	1111	67	.108	011	102	.165
4	11010000	25	.040	1011	46	.074	000	71	.115
5	11110101	24	.039	0000	45	.073	110	39	.063
6	00000000	23	.037	0001	40	.065	010	33	.053
7	00010000	16	.026	0100	38	.061	101	29	.047
8	11111011	16	.026	1001	38	.061	100	21	.034
9	00000001	14	.023	0011	33	.053	-	-	-
10	01010000	13	.021	0110	32	.052	-	-	-
	合計	308	.497		521	.840		620	1.000

表 5.22 アルメニアデータにおける各アトリビュート領域の習得パタン

順位	数			図形と測定			資料の活用		
	パタン	人数	割合	パタン	人数	割合	パタン	人数	割合
1	11101111	71	.121	0111	133	.227	111	144	.246
2	00001000	44	.075	0101	80	.137	000	120	.205
3	11111111	34	.058	1111	73	.125	010	78	.133
4	11101011	27	.046	0001	69	.118	101	72	.123
5	11101001	19	.032	0011	59	.101	110	52	.089
6	00101100	16	.027	1001	38	.065	100	51	.087
7	11101101	15	.026	1011	32	.055	001	45	.077
8	11111101	14	.024	0000	16	.027	011	24	.041
9	11011111	13	.022	1010	15	.026	-	-	-
10	10101111	11	.019	0110	14	.024	-	-	-
	合計	264	.451		529	.903		586	1.000

表 5.23 カタールデータにおける各アトリビュート領域の習得パタン

順位	数			図形と測定			資料の活用		
	パタン	人数	割合	パタン	人数	割合	パタン	人数	割合
1	10000000	260	.260	0000	280	.280	000	334	.334
2	11000000	80	.080	0100	116	.116	010	253	.253
3	01000000	64	.064	0010	108	.108	101	93	.093
4	00000000	52	.052	0001	91	.091	110	83	.083
5	10010000	37	.037	0101	74	.074	100	82	.082
6	11000001	30	.030	0011	60	.060	111	72	.072
7	10100000	29	.029	0110	55	.055	001	49	.049
8	10100001	25	.025	0111	45	.045	011	33	.033
9	00110000	22	.022	1000	41	.041	-	-	-
10	10000001	20	.020	1100	25	.025	-	-	-
	合計	619	.620		895	.896		999	1.000

表 5.24 イエメンデータにおける各アトリビュート領域の習得パタン

順位	数			図形と測定			資料の活用		
	パタン	人数	割合	パタン	人数	割合	パタン	人数	割合
1	10000000	162	.194	0000	305	.365	000	430	.514
2	10010000	53	.063	0100	127	.152	010	201	.240
3	10000100	41	.049	0010	83	.099	100	57	.068
4	01000010	34	.041	1000	73	.087	001	52	.062
5	10010100	34	.041	0110	40	.048	110	47	.056
6	11000000	24	.029	0001	37	.044	011	25	.030
7	10110000	23	.028	0101	29	.035	101	16	.019
8	10000001	22	.026	1010	29	.035	111	8	.010
9	10000010	22	.026	0111	28	.033	-	-	-
10	00000000	19	.023	0011	27	.032	-	-	-
	合計	434	.519		778	.931		836	1.000

表 5.25 アメリカデータにおける 2PL モデルを用いた潜在特性値を独立変数，R-RUM で推定した各アトリビュートの習得確率を従属変数としたときのロジスティック回帰分析の結果

従属変数 (アトリビュート)	推定値 上段：切片 下段：回帰係数	95%CI [下限，上限]	オッズ比	オッズ比95%CI [下限，上限]	p	R^2
数1	0.003	[-0.124 , 0.130]			.964	.205
	1.092	[0.921 , 1.264]	2.981	[2.511 , 3.540]	<.001	
数2	1.472	[1.229 , 1.716]			<.001	.659
	4.128	[3.623 , 4.632]	62.038	[37.449 , 102.770]	<.001	
数3	0.571	[0.428 , 0.713]			<.001	.259
	1.555	[1.349 , 1.762]	4.736	[3.853 , 5.822]	<.001	
数4	-0.146	[-0.278 , -0.014]			.030	.273
	1.329	[1.142 , 1.516]	3.777	[3.135 , 4.552]	<.001	
数5	-0.416	[-0.551 , -0.281]			<.001	.272
	1.356	[1.166 , 1.546]	3.879	[3.209 , 4.691]	<.001	
数6	1.159	[1.007 , 1.310]			<.001	.193
	1.073	[0.884 , 1.262]	2.924	[2.421 , 3.531]	<.001	
数7	-1.298	[-1.472 , -1.125]			<.001	.246
	1.751	[1.517 , 1.984]	5.758	[4.558 , 7.274]	<.001	
数8	0.776	[0.640 , 0.912]			<.001	.104
	0.990	[0.815 , 1.164]	2.690	[2.260 , 3.202]	<.001	
図形と測定9	-0.571	[-0.695 , -0.447]			<.001	.048
	0.463	[0.313 , 0.613]	1.588	[1.367 , 1.845]	<.001	
図形と測定10	-0.314	[-0.443 , -0.186]			<.001	.122
	1.079	[0.907 , 1.251]	2.942	[2.477 , 3.495]	<.001	
図形と測定11	-0.036	[-0.154 , 0.081]			.543	.011
	0.290	[0.149 , 0.431]	1.337	[1.161 , 1.539]	<.001	
図形と測定12	0.441	[0.312 , 0.571]			<.001	.188
	1.005	[0.836 , 1.175]	2.733	[2.307 , 3.238]	<.001	
資料の表現13	1.213	[1.056 , 1.371]			<.001	.227
	1.226	[1.028 , 1.425]	3.409	[2.795 , 4.158]	<.001	
資料の表現14	0.516	[0.374 , 0.658]			<.001	.295
	1.591	[1.382 , 1.799]	4.907	[3.983 , 6.044]	<.001	
資料の表現15	1.391	[1.236 , 1.547]			<.001	.106
	0.756	[0.571 , 0.940]	2.129	[1.770 , 2.561]	<.001	

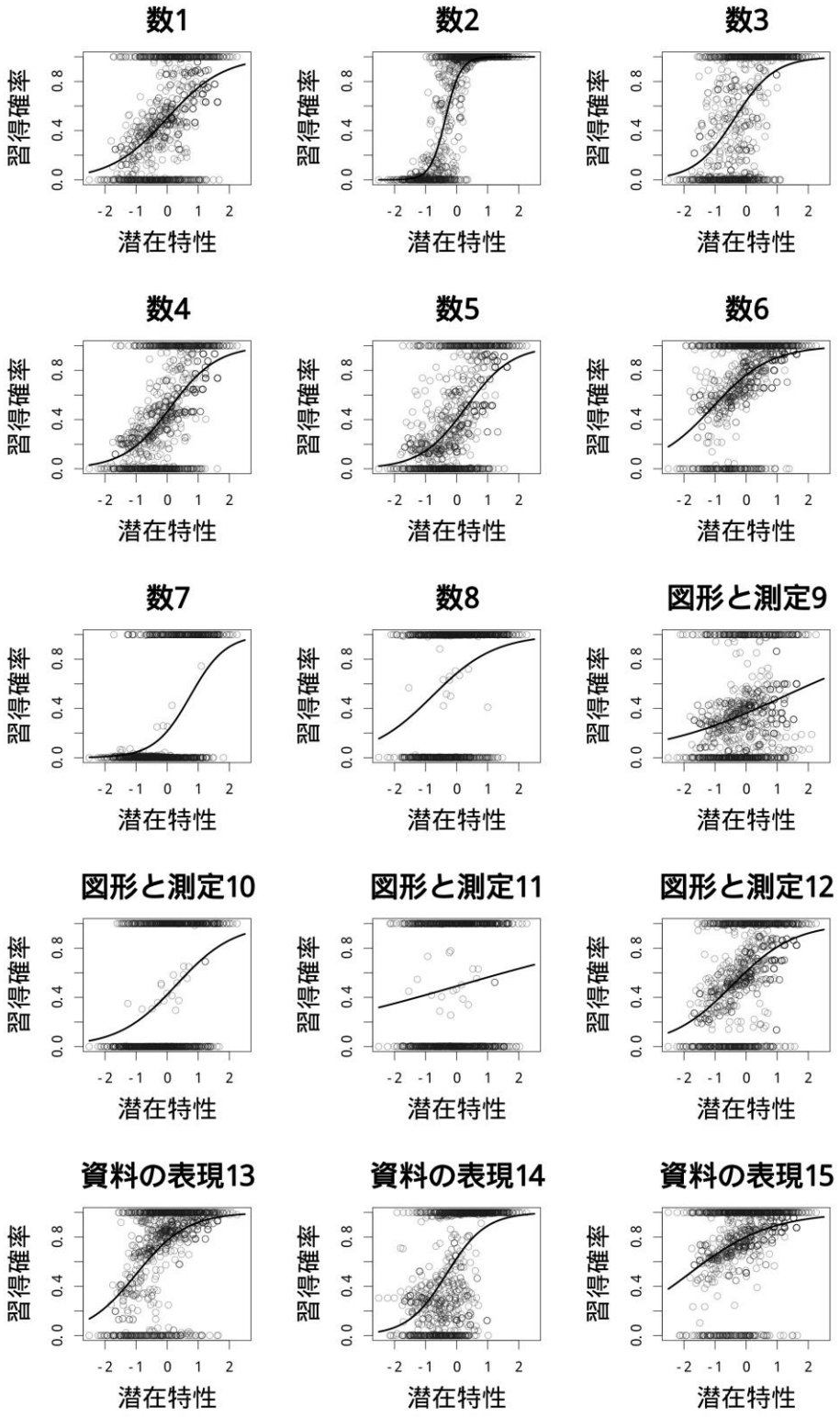


図 5.1 アメリカデータにおける各アトリビュートの習得確率と 2PL モデルで推定した潜在特性との関係

表 5.26 香港データにおける 2PL モデルを用いた潜在特性値を独立変数，LLM で推定した各アトリビュートの習得確率を従属変数としたときのロジスティック回帰分析の結果

従属変数 (アトリビュート)	推定値 上段：切片 下段：回帰係数	95%CI [下限，上限]	オッズ比	オッズ比95%CI [下限，上限]	p	R^2
数1	1.933	[1.639 , 2.228]			<.001	.263
	1.466	[1.140 , 1.792]	4.332	[3.128 , 5.999]	<.001	
数2	1.519	[1.238 , 1.800]			<.001	.460
	2.290	[1.879 , 2.701]	9.876	[6.550 , 14.892]	<.001	
数3	0.616	[0.422 , 0.811]			<.001	.196
	1.168	[0.905 , 1.430]	3.215	[2.473 , 4.180]	<.001	
数4	-0.211	[-0.382 , -0.040]			.016	.020
	0.311	[0.101 , 0.522]	1.365	[1.106 , 1.685]	.004	
数5	1.375	[1.135 , 1.615]			<.001	.257
	1.354	[1.061 , 1.648]	3.875	[2.888 , 5.199]	<.001	
数6	1.100	[0.896 , 1.304]			<.001	.085
	0.666	[0.426 , 0.906]	1.946	[1.532 , 2.473]	<.001	
数7	1.517	[1.225 , 1.809]			<.001	.389
	2.635	[2.172 , 3.099]	13.950	[8.777 , 22.173]	<.001	
数8	0.998	[0.802 , 1.195]			<.001	.052
	0.558	[0.326 , 0.791]	1.748	[1.386 , 2.205]	<.001	
図形と測定9	1.032	[0.838 , 1.227]			<.001	.021
	0.326	[0.097 , 0.555]	1.385	[1.102 , 1.742]	.005	
図形と測定10	0.837	[0.630 , 1.044]			<.001	.163
	1.322	[1.043 , 1.601]	3.751	[2.837 , 4.960]	<.001	
図形と測定11	0.483	[0.294 , 0.671]			<.001	.117
	1.082	[0.828 , 1.336]	2.951	[2.289 , 3.805]	<.001	
図形と測定12	2.696	[2.303 , 3.090]			<.001	.218
	1.307	[0.936 , 1.678]	3.696	[2.550 , 5.355]	<.001	
資料の表現13	0.612	[0.427 , 0.796]			<.001	.090
	0.707	[0.478 , 0.935]	2.028	[1.613 , 2.548]	<.001	
資料の表現14	0.451	[0.263 , 0.638]			<.001	.147
	1.047	[0.796 , 1.298]	2.848	[2.216 , 3.661]	<.001	
資料の表現15	2.251	[1.924 , 2.578]			<.001	.224
	1.291	[0.959 , 1.624]	3.638	[2.609 , 5.072]	<.001	

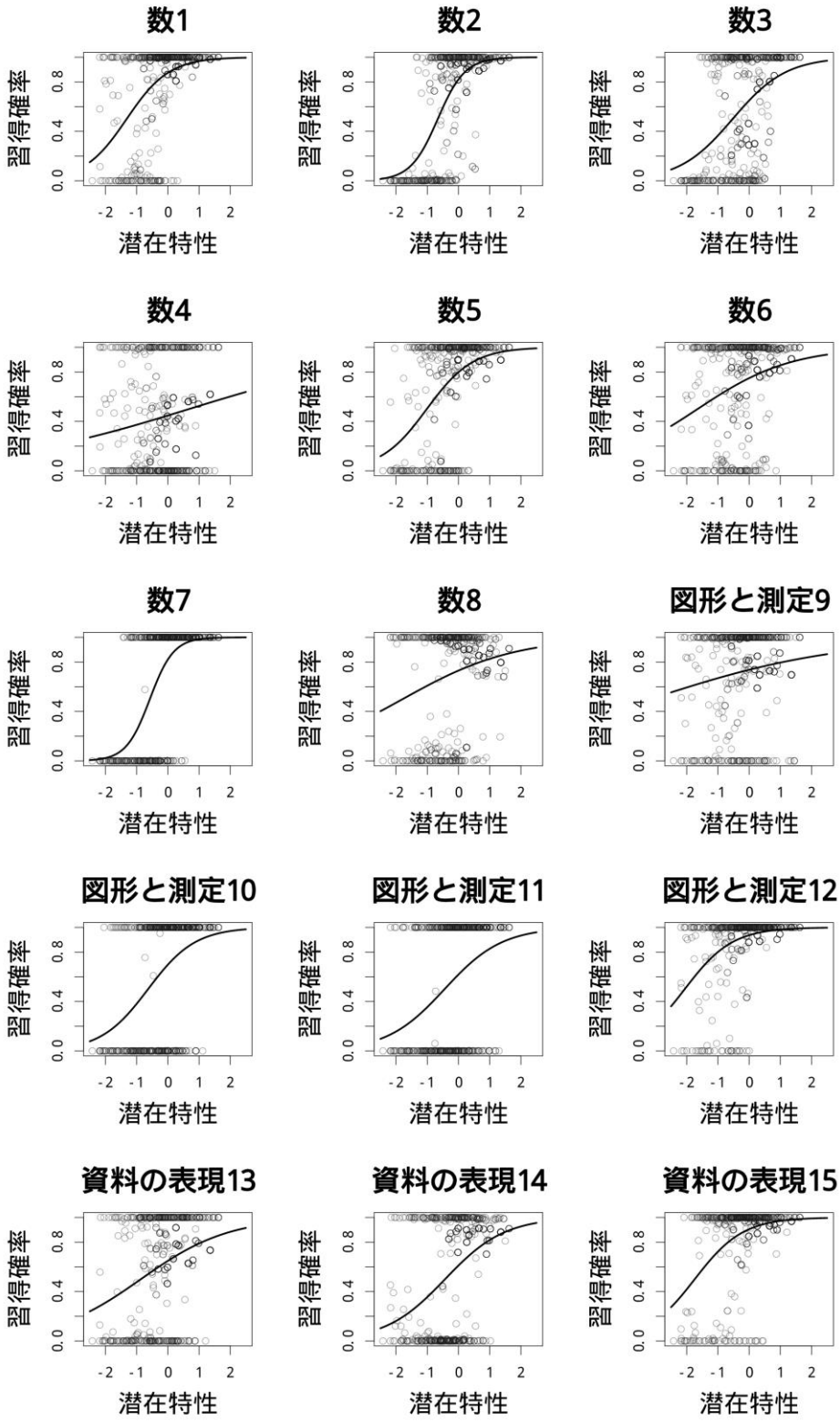


図 5.2 香港データにおける各アトリビュートの習得確率と 2PL モデルで推定した潜在特性との関係

表 5.27 シンガポールデータにおける 2PL モデルを用いた潜在特性値を独立変数，ACDM で推定した各アトリビュートの習得確率を従属変数としたときのロジスティック回帰分析の結果

従属変数 (アトリビュート)	推定値 上段：切片 下段：回帰係数	95%CI [下限，上限]	オッズ比	オッズ比95%CI [下限，上限]	p	R^2
数1	2.509	[2.181 , 2.836]			<.001	.358
	1.608	[1.299 , 1.918]	4.995	[3.666 , 6.807]	<.001	
数2	8.050	[6.108 , 9.992]			<.001	.845
	7.412	[5.565 , 9.259]	1655.392	[261.066 , 10496.675]	<.001	
数3	1.419	[1.171 , 1.666]			<.001	.455
	2.555	[2.171 , 2.939]	12.871	[8.767 , 18.897]	<.001	
数4	-0.110	[-0.263 , 0.043]			.159	.111
	0.728	[0.531 , 0.924]	2.071	[1.701 , 2.520]	<.001	
数5	1.860	[1.589 , 2.130]			<.001	.488
	2.103	[1.770 , 2.436]	8.190	[5.871 , 11.427]	<.001	
数6	2.042	[1.777 , 2.307]			<.001	.288
	1.345	[1.077 , 1.614]	3.839	[2.935 , 5.021]	<.001	
数7	0.661	[0.456 , 0.866]			<.001	.479
	2.547	[2.173 , 2.921]	12.770	[8.788 , 18.556]	<.001	
数8	1.239	[1.041 , 1.437]			<.001	.161
	1.235	[1.000 , 1.469]	3.437	[2.719 , 4.343]	<.001	
図形と測定9	0.068	[-0.094 , 0.229]			.413	.234
	1.198	[0.972 , 1.425]	3.314	[2.643 , 4.156]	<.001	
図形と測定10	1.640	[1.387 , 1.893]			<.001	.358
	2.160	[1.825 , 2.495]	8.673	[6.203 , 12.127]	<.001	
図形と測定11	0.297	[0.142 , 0.451]			<.001	.056
	0.685	[0.495 , 0.875]	1.984	[1.640 , 2.400]	<.001	
図形と測定12	1.042	[0.852 , 1.233]			<.001	.285
	1.328	[1.090 , 1.566]	3.772	[2.973 , 4.786]	<.001	
資料の表現13	1.433	[1.209 , 1.657]			<.001	.396
	1.715	[1.435 , 1.995]	5.557	[4.198 , 7.355]	<.001	
資料の表現14	0.873	[0.684 , 1.062]			<.001	.369
	1.585	[1.325 , 1.844]	4.877	[3.762 , 6.323]	<.001	
資料の表現15	2.580	[2.247 , 2.913]			<.001	.307
	1.466	[1.160 , 1.772]	4.332	[3.191 , 5.880]	<.001	

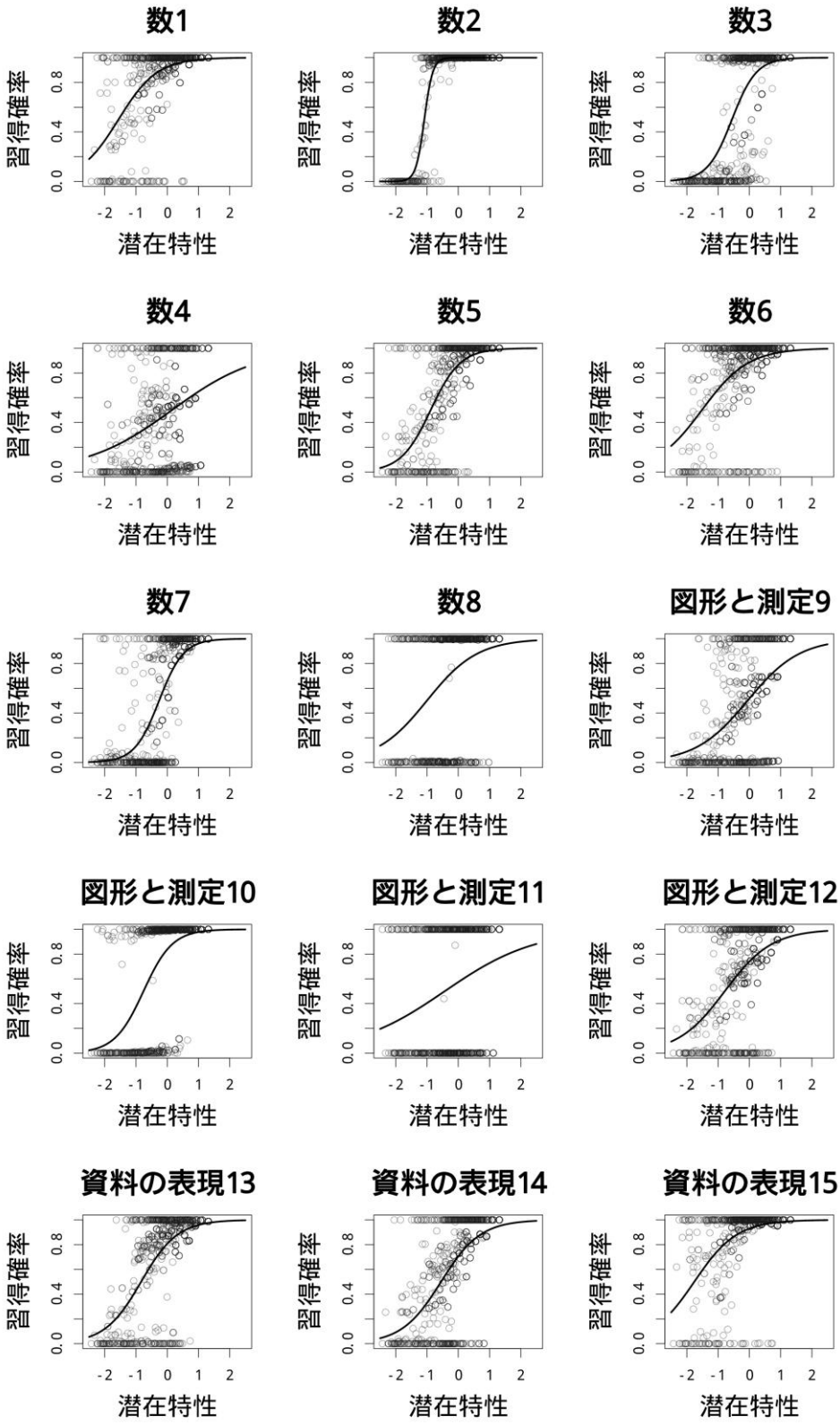


図 5.3 シンガポールデータにおける各アトリビュートの習得確率と 2PL モデルで推定した潜在特性との関係

表 5.28 スロベニアデータにおける 2PL モデルを用いた潜在特性値を独立変数，R-RUM で推定した各アトリビュートの習得確率を従属変数としたときのロジスティック回帰分析の結果

従属変数 (アトリビュート)	推定値 上段：切片 下段：回帰係数	95%CI [下限，上限]	オッズ比	オッズ比95%CI [下限，上限]	p	R^2
数1	1.160	[0.934 , 1.385]			<.001	.363
	1.775	[1.462 , 2.087]	5.898	[4.316 , 8.060]	<.001	
数2	1.745	[1.416 , 2.073]			<.001	.579
	3.357	[2.812 , 3.903]	28.717	[16.638 , 49.566]	<.001	
数3	1.053	[0.822 , 1.284]			<.001	.333
	2.136	[1.785 , 2.487]	8.467	[5.962 , 12.024]	<.001	
数4	1.805	[1.518 , 2.091]			<.001	.389
	2.003	[1.642 , 2.364]	7.408	[5.163 , 10.628]	<.001	
数5	-0.048	[-0.221 , 0.125]			.586	.204
	1.148	[0.909 , 1.386]	3.150	[2.483 , 3.998]	<.001	
数6	-0.598	[-0.781 , -0.415]			<.001	.224
	1.201	[0.950 , 1.452]	3.324	[2.587 , 4.271]	<.001	
数7	-1.593	[-1.844 , -1.342]			<.001	.216
	1.575	[1.251 , 1.899]	4.830	[3.493 , 6.680]	<.001	
数8	0.938	[0.734 , 1.141]			<.001	.198
	1.450	[1.176 , 1.725]	4.265	[3.241 , 5.613]	<.001	
図形と測定9	-0.485	[-0.650 , -0.320]			<.001	.039
	0.447	[0.246 , 0.647]	1.563	[1.279 , 1.911]	<.001	
図形と測定10	0.290	[0.125 , 0.455]			<.001	.058
	0.659	[0.453 , 0.864]	1.932	[1.573 , 2.373]	<.001	
図形と測定11	0.164	[0.004 , 0.325]			.045	.023
	0.425	[0.230 , 0.619]	1.529	[1.259 , 1.858]	<.001	
図形と測定12	0.718	[0.532 , 0.904]			<.001	.212
	1.176	[0.930 , 1.422]	3.242	[2.534 , 4.147]	<.001	
資料の表現13	-0.131	[-0.297 , 0.035]			.122	.118
	0.799	[0.585 , 1.012]	2.223	[1.796 , 2.752]	<.001	
資料の表現14	0.598	[0.403 , 0.792]			<.001	.266
	1.640	[1.353 , 1.927]	5.155	[3.870 , 6.866]	<.001	
資料の表現15	1.492	[1.241 , 1.743]			<.001	.261
	1.810	[1.483 , 2.137]	6.109	[4.404 , 8.474]	<.001	

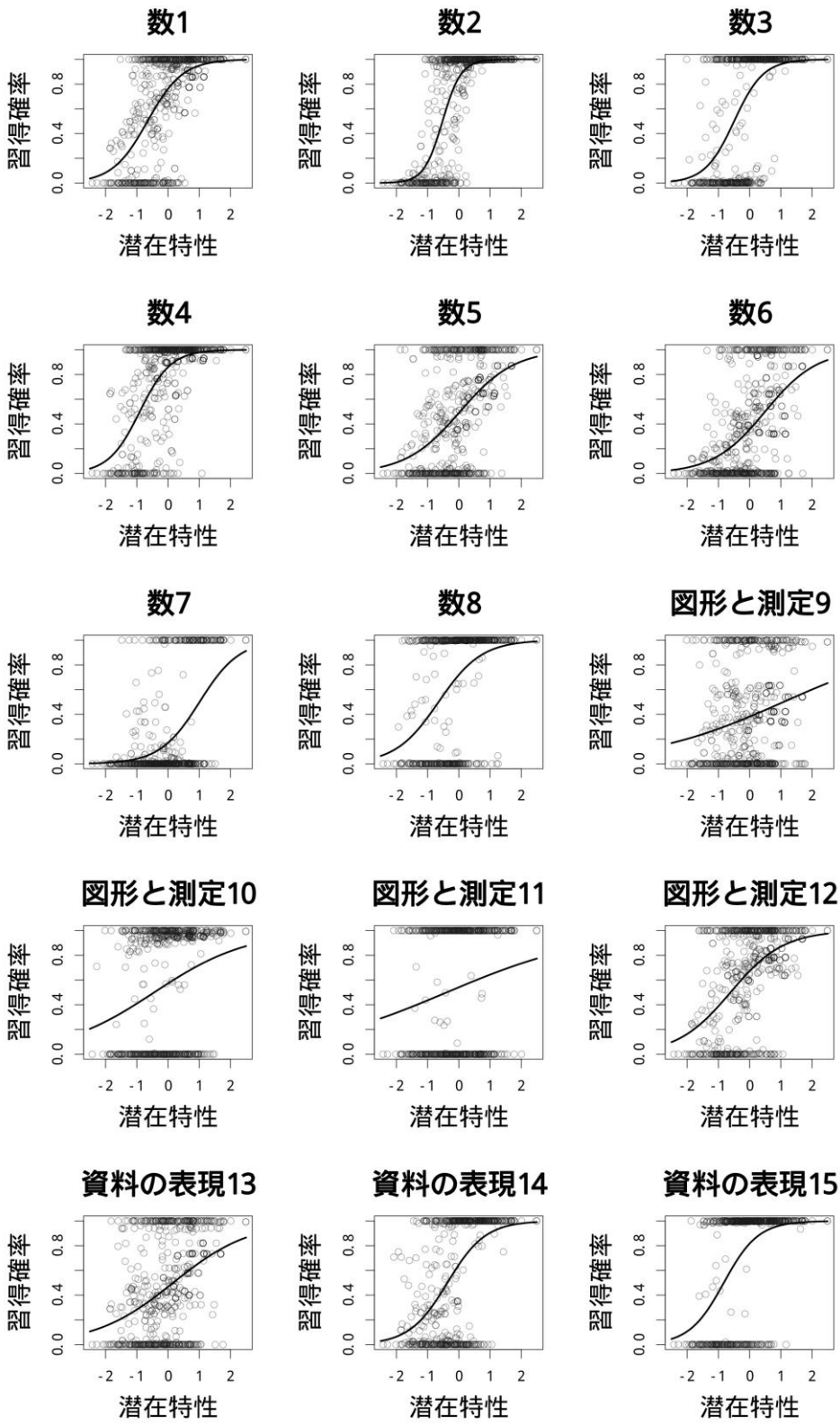


図 5.4 スロベニアデータにおける各アトリビュートの習得確率と 2PL モデルで推定した潜在特性との関係

表 5.29 アルメニアデータにおける 2PL モデルを用いた潜在特性値を独立変数，A-CDM で推定した各アトリビュートの習得確率を従属変数としたときのロジスティック回帰分析の結果

従属変数 (アトリビュート)	推定値 上段：切片 下段：回帰係数	95%CI [下限，上限]	オッズ比	オッズ比95%CI [下限，上限]	p	R^2
数1	0.500	[0.318 , 0.683]			<.001	.208
	1.113	[0.866 , 1.359]	3.042	[2.377 , 3.894]	<.001	
数2	0.387	[0.137 , 0.637]			.002	.628
	3.599	[3.022 , 4.176]	36.544	[20.522 , 65.073]	<.001	
数3	1.019	[0.792 , 1.246]			<.001	.345
	1.905	[1.571 , 2.238]	6.717	[4.813 , 9.373]	<.001	
数4	-0.469	[-0.639 , -0.299]			<.001	.059
	0.538	[0.322 , 0.755]	1.713	[1.380 , 2.127]	<.001	
数5	1.736	[1.509 , 1.963]			<.001	.000
	0.012	[-0.266 , 0.290]	1.012	[0.766 , 1.337]	.933	
数6	0.028	[-0.137 , 0.194]			.737	.059
	0.514	[0.305 , 0.723]	1.672	[1.357 , 2.060]	<.001	
数7	-0.101	[-0.299 , 0.097]			.316	.283
	1.929	[1.605 , 2.253]	6.881	[4.977 , 9.512]	<.001	
数8	1.002	[0.787 , 1.217]			<.001	.322
	1.569	[1.272 , 1.867]	4.804	[3.567 , 6.469]	<.001	
図形と測定9	-0.501	[-0.669 , -0.334]			<.001	.005
	0.152	[-0.054 , 0.359]	1.165	[0.948 , 1.432]	.147	
図形と測定10	0.489	[0.283 , 0.695]			<.001	.305
	1.987	[1.654 , 2.320]	7.293	[5.229 , 10.173]	<.001	
図形と測定11	0.372	[0.190 , 0.554]			<.001	.155
	1.202	[0.950 , 1.454]	3.328	[2.586 , 4.282]	<.001	
図形と測定12	1.261	[1.056 , 1.465]			<.001	.085
	0.669	[0.421 , 0.916]	1.952	[1.523 , 2.500]	<.001	
資料の表現13	0.142	[-0.043 , 0.327]			.133	.301
	1.445	[1.172 , 1.717]	4.240	[3.230 , 5.567]	<.001	
資料の表現14	0.028	[-0.142 , 0.199]			.746	.090
	0.840	[0.615 , 1.065]	2.316	[1.849 , 2.901]	<.001	
資料の表現15	-0.150	[-0.322 , 0.021]			.085	.141
	0.850	[0.624 , 1.077]	2.341	[1.866 , 2.937]	<.001	

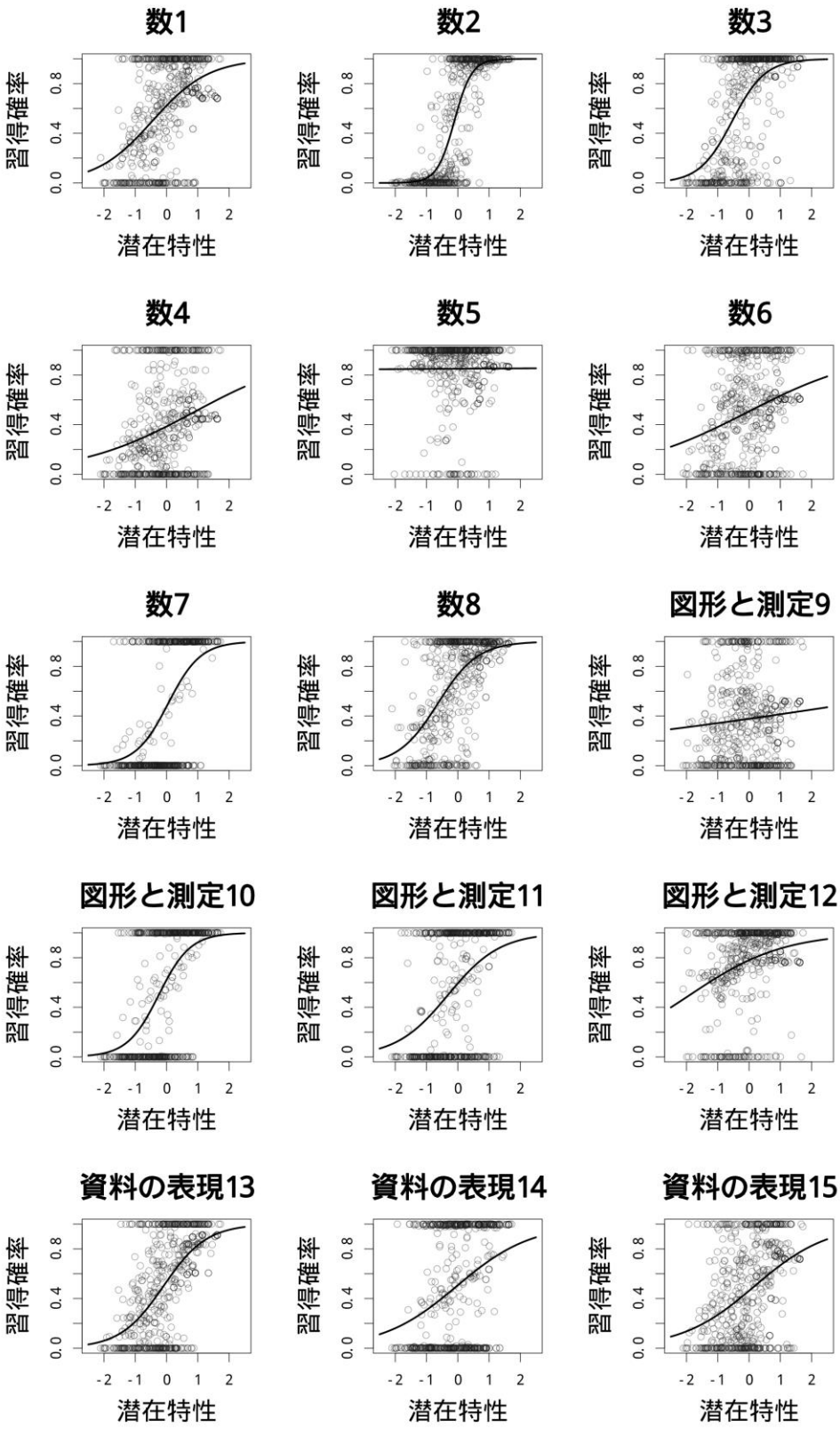


図 5.5 アルメニアデータにおける各アトリビュートの習得確率と 2PL モデルで推定した潜在特性との関係

表 5.30 カタールデータにおける 2PL モデルを用いた潜在特性値を独立変数，LLM で推定した各アトリビュートの習得確率を従属変数としたときのロジスティック回帰分析の結果

従属変数 (アトリビュート)	推定値 上段：切片 下段：回帰係数	95%CI [下限，上限]	オッズ比	オッズ比95%CI [下限，上限]	p	R^2
数1	0.644	[0.499 , 0.789]			<.001	.235
	1.662	[1.365 , 1.958]	5.268	[3.916 , 7.085]	<.001	
数2	-0.838	[-0.976 , -0.699]			<.001	.033
	0.611	[0.405 , 0.817]	1.842	[1.499 , 2.263]	<.001	
数3	-1.066	[-1.215 , -0.917]			<.001	.099
	0.932	[0.711 , 1.154]	2.541	[2.036 , 3.170]	<.001	
数4	-0.580	[-0.710 , -0.450]			<.001	.015
	-0.320	[-0.527 , -0.114]	0.726	[0.591 , 0.892]	.002	
数5	-1.570	[-1.735 , -1.405]			<.001	.001
	0.073	[-0.175 , 0.320]	1.075	[0.840 , 1.377]	.565	
数6	-1.636	[-1.809 , -1.462]			<.001	.058
	0.620	[0.384 , 0.856]	1.859	[1.468 , 2.354]	<.001	
数7	-2.228	[-2.438 , -2.017]			<.001	.007
	-0.338	[-0.683 , 0.007]	0.713	[0.505 , 1.007]	.055	
数8	-0.840	[-0.987 , -0.692]			<.001	.223
	1.434	[1.184 , 1.684]	4.196	[3.267 , 5.389]	<.001	
図形と測定9	-1.162	[-1.315 , -1.008]			<.001	.130
	0.950	[0.726 , 1.174]	2.586	[2.067 , 3.235]	<.001	
図形と測定10	-0.592	[-0.727 , -0.457]			<.001	.055
	0.850	[0.638 , 1.063]	2.340	[1.893 , 2.894]	<.001	
図形と測定11	-0.597	[-0.728 , -0.466]			<.001	.009
	0.334	[0.138 , 0.531]	1.397	[1.148 , 1.700]	<.001	
図形と測定12	-0.215	[-0.341 , -0.089]			<.001	.015
	0.327	[0.134 , 0.520]	1.387	[1.144 , 1.682]	<.001	
資料の表現13	-0.306	[-0.432 , -0.180]			<.001	.006
	-0.203	[-0.398 , -0.008]	0.817	[0.672 , 0.992]	.041	
資料の表現14	-0.292	[-0.424 , -0.160]			<.001	.095
	1.077	[0.850 , 1.304]	2.935	[2.339 , 3.682]	<.001	
資料の表現15	-0.793	[-0.930 , -0.656]			<.001	.041
	0.590	[0.386 , 0.795]	1.804	[1.471 , 2.214]	<.001	

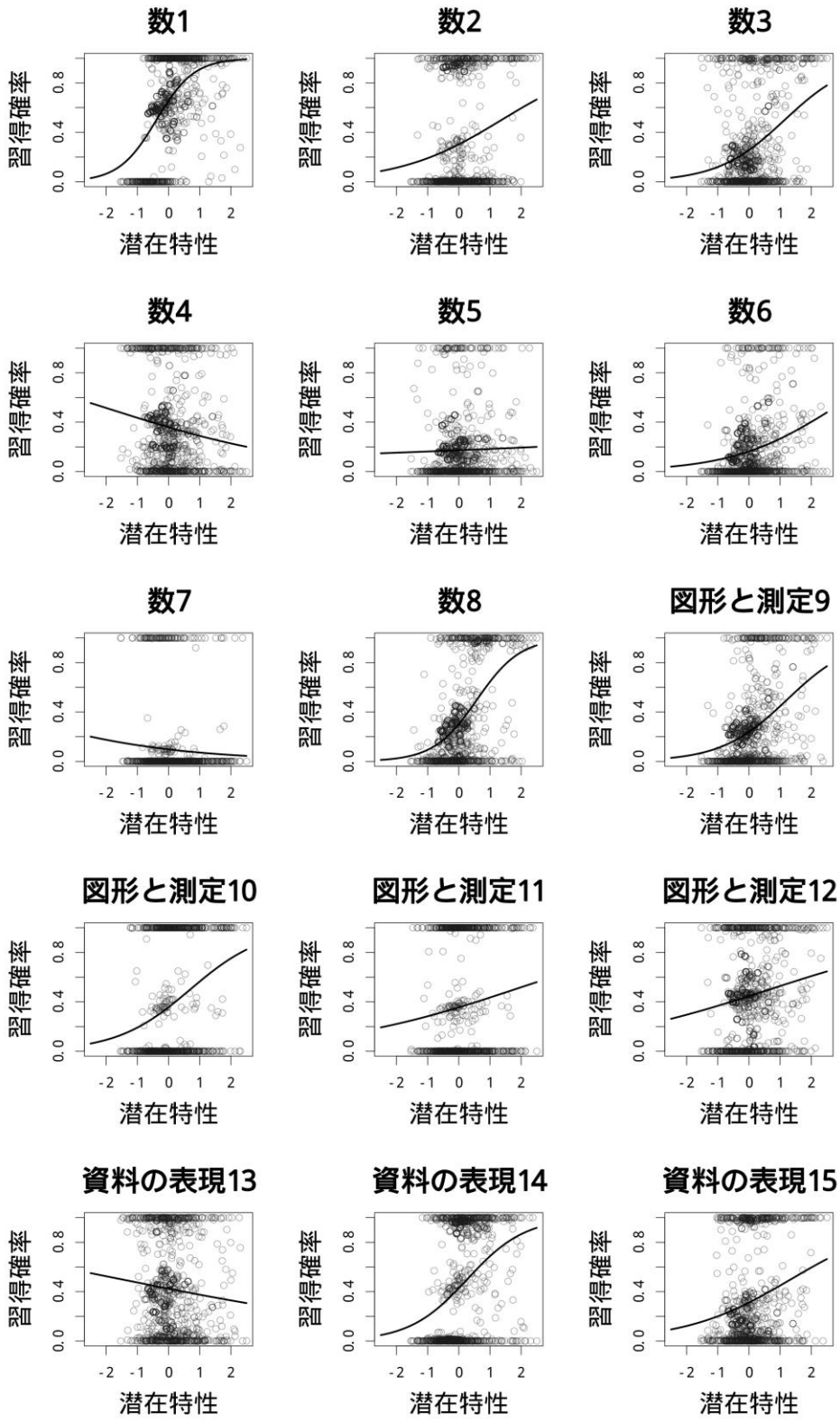


図 5.6 カタールデータにおける各アトリビュートの習得確率と 2PL モデルで推定した潜在特性との関係

表 5.31 イエメンデータにおける 2PL モデルを用いた潜在特性値を独立変数，A-CDM で推定した各アトリビュートの習得確率を従属変数としたときのロジスティック回帰分析の結果

従属変数 (アトリビュート)	推定値 上段：切片 下段：回帰係数	95%CI [下限，上限]	オッズ比	オッズ比95%CI [下限，上限]	p	R^2
数1	0.369	[0.227 , 0.511]			<.001	.074
	0.752	[0.519 , 0.985]	2.122	[1.681 , 2.678]	<.001	
数2	-1.222	[-1.390 , -1.055]			<.001	.048
	-0.834	[-1.114 , -0.553]	0.434	[0.328 , 0.575]	<.001	
数3	-1.015	[-1.181 , -0.848]			<.001	.192
	1.267	[1.007 , 1.526]	3.549	[2.738 , 4.602]	<.001	
数4	-0.379	[-0.519 , -0.239]			<.001	.033
	0.451	[0.243 , 0.660]	1.570	[1.275 , 1.935]	<.001	
数5	-0.880	[-1.035 , -0.726]			<.001	.086
	0.705	[0.480 , 0.930]	2.024	[1.617 , 2.534]	<.001	
数6	-0.858	[-1.014 , -0.702]			<.001	.127
	0.910	[0.676 , 1.144]	2.485	[1.967 , 3.140]	<.001	
数7	-1.897	[-2.150 , -1.645]			<.001	.290
	-2.858	[-3.339 , -2.377]	0.057	[0.035 , 0.093]	<.001	
数8	-1.204	[-1.376 , -1.032]			<.001	.156
	1.071	[0.821 , 1.320]	2.918	[2.274 , 3.745]	<.001	
図形と測定9	-0.832	[-0.980 , -0.683]			<.001	.005
	0.176	[-0.039 , 0.391]	1.193	[0.962 , 1.479]	.108	
図形と測定10	-0.805	[-0.964 , -0.645]			<.001	.137
	1.298	[1.038 , 1.559]	3.662	[2.823 , 4.752]	<.001	
図形と測定11	-0.912	[-1.065 , -0.759]			<.001	.019
	0.452	[0.233 , 0.670]	1.571	[1.262 , 1.955]	<.001	
図形と測定12	-0.953	[-1.111 , -0.796]			<.001	.086
	0.752	[0.524 , 0.980]	2.121	[1.689 , 2.665]	<.001	
資料の表現13	-0.990	[-1.143 , -0.836]			<.001	.001
	-0.070	[-0.297 , 0.157]	0.933	[0.743 , 1.170]	.547	
資料の表現14	-0.528	[-0.670 , -0.385]			<.001	.012
	0.346	[0.138 , 0.554]	1.413	[1.148 , 1.740]	.001	
資料の表現15	-2.219	[-2.467 , -1.971]			<.001	.122
	1.215	[0.919 , 1.511]	3.371	[2.507 , 4.532]	<.001	

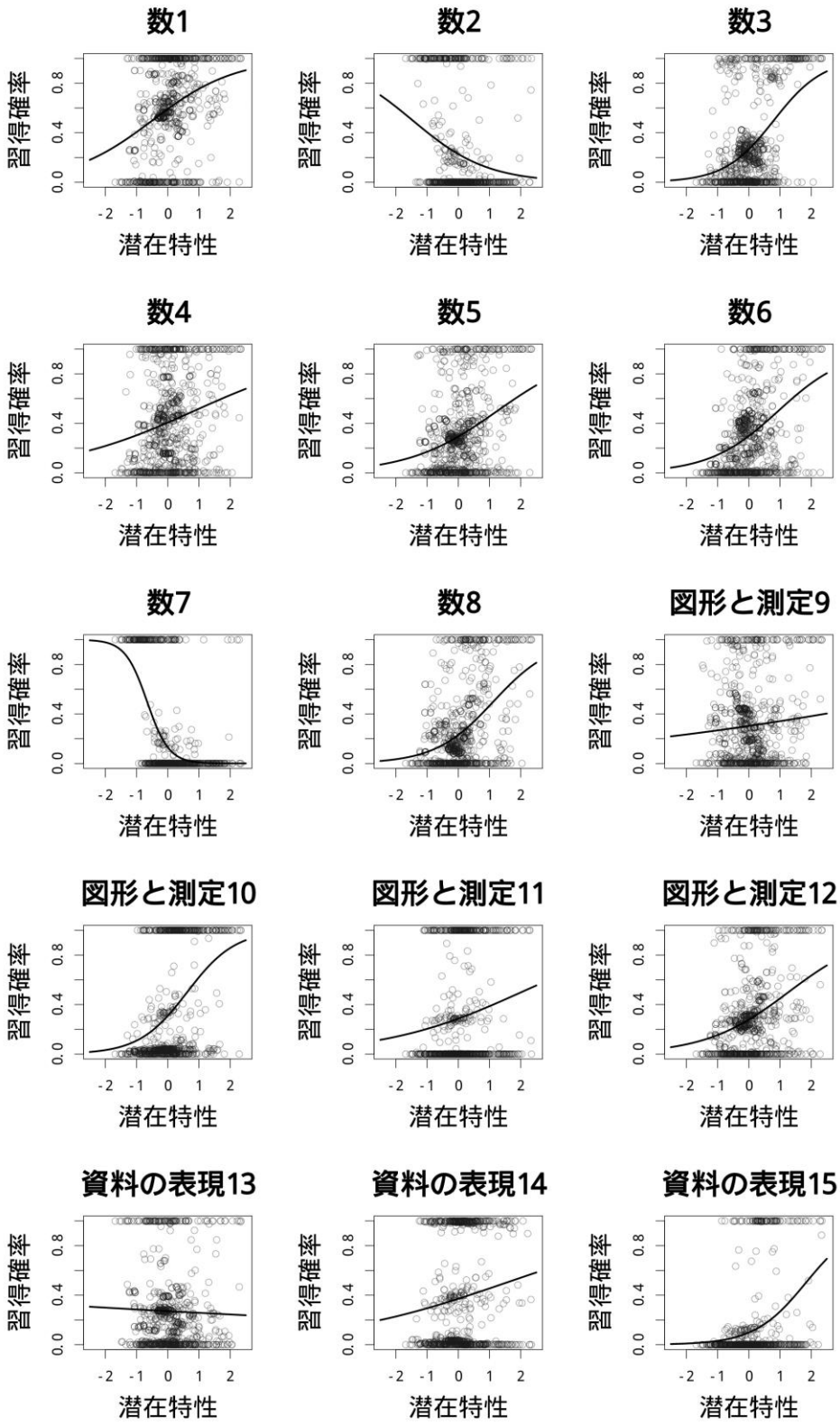


図 5.7 イエメンデータにおける各アトリビュートの習得確率と 2PL モデルで推定した潜在特性との関係

表 5.32 各国，各領域における 2PL モデルで推定した潜在特性値と習得アトリビュート数の相関係数

	数		図形と測定		資料の表現		全て	
国	*r*	95% CI	*r*	95% CI	*r*	95% CI	*r*	95% CI
アメリカ	.878	[.864, .891]	.558	[.517, .597]	.517	[.636, .701]	.913	[.903, .922]
香港	.850	[.825, .872]	.591	[.534, .643]	.534	[.588, .687]	.906	[.890, .920]
シンガポール	.879	[.861, .895]	.751	[.717, .781]	.717	[.730, .791]	.913	[.900, .925]
スロベニア	.895	[.878, .910]	.516	[.455, .571]	.455	[.647, .729]	.919	[.905, .930]
アルメニア	.853	[.829, .873]	.649	[.600, .694]	.600	[.618, .709]	.908	[.893, .921]
カタール	.534	[.488, .577]	.402	[.349, .453]	.349	[.222, .336]	.612	[.572, .650]
イエメン	.441	[.385, .494]	.426	[.368, .480]	.368	[.166, .294]	.531	[.481, .578]

表 5.33 TIMSS2007 の公式到達度スコアと各領域の平均アトリビュート習得数との相関

	r	*p*	95% CI
数	.962	<.001	[.758, .995]
図形と測定	.941	.002	[.645, .991]
資料の表現	.969	<.000	[.800, .996]
全て	.984	<.000	[.892, .998]

Note. N = 7(国の数)

表 5.34 アメリカデータにおける R-RUM の項目パタメタの推定値

			数								図形と測定				資料の活用		
項目番号	項目名	π	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	M041052	.919	1.000	.670													
2	M041056	1.000	.000				.000										
3	M041069	1.000	1.000	.969		.000	1.000										
4	M041076	1.000	.000		1.000			.000									
5	M041281	.960	1.000	.410	.678					1.000							
6	M041164	1.000	.862									.945		.862			
7	M041146	1.000	.015								1.000	1.000		.015			
8	M041152	1.000	.000	1.000	1.000							1.000	1.000				
9	M041258A	.895	.828									.828					
10	M041258B	.108	.713								.881	.713					
11	M041131	.857	.000	1.000	1.000	1.000					.000						
12	M041275	1.000	1.000												1.000		.000
13	M041186	1.000	1.000	1.000		1.000									.128		
14	M041336	.806	.904	.701			.951	.902							1.000	.000	
15	M031303	.918	.925	.543	.925												
16	M031309	.849	.795	.102	.795												
17	M031245	1.000	.000	.930					.000								
18	M031242A	.983	1.000	1.000	.328					1.000							
19	M031242B	1.000	.016	1.000	.006											.016	
20	M031242C	1.000	1.000	.999	.997					.003						1.000	
21	M031247	.358	1.000	.201	.518				1.000								
22	M031219	1.000	.999									.998	.001	.999			
23	M031173	.923	.828	.353	.828												
24	M031085	1.000	.000									.000					
25	M031172	1.000	.999	.689											.633		1.000

Note. GDINA関数ではR-RUMモデルの標準誤差が算出されないためここでは項目パラメタの推定値のみを示した。

表 5.35 香港データにおける LLM の項目パラメタの推定値および標準誤差

項目番号	項目名	切片	数								図形と測定				資料の表現		
			1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	M041052	-7.901	16.387	.724													
		(294.184)	(261.102)	(442.443)													
2	M041056	.233	8.977				8.977										
		(.331)	(401.071)				(401.071)										
3	M041069	-1.828	10.680	.358		-.055	10.680										
		(.660)	(262.009)	(.892)		(.922)	(262.009)										
4	M041076	-3.154	7.262		5.103			7.262									
		(6.889)	(8.952)		(6.941)			(8.952)									
5	M041281	-.599	.209	3.041	1.184					.209							
		(.561)	(.503)	(.641)	(.584)					(.503)							
6	M041164	3.716	1.071									4.424		1.071			
		(2.020)	(20.131)									(210.405)		(20.131)			
7	M041146	-7.648	16.060								.332	.466		16.060			
		(172.254)	(138.767)								(359.062)	(344.982)		(138.767)			
8	M041152	.119	.865	2.337	2.227							.285	-.738				
		(.681)	(.664)	(.780)	(.806)							(.731)	(.668)				
9	M041258A	2.079	-.163									-.163					
		(.423)	(.486)									(.486)					
10	M041258B	-8.199	.148								17.262	.148					
		(154.664)	(131.625)								(109.639)	(131.625)					
11	M041131	-9.210	.000	.000	.000	18.421					.000						
		(138.007)	(119.980)	(119.562)	(109.663)	(90.873)					(119.980)						
12	M041275	-1.343	.987												-.248		9.566
		(1.467)	(1.350)												(1.153)		(291.860)

表 5.35 つづき

項目番号	項目名	切片	数								図形と測定				資料の表現		
			1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
13	M041186	-9.181	.000	.030		.000									18.361		
		(143.355)	(141.955)	(131.404)		(87.094)									(105.921)		
14	M041336	-8.905	-.306	6.469			.000	4.485							.840	6.321	
		(9.818)	(1.128)	(9.623)			(1.046)	(1.372)							(1.251)	(9.796)	
15	M031303	.307	1.256	7.648	1.256												
		(.367)	(.541)	(99.898)	(.541)												
16	M031309	.124	1.664	4.049	1.664												
		(.407)	(.538)	(.787)	(.538)												
17	M031245	-8.677	16.718	1.169					16.718								
		(146.655)	(223.335)	(227.340)					(223.335)								
18	M031242A	-4.571	5.733	2.445	5.604					5.733							
		(1.356)	(1.421)	(.767)	(1.347)					(1.421)							
19	M031242B	-8.588	8.929	4.463	4.406											8.929	
		(14.391)	(14.339)	(11.317)	(5.677)											(14.339)	
20	M031242C	-6.560	-.328	6.687	.738					8.345						-.328	
		(71.805)	(.596)	(71.832)	(.609)					(71.901)						(.596)	
21	M031247	-9.210	.278	1.689	6.470				.278								
		(103.629)	(.456)	(.631)	(103.684)				(.456)								
22	M031219	-9.111	.632									.480	17.209	.632			
		(159.334)	(132.944)									(321.221)	(161.289)	(132.944)			
23	M031173	-1.877	4.687	5.234	4.687												
		(.593)	(.897)	(.826)	(.897)												
24	M031085	-9.210	18.421									18.421					
		(185.677)	(122.744)									(122.744)					
25	M031172	-3.638	1.390	4.582											2.491		4.386
		(1.155)	(1.031)	(1.330)											(1.610)		(1.260)

Note. カッコ内は推定値の標準誤差を意味する。

表 5.36 シンガポールデータにおける A-CDM の項目パラメタの推定値および標準誤差

項目番号	項目名	切片	数								図形と測定				資料の表現		
			1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	M041052	.001	.999	.000													
		(.282)	(.375)	(.255)													
2	M041056	.000	1.000				1.000										
		(.228)	(.353)				(.353)										
3	M041069	.310	.136	.511		.026	.136										
		(.079)	(.059)	(.089)		(.023)	(.059)										
4	M041076	.000	1.000		.000			1.000									
		(.181)	(.249)		(.151)			(.249)									
5	M041281	.410	.014	.520	.010					.014							
		(.083)	(.038)	(.085)	(.037)					(.038)							
6	M041164	.923	.000									.077		.000			
		(.038)	(.072)									(.149)		(.072)			
7	M041146	.000	1.000								.000	.000		1.000			
		(.165)	(.243)								(.087)	(.120)		(.243)			
8	M041152	.355	.182	.324	.091							-.030	.028				
		(.082)	(.090)	(.099)	(.056)							(.051)	(.036)				
9	M041258A	.682	.222									.222					
		(.057)	(.061)									(.061)					
10	M041258B	.000	.000								.996	.000					
		(.170)	(.142)								(.241)	(.142)					
11	M041131	.000	.000	.001	.001	.971					.000						
		(.258)	(.080)	(.235)	(.119)	(.309)					(.080)						
12	M041275	.009	.001												.000		.990
		(.212)	(.209)												(.175)		(.338)

表 5.36 つづき

項目番号	項目名	切片	数								図形と測定				資料の表現		
			1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
13	M041186	.000	.000	.215		.000									.785		
		(.297)	(.257)	(.207)		(.055)									(.096)		
14	M041336	.000	.000	.001			.000	.000							.000	.999	
		(.241)	(.204)	(.254)			(.171)	(.174)							(.161)	(.257)	
15	M031303	.437	.021	.529	.021												
		(.058)	(.021)	(.062)	(.021)												
16	M031309	.448	.002	.511	.002												
		(.063)	(.024)	(.066)	(.024)												
17	M031245	.038	.619	.344					.619								
		(.055)	(.048)	(.067)					(.048)								
18	M031242A	.462	.024	.088	.420					.024							
		(.072)	(.030)	(.076)	(.050)					(.030)							
19	M031242B	.000	.001	.033	.865											.001	
		(.103)	(.063)	(.116)	(.037)											(.063)	
20	M031242C	.001	.000	.002	.007					.990						.000	
		(.111)	(.096)	(.105)	(.082)					(.195)						(.096)	
21	M031247	.000	.533	.000	.298				.533								
		(.109)	(.056)	(.110)	(.060)				(.056)								
22	M031219	.000	.000									.000	1.000	.000			
		(.088)	(.091)									(.063)	(.171)	(.091)			
23	M031173	.291	-.013	.709	-.013												
		(.057)	(.031)	(.067)	(.031)												
24	M031085	.025	.969									.969					
		(.023)	(.024)									(.024)					
25	M031172	.277	.093	.371											.259		.000
		(.083)	(.055)	(.073)											(.062)		(.100)

Note. カッコ内は推定値の標準誤差を意味する。

表 5.37 スロベニアデータにおける R-RUM の項目パラメタの推定値

			数								図形と測定				資料の活用		
項目番号	項目名	π	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	M041052	1.000	.178	1.000													
2	M041056	1.000	.000				.000										
3	M041069	.165	.135	.016		.284	.135										
4	M041076	1.000	.000		.513			.000									
5	M041281	.901	1.000	.468	.933					1.000							
6	M041164	1.000	.973									.980		.973			
7	M041146	1.000	.000								1.000	1.000		.000			
8	M041152	.399	.774	.322	.791							.614	1.000				
9	M041258A	.791	.849									.849					
10	M041258B	.033	.044								.070	.044					
11	M041131	1.000	.045	1.000	1.000	1.000					.045						
12	M041275	1.000	.651												1.000		1.000
13	M041186	1.000	1.000	.997		.002									1.000		
14	M041336	1.000	1.000	1.000			1.000	1.000							.056	1.000	
15	M031303	.930	.802	.375	.802												
16	M031309	.821	.773	.422	.773												
17	M031245	1.000	.000	.286					.000								
18	M031242A	.993	.984	1.000	.000					.984							
19	M031242B	1.000	.001	1.000	.323											.001	
20	M031242C	1.000	1.000	1.000	1.000					.004						1.000	
21	M031247	.380	.528	.231	.412				.528								
22	M031219	1.000	1.000									1.000	.000	1.000			
23	M031173	.950	.899	.214	.899												
24	M031085	1.000	.074									.074					
25	M031172	1.000	1.000	.999											1.000		.009

Note. GDINA関数ではR-RUMモデルの標準誤差が算出されないためここでは項目パラメタの推定値のみを示した。

表 5.38 アルメニアデータにおける A-CDM の項目パラメタの推定値および標準誤差

項目番号	項目名	切片	数								図形と測定				資料の表現		
			1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	M041052	.000	1.000	.000													
		(.157)	(.312)	(.141)													
2	M041056	.000	1.000				1.000										
		(.234)	(.336)				(.336)										
3	M041069	.216	.076	.626		-.009	.076										
		(.075)	(.072)	(.059)		(.056)	(.072)										
4	M041076	.705	.102		.193			.102									
		(.068)	(.090)		(.072)			(.090)									
5	M041281	.123	.840	.037	.000					.840							
		(.136)	(.185)	(.124)	(.148)					(.185)							
6	M041164	.022	.978									.000		.978			
		(.152)	(.195)									(.136)		(.195)			
7	M041146	.541	.089								.235	.135		.089			
		(.097)	(.086)								(.077)	(.081)		(.086)			
8	M041152	.445	.113	.222	-.015							.044	.078				
		(.072)	(.069)	(.081)	(.071)							(.070)	(.072)				
9	M041258A	.742	.090									.090					
		(.063)	(.079)									(.079)					
10	M041258B	.000	.000								1.000	.000					
		(.181)	(.166)								(.273)	(.166)					
11	M041131	.000	.000	.000	.000	1.000					.000						
		(.206)	(.128)	(.144)	(.157)	(.348)					(.128)						
12	M041275	.000	.000												.000		1.000
		(.194)	(.193)												(.180)		(.231)

表 5.38 つづき

項目番号	項目名		数								図形と測定				資料の表現		
		切片	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
13	M041186	.000	.000	.000		.000									1.000		
		(.133)	(.137)	(.137)		(.146)									(.285)		
14	M041336	.000	.000	.000			.000	.927							.022	.051	
		(.143)	(.099)	(.098)			(.155)	(.232)							(.104)	(.138)	
15	M031303	.395	.266	.309	.266												
		(.046)	(.060)	(.051)	(.060)												
16	M031309	.631	.308	.054	.308												
		(.042)	(.052)	(.032)	(.052)												
17	M031245	.000	.997	.003					.997								
		(.062)	(.190)	(.103)					(.190)								
18	M031242A	.001	.228	.000	.772					.228							
		(.088)	(.129)	(.093)	(.078)					(.129)							
19	M031242B	.068	.008	-.068	.861											.008	
		(.068)	(.053)	(.063)	(.079)											(.053)	
20	M031242C	.000	.985	.000	.000					.015						.985	
		(.102)	(.198)	(.104)	(.109)					(.145)						(.198)	
21	M031247	.000	.077	.086	.055				.077								
		(.039)	(.039)	(.041)	(.044)				(.039)								
22	M031219	.001	.002									.001	.997	.002			
		(.118)	(.135)									(.068)	(.166)	(.135)			
23	M031173	.119	.132	.692	.132												
		(.040)	(.057)	(.051)	(.057)												
24	M031085	.000	1.000									1.000					
		(.073)	(.110)									(.110)					
25	M031172	.143	-.011	.280											.236		.335
		(.048)	(.076)	(.062)											(.083)		(.070)

Note. カッコ内は推定値の標準誤差を意味する。

表 5.39 カタールデータにおける LLM の項目パラメタの推定値および標準誤差

項目番号	項目名	切片	数								図形と測定				資料の表現		
			1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	M041052	-7.710	16.920	-1.036													
		(178.780)	(157.057)	(289.114)													
2	M041056	-9.210	16.071				16.071										
		(101.828)	(88.386)				(88.386)										
3	M041069	-7.864	-1.346	2.271		6.644	-1.346										
		(28.693)	(.627)	(.517)		(28.736)	(.627)										
4	M041076	-9.210	15.841		.802			15.841									
		(89.068)	(131.980)		(374.974)			(131.980)									
5	M041281	-8.437	10.131	-.706	7.516					10.131							
		(316.849)	(316.906)	(.649)	(316.886)					(316.906)							
6	M041164	-9.210	17.832									.582		17.832			
		(212.445)	(283.747)									(506.043)		(283.747)			
7	M041146	-7.004	2.172								10.310	-2.206		2.172			
		(18.377)	(8.170)								(20.088)	(4.660)		(8.170)			
8	M041152	-2.232	-.633	.687	2.215							-.994	-.640				
		(.419)	(.334)	(.423)	(.412)							(.424)	(.373)				
9	M041258A	-1.467	.162									.162					
		(.194)	(.349)									(.349)					
10	M041258B	-9.210	2.639								5.207	2.639					
		(90.349)	(1.350)								(90.330)	(1.350)					
11	M041131	-9.210	.242	.394	1.524	7.993					.242						
		(113.437)	(.476)	(.543)	(.449)	(113.506)					(.476)						
12	M041275	-5.057	.875												-4.153		6.749
		(8.123)	(2.358)												(2.581)		(8.522)

表 5.39 つづき

項目番号	項目名	切片	数								図形と測定				資料の表現		
			1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
13	M041186	-8.857	.048	-.354		6.417									3.820		
		(8.590)	(.633)	(.989)		(8.530)									(.685)		
14	M041336	-4.488	.308	.554			3.052	-1.608							3.344	-2.488	
		(.801)	(.419)	(.501)			(.552)	(.826)							(.740)	(.555)	
15	M031303	-3.494	-3.763	8.336	-3.763												
		(.616)	(2.193)	(2.220)	(2.193)												
16	M031309	-5.885	2.827	2.773	2.827												
		(.556)	(.527)	(.486)	(.527)												
17	M031245	-9.210	14.657	.921					14.657								
		(187.743)	(208.602)	(110.105)					(208.602)								
18	M031242A	-9.210	8.192	.635	5.984					8.192							
		(19.533)	(19.464)	(.492)	(19.420)					(19.464)							
19	M031242B	-9.210	3.490	2.887	4.722											3.490	
		(1.018)	(.550)	(.609)	(.775)											(.550)	
20	M031242C	-5.661	9.641	1.793	.124					3.314						9.641	
		(3.930)	(6.629)	(1.061)	(.811)					(3.775)						(6.629)	
21	M031247	-6.010	-.778	3.673	-2.422				-.778								
		(1.299)	(7.278)	(1.336)	(1.304)				(7.278)								
22	M031219	-9.208	.847									-.002	16.742	.847			
		(84.835)	(224.336)									(162.522)	(135.058)	(224.336)			
23	M031173	-6.185	5.753	4.249	5.753												
		(1.310)	(1.310)	(1.281)	(1.310)												
24	M031085	-9.210	18.241									18.241					
		(159.756)	(82.319)									(82.319)					
25	M031172	-8.700	-.510	2.119											4.574		6.573
		(3.198)	(1.003)	(1.162)											(1.736)		(2.798)

Note. カッコ内は推定値の標準誤差を意味する。

表 5.40 イエメンデータにおける A-CDM の項目パラメタの推定値および標準誤差

項目番号	項目名	切片	数								図形と測定				資料の表現		
			1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	M041052	-7.485	16.695	-1.082													
		(165.854)	(160.942)	(359.425)													
2	M041056	-9.210	10.055				10.055										
		(21.601)	(21.604)				(21.604)										
3	M041069	-8.761	3.873	-.450		7.247	3.873										
		(22.433)	(1.057)	(1.849)		(22.466)	(1.057)										
4	M041076	-9.210	8.779		4.180			8.779									
		(27.926)	(27.950)		(1.758)			(27.950)									
5	M041281	-9.210	16.697	.748	.940					16.697							
		(136.349)	(251.025)	(437.051)	(409.740)					(251.025)							
6	M041164	-9.210	16.809									.825		16.809			
		(127.948)	(191.805)									(372.225)		(191.805)			
7	M041146	-9.210	.467								7.141	1.890		.467			
		(60.219)	(.716)								(60.273)	(.729)		(.716)			
8	M041152	-3.008	-.975	1.936	4.329							-1.662	.091				
		(.879)	(.476)	(.867)	(.950)							(.513)	(.444)				
9	M041258A	-2.101	1.322									1.322					
		(.315)	(.455)									(.455)					
10	M041258B	-9.210	4.968								4.736	4.968					
		(16.765)	(15.865)								(4.864)	(15.865)					
11	M041131	-5.677	4.613	-1.017	-2.516	5.132					4.613						
		(4.979)	(4.912)	(.889)	(4.641)	(5.000)					(4.912)						
12	M041275	-9.210	3.453												3.567		2.023
		(33.629)	(31.829)												(13.334)		(1.110)

表 5.40 つづき

項目番号	項目名	切片	数								図形と測定				資料の表現		
			1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
13	M041186	-6.233 (74.156)	-.825 (3.265)	-.065 (1.655)		-2.087 (4.323)									10.427 (76.143)		
14	M041336	-8.552 (158.309)	-.369 (.729)	-.193 (.814)			1.771 (.657)	7.705 (158.441)							-.051 (.839)	-.046 (.558)	
15	M031303	-.614 (.128)	.196 (.264)	.093 (.219)	.196 (.264)												
16	M031309	-9.210 (34.548)	8.281 (34.586)	3.802 (6.918)	8.281 (34.586)												
17	M031245	-9.210 (186.127)	15.828 (251.396)	1.196 (145.471)					15.828 (251.396)								
18	M031242A	-4.559 (.847)	2.178 (.639)	.525 (.618)	3.413 (.836)					2.178 (.639)							
19	M031242B	-4.439 (3.160)	-4.771 (63.913)	1.373 (1.489)	2.598 (3.404)											-4.771 (63.913)	
20	M031242C	-7.877 (20.201)	10.403 (20.706)	1.085 (2.456)	2.610 (10.007)					2.988 (15.235)						10.403 (20.706)	
21	M031247	-9.210 (142.024)	.245 (1.537)	1.772 (1.988)	5.791 (142.933)				.245 (1.537)								
22	M031219	-9.210 (87.949)	.685 (341.016)									.433 (201.777)	16.611 (128.566)	.685 (341.016)			
23	M031173	-8.994 (120.879)	-.216 (328.437)	16.811 (209.763)	-.216 (328.437)												
24	M031085	-9.210 (383.712)	11.826 (384.073)									11.826 (384.073)					
25	M031172	-9.210 (232.972)	.054 (95.685)	.265 (102.353)											.790 (183.880)		14.879 (266.958)

第6章　総合考察

本研究では，第 2 章で現在までに提案されている CDM の最近の展開を，発展的なモデルを含めて網羅的に整理・検討し，モデルの再分類を行った。また，CDM において重要な Q 行列の推定や誤設定についても検討を行った。誤設定については第 3 章でシミュレーション実験を用いてさらに詳細な検討を行った。第 4 章・第 5 章では，CDM どうしの比較に重点をおき，特に CDM の複雑さの観点から分類した倹約モデル，主効果モデル，飽和モデルのどのモデルが選択されるのかを検討した。それらのモデルから得られるアトリビュート習得パタンや項目パラメタの推定結果を通じたモデルの評価を行った。

本章では第 2 章から第 5 章までで得られた知見について，6-1 で各章ごとに概観する。さらに，本研究の限界と展望について 6-2 で議論を行う。特に，限界点・展望としては，比較に用いたモデルについて，TIMSS データの特殊性，アトリビュートの妥当性，Q 行列の設定の妥当性，モデル比較に関して最近のベイズ統計学の知見を用いた議論を行う。最後に，今後の発展的な課題として，理論的には新たに開発が望まれる CDM について，応用的には TIMSS データを分析するにあたり必要となりうる各国の教育制度と CDM を用いた結果の対応関係・解釈についての議論を行う。

6-1 本研究で得られた知見

6-1-1　第 2 章で得られた知見

第 2 章では，一般的に行われる CDM の区別であるアトリビュートの補償・非補償関係によるモデルの区別ではなく，モデルのパラメタ数に着目した 3 つの分類である倹約モデル・主効果モデル・飽和モデルの 3 つの分類を提案した。これまでに多くなされていたモデル分類の観点としては，アトリビュートの補償・非補償関係に注目したモデルの区分があるが，これは多次元 IRT モデルにもみられるモデルの区分でもある。一方，本研究で注目した，主効果モデルなどの区分は G-DINA モデルにおける各パラメタの意味に注目したモデルの区分である。主効果や交互作用といった用語は線形モデルや分散分析モデルでも通常使われる用語であり，こうした用語を利用することにより CDM の各パラメタの特徴を容易に理解することができる。また，CDM の交互作用項はアトリビュートの非補償関係として解釈することもできる。本論文で行ったモデルの分類は交互作用の有無や制約の量などモデルの複雑さを反映したものである。補償・非補償関係が認知モデルやテストの理論的想定から導かれるアトリビュートについての関心を反映している一方で，倹約モデル・主効果モ

デル・飽和モデルといった区分はパラメタの多寡に注目したシンプルな分類であるといえる。理論的には G-DINA モデルや LCDM といった飽和モデルによって包括的なモデルが提案されたと考えられるが，具体的なモデルを使用する文脈でモデルの性質を適切に反映した分類として，倹約モデル，主効果モデル，飽和モデルという観点は有用であると考えられる。

本研究で経験的な検討の対象としたのは，応用場面で最も利用される可能性の高い 2 値反応変数に適したモデル群であった。しかし，多肢選択肢もまたテスト形式としては広く採用されているものであり，正答以外の選択肢も診断に有用な情報を含んでいる可能性がある。2 値データが含む情報よりもこうした多肢選択肢などにより補助的な情報を増やすということは，診断を正確に行うにあたり有用である可能性がある。ただし，そういった補助情報を含む解答反応のモデル化には選択肢も含めた問題セットの性質を理解していなければならない点に留意が必要である。診断に適した情報をテストに盛り込むためには，テスト項目自体の解答形式の設計や内容的な側面も適切に設計する必要があり，テストの内容，反応モデル，解答プロセスの全容を理解している必要が生じる。

このような多値反応に適した CDM は有用であるものの，2 値反応モデルに比して適用が難しい可能性がある。2 値反応モデルの方が Q 行列の作成などが比較的容易だが 1 項目に含まれる情報が相対的に少ない可能性がある。その一方，多値項目モデルの方が 2 値反応モデルよりも 1 項目に含まれる情報が多くなる一方で，解析や結果の解釈を行うのが難しくなるといえる。こうした観点も含めて，どのようなモデルを利用するのかをテストの目的に則して判断する必要がある。

また，本研究で検討したように 2 値反応モデルの中にも多様なモデルが開発されており，多値反応モデルや他の拡張モデルでなくとも一定の実用性を有しているモデルも多いと考えられる。統計的なモデル選択の観点だけでなく，選択されたモデルの解釈ができることも必要であり，2 値反応モデルは解釈のしやすさからも重要なモデル群であると考えられる。IRT モデルでも実用上は 1~3PL モデルを利用することが多いのと同様に，CDM においても実データ解析に利用されるモデルとして，2 値反応に適したモデルの有用性の検討が必要である。

6-1-2　第 3 章で得られた知見

第 3 章では現実的なテストの状況で生じるアトリビュート階層構造を考慮した場合の Q

行列における誤設定の影響を検討した。これにより，1. 分岐型での誤設定でアトリビュート習得パタンの推定に大きな問題が生じうること，2. 直線型や収束型などの誤設定の影響は若干見られるもの相対的にみて小さいこと，3. アトリビュートの設定の正しい項目を増やし誤設定の Q 行列の大きさに対する割合を小さくすることにより誤設定の影響を緩和できること，4. 特定の構造において誤設定は設定の正しい項目のパラメタにも影響を及ぼすこと，という 4 点が明らかになった。これらのことから，診断を正確にするためにはアトリビュートの設定が正しい項目を増やすことが重要であると示唆された。

このように，アトリビュートの階層構造を考慮した場合，誤設定の影響を受けにくい構造と誤設定の影響が大きい構造がある可能性が示唆された。シミュレーション研究の結果から，アトリビュート階層構造を考慮し，誤設定に脆弱な構造を避けるようなテストの構成を行うことにより，誤設定の影響を軽減する必要があるという点を示唆した。

また，今回の誤設定研究は DINA モデルを用いていたため，誤設定の影響が大きくなってしまった可能性もある。DINA モデルでは 1 つの項目で 2 種類のアトリビュート習得パタンの正答確率しか区別ができない。一方，主効果モデルや飽和モデルではアトリビュートの付与のされ方によっては 1 つの項目であっても複数のアトリビュート習得パタンを区別出来る。こうしたことから誤設定が DINA モデル以外の主効果モデルや飽和モデルであった場合には影響が小さい可能性もある。今後の研究としては，DINA モデル以外のモデルでの誤設定の影響についても検証する必要があるだろう。

6-1-3 第 4 章で得られた知見

第 4 章では，TIMSS2007 の 4 年生データの日本人データを用いてモデル比較を行った結果，IRT モデルよりも CDM が選択され，主効果モデルの 1 つである R-RUM モデルの適合が比較的よいことが示された。また，CDM で仮定するアトリビュートは一般的な算数能力を十分に反映したものであることが示唆された。さらに，使用する CDM によって，推定されるアトリビュート習得パタンにはかなりの違いが生じる可能性が示唆された。項目パラメタの推定結果から，日本の TIMSS 2007 の算数データにおいては，複数のアトリビュートが必要な項目とそうではない項目が混在することが示された。また，G-DINA モデル，R-RUM，DINA モデルの 3 種類のモデルを適用した場合のアトリビュート習得パタンを検討し，利用するモデルによって，推定されるアトリビュート習得パタンが異なっている場合があることが示された。

6-1-4　　第 5 章で得られた知見

第 5 章では，日本人データで得られた結果の一般化がどれほど可能なのかを検討するために，TIMSS の公式到達度スコアが高い国から低い国まで様々なタイプの解答者の解答に対し，それらに適したモデルを探索し，その傾向を検討することで，どのような形式の CDM が広く有用であるのか検討した。その結果として，TIMSS データにおいては，いずれも主効果モデルである R-RUM や A-CDM，LLM が選択される傾向がみられた。この結果から，TIMSS に適合するモデルは TIMSS のスコアの高低によらず一貫していることが示唆された。また日本人データと同様に，IRT モデルによって推定された潜在特性や TIMSS の公式スコアとアトリビュートとの相関分析から，アトリビュートは全般的な算数能力を反映している可能性が示唆された。更に，日本人データと同様に，項目パラメタの推定値から，複数のアトリビュートが項目の正答に必要な項目と，単一のアトリビュートのみで正答可能な項目の相違が示唆された。また，国によらず共通のアトリビュートが必要である項目と，国によって異なったアトリビュートが必要な項目が存在する可能性も示された。各国のアトリビュート習得パタンについて，TIMSS の公式スコアと関連して習得しているアトリビュート数が異なっている傾向がみられた。また国によって，習得されている割合が多いアトリビュートとそうでないアトリビュートに違いがある傾向がみられた。

以上より，モデル比較の観点についてまとめると，TIMSS の小学校 4 年生データでは，国によって採用されるモデルが若干異なっているものの，主に R-RUM，A-CDM，LLM といった主効果モデルが支持される傾向が示された。一方で，これまで CDM の代表的なモデルとされてきた DINA モデルや DINA モデルを補償モデルとして定式化した DINO モデルは倹約モデルであり，これらの倹約モデルは概して適合が悪かった。また，理論的には補償モデル，非補償モデルを包含する飽和モデルは確かに尤度においては事前に予測されたように最もよい対数尤度の値であったものの，情報量規準 AIC，BIC の観点からは選択されないという結果が得られた。

今回相対指標を用いて選択されたモデルは全て主効果モデルであり，項目に必要な各アトリビュートがどの程度問題正答確率に寄与するかを検討できる。しかし，それぞれの国で選択されたモデルが異なっていたという点に注意が必要といえよう。そのため，第 5 章で述べたように，項目パラメタの推定値の直接的な大小の比較というより，それぞれの項目でどのアトリビュートが正答に大きく寄与していたのかという傾向についての比較が望ましい。

また，今回得られた項目パラメタの結果については，推定値の標準誤差が極めて大きいものもあり，解釈には注意が必要である点は改めて述べておく。

こうした注意点はあるものの，CDM を適用することによってこれまでは十分に検討されていなかったアトリビュートと問題の正答確率の関係を検証できるようになった点は重要であろう。

6-2 限界・展望

本節では本研究の限界と展望を示す。まず，本研究の限界点として，比較に用いたモデルの限定性について議論を行う。さらに，本研究で利用した実データが TIMSS に限定されていた点についての考察を行う。これに加え，本研究の経験的検討で用いた Lee et al.（2011）の用いた Q 行列の設定についても考察を行う。本研究の経験的検討で用いた Q 行列はその作成手続きなどの観点からは妥当なものであると考えられる一方で，実データ解析の結果からは必ずしも適切でない可能性が指摘できる。先に議論したアトリビュートの設定と合わせて，Q 行列の設定についても議論が必要である。

また，今後の展望としては，TIMSS で収集されている各国の学習環境や教育制度と CDM の結果の対応関係について議論する必要性を述べる。CDM は計量心理学的なモデルであり，そのモデルがどのような意味を持つのかを議論するためには，解答者の教育背景や教科教育の状況についての議論も必要と考えられる。さらに，モデル比較に関しては，ベイズ統計学の最近の知見を用いることもできる。今後の研究で，こうした知見を利用するために，どのような観点からのモデル比較が実行可能であるのか議論を行う。最後に，これまでの議論を踏まえて，今後の CDM で必要なモデル開発の方向性について議論を行う。今回の研究は既存のモデルのうちどれが有用であるかを検討したものであり，これらの知見を踏まえた新しいモデルを開発する必要が考えられるため，この点について議論を行う。

6-2-1 比較モデルの限定性

本研究の経験的検討で用いた CDM は本研究で提案したパラメタ数に基づく分類に含まれるモデル群に限定されていた。これらのモデル群は G-DINA モデルとその下位モデルを含む。そのため，応用上利用しやすいモデル群である。しかし，第 2 章で検討したように，CDM を冠するモデルは無数に存在する。モデル比較研究ではどのようなモデルを比較対象にするのかによって，異なった知見が得られうる。本研究で利用しなかった他のモデルが

TIMSS データに適合する可能性もある。

しかしながら, CDM を利用する場面で常に全てのモデルを比較してモデル選択を行うということは現実にはない。また特定のモデルを志向したテスト設計を行っていれば, 比較するべきモデルは限定される。ただし, 診断目的に作成されたテストではないテストから CDM を用いて診断情報を抽出する際には背後にある認知モデルが明示的でないこともある。こういった場合には可能な限り広くモデル比較を行い，適切なモデルを選択することが適切な診断のために必要である。特に，既存のテストに CDM を適用する場合には特殊な目的がない限り，G-DINA とその下位モデルを比較することが第 1 の選択肢として挙げられる。これは，ソフトウェアでの実行が容易であるという観点からも現実的な選択肢と考えられる。特定の目的がある場合には既存の CDM ではなく, 専用のモデルを開発する必要があるかもしれず, その場合にも比較候補となるモデルは非常に限定的である。このような観点からいえば，本研究が行った分析は応用的な分析に際して実際に行われうる比較的一般的なモデル比較の手続きを行ったと見なすことができ，実データでの項目パラメタの推定値やアトリビュート習得パタンの推定値を知るための研究としては目的を達成していると考えることができる。

6-2-2 TIMSS データの特殊性

先にも述べたように，TIMSS データは基本的に IRT モデルを適用することを目指して作成されたテストであり，もともとは一次元の潜在特性を想定したものである。このため，結果の一般化には注意が必要である。しかしながら, テストが診断向けに作られていないことは，CDM を当該テストに適用する意義がないことを意味しない。一次元性が高いテストに対して CDM を適用した場合にはたしかにアトリビュート間に相関が生じ, すべてのアトリビュートを習得している習得パタンと一つもアトリビュートを習得していないパタンの 2 つのパタンに分類される解答者が多くなる。その一方で, 中間的なアトリビュートの習得パタン，つまり特定のアトリビュートは習得しているが別のアトリビュートは習得していないといった通常期待されるアトリビュート習得パタンを示す解答者も存在する。こういった解答者が一定数存在するのであれば，CDM によって習得パタンを推定する意味があるだろう。

また，解答者のみならずテストそれ自体を精緻化する目的として，CDM を適用することもあるだろう。例えば Q 行列を作成することによって，テストで必要とされるアトリビュ

ートにどのような種類があるのか，何個のテスト項目でそれらのアトリビュートが測定されているのかといったことを明示できる。そのため，新たに項目を開発する際に測定しているアトリビュートを定義することで，身につけるべき学習内容や認知的な要素のバランスをとることが可能となる。このような指標を得るために，CDM を利用することができるだろう。

また TIMMS などの大規模テストは広い範囲の学習内容を問うために，計画欠損を利用している。このため，解答者ごとに異なったアトリビュートセットのテストを受験している可能性もある。こういった場合には，もちろんアトリビュートの習得パタンについての比較は行うことができない。今回の結果は Lee et al.（2011）の結果に依拠しており，特定のブックレットで測定しているアトリビュートについて，それらのブックレットに解答した解答者の結果であるという点については留意する必要があろう。

今後は，PISA データやその他の公開データでの検討を行い，テスト内容と最適な CDM の組み合わせについての知見の蓄積が必要不可欠であるといえる。例えば，Chen & de la Torre（2014）のように PISA 調査の英語テストを用いることが考えられる。さらに，同じ TIMSS の算数や数学でも内容の領域による違いや，経年変化によるアトリビュートの習得状況の変化や最適なモデルの変化といったことの検討も知見の蓄積としては必要な観点であろう。

さらに，TIMSS は算数や数学の領域を対象にした調査であり，本研究の知見の英語やその他の教科への一般化には留意する必要がある。

6-2-3　　アトリビュートの妥当性

アトリビュートの妥当性については，後述する Q 行列の妥当性についての議論とも類似する点がある。アトリビュートは統計的には離散的潜在変数であり，アトリビュートの関数によって潜在クラスが生成される。一般にアトリビュートはその解釈の容易さから，習得・未習得の 2 値変数として表されることが多い。このような解釈を行うためには，因子分析モデルにおける因子のようにアトリビュートがどのような項目で測定されているのか注意深く検討する必要がある。つまり，アトリビュートの解釈は理論的な背景に加えて測定項目に依存すると考えられる。因子分析との相違点を上げるのであれば，因子分析においては因子と測定項目の関係は因子負荷量として推定する一方，CDM においては Q 行列を分析者の先見知識としてモデルに組み込む点にある。そのため，データからアトリビュートと項目の関係を推定することは通常は行わない。このような観点からいえば，アトリビュートの設定は

理論的背景に依存して決定され，さらに想定したアトリビュートがその項目で測定されているということを保証する必要もある。

言い換えると，アトリビュートには解釈を行えるような実質的な意味をあるものを選ぶ必要であるということである。CDM では解答スキルや認知方略としてもアトリビュートを解釈するが，アトリビュートが離散的な変数であるという点を考慮するとアトリビュートとして扱うことができるのはある知識的な要素や，数学における計算の演算の能力のように，知っているか・知っていないか，というはっきりと 2 値的に分離できる変数に限定される。応用研究では，明らかに連続値として扱うことが望ましい潜在特性であってもアトリビュートとして 2 値的なものとして扱うこともあるが，これは必ずしも好ましくない可能性がある。例えば，「読解能力」など高低で表すことが望ましいと直観的に思われる能力をアトリビュートとした場合には，「読解能力を習得した状態」が何を意味するのか明確ではなく CDM の結果の解釈は難しくなる。

本研究の経験的検討で用いた Q 行列のアトリビュートは Lee et al.（2011）で言及されているように，アトリビュートの内容的には非常に入念な調査と専門家の協議を行って決定されているため内容的な意味で一定の妥当性があると判断できる。しかし，認知スキルとしては連続的に変化する潜在特性とみなることが好ましいものも存在している。例えば，「3. 実生活の文脈での問題解決（例：測定，お金の問題）を含む問題解決」（数 3 アトリビュート）などはその例である。このアトリビュートも実生活での特定の問題に対して定義されていれば，習得・未習得の 2 値のアトリビュートと見なすこともできると考えられる。しかし，このアトリビュートは実生活のどのような問題解決を問題にしているのかが明確ではなく，習得しているという状態がどのような状態なのか明確ではない。このような観点も考慮すると，アトリビュートとみなすことができる認知要素や解答スキルは一部に限定的であり，アトリビュートの定義については詳細な検討が必要と考えられる。

今回使用された項目とアトリビュートの関係を見ると，比較的冗長な関係が見られる。幾つか具体例を挙げて考察する。まず，項目 3（M041069）には，「数 2」，「数 4」および「数 5」のアトリビュートが必要とされており，これらのアトリビュートは基本的な計算と比率を含む問題解決および，単純な分数の問題解決となっている。しかし，比率を含む問題解決と単純な分数の問題解決の間にどのような違いがあるのかは明確でなく，比率を含む問題解決の中にはすでに基本的な計算が含まれている可能性があり，アトリビュートの設定には大きな重複があるといわざるをえない。この項目 3 についてはアトリビュートの定義的

に「数 4」のアトリビュートのみで十分に問題正答に必要な能力を被覆していると考えることもできよう。

項目 5（M041281）は「問題設定を数式に変換する」ことが，「所与の整数の組からルールを同定する」というアトリビュートに対応するが，これにはやや無理がある可能性がある。むしろ，現実的な状況を数学的な表現に直すという観点は「数 3」のアトリビュートが意図している実生活の文脈での問題解決といって差し支えないと考えられる。

項目 6（M041164）に必要な図形 12 のアトリビュートは「12. 図形の描画と認識およびその動きを認識するためのインフォーマルな座標での場所の点の認識」であり，定義自体が非常に曖昧なものであり，認知スキルとしてこれが何を反映しているのかは定かではない。項目 7（M041146）は長方形の性質のみがわかっていれば問題に正答できる可能性も多分にあり，他の直線の認識やインフォーマルな座標の認識などは，幾何図形の性質の理解にも含まれている可能性が考えられる。

項目 8（M041152）の問題では，問題の趣旨としてはフェンスの面積を計算する問題であるが，常識的にフェンスは長方形とみなされている可能性もある。そのため，与えられたアトリビュート（「数 1，2，3」および「図形と測定 10，11」）は過剰であると考えられる。「図形と測定 11」は面積の計算ができるかどうかのアトリビュートであり，この項目 8 ではこのアトリビュートが実生活場面で使えるかどうかを問うているに過ぎず，整数の表現などの知識が必要とは考えにくい。

項目 9（M041258A），項目 10（M041258B）は問題の正答には記述が必要であり図形の性質を単に知っているのみならず，説明できるほど理解していなければならない可能性がある。そのため，ここではアトリビュートを習得している状況に，このような説明もできることを暗に含んでいるものの，説明能力は図形の性質において特異に必要な能力とも考えにくい。また，項目 11（M041131）は図形の理解というよりも，倍数の知識があれば解答できる問題であり，これも不要なアトリビュートが付与されている例と考えられる。

項目 13（M041186）は数値の計算を行うというより，りんごのシンボルが実際のりんご何個分を表しているのかということを，提示されている表の中から発見できるか問うている可能性がある。この意味では，数値の計算ではない能力によって問題の正答確率が変化している可能性がある。

項目 14（M041336）でも他の問題と同様に，アトリビュートの定義の重複が見られる。この項目は 6 つのアトリビュートが必要であるものの，「資料の表現 14」のアトリビュート

は「データからの情報の使い方の理解と計算」であり，実際にはこのアトリビュートのみで問題に正答できる可能性がある。ただし，このようなアトリビュートを想定していることから，計算ができるかどうかがアトリビュート自体に包含されており，純粋な数値の計算を習得していなければこのアトリビュートを習得できないなどの階層関係を想定していると考えることも可能であるものの，このように考える根拠は必ずしも明確ではない。

このように，項目を細かく調査すると，Lee et al.（2011)が想定したアトリビュートがそもそもどれだけ妥当であったのか，一考の余地がある可能性がある。ただし，ここでの指摘は素朴なものであり，教科教育の専門的な文脈でこれら問題解決方略がどのように議論されているのかを慎重に調査する必要がある。しかし，項目パラメタの分析結果を見ると，アトリビュートが付与されている状況は比較的冗長であったとも考えられる。このようなアトリビュートの定義に関しては，テスト解答者へのインタビューや発話思考などのデータを含めた妥当性の証拠を収集することが必要である。

アトリビュートの定義としては，複数のアトリビュートを想定する場合に互いに重複がないように定義することは難しい可能性があるものの，どれくらいまでの定義の重複が許容されるのか，また互いに背反なアトリビュートを想定できるのかといったことについても経験的な蓄積が望まれる。

また，アトリビュートの設定と問題項目によって，最適なモデルは変化しうる。このことから，本研究の第 4 章・第 5 章の結果は Lee et al.（2011）の設定での結果ということは注意が必要である。アトリビュートや Q 行列の設定もモデル選択に影響があるという点には留意が必要であるものの，アトリビュートと Q 行列の設定を固定した場合に，どのようなモデルが選択されるのかを検討し，その知見の蓄積が必要と考えられる。今後は類似した項目であってもアトリビュートの設定が異なった場合に，選択されるモデルも変化するのかといった点についての経験的な検討が必要となる。

6-2-4　Q 行列の設定の妥当性

本研究で利用した Q 行列は Lee et al.（2011）で使用されたものである。これは先行研究との比較可能性を担保することを優先したためである。本研究の目的のためには先行研究と同様の Q 行列を使用せざるを得なかったものの，その Q 行列ではアトリビュート数が 15 であり，CDM が適用される場面一般の中では，比較的多い数であったといえよう。同じく TIMSS テストに CDM を利用した研究に目を向けると，Tatsuoka et al.（2004）などでは，23

のアトリビュートを仮定している。このように，TIMSS での CDM の設定では比較的多くのアトリビュートが設定されている。一方で，他のテストでの CDM の適用例では，Templin & Hoffman（2013）は ECPE（Examination for the Certificate of Proficiency in English）の文法セクションのデータを用い，アトリビュート数が 3，Kim（2015）は成人向け ESL プログラムの中で用いられた読解セクションのテストを分析し，アトリビュート数 10 を仮定した。このように，TIMSS テストで想定されているアトリビュートは比較的多いといえよう。アトリビュートに相関構造などを仮定しない場合には，2^{15}通りのアトリビュート習得パタンが存在すると仮定するものの，これは得られたサンプルサイズに比して非常に大きいと考えられる。つまり，解答者がほとんどいない習得パタンもかなり多く存在すると考えられる。このため，項目パラメタ，アトリビュート習得パタンの推定の安定性の観点から，アトリビュート数をより重要なものだけに限定したり，習得順序を仮定して，推定するパタンを減じるといった方策をとる必要もあると考えられる。

こうした不要なアトリビュート習得パタンの数を制御するための統計的な方法としては，正則化潜在クラス分析（regularized latent class analysis; Chen, Li, Liu, & Ying, 2017）が提案されている。このような罰則付き（正則化項つき）最尤推定法を使った推定はスパース推定法として最近盛んに研究が行われており，例えば共分散構造分析での応用は Huang, Chen, & Weng（2017）で行われている。正則化は変数選択（モデル選択）とパラメタの推定を同時に行い，不要なパラメタをゼロに縮小した推定を行うものであり，LCA の文脈でいえば不要なクラスの混合比率パラメタを 0 に縮小するものとみなすことができる。類似した方法論として，ベイズ推定法の一種とみなされる MAP 推定や EAP 推定などは事前分布を罰則とした罰則付き最尤推定の一種として考えることも可能であり心理統計・教育測定の文脈では，例えば IRT モデルの項目パラメタ推定などに主に利用されてきた。CDM で罰則付き推定を実行することで，不必要なアトリビュート習得パタンの混合比率の推定値を 0 に縮約することができれば，解釈上意味のあるアトリビュート習得パタンのみが残ることが期待できる。こういった新しい推定方法を積極的に用いる必要もあるだろう。ただし，罰則付き最尤推定を実行するためには，罰則の強さを決めるハイパーパラメタの設定などは交差検証（クロスバリデーション）などによって選択する必要がある点などは応用上の注意が必要である。

アトリビュート数が多いことによってきめ細やかな診断を行うことができる一方，誤設定の問題も生じやすくなる。本研究の項目パラメタの推定結果からは不要なアトリビュー

トがある可能性も示された。このように，Lee et al.（2011）の Q 行列では誤設定の可能性が指摘できる。また，アトリビュートや Q 行列の設定によって，最適なモデルや診断の結果が変わりうるという点は非常に重要である。設定したアトリビュートの妥当性や Q 行列の妥当性を担保するために，統計的な評価以外にも内容的な点に注意すべきであるというのは，CDM を応用する際にはいつでも注意が必要である。

このように，問題の正答に必要な少数のアトリビュートのみを用いて Q 行列を構成する方が，不要なパラメタを推定する必要もなく，CDM が TIMSS データにより適合する可能性がありうる。今後の研究としては，アトリビュートの定義を洗練させ，少数のアトリビュートのみで，Q 行列を再構成することで，CDM が TIMSS により適合する可能性について，検討が必要であろう。

さらに，第 2 章で述べたようにアトリビュート数が多すぎることによって CDM における識別性の条件を満たさなくなる危険性が指摘できる。今回用いた Q 行列は完備（complete）ではなく，CDM のどのモデルにおいても識別性を満たしていない。Q 行列が完備であるとは，各アトリビュートのみを測定している項目が Q 行列に含まれていることを意味し，そうでなければ Q 行列は完備ではないという（Xu, 2017）。Xu & Zhang（2016），Xu（2017），Xu & Shan（2017）などでは CDM における識別性の条件として Q 行列の完備性を指摘したが，Lee et al.（2011）の Q 行列はその条件を満たしていないことが確認できる。その結果として，項目パラメタとアトリビュート習得パタンの推定値に不定性が生じるとされている。これは結果を解釈する上で重要な問題であるものの，このことは本研究のモデル比較の結果には影響がないと考えられる。なぜならば，Xu & Zhang（2016）の示した不定性は項目パラメタとアトリビュート習得パタンの推定に関連したものであり，尤度それ自体には影響を及ぼさないからである。本研究の経験的検討で用いた相対的なモデル比較指標は尤度にもとづくものであり直接不定性の問題は受けないと考えることができる。ただし，本研究で示した項目パラメタやアトリビュート習得パタンについては不定性の影響を受けている可能性がある。そのため，本研究の結果がどれほど識別性の影響を受けているのか，今後は詳細な検討が必要である。より具体的には，識別性がない Q 行列を用いた場合にどれほど結果が変化してしまうのか，複数の最大尤度を示した結果を比較することなどが考えられる。識別性が現実的な問題となる程度について明らかにすることは CDM を応用する上で非常に重要であると考えられる。

また，識別性に関連して CDM の局所解についても，慎重に考える必要がある。本研究

では，複数の初期値から何度か推定を行い最大になった対数尤度を示した場合の結果を報告し，局所解を避ける工夫を行った。しかしながら，CDM は観測変数・潜在変数のどちらも離散的であり，局所解が多い可能性がある。そのため，本研究の結果も大域解であるという絶対的な保証は難しい。今後はこうした局所解の問題も合わせて実データでの応用に際して問題が生じるのか，明らかにする必要がある。CDM の実データ解析では局所解を避けて複数の最大対数尤度を比較する必要があるが，しかし，こうした解析はアトリビュートが多かったり，サンプルサイズが大きいテストの場合には推定に非常に時間がかかる場合もあり，必ずしも容易に実行できない可能性もある。

こうした問題点があるため，今回得られた項目パラメタの推定値がどれだけ意味のあるものであったのかについては慎重な解釈が必要であろう。項目パラメタの不定性の問題のみならず Q 行列の誤設定の問題もある可能性は先に指摘したが，さらに，それに加えて項目パラメタの推定値の境界問題（boundary problem; e.g., DeCarlo, 2011; Garre & Vermunt, 2006）と呼ばれる問題が存在するからである。境界問題はもともと LCA で発見された問題であり，真値がパラメタの境界値である 0 や 1 ではないにも関わらず，パラメタの推定値が 0 や 1 といった境界の値として推定される現象である。CDM においても，サンプルサイズが小さい場合にはこのような現象は経験的に知られている（e.g., Philipp, Carolin, de la Torre, & Achim, 2017）。また，その対応方法としては MAP 推定などベイズ推定が有効であるともいわれている（e.g., DeCarlo, 2011; Garre & Vermunt, 2006）。LCA での境界問題は特定のクラスに所属する人がいない場合に生じるとされているが，CDM ではどのような条件によってこれらの境界問題が生じるのかは未だ詳細な検討は少ない。CDM のアプリケーション論文である Lee et al.（2011）の Table 1 の項目パラメタの推定値を見ても，DINA モデルの項目パラメタ s, gの推定値が 0 となっている項目も散見される。また，本研究のように項目パラメタの推定値が.999 になっている場合もある。このような境界問題の可能性も指摘できるため，Q 行列の誤設定のみならず項目パラメタの推定上の問題についても，検討する必要があるだろう。また，出版されている論文であってもこのような問題がある可能性は捨てきれないため，注意する必要があろう。こうした観点からも，Q 行列の設計には注意する必要がある。

6-2-5　ベイズ統計学の観点からのモデル比較の必要性

本研究の経験的検討で用いたモデル比較に用いた指標は比較的古典的な情報量規準であることが，本研究の限界の 1 つとして挙げられる。AIC，BIC は正則なモデルに対して利用

されるものである。また，AIC はモデルによる予測の観点を重視し経験的には比較的複雑なモデルを選択するといわれる。BIC はベイズファクターの近似であり，AIC とは異なった観点からのモデル比較基準である。BIC は罰則項にサンプルサイズが含まれているため，BIC の方が AIC よりも罰則が強い情報量規準である。

近年のベイズ統計学の成果としては, AIC や BIC に変わる新たな情報量規準として WAIC（Watanabe, 2010）やWBIC（Watanabe, 2013）といった指標も提案されている。WAICは leave-one-outクロスバリデーションと漸近的に同じ平均と分散をもつ指標であり（Vehtari & Gelman, 2014），事後分布の正規性が成り立たない特異モデルなどでも利用できる指標とされる（Watanabe, 2010）。つまり，WAICはモデルの汎化性能からモデル選択を行う指標であり，CDMの比較においても有効であると考えられる。なぜならば，CDMのような識別性に問題がある場合があるモデルであってもWAICを利用できるためである。WBICはベイズ自由エネルギーの推定量であり，真の構造を推定するのに適した指標であるとされる（Watanabe, 2013）。いずれの指標もMCMCサンプルから事後対数尤度を用いて計算することが可能であり，今後のモデル選択基準として期待が持てる指標である。

これらの指標を CDM に適用することを考えると，WBIC を用いることにより，解答者の問題解答行動が CDM のモデルで表されるものとどれほど乖離があるのかを検討することができる。このような検討が可能になれば新しいモデル開発の示唆を得ることができるだろう。

また，情報量規準に限らず，ベイズ推定法を用いることにより事後予測分布を自然に得ることができる。これらの事後予測分布を用いて，実データがモデルによってよく記述されているのかを調べる事後予測チェック（Sinharay & Almond, 2007）などもモデルの性能を評価するために必要であろう。

ベイズ統計学の知見を用いたモデル評価は上記の WAIC や WBIC などの情報量規準や予測の観点からの評価を自然に導入できるため，非常に有用である。その一方で，最尤推定法による方法においても，クロスバリデーション法によるモデルの汎化誤差の推定が実行できる。こういった汎化誤差を考慮したモデル比較の方法も合わせて検討することでより多面的なモデル評価を行い，モデル選択を実行することも必要である。

6-2-6　モデル開発の必要性

第 2 章でも述べたように，近年の CDM モデル開発は非常に活発であり新しいテスト分

析モデルとして急速に認知されてきている。本研究ではその現状を受けて，現在提案されている代表的な認知診断モデルの経験的な比較を行いどのようなモデルが有用であるのかを検討したものであった。このことはモデル開発の流れを改めて考える一つのきっかけになると考えられる。先に述べたようにCDMのモデル開発としてはより一般的なモデルや様々なテストフォーマットに対応できるように一般化・複雑化されている。その一方で，本研究の結果からは，現実的にはそこまで複雑なモデルや単純すぎるモデルよりもアトリビュートが加法的に影響するという単純なモデルが有用である可能性が示唆された。つまり，複雑すぎるモデルは必ずしも必要ではない可能性も示唆される。IRTモデルも最初期に開発された1〜3パラメタモデルや段階反応モデルが未だに利用されているように，CDMにおいても実用上はA-CDMなどの比較的シンプルなモデルが普及する可能性もある。このような意味で，応用しやすいモデルについての知見の蓄積は今後も必要である。

これらを踏まえて改めてモデル開発の方向性を考えると，複雑なモデルではなくどのようなテストでも汎用的に利用しやすく，安定した推定が可能で，かつ結果が直感的に解釈しやすいモデルが応用上は望まれるだろう。もちろん学問的には精緻なモデルを開発しその有用性を検討することは重要である一方，診断に使いやすいモデルに目を向けた開発も必要だろう。特に，CDMが広く利用されるためには少ない仮定で使いやすいモデルが必要と考えられる。

ただし，先に述べたように本研究の結果はあくまで一連のモデル群の中での相対的な適合の比較しか行っていない。R-RUM，A-CDM，LLMの適合は相対的にみて他のCDMよりもよかったものの，これら主効果モデルよりもよい適合を示すモデルを開発することは可能だろう。研究においては，解答者の問題解決プロセスをより精緻に反映したモデルを開発していくことは今後も必要である。

また，CDMベースでテストを作成・実施するための方法論の整備も必要であろう。“問題解答に際して複数のアトリビュートが必要”という表現は必ずしも厳密ではなく，解答者が問題解決を行う際に，同時にアトリビュートが必要になるという状況がどのようなものなのか，改めて検討する必要があると考えられる。あくまでそれが概念上の操作なのか，実際に行われている行動にもとづいたことなのか，詳細な検討が必要である。

このほか，IRT でのテスト運用の類推でいえば項目プールの作成や保持の仕方についても今後開発が望まれる。例えば e-ラーニングでは問題の解答パタンから次に学習するべき学習内容を提示することが必要となるが，CDMを使うことによりこうしたことを効果的に

実施できる可能性がある。このためには，どういった項目をどのようにデータベースとして保存するのが適切なのかといった研究なども必要である。

統計的な興味としては，アトリビュート数や項目数，Q 行列，サンプルサイズなど統計的な観点からの評価方法に関しては未だ課題が山積しているといえよう。例えば，CDM が利用されうるクラスルームや学級単位といった比較的小さいサンプルサイズの状況であっても正確に診断を行う方法論の開発などは今後の検討課題といえよう。

さらに，CDM の一般形式の定義も今後行なっていくことが望ましいだろう。また，第 2 章で検討したように，CDM には，2 値データ以外の多値・複数選択枝に適したモデルや拡張的なモデルも存在する。これまでに開発されてきたモデルにはそれぞれ適切な使用状況があるものの, モデルの表現として共通するのは, アトリビュートを仮定し, Q 行列を用い, アトリビュートの習得パタンを用いた分類を試みている点であろう。一般性の高いモデルは項目パラメタの数を増やし，アトリビュート習得パタンの相違を表現できるように設計されている。本研究で主に用いた G-DINA モデルや他の一般性の高い LCDM，GDM などモデル開発がなされているものの，それらは Q 行列それ自体をモデルの一部として明示的に取り扱っていない。しかし, Q 行列はあくまで分析者の設定である以上はモデルの一部と見なす必要がある。例えば，先に紹介した多肢選択式問題への解答モデルや，複数の解答方略を許容するモデルでは Q 行列自体の設定を変更する提案がなされている。これまでは，Q 行列も項目反応確率を規定する要素であるものの，Q 行列は所与の要素として扱われてきた。この Q 行列もモデルのパラメタとして同時に推定するモデルを構成することは CDM の利便性を高めるためには必要であろう。例えば，DeCarlo（2012）は DINA モデルを用いて Q 行列の一部を確率変数としてベイズ推定を行う方法を示したが，こうした考え方を利用しモデルを構成出来る可能性がある。こうした Q 行列をパラメタとみなしたモデルの開発も必要であると考えられる。

このように考えると，Q 行列の誤設定は広くはモデルの誤設定問題の特殊な場合と見なすことができよう。本研究を含めたモデル比較研究や Q 行列を推定する問題では，Q 行列を固定して適切なモデルを選択する問題と項目反応関数を固定して Q 行列を推定する問題の 2 つに分かれて研究がなされてきた。ところが，真の項目反応関数も真の Q 行列も本来はわからず，それらを片方だけ真のものと固定することができる合理的な根拠は本来少なく，際にはどちらの要素も不確定なものであり，“項目反応確率を規定する要素”であり，モデルと解釈する必要があると考えられる。Q 行列は項目の内容分析や先行研究の知見，専

門家の見解など定量的ではないデータによってその妥当性を担保しようとしてきたものの，便宜的な手続きにすぎない。本研究の結果から，TIMSS 2007 の 4 年生の算数データでは Lee et al.（2011）が採用した Q 行列の設定が適切ではない可能性も示唆されてきている。これらの観点からも Q 行列と項目反応関数をそれぞれ独立の要素として考えるのではなく，それらを統合したモデルの開発が望まれるだろう。

さらに，今回用いた CDM の一群はアトリビュートの習得・未習得のパタンを推定するいわば静的なモデルであり，学習が進むにつれてアトリビュート自体が変化することなどは想定していない。今後のモデル開発としては，こうした動的な点にも注目し，解答者の認知プロセスをより直接的に反映させた CDM を開発することが望まれる。これにより，詳細に解答者の認知的な状態を把握することに繋がり，解答者のつまづきを解消することができるだろう。

6-2-7　各国の教育背景との診断結果の関係性

本研究の結果はあくまでも教育測定モデルの実データを用いた比較によるものであり，CDM を適用した結果が各国で異なっていた点について，各国特有の背景を踏まえた詳細な検討を行うことはできていない。しかし，各国での教育制度や教科教育で何が教授されているのかという観点は国際比較の結果を解釈するためには有用であると考えられる。教育制度による項目反応関数の違い，あるいはアトリビュート習得パタンの違いがどのように生じているのかということに関しては今後検討する必要がある話題であろう。

また，教育制度に関連して，学校の教員が実際にどのような指導を行っているか，家庭の教育環境がどのようであるのかといった情報もまた，アトリビュート習得パタンを用いた診断情報を深く解釈するためには必要であろう。TIMSS ではこういった変数は質問紙によって収集されておりデータとしても利用可能なものである。外部情報をより積極的に用いた分析によって，推定結果を理解することにより，新しいモデル開発にもつながる可能性があるだろう。

もちろん，項目反応パタン以外の情報として最も素朴なものとしては，解答者自身の学習習慣やその学習への意欲などが挙げられる。教師や家庭の変数に加えて，解答者自身の変数についても利用できるものは多く存在する。これらの外的な変数と CDM の分析結果の関係や外部情報を一度に分析する方法論の開発は非常に重要である一方，研究は十分ではない。現状では，CDM は項目反応パタンのみを利用した分析にとどまっている。今後の研究とし

てはこれらの外部情報を積極的に取り込んだ分析を行うことによって，解答者の学習パタンを分析することが，CDM の応用として非常に重要な研究と考えられる。

また，本研究で選択した国は，TIMSS の学力の高低による国の異質性の側面のうちの 1 つである点には留意が必要である。その他の基準としては，ここで述べたような教育制度や教育方法の違いなども考えられる。こうした別の異質性の基準を用いたモデルの比較も今後の研究として必要であると考えられる。

関連論文

第 2 章：山口一大・岡田謙介（2017a）. 近年の認知診断モデルの展開 行動計量学, *44*, 181-198.

第 3 章 : 山口一大（2017）. 認知診断モデルにおける Q 行列の誤設定が診断精度に与える影響–認知の階層構造を考慮した場合の検討– 日本テスト学会誌, *13*, 17-32.

第 4 章：山口一大・岡田謙介（2017b）. TIMSS2007 の日本人サンプルを用いた認知診断モデルと項目反応理論モデルの比較 日本テスト学会誌, *13*, 1-16

第 5 章：Yamaguchi, K., & Okada, K. (in press). Comparison among cognitive diagnostic models for the TIMSS 2007 fourth grade mathematics assessment *PLOS ONE*

引用文献

Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessmen*t. London, UK: Continuum.

Almond, R. G., DiBello, L. V., Moulder, B., & Zapata‐Rivera, J. D. (2007). Modeling diagnostic assessments with Bayesian networks. *Journal of Educational Measurement, 44*, 341-359. doi: 10.1111/j.1745-3984.2007.00043.x.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19,* 716-723. doi: 10.1109/TAC.1974.1100705.

Ayers, E., Rabe-Hesketh, S., & Nugent, R. (2013). Incorporating student covariates in cognitive diagnosis models. *Journal of Classification, 30,* 195-224. doi: https://doi.org/10.1007/s00357-013-9130-y.

Baker, F. B. (1993). Sensitivity of the linear logistic test model to misspecification of the weight matrix. *Applied Psychological Measurement, 17,* 201-210. doi: https://doi.org/10.1177/014662169301700301.

Barnes, T. (2005, July). The Q-matrix method: Mining student response data for knowledge. In *American Association for Artificial Intelligence 2005 Educational Data Mining Workshop* (pp. 1-8). Available from: https://www.aaai.org/Papers/Workshops/2005/WS-05-02/WS05-02-006.pdf.

Beheshti, B., Desmarais, M. C., & Naceur, R. (2012). Methods to find the number of latent skills. *International Educational Data Mining Society*, 81-86. Available from: http://www.professeurs.polymtl.ca/michel.desmarais/Papers/EDM2012/nskills-edm2012.pdf.

Belov, D. I. (2008). Uniform test assembly. *Psychometrika, 73*, 21-38. doi: https://doi.org/10.1007/s11336-007-9025-0.

Bergman, L. R. & Lundh, L. G. (2015). Introduction: The person-oriented approach: Roots and roads to the future. *Journal for Person-Oriented Research, 1*, 1-6. doi: 10.17505/jpor.2015.01.

Bergman, L. R. & Magnusson, D. (1997). A person-oriented ap- proach in research on developmental psychopathology. *Development and Psychopathology, 9,* 291-319. doi: 10.1017/S095457949700206X.

Bolt, D. M., & Lall, V. F. (2003). Estimation of compensatory and noncompensatory multidimensional item response models using Markov chain Monte Carlo. *Applied Psychological Measurement,*

27, 395-414. doi: https://doi.org/10.1177/0146621603258350.

Cai, J., Mok, I. A., Reddy, V., & Stacey, K. (2016). International comparative studies in mathematics: Lessons for improving students' learning. In Kaiser, G. (Series ed.). *Part of the series ICME-13 Hamburg, Topical Surveys* (1–36), 24–31 july 2016. Springer Open. doi: 10.1007/978-3-319-42414-9_1.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., ... & Riddell, A. (2016). Stan: A probabilistic programming language. *Journal of Statistical Software, 20*, 1-37. doi: 10.18637/jss.v076.i01.

Chen, J., & de la Torre, J. (2013). A general cognitive diagnosis model for expert-defined polytomous attributes. *Applied Psychological Measurement, 37,* 419-437. doi: https://doi.org/10.1177/0146621613479818.

Chen, J., & de la Torre, J. (2014). A procedure for diagnostically modeling extant large-scale assessment data: The case of the programme for international student assessment in reading. *Psychology, 5*, 1967-1978. doi: 10.4236/psych.2014.518200.

Chen, Y., Li, X., Liu, J., & Ying, Z. (2017). Regularized latent class analysis with application in cognitive diagnosis. *Psychometrika, 82*, 660-692. doi: 10.1007/s11336-016-9545-6.

Cheng, Y. (2009). When cognitive diagnosis meets computerized adaptive testing: CD-CAT. *Psychometrika, 74*, 619-632. doi: https://doi.org/10.1007/s11336-009-9123-2.

Chiu, C. Y. (2013). Statistical refinement of the Q-matrix in cognitive diagnosis. *Applied Psychological Measurement, 37*, 598-618. doi: 10.1177/0146621613488436.

Chiu, C. Y., & Douglas, J. (2013). A nonparametric approach to cognitive diagnosis by proximity to ideal response patterns. *Journal of Classification, 30*, 225-250. doi: https://doi.org/10.1007/s00357-013-9132-9.

Chiu, C. Y., Douglas, J. A., & Li, X. (2009). Cluster analysis for cognitive diagnosis: Theory and applications. *Psychometrika, 74*, 633-665. doi: https://doi.org/10.1007/s11336-009-9125-0.

Chiu, C. Y., & Köhn, H. F. (2015). A general proof of consistency of heuristic classification for cognitive diagnosis models. *British Journal of Mathematical and Statistical Psychology, 68*, 387-409. doi: 10.1111/bmsp.12055.

Chiu, C. Y., & Köhn, H. F. (2016). The reduced RUM as a logit model: Parameterization and constraints. *Psychometrika, 81*, 350–370. doi: 10.1007/s11336-015-9460-2.

Chung, M. T. (2014). *Estimating the Q-matrix for cognitive diagnosis models in a Bayesian framework.* (Doctoral dissertation, Columbia University). doi: https://doi.org/10.7916/D857195B.

Choi, K. M., Lee, Y. S., & Park, Y. S. (2015). What CDM can tell about what students have learned: An analysis of TIMSS eighth grade mathematics. *Eurasia Journal of Mathematics, Science & Technology Education, 11,* 1563–1577. doi: 10.12973/eurasia.2015.1421a.

Cichocki, A., Zdunek, R., Phan, A. H., & Amari, S. I. (2009). *Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation*. UK: John Wiley & Sons.

Collins, L. M. & Lanza, S. T. (2010). *Latent class and latent transition analysis: with applications in the social, behavioral, and health sciences*. Hoboken, NJ: Wiley.

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*, 8-26. doi: 10.1177/0146621610377081.

DeCarlo, L. T. (2012). Recognizing uncertainty in the Q-matrix via a Bayesian extension of the DINA model. *Applied Psychological Measurement, 36*, 447–468. doi: https://doi.org/10.1177/0146621612449069.

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*, 343–362. doi: 10.1111/j.1745-3984.2008.00069.x.

de la Torre, J. (2009a). A cognitive diagnosis model for cognitively based multiple-choice options. *Applied Psychological Measurement, 33*, 163-183. doi: https://doi.org/10.1177/0146621608320523.

de la Torre, J. (2009b). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics, 34*, 115-130. doi: https://doi.org/10.3102/1076998607309474.

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika, 76*, 179-199. doi: http://dx.doi.org/10.1007/s11336-011-9214-8.

de la Torre, J., & Chiu, C. Y. (2016). A general method of empirical Q-matrix validation. *Psychometrika, 81*, 253–273. doi: 10.1007/s11336-015-9467-8.

de la Torre, J. & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika, 69*, 333–353. doi: https://doi.org/10.1007/BF02295640.

de la Torre, J., & Douglas, J. A. (2008). Model evaluation and multiple strategies in cognitive diagnosis: An analysis of fraction subtraction data. *Psychometrika, 73*, 595-624. doi: https://doi.org/10.1007/s11336-008-9063-2.

de la Torre, J., & Karelitz, T. M. (2009). Impact of diagnosticity on the adequacy of models for cognitive diagnosis under a linear attribute structure: A simulation study. *Journal of Educational Measurement, 46,* 450-469. doi: 10.1111/j.1745-3984.2009.00092.x.

Desmarais, M.C. (2011). Conditions for effectively deriving a q-matrix from data with non-negative matrix factorization. In C. Conati, S. Ventura, T. Calders, & M. Pechenizkiy (Eds.). *4th International Conference on Educational Data Mining* (pp. 41-50). Eindhoven, Netherlands. Available from: http://educationaldatamining.org/EDM2011/wp-content/uploads/proc/edm2011_paper35_full_Desmarais.pdf.

Desmarais, M., Beheshti, B., & Naceur, R. (2012). Item to skills mapping: deriving a conjunctive Q-matrix from data. In *Intelligent tutoring systems* (pp. 454-463). Springer Berlin/Heidelberg. Available from: http://www.professeurs.polymtl.ca/michel.desmarais/Papers/ITS2012/its2012.pdf.

Desmarais, M., Beheshti, B., & Xu, P. (2014, July). The refinement of a Q-matrix: Assessing methods to validate tasks to skills mapping. In Educational Data Mining 2014. Available from: https://pdfs.semanticscholar.org/a113/f0ebd7c2708302459e82e55051fff952a62a.pdf.

DiBello, L. V., Henson, R. A., & Stout, W. F. (2015). A family of generalized diagnostic classification models for multiple choice option-based scoring. *Applied Psychological Measurement, 39*, 62-79. doi: https://doi.org/10.1177/0146621614561315.

DiBello, L. V., Roussos, L. A., & Stout, W. (2006). Review of cognitively diagnostic assessment and a summary of psychometric models. In C. Rao & S. Sinharay (Eds.), *Handbook of Statistics, Vol. 26: Psychometrics.* North Holland: Elsevier, pp. 979–1030. doi: https://doi.org/10.1016/S0169-7161(06)26031-0.

DiBello, L. V., & Stout, W., & Roussos, L. A. (1995). Unified cognitive/psychometric diagnostic assessment likelihood-based classification techniques. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.) (1995). *Cognitively diagnostic assessment.* New York: Routledge.

Dimitrov, D. M. (2007). Least squares distance method of cognitive validation and analysis for binary items using their item response theory parameters. *Applied Psychological Measurement, 31*, 367-

387. doi: https://doi.org/10.1177/0146621606295199.

Dimitrov, D. M., & Atanasov, D. V. (2012). Conjunctive and disjunctive extensions of the least squares distance model of cognitive diagnosis. *Educational and Psychological Measurement, 72*, 120-138. doi: https://doi.org/10.1177/0013164411402324.

Dimitrov, D. M., & Raykov, T. (2003). Validation of cognitive structures: A structural equation modeling approach. *Multivariate Behavioral Research, 38*, 1-23. doi: http://dx.doi.org/10.1207/S15327906MBR3801_1.

Embretson, S. E. (1984). A general latent trait model for response processes. *Psychometrika, 49*, 175–186. doi: https://doi.org/10.1007/BF02294171.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods, 3*, 380. doi: 10.1037/1082-989X.3.3.380.

Embretson, S. E. (2015). The multicomponent latent trait model for diagnosis applications to heterogeneous test domains. *Applied Psychological Measurement, 39*, 16-30. doi: https://doi.org/10.1177/0146621614552014.

Embretson, S. E. & Yang, X. (2013). A multicomponent latent trait model for diagnosis. *Psychometrika, 78*, 14–36. doi: 10.1007/s11336-012-9296-y.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta psychologica, 37*, 359-374. Available from: https://www.researchgate.net/profile/Gerhard_Fischer3/publication/281345292_The_Linear_logistic_test_model_as_an_instrument_in_educational_research/links/54a2da990cf257a63604da16/The-Linear-logistic-test-model-as-an-instrument-in-educational-research.pdf.

García, P. E., Olea, J., & de la Torre, J. (2014). Application of cognitive diagnosis models to competency-based situational judgment tests. *Psicothema, 3*, 372-377. doi:10.7334/psicothema2013.322.

Garre, F. G., & Vermunt, J. K. (2006). Avoiding boundary estimates in latent class analysis by Bayesian posterior mode estimation, *Behaviormetrika, 33*, 43–59. https://doi.org/10.2333/bhmk.33.43.

Geman, S. & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 6*, 721-741, doi: 10.1109/TPAMI.1984.4767596.

George, A. C., & Robitzsch, A. (2014). Multiple group cognitive diagnosis models, with an emphasis on differential item functioning. *Psychological Test and Assessment Modeling, 56*, 405-432. Available from: http://www.psychologie-aktuell.com/fileadmin/download/ptam/4-2014_20141222/06_George.pdf.

George, A. C., Robitzsch, A., Kiefer, T., Gross, J., & Uenlue, A. (2016). The R Package CDM for cognitive diagnosis models. *Journal of Statistical Software, 74*, 1-24. doi:10.18637/jss.v074.i02.

Gierl, M. J. (2007). Making diagnostic inferences about cognitive attributes using the rule-space model and attribute hierarchy method. *Journal of Educational Measurement, 44*, 325-340. doi: 10.1111/j.1745-3984.2007.00042.x.

Gierl, M. J., Cui, Y., & Zhou, J. (2009). Reliability and attribute based scoring in cognitive diagnostic assessment. *Journal of Educational Measurement, 46*, 293 − 313. doi: 10.1111/j.1745-3984.2009.00082.x.

Gierl, M. J., & Haladyna, T. M. (Eds.). (2012). *Automatic item generation: Theory and practice*. New York, NY: Routledge.

Gierl, M. J., Wang, C., & Zhou, J. (2008). Using the attribute hierarchy method to make diagnostic inferences about examinees' cognitive skills in algebra on the SAT. *Journal of Technology, Learning, and Assessment, 6*, 1-52. Available from: http://files.eric.ed.gov/fulltext/EJ838616.pdf.

Gierl, M. J., Zheng, Y., & Cui, Y. (2008). Using the attribute hierarchy method to identify and interpret cognitive skills that produce group differences. *Journal of Educational Measurement, 45*, 65-89. doi: 10.1111/j.1745-3984.2007.00052.x.

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

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement, 26*, 301-321. doi: 10.1111/j.1745-3984.1989.tb00336.x.

Hartz, S., & Roussos, L. (2008). The fusion model for skills diagnosis: Blending theory with practicality. *ETS Research Report Series,* 1-57. doi: http://dx.doi.org/10.1002/j.2333-8504.2008.tb02157.x.

服部環（2011). 心理・教育のためのRによるデータ解析 福村出版

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

https://doi.org/10.1007/s11336-008-9089-5.

Hoffman, M. D., & Gelman, A. (2014). The No-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research, 15*, 1593-1623. doi: http://dl.acm.org/citation.cfm?id=2627435.2638586.

星野崇宏・岡田謙介・前田忠彦（2005）. 構造方程式モデリングにおける適合度指標とモデル改善について：展望とシミュレーション研究による新たな知見. 行動計量学,.*32,* 209-235. doi: 10.2333/jbhmk.32.209.

Huang, P., Chen, H., & Weng, L. (2017). A penalized likelihood method for structural equation modeling. *Psychometrika, 82*(2), 329–354. https://doi.org/10.1007/s11336-017-9566-9

Huang, H. Y., & Wang, W. C. (2014). The random effect DINA model. *Journal of Educational Measurement, 51*, 75-97. doi: 10.1111/jedm.12035.

Huo, Y., & de la Torre, J. (2014). Estimating a cognitive diagnostic model for multiple strategies via the EM algorithm. *Applied Psychological Measurement, 38*. 464-485. doi: https://doi.org/10.1177/0146621614533986.

Im, S. & Corter, J. E. (2011). Statistical consequences of attribute misspecification in the rule space method. *Educational and Psychological Measurement, 71*, 712–731. doi: 10.1177/0013164410384855.

Jang, E. E. (2009). Demystifying a Q-matrix for making diagnostic inferences about L2 reading skills. *Language Assessment Quarterly, 6*, 210-238. doi: http://dx.doi.org/10.1080/15434300903071817.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25*, 258-272. doi: https://doi.org/10.1177/01466210122032064.

Kelley, K. (2017). MBESS: The MBESS R Package [Computer software]. R package version 4.3.0. Retrieved from: https://CRAN.R-project.org/package=MBESS

Kim, A. Y. (2015). Exploring ways to provide diagnostic feedback with an ESL placement test: Cognitive diagnostic assessment of L2 reading ability. *Language Testing, 32*, 227-258. doi: 10.1177/0265532214558457.

国立教育政策研究所（2009）. TIMSS2007 算数・数学教育の国際比較 −国際数学・理科教育動向調査の 2007 年調査報告書− （改訂版）. Available from:

http://www.nier.go.jp/timss/2007/report_math.pdf.

Kunina‐Habenicht, O., Rupp, A. A., & 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*, 59-81. doi: 10.1111/j.1745-3984.2011.00160.x.

倉元直樹・スコット寿美・笠居昌弘（2003）．日本語語彙理解力テストの妥当性についての検討．心理学研究, *51*, 413-424. doi: http://doi.org/10.5926/jjep1953.51.4_413.

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis.* Boston, MA: Houghton Mifflin.

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, *401(6755)*, 788. doi: 10.1038/44565.

Lee, D. D., & Seung, H. S. (2001). Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems* (pp. 556-562). Available from: https://papers.nips.cc/paper/1861-algorithms-for-non-negative-matrix-factorization.pdf.

Lee, M. D., & Wagenmakers, E. J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge: Cambridge University Press.

Lee, Y. S., Park, Y. S., & Taylan, D. (2011). A cognitive diagnostic modeling of attribute mastery in Massachusetts, Minnesota, and the U.S. national sample using the TIMSS 2007. *International Journal of Testing, 11*, 144-177. doi: http://dx.doi.org/10.1080/15305058.2010.534571.

Lee, Y. W., & Sawaki, Y. (2009). Cognitive diagnosis approaches to language assessment: An overview. *Language Assessment Quarterly, 6*, 172-189. doi: http://dx.doi.org/10.1080/15434300902985108.

Leighton, J., & Gierl, M. (Eds.). (2007). *Cognitive diagnostic assessment for education: Theory and applications*. New York, NY: Cambridge University Press.

Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The attribute hierarchy method for cognitive assessment: A variation on Tatsuoka's rule-space approach. *Journal of Educational Measurement, 41*, 205–237. doi: 10.1111/j.1745-3984.2004.tb01163.x.

Li, F., Cohen, A., Bottge, B., & Templin, J. (2016). A latent transition analysis model for assessing change in cognitive skills. *Educational and Psychological Measurement, 76*, 181-204. doi: https://doi.org/10.1177/0013164415588946.

Li, H. (2011). A cognitive diagnostic analysis of the MELAB reading test. *Second or Foreign Language Assessment, 9,* 17–46. Available from:

https://www.researchgate.net/profile/Hongli_Li4/publication/264860933_A_cognitive_diagnostic_analysis_of_the_MELAB_reading_test/links/53f3a50b0cf256ab87b4894a.pdf.

Li, H., Hunter, C. V., & Lei, P. W. (2016). The selection of cognitive diagnostic models for a reading comprehension test. *Language Testing, 33*, 391-409. doi: https://doi.org/10.1177/0265532215590848.

Liu, H-Y., You, X-F., Wang, W-Y., Ding, S-L., & Chang, H-H. (2013). The development of computerized adaptive testing with cognitive diagnosis for an English achievement test in China. *Journal of Classification, 30*, 152-172. doi: https://doi.org/10.1007/s00357-013-9128-5.

Liu, J., Xu, G., & Ying, Z. (2012). Data-driven learning of Q-matrix. *Applied psychological measurement, 36*, 548-564. doi: 1177/0146621612456591.

Liu, J., Xu, G., & Ying, Z. (2013). Theory of the self-learning Q-matrix. *Bernoulli: official journal of the Bernoulli Society for Mathematical Statistics and Probability, 19*, 1790–1817. doi: 10.3150/12-BEJ430.

Liu, R. (2017). Misspecification of attribute structure in diagnostic measurement. *Educational and Psychological Measurement,* doi: 10.1177/0013164417702458.

Liu, R., & Huggins-Manley, A. C. (2016). The specification of attribute structures and its effects on classification accuracy in diagnostic test design. In Millsap, R.E., Bolt, D.M., van der Ark, L.A., & Wang, W.C. (Eds.) *Quantitative psychology research* (pp. 243-254). Springer International Publishing. doi: 10.1007/978-3-319-38759-8_18.

Liu, Y., Douglas, J. A., & Henson, R. A. (2009). Testing person fit in cognitive diagnosis. *Applied psychological measurement, 33*, 579-598. doi: 10.1177/0146621609331960.

Loehlin, J. C. (2004). *Latent variable models: An introduction to factor, path, and structural equation analysis* (4th ed.). Mahwah, NJ: Erlbaum.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS -A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing, 10*, 325-337. doi: https://doi.org/10.1023/A:1008929526011.

Luo, F., Ding, S., Wang, W., & Xiong, J. (2016). Application study on online multistage intelligent adaptive testing for cognitive diagnosis. In L. A. van der Ark, D. M. Bolt, W-C. Wang, J. A.

Douglas, & M. Wiberg. (Eds.), *Quantitative Psychology Research*. New York: Springer.

Ma, W. & de la Torre, J. (2017). GDINA: The generalized DINA model framework [Computer software]. R package version 1.4.2. Retrived from https://CRAN.R-project.org/package=GDINA.

Magnusson, D. (1999). On the individual: A person-oriented approach to developmental research. *European Psychologist, 4*, 205–218. doi: http://dx.doi.org/10.1027//1016-9040.4.4.205.

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika, 64*, 187-212. doi: https://doi.org/10.1007/BF02294535.

Martin, M. O., Mullis, I. V., & Foy, P. (2008). *TIMSS 2007 international mathematics report. Findings from IEA's trends in international mathematics and science study at the fourth and eighth grades*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. Available from http://timss.bc.edu/TIMSS2007/PDF/TIMSS2007_InternationalMathematicsReport.pdf.

Maydeu-Olivares, A. (2013). Goodness-of-fit assessment of item response theory models. *Measurement, 11*, 71–101. doi: 10.1080/15366367.2013.831680.

McLaclen, G., & Peel, D. (2000). *Finite mixture models*. New York: Wiley.

Meijer, R. R., & Nering, M. L. (1999). Computerized adaptive testing: Overview and introduction. A*pplied Psychological Measurement, 23*, 187–194 doi: 10.1177/01466219922031310.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741-749. doi: http://dx.doi.org/10.1037/0003-066X.50.9.741.

Mislevy, R. J. (1994). Evidence and inference in educational assessment. *Psychometrika, 59*, 439-483. doi: https://doi.org/10.1007/BF02294388.

Mislevy, R. J., & Haertel, G. D. (2006). Implications of evidence centered design for educational testing. *Educational Measurement, 25*, 6-20. doi: 10.1111/j.1745-3992.2006.00075.x.

村山航（2012）. 妥当性概念の歴史的変遷と心理測定学的観点からの考察 教育心理学年報 , 51, 118‐130. doi: 10.5926/arepj.51.118.

Muthén, L.K., & Muthén, B.O. (1998-2017). Mplus user's guide [Computer software]. (8 ed). Los Angeles, CA: Muthén & Muthén. Available from: https://www.statmodel.com/download/usersguide/MplusUserGuideVer_8.pdf.

Nichols, P. D., Chipman, S. F., & Brennan, R. L. (Eds.) (1995). *Cognitively diagnostic assessment*. New York, NY: Routledge.

二宮衆一（2014）. 教育評価の機能 西岡加名恵・石井英真・田中耕治（編） 新しい教育評価入門 人を育てる評価のために（pp. 51-75） 有斐閣

Ntzoufras, I. (2009). *Bayesian modeling using WinBUGS.* NJ :Wiley, Hoboken.

Okada, K., & Lee, M.D. (2016). A Bayesian approach to modeling group and individual differences in multidimensional scaling. *Journal of Mathematical Psychology, 70*, 35-44. doi: https://doi.org/10.1016/j.jmp.2015.12.005.

Ozaki, K. (2015). DINA models for multiple-choice items with few parameters considering incorrect answers. *Applied Psychological Measurement, 39*, 431-447. doi: https://doi.org/10.1177/0146621615574693.

Patz, R. J., & Junker, B. W. (1999). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics, 24*, 146-178. doi: https://doi.org/10.3102/10769986024002146.

Preacher, K. J., Wichman, A. L., MacCallum, R. C., & Briggs, N. E. (2008). *Latent growth curve modeling*. Quantitative applications in the social sciences. Sage Publications.

Philipp, M., Carolin, S., de la Torre, J., & Achim, Z. (2017). On the estimation of standard errors in cognitive diagnosis models. *Journal of Educational and Behavioral Statistics*, 1–28. https://doi.org/10.3102/1076998617719728.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In K. Hornik, F. Leisch, & A. Zeileis (Eds.), *Proceedings of the 3rd International Workshop on Distributed Statistical Computing, March 20-22*. Vienna, Austria: DSC. Retrieved from http://www.ci.tuwien.ac.at/ Conferences/DSC-2003/Proceedings/

Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement, 9*, 401-412. doi: https://doi.org/10.1177/014662168500900409.

Reckase, M. D. (2009). *Multidimensional item response theory*. New York, NY: Springer.

Reckase, M. D., & McKinley, R. L. (1991). The discriminating power of items that measure more than one dimension. *Applied Psychological Measurement, 15*, 361-373. doi: https://doi.org/10.1177/014662169101500407.

Revelle, W. (2017) psych: Procedures for Personality and Psychological Research [Computer

software], Retrieved from: https://CRAN.R-project.org/package=psych Version 1.7.5.

Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response theory analyses. *Journal of Statistical Software*, *17*, 1–25. Available from https://www.ime.usp.br/~mbranco/Rpackage_TRI.pdf.

Roberts, M. R., & Gierl, M. J. (2010). Developing score reports for cognitive diagnostic assessments. *Educational Measurement: Issues and Practice, 29*, 25-38. doi: 10.1111/j.1745-3992.2010.00181.x.

Robitzsch, A., Kiefer, T., & Wu, M. (2017). TAM: Test analysis modules [Computer software]. R package version 2.2-49. Available from https://CRAN.R-project.org/package=TAM.

Roussos, L. A., DiBello, L. V., Stout, W., Hartz, S. M., Henson, R. A., & Templin, J. (2007). The fusion model skills diagnosis system. In J. Leighton & M. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications.* Cambridge: Cambridge University Press, pp. 275–318.

Rupp, A. A. (2007). The answer is in the question: A guide for describing and investigating the conceptual foundations and statistical properties of cognitive psychometric models. *International Journal of Testing, 7*, 95–125. http://dx.doi.org/10.1080/15305050701193454.

Rupp, A. A. (2009). Software for calibrating diagnostic classification models: An overview of the current state-of-the-art, Available from: http://www.education.umd.edu/EDMS/fac/Rupp/AERA-SIG%20Software%20Symposium%20(Handout%20Package).pdf.

Rupp, A. A., & Templin, J. L. (2008a). The effects of Q-matrix misspecification on parameter estimates and classification accuracy in the DINA model. *Educational and Psychological Measurement, 68*, 78-96. doi: https://doi.org/10.1177/0013164407301545.

Rupp, A. A., & Templin, J. L. (2008b). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement, 6*, 219-262. doi: http://dx.doi.org/10.1080/15366360802490866.

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York, NY: Guilford Press.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*, 461-464. Available from: http://www.jstor.org/stable/2958889.

繁桝算男・植野真臣・本村陽一（2006）．ベイジアンネットワーク概説 培風館

荘島荘二郎（2003）．複数の項目反応モデルの母数の同時推定−出題形式が混在しているときの潜在特性尺度構成− 豊田秀樹（編） 共分散構造分析［技術編］ −構造方程式モデリング− （pp. 221-232） 朝倉書店

Sinharay, S., & Almond, R. G. (2007). Assessing fit of cognitive diagnostic models a case study. *Educational and Psychological Measurement, 67,* 239–257. doi: https://doi.org/10.1177/0013164406292025.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal and structural equation models*. Boca Raton, FL: Chapman & Hall/CRC.

Sorrel, M. A., Olea, J., Abad, F. J., de la Torre, J., Aguado, D., & Lievens, F. (2016). Validity and reliability of situational judgement test scores: A new approach based on cognitive diagnosis models. *Organizational Research Methods, 19*, 506–532. doi: 10.1177/1094428116630065

孫媛・井上俊哉（1995）．アメリカにおける差異項目機能（DIF）研究. 学術情報センター紀要, *7,* 193-216.

孫媛・島田めぐみ・谷部弘子（2015）．診断的日本語語彙テストの開発 李在鎬（編） 日本語教育のための言語テストガイドブック（pp. 175-194）くろしお出版

Spiegelhalter, D. J., Thomas, A., & Best, N. G. (1999). WinBUGS version 1.2 user manual [Computer software]. MRC Biostatistics Unit, 83. Available from: https://cdn.uclouvain.be/public/Exports%20reddot/stat/documents/manual14.pdf.

Spiegelhalter, D., Thomas, A., Best, N., & Lunn, D. (2010). OpenBUGS user manual. Retrieved from: http://www.openbugs.info.

Sun, Y., Suzuki, M., & Kakinuma, S. (2012). Effective feedback for self-regulated learning: Applying cognitive diagnostic assessment. *Advances in Education Research, 7*, 140-145.

Sun, Y., Ye, S., Inoue, S., & Sun, Y. (2014, July). Alternating recursive method for Q-matrix learning. In Educational Data Mining 2014. Available from: https://pdfs.semanticscholar.org/e2b8/d77a96e64890690c17202476a873a8f63e0d.pdf?_ga=2.258799882.1896703624.1505451579-813043958.1505451579.

Sun, Y., Ye, S., Shi, H., Wang, H., & Sun, Y. (2014, October). Maximum likelihood estimation based DINA model and Q-matrix learning. *In Behavior, Economic and Social Computing (BESC), 2014 International Conference on (pp. 1-6)*. IEEE. doi: 10.1109/BESC.2014.7059511.

Sun, Y., Ye, S., Sun, Y., & Kameda, T. (2015, October). Improved algorithms for exact and

approximate boolean matrix decomposition. In *Data Science and Advanced Analytics (DSAA), 2015. 36678 2015. IEEE International Conference on (pp. 1-10)*. IEEE. Available from: https://arxiv.org/pdf/1512.08041v1.pdf.

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―教研式標準学力検査 NRT「中学 1 年数学」への適用― 日本テスト学会誌, *11*, 81-97.

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20*, 345–354. doi: 10.1111/j.1745-3984.1983.tb00212.x.

Tatsuoka, K. K. (2009). *Cognitive assessment: An introduction to the rule space method.* New York, NY: Routledge.

Tatsuoka, K. K., Corter, J. E., & Tatsuoka, C. (2004). Patterns of diagnosed mathematical content and process skills in TIMSS-R across a sample of 20 countries. *American Educational Research Journal, 41*, 901-926. doi: 10.3102/00028312041004901.

龍岡菊美・林篤裕（2001）. 個人の潜在的知識ステートを診断する統計的方法論 計測と制御 , *40*, 561‐567. doi: http://doi.org/10.11499/sicejl1962.40.561.

龍岡菊美・倉元直樹（2006）. ルールスペース法(RSM) −テストの診断的利用に対する測定論的アプローチ− 日本テスト学会誌, *2*, 127–142.

Tatsuoka, K. K., & Tatsuoka, M. M. (1987). Bug distribution and statistical pattern classification. *Psychometrika, 52*, 193-206. doi: https://doi.org/10.1007/BF02294234.

Templin, J. L. (2006). *CDM Cognitive diagnosis modeling with Mplus user guid*e., Mplus. Available from: http://jtemplin.coe.uga.edu/files/dcm/cdm/CDM_user_guide.pdf.

Templin, J. L., & Bradshaw, L. (2013). Measuring the reliability of diagnostic classification model examinee estimates. *Journal of Classification, 30*, 251-275. 10.1007/s00357-013-9129-4.

Templin, J. L., & Bradshaw, L. (2014). Hierarchical diagnostic classification models: A family of models for estimating and testing attribute hierarchies. *Psychometrika, 79*, 317-339. doi: https://doi.org/10.1007/s11336-013-9362-0.

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11*, 287-305. doi : 10.1037/1082-989X.11.3.287.

Templin, J. L., Henson, R. A., Templin, S. E., & Roussos, L. (2008). Robustness of hierarchical

modeling of skill association in cognitive diagnosis models. *Applied Psychological Measurement, 32*, 559-574. doi: https://doi.org/10.1177/0146621607300286.

Templin, J. L., & Hoffman, L. (2013). Obtaining diagnostic classification model estimates using Mplus. *Educational Measurement, 32*, 37-50. doi: 10.1111/emip.12010.

Tjoe, H., & de la Torre, J. (2013). Designing cognitively-based proportional reasoning problems as an application of modern psychological measurement models. *Journal of Mathematics Education, 6*, 17-26. Available from: http://educationforatoz.net/images/2_Dec_2013.pdf.

Tjoe, H. & de la Torre, J. (2014). The identification and validation process of proportional reasoning attributes: an application of a cognitive diagnosis modeling framework. *Mathematics Education Research Journal, 26,* 237–255. doi: https://doi.org/10.1007/s13394-013-0090-7.

Traub, R. (1997). Classical test theory in historical perspective. *Educational Measurement, 16*, 8-14. doi: 10.1111/j.1745-3992.1997.tb00603.x.

植野真臣（2010）. 知識観の変遷と評価理論 植野真臣・荘島宏二郎（2010）. 学習評価の新潮流（pp.1-36.） 朝倉書店

van der Linden, W. & Glas, C. A. W. (2000). *Computerized adaptive testing: Theory and practice*. New York: Springer.

Vehtari, A., & Gelman, A. (2014). WAIC and cross-validation in Stan. Retrieved from: http://www.stat.columbia.edu/~gelman/research/unpublished/waic_stan.pdf

von Davier, M. (2006). Multidimensional latent trait modeling(*mdltm*) [Computer software]. Princeton, NJ: Educational Testing Service.

von Davier, M. (2007a). Hierarchical general diagnostic models. *ETS Research Report Series, 2007*, i-19. doi: 10.1002/j.2333-8504.2007.tb02061.x.

von Davier, M. (2007b). Mixture distribution diagnostic models. *ETS Research Report Series, 2007*, i-21. doi: 10.1002/j.2333-8504.2007.tb02074.x.

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology, 61*, 287-307. doi: 10.1348/000711007X193957.

von Davier, M. (2014a). The DINA model as a constrained general diagnostic model: Two variants of a model equivalency. *British Journal of Mathematical and Statistical Psychology, 67*, 49-71. doi: 10.1111/bmsp.12003.

von Davier, M. (2014b). The log-linear cognitive diagnostic model (LCDM) as a special case of the

general diagnostic model (GDM). *ETS Research Report Series, 2014*, 1-13. doi: http://dx.doi.org/10.1002/ets2.12043.

Wang, C. & Gierl, M. J. (2011). Using the attribute hierarchy method to make diagnostic inferences about examinees' cognitive skills in critical reading. *Journal of Educational Measurement, 48*, 165−187. doi: 10.1111/j.1745-3984.2011.00142.x.

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research, 11*, 3571–3594. Retrieved from: http://www.jmlr.org/papers/volume11/watanabe10a/watanabe10a.pdf

Watanabe, S. (2013). A widely applicable Bayesian information criterion. *Journal of Machine Learning Research, 14,* 867–897. Retrieved from: http://www.jmlr.org/papers/volume14/watanabe13a/watanabe13a.pdf

Whitely, S. (1980). Multicomponent latent trait models for ability tests. *Psychometrika, 45*, 479–494. doi: https://doi.org/10.1007/BF02293610.

Xiang, R. (2013). *Nonlinear penalized estimation of true Q-Matrix in cognitive diagnostic models*. (Doctoral dissertation, Columbia University). Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.976.6728&rep=rep1&type=pdf.

Xu, G. (2017). Identifiability of restricted latent class models with binary response. *Annals of Statistics, 45*, 675-707. doi: 10.1214/16-AOS1464.

Xu, G., & Shang, Z. (2017). Identifying latent structures in restricted latent class models. *Journal of the American Statistical Association,* doi: 10.1080/01621459.2017.1340889.

Xu, G., & Zhang, S. (2016). Identifiability of diagnostic classification models. *Psychometrika, 81*, 625-649. doi: https://doi.org/10.1007/s11336-015-9471-z.

Xu, X., & von Davier, M. (2008). Comparing multiple-group multinominal log-linear models for multidimensional skill distributions in the general diagnostic model. *ETS Research Report Series, 2008*, i-14. doi: http://dx.doi.org/10.1002/j.2333-8504.2008.tb02121.x.

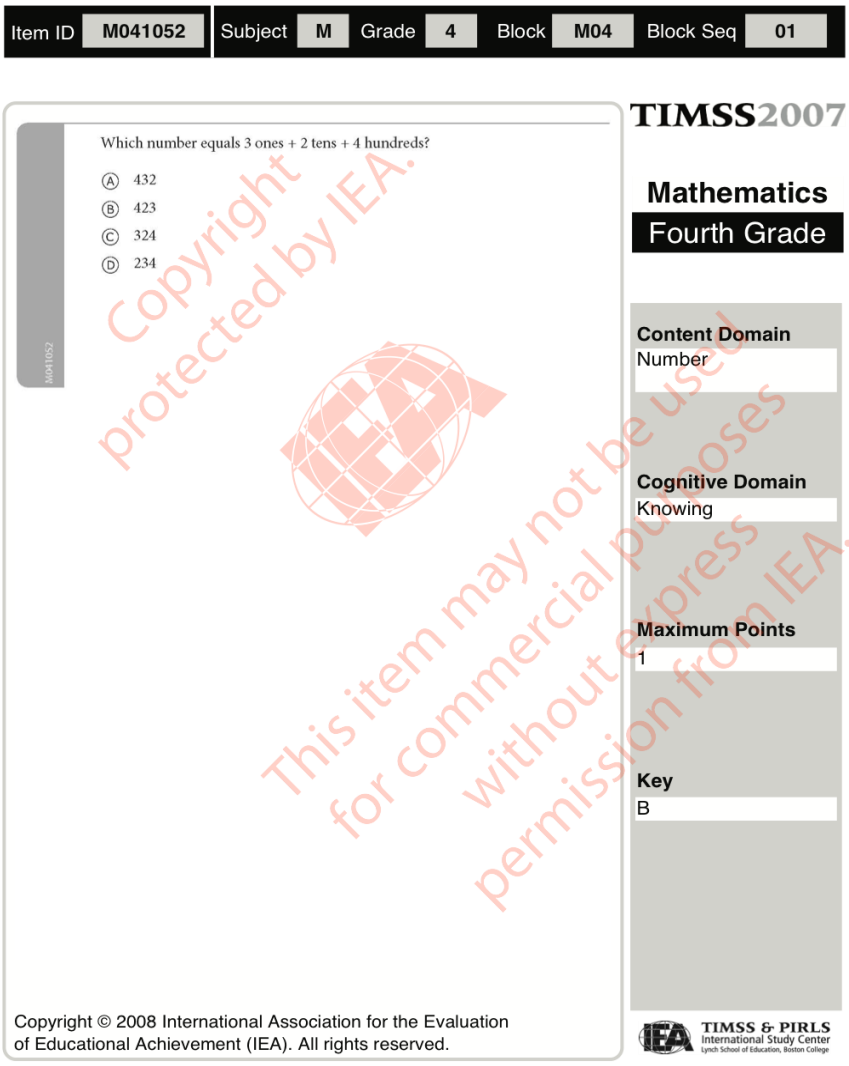
山口一大（2016）. 学習ログとテスト解答データの関連の検討：認知診断モデルを用いたアプローチ 日本行動計量学会第 44 回大会抄録集, 82-83.

Yang, X. & Embretson, S. E. (2007). *Construct validity and cognitive diagnostic assessment*. In J. Leighton & M. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and*

applications. Cambridge: Cambridge University Press, pp. 119–145. Available from: https://smartech.gatech.edu/bitstream/handle/1853/34250/embretson_CDAE_2007.pdf.

Zheng, Y., & Chiu, C. (2016). NPCD: Nonparametric methods for cognitive diagnosis [Computer software]. R package version 1.0-10. Retrieved from: https://CRAN.R-project.org/package=NPCD.

付録：第 4 章，第 5 章で用いた問題項目



Item ID M041052 | Subject M | Grade 4 | Block M04 | Block Seq 01

Which number equals 3 ones + 2 tens + 4 hundreds?

Ⓐ 432
Ⓑ 423
Ⓒ 324
Ⓓ 234

M041052

TIMSS2007

Mathematics
Fourth Grade

Content Domain
Number

Cognitive Domain
Knowing

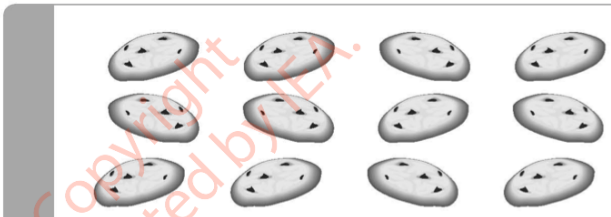
Maximum Points
1

Key
B

TIMSS & PIRLS International Study Center Lynch School of Education, Boston College

項目 1（M041052）の問題および正答

Item ID	M041056	Subject	M	Grade	4	Block	M04	Block Seq	02



There are 12 cookies. Draw a circle around $\frac{1}{3}$ of the cookies.

M041056

TIMSS2007

Mathematics

Fourth Grade

Content Domain
Number

Cognitive Domain
Knowing

Maximum Points
1

Key
See scoring guide

TIMSS & PIRLS International Study Center Lynch School of Education, Boston College

項目 2（M041056）の問題および正答

Item ID	M041069	Subject	M	Grade	4	Block	M04	Block Seq	03

TIMSS2007

Mathematics
Fourth Grade

Content Domain
Number

Cognitive Domain
Knowing

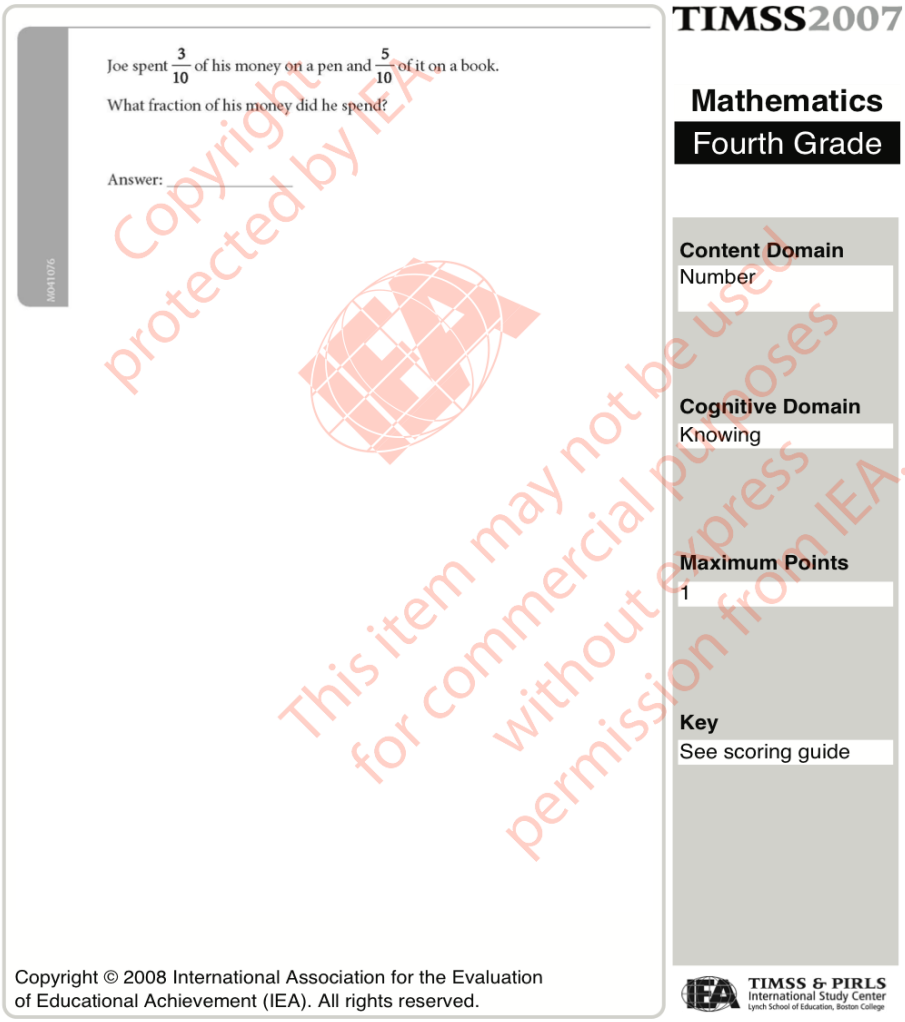
Maximum Points
1

Key
C

TIMSS & PIRLS International Study Center
Lynch School of Education, Boston College

項目 3（M041069）の問題および正答

Item ID	M041076	Subject	M	Grade	4	Block	M04	Block Seq	04



Joe spent $\frac{3}{10}$ of his money on a pen and $\frac{5}{10}$ of it on a book.

What fraction of his money did he spend?

Answer: _____

M041076

TIMSS2007

Mathematics

Fourth Grade

Content Domain
Number

Cognitive Domain
Knowing

Maximum Points
1

Key
See scoring guide

TIMSS & PIRLS International Study Center
Lynch School of Education, Boston College

項目 4（M041076）の問題および正答

Item ID	M041281	Subject	M	Grade	4	Block	M04	Block Seq	05

TIMSS2007

Mathematics

Fourth Grade

Content Domain
Number

Cognitive Domain
Applying

Maximum Points
1

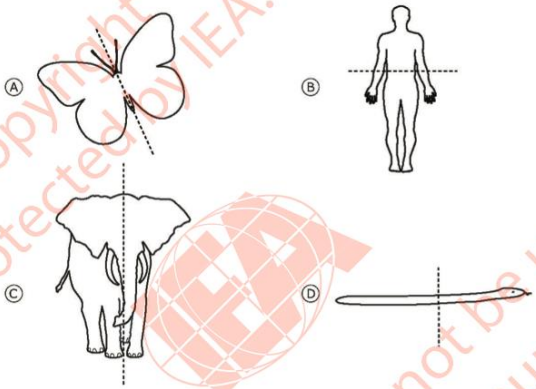
Key
D

TIMSS & PIRLS International Study Center
Lynch School of Education, Boston College

項目 5（M041281）の問題および正答

Item ID	M041164	Subject	M	Grade	4	Block	M04	Block Seq	06

In which of these drawings is the dotted line a line of symmetry?



M041164

TIMSS2007

Mathematics

Fourth Grade

Content Domain
Geometric Shapes and Measures

Cognitive Domain
Knowing

Maximum Points
1

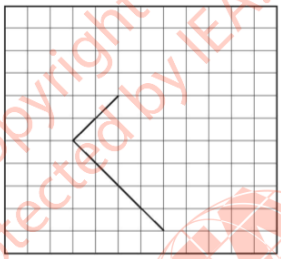
Key
A

TIMSS & PIRLS International Study Center Lynch School of Education, Boston College

項目 6（M041164）の問題および正答

Item ID	M041146	Subject	M	Grade	4	Block	M04	Block Seq	07

Here are two sides of a rectangle. Draw the other two sides.



M041146

TIMSS2007

Mathematics
Fourth Grade

Content Domain
Geometric Shapes and Measures

Cognitive Domain
Applying

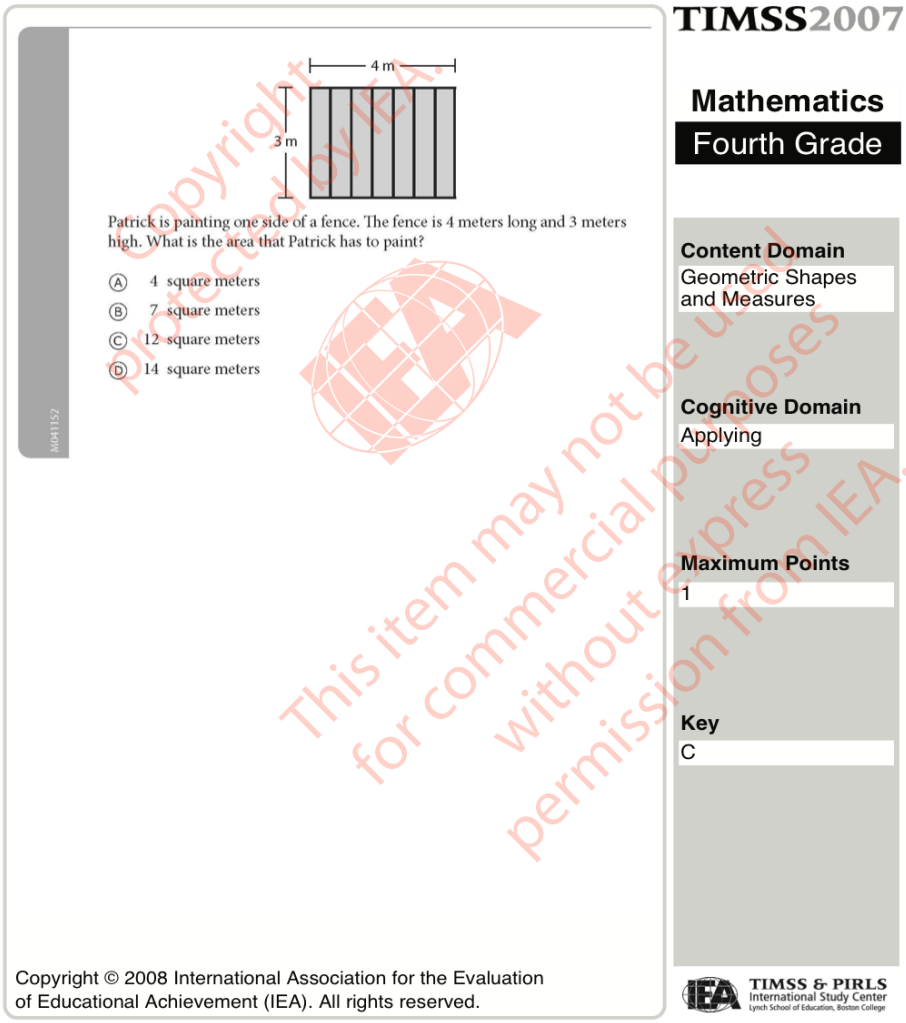
Maximum Points
1

Key
See scoring guide

TIMSS & PIRLS International Study Center Lynch School of Education, Boston College

項目 7（M041146）の問題および正答

Item ID	M041152	Subject	M	Grade	4	Block	M04	Block Seq	08



項目 8（M041152）の問題および正答

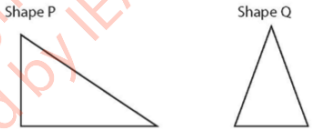
Item ID	M041258A	Subject	M	Grade	4	Block	M04	Block Seq	09

TIMSS2007

Mathematics

Fourth Grade

Two shapes are shown below. Describe one way they are the same and one way they are different.



A. Same

B. Different

M041258

Content Domain
Geometric Shapes and Measures

Cognitive Domain
Reasoning

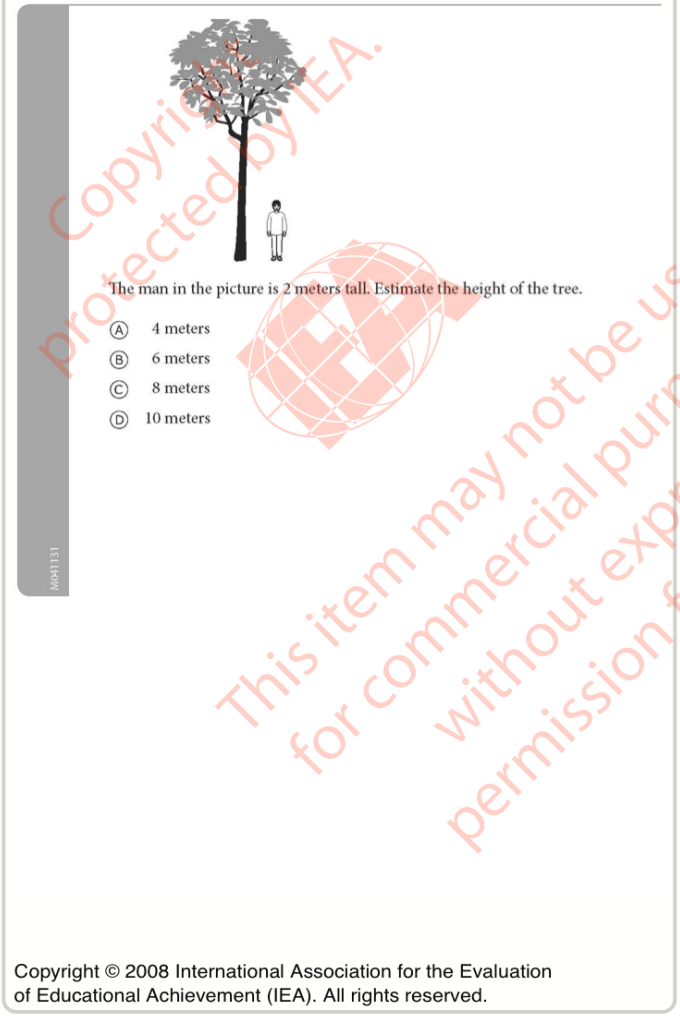
Maximum Points
1

Key
See scoring guide

TIMSS & PIRLS International Study Center
Lynch School of Education, Boston College

項目 9（M041258A），項目 10（M041258B）の問題および正答

Item ID	M041131	Subject	M	Grade	4	Block	M04	Block Seq	10



The man in the picture is 2 meters tall. Estimate the height of the tree.

Ⓐ 4 meters
Ⓑ 6 meters
Ⓒ 8 meters
Ⓓ 10 meters

M041131

TIMSS2007

Mathematics
Fourth Grade

Content Domain
Geometric Shapes and Measures

Cognitive Domain
Knowing

Maximum Points
1

Key
C

TIMSS & PIRLS International Study Center
Lynch School of Education, Boston College

項目 11（M041131）の問題および正答

Item ID	M041275	Subject	M	Grade	4	Block	M04	Block Seq	11

TIMSS2007

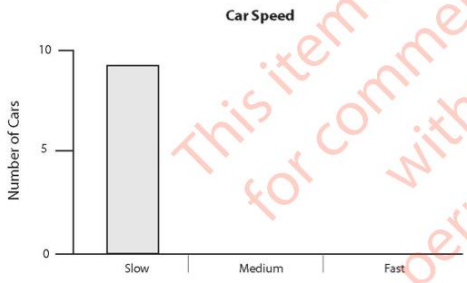
Mathematics
Fourth Grade

Several students were collecting information about how fast cars were driving by their school. The table below shows the results for 20 cars.

Car	Slow	Medium	Fast
1		X	
2	X		
3	X		
4			X
5			X
6	X		
7		X	
8		X	
9	X		
10	X		
11	X		
12		X	
13	X		
14			X
15			X
16	X		
17		X	
18	X		
19		X	
20			X

To make the results easier to read, the students started to put the information into the bar graph.

Complete the bar graph.



M041275

Content Domain
Data Display

Cognitive Domain
Applying

Maximum Points
2

Key
See scoring guide

TIMSS & PIRLS International Study Center
Lynch School of Education, Boston College

項目 12（M041275）の問題および正答

Item ID	M041186	Subject	M	Grade	4	Block	M04	Block Seq	12

TIMSS2007

Mathematics

Fourth Grade

The graph shows the number of apples John picked each day.

each stands for 10 apples

Monday	
Tuesday	
Wednesday	
Thursday	

On which day did John pick 5 apples?

(A) Monday
(B) Tuesday
(C) Wednesday
(D) Thursday

M041186

Content Domain
Data Display

Cognitive Domain
Knowing

Maximum Points
1

Key
D

TIMSS & PIRLS International Study Center Lynch School of Education, Boston College

項目 13（M041186）の問題および正答

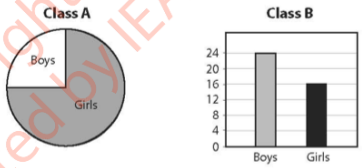
Item ID	M041336	Subject	M	Grade	4	Block	M04	Block Seq	13

TIMSS2007

Mathematics

Fourth Grade

Class A and B each have 40 students.



There are more girls in Class A than in Class B. How many more?

Ⓐ 14
Ⓑ 16
Ⓒ 24
Ⓓ 30

M041336

Content Domain
Data Display

Cognitive Domain
Reasoning

Maximum Points
1

Key
A

TIMSS & PIRLS International Study Center Lynch School of Education, Boston College

項目 14（M041336）の問題および正答

Item ID	M031303	Subject	M	Grade	4	Block	M05	Block Seq	01

M031303

There are 9 rows of chairs. There are 15 chairs in each row. Which of these gives the total number of chairs?

Ⓐ 15 ÷ 9
Ⓑ 15 − 9
Ⓒ 15 × 9
Ⓓ 15 + 9

TIMSS2007

Mathematics
Fourth Grade

Content Domain
Number

Cognitive Domain
Applying

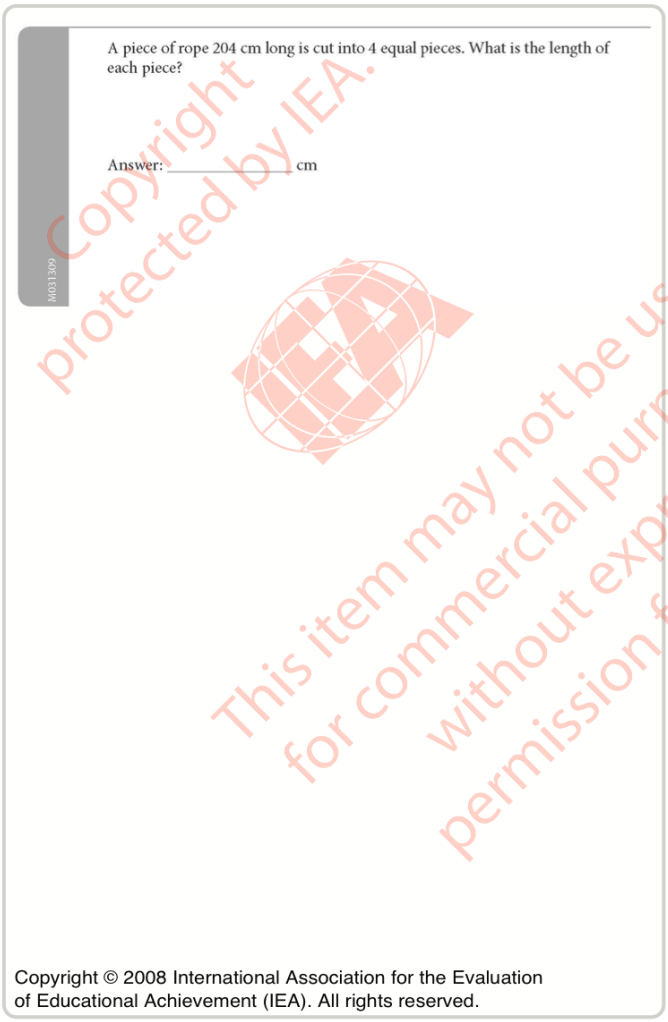
Maximum Points
1

Key
C

TIMSS & PIRLS International Study Center
Lynch School of Education, Boston College

項目 15（M031303）の問題および正答

Item ID	M031309	Subject	M	Grade	4	Block	M05	Block Seq	02



TIMSS2007

Mathematics

Fourth Grade

Content Domain
Number

Cognitive Domain
Applying

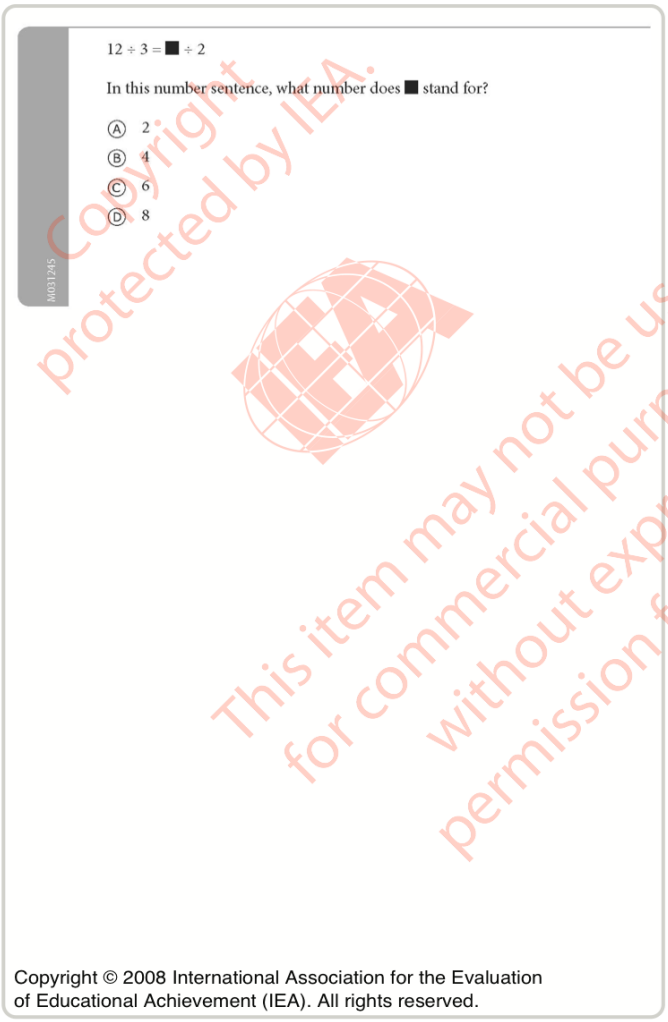
Maximum Points
1

Key
See scoring guide

TIMSS & PIRLS International Study Center
Lynch School of Education, Boston College

項目 16（M031309）の問題および正答

Item ID	M031245	Subject	M	Grade	4	Block	M05	Block Seq	03



TIMSS2007

Mathematics

Fourth Grade

Content Domain
Number

Cognitive Domain
Applying

Maximum Points
1

Key
D

TIMSS & PIRLS International Study Center
Lynch School of Education, Boston College

項目 17（M031245）の問題および正答

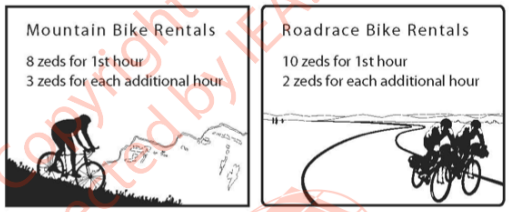
Item ID	M031242A	Subject	M	Grade	4	Block	M05	Block Seq	04

TIMSS2007

Mathematics

Fourth Grade

Posters for two sports clubs that rent bikes are shown below.



A. Use the information in the posters to complete the tables.

Mountain Bike Rentals	
Hours	Cost (zeds)
1	8
2	11
3	
4	
5	
6	

Roadrace Bike Rentals	
Hours	Cost (zeds)
1	10
2	12
3	
4	
5	
6	

B. For what number of hours are the rental costs the same at the two clubs?

Answer: _____

C. From which club does it cost less to rent a bike for 12 hours?

Ⓐ Mountain Bike Rentals
Ⓑ Roadrace Bike Rentals
Ⓒ They are both the same
Ⓓ It cannot be worked out

M031242

Content Domain
Number

Cognitive Domain
Applying

Maximum Points
1

Key
See scoring guide

TIMSS & PIRLS International Study Center
Lynch School of Education, Boston College

項目 18（M031242A），項目 19（M031242B），項目 20（M031242C）の問題および正答

Item ID	M031247	Subject	M	Grade	4	Block	M05	Block Seq	05

A man took his 3 children to a fair. Tickets cost twice as much for adults as for children. The father paid a total of 50 zeds for the 4 tickets.

How many zeds did each child's ticket cost? Show your work.

Answer: _____

M031247

TIMSS2007

Mathematics
Fourth Grade

Content Domain
Number

Cognitive Domain
Reasoning

Maximum Points
2

Key
See scoring guide

TIMSS & PIRLS International Study Center
Lynch School of Education, Boston College

項目 21（M031247）の問題および正答

Item ID	M031219	Subject	M	Grade	4	Block	M05	Block Seq	06

TIMSS2007

Mathematics
Fourth Grade

Jill had a rectangular piece of paper.

She cut her paper along the dotted line and made an L shape like this.

Which of these statements is true?

Ⓐ The area of the L shape is greater than the area of the rectangle.
Ⓑ The area of the L shape is equal to the area of the rectangle.
Ⓒ The area of the L shape is less than the area of the rectangle.
Ⓓ You cannot work out which area is greater without measuring.

M031219

Content Domain
Geometric Shapes and Measures

Cognitive Domain
Knowing

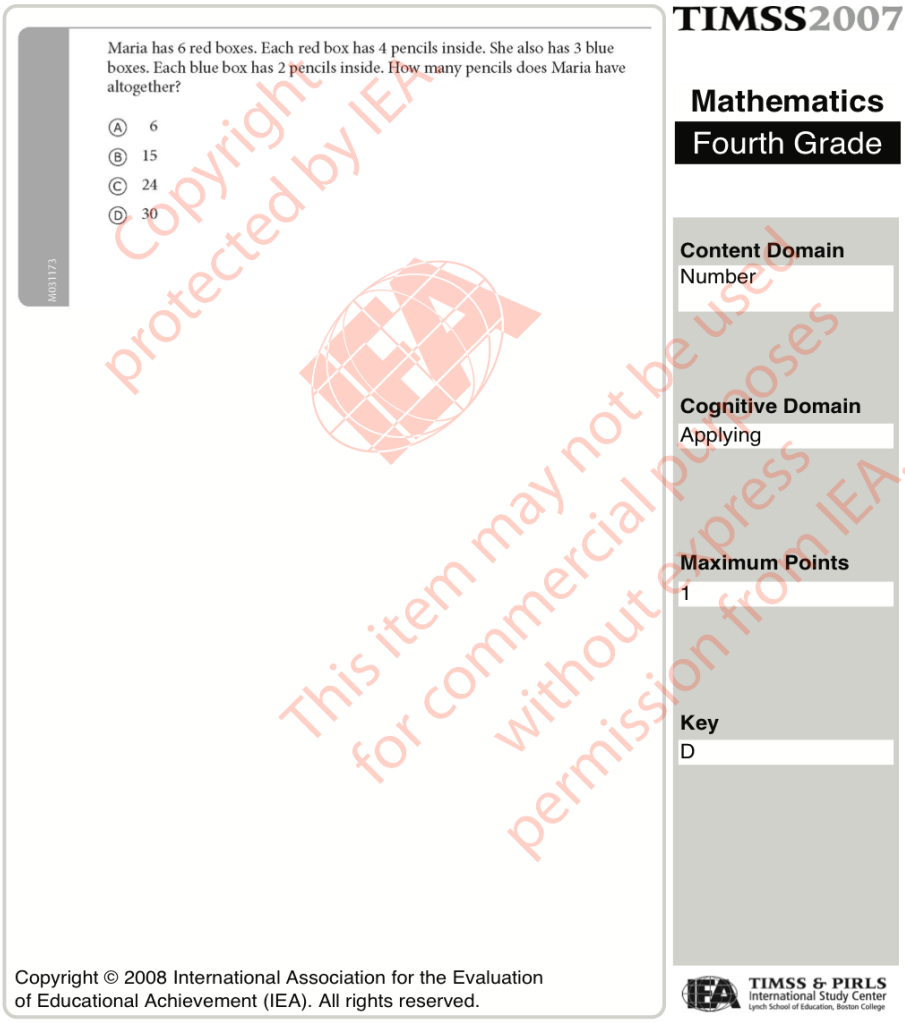
Maximum Points
1

Key
B

TIMSS & PIRLS International Study Center
Lynch School of Education, Boston College

項目 22（M031219）の問題および正答

Item ID	M031173	Subject	M	Grade	4	Block	M05	Block Seq	07



Maria has 6 red boxes. Each red box has 4 pencils inside. She also has 3 blue boxes. Each blue box has 2 pencils inside. How many pencils does Maria have altogether?

Ⓐ 6
Ⓑ 15
Ⓒ 24
Ⓓ 30

M031173

TIMSS2007

Mathematics
Fourth Grade

Content Domain
Number

Cognitive Domain
Applying

Maximum Points
1

Key
D

TIMSS & PIRLS International Study Center
Lynch School of Education, Boston College

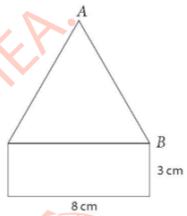
項目 23（M031173）の問題および正答

Item ID	M031085	Subject	M	Grade	4	Block	M05	Block Seq	08

TIMSS2007

Mathematics

Fourth Grade



The figure above is made from a rectangle and a triangle with three equal sides. What is the length, in centimeters, of side AB?

Ⓐ 8
Ⓑ 9
Ⓒ 10
Ⓓ 11

M031085

Content Domain
Geometric Shapes and Measures

Cognitive Domain
Knowing

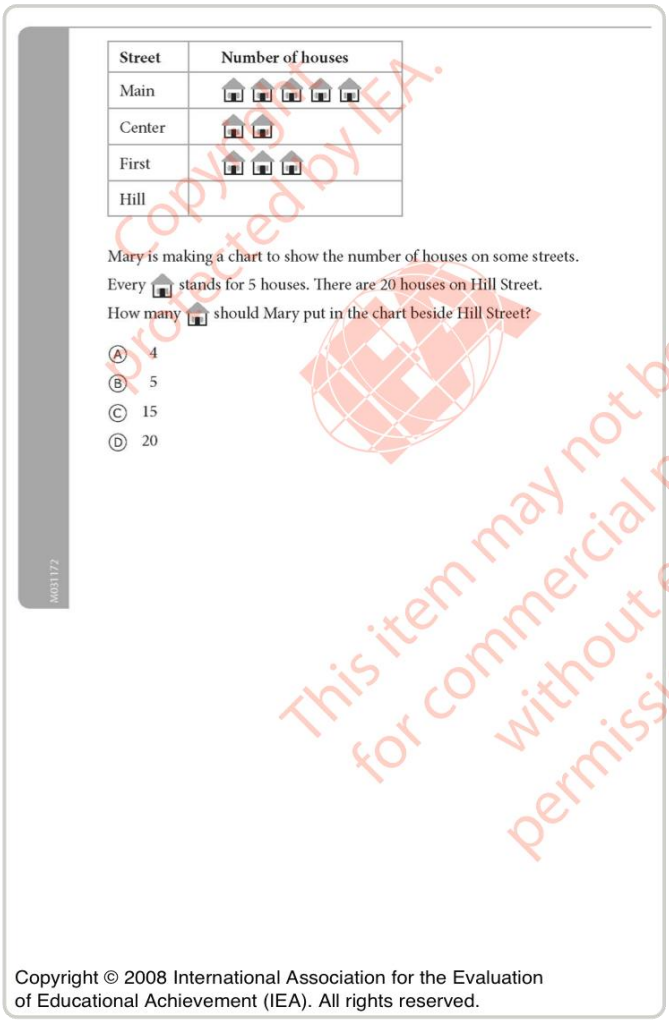
Maximum Points
1

Key
A

TIMSS & PIRLS International Study Center Lynch School of Education, Boston College

項目 24（M031085）の問題および正答

Item ID	M031172	Subject	M	Grade	4	Block	M05	Block Seq	09



Street	Number of houses
Main	
Center	
First	
Hill	

Mary is making a chart to show the number of houses on some streets.
Every stands for 5 houses. There are 20 houses on Hill Street.
How many should Mary put in the chart beside Hill Street?

Ⓐ 4
Ⓑ 5
Ⓒ 15
Ⓓ 20

M031172

TIMSS2007

Mathematics
Fourth Grade

Content Domain
Data Display

Cognitive Domain
Applying

Maximum Points
1

Key
A

TIMSS & PIRLS International Study Center
Lynch School of Education, Boston College

項目 25（M031172）の問題および正答

謝辞

博士論文を執筆するに当たり，多くの方のご支援とご指導を賜りました。著者の指導教員である東京大学大学院教育学研究科教授の南風原朝和先生には全体の構成や詳細な表現についてなど，多くの指導をいただきました。また，南風原先生には修士課程から 5 年間指導をいただきました。研究の観点や方向性について私自身でしっかりと考える機会を下さったことに深く感謝をしております。今後も研究者として，知見を生み出し続けられるように日々精進して参りたいと思います。

お忙しい中，博士論文の審査委員をお引き受け頂いた，東京大学大学院教育学研究科の針生悦子教授，藤村宣之教授，東京大学高大接続研究開発センター追跡調査部門の宇佐美慧准教授，専修大学人間科学部の岡田謙介准教授にもこの場を借りて御礼申し上げます。特に宇佐美先生には，様々な議論を通じて研究の展開の可能性について示唆をいただきました。岡田先生は共同研究を通じて具体的な研究の進め方などについてもご教授下さっただけでなく，常に前向きな姿勢を示し，温かいお言葉をかけてくださり挫けそうになる私を勇気づけてくださいました。本当に感謝しております。

また，教育学研究科教育心理学コースの皆様が研究に邁進されている様を見て，自分もがんばろうという気持ちをいただきました。5 年間で多少なりとも私が成長できたのも教育心理学コースの皆様のおかげだと感じております。

最後になりましたが，ここまでの私の人生を応援し，暖かく見守ってくれた家族に感謝の意を述べたいと思います。