異なる難易度のテスト項目の IRT 垂直尺度化

―尺度化テストデザインによる垂直尺度構成―

B7PM2102　澁谷拓巳

はじめに

学力テストによる客観的な評価においては，かつては相対評価のような集団準拠の評価が中心であり，常に集団のレベルに依存して個人の能力や成績が決定されてきた。一方，現在では目標に準拠した評価で個人の到達度を確認し，個に応じた指導が求められるようになっている。しかし従来の学習指導要領や教科書に基づいた学力テストでは特定の学習分野における現時点での学力を評価するに過ぎない。国語や算数・数学，英語といった教科は，たとえ単元や学期，学年をまたいでも学習内容は継続して，密接に関わり合っていることが多い。そのような教科では一時点での到達度だけでなく，これまでの成績も考慮して，学力の伸びを評価する方が適切ではないだろうか。児童生徒個人の学力の伸びを評価することができれば，学校教育においてどのように学力が身につき，児童生徒が成長するのかという学力発達のモデルを明らかにする手がかりを得ることができるだろう。それだけでなく個人の現在の習熟度に最適な学習を提供することにもつながり，さらに外的な基準に基づいた有用なテスト得点を算出することができる。

そのためには何よりもまず学力を測定するための尺度が必要になり，その尺度を構成するためのツールのひとつが IRT (Item Response Theory) モデルと垂直尺度化（vertical scaling）である。IRT モデルは受検者の潜在的な特性値とテストや質問項目のパラメタをモデル上で独立に定義し，それらの関数として項目への反応確率をモデリングしている。このモデルは心理尺度構成や学力テスト分析などに広く用いられているが，学力の伸びを測定するために適した特徴も備えている。たとえば間隔尺度水準のテスト得点が利用可能であることや，不変性により特定の項目だけに依存せずに受検者の能力を推定できることなどが挙げられるだろう。そして尺度の不定性を活かした柔軟な対応づけが実行できることも IRT モデルを利用する大きな強みである。

垂直尺度化は異なる難易度のテスト得点を比較するための手法であり，IRT の応用としての側面が強く，等化と非常に近い概念である。比較したい異なる難易度のテストは，たとえば算数の小学 5 年生のテストと小学 6 年生のテストといったように，測定する構成概念が類似していることが前提となっている。イメージとしては 30 センチの物差しの上半分に，同じく 30 センチの物差しをくっつけて，45 センチの物差しを作るようなものである。ただしくっつけたい物差し同士の目盛りの間隔が違うので，直接重なり合う 15 センチの部分の情報を利用して，その目盛りを振り直す作業もしなくてはならない。垂直尺度構成に適用できる IRT モデルや実行可能な推定方法，データ収集デザインは数多く存在するが，特に全学年共通の項目を含む尺度化テストデザインに着目した研究は皆無であるため，最適なモデルや推定方法について未検討である。

本稿ではいくつかの IRT モデルとその推定方法およびテスト得点の対応づけの手法，尺度調整法についてレビューをおこない，実験として，シミュレーションデータを用いて尺度化テストデザインに適した尺度調整方法について検討し，続いて，実際に収集された学力テストのデータを利用して，尺度化テストデザインにもとづく垂直尺度を構成する。この手法は単一の尺度を構成するだけにとどまらず，縦断的に拡張することで効率よく個人や集団の学力の伸びを推定することができる。データの制約により，本稿では個人や集団の学力の伸びを推定することはでき

ないため，基本的に単一年度・単一集団における垂直尺度化を主題とし，垂直尺度構成のための推定方法の比較や，構成された尺度の特徴に焦点を当てて分析をおこなっていく。

目次

図リスト

表リスト

1　心理計量モデル

1.1　心理計量モデルによる測定

ここでは垂直尺度構成の基盤的な役割を果たす心理計量モデルについて説明をおこなう。主に尺度水準，古典的テスト理論モデル，項目反応理論 (IRT) モデルについて述べていく。

1.1.1　構成概念と学力の発達

何らかの物体の量を測定するためには，その量に応じた尺度，物差しが必要である。物理量において，例えば温度であれば摂氏（℃）が日本で最も一般的に使われる尺度である。あるいは物体の長さであればセンチメートル（cm）やメートル（m）などが使用されることもあるだろう。一方欧米では華氏やヤードといった，異なる目盛りの指標が使われる事がある。

心理学や教育学では人の潜在的な特性や学力といった，本来は目で見えないものを伸ばしたり，評価，測定したりすることが多い。このときの評価の方法は，得点をつけたり，順位をつけたりと様々であるが，すべてに共通するのは測定したい対象があり，測定のための尺度が不可欠であるということだ。このときの測定の対象としたい特性の仮説的，理論的な概念を構成概念（construct）と呼ぶ。構成概念自体は実態を持たず，温度や長さのように直接そのものを観測するできないから，我々は観測可能な現象と構成概念の関連を考えることになる。たとえばいま，数学の学力という構成概念を定義したとして，数学の幅広い領域から小数の項目を選択し，項目の正答率や正答数で能力を測ることは，まさに測定のための尺度を構成し，誰かの数学の能力に数値を割り当て，測定することである。

測定に用いられる尺度には水準と呼ばれるいくつかの分類がある。Stevens (1946) の定義に従えば，測定のための尺度は，名義 (nominal) 尺度，順序 (ordinal) 尺度，間隔 (interval) 尺度，比 (ratio) 尺度の 4 つの水準に分類ができる。これら水準の中でも心理学の分野で対象とされるのはとりわけ順序尺度と間隔尺度である。順序尺度は物事の大小関係や順序を表す尺度であり，数学的な性質は入力値に対しての単調増加の関数であると定義される。統計的な特徴は中央値とパーセンタイルを定義できるが，平均や分散を定義することはできず，2 つ以上の値の差の意味を考えることができない。一般にこの尺度水準にあるデータの一部をカテゴリカルデータと呼ぶ。間隔尺度は量的なデータに対する尺度であり，線形変換可能な数学的性質を持つ。統計的にも扱いやすい値であり，平均や分散，相関係数などを計算できる。目盛りが等間隔に設定されており，原点が不定であることが特徴である。一方，同様に量的なデータに対する尺度である比尺度は絶対的な原点が 0 として定められている。

次に異なる 2 つの時点での測定結果を比較して，潜在的な特性の変化を測定することを考える。学力テストなどで測定される特性であれば伸びという言葉を使用し，語彙理解尺度や知能検査で測定される構成概念であれば発達という言葉を用いるのがふさわしいかもしれないが，この場合は特定の構成概念を想定せず，あくまでも変化と呼ぶことにする。潜在的な特性の 2 時点

間の変化を測定したいときに必要となるのは，厳密には間隔尺度か比尺度である。いま，20 問からなる数学の学力テストを考える。このテストは，前半は易しい問題だが，後半は応用的な難しい問題が含まれるという風な構成である。採点者は問題の難しさを考慮して，易しい問題は 1 点だが難しい問題は最大で 5 点とするように，あらかじめ項目得点に傾斜をかけ，合計で 100 点となるようにしている。このテストを尺度とみなしたとき，果たして順序尺度，あるいは比尺度の水準にあると言えるだろうか。確かに原点は 0 であるが，項目の難しさを等間隔に数値化できておらず，厳密には順序尺度水準にあり，用途は得点の大小比較にとどまるだろう。ただし，これでも実用上は平均や分散，相関係数を計算することが多く，間隔尺度か比尺度のように扱われる。

仮に，項目数をもっと増やして目盛りの間隔を非常に細かくすれば，項目ごとの難しさのばらつきは希薄になり，実用上は平均や分散を計算しても問題はないかもしれない。しかし 2 点間の特性の変化を測定したいときに問題となるのは尺度水準の問題だけではない。たとえば，身長の変化を測る際にセンチメートル単位の物差用いたとすれば，そのセンチメートル単位の数値を，フィート単位で測定した別の数値と直接比較することはできない。このことから言えることは，変化を測定したいのであれば，あくまでも測定の尺度は不変でなくてはならないということである。その点で先ほどの数学のテストの例を考えると，同一構成概念を測定していて，かつ変化を測定するためには，理論上全く等質なテストを 2 つ作成しておく必要がある。このときの等質と言うのは尺度の目盛りの単位や原点が揃っていることを意味しており，問題の出題傾向を大きく変更することや，難易度や教科の全く異なる問題などを出題することは，別の新たな尺度で測定することになるので，正確な測定のため避けるべきである。

つまり，潜在的な特性の変化を測定するためには，1）尺度水準は間隔尺度水準であり，2）尺度の単位や原点を揃えた共通尺度がなければならない，といえる。もちろんこの条件は精度の良い測定のために必要な条件である。実際，教育の現場で使用されるテストの多くはこの条件を満たしていないが，児童生徒が学力の伸びを実感したり，教師がそれを判断したりするための材料としては，尺度水準や尺度の単位などといった問題はまず意識されない。次は，潜在的な特性を測定する際の問題を別の観点から考察する。

1.1.2 個人間の差と個人内の差

知能測定の歴史を振り返ると，必ずしも先ほど確認した尺度に関する問題は意識されないまま，様々な知能検査が作成されてきたことが分かる。おそらく知能検査で最も有名なビネー式の知能検査は，ある個人がどの程度の発達段階にいるのかを知るために知能指数（IQ）を測定する道具である。

この IQ の利用方法はさておき，IQ とは「精神年齢÷暦年齢×100」で計算できる数値であり，テストで測定できるのは精神年齢の部分である。IQ を測定するための検査を知能検査と呼ぶが，この項目はたんに記憶力を試すものではなく，いわゆる日常的な物事に対する判断や理解力を試すようなものである。知能検査のテスト項目は，あらかじめ様々な年齢の児童に対して試行さ

れた項目を通過率で分析しており，通過率がちょうど 50 パーセントになるあたりの年齢を当該項目に回答するにふさわしい年齢としている。つまり尺度の目盛りを項目に集団のおよそ半数が正しく回答できるであろう年齢に依存させたのである。これを年齢尺度と呼ぶ(市川, 1991)。こうして項目に回答するにふさわしい年齢の情報が得られた後に児童に対して知能検査は実施され，低い年齢から順に出題していき，やがて正しい回答が得られなくなった時点で試験を打ち切り，その項目の年齢から精神年齢を得ることができる。IQ は定数項を除けば年齢 1 あたりの精神年齢の数値と考えることができるので，年齢に比例して精神年齢も線形的に変化するという仮定が満たされるのであれば，先ほどの Stevens の尺度水準の間隔尺度に当てはまるであろう。

IQ は確かに客観的に児童の知能を測定することに一役買っているが，この指数はいわば偏差値のようなものであり，その児童が属している年齢集団の中で相対的にどれくらい成長が早いか，遅いかを知ることができるに過ぎない (永野, 2001)。すなわち個人間の特性の差を知ることはできるが，個人内の発達の軌跡を明らかにすることはできないといえる。個人の異なる時点での測定値を比較するためには集団準拠ではなく，何らかの外的な基準に準拠した尺度や評価基準が必要である。

小学校 6 年間や中学校の 3 年間といった連続した学年にまたがって，児童生徒の学力の達成 (発達) を測定するための尺度を発達得点尺度 (developmental score scale) とか単に発達尺度 (developmental scale, Young & Tong, 2016) などと呼ぶ。あるいは尺度を構成する際に，レベルが等しいテストを水平的に比較するのではなく，異なる難易度のテストを垂直的に比較する必要があることから，垂直尺度 (vertical scale) とも呼ばれる。この尺度の構成のために用いられるのが心理計量モデル (psychometric model) である。代表的な心理計量モデルには因子分析 (Factor Analysis, FA) モデルや項目反応理論モデルなどがある。これらの心理計量モデルはそれぞれ単独の理論として扱われることもあるが，それらに共通する点はアンケートや学力テストなどのテスト全般を構成する為のモデルであるため，広くテスト理論という枠組みで扱われることもある。次はこのテスト理論という枠組みで潜在変数を測定するためのモデルについて概観する。

1.1.3 モデルで測定できるもの

先ほどの数学のテストでは，まず数学の構成概念を仮定し，その構成概念を測定していると想定される小数の項目の集合を尺度として考えていた。しかし，理論上の数学の構成概念とは計算する能力や数学的な考え方についての概念であるはずで，選択された小数の項目に回答する能力そのものではない。我々が直接観測できるデータはあくまでも項目に対する反応であるが，それは様々な要因が複雑に関係しながら生じた結果であると解釈することができる。たとえば，伝統的な，紙に書かれた問題用紙を読み，それに鉛筆で紙に記述して回答する形式のテストであれば，純粋に数学的な思考を測定する以外に数多くの外的な要因が関わっている。それは問題文を読む力や解答用紙に記入する技術，あるいは回答する空間の気温や音など集中力に影響する環境要因も考えられるだろう。そう考えると我々が観測しているデータは，本当に測定したい能力

とそれ以外のノイズ，誤差が加わった結果であると解釈することが妥当である。

この考え方は潜在変数 (latent variable) を扱う統計的なモデルの基本であるとともに，古典的テスト理論 (Classical Test Theory, CTT) モデルの考え方そのものでもある。このノイズのことを誤差 (error) と呼ぶが，直接観測することは不可能であるが，特定の条件下で推定をすることは可能である。誤差にはいくつかの種類があり，特にテストをおこなう中で生じるであろう誤差を野口・大隅 (2014)は (1) 採点に関わる誤差，(2) 受検者の回答に関わる誤差，(3) 受検者に内在する誤差, (4) テスト項目の抽出における誤差, の 4 つに分類している（一部表現を改めている）。CTT では観測されたテスト得点をX，真の得点をTとし，さらにそこに加わる様々な誤差をEとして，以下のような基本式をおいている。すなわち，

$$X = T + E, \tag{1.1}$$

と表現される。この古典的テスト理論については後ほど詳細に取り扱う。

ここで重要なのは，我々が本当に知りたい値は式（1）におけるTやEであり，Xではないと言うことである。したがって何らかの方法でこの潜在的な変数を推定したり，あるいは様々な制約をおいた上でXの値を解釈するのである。また注意すべき点として，測定したいTには実態がないということが挙げられる。

1.1.4 潜在変数の測定

潜在変数を推定するためには，分析者の仮説や得られたデータにふさわしい心理計量モデルが必要になる。データからいくつかの潜在変数のまとまりを取り出したいのであれば因子分析モデルを適用するのが良い。ある集団における複数の変量間の関係性を少数の仮説的な因子 (factor) に次元圧縮をおこなう因子分析モデルは, 潜在変量を扱う心理計量モデルの一種である。一般的な因子分析モデルは,

$$\boldsymbol{Z} = \boldsymbol{F}\boldsymbol{A}' + \boldsymbol{U}\boldsymbol{D}, \tag{1.2}$$

のように表現される。このとき，$\boldsymbol{Z}$は行にN個の個体のデータを，列にm個の変量のデータを持つ標準化されたデータ行列である。そして，$\boldsymbol{F}$は因子スコア行列，$\boldsymbol{A}$は因子負荷行列，$\boldsymbol{U}$は独自因子スコア行列，$\boldsymbol{D}$は独自因子スコアにかけられる重み行列である。このとき変量データ行列$\boldsymbol{X}$は間隔尺度や比尺度水準にある連続量のデータからなる行列である。因子分析モデルはデータ行列$\boldsymbol{Z}$におけるm個の変量の変動を，p個の仮説的な因子で説明するためのモデルである。このときの$\boldsymbol{U}\boldsymbol{D}$成分は，明示的に定められたモデルの因子によって説明されない，変量固有の変動に関わる部分であり，いわばモデルの誤差のようなものでもある (芝, 1981)。

得られた反応データからこのモデルの諸変数を推定することは，実質的には因子分析モデルにおける因子スコア行列や因子負荷行列といったモデルのパラメタ（母数）を解析的，あるいは

数値的に推定する作業である。構成概念を定義する部分でも述べたことと同様に，この因子についてもあくまでも仮説的な概念に過ぎず，実態のあるものではない。その実態のない概念を合理的なモデルを用いてデータから如何に推論をおこなうことが，潜在変数を測定することになる。言い換えれば，分析者が持つ仮説的なデータ生成モデルを実際のデータに当てはめてパラメタを推定することが，モデルを通して構成概念を間接的に数値化する方法である。なお，潜在変数と顕在変数との関係を考えたモデルは，因子分析を内包するより一般的な分析手法である共分散構造分析の文脈における測定方程式にあたる。

1.2 古典的テスト理論

1.2.1 モデルと基本的な仮定

式 (1.1) で示したように，古典的テスト理論では観測値のテスト得点を真のテスト得点と誤差に分解して考える。ただしいくつかの仮定をおくことで，後の信頼性 (Reliability) に関する議論をスムーズにできるため，ここで重要な仮定を示しておく。まず受検者に関する添え字をi，繰り返しの試行に関する添え字をkとする。いまN人の受検者集団に対してM回の繰り返しの測定をおこなったとすると，誤差に関してはN人の受検者とM回の試行両方について，平均は 0 となる。すなわち，

$$\bar{E}_i = \frac{1}{M}\sum_{k=1}^{M} E_{ik} = 0, \tag{1.3}$$

$$\bar{E}_k = \frac{1}{N}\sum_{i=1}^{N} E_{ik} = 0, \tag{1.4}$$

であり，したがって，

$$\bar{X}_i = T_i, \tag{1.5}$$

$$\bar{X} = T, \tag{1.6}$$

である。式 (1.5) は同一受検者に対する繰り返しの試行の期待値は真値に一致することを表しており，式 (1.6) は十分大きな集団の観測値の期待値は真値に一致することを表している。なお，$\bar{X}$は期待値E[X]と表現されることもある。

次に，この観測値の分散について考える。通常であれば真値の分散と誤差の分散，そして真値と誤差の共分散の 3 つの項に分解されるが，ここでは真値と誤差は直交すると仮定するため，分散$V[T,E]=0$を用いて，

$$V[X] = V[T] + V[E], \tag{1.7}$$

となり，観測値の分散は真値と誤差の分散のみに分解される。

テストが測定ツールである以上，体重計や物差しのように測定の誤差が存在し，測定の精度 (accuracy) を知る必要があるだろう。その誤差のばらつきについての指標を測定の標準誤差 (Standard Error of Measurement, SEM) と呼ぶ。この標準誤差が小さければ測定値は真値の周辺に分布することとなり，測定の精度が高いと言える。また，測定の精度についての指標はもうひとつ存在する。それが信頼性係数 (reliability coefficient) である。式 (7) の両辺を$V[X]$で割り，

$$1 = \frac{V[T]}{V[X]} + \frac{V[E]}{V[X]}, \tag{1.8}$$

を得る。この第一項は測定値の分散に占める真値の分散の大きさを表しており，これを信頼性係数と呼び，

$$\rho(X) = \frac{V[T]}{V[X]}, \tag{1.9}$$

と表記する。信頼性係数は$V[E] \geq 0$である事から，その最大値は 1 であり，誤差しか測定できていないと仮定すると最小値は 0 となる。測定の標準誤差は誤差分散の正の平方根であるから，

$$\sigma(E) = \sqrt{V[E]} \tag{1.10}$$

である。信頼性係数の計算には真の得点の分散が必要になるが，これは観測されない値であるため，何らかの方法で推定する必要がある。

次にふたつの異なるテストの平行測定について考える。ひとつの集団にふたつの異なるテストを実施したときに，どちらのテストもすべての受検者に対して真の得点が等しく，測定値の誤差分散が等しいような場合に，このふたつのテストは平行であると言う。平行なふたつのテストを$X_A = T + E_A$，$X_B = T + E_B$とおくと，どちらのテストも誤差の期待値は 0 であるのは式(3)で示したとおりなので，

$$\bar{X}_A = \bar{X}_B \tag{1.11}$$

である。各テストの分散に関して，誤差分散は等しいと仮定したので，

$$V[X_A] = V[X_B] \tag{1.12}$$

となり，ふたつのテストの共分散は真値と誤差が直交するという仮定と，異なるテストの誤差は直交するという仮定をおくことで，

$$V[X_A, X_B] = V[T] + V[T, E_A] + V[T, E_B] + V[E_A, E_B] \tag{1.13}$$

となり，第一項以外はすべて 0 になるので，結局，

$$V[X_A, X_B] = V[T] \tag{1.14}$$

である。

上で示したように，ふたつの測定の真値が等しく，誤差分散も等しいという非常に強い仮定をおく測定を強平行測定と呼び，一方この仮定を緩めて，真値が等しいという条件のみをおくものを τ 等価な測定，真値が等しくなくとも，その差が等しければよいとする測定を弱平行測定というように区別することもある。

信頼性係数はこの平行測定の仮定を利用して推定する必要がある。平行測定の仮定を利用すれば，ふたつの平行なテストの相関係数から信頼性係数を導出することができる。

$$r(X_A, X_B) = \frac{V[X_A, X_B]}{\sqrt{V[X_A]V[X_B]}} = \frac{V[T]}{V[X_A]} = \frac{V[T]}{V[X_B]} \tag{1.15}$$

しかし実際のテストでは完全に平行なテストを作ることも，平行であることを保証することも困難である。一般的なテストにおいては，平行なテストなどで信頼性係数を測定せずに，一回の測定のみで信頼性係数を推定できることが望ましい。そこで一回の測定を 2 つに折半することで擬似的な平行テストを作ることを考える。ここでは便宜上偶数個の項目からなるテスト X を想定する。項目数は2m個であるとすると折半の数は合計で $l = \frac{1}{2}{}_{2m}\mathrm{C}_m$ 個存在するため，この組み合わせすべてでの相関係数を考え，そのテストの信頼性係数として考えることができるだろう。導出は割愛するが，このようにして求められる信頼性係数の推定値をクロンバックの α 係数 (Cronbach, 1951) と呼び，項目数を J 個とすると，

$$\alpha = \frac{\mathrm{m}}{m-1}\left\{1 - \frac{\sum_{j=1}^{J} V[X_j]}{V[X]}\right\} \tag{1.16}$$

と表現できる。ただし X_j は項目jの項目得点である。一般にこのクロンバックの α 係数は内的整合性や内的一貫性の指標として用いられるが，要するにすべての折半の方法を考慮したときの信頼性係数の推定値の平均値である。注意したいのは2点であり，まずクロンバックの α 係数はあくまでも真の信頼性係数の推定値に過ぎないことであり，さらにクロンバックの α 係数は真

の信頼性係数の下界の最大値を与えている (Lord, Novick & Birnbaum, 1968; 2008) ことである。なお，クロンバックの α 係数は項目数が大きければ大きな値をとる。信頼性係数の推定値についてのより詳細な議論については岡田 (2015)などを参照されたい。

この信頼性係数の推定値を利用して求められるのが測定の標準誤差 (SEM),

$$\mathrm{SEM} = \sigma_X\sqrt{1-\hat{\rho}}, \tag{1.17}$$

である。ただしσ_XはテストXの得点の標準偏差である。

CTTでは平均や分散，信頼性係数といった統計量でテスト得点を分析することができるが，そのほかに項目ごとに困難度や識別力を推定することができる。もちろん後述するIRTの困難度と識別力とは異なる値ではあるが，比較的に簡単な計算により求めることができる場合が多く，IRTにおける分析の初期段階として用いられることが多い。項目困難度は，通過率とも呼ばれ，全受検者に占めるその項目に正答した受検者の割合である。すなわち，受検者が項目jに正答していれば1を，誤答していれば0が与えられている反応ベクトルを$\mathbf{u_j}$とすれば,

$$\mathrm{p_j} = \frac{1}{N}\mathbf{u_j}\mathbf{1}^{\mathrm{T}} \tag{1.18}$$

と表現されるのが項目困難度である。ただし$\mathbf{1}$は単位ベクトルであり，列ベクトルである。Tは行列の転置を表す記号である。項目識別力は，古典的テスト理論では全体のテスト得点と項目得点の相関係数の値の事を指し，0から1の範囲の値をとるため，項目テスト得点相関，I-T相関などとも呼ばれる。テストの合計得点ベクトルを$\mathbf{X}$とすると,

$$\mathrm{cor}(\boldsymbol{X},\mathbf{u}) = \frac{\frac{1}{\mathrm{n}}(\boldsymbol{X}-\bar{\boldsymbol{X}})\left(\mathbf{u_j}-\bar{\mathbf{u}}_{\mathbf{j}}\right)^{\mathrm{T}}}{\sqrt{\boldsymbol{V}[\boldsymbol{X}]\boldsymbol{V}^{\mathrm{T}}[\mathbf{u}_{\boldsymbol{j}}]}}, \tag{1.19}$$

と表すことができる。ただし，分母はテスト得点ベクトルと項目得点ベクトルの標準偏差の積である。テスト得点は多値のデータであるのに対し，この場合項目得点は二値のデータであるが，このようなベクトルの相関係数を点双列相関係数(point biserial correlation coefficient)と呼ぶ事もある。

CTTは誤差と真値の平均や共分散に関する仮定しか持たない。それらは後述するIRTの仮定に比べて弱く，一般的にCTTは弱い理論，モデル (weak theory, model) などと呼ばれることがある。それ故，分析に必要なサンプルサイズは比較的少数でもよく，統計量の計算も解析的に求める事ができ，解釈も容易である(Hambleton and Jones, 2005)。

1.2.2 古典的テスト理論の制約

古典的テスト理論のテスト得点は素点，つまり項目数に依存する。したがって異なる項目数のテスト得点をそのまま比較することはできない。また，尺度水準も厳密に言えば順序尺度水準の得点である。いま，テスト得点 Xを十分項目数の多いテストのものであると仮定すれば，実用上は間隔尺度水準の尺度得点として扱うことができるため，線形変換可能になる。これを利用して前者の問題を解決するために得点の標準化 (standardization) をおこなう。得点の標準化は平均と標準偏差を用いて，

$$Z = \frac{X - E[X]}{\sqrt{V[X]}} \tag{1.20}$$

とすれば良い。こうすることで得点の分布は平均 0，標準偏差 1 の尺度に変換される。これを標準得点と呼び，例えば同一集団が受検する異なるテストにおける個人の相対的な順位の比較や変化の追跡には十分である。

しかし CTT は真値と誤差の平均と分散，共分散の仮定のみをおく弱い理論，モデルであるため，現実のテストデータに合わせやすいという利点がある反面，いくつかの限界がある (Hambleton and Jones, 2005)。例えば標準化をおこなってもなお，その得点は集団あるいはテスト項目に依存してしまっている点である。異なる学力水準にある受検者集団が同一のテスト項目を受けても，標準得点はその集団での相対的な位置を示すに過ぎないため，能力が同程度の受検者であっても高い水準の集団にいる方が低い得点になることもある。逆に，単一集団であっても難しいテストと易しいテストの得点を比較した場合，易しいテストは困難度，識別力ともに低く受検者全体の得点が高い可能性がある。このように CTT では受検者のテスト得点が項目困難度に依存したり，項目困難度や識別力が標本集団の能力水準に依存したりする特徴がある。

このような特徴は単一のテストを用いる受検者の選抜や調査では大きな問題にはならないが，尺度を作成する過程で複数の集団にテストを受検させ，複数のレベルのテストを共通尺度化したい場合に不都合が生じやすい。また CTT のモデルは受検者の観測されたテスト得点を真値と誤差に分解したモデルに過ぎず，その得点が得られた背景（潜在変数）をモデリングしているわけではない。そのためモデルとデータの適合度という観点から分析をおこなうことができないなどモデルの柔軟性に乏しい。

1.3 項目反応理論

CTT ではテスト得点全体を考えていたが，FA モデルでは変量（項目）ごと，あるいは潜在変数ごとの関係性を考えていた。IRT でも FA 同様に項目ごとの反応と受検者の潜在的な能力との関係を考えることとなる。現在主流の IRT モデルは概ねロジスティックシグモイド関数を用いるモデルであり，統計学の一般的な枠組みではリンク関数にロジスティックシグモイド関数を用いて，線型モデルを非線形の出力に変換する一般化線形モデルの一部とみなすことができる。

ただし後述するように誤差に正規分布を仮定するため，より正しくは一般線形モデルである。さらに IRT モデルの 2 母数正規累積モデル (2-parameter normal ogive model) は変数が順序尺度水準のように質的なデータである場合の因子分析，すなわちカテゴリカル因子分析と数理的には同一である (豊田, 1998)。しかし実際に IRT モデルのパラメタを推定し，それに付随する統計量を分析する手法は IRT 研究の文脈で発達してきたものが多く，一般化して語るメリットがないため，本稿では IRT モデルと FA モデルの関連を述べるものの，それ以上の一般化の議論については触れない。そして本節おける IRT モデルは特別に表記がある場合を除いてすべて一次元 (unidimensional) IRT モデルのことを指す。

1.3.1 二値型モデル

はじめに最も基本的な二値(dichotomous, binary)型のデータに対する IRT モデルについて説明する。また，ここでは Lord (1980, pp30-34) と村木 (2011, pp42-45) を参考に正規累積モデルとロジスティックモデルの考察も兼ねることとする。はじめに受検者の潜在的な一次元能力の尺度θを考える。この尺度は$-\infty < \theta < \infty$の範囲にわたって存在するものとする。次にこの能力と項目反応$u_j = \{0,1\}$の関係について考える。Lord (1952) はこの関係を正規累積分布関数で表すことを考えた。この尺度上で平均μ，分散σ^2の正規分布は，

$$N(\theta|\mu_j, \sigma_j^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\left(\frac{\theta-\mu_j}{\sigma_j}\right)^2\right\} \tag{1.21}$$

と表される。この累積分布関数が当初の研究 (Lord, 1952; Lord, 1953) などで用いられている正規累積モデル (normal ogive model) であり，

$$P_j(\theta) = \int_{-\frac{\theta-\mu_j}{\sigma_j}}^{\infty} N(\theta|\mu_j, \sigma_j^2)\mathrm{d}\theta \tag{1.22}$$

と表現されるモデルである。このときのσ_jは尺度の単位を定め，μ_jは原点を定めるパラメタの役割を果たす。式 (1.22) は受検者がある項目jに正答する確率をモデリングしている。

ここで新たに受検者の項目jにおける反応を決定する潜在変数Y_j'について考える。この潜在変数Y_j'が項目ごとの閾値定数γ_jを超えていれば 1 を，下回っていれば 0 という反応を得るとする。ただしこのY_j'はθについての関数であり，以下の 3 つの仮定がおかれている。1）Y_j'のθへの回帰関数$\mu'_{j|\theta}$は線形であり，2）この回帰についてY_j'のちらばり (scatter) は等分散 (homoscedastic) であり，3）Y_j'のθについての条件付き分布は正規分布である。ただしこの条件に示した回帰関数μ'は正規分布のパラメタとは全く異なるものであり，区別するために「′」を添えている。

この回帰関数とY_j'の条件付き分布，閾値定数γ_jを図示した (図 1.1) 。図 1.1 における正規分布

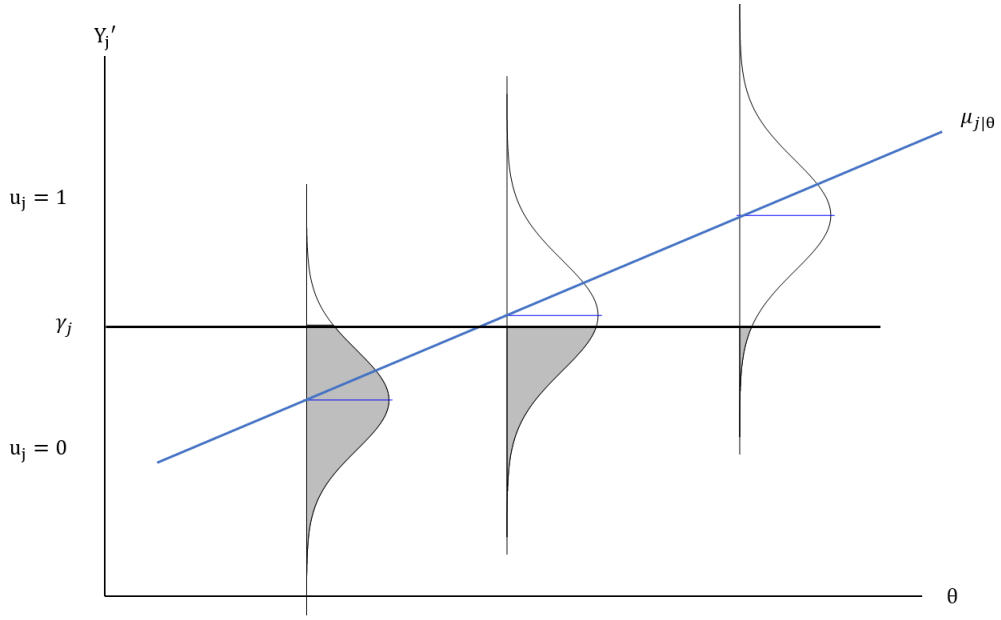

図 1.1 Y_j'のθへの回帰関数

はθで条件付けられたY_j'の確率分布である。この正規分布の平均は回帰直線$\mu_{j|\theta}$で与えられており，図中では便宜上 3 つの点での条件付き分布を示している。正規分布のグレーで塗りつぶされている部分の面積は$u_j = 0$となる確率を，反対に白い部分は$u_j = 1$となる確率を表現している。この確率の大きさはθの値と閾値定数γ_jによって定まる。

回帰直線$\mu_{j|\theta}$の傾きをρ_jとするとこの回帰直線は，

$$\mu_{j|\theta} = \rho_j \theta, \tag{1.23}$$

である。このときρ_jはふたつの潜在変数Y_j'とθの相関係数と見なせる。この傾きを利用して，かつθの尺度を平均 0，標準偏差 1 とすると，条件付き分布の標準偏差をZ_jとおいて，その符号を逆にしたものは式 (1.19) より，

$$-Z_j = \frac{\mu_j - \theta}{\sigma_j} = \frac{\gamma_j - \mu_{j|\theta}}{\sigma_{j|\theta}} = \frac{\gamma_j - \mu_{j|\theta}}{\sqrt{1-\rho_j^2}} = -\frac{\rho_j}{\sqrt{1-\rho_j^2}}\left(\theta - \frac{\gamma_j}{\rho_j}\right), \tag{1.24}$$

とであり，この表現においてμとσはρとγに置き換えられる。

ここまでの説明で用いてきた正規累積モデルには積分計算が入っているため，今後パラメタ推定をする際に微分，積分をおこなう上で非常に不便である。そこで Lord et al. (2008, p. 399) はロジスティック関数を用いてこれを近似する方法を示している。つまり，

$$P_j(\theta) \cong \frac{\exp(DZ_j)}{1+\exp(DZ_j)} = \frac{1}{1+\exp(-DZ_j)}, \tag{1.25}$$

ただし，D は尺度定数であり D=1.7 であるときに全域にわたって正規累積分布関数とロジスティック関数の誤差が 0.01 以下になることが知られている (Haley, 1952; Camilli, 1994)。しかし Kullback-Leibler 情報量を最小にするように D を最適化すると D=1.749 であるため，こちらを支持する意見もある (Savalei, 2006)。2 種類の D のロジスティック関数と正規累積分布関数の比較を図 1.2 に示した。

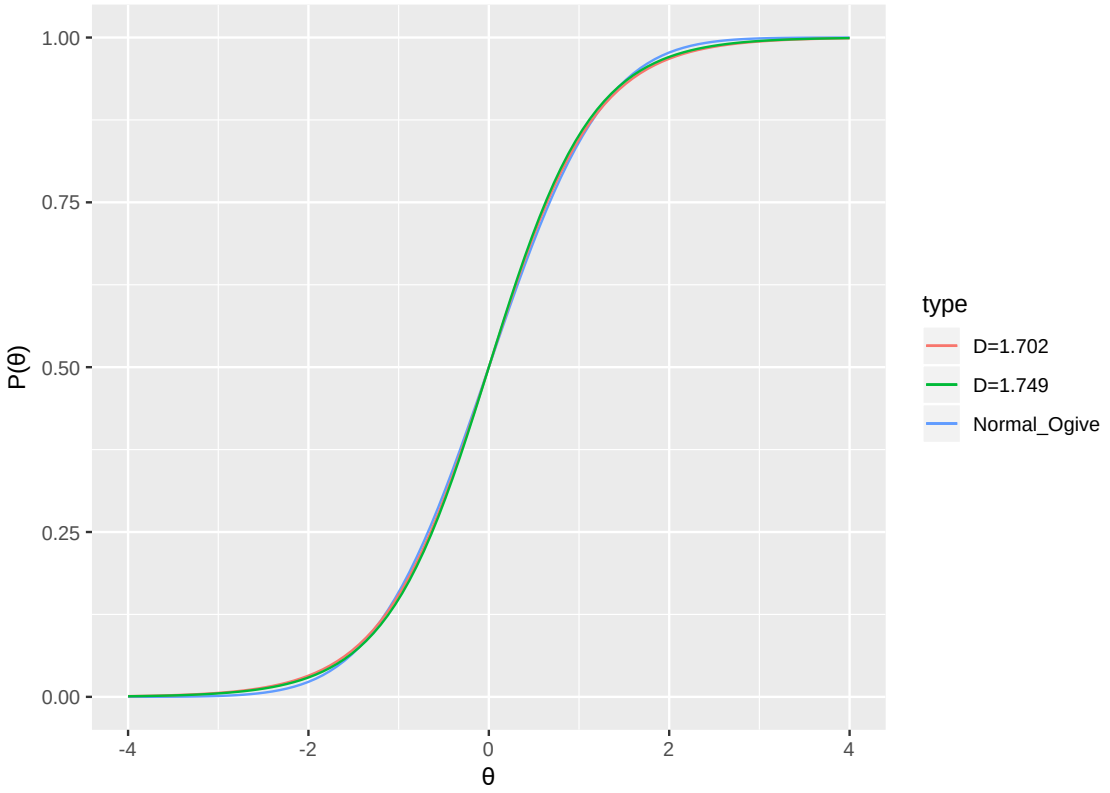

図 1.2 ふたつの尺度定数による正規累積分布関数の近似

次に，式 (1.24) の表記をより簡単にする。具体的には

$$a_j = \frac{\rho_j}{\sqrt{1-\rho_j^2}}, \tag{1.26}$$

$$b_j = \frac{\gamma_j}{\rho_j}, \tag{1.27}$$

とおくと式 (1.25) は,

$$P_j(\theta) = \frac{1}{1 + \exp\left(-Da_j(\theta - b_j)\right)}, \tag{1.28}$$

と置き換えることができる。これが現在一般的に用いられる 2 パラメタ・ロジスティックモデル (2-Parameter Logistic Model, 2PLM) の項目特性曲線 (Item Characteristic Curve, ICC) である (図

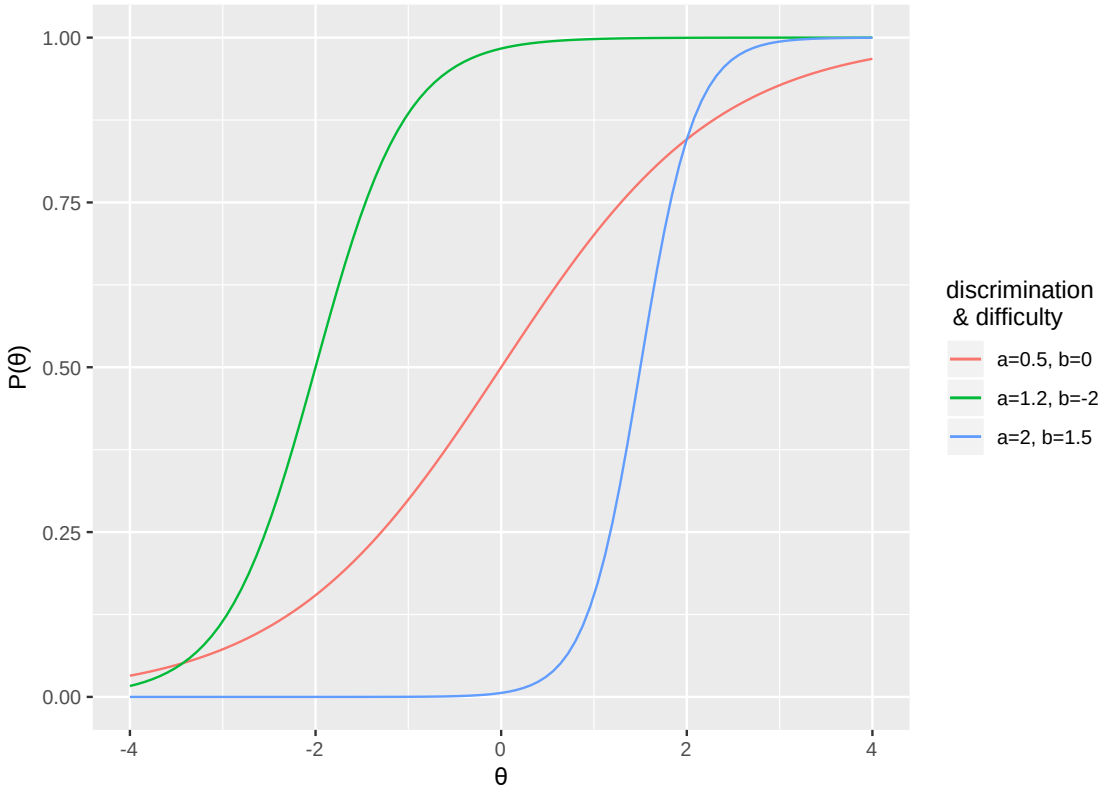

図 1.3 2PLM の項目特性曲線

1.3)。ただし，$D = 1.702$としている。ICC は項目反応確率のθについての関数とみなすこともできるため，項目反応関数 (Item Response Function, IRF) と呼ばれることもある。式中のaは識別力パラメタ (discrimination parameter)，bは困難度パラメタ (difficulty parameter) と呼ばれる。式 (1.28) ではDとaが単純な積であるため，教科書や一部研究で使われるモデルではDが省略されることがある。たとえ省略しても尺度が変わるだけで推定には支障をきたさないが，尺度調整 (calibration) や等化 (equating) の際には揃えておく必要があり，注意すべき点である。また式 (1.26) と (1.27) は簡易推定法 (heuristic method) と呼ばれ，実際に推定値として用いられはしないものの，数値計算の初期値として用いられる。その場合ρ_jを点双列相関係数でおきかえ，γ_jを，

$$\gamma_j = N^{-1}\left(1 - p_j \middle| \mu_j, \sigma_j^2\right), \tag{1.29}$$

のように項目通過率の値を利用した正規分布の逆関数から計算した値で置き換えればよい。

2PLM における識別力パラメタaはロジスティック曲線のちょうど縦軸が 0.5 にあたる部分の接線の傾きを決定する関数である。このパラメタをテスト全体で 1 とする，つまり母数自体を固定したモデルを 1 パラメタ・ロジスティックモデル (1-Parameter Logistic Model, 1PLM) と呼び，1 ではなく任意の 0 より大きい実数で固定したモデルを 1.5 パラメタ・ロジスティックモデル (1.5-parameter logistic model) などと呼ぶ。$D = 1.702$で固定した 1PLM の ICC は以下の通りである（図 1.4）。

$$P_j(\theta) = \frac{1}{1 + \exp\left(-Da\left(\theta - b_j\right)\right)}, \tag{1.30}$$

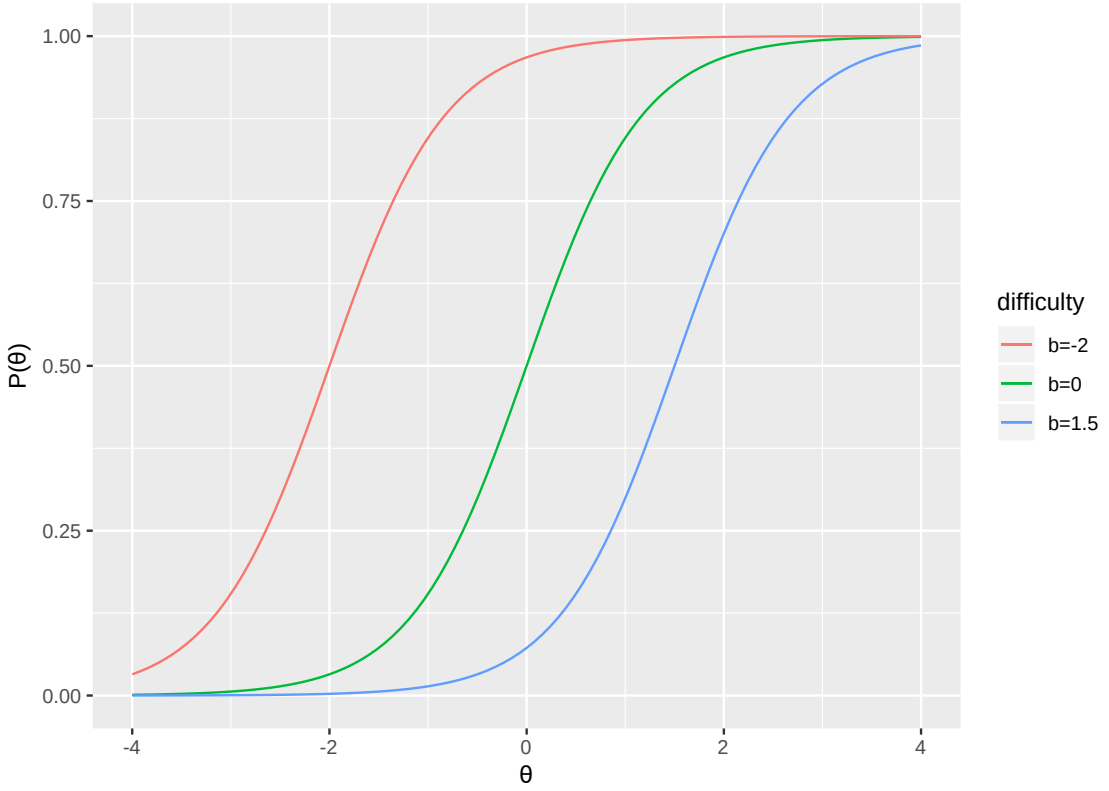

図 1.4 1PLM の項目特性曲線

数理的に 1PLM は，全く別の歴史的背景で開発された Rasch モデルと同じであるため，Rasch モデルと呼ばれることもある。しかし，1PLM と Rasch モデルは区別された議論されることがある。前者は 2PLM など他のモデルと比較して現実のデータセットに最もふさわしいモデルを選択するという観点から使用する立場であるのに対し，Rasch モデルは，モデルありきで，モデルに適合するように現実のデータをサンプリングしてくるべきという立場をとる (靜, 2007)。しかし実用上はそこまで厳密に線引きをしている訳ではなく，基本的にはどちらも IRT のひとつのモデルとして同一に扱われていることも多い。

ここで一度パラメタの解釈について整理しておく。困難度パラメタbはロジスティック曲線における確率 0.5 の位置を左右に調整する母数である。そのため位置母数 (location parameter) とも呼ばれる。困難度と受検者のθが等しいとき，指数関数の内部は 0 となり反応確率はちょうど 0.5 になる。大きい値ほどその項目に正答する確率が 50 パーセントになるために必要なθの値が大きいことを意味している。識別力パラメタaは先ほど説明したように$b = \theta$となる点における接線の傾きであるため，傾き母数 (slope parameter) とも呼ばれる。この母数が大きいほど正答確率がより狭い区間で大きく上昇するため，項目が正答に必要な能力を有しているかどうかをどれだけはっきりと識別できるかを表しいると解釈できる。

2PLM のさらに発展的なモデルとして 3 パラメタ・ロジスティックモデル (3-Parameter Logistic Model, 3PLM) も存在する。3PLM では下方漸近パラメタ (lower asymptote parameter) と呼ばれる母数により，受検者の項目に対する当て推量 (guessing) をモデリングしている。3PLM の ICC は，

$$P_j(\theta) = c_j + (1 - c_j)\frac{1}{1 + \exp\left(-Da_j(\theta - b_j)\right)}, \tag{1.31}$$

であり（図 1.5），当て推量パラメタは 0 から 1 の範囲の値をとる。テストによっては全項目で当

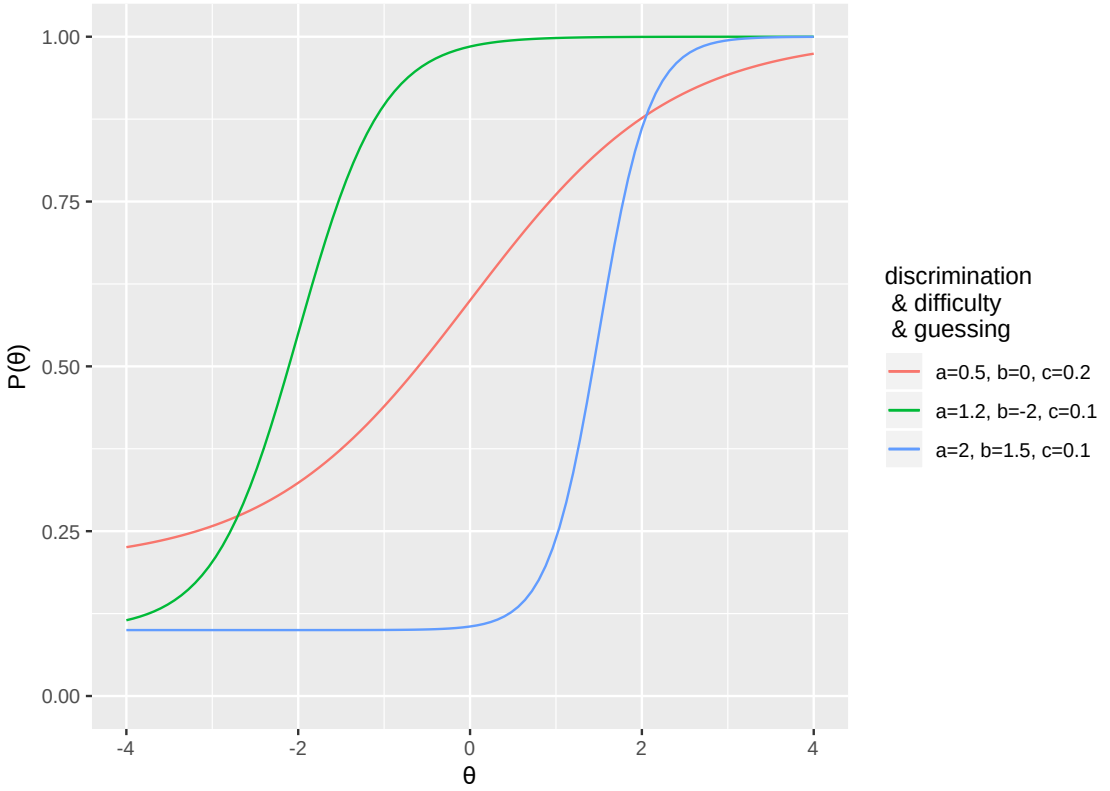

図 1.5 3PLM の項目特性曲線

て推量パラメタを固定することもある。3PLM は多肢選択式などで能力の低い受検者であっても適当に選択肢を選ぶことで一定確率正答してしまう状況をモデリングしたモデルであると解釈される場合がある (豊田, 2012)。しかし実際のところ当て推量パラメタはθ全域にわたって正答確率を底上げするパラメタであり，必ずしもそのような状況を正確にモデリングできているとは言いがたい。加藤・山田・川端 (2014) はこのモデルについて「項目の正答確率からcの分を取り去った残りである$1-c$に対して，2PLM の ICC を当てはめているのが 3PLM」であると述べている。つまり，3PLM の識別力と困難度は当て推量パラメタの分だけ圧縮されているため，2PLM などと同様の解釈をすることはできない。

この他に 4 パラメタ，5 パラメタのロジスティックモデルが存在し，それぞれ上方漸近線，対称性に関する母数を式 (1.31) に追加することで得られるが，実際のテストではこれまでに紹介した 1〜3 パラメタのモデルを使用することがほとんどである。

ここまでのモデルでは受検者が正答する確率のみをモデリングしてきたが，誤答する場合は確率であるため単純に，

$$Q_j(\theta) = 1 - P_j(\theta), \tag{1.32}$$

とすればよい。

1.3.2 多値型モデル

二値型の IRT モデルでは受検者の項目反応に正答か誤答という 2 種類の反応のみを考えていたが，実際の学力テストでは部分点を設けることもある。二値以上の反応データを扱うモデルを多値型 IRT モデルと呼び，そのなかでも学力テストの部分点を扱えるように (Muraki, 1992) が 2PLM を拡張したものを一般化部分採点モデル (Generalized Partial Credit Model, GPCM) と呼ぶ。部分点をK (1,2, … , k, … , K) 個のカテゴリとみなし，項目jにおけるカテゴリk番目の反応確率は隣接する$k-1$番目の反応カテゴリとともに考えられる。すなわちカテゴリkあるいは$k-1$のいずれかに反応するときに，kの反応を得る確率を考え，それ以外のカテゴリは考慮しないものとする。この状況でカテゴリkとなる遷移確率 (transition probability) に 2PLM を当てはめると仮定し，項目jのカテゴリkに反応する確率を$P_{jk}(\theta)$とすると，

$$C_{jk} = \frac{P_{jk}(\theta)}{P_{j,k-1}(\theta) + P_{jk}(\theta)} = \frac{\exp\left(Da_j\left(\theta - b_{jk}\right)\right)}{1 + \exp\left(Da_j\left(\theta - b_{jk}\right)\right)}, \tag{1.33}$$

と表せる。簡単のため$P_{jk}(\theta)$のθを省略し，式 (1.33) を整理すると，

$$P_{jk}\left(1 + \exp\left(Da_j\left(\theta - b_{jk}\right)\right)\right) = \exp\left(Da_j\left(\theta - b_{jk}\right)\right)\left(P_{j,k-1} + P_{jk}\right), \tag{1.34}$$

さらに展開して，

$$P_{jk} + P_{jk}\exp\left(Da_j(\theta - b_{jk})\right) = P_{j,k-1}\exp\left(Da_j(\theta - b_{jk})\right) + P_{jk}\exp\left(Da_j(\theta - b_{jk})\right), \quad (1.35)$$

となり，さらに両辺の第二項は等しいため消去して，

$$P_{jk} = P_{j,k-1}\exp\left(Da_j(\theta - b_{jk})\right), \quad (1.36)$$

となる。このとき，パラメタb_{jk}は左右のどちらの確率が大きくなるかに関するパラメタであり，遷移点 (transition point) と呼ばれる。式 (1.36) を整理すると，2 つのカテゴリに反応する確率のオッズは，

$$\frac{P_{jk}(\theta)}{P_{j,k-1}(\theta)} = \exp\left(Da_j(\theta - b_{jk})\right), \quad (1.37)$$

となる。カテゴリ 1 からKまでのこのオッズの積をとると，隣接するカテゴリのオッズの分子と分母で打ち消し合うため，最終的に

$$\prod_{k=1}^{K}\frac{P_{jk}(\theta)}{P_{j,k-1}(\theta)} = \frac{P_{jK}(\theta)}{P_{j0}(\theta)} = \exp\sum_{k=1}^{K} Da_j(\theta - b_{jk}), \quad (1.38)$$

が得られる。

カテゴリ 0 への反応確率は 1 から 0 以外のすべての確率の和を引いたものとして計算されるため，ある定数Gを用いて，

$$P_{j0}(\theta) = \frac{1}{G}, \quad (1.39)$$

とすると，カテゴリkへの反応確率は，

$$P_{jk}(\theta) = \frac{\exp\left(\sum_{v=1}^{k} Da_j(\theta - b_{jv})\right)}{G}, \quad (1.40)$$

と表現できる。ただし，表現の都合上指数関数の内部の添え字はkではなくvに変更している。

次に，すべてのカテゴリへの反応確率の和は 1 である，

$$P_{j0} + P_{j1} +, \dots, + P_{JK} = 1, \quad (1.41)$$

という制約をおくと，

$$\frac{1+\sum_{k=1}^{K}\exp\left(\sum_{v=1}^{k}Da_j(\theta-b_{jv})\right)}{G}=1, \tag{1.42}$$

となり，

$$G=1+\sum_{k=1}^{K}\exp\left(\sum_{v=1}^{k}Da_j(\theta-b_{jv})\right), \tag{1.43}$$

が得られる。したがってカテゴリkへの反応確率を，

$$P_{jk}(\theta)=\frac{\exp\left(\sum_{v=0}^{k}Da_j(\theta-b_{jv})\right)}{\sum_{k=0}^{K}\exp\left(\sum_{v=0}^{k}Da_j(\theta-b_{jv})\right)}, \tag{1.44}$$

と表現することができる。ただし，このとき$\theta-b_{j0}$は常に 0 である。この表現は遷移点を定義していないカテゴリ 0 をシグマに内包しているが，別の表現では

$$P_{jk}(\theta)=\frac{1+\exp\left(\sum_{v=1}^{k}Da_j(\theta-b_{jv})\right)}{1+\sum_{k=1}^{K}\exp\left(\sum_{v=1}^{k}Da_j(\theta-b_{jv})\right)}, \tag{1.45}$$

とも書ける。式 (1.44) や (1.45) を項目カテゴリ反応関数 (Item Category Response Function, ICRF) と呼ぶ。a_jは勾配パラメタであり，これを 1 に固定したものが部分採点モデル (Partial Credit Model, PCM) である。

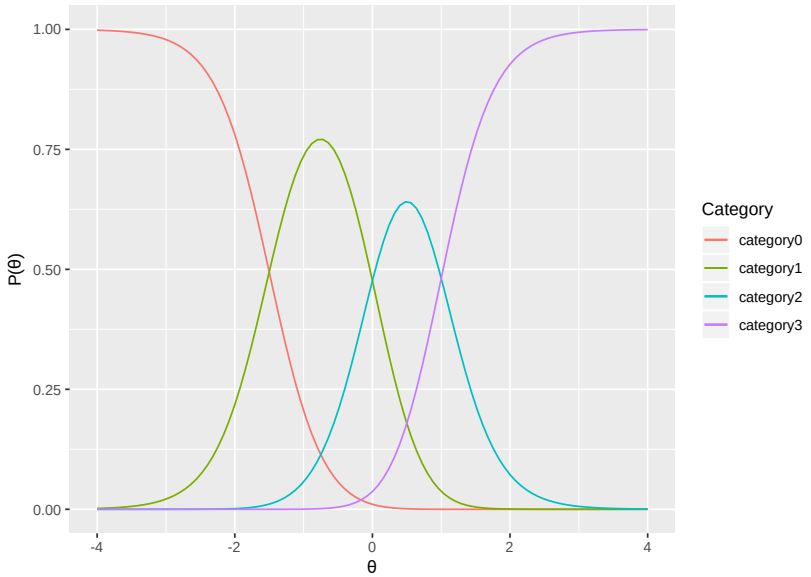

図 1.6 GPCM の項目特性曲線

GPCM のカテゴリ反応曲線を図 1.6 に示す。このグラフは勾配パラメタ$a_{\mathrm{j}} = 1.5$，項目カテゴリパラメタ$b_j = \{-1.5, 0, 1\}$としたカテゴリ反応曲線である。隣接するカテゴリの反応曲線との交点はカテゴリパラメタの値と一致していることが分かる。

1.3.3 多次元項目反応モデル

これまで二値型と多値型の IRT モデルを説明してきたが，どちらも単一の構成概念を扱う心理計量モデルであった。しかし，たとえば複数の因子を想定する因子分析モデルが存在するように，IRT でも複数の構成概念測定するモデルは存在する。これを多次元項目反応モデル (Multidimensional Item Response Theory model, MIRT model) と呼ぶ。Bock & Aitkin (1981) は 2 因子のj変量直交因子モデル,

$$y_j = \alpha_{j1}\theta_1 + \alpha_{j2}\theta_2 + \varepsilon_j, \tag{1.46}$$

ただし,

$$y_j \sim N(0,1), \tag{1.47}$$

$$\begin{bmatrix}\theta_1 \\ \theta_2\end{bmatrix} \sim \boldsymbol{N}\left(\begin{bmatrix}0 \\ 0\end{bmatrix}, \begin{bmatrix}1 & 0 \\ 0 & 1\end{bmatrix}\right), \tag{1.48}$$

$$\varepsilon_j \sim N\left(0, 1 - \alpha_{j1}^2 - \alpha_{j2}^2\right), \tag{1.49}$$

を利用して二次元 IRT モデルへの拡張を提案した。ここで一次元の正規累積モデルの場合と同様に閾値母数γ_jよりもy_jが大きければ受検者は 1 と反応し，そうでなければ 0 と反応すると仮定する。すると二次元正規累積モデルは,

$$P\left(\mathrm{u_j} = 1 \middle| \theta_1, \theta_2\right) = \int_{-z_j(\theta)}^{\infty} \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{1}), \tag{1.50}$$

ただし,

$$-z_j(\theta) = \frac{\gamma_j - \alpha_{j1}\theta_1 + \alpha_{j2}\theta_2}{\sigma_j}, \tag{1.51}$$

$$\sigma_j = \sqrt{1 - \alpha_{j1}^2 - \alpha_{j2}^2}, \tag{1.52}$$

と表現できる。ただし，このままでは因子分析モデルのパラメトリゼーションであるため,

$$d_j = -\frac{\gamma_j}{\sigma_j}, \tag{1.53}$$

$$a_{j1} = \frac{\alpha_{j1}}{\sigma_j}, \tag{1.54}$$

$$a_{j2} = \frac{\alpha_{j2}}{\sigma_j}, \tag{1.55}$$

と変換する。Bock & Aitkin (1981) の示した推定方法は Bock, Gibbons & Muraki (1988) によってm次元に一般化された。彼らはこの方法を完全情報項目因子分析 (Full-Information Item Factor Analysis, FIFA) と呼んだ。

さらにこのモデルのロジスティックシグモイド型は,

$$P\left(\mathrm{u}=1\middle|\mathrm{z_j}(\boldsymbol{\theta})\right) = \frac{\exp\left(z_j(\boldsymbol{\theta})\right)}{1+\exp\left(z_j(\boldsymbol{\theta})\right)}, \tag{1.56}$$

と表すことができる(McKinley and Reckase, 1982)。このモデルにおける識別力は各次元の識別力パラメタの平方和の正の平方根として与えられる。

$$\eta_{\mathrm{j}} = \sqrt{\boldsymbol{a}'\boldsymbol{a}}, \tag{1.57}$$

同様に多次元困難度も,

$$\beta_j = -\frac{d_j}{\eta_j}, \tag{1.58}$$

と表される。

MIRT モデルには一次元 IRT モデル以上に様々なモデルが存在する。例えば複数の能力のうち，どれかひとつでも閾値に達していれば正答できると考えるモデル（補償型モデル, compensatory model）とすべての次元で能力が閾値に達している必要があるとするモデル（非補償型モデル, non-compensatory model）がある。より詳細な議論は Reckase (2009)などを参照されたい。

1.3.4 一般項目反応モデル

Lord (1980)の正規累積モデルでは，受検者の能力値θを所与としたときに項目に対しての能力値である$\mathrm{Y_j}'$は正規分布が仮定されていた (図 1.1)。ここで中村・豊田 (1991) にしたがって，項目反応理論ではなく Thurstone (1927) の比較判断の法則 (law of comparative judgement) ならびに Torgerson (1958) のカテゴリ判断の法則 (law of categorical judgement) の観点から受検者の項目に対する反応を考える。

比較判断の法則は n 個の異なる刺激$S_1, S_2, \ldots, S_{\mathrm{n}}$が与えられたときに，例えば$S_1$と$S_2$という刺激の大小比較についての情報$X_1$と$X_2$を基に一次元の心理尺度を構成するための法則である。比較判断の法則は刺激と刺激の比較のみを扱うが，これを質問項目と刺激，つまり項目に対する反応（0, 1,…）の境界値（閾値）と刺激の関係に拡張したものがカテゴリ判断の法則である。比較判

断の法則やカテゴリ判断の法則を用いて尺度を構成する場合，必要となるデータ行列は項目数を m とすると m 行 m 列の正方行列である。テストデータの場合，受検者の項目に対する反応というデータを扱うため，受検者と項目のデータをどちらも刺激とみなし，受検者数を N とすれば，N+m 行 N+m 列のデータ行列となる。当然受検者と受検者の比較，あるいは項目と項目の比較のブロックは欠測である。

項目反応理論とカテゴリ判断の法則にはいくつかの共通点がある。それは受検者θの特性値と項目の特性値bが同一の尺度上に位置づけられることや，あるカテゴリの反応を得るためには受検者の能力が項目の閾値を超えなくてはならないという仮定に表れている。ただし，項目反応理論の正規累積モデルではθは各受検者につき唯一の値をとるものと仮定されていたのに対し，比較判断の法則の観点ではθにもばらつきϕ^2を仮定する点が異なる。

確かに既存の 2PLM では受検者の項目に対する反応確率はパラメタb_jとa_j，θで決定されるが，このとき式 (1.21) で表される条件付き分布の標準偏差には項目に関する情報としての式（1.23）のみが与えられており，受検者の能力の分散は考慮されていない。Torgerson は受検者の特性値についても確率的な分布を導入し，これにより正規累積モデルをより一般化した。さらにこれをロジスティック関数で近似した孫・芝 (1990) によるモデルは，

$$P_j(\mathrm{u}|\theta,\alpha,a_j,b_j)=\frac{1}{1+\exp\left(-D\dfrac{\alpha a_j}{\sqrt{\alpha^2+a_j^2}}(\theta-b_j)\right)}, \tag{1.59}$$

である。ただし，受検者の能力分布の誤差分散がϕ^2であるのに対して，

$$\alpha=\frac{1}{\phi}, \tag{1.60}$$

である。中村・前川 (1993) はαではなくϕを用いて，

$$P_j(\mathrm{u}|\theta,\alpha,a_j,b_j)=\frac{1}{1+\exp\left(-D\dfrac{a_j}{\sqrt{1+\phi^2 a_j^2}}(\theta-b_j)\right)}, \tag{1.61}$$

と表現している。式 (1.61) においていくつかϕの値を変化させた ICC を図 1.7 に示す。

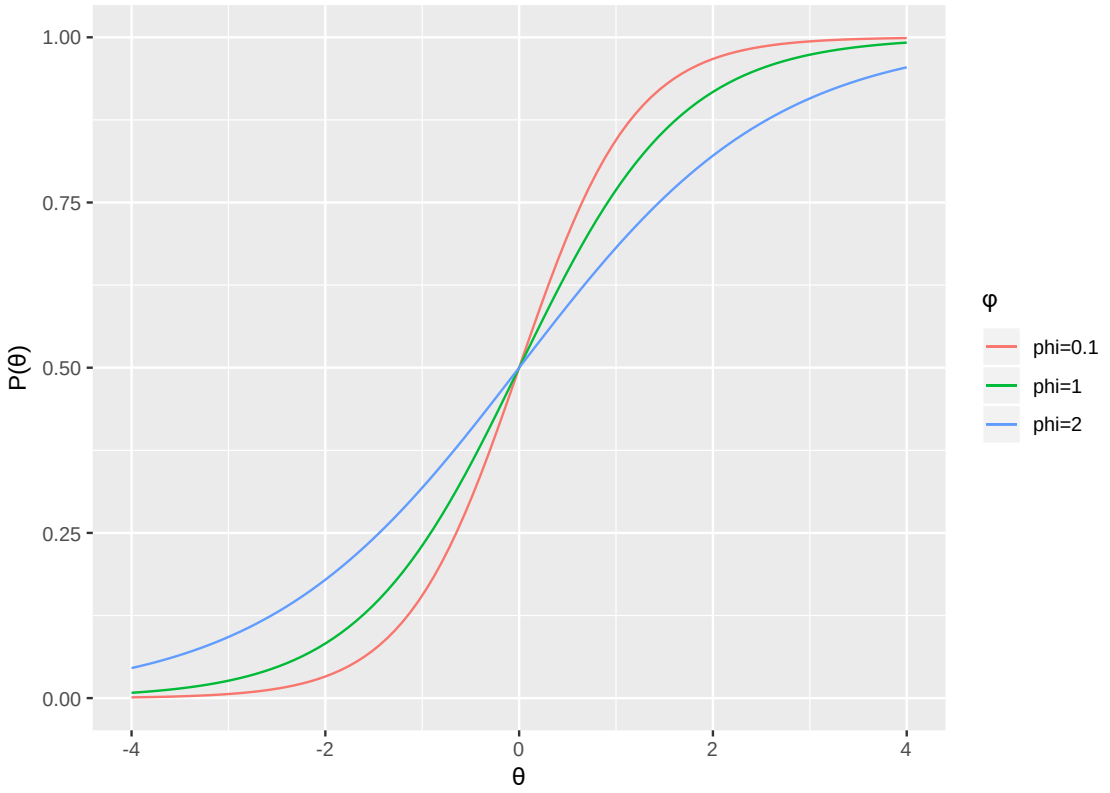

図 1.7 GIRT モデルの項目特性曲線

これらを一般項目反応モデル (General Item Response Model, GIRT) モデルと呼ぶ。

このモデルの優れている点は，テストで測りたい能力が一次元であるのに対し，実際は多次元的な能力が回答に必要とされるような状況において，MIRT モデルのように次元ごとに区別するのではなく，(1) 第一の次元以外の攪乱能力次元の情報をつぶして，各受検者の能力分布として扱うことができ (孫, 1997)，(2) 項目と受検者のパラメタに尺度の不定性があるため，従来の一次元 IRT モデルと同様の等化手法を用いることができる点である (柴山・繁桝, 1994)。

1.4 IRT の仮定

IRT は受検者の項目に対する反応確率をモデリングしている。CTT に比べてひとつのテストから項目と受検者に関する情報をより多く抽出することができるが，その情報もいくつかの仮定を前提として得られていることに注意しなければならない。ここでは IRT において重要である潜在変数の次元に関してと，局所独立性について説明する。

1.4.1 測定の一次元性

一般的なテストは一種類のスコアを返すことが多い。中には下位尺度を設けて，例えば TOFLE のように英語の技能ごとの得点を出すようなテストもあるが，それでもトータルのスコアが提示される。このようなテストではテストの測りたい構成概念がひとつに決定されている。このよ

うな前提を測定の一次元性 (unidimensionality) と呼び, 一次元 IRT モデルで最も重要な仮定のひとつである。つまり一次元 IRT モデルで分析をおこなう場合は，まずはこの前提が成り立っているかを確認しなくてはならない。

一次元性を確認する手立てはいくつかある。例えば反応データの相関行列を計算し，そのデータ行列の固有値 (eigenvalues) の減衰状況を確認する方法である。通常の探索的因子分析において仮説的な因子の数を決定する方法として最もポピュラーな方法がこれである。正確にはこの固有値が 1 以上の固有値の数を因子数とするガットマン基準や，固有値を大きいものから順にプロットしていき，勾配がなだらかになる直前までの固有値の数を因子数とするスクリープロットと呼ばれる手法の基本となっているのが，この固有値計算である。

しかし二値型のデータの場合, 相関係数が項目通過率に依存する。そのため本来の相関係数よりも低く推定される可能性がある。これを証明するためにはまず，ピアソンの積率相関係数の定義式,

$$r_{jl} = \frac{\sigma_{jl}}{\sigma_j \sigma_l}, \tag{1.62}$$

から出発する。式 (1.62) においてi，jは項目についての添え字であり，σは分散であるとする。0 か 1 しかとらないデータ$\mathrm{u_{ij}}$の場合，分散は式 (1.17) の項目通過率p_jを用いて,

$$\sigma_j^2 = \frac{1}{N}\sum_{i=1}^{N}\left(\mathrm{u}_{ij} - p_j\right)^2, \tag{1.63}$$

となるところ，分散の定義式におけるシグマ記号内を展開して整理すると,

$$\sigma_j^2 = \frac{1}{N}\sum_{i=1}^{N}{\mathrm{u}_{ij}}^2 - p_j^2 = p_j - p_j^2 = \left(1 - p_j\right)p_j, \tag{1.64}$$

が得られる。共分散の場合も同様にして,

$$\sigma_{jl}^2 = p_{\mathrm{jl}} - p_j p_l, \tag{1.65}$$

である。ただし，p_{jl}は両方の項目に正答した受検者の割合である。これらを用いて,

$$r_{jl} = \frac{p_{jl} - p_j p_l}{\sqrt{\left(1 - p_j\right)p_j(1 - p_l)p_l}}, \tag{1.66}$$

というように相関係数の式が変形できた。これをϕ係数と呼ぶ。

ところで式 (1.64) における最大値は通過率p_jが 0.5 のときに 0.25 であることは明らかである。これが分母に来ることから相関係数の大小は項目通過率に一部依存する。相関係数が項目通過率に依存するということは，この相関行列で因子分析をおこなった場合に通過率，すなわち困難度の因子を捉えてしまう可能性がある。

ちなみに，片方の項目に正答している受検者はもう片方の項目に必ず誤答している条件（$p_{jl} = 0$），かつ項目通過率が等しいとき（$\mathrm{p}_j = \mathrm{p}_l$）に$\phi$係数は-1 をとり，逆に$p_{jl} = 1$で通過率が等しい場合には$\phi$係数は 1 となる。一般的な連続量で相関係数が 1 や-1 をとることはほとんどないが，易しい項目同士や難しい項目同士のϕ係数を求める場合に極端な相関係数をとりやすい。

この問題を解消するために一般的に用いられるのが四分位相関係数 (tetrachoric correlation coefficient) である。1.3.1 での仮定と同様に，ある閾値を超えたら 1 を，下回ったら 0 という反応を得ることが想定できるモデルの場合，その背後にはY_j'のような連続量の潜在変数が想定できる。これが両方の項目において 2 変量正規分布をなしていると仮定すれば，この分布の相関母数を求める事で四分位相関係数を推定することができる (Olsson, 1979; 豊田, 1998)。しかしこの行列が非負定値である保証はないため固有値の計算に支障をきたす場合もある (柳井・前川・繁桝・市川, 1990)。

この他にもいくつかの一次元性を確認する手法が考案されており，(Hattie, 1985) や Stout, Nandakumar & Habing (1996) に詳しい。しかし実用上では四分位相関係数行列から固有値を求める方法でも問題はないだろう。

最後に一次元性の仮定が保たれない場合について考える。例えば明らかに複数の種類の能力を測定していると考えられるテストを，同時に一次元 IRT モデルで分析することは許容されない。その場合は MIRT モデルを使用するか，そもそも一次元 IRT モデルで分析することを諦める必要がある。しかし，現実的な場面ではテストに回答するために必要な能力は潜在的な要素であり，その数を特定することは非常に困難である。

1.4.2 局所独立性

一般的なテストでは 10〜50 個程度の項目が出題される。IRT において，ある特定のθの個人がJ項目のテストに回答したときに反応パタン$X = \{\mathrm{u}_1, \mathrm{u}_2, \ldots, \mathrm{u}_J\}$を得る確率は，

$$P(\mathrm{X}|\theta) = \prod_{j=1}^{J} P_j(\theta)^{\mathrm{u}_j} Q_j(\theta)^{(1-\mathrm{u}_j)}, \tag{1.67}$$

と考えられる。このとき一次元 IRT モデルであれば，正答確率に影響する要因は項目が固定されている場合，受検者の潜在的な能力値θのみである。したがってθを固定してしまえば項目間には相関は生じない。これが局所独立性 (local independence) と呼ばれる性質である。(Lord et al.,

2008) の定義に従えば「局所独立とは同じθによって特徴付けられた受検者のいかなる集団内においても, 項目得点の条件付き分布は互いにすべて独立である」ということである。加藤ら (2014, p.144) に則って具体的に説明すると，項目 1 に正答した受検者が項目 2 に正答する割合と，項目 1 に誤答した受検者が項目 2 に正答する割合が，誤差の範囲で等しくなる事を意味する。つまりϕ係数がほぼ 0 になる。より詳細な議論は (南風原, 2000) を参照されたい。

ところで式（1.67）では各項目間の反応確率を単純な積で表現しているが，このように扱うことができるのはθを固定したときに項目間の反応確率が独立になるからであり，IRT における反応確率の計算の基本的な根拠となっているのが，この局所独立性の仮定であることがわかる。

局所独立の仮定が侵される典型的な，例は項目間の依存関係と測定対象外の攪乱因子の存在である。項目間の依存関係とは，例えば大問形式の項目の場合途中の項目への正答，誤答が，後半の項目正答に影響する。測定対象外の攪乱因子というのは，例えば数学の割合を計算する項目で，野球選手の打率を計算するような設問にした場合，本来測定した割合を計算する能力のほかに，野球に詳しいという要因が作用する可能性があるということである。そのほかに測定すべき能力の多次元性や疲労，テストの受験環境など様々な要因が考えられる (Yen, 1993)。

局所独立が侵されている状況を局所依存 (local dependence) ，あるいは局所項目依存 (Local Item Dependence, LID) と呼ぶが，この状況が IRT モデルの分析にどのような影響を与えるのかについてはいくつか研究 (Yen, 1993; 登藤, 2012) がおこなわれているが，推定値にバイアスが入るため局所項目依存はなるべく影響を取り除く方がよい。

LID を診断するための指標はいくつか提案されている。後述する項目適合度として用いられるχ^2統計量やG^2統計量を用いることもあるが，ここでは残差得点の相関係数であるQ_3統計量について説明する。まず，残差得点とは実際の項目反応からモデル上の反応確率を引いた，

$$\boldsymbol{d_j} = \mathbf{u_j} - P_j(\boldsymbol{\theta}), \tag{1.68}$$

と定義される。式 (1.68) 中の表現は全受検者のベクトルである。これの項目間の相関係数がQ_3 統計量,

$$Q_{jl}^3 = r(\boldsymbol{d_j}, \boldsymbol{d_l}), \tag{1.69}$$

である。残差得点自体はモデルフィットの指標としても使われるが, 簡単に言えばモデルとデータの乖離具合を表している。局所独立の仮定が守られているのであれば, 項目間の残差得点の相関は 0 に近くなるはずである。このQ_3統計量，χ^2統計量，G^2統計量のほか，局所依存を認めた IRT モデルを立てて Rao のスコア統計量やラグランジュ乗数統計量を計算する方法もある。これらの指標やそれぞれの比較にについては Chen & Thissen (1997) や Liu & Maydeu-Olivares (2012) に詳しい。

1.5　IRT における測定精度

古典的テスト理論におけるテストの測定精度の概念が信頼性であり，その推定値のひとつがクロンバックの α 係数とそれに基づく測定の標準誤差であった。しかし，これらの指標はテスト単位でしか求める事ができず，CTT の範疇で推定されるため，受検者集団に依存する指標であった。一方，IRT モデルはモデル上で受検者と項目のパラメタを分離しているため，このパラメタに基づいて項目ごとに，そして標本依存性を克服した測定精度の指標を与えてくれる。

IRT における測定精度は推定値（例えば，後述する最尤推定値）のフィッシャー情報量 (Fisher information) の逆数として与えられる。フィッシャー情報量は測定精度を知りたいパラメタの関数として与えられ，その逆数は不偏推定量 $\hat{\theta}$の分散の下限を与えることがクラメール・ラオの不等式によって知られている (芝・渡部・石塚, 1984)。θの最尤推定量は項目数が多いときに近似的に不偏推定量となるため (南風原, 1991)，フィッシャー情報量の逆数はθの推定量の分散について評価しており，すなわちこれを測定精度とみなしている。この情報量を適当なθの区間における関数，つまり情報関数とみなすこともでき，項目について求めた情報関数を項目情報関数 (Item Information Function, IIF) と呼ぶ。同様にテストの測定精度はテスト情報関数 (Test Information Function, TIF) と呼ばれ，局所独立性の仮定が満たされている場合 IIF には加法性があるため，単にテストの全項目の IIF の和をとれば良い。

フィッシャー情報量は一般的に二階偏微分可能な関数においては，「対数尤度関数の二階偏微分の負の期待値」である。ここでの対数尤度 (log likelihood) 関数とは IRT における項目反応パタンと項目パラメタが所与のときのθについての関数であり，具体的には項目反応ベクトル$\mathbf{u}$と項目パラメタベクトル$\boldsymbol{\delta}$が与えられている場合の尤度関数，

$$L(\theta|\mathbf{u},\boldsymbol{\delta}) = \prod_{j=1}^{J} P_j(\theta)^{\mathrm{u}_j}\ Q_j(\theta)^{(1-\mathrm{u}_j)}, \tag{1.70}$$

の対数をとった，

$$\ln L(\theta|\mathbf{u},\boldsymbol{\delta}) = \sum_{j=1}^{J} \mathrm{u}_j P_j(\theta) + \big(1-\mathrm{u}_j\big) Q_j(\theta)\,, \tag{1.71}$$

である。

なお，尤度とはデータから推定される母数についての尤もらしさの値である。数式としてはすべてのパラメタが所与のときの項目反応パタン$\mathbf{u}$を得る確率密度関数と同じであるが，頻度主義において母数は確率変数ではなく，また尤度関数を適当な区間で積分しても 1 にはならず確率としての条件を満たさないため，これを尤度関数と呼んでいる。

IIF はθについての ICC のθについての偏微分を用いて，

$$I_j(\theta) = \frac{\left\{\frac{\partial P_j(\theta)}{\partial \theta}\right\}^2}{P_j(\theta)Q_j(\theta)}, \tag{1.72}$$

と表される，特に 3PLM の場合は，

$$I_j(\theta) = \frac{\left\{\frac{D^2 a_j^2 Q_j(\theta)\left(P_j(\theta) - c_j\right)}{\left(1 - c_j\right)}\right\}^2}{\left(1 - c_j\right)^2 P_j(\theta)}, \tag{1.73}$$

である。2PLM の場合はcパラメタを 0 とすればよく，1PLM の場合はさらにaパラメタを 1 とすれば求める事ができる。テスト情報関数は繰り返しになるが IIF の和であるので，式(1.73)より，

$$I_X(\theta) = \sum_{j=1}^{J} I_j(\theta), \tag{1.74}$$

と求めることができる。TIF は二値型の一次元 IRT モデルにおいては単峰型の曲線を描くが，情報関数が大きいほど，その位置のθの値に関して分散が小さいことを意味する。1PLM と 2PLM の場合，IIF の最大点は，$P_j(\theta)Q_j(\theta)$，つまり二項確率の分散が最大となる点と等しいため，正答確率が 0.5 となる項目困難度の値と等しくなる。そして 2PLM の場合，同じθであれば識別力a_jの高い項目ほど情報量は多くなる。すなわち識別力の高さが測定精度を左右するのである。

IIF や TIF は項目やテストがどの能力の受検者層に適しているのかを決定する上で役に立つ。またひとつのテストをふたつの等質なテストに分割したい場合などにも，TIF の形状がおおむね一致することを方針として組み立てるものひとつの方法である。

1.6 適合度

IRT モデルを利用する上で，適用したいテストデータが使用する IRT モデルに適合しているかどうかを評価することは重要である。不適合なモデル（あるいはデータ）を使用して分析をおこなうことは，パラメタの推定を不安定にし，測定の精度を悪化させる一因となる。ここでは項目ごとに適合度を評価する方法を説明する。

1.6.1 項目適合度

ひとつの IRT モデルの適合度を確認する手法は 3 種類に大別される。個人適合度，項目適合度，全体適合度である。本稿の実験でも，垂直尺度化された項目がデータに適合しているかどうかを確かめるため，項目適合度を利用している。個人適合度は受検者一人あたりの反応パタンが

どれだけ IRT モデルにフィットしているかを評価できる。全体適合度はテストで観測されたすべての反応パタンとその受検者度数から計算することができる。全体適合度については本稿では扱わないが，Bock & Aitkin (1981)のG^2統計量が有名である。

（1）受検者を適当な下位集団に分割して推定する方法

項目適合度の考え方は非常にシンプルであり，実際のデータの正答率とモデルから予測された正答率のズレを評価する指標である。代表的な項目適合度としてはχ^2統計量の考え方にもとづいたX^2統計量と，尤度比の考え方にもとづいたG^2統計量がある。X^2統計量の一種として Yen (1981)のQ_1統計量とX^2統計量が，G^2統計量の一種として McKinley & Mills (1985) のものが代表的である。それぞれ一般形は，

$$X_j^2 = \sum_{k=1}^{m} \frac{N_k\left(O_{jk} - E_{jk}\right)^2}{E_{jk}\left(1 - E_{jk}\right)}, \tag{1.75}$$

$$G_j^2 = 2\sum_{k=1}^{m} N_k \left[O_{jk} \ln\left(\frac{O_{jk}}{E_{jk}}\right) + \left(1 - O_{jk}\right) \ln\left(\frac{1 - O_{jk}}{1 - E_{jk}}\right)\right], \tag{1.76}$$

である。ただし，kは下位集団の分割点（ただし$k = \{1,2,\ldots,m\}$），jは項目数，Nは下位集団内の受検者数，Oは観測値にもとづく正答率，Eはモデルによる正答率を表す。Q_1とG^2においてEは下位集団に含まれる受検者のθを ICC に代入して得られる正答確率の平均と定義され，Bock (1972)のX_j^2では中央値で定義される。なお，θの計算方法には特に決まりはなく，後述する最尤推定やベイズ推定法を用いれば良い。

Yen (1981) のQ_1と Mckinley & Mills (1985) の統計量の下位集団の数は 10 個であり，各下位集団の中に含まれる人数ができるかぎり等しくなるように分割される。Bock (1972) のX_j^2統計量は，Q_1統計量とは異なり任意の下位集団の分割点を設定し，正答率の代表値として中央値を用いている。どちらの統計量もモデルが真の時に自由度＝下位集団の分割点 (m) －モデルの項目パラメタ数 (1~3) のχ^2分布に近似的にしたがうことが分かっており，χ^2検定によりズレの大きな項目を検出することが可能である。

しかし，これらの統計量には 2 つの理論的な問題がある。この下位集団の分割点は，得られたデータ（サンプル）に強く依存してしまう。また，1PLM 以外の IRT モデルのθは正答数得点ではなく項目反応パタンによって値が変化するため，直接観測値にもとづく正答率とモデルによる正答率を比較することは厳密に妥当とは言えない。

（2）Lord & Wingersky (1984) の recursion formula をアルゴリズムに基づく方法

そこで Lord & Wingersky (1984) の recursion formula をアルゴリズムとして求めた復元得点分布 (IRT observed score distribution) にもとづいてモデルの正答率を算出し，観測値による正答率

と比較することが推奨される (Orlando & Thissen, 2000)。

ところで, IRT モデルにもとづく得点には真の得点 (true score) と復元得点 (observed score) の 2 種類がある。真の得点の求め方は, 各受検者の全項目についての正答確率の総和をとればよい。能力値θの受検者が取り得る真の得点xの分布関数を$f(\mathrm{x}|\theta)$とすると, まさに受検者全体の能力分布に等しい。真の得点が受検者一人に対してただひとつ求められる期待テスト得点であるのに対し, 復元得点は, すべての項目反応パタンにもとづいて能力値θ_i受検者が取り得る正答数得点の確率分布を計算し, 全受検者についてその和をとるものである。recursion formula のアルゴリズムと具体的な計算手順については Lord & Wingersky (1984) や柴山ら (2018) の第 2 節に詳しい。いま, 項目数nのテストにおけるθ_iの受検者がとりうる得点の確率分布は項目jに正答する確率をP_jとすると, 以下のように与えられる。

$$f_n(\mathrm{x}|\theta_i) = \begin{cases} f_{n-1}(\mathrm{x}|\theta_i)(1-P_n)\,, & (\mathrm{x}=0), \\ f_{n-1}(\mathrm{x}|\theta_i)(1-P_n) + f_{n-1}(\mathrm{x}-1|\theta_i)P_n\,, & (0<\mathrm{x}<n)\,, \\ f_{n-1}(\mathrm{x}|\theta_i)P_n\,, & (\mathrm{x}=n). \end{cases} \tag{1.77}$$

さらにこの分布関数をθについて周辺化することで,

$$f(\mathrm{x}|\theta) = \int_\theta f(\mathrm{x}|\theta)\psi(\theta)\,d\theta\,, \tag{1.78}$$

によって受検者全体の得点分布を求めることができる。ここで$\psi(\theta)$は受検者の能力分布を表しており, これには事前に最尤推定法 (MLE) やベイズ推定法 (EAP, MAP) などにより求められた能力パラメタベクトルや適当な確率分布などを用いれば良い。ただし実際の受検者数は有限であるため, 得点分布は,

$$f(\mathrm{x}|\theta) = \sum_{i=1}^{\mathrm{N}} f(\mathrm{x}|\theta_i)\psi(\theta_i)\,, \tag{1.79}$$

のように離散近似をして求めることになる。

式 (1.78) より, 正答数得点k'のグループに属し, かつ項目jに正答した受検者の割合がモデルによる正答率$E_{jk'}$であり, それは,

$$E_{jk'} = \frac{\int_\theta P_j\, f^{*j}(k'-1|\theta)\psi(\theta)\,d\theta}{\int_\theta f(k'|\theta)\psi(\theta)\,d\theta}, \tag{1.80}$$

で与えられる。ただし, $f^{*j}(k'-1|\theta)$は正答数得点k'の集団の中で項目jを除いたときの正答数得

点が$k'-1$点となる受検者の分布関数である。つまり式 (1.80) の分母は正答数得点k'の下位集団の度数を，分子はそのなかでも項目jに正答している受検者の度数を表している。ただし，この場合の度数は整数値とは限らない。

式 (1.80) を式 (1.75) と (1.76) に当てはめると，

$$S-Q_{1j}=\sum_{k'=1}^{n-1}\frac{N_k\left(O_{jk'}-E_{jk'}\right)^2}{E_{jk'}\left(1-E_{jk'}\right)}, \tag{1.81}$$

$$S-G_j^2=2\sum_{k'=1}^{n-1}N_k\left[O_{jk'}\ln\left(\frac{O_{jk'}}{E_{jk'}}\right)+\left(1-O_{jk}\right)\ln\left(\frac{1-O_{jk'}}{1-E_{jk'}}\right)\right], \tag{1.82}$$

が得られる。注意すべき点は，正答数得点が 0 の受検者集団は全問不正解しているため当然正答確率は 0 になり，逆に全問正解の受検者集団の正答率は 1 になるため，下位集団からは除外されている点である。そのためシグマ記号の範囲は$k'=\{1,2,\ldots,n-1\}$となる。

このような発想にもとづいて修正されたX_2統計量とG_2統計量はそれぞれ$S-X_2$，$S-G_2$と呼ばれ，Yen (1981) のQ_1統計量や McKinley & Mills (1985)のG_2統計量では 10 個に区切られていた下位集団は正答数得点ごとの集団に置き換えられる。

（3）EM アルゴリズムで計算される期待度数を用いる方法

式（1）（2）中のO_{jk}は項目パラメタ推定アルゴリズムとして広く用いられる EM アルゴリズムの E ステップに表れる受検者の期待度数N_{jm}と正答する受検者の期待度数r_{jm}（m はいずれも E ステップにおける分点に関する添え字）を，

$$O_{jm}=\frac{r_{jm}}{N_{jm}}, \tag{1.83}$$

のように用いることでも適合度を計算することができる。ただしこの方法を用いる場合にはグループの分割が分点数によって左右されてしまうことと，$\chi 2$ 統計量の計算はおこなわず，残差のみを数値的に確認するのみにとどまる(前川, 1991)。

（4）受検者ひとりごとで残差を計算する方法

このほかに，OUTFIT 指標 (Wright & Stone, 1979)，INFIT 指標 (Wright & Masters, 1982) と呼ばれる，下位集団に分割しないで項目の適合度を測る指標も存在する。式 (1.75) ではデータとモデルのズレ（残差）を二項確率の標準偏差の推定値で割ることで標準化した値（標準残差）を用いており，これを下位集団ごと計算し，和を取っていた。OUTFIT 指標は受検者一人ずつにこの標準残差を求めて，平均をとる。受検者ひとりの標準残差は，

$$z_{ij} = \frac{\mathrm{u}_{ij} - P_j(\theta_i)}{\sqrt{P_j(\theta_i)\left(1 - P_j(\theta_i)\right)}}, \tag{1.84}$$

で与えられる。ただしuは項目反応パタンであり，二値型モデルの場合は 0 か 1 をとる。項目jにおけるこの標準残差の平均平方，

$$v_j = \frac{\sum_{i=1}^{N} z_{ij}^2}{N-1}, \tag{1.85}$$

は自由度$N-1$の F 分布にしたがう (Wright & Stone, 1979)。これが OUTFIT 指標である。あるいは，

$$t_j = \left(\ln(v_j) + v_j - 1\right)\sqrt{\frac{N-1}{8}}, \tag{1.86}$$

というように t 分布や標準正規分布に近似的にしたがうように変換した，標準化された OUTFIT 指標もある。

INFIT 指標は OUTFIT 指標を何らかの情報で重み付けしたものであり，

$$v_j' = \frac{\sum_{i=1}^{N} z_{ij}^2 w_{ij}}{\sum_{i=1}^{N} w_{ij}}, \tag{1.87}$$

と与えられる。重みは正答確率の分散（二次のモーメント）であり，二値型のデータの場合，

$$w_{\mathrm{ij}} = P_j(\theta_i)\left(1 - P_j(\theta_i)\right), \tag{1.88}$$

と置くことができる。これを標準化するためには尖度（第 4 次のモーメント），

$$m_{ij} = P_j(\theta_i)\left(1 - P_j(\theta_i)\right)\left(1 - 3P_j(\theta_i)\left(1 - P_j(\theta_i)\right)\right), \tag{1.89}$$

を用いて重み付け平均平方の分散を，

$$q_j^2 = \frac{\sum_{i=1}^{N} m_{ij} - w_{ij}}{\left(\sum_{i=1}^{N} w_{ij}\right)^2}, \tag{190}$$

とおき，

$$t_j' = \frac{3}{q_j}\left(\left(v_j'\right)^{\frac{1}{3}} - 1\right) + \frac{q_j}{3}, \tag{1.91}$$

と t 統計量を求めればよい。

適合度の目安としては，v_jおよびv_j'は 0.75~1.3，t_jおよびt_j'は-2.0~2.0 であればモデルに適合しているといわれている (村木, 2011)。熊谷 (2009) の Easy Estimation では，式 (1.85), (1.86), (1.87), (1.91) のことをそれぞれ”OutFit”, “StdOutFit”, “InFit”, “StdInFit”と表記している。

1.7 他の計量モデルとの比較

1.7.1 古典的テスト理論と IRT

IRT は CTT よりも厳しい仮定をおく代わりに，項目と受検者についてより細かな情報を提供してくれるが，無条件にすべてのテストを IRT で分析，運用することはできない。IRT ではパラメタの推定精度とサンプリングコストがトレードオフの関係にある。多変量解析一般に当てはまることだが，最小二乗法や最尤推定法でモデルのパラメタを推定する際には適当な大きさの標本を集めなくてはならない。それは推定結果が収束するかどうかにも関わるが，それ以上に推定されたパラメタがどの程度母集団の推論に使えるかどうかを左右する問題でもある。IRT においては，比較的小数のテスト項目であっても数百から数千の受検者数が安定したパラメタ推定に必要となる(大友, 1996)。多次元 IRT モデルなどの複雑でパラメタの多いモデルであれば，さらに標本数が必要とされるだろう。さらにその推定方法や推定されるパラメタの解釈は一様ではなく，専門家でなくては正確に判断，解釈をおこなうことが難しい。一方で，CTT は単純な素点の合計点で個人得点を算出し，その正答率や平均などの比較的単純な統計量をもとにテストやテスト項目について解釈をおこなうことができる。

そして IRT でテストを分析し，運用していくためには多くのコストがかかることも指摘できる(豊田・岩間・中村・齋藤, 2015)。専門的な知識を持つ研究者と推定結果の保証されたプログラム，精度のよいテスト作成に必要なプレテストの実施，さらにはパラメタの変化を防ぐための項目バンクの構築やセキュリティの問題など，IRT に基づくテストの運用は，従来の一回実施したら問題を破棄するテストの運用とは方針が大きく異なる。これが数万人の受検者を対象とする大規模なテストであればなおさらである。

すなわち CTT と IRT はそれぞれ使い分けられるべきテスト理論である。教師がクラスの進捗状況を確認するため用いる教師作成テストに IRT は不要であり，CTT で十分であるが，政策分析に必要となるような学力分布の変化を知るための学力調査には，正確な能力分布の推定とその測定精度などの付随する情報を提供可能な IRT が適している。

真値についての解釈という点でも CTT と IRT は異なる。式 (1.1) より，古典的テスト理論における観測されるテスト得点は真の得点に確率的に変動する誤差を加えたものと解釈される。そして同一受検者に多数回の試行をおこない，得られた結果の期待値は真値に一致することは式 (1.5) でも示したとおりである。

一方，IRT モデルにおいてある受検者のθが得られたときに，全項目の正答確率を足し上げた得点を期待テスト得点と呼ばれ，古典的テスト理論における真値と似たような解釈ができる。ところで IRT モデルにおける正答確率の解釈について (南風原, 1991) は，(1) 特性値がθであるこ

この受検者の正答確率，(2) 特性値がθである受検者母集団における正答者の比率，というふたつの解釈可能性を提示し，後者のほうが「より一般的で無理のない解釈」だと述べている。したがって IRT の期待テスト得点は母集団における真の得点の平均であると解釈されるため，古典的テスト理論における真値とは異なる得点であることに注意するべきである。

1.7.2 因子分析と IRT

カテゴリカルデータの因子分析と IRT の関係について Takane & De Leeuw (1987) や柳井ら 1990), 豊田 (1998) などによって, 数理的には同一のモデルであることが知られており, 式（1.46）から式 (1.55) でも示したとおりである。

しかし両者では分析の目的が異なる (Reckase, 2009)。IRT は項目のパラメタ（識別力や困難度）を推定したり，あるいは受検者の能力値を推定したりすることが主な目的である。そのほか，テスト情報量などの統計量を推定することや, 後述する等化, 垂直尺度化といった応用が目的だったりもする。一方で, 因子分析の主な目的は多数の項目の背後にある潜在的な因子のまとまりを発見することであり, 項目ごとの特性を分析するために用いられることは滅多にない。そのため通常の因子分析では項目反応データの相関行列を計算し，標準化してしまう。

2　IRT におけるパラメタ推定方法

この節では IRT の分析にかかる労力の多くを占めるパラメタの推定について説明する。前節では一次元 IRT モデル以外にも MIRT モデルや GIRT モデル，GPC モデルなどを扱ったが，ここからは一次元 IRT モデルのみを扱っていく。

母数の推定には多くの数学的，統計的な計算を必要とする。そのため，まずは確率密度関数や尤度関数，ベイズの定理などのいくつかの基本的な式と定理について説明をおこなっておく。次に，受検者の能力パラメタの推定を扱う。実際のデータを分析する過程では，項目パラメタを先に推定する必要があるが，能力パラメタの推定法は一変数の最適化計算であり数理的に簡単であるため，こちらを優先して説明する。最後に，項目パラメタの推定について扱うが，ここでは通常のパラメタ推定に加えて，多母集団モデルのためのパラメタ推定についても説明する。

2.1　基本的な定理と方程式

2.1.1　確率密度関数

確率変数 (random variable) とは確率にしたがって異なる値をとる変数のことである。たとえば IRT モデルは反応確率をモデリングしているが，それは項目反応という確率変数が得られる確率を計算している。確率変数は離散変数 (discreate variable) と連続変数 (continuous variable) の 2 つに大別できる。離散変数とは適当な区間の中で取り得る値が飛び飛びの値をとる変数である。具体的には二値型の項目反応のように，0 か 1 をとる変数やサイコロのでる目などが離散変数にあたる。連続変数は離散変数とは異なり飛び飛びの値をとらない変数である。物体の長さや IRT における能力値はこの連続変数であるが，実際に計算して得られる数値は適当な桁数で打ち切られているため，離散的に扱うことができる。逆に厳密には離散的な値であっても区間内に十分に多くの値が取り得る値として存在する場合には連続変数のように扱うこともある。

この連続確率変数$X = x$が与えられたときに，その確からしさの測度$p(x)$を確率密度 (probability density) と呼ぶ。またこれを関数とみなしたものを確率密度関数 (probability density function) と呼び，$f(x)$とおく。確率変数が離散変数である場合には確率質量関数 (probability mass function) と呼ばれる。確率密度関数は確率変数の確率密度を返す関数であるから，またの名を確率分布 (probability distribution) と呼ぶ。確率密度関数すべての区間に対して$f(x) \geq 0$であり，その区間における積分は 1 と定義される。いま確率変数Xの区間を$-\infty \leq X \leq \infty$としたときに，$p(X \leq x)$を与える関数を累積分布関数 (cumulative density function) と呼ぶ。

確率分布は確率密度関数で定式化される分布であるが，この分布の形状を特定するための値をパラメタ (parameter) とか母数と呼ぶ。

2.1.2 尤度

ある確率変数XがパラメタΘの確率分布$f(X;\Theta)$に従うとする。いまこの確率変数の実現値として$X = x_1, x_2$というふたつの値が得られたとする。このふたつの値が得られる同時確率 (joint probability) は確率の乗法定理から,

$$p(x_1, x_2|\Theta) = p(x_1|\Theta)p(x_2|\Theta), \tag{2.1}$$

と表すことができる。しかし現実的にはx_1, x_2というデータは既知であるが，このデータが得られた確率分布のパラメタΘは未知であることがほとんどである。ここで確率密度関数とは逆に，データを既知のものとして固定したときのパラメタΘの関数を考える。すなわちデータが与えられたときに確率分布のパラメタΘの尤もらしさの測度である。これを,

$$L(\Theta|x_1, x_2) = p(x_1, x_2|\Theta), \tag{2.2}$$

と表し，尤度 (likelihood) と呼ぶ。また，尤度をΘの関数とみなした場合これを尤度関数 (likelihood function) と呼ぶ。数式としては式(2.1)の同時確率と同一であるが，Θの関数となっており，その積分は一般に 1 となるわけではない。

データから確率分布のパラメタを推定する際に，尤度が最も高くなる点をパラメタの推定値とする方法がある。この方法を最大尤度となるパラメタを推定値とすることから最尤推定法 (maximum likelihood estimation) と呼び，推定値を最尤推定値と呼ぶ。最尤推定値の求めるときはパラメタの値で偏微分した関数を＝0 とおいて解く。しかし IRT モデルの尤度関数の導関数は解析には解けず，数値的に最適解を探し出す必要がある。

2.1.3 ベイズの定理

確率変数XとYの条件付き確率(conditional probability)について考える。条件付き確率とは一方が所与のときの他方の確率であり，すなわち$p(X|Y)$はYを所与としたXの確率である。XとYの同時確率をこの条件付き確率を用いて表現すると,

$$p(X, Y) = p(X|Y)p(Y), \tag{2.3}$$

となる。一方で，同時確率はXとYの条件部分をひっくり返しても成立するので,

$$p(X, Y) = p(Y|X)p(X), \tag{2.4}$$

と表すこともできる。以上より，XについてのYの条件付き確率$p(X|Y)$は,

$$p(X|Y) = \frac{p(Y|X)p(X)}{p(Y)}, \tag{2.5}$$

と表現できる。分子の$p(Y)$はYについてのXの周辺確率 (marginal probability) であり，

$$p(Y) = \int p(X,Y)\mathrm{d}X = \int p(Y|X)p(X)\mathrm{d}X, \tag{2.6}$$

と置き換えることができる。周辺確率を計算する際に，確率変数Xは積分により同時確率分布から消去される。このことを積分消去 (integrate out)，あるいは周辺化 (marginalization) という。式(2.5)で表される形がベイズの定理 (Bayes' theorem) である。ベイズ統計学では$p(X)$を事前分布 (prior distribution)，$p(X|Y)$をYで更新した事後分布 (posterior distribution) と呼ぶ。ここで仮に確率変数Xを確率分布のパラメタに置き換えると式 (2.5) は，データを所与としたときの条件付き確率（尤度）とパラメタについての事前分布，そしてデータについての周辺確率の 3 つによって，パラメタの事後確率を定式化しているといえる。ベイズ統計学ではパラメタも確率変数とみなすが，数式としては，

$$p(X|Y) = \frac{L(X|Y)p(X)}{p(Y)}, \tag{2.7}$$

と表現しても問題ない。

ベイズの定理は心理学におけるモデルのパラメタ推定のほか，工学，情報学分野における機械学習 (machine learning) のパラメタ学習にも用いられる非常に汎用的な定理である。しかし定理そのものは同時確率と条件付き確率のみを用いて非常に平易に表すことができる。

このベイズの定理を原則として，事後分布に関しての統計的推論をおこなう体系をベイズ統計学という。この事後分布の代表値を計算することもでき，期待値をとる方法を EAP (expected a posteriori) 推定法，最頻値をとる方法を MAP (maximum a posteriori) 推定法と呼ぶ。

2.2 能力パラメタの推定

ここでは IRT モデルの，受検者の能力パラメタ推定の方法について説明する。基本的に想定するモデルは 3PLM である。2PLM に適用したい場合には当て推量のパラメタを 0 とおくだけで良い。IRT モデルのパラメタを推定する作業は，テストの項目反応データが IRT モデルによって確率的に生成された確率変数であるとみなし，得られたデータから確率モデルの尤もらしいパラメタを推定したり，パラメタの事後分布の代表値を点推定したりするものである。

IRT のパラメタの推定をするための目的関数は解析的に解くことができないため，数値最適化 (numerical approximation) や区分求積などの数値解析の手法を用いる。そのため実際の計算には比較的小規模なテストであっても専用の PC ソフトウェアを用いるほかない。有名なものとして

は有償ソフトウェアの BILOG-MG (Zimowski, Muraki, Mislevy & Bock, 2003) や IRTPRO (Cai et al., 2011) があるほか，共分散構造分析用のソフトウェアである Mplus(Muthén & Muthén, 2006)でも推定ができる。無償で利用可能なものとしては熊谷 (2009) の Easy Estimation がある。そのほかフリーに利用可能なプログラミング言語 R を通して利用できるパッケージとしては ltm (Rizopoulos, 2006) や lazy.irt (Mayekawa, 2016)がある。

2.2.1 最尤推定法

いまひとりの受検者のJ項目のテストにおける項目反応データ$\mathbf{u} = \{\mathrm{u}_1, \mathrm{u}_2, \ldots, \mathrm{u}_J\}$が得られており，既にそれら項目の項目パラメタは既知であるとする。この項目パラメタベクトルは，すべてまとめて$\boldsymbol{\delta}$でおくものとする。したがって項目反応データと項目パラメタが所与の状態での，パラメタθの尤度関数は正答反応確率$P_j(\theta)$と誤答反応確率$Q_j(\theta)$によって，

$$L(\theta|\mathbf{u}, \boldsymbol{\delta}) = \prod_{j=1}^{J} P_j(\theta)^{\mathrm{u}_j} Q_j(\theta)^{1-\mathrm{u}_j}, \tag{2.8}$$

と表される。ここでは二値型のモデルを扱うため受検者の反応パタンは 0 か 1 のいずれかであるため，その場合分けを項目反応u_jのべき乗を使って表現している。次に最適化をおこなうためにパラメタθで偏微分をするのだが，それには尤度関数の対数をとった，

$$\ln L(\theta|\mathbf{u}, \boldsymbol{\delta}) = \sum_{j=1}^{J} \mathrm{u}_j \ln P_j(\theta) + \left(1 - u_j\right) \ln Q_j(\theta), \tag{2.9}$$

対数尤度関数 (log likelihood function) を利用する。対数をとる意味は大きくふたつである。ひとつは計算機のアンダーフローを防ぐためである。テストでは 20 項目や 30 項目ほどの項目が用意されるが，1 項目ごとの反応確率は 0 から 1 までの間の実数をとるため，全項目についての席である尤度関数は非常に小数点以下の桁数が多くなる可能性がある。非常に 0 に近い小さな数は計算機の性質上，正確に表現できなくなる可能性があるため，これを自然対数（ネイピア数を底にとる対数）で置き換えることをする。ネイピア数とは

$$e = \lim_{n \to \infty} \left(1 + \frac{1}{n}\right)^n = 2.7182818 \ldots, \tag{2.10}$$

で与えられる定数である。

もうひとつは導関数の導出の簡便さのためである。積の微分はやや複雑な形に変形されるが，

和の微分であれば，単純に関数同士の微分を足し合わせるだけで良い。なお，対数をとる前後の関数においては以下の関係が成り立つので，

$$f(x) \propto \ln f(x), \tag{2.11}$$

対数をとっても関数の最大となる一点は変化しないため，対数尤度関数の最大点はもとの尤度関数の最大点と一致する。

対数尤度関数の導関数を求める。まず対数をとった 3PLM の項目反応確率のパラメタによる偏微分は，合成微分律により，

$$\frac{\partial \ln P_j(\theta)}{\partial \theta} = \frac{1}{P_j(\theta)} \frac{P_j(\theta)}{\partial \theta} = \frac{D a_j \{P_j(\theta) - c_j\} Q_j(\theta)}{P_j(\theta)(1 - c_j)}, \tag{2.12}$$

である。ここでは省略したが$P_j(\theta)$の導関数は指数関数の中を適当な文字に置き換える工夫をすることで，比較的容易に展開することができる。詳しくは豊田（2005, p2）を参照されたい。なお$\ln Q_j(\theta)$の導関数については式(2.12)にマイナスの符号をつけ，分母の$P_j(\theta)$を$Q_j(\theta)$に置き換えるだけで良い。式 (2.12) を式 (2.9) の導関数に代入する。積の微分の公式に当てはめるが，u_jをθで微分すると 0 になるので単純に

$$\frac{\partial \ln L(\theta|\mathbf{u}, \boldsymbol{\delta})}{\partial \theta} = \sum_{j=1}^{J} u_j \frac{D a_j \{P_j(\theta) - c_j\} Q_j(\theta)}{P_j(\theta)(1 - c_j)} + (1 - u_j) \left\{ - \frac{D a_j \{P_j(\theta) - c_j\}}{1 - c_j} \right\}, \tag{2.13}$$

となり，整理すると，

$$\frac{\partial \ln L(\theta|\mathbf{u}, \boldsymbol{\delta})}{\partial \theta} = D \sum_{j=1}^{J} u_j \frac{a_j \{\mathrm{u}_j - P_j(\theta)\} \{P_j(\theta) - c_j\}}{P_j(\theta)(1 - c_j)}, \tag{2.14}$$

が得られる。これを特に尤度方程式と呼ぶ。

これを＝0 とおいた方程式を解くことで最尤推定値が求まる。しかし式 (2.14) は解析的に解くことができないため適当な初期値 (initial value) を与えた後に，何らかの方針にしたがって繰り返し計算をしていくことで，漸近的に最適な解を求めるという反復数値計算の方法をとる。この方法として有名なものはニュートン・ラフソン (Newton-Raphson) 法とフィッシャースコアリング (Fisher scoring) という手法がある。二階偏微分つまり二次導関数を用いて逐次計算をおこなう。t+1 回目の更新された解は，

$$\theta_{t+1} = \theta_t - \frac{f'(\theta_t)}{f''(\theta_t)}, \tag{2.14}$$

と求めることができる。ただし，$f'(\theta_t)$は一階偏微分，$f''(\theta_t)$は二階偏微分を表す。一方でフィッシャースコアリングにおける t+1 回目の更新された解は IRT におけるフィッシャー情報量であるテスト情報関数$I(\theta_t)$を用いて，

$$\theta_{t+1} = \theta_t + \frac{f'(\theta_t)}{I(\theta_t)}, \tag{2.15}$$

と計算をおこなう。どちらの手法も近似的に最適解に近づくことはできるが，理論的には無限に繰り返しを行えてしまうため，一定の基準を設けて，その基準を満たす推定値が得られた時点で収束を判断しなければならない。一般的な収束基準としてはパラメタや対数尤度の変化量が十分小さくなることを条件とすることが多い。実用的にはパラメタの変化量が$1e-3$か$1e-4$を下回る程度で十分である。

ニュートン・ラフソン法の仕組みについて簡単に説明すると，上に凸な尤度関数の尤度方程式は，尤度関数のある点における接線の傾きについての関数である。つまり尤度方程式は単調減少な，必ず 0 を通る曲線である。尤度関数の二次導関数は尤度方程式の曲線上の適当な点における接線であるから，必ず負の値をとる。この二次導関数が$f'(\theta_t) = 0$の軸とぶつかる点が更新された値である。ニュートン・ラフソン法の更新は二次で収束するので，適切な初期値を選択すれば繰り返しが多くなるほど最適解に近づく速度は増える。ここに尤度関数の二次導関数を示す。

$$\frac{\partial^2 \ln L(\theta|\mathbf{u},\boldsymbol{\delta})}{\partial \theta^2} = D^2 \sum_{j=1}^{J} u_j \frac{a_j^2 \{u_j c_j - P_j(\theta)^2\}\{P_j(\theta) - c_j\} Q_j(\theta)}{P_j(\theta)^2 \left(1 - c_j\right)^2}, \tag{2.16}$$

式 (2.15) で示されるフィッシャースコアリングもニュートン・ラフソン法と同様に最適解に収束する。収束した最適解におけるテスト情報量の逆数が最尤推定値の漸近分散を与えることが知られている (前川，1991)。ちなみに 2PLM の場合，テスト情報関数と尤度関数の二次導関数は全く同じ形をとるため，2 つの方法は同一であるとみなせる。ニュートン・ラフソン法はフィッシャースコアリングに比べて最適解に近づく速度は速いものの，実際用いられる収束基準においては両者の反復回数に大きな差はない。この他により簡便な方法として二分法 (bisection method) を用いることもできる。二分法は最適解を間に含む適当な 2 点を初期値として設定する必要があるが，一次導関数のみで計算を実行することができる。数値最適化の手法は目的関数によってうまく収束することもあれば，収束しないこともある。万全を期すのであればいくつかの

最適化の手法を目的関数に対して用意しておくべきである。

ニュートン・ラフソン法をはじめとする多くの数値最適化の手法にはいくつかの欠点がある。それは収束する解が初期値に依存しやすいことと，局所的な最適解を求めてしまうことがある点である。初期値依存と局所的最適解が問題になるケースのほとんどは，目的関数が多峰型の形状をしているときである。一次元 IRT モデルにおいては 3PLM において二山の尤度関数が与えられるため，初期値や最適化の手法によっては局所的最適解に陥ってしまう。

視覚的に反復計算の状況を示すために，乱数により 30 項目分の項目パラメタと項目反応データを発生させ，その対数尤度関数，尤度方程式，二次導関数，テスト情報関数を図示した (図 2.1~2.4)。なおテスト情報関数は二次導関数の関数値の符号と揃えるためにマイナスをかけている。

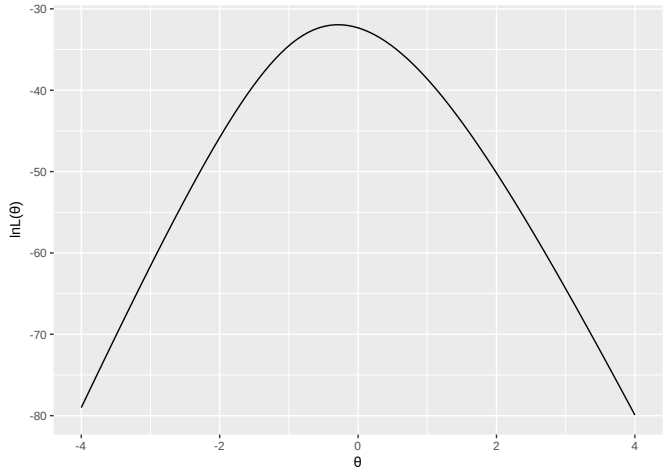

図 2.1 仮想的な 30 項目のパラメタによる 2PLM の対数尤度関数

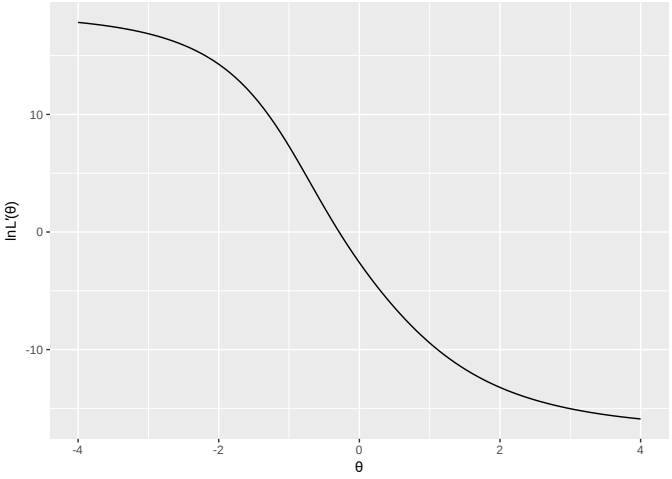

図 2.2 対数尤度関数の一階偏微分（尤度方程式）

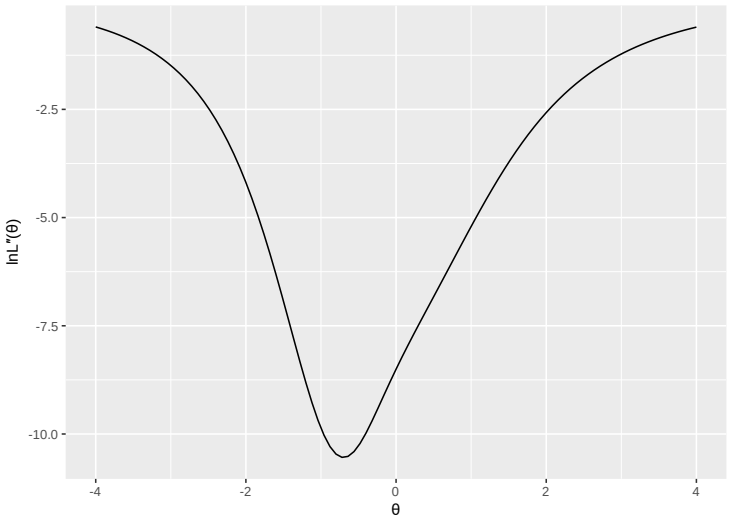

図 2.3 対数尤度関数の二階偏微分（二次導関数）

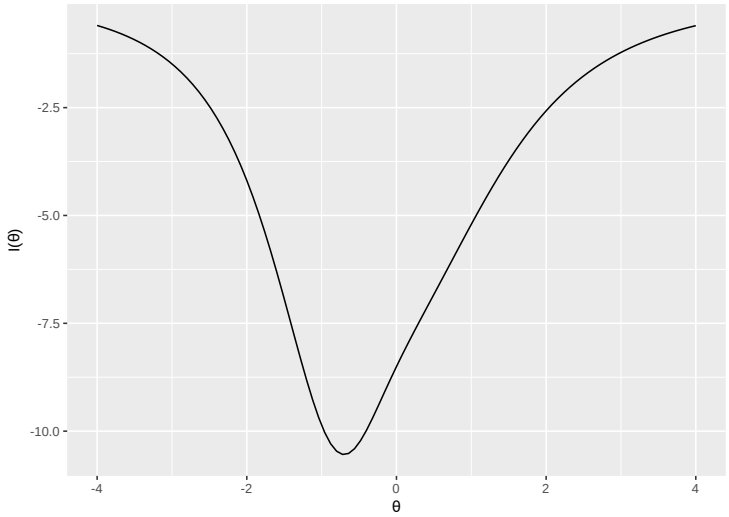

図 2.4 負のテスト情報関数

IRT モデルは 2PLM であるため，項目反応パタンが全問正答，誤答でなければ尤度関数は必ず定義域中のどこかで極大となる単峰型の関数になる。ニュートン・ラフソン法およびフィッシャースコアリングによるθの反復計算の様子を図 2.5 と図 2.6 に示す。初期値は－2 とし，収束基準は項目パラメタの差の絶対値が 0.001 を下回ることとした。収束プロセスのグラフでは，黒い実線（一次導関数）が 0 となる位置が最適解であるが，色のついた接線（二次導関数）が徐々に最適解付近の接線へと変化していっているプロセスを表現している。尤度方程式と接線の交点が更新されたθの値を示しており，その接線が縦軸＝ 0 となる位置のθが次に更新される値を示している。先ほども述べたように 2PLM において両者はまったく同じ反復計算の過程を経る。

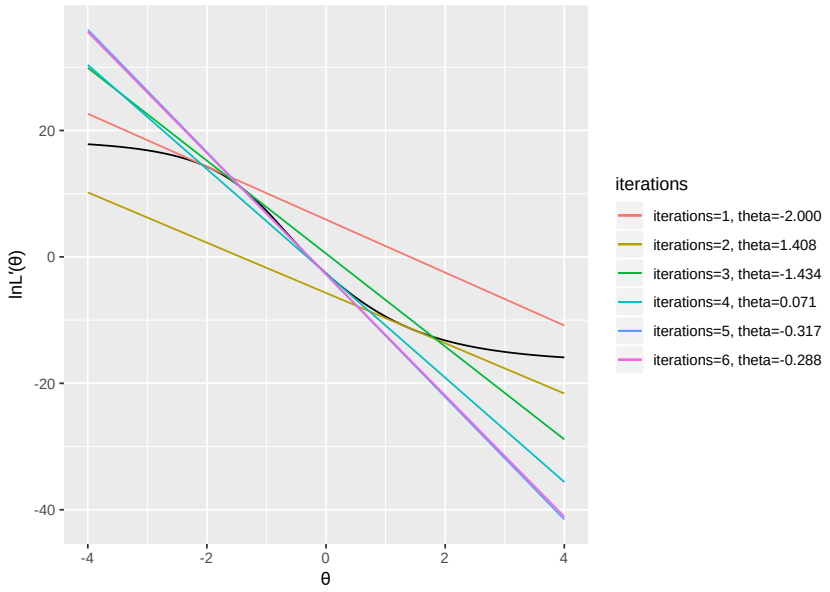

図 2.5 ニュートン・ラフソン法による反復計算のプロセス（2PLM）

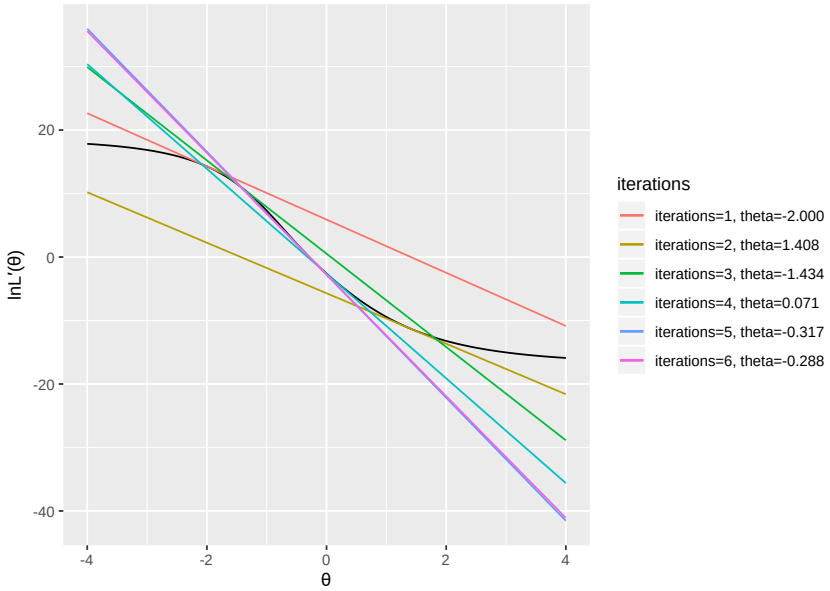

図 2.6 フィッシャースコアリングによる反復計算のプロセス（2PLM）

ニュートン・ラフソン法とフィッシャースコアリングの比較のために，3PLM での反復計算の様子を示す (図 2.7~2.12)。こちらは最適解が 0 付近であることが尤度方程式から視覚的に観察できたので，初期値は 0 を設定し，収束基準は 2PLM のケースと同様の設定を用いた。ニュートン・ラフソン法は 3 回で，フィッシャースコアリングは 4 回で収束したと判断されている。

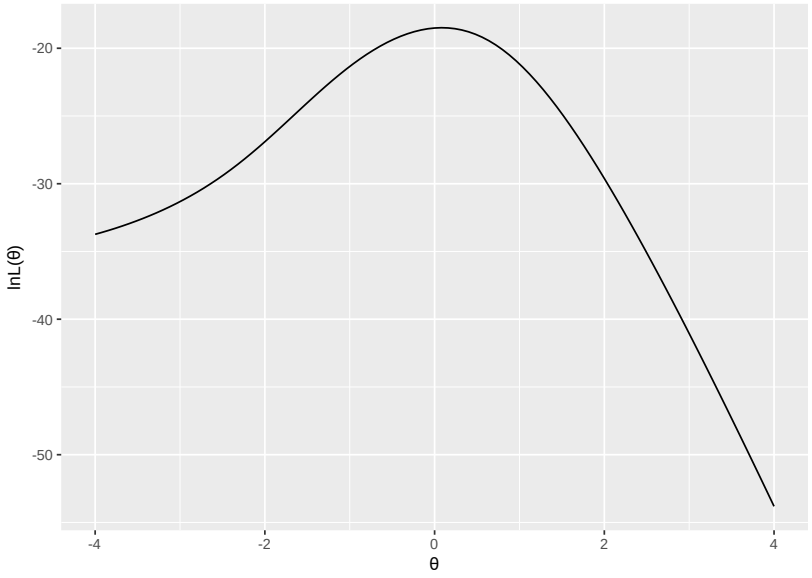

図 2.7 3PLM の対数尤度関数

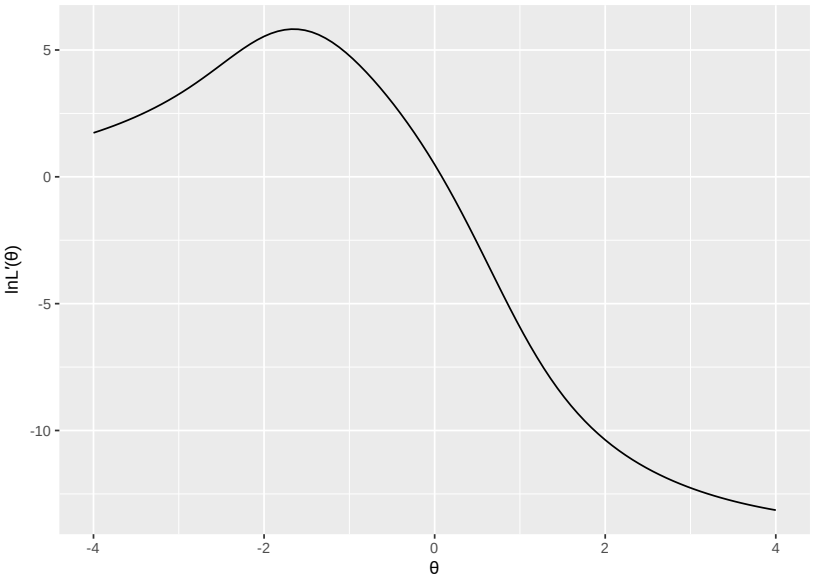

図 2.8 3PLM の尤度方程式

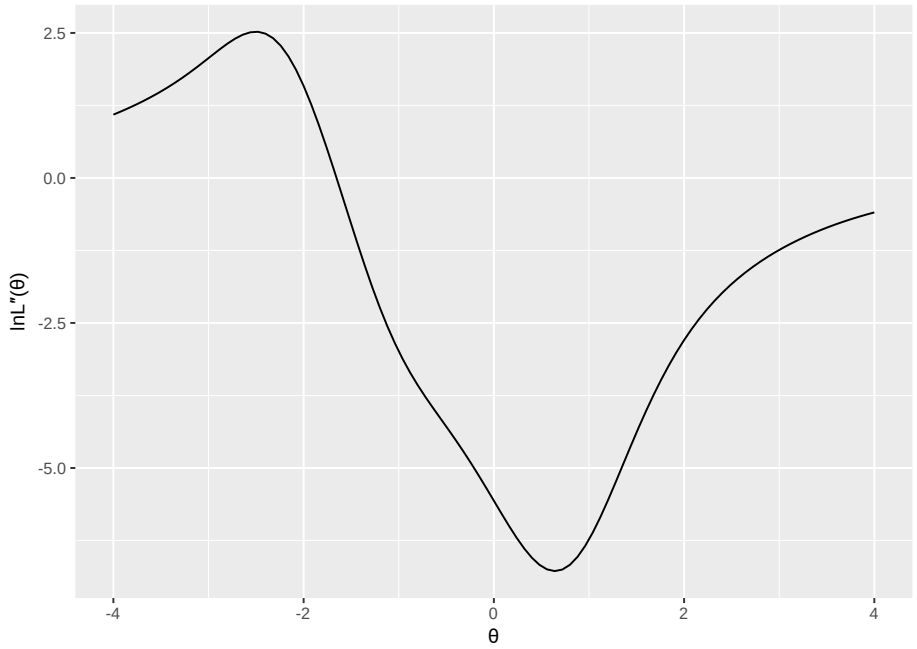

図 2.9 3PLM の二次導関数

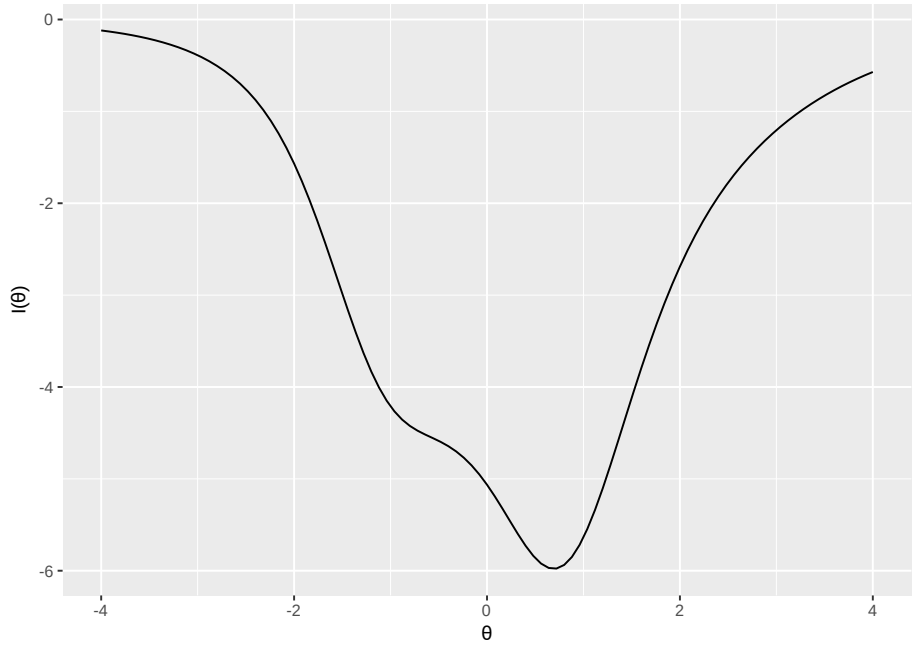

図 2.10 3PLM の負のテスト情報関数

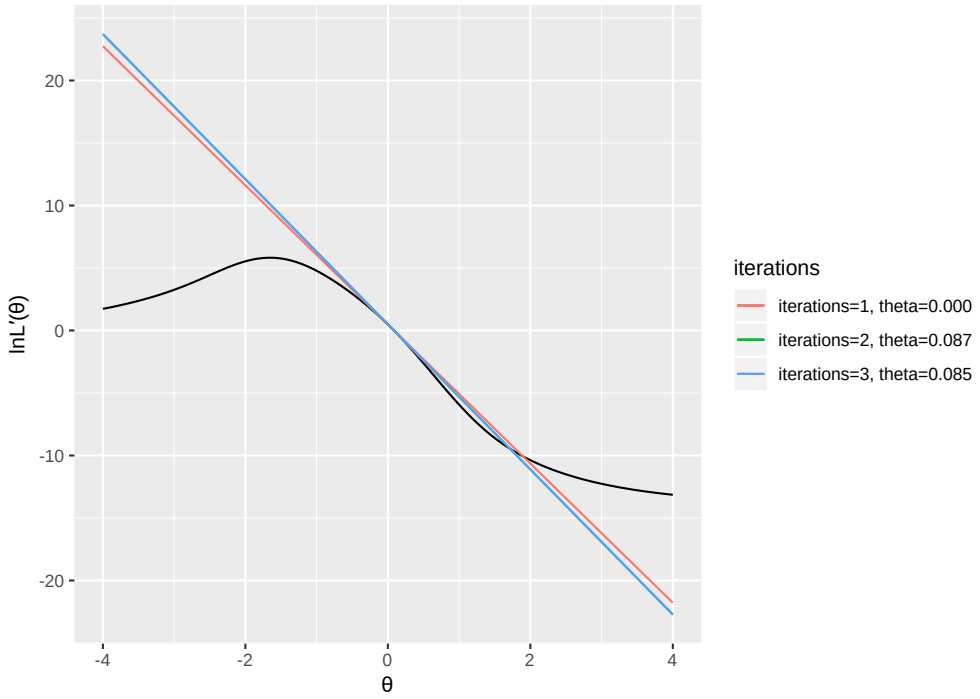

図 2.11 ニュートン・ラフソン法による反復計算のプロセス（3PLM）

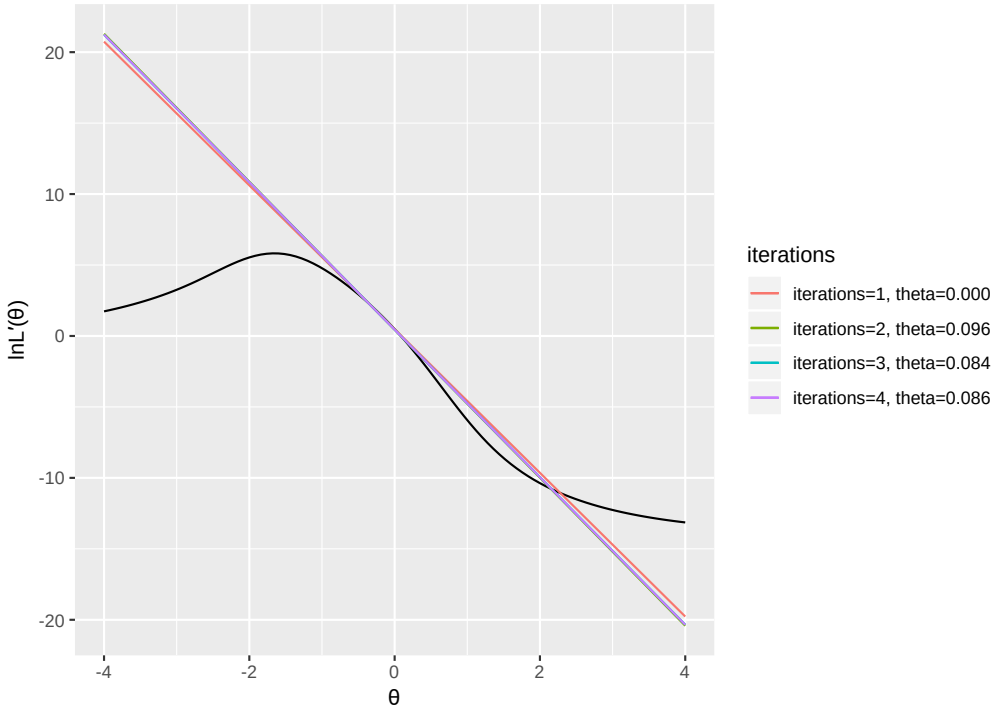

図 2.12 フィッシャースコアリングによる反復計算のプロセス（3PLM）

2.2.2 最大事後確率推定（ベイズ最頻値）

最尤推定法は原理的に全問正答，誤答の受検者の尤度関数が発散するため推定値を計算することができない。こうしたケースで利用されるのがベイズ推定法 (Bayesian estimation) である。ベイズ推定法では先述したように EAP 推定値と MAP 推定値が推定値として利用可能であり，上記のようなケースでも推定値が得られ，事後分布の標準偏差を求めることで推定値の事後標準偏差を計算することができる。

最尤推定法と方法的に非常に似ているので，まずは MAP 推定値の求め方から説明する。まずベイズの定理で用いる尤度関数は式 (2.8) をそのまま用いれば良いので，事前分布を導入する。事前分布を$p(\theta)$, 項目パラメタベクトルを$\boldsymbol{\delta}$とすると，ベイズの定理により事後分布$g(\theta|\mathbf{u};\boldsymbol{\delta})$は，

$$g(\theta|\mathbf{u};\boldsymbol{\delta}) = \frac{L(\theta|\mathbf{u};\boldsymbol{\delta})p(\theta)}{p(\mathbf{u})}, \tag{2.17}$$

と表すことができる。ただしここでも項目パラメタは所与のものとする。MAP 推定値はこの事後分布の最頻値, つまり分布の最大点を求める推定方法である。ここで右辺の分母は規格化のための定数項であるため最適化の計算の際は無視できるので，最大化の目的関数を分子だけの形である，

$$g(\theta|\mathbf{u};\boldsymbol{\delta}) \propto L(\theta|\mathbf{u};\boldsymbol{\delta})p(\theta), \tag{2.18}$$

とする。あとは最尤推定法のときと同様にパラメタθで偏微分した関数を＝0 とおいて解くのだが，ここでもやはり目的関数は解析的には解くことができないため先ほど説明したニュートン・ラフソン法やフィッシャースコアリングなどの数値最適化の手法を用いることになる。

導関数を導出する前に事前分布を具体的な関数の形に定式化する必要がある。一般的な学力テストでは受検者の素点のヒストグラムは単峰型の分布をすることが多い。もちろん CTT の範疇のヒストグラムなので，テスト項目の設定や受検者集団のレベルによっても分布は変化することには注意が必要であるが，このことを踏まえて能力パラメタの事前分布には正規（ガウス）分布 (normal distribution, Gaussian distribution) が用いられることが多い。したがってここでも，

$$p(\theta) = N(\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\,\sigma}\exp\left\{-\frac{1}{2}\left(\frac{\theta-\mu}{\sigma}\right)^2\right\}, \tag{2.19}$$

を事前分布とする。式 (2.19) を式(2.18) に代入して対数をとると，指数関数の中はそのまま外にくくり出され,

$$\ln g(\theta|\mathbf{u};\boldsymbol{\delta}) = L(\theta|\mathbf{u};\boldsymbol{\delta}) - \frac{1}{2}\left(\frac{\theta-\mu}{\sigma}\right)^2 + \ln\frac{1}{\sqrt{2\pi}\,\sigma}, \tag{2.20}$$

となり，これをθで偏微分すると，第一項は式 (2.14) で与えられる尤度方程式であり，第 3 項は消え，第 2 項は商の微分の法則と合成微分律により変形され，最終的に一次導関数は以下の形になり，

$$\frac{\partial \ln g(\theta|\mathbf{u};\boldsymbol{\delta})}{\partial\theta} = \frac{\partial L(\theta|\mathbf{u};\boldsymbol{\delta})}{\partial\theta} - \frac{\theta-\mu}{\sigma^2}, \tag{2.21}$$

二次導関数は，

$$\frac{\partial^2 \ln g(\theta|\mathbf{u};\boldsymbol{\delta})}{\partial\theta^2} = \frac{\partial^2 L(\theta|\mathbf{u};\boldsymbol{\delta})}{\partial\theta^2} - \frac{1}{\sigma^2}, \tag{2.22}$$

となる。

ベイズ統計学の枠組みでは事前分布に明確な根拠がない場合には広い分散をもつ正規分布や区間の広い一様分布などの無情報事前分布 (noninformative prior) を用いることが推奨される (松浦, 2016)。一様分布は確率変数の区間[a. b]において確率密度関数が，

$$p(x) = \frac{1}{b-a}, \tag{2.23}$$

で与えられるような確率分布である。確率変数xを右辺に含まないため，仮に偏微分したとすれば 0 となる。したがって一様分布を事前分布としたときの MAP 推定値は最尤推定値に一致する。

式 (2.21) を見ると，第 2 項は事前分布の影響を表す項であり，尤度方程式全体をμの方向にσ^2で標準化したθとμの偏差の分だけずらす働きをしていることが分かる。これは統計学や機械学習の分野における正則化 (regularization)と同じ働きである。データに対する過剰な適合を抑制するための罰則が事前分布の働きとも考えられるだろう。図 2.13 は図 2.8 のパラメタセットと全く同じ条件で，対数尤度関数の一次導関数に正規分布の偏微分を加えたものである対数事後分布の一次導関数が，正規分布のパラメタによってどのように変化するのかを視覚化したものである。

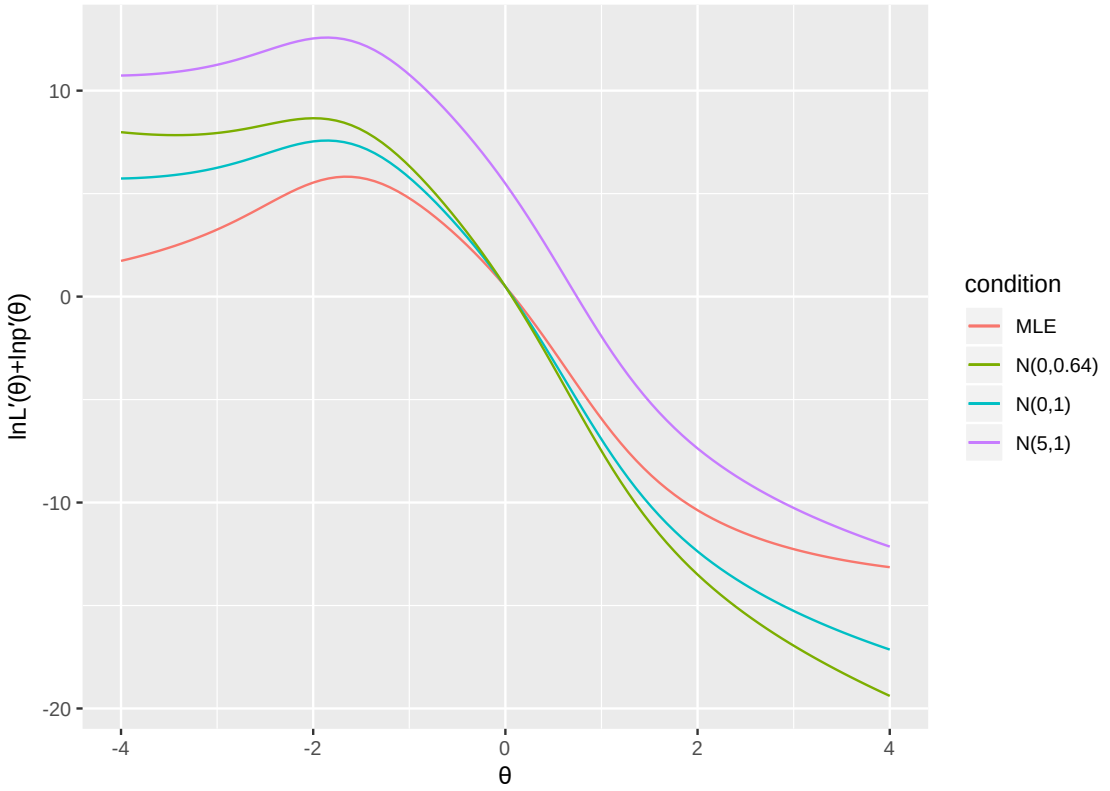

図 2.13 事前分布による事後分布の一次導関数の変化（3PLM）

この図からは仮に事前分布の平均を大きくずらしても，100%そのズレが推定値にも反映されるわけではなく，あくまでもデータの情報を補正する働きしかしないことや，分散のパラメタを変化させると平均の方向にわずかに収縮 (shrinkage) することが視覚的に分かる。

項目数を増やせばその分だけデータの持つ情報量が多くなるため，事前分布が与える収縮の影響は小さくなる。また事前分布の分布族やパラメタの設定によって同じ反応パタンでも MAP 推定値は変わるため，テストの客観性や公平性という観点からするとベイズ推定法は必ずしも良い結果をもたらすわけではない (加藤ら，2014)。

2.2.3 期待事後平均推定

EAP 推定値は最尤推定値や MAP 推定値と違い，反復計算をともなう推定法ではない。EAP という名の通り，事後分布の期待値を推定値とするが，やはりこの期待値も解析的に解くことはできない。解析的に解くことができない確率分布の期待値を求める方法としては，期待値計算ができる指数分布族などで確率分布を近似する方法がある。具体的にはラプラス近似や変分ベイズ (Variational Bayes, VB) 法などがこの方法にあたる。また確率分布に従う疑似乱数を大量に発生させて近似推論することも考えられる。棄却サンプリング (rejection sampling) の他，近年注目されているマルコフ連鎖モンテカルロ (Markov Chain Monte Carlo, MCMC) 法の一種であるギブスサンプリング (Gibbs sampling) やスライスサンプリング (slice sampling) といった様々な手法が

ある。これら近似推論の方法や高次元や複雑な形状の確率分布において有効な手法として注目されているが, IRT の受検者の能力分布は一次元の比較的単純な形状をした確率分布であるため, ここまで複雑な方法を用いなくても期待値を計算することができる。

ここでは近似積分計算の代表的な手法である区分求積法の等区分求積を用いる。等区分求積法は確率分布の区間を任意の実数の範囲に限定し，さらに範囲を有限の実数で等間隔に分割する。分割した点 (node) に対して分点ごとの適当な重み (weight) をかけ，それらを足し上げることで確率分布の積分を離散近似する手法である。離散近似の精度は分点数が増えるにしたがって向上していくが, より厳密な方法を用いるのであれば, 正規分布の形状が中央に密度が集中し, 裾は薄く広がっている特徴を考慮して，ガウスエルミート求積法 (Gauss-Hermite quadrature) を用いることもできる。これにより少ない求積点でも精度の良い近似が実行可能である。いずれにしても分点の数は 40 もあれば十分である (村木，2011)。

事後分布の期待値は,

$$\mathrm{E}[g(\theta|\mathbf{u};\boldsymbol{\delta})] = \int_{-\infty}^{\infty} \theta g(\theta|\mathbf{u};\boldsymbol{\delta})\mathrm{d}\theta = \frac{\int_{-\infty}^{\infty} L(\theta|\mathbf{u};\boldsymbol{\delta})p(\theta)\mathrm{d}\theta}{\int_{-\infty}^{\infty} p(\mathbf{u}|\theta)\mathrm{d}\theta}, \tag{2.24}$$

と定義される。この積分を離散近似するためには，確率変数θの区間を適当な実数，例えば-4 から 4 の区間に限定し，M個の等間隔の分点に分割する。ここで積分の対象になっているのは事前分布であるから，事前分布の区間におけるある分点mにおける求積点の値をX_m，それに対応する重みをW_mとすると,

$$\mathrm{E}[g(\theta|\mathbf{u};\boldsymbol{\delta})] \approx \frac{\sum_{m=1}^{M} X_m L(X_m|\mathbf{u};\boldsymbol{\delta})W_m}{\sum_{l=1}^{M} L(X_l|\mathbf{u};\boldsymbol{\delta})W_l}, \tag{2.25}$$

と表せる。実際の計算アルゴリズムとしては分母の定数項を計算してから分子を計算しても良いし，あるいは式 (2.25) を整理して,

$$\mathrm{E}[g(\theta|\mathbf{u};\boldsymbol{\delta})] \approx \sum_{m=1}^{M} X_m \frac{L(X_m|\mathbf{u};\boldsymbol{\delta})W_m}{\sum_{l=1}^{M} L(X_l|\mathbf{u};\boldsymbol{\delta})W_l}, \tag{2.26}$$

求積点と事後分布の重みという様にまとめて計算することもできる。

EAP 推定法は式 (2.25) や，あるいは式 (2.26) の計算だけで確実に計算することができる。この点は反復計算を必要とし，ときに局所的最適解に陥る可能性のある最尤推定法や MAP 推定法にはないメリットを備えているとも言える。しかし精度を上げるために分点の数を増やすと，その分計算コストも増加する。尤度関数の形状が明らかに単峰型ですみやかにニュートン・ラフソン法により最適解が計算できる場合には，そちらの方が計算コストはかからない。

2.3 項目パラメタ推定

これまで所与のものとして扱ってきた項目パラメタの推定方法について述べる。項目パラメタの推定方法にはいくつかの種類があるが，様々な項目パラメタの推定方法は Baker & Kim (2004)や村木 (2011), 加藤ら (2014), 豊田 (2005) などに詳しい。ここでは同時最尤推定法 (Joint Maximum Likelihood Estimation, JMLE) と周辺最尤推定法 (Marginal Maximum Likelihood Estimation, MMLE) および，EM アルゴリズムを用いる MMLE(MMLE via EM algorithm, MMLE-EM)，周辺ベイズ推定法 (Marginal Bayes Estimation, MBE)，階層ベイズ推定法 (Hierarchical Bayes Estimation, HBE) について説明する。理論的な観点から JMLE や MMLE の説明を簡単におこなうが，実際に本稿の実験における項目パラメタの推定で使用されるのは MMLE-EM および MBE である。

2.3.1 同時最尤推定法

JMLE は対数尤度関数，

$$\ln L(\boldsymbol{\delta}, \theta|\mathbf{U}) = \sum_{i=1}^{N}\sum_{j=1}^{J} u_{ij}\ln P_j(\theta_i) + \left(1-u_{ij}\right)\ln Q_j(\theta_i), \tag{2.27}$$

を最大化するパラメタ$\boldsymbol{\delta}, \theta$を同時に推定する方法である (Baker and Kim, 2004; de Ayala, 2009; Wingersky et al., 1982)。ただし能力パラメタの推定のときとは異なり，項目反応パタンはベクトルではなく，N × Jの行列である。この目的関数を一度に最大化することは困難であるため，まず各パラメタの尤度方程式を導出し，2 つのステップを交互に繰り返すことで漸近的に最適解を得る。具体的には，(1) 項目パラメタの初期値$\boldsymbol{\delta_0}$を所与として$\boldsymbol{\theta_0}$の MLE 推定値を得て，(2) $\boldsymbol{\theta_0}$を所与として，今度は項目パラメタベクトルを更新して$\boldsymbol{\delta_1}$を得る，というステップを，パラメタか式 (2.27) の対数尤度の変化率が十分小さくなるまで繰り返す。

2PLM 以上のモデルでは項目パラメタが多変数になるため，先に説明したニュートン・ラフソン法やフィッシャースコアリングを多変数に拡張する必要がある。多変数関数のニュートン・ラフソン法には，対数尤度関数のパラメタベクトル$\boldsymbol{\delta}$についての一次導関数ベクトル$\mathbf{d}(\boldsymbol{\delta})$ (勾配ベクトル，gradient vector) と，そのヘッセ行列$\mathbf{H}(\boldsymbol{\delta})$の逆行列を用いる。つまりt + 1回目の更新されたパラメタは，

$$\boldsymbol{\delta}_{\mathbf{t+1}} = \boldsymbol{\delta}_{\mathbf{t}} - \mathbf{H}(\boldsymbol{\delta})^{-1}\mathbf{d}(\boldsymbol{\delta}), \tag{2.28}$$

と求める事ができる。局所独立の仮定が成り立つのであれば，θが所与のときの各項目反応は独立であることが約束されるため，次のステップでの数値計算はパラメタ数をpとすれば，p × p行列のヘッセ行列の逆行列を解くことで実行できる。

JMLE のパラメタ更新では尺度の不定性 (scale invariance, indeterminacy) が問題となる。IRT モ

デルの反応確率はパラメタに依存するが，反応確率自体はパラメタの尺度の単位と原点に依存しない。つまり，ひとつの正答確率Pをとるパラメタの組み合わせが無数に存在する。これは得られた全受検者の対数尤度関数ををN＋J×p元連立方程式とみなしたときに，その次数は 1 次にとどまるため，解が不定となることからも明らかである。そこで JMLE では 1 サイクルの更新が終了した段階で項目パラメタか能力パラメタのどちらかの尺度の単位と原点を適当な値に揃える必要がある。この線形変換には等化係数 (equating coefficient) と呼ばれる係数が必要になるが，詳しくは後ほど等化の節で取り扱う。

2.3.2 周辺最尤推定法

同時最尤推定法においては項目パラメタだけでなく能力パラメタも漸次最適な値に更新しなくてはならなかった。しかし項目パラメタがテストを実施する段階ではじめから固定されている構造 (structure) 母数であることに対して，能力パラメタはサンプルサイズを大きくするほど増加してしまう付随 (incidental) 母数である。Neyman & Scott (1948) が指摘したように，同時推定では付随母数は一致推定量とならない統計的に望ましくない性質があるため，これを改善した方法を Bock & Lieberman (1970) が提案した。それが MMLE である。

項目パラメタ推定において尤度関数における能力パラメタは局外 (nuisance) 母数である。MMLE ではベイズ統計の考え方を利用して，この局外母数を周辺化 (marginalization) することで一致性の問題を解決している。能力パラメタの事前分布を$h(\theta)$とする。仮にこの分布が何らかのパラメタΘを持つパラメトリックな分布であるとすれば$h(\theta|\Theta)$と書ける。これにより周辺化された周辺尤度関数 (marginal likelihood function)，および周辺対数尤度関数$\ln L_M$は，

$$L_M(\boldsymbol{\delta}|\mathbf{U}) = \prod_{i=1}^{N} \int \left\{ \prod_{j=1}^{J} P_j(\theta_i)^{u_{ij}} Q_j(\theta_i)^{1-u_{ij}} \right\} h(\theta_i) \mathrm{d}\theta_i \,, \tag{2.29}$$

$$\ln L_M(\boldsymbol{\delta}|\mathbf{U}) = \sum_{i=1}^{N} \ln \left[\int \left\{ \prod_{j=1}^{J} P_j(\theta_i)^{u_{ij}} Q_j(\theta_i)^{1-u_{ij}} \right\} h(\theta_i) \mathrm{d}\theta_i \right], \tag{2.30}$$

と表される。積分計算の離散近似には Bock & Lieberman (1970) にしたがってガウスエルミート求積法を用いればよい。

計算された対数尤度関数は，この後$\boldsymbol{\delta}$の各要素で偏微分される。しかし項目について独立に計算することができないため計算式は非常に複雑であり，数値最適化のための逆行列計算はpJ × pJという大きなサイズの逆行列を計算しなくてはならない。

2.3.3 EM アルゴリズムによる周辺尤度の最大化

そこで Bock & Aitkin (1981) は EM アルゴリズム (Dempster, Laird & Rubin, 1977) を用いる計

算方法を提案することにより計算効率を大幅に向上させている。EM アルゴリズム (Expected-Maximum algorithm) は潜在変数や欠測値を含む統計モデルとデータにおけるパラメタ推定方法の一般的な枠組みである。EM アルゴリズムにとよる最適化では潜在変数をデータから欠測した情報とみなし，欠測データの情報を確率分布で暫定的に補った尤度関数を最大化する。Bock & Aitkin (1981) が提案した EM アルゴリズムによる周辺最尤推定法は一般的な EM アルゴリズムの導出とはやや異なる。

一般的な説明のためデータ行列を$\mathbf{U}$，潜在変数ベクトルを$\boldsymbol{\theta}$，推定したいパラメタベクトルを$\boldsymbol{\delta}$とおく。まず一般的な EM アルゴリズムでは対数尤度関数$\ln L(\boldsymbol{\delta}, \boldsymbol{\theta}|\mathbf{U})$について暫定的に$\boldsymbol{\theta}$が観測されたとして得られる完全データ対数尤度関数 (complete data likelihood function),

$$\ln L_C(\boldsymbol{\delta}|\mathbf{U}, \boldsymbol{\theta}), \tag{2.31}$$

について考える。さらにこの関数に初期値$\boldsymbol{\delta_0}$を与えたときの条件付き期待値,

$$\mathrm{E}[\ln L_C(\boldsymbol{\delta}|\mathbf{U}, \boldsymbol{\theta})|\boldsymbol{\delta_0}], \tag{2.32}$$

を求めるステップが E ステップである。この期待値は潜在変数ベクトル$\boldsymbol{\theta}$の事後分布$g(\boldsymbol{\theta}|\mathbf{u}, \boldsymbol{\delta_0})$により,

$$\mathrm{E}[\ln L_C(\boldsymbol{\delta}|\mathbf{U}, \boldsymbol{\theta})|\mathbf{U}; \boldsymbol{\delta_0}] = \int_{-\infty}^{\infty} \ln L_{\mathrm{C}}(\boldsymbol{\delta}|\mathbf{U}, \boldsymbol{\theta}) g(\boldsymbol{\theta}|\mathbf{U}, \boldsymbol{\delta_0}) \mathrm{d}\boldsymbol{\theta} \equiv Q(\boldsymbol{\delta}|\boldsymbol{\delta_0}), \tag{2.33}$$

と計算される。ただし$\boldsymbol{\theta}$の積分区間は$-\infty$から∞である。この関数は$\boldsymbol{\delta_0}$と$\mathbf{U}$によって与えられる潜在変数$\boldsymbol{\theta}$の事後分布によって，本来は観測され得ない潜在変数に関する情報が暫定的に補われた,$\boldsymbol{\delta}$についての関数であるとみなすことができる。これを一般に Q 関数，あるいは期待対数完全データ尤度関数 (expected log complete likelihood function) と呼ぶ。

期待値を計算し，Q 関数を導出できれば，次は Q 関数を最大化するような$\boldsymbol{\delta}$を何らかの手段で推定し，更新された推定値を得る。このステップを M ステップと呼ぶ。Q 関数はもとの潜在変数を含む対数尤度関数に比べていくらか簡単な表現になっていることが多いため，ときにこの最大化の問題は解析的に解くことができる場合がある。解析的に解けない場合にはニュートン・ラフソン法などの反復計算により最適解を計算する必要があるが，その場合に EM アルゴリズムが最適解にたどり着くことができるかどうかは，M ステップが適切に最適化されているかどうかに依存してしまう。

EM アルゴリズムによる反復が期待対数完全データ尤度関数を扱っているのにもかかわらず,もとの対数尤度関数を単調増加させることができる理論的根拠を，樺島・上田 (2003)に則って示す。まず，t回目の反復における対数尤度は，$\boldsymbol{\delta_t}$をその時点までに得られている暫定的な推定

値とすれば，潜在変数の事後分布を乗算，除算することで以下のように変形できる。

$$\ln L(\boldsymbol{\delta}|\mathbf{U}) = \ln \int g(\boldsymbol{\theta}|\mathbf{U}, \boldsymbol{\delta}_\mathbf{t}) \frac{L(\boldsymbol{\delta}, \boldsymbol{\theta}|\mathbf{U})}{g(\boldsymbol{\theta}|\mathbf{U}, \boldsymbol{\delta}_\mathbf{t})} \mathrm{d}\boldsymbol{\theta} = \ln E\left[\frac{L(\boldsymbol{\delta}, \boldsymbol{\theta}|\mathbf{U})}{g(\boldsymbol{\theta}|\mathbf{U}, \boldsymbol{\delta}_\mathbf{t})} \,\middle|\, \boldsymbol{\delta}_t \right]. \tag{2.34}$$

これは事後分布$g(\boldsymbol{\theta}|\mathbf{U}, \boldsymbol{\delta}_\mathbf{t})$についての期待値と見なせる。ここで Jensen 不等式

$$\ln E[f(x)] \geq E[\ln f(x)]$$

を利用すると，

$$\ln E\left[\frac{L(\boldsymbol{\delta}, \boldsymbol{\theta}|\mathbf{U})}{g(\boldsymbol{\theta}|\mathbf{U}, \boldsymbol{\delta}_\mathbf{t})} \,\middle|\, \boldsymbol{\delta}_t \right] \geq E\left[\ln \frac{L(\boldsymbol{\delta}, \boldsymbol{\theta}|\mathbf{U})}{g(\boldsymbol{\theta}|\mathbf{U}, \boldsymbol{\delta}_\mathbf{t})} \,\middle|\, \boldsymbol{\delta}_t \right] = \int g(\boldsymbol{\theta}|\mathbf{U}, \boldsymbol{\delta}_\mathbf{t}) \ln \frac{L(\boldsymbol{\delta}, \boldsymbol{\theta}|\mathbf{U})}{g(\boldsymbol{\theta}|\mathbf{U}, \boldsymbol{\delta}_\mathbf{t})} \mathrm{d}\boldsymbol{\theta}, \tag{2.35}$$

というように不等式で表現できる。このとき事後分布$g(\boldsymbol{\theta}|\mathbf{U}, \boldsymbol{\delta}_\mathbf{t})$はベイズの定理により自然な形として，

$$g(\boldsymbol{\theta}|\mathbf{U}, \boldsymbol{\delta}_\mathbf{t}) = \frac{L(\theta|\mathbf{U}, \boldsymbol{\delta}) q(\boldsymbol{\theta})}{\int L(\theta|\mathbf{U}, \boldsymbol{\delta}) q(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}}, \tag{2.36}$$

と得られる。以上により, Jensen 不等式から対数尤度の下限値を代数的に表現することができた。式 (2.35) の最右辺を変分下限と呼ぶこともある。この下限値の式を$\boldsymbol{\delta}_\mathbf{t}$を所与としたときの$\boldsymbol{\delta}$の式$F(\boldsymbol{\delta}|\boldsymbol{\delta}_t)$とおいて，対数尤度$\ln L(\boldsymbol{\delta}|\mathbf{U})$と$F(\boldsymbol{\delta}|\boldsymbol{\delta}_t)$の差をとると，

$$\ln L(\boldsymbol{\delta}|\mathbf{U}) - F(\boldsymbol{\delta}|\boldsymbol{\delta}_t) = \ln L(\boldsymbol{\delta}|\mathbf{U}) - \int g(\boldsymbol{\theta}|\mathbf{U}, \boldsymbol{\delta}_\mathbf{t}) \ln \frac{L(\boldsymbol{\theta}|\mathbf{U}, \boldsymbol{\delta})}{g(\boldsymbol{\theta}|\mathbf{U}, \boldsymbol{\delta}_\mathbf{t})} \mathrm{d}\boldsymbol{\theta}, \tag{2.37}$$

となり，右辺第一項に$\int g(\boldsymbol{\theta}|\mathbf{U}, \boldsymbol{\delta}_\mathbf{t}) \mathrm{d}\boldsymbol{\theta} = 1$（確率分布なので）を乗じ，第二項の対数の分子を（尤度だが数式としては確率$p(\mathbf{U}, \boldsymbol{\theta}|\boldsymbol{\delta})$ と同じなので）確率の乗法定理により分解すると式 (2.37) は，

$$= \ln L(\boldsymbol{\delta}|\mathbf{U}) \int g(\boldsymbol{\theta}|\mathbf{U}, \boldsymbol{\delta}_\mathbf{t}) \mathrm{d}\boldsymbol{\theta} - \int g(\boldsymbol{\theta}|\mathbf{U}, \boldsymbol{\delta}_\mathbf{t}) \ln \frac{L(\boldsymbol{\theta}|\mathbf{U}, \boldsymbol{\delta}) L(\boldsymbol{\delta}|\mathbf{U})}{g(\boldsymbol{\theta}|\mathbf{U}, \boldsymbol{\delta}_\mathbf{t})} \mathrm{d}\boldsymbol{\theta}, \tag{2.38}$$

となるが，第一項の$\ln L(\boldsymbol{\delta}|\mathbf{U})$は$\boldsymbol{\theta}$を含まない関数であるため積分の中に戻してもよい。さらに第二項の対数記号を分数中のそれぞれの関数に分配すると，さらに，

$$\int g(\boldsymbol{\theta}|\mathbf{U}, \boldsymbol{\delta}_\mathbf{t}) \ln L(\boldsymbol{\delta}|\mathbf{U}) \mathrm{d}\boldsymbol{\theta} - \int g(\boldsymbol{\theta}|\mathbf{U}, \boldsymbol{\delta}_\mathbf{t}) \left[\ln L(\boldsymbol{\theta}|\mathbf{U}, \boldsymbol{\delta}) + \ln L(\boldsymbol{\delta}|\mathbf{U}) - \ln g(\boldsymbol{\theta}|\mathbf{U}, \boldsymbol{\delta}_\mathbf{t}) \right] \mathrm{d}\boldsymbol{\theta}, \tag{2.39}$$

と変形される。ここで積分記号の中をまとめると式 (2.39) は,

$$\int g(\boldsymbol{\theta}|\mathbf{U},\boldsymbol{\delta}_\mathbf{t})\,[\ln L(\boldsymbol{\delta}|\mathbf{U}) - \{\ln L(\boldsymbol{\theta}|\mathbf{U},\boldsymbol{\delta}) + \ln L(\boldsymbol{\delta}|\mathbf{U}) - \ln g(\boldsymbol{\theta}|\mathbf{U},\boldsymbol{\delta}_\mathbf{t})\}]\mathrm{d}\boldsymbol{\theta}$$
$$= \int g(\boldsymbol{\theta}|\mathbf{U},\boldsymbol{\delta}_\mathbf{t})\,[\ln L(\boldsymbol{\theta}|\mathbf{U},\boldsymbol{\delta}) - \ln g(\boldsymbol{\theta}|\mathbf{U},\boldsymbol{\delta}_\mathbf{t})]\mathrm{d}\boldsymbol{\theta}, \tag{2.40}$$

と変形することができる。ここで確率分布の差異についての指標である KL ダイバージェンス,

$$\mathrm{KL}(P,Q) = \int P(\boldsymbol{\theta}) \ln \frac{P(\boldsymbol{\theta})}{Q(\boldsymbol{\theta})}\mathrm{d}\boldsymbol{\theta}, \tag{2.41}$$

を利用して KL 項で式 (2.40) を置き換えると式 (2.37) は,

$$\ln L(\boldsymbol{\delta}|\mathbf{U}) - F(\boldsymbol{\delta}|\boldsymbol{\delta}_t) = \mathrm{KL}\big(g(\boldsymbol{\theta}|\mathbf{U},\boldsymbol{\delta}_\mathbf{t}), L(\boldsymbol{\theta}|\mathbf{U},\boldsymbol{\delta})\big), \tag{2.42}$$

となる。さらに変形すると,

$$\ln L(\boldsymbol{\delta}|\mathbf{U}) = F(\boldsymbol{\delta}|\boldsymbol{\delta}_t) + \mathrm{KL}\big(g(\boldsymbol{\theta}|\mathbf{U},\boldsymbol{\delta}_\mathbf{t}), L(\boldsymbol{\theta}|\mathbf{U},\boldsymbol{\delta})\big), \tag{2.43}$$

が得られる。このとき式 (2.43) で$\boldsymbol{\delta} = \boldsymbol{\delta}_t$とおいたものとの差をとると，$\boldsymbol{\delta} = \boldsymbol{\delta}_t$のときの KL 項は $\log 1 = 0$より消去され,

$$\ln L(\boldsymbol{\delta}|\mathbf{U}) - \ln L(\boldsymbol{\delta}_t|\mathbf{U}) = F(\boldsymbol{\delta}|\boldsymbol{\delta}_t) - F(\boldsymbol{\delta}_t|\boldsymbol{\delta}_t) + \mathrm{KL}\big(g(\boldsymbol{\theta}|\mathbf{U},\boldsymbol{\delta}_\mathbf{t}), L(\boldsymbol{\theta}|\mathbf{U},\boldsymbol{\delta})\big), \tag{2.44}$$

となるが，KL 項は非負であるので,

$$F(\boldsymbol{\delta}|\boldsymbol{\delta}_t) \geq F(\boldsymbol{\delta}_t|\boldsymbol{\delta}_t)$$

という条件さえ成り立てば,

$$\ln L(\boldsymbol{\delta}|\mathbf{U}) \geq \ln L(\boldsymbol{\delta}_t|\mathbf{U})$$

が保証される。ここで式 (2.33) および式 (2.34) より変分下限$F(\boldsymbol{\delta}|\boldsymbol{\delta}_t)$は,

$$F(\boldsymbol{\delta}|\boldsymbol{\delta}_t) = \int g(\boldsymbol{\theta}|\mathbf{U},\boldsymbol{\delta}_\mathbf{t}) \ln L(\boldsymbol{\delta},\boldsymbol{\theta}|\mathbf{U})\mathrm{d}\boldsymbol{\theta} - \int g(\boldsymbol{\theta}|\mathbf{U},\boldsymbol{\delta}_\mathbf{t}) \ln g(\boldsymbol{\theta}|\mathbf{U},\boldsymbol{\delta}_\mathbf{t})\mathrm{d}\boldsymbol{\theta}, \tag{2.45}$$

と変形される。第一項は式 (2.33) の Q 関数そのものであり，第二項はパラメタ$\boldsymbol{\delta}$とは無関係の項であるため, この式により Q 関数の最大化が$F(\boldsymbol{\delta}|\boldsymbol{\delta_t})$の最大化と等価であることが示せたため, Q 関数の最大化をすることで対数尤度$\ln L(\boldsymbol{\delta}|\mathbf{U})$も最大化できることが証明できた。

EM アルゴリズムは E ステップと M ステップを，適当な収束基準に達するまで繰り返す推定アルゴリズムである。一般にニュートン・ラフソン法などの計算方法に比べて収束に至るまでの反復回数は非常に多いため，いくつかの高速化の手法が提案されている (たとえば Ramsay (1975) など)。

2.3.4 Bock & Aitkin による解法

これを IRT の項目パラメタ推定に置き換えて説明すると, いま完全データ対数尤度関数 (Q 関数) は，式 (2.30) の積分の対象となっている尤度関数の対数をとり，さらにそこから全受検者について和をとったもの,

$$\ln L_C(\boldsymbol{\delta}|\mathbf{U},\boldsymbol{\theta}) = \sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} \ln L_C\left(\boldsymbol{\delta_j}\middle|\mathbf{u_i},\theta_i\right), \tag{2.46}$$

ただし,

$$\ln L_C\left(\boldsymbol{\delta_j}\middle|\mathbf{u_i},\theta_i\right) = \sum_{\boldsymbol{j=1}}^{\boldsymbol{J}} u_{ij}\ln P_j(\theta_i) + \left(1-u_{ij}\right)\ln Q_j(\theta_i), \tag{2.47}$$

である。$\mathbf{u_i}$は受検者一人分の項目反応データであり，データ行列$\mathbf{U}$の行ベクトルの要素にあたる潜在変数θ_iの事後分布$g(\theta_i|\mathbf{u_i},\boldsymbol{\delta_0})$は, θにサブスクリプトがここから追加されることに注意して,

$$g(\theta_i|\mathbf{u_i},\boldsymbol{\delta_0}) = \frac{L(\boldsymbol{\delta_0},\theta_i|\mathbf{u}_i)h(\theta_i)}{\int L(\boldsymbol{\delta_0},\theta_i|\mathbf{u}_i)h(\theta_i)\mathrm{d}\theta_i}, \tag{2.48}$$

と表現される。なお, 分子は適当な求積点X_qと求積点に対応する重み$\phi_q = h(X_q)$で離散近似され,

$$g(\theta_i|\mathbf{u_i},\boldsymbol{\delta_0}) \approx \frac{L(\boldsymbol{\delta_0},\theta_i|\mathbf{u}_i)h(\theta_i)}{\sum_q L\left(\boldsymbol{\delta_0},X_q\middle|\mathbf{u}_i\right)\phi_q}, \tag{2.49}$$

これらを式 (2.33) に代入すると Q 関数は, 期待値計算は受検者ひとりずつおこなわれることと, パラメタ$\boldsymbol{\delta}$と$\boldsymbol{\delta_0}$の違いに注意して,

$$\mathrm{E}[\ln L_{\boldsymbol{C}}(\boldsymbol{\delta}|\mathbf{U},\boldsymbol{\theta})|\mathbf{u};\boldsymbol{\delta_0}] = \sum_{i=1}^{N}\int_{-\infty}^{\infty} \ln L_C\left(\boldsymbol{\delta_j}\middle|\mathbf{u_i},\theta_i\right)g(\theta_i|\mathbf{u_i},\boldsymbol{\delta_0})\mathrm{d}\theta, \tag{2.50}$$

となる。これが IRT モデルの項目パラメタ推定における Q 関数になる。なお，積分計算はこれまで同様に区分求積法で離散近似すればよい。

Bock & Aitkin (1980) はこの Q 関数を取り扱いやすいように式変形している。具体的には，式 (2.50) における$g(\theta_i|\mathbf{u_i},\boldsymbol{\delta_0})$の離散近似に用いる求積点を$Y_r$，それに対応する重みを$W_r = g(Y_r)$とすると，受検者一人あたりの事後分布の求積点$r$の重みは

$$G_{ir} = \frac{L(\boldsymbol{\delta_0}, Y_r|\mathbf{u}_i)W_r}{\sum_q L(\boldsymbol{\delta_0}, X_q|\mathbf{u}_i)\phi_q}, \tag{2.51}$$

であるので，これにより式 (2.50) の離散近似は以下のように実行される。

$$\mathrm{E}[\ln L_{\boldsymbol{C}}(\boldsymbol{\delta}|\mathbf{U},\boldsymbol{\theta})|\mathbf{U};\boldsymbol{\delta_0}] \approx \sum_{i=1}^{N}\sum_{r=1}^{R}\left[\ln L\left(\boldsymbol{\delta_j}, Y_r\middle|\mathbf{u_i}\right) + \ln W_r\right]G_{ir}\,, \tag{2.52}$$

さらにここから M ステップで不要となるパラメタ$\boldsymbol{\delta}$とは無関係な項を除いて，和をとる順番を項目＞離散近似＞受検者という順に整理し直すと，

$$= \sum_{j=1}^{J}\sum_{r=1}^{R}\sum_{i=1}^{N}\left[u_{ij}\ln P_j(Y_r) + \left(1-u_{ij}\right)\ln Q_j(Y_r)\right]G_{ir}, \tag{2.53}$$

となる。ここで受検者の事後分布の重みであるG_{ir}について考えると，確率密度であるので分点についての和をとれば 1 になるが，受検者について和をとると，その分点の能力値である受検者の期待度数，

$$\widehat{N_r} = \sum_i G_{ir}\,, \tag{2.54}$$

となり，また，これに受検者の項目反応パタンu_{ij}を書けたものの，同じく受検者についての和は，その分点にいる受検者のうち，その項目に正答する受検者の期待度数，

$$\widehat{r_{jr}} = \sum_i u_{ij}G_{ir}\,, \tag{2.55}$$

となる。これを，式 (2.53) を展開した形に代入して整理すると，最終的に，

$$\mathrm{E}[\ln L_{\boldsymbol{C}}(\boldsymbol{\delta}|\mathbf{U},\boldsymbol{\theta})|\mathbf{U};\boldsymbol{\delta_0}] = \sum_{j=1}^{J}\sum_{r=1}^{R}\left[\widehat{r_{jr}}\ln P_j(Y_r) + \left(\widehat{N_r} - \widehat{r_{jr}}\right)\ln Q_j(Y_r)\right], \tag{2.56}$$

という形で E ステップの計算をおこなうことができる。E ステップの形式を 2.3.3 で説明した展開ではなく，求積点の期待受検者度数と期待正答受検者度数の計算に実質的に置き換えたのは Bock & Aitkin (1981) であり，IRT における MMLE-EM 法の普及は Bock と Aitkin の功績によるところが大きい (村木, 2011)。

M ステップでは$\mathrm{E}[\ln L_{\boldsymbol{C}}(\boldsymbol{\delta}|\mathbf{U},\boldsymbol{\theta})|\mathbf{U};\boldsymbol{\delta_0}]$を$\boldsymbol{\delta}$で偏微分したものを＝ 0 とおいて極大点を求めれば良い。多変数の最適化のためニュートン・ラフソン法およびフィッシャースコアリングにはヘッセ行列か情報行列が必要となる。M ステップで必要となる勾配ベクトルと行列の具体的な形については Baker & Kim (2004), 加藤ら (2014) や豊田 (2005) などに詳しい。

MMLE-EM 法においても JMLE と同様に尺度の原点と単位を固定しなくては収束を判定することが困難になる。いくつかの方法があるが，最もポピュラーな方法は事前分布の平均と分散を特定の値に固定してしまう方法である。

2.3.5 階層ベイズ推定法

JMLE や MMLE では全受検者が正答，もしくは誤答した項目のパラメタを推定できない。また，ときおり推定値が異常に高い，あるいは低い値になるケースがある。最尤推定の原理ではパラメタの推定の標準誤差を計算することができるが，漸近的な仮定をおく必要があるため，標本数が少ない場合などに適用することが望ましくない。そこで HBE 推定法を使用することで，これらの問題に対処できる。

ベイズ統計学の枠組みでは母数にも事前分布を仮定することで，母数の事後分布を推定することが可能である。そして事後分布のベイズ確信区間を推定すればより直感的なパラメタの推定精度を評価する事も可能である。確率変数Xの確率分布を$p(X)$と表す場合，項目パラメタの推定に必要な同時事後分布 (joint posterior distribution) は，ベイズの定理と$L(\boldsymbol{\theta},\boldsymbol{\delta}|\mathbf{U}) = p(\mathbf{U}|\boldsymbol{\theta},\boldsymbol{\delta})$より，

$$p(\boldsymbol{\theta},\boldsymbol{\delta}|\mathbf{U}) = \frac{p(\mathbf{U}|\boldsymbol{\theta},\boldsymbol{\delta})p(\boldsymbol{\theta},\boldsymbol{\delta})}{p(\mathbf{U})} \propto L(\boldsymbol{\theta},\boldsymbol{\delta}|\mathbf{U})p(\boldsymbol{\theta},\boldsymbol{\delta}), \tag{2.57}$$

となる。各パラメタの MAP 推定値を求めるだけであれば$p(\mathbf{U})$を計算する必要はなく，最右辺だけでよい。すべてのパラメタの同時分布である$p(\boldsymbol{\theta},\boldsymbol{\delta})$は，パラメタ間が互いに独立であると仮定すると，

$$p(\boldsymbol{\theta},\boldsymbol{\delta}) = \prod_{i=1}^{N} p(\theta_i)\prod_{j=1}^{J} p(a_j)p(b_j)p(c_j), \tag{2.58}$$

と表現できる。式 (2.57) の対数をとり，これまで同様に適当な数値最適化の手法を用いれば各パラメタの MAP 推定値を計算できる。多変数の偏微分の導出が困難な場合には数値微分を用いるか，勾配の計算を必要としない最適化の手法である Nelder-Mead 法 (Nelder and Mead, 1965) を用いる。このようにパラメタの事前分布を階層的に設定する方法が階層ベイズ推定法である。HBE に必要な事前分布の設定としては識別力に対数正規分布や切断正規分布，困難度に正規分布，当て推量にはベータ分布などが提案されている (Swaminathan & Gifford, 1986)。事前分布の設定やその違いが推定結果に及ぼす影響については Fox (2010) が広範にレビューしている。

2.3.6 周辺ベイズ推定法

HBE では同時事後分布を最大化していた。同時事後分布をそのまま最大化することは同時最尤推定法と同様に，一致性の問題を生じさせてしまう。そこで Mislevy (1986) が提案するように，θについて周辺化した周辺事後分布を最大化することで各項目パラメタの MAP 推定値を得る方法を説明する。この方法は MBE と呼ばれ，最大化したい周辺事後分布は，

$$p(\boldsymbol{\delta}|\mathbf{U}) = \int L(\boldsymbol{\theta}, \boldsymbol{\delta}|\mathbf{U}) p(\boldsymbol{\theta}) p(\boldsymbol{\delta}) \mathrm{d}\boldsymbol{\theta}\,, \tag{2.59}$$

である。この方法では MMLE-EM の M ステップにおける目的関数に，各項目パラメタの対数事前分布を加えるだけでよい。なぜなら E ステップは潜在変数θについてのみ期待値をとっているため，項目パラメタの事前分布には影響しないからである。M ステップの数値最適化を実行するためには対数事後分布の一階偏微分，あるいは二階偏微分が必要になる。一部の分布の導関数については豊田 (2005) が詳しく展開をおこなっている。MBE では，階層ベイズ推定法と同様のメリットがある。例えば標本数が少ない場合であっても，適切な事前分布を設定することで安定したパラメタ推定を実行できる。

2.3.7 多母集団推定

項目パラメタを推定するために集められた被検者が複数の異なる母集団からのサンプリングを想定し，さらに集団ごとに部分的に異なる項目に回答するようなケースを想定する。例えば学力が高い群が一部の難しい項目に回答し，低い群はその項目には回答せず，易しい項目だけに回答するようなケースである。項目パラメタの推定には MMLE-EM 法を用いれば良いが，周辺化する際の事前分布は，母集団ごとに定める必要がある。この項目パラメタの推定方法は多母集団（他群）モデルを扱う推定方法として知られており，この推定方法に対応しているプログラムとしては，BILOG-MG や Easy Estimation, lazy.irtx などがある。

いま母集団についての変数$g = \{1, 2, \ldots, G\}$を導入し，さらにその集団がどの項目を受検しているのかを判断するためにデザインマトリックス (design matrix) を導入する。デザインマトリッ

クスは母集団以外にも欠測値に対応した推定に用いることができる。デザインマトリックスの具体的な内容は集団×受検者×項目という 3 次元の配列 (array) であり，

$$d_{gij} = \begin{cases} 1, & \text{母集団}g\text{に属する受検者}i\text{の項目}j\text{への反応が観測される場合} \\ 0, & \text{母集団}g\text{に属する受検者}i\text{の項目}j\text{への反応が欠測している場合} \end{cases}$$

という 3 次元配列である。デザイン行列を**D**とおくと，対数尤度関数は，

$$\ln L(\boldsymbol{\delta}, \boldsymbol{\theta}|\mathbf{U}, \mathbf{D}) = \sum_{g=1}^{G}\sum_{i=1}^{N}\sum_{j=1}^{J} u_{ij}d_{gij}\ln P_j(\theta_i) + \left(1 - u_{ij}\right)d_{gij}\ln Q_j(\theta_i), \tag{2.60}$$

と書き換えられる。これにより受検者が回答していない項目における尤度を正しく欠測として処理できる。さらに MMLE-EM の E ステップにおける計算も，母集団ごとに異なるパラメタの事前分布を設定する必要がある。平均と標準偏差によって決定される事前分布を仮定したとき，その事前分布は，

$$h\left(\theta|\mu_g, \sigma_g\right), \tag{2.61}$$

とおくものとする。すなわち E ステップの期待対数完全データ尤度関数は，

$$\mathrm{E}[\ln L_{\boldsymbol{C}}(\boldsymbol{\delta}|\mathbf{U}, \mathbf{D}, \boldsymbol{\theta})|\mathbf{U}, \mathbf{D}; \boldsymbol{\delta_0}] = \sum_{g=1}^{G}\sum_{i=1}^{N}\int_{-\infty}^{\infty} \ln L_C\left(\boldsymbol{\delta_j}|\mathbf{u_i}, \boldsymbol{d_{gi}}, \theta_i\right)g\left(\theta_i|\mathbf{u_i}, \boldsymbol{\delta_0}, \mu_g, \sigma_g\right)\mathrm{d}\theta, \tag{2.62}$$

と書き換えられ，最終的に

$$\mathrm{E}[\ln L_{\boldsymbol{C}}(\boldsymbol{\delta}|\mathbf{U}, \mathbf{D}, \boldsymbol{\theta})|\mathbf{U}, \mathbf{D}; \boldsymbol{\delta_0}] = \sum_{g=1}^{G}\sum_{j=1}^{J}\sum_{r=1}^{R}\left[\widehat{r_{gjr}}\ln P_j\left(Y_{gr}\right) + \left(\widehat{N_{gr}} - \widehat{r_{gjr}}\right)\ln Q_j\left(Y_{gr}\right)\right], \tag{2.63}$$

を計算する。ただし

$$\widehat{N_{gr}} = \sum_i d_{gij}G_{ir}, \tag{2.64}$$

$$\widehat{r_{jr}} = \sum_i u_{ij}d_{gij}G_{ir}, \tag{2.65}$$

である。ここでは E ステップの積分の離散近似の分点は十分に幅の広い分点をとることで全母集団共通のものを使用すると仮定する。

多母集団モデルの項目パラメタ推定では各 EM サイクルが終了した時点で集団ごとの平均と標準偏差を計算し，次回の E ステップで使用する分点の重みを再計算する必要がある。分点の重みを計算する場合，正規分布などの分布族を仮定するほか，分布の関数形を指定せずに式 (2.64) および式 (2.65) により計算される受検者の分点ごとのヒストグラムを規格化し，そこから推定される多項分布を使用することもできる (前川, 1991)。

多母集団モデルの推定においても項目パラメタの推定値が発散するのを避けるため，尺度の単位と原点を固定しなくてはならない。ただしこの場合，いずれかの母集団の平均と標準偏差を EM サイクルの更新のたびに固定し続け，さらに他の母集団の事前分布のパラメタも尺度の不定性を利用して線形変換する必要がある。尺度の不定性については後述する。

3　垂直尺度化（Vertical Scaling）

これまでは尺度に用いられる心理計量モデルについて紹介し，一部のモデルのパラメタの推定方法の数理的手法について述べてきた。潜在的な特性の測定が一度きりで，測定したい能力の範囲が比較的限定されているのであれば，ひとつのテストを実施し，そのパラメタを推定するだけで十分である。しかし能力の変化を測定するためには複数のテストの実施が必要で，垂直尺度のように複数の学年をまたぐような尺度を構成する場合には，異なるテスト得点を比較可能なものとし，単一の尺度としてまとめ上げる必要がある。特に垂直尺度を構成する為の手法を垂直尺度化 (vertical scaling)と呼称する。しかし本来，尺度化 (scaling) というものが学力に限らず，広く「能力，信念，嗜好，感覚など，物理，科学器計で直接計測できないものを数量として表そうという試みを指す (印東, 1995，p. 135)」ものだとすれば，垂直尺度化は尺度の構成方法のひとつでもあり，さらにテスト得点の対応づけの手法のひとつでもあるといえる。

垂直尺度化とは一言で述べれば，同じ構成概念を測定しているがテストの難易度が異なるような複数のテストスコアを共通尺度上に位置づける手法である。この手法は，例えば小学校や中学校で扱う国語や算数・数学のように，連続する学年で学ばれるものの，評価は学期や学年ごとで断続的に実施されているために，児童・生徒の学力の伸びを適切に評価することが困難であるという問題にひとつの解決策を与えることができる。あるいは，垂直尺度にしたがって学力の縦断的な変化を測定すれば，順位や平均点といった相対的で個人内の比較が難しい指標に依らずに評価可能である。

垂直尺度化のアイディア自体は 20 世紀中頃には既に存在するものの，時代や地域によって微妙に異なる定義がなされているために，周辺の概念と明確な区別がなされてこなかった。ここでは，まず垂直尺度化と密接な関わりを持つ対応づけ (linking) や等化 (equating) といった概念との比較を通して，垂直尺度化とは何であるのかを定義することを試みる。

次に，基本的な垂直尺度化のための手順を考える。垂直尺度化に限らず異なるテストのテスト得点を比較するためには何らかの共通情報を直接的，あるいは間接的に設ける必要がある。そしてデータ収集デザインとならんで重要であるのがテストデザインである。データ収集デザインがテスト全体についてのフレームを決めるとすれば，テストデザインは共通情報を持つふたつのテスト間に，どのように共通情報を配置するかについて決定する。そのためのデータ収集デザイン，テストデザインおよびそれらの長所・短所についてまとめ，さらに尺度調整 (calibration) 方法についても議論する。尺度調整方法はデータ収集デザインと関連して選択されるが，方法によって推定値が変化し，計算にかかるコストも違う。いくつかの尺度調整法の比較および，先行研究から得られた知見をまとめ，最後にこれまでの研究で指摘されている垂直尺度化の問題点や課題を指摘する。

3.1　基本的な概念

3.1.1　対応づけ

垂直尺度化が何であって，何ではないのかを明確にするために，類似する周辺手法について整理する必要がある。垂直尺度化は，対応づけ (linking) と呼ばれる異なるテストの得点を交換 (exchange) あるいは比較可能 (comparable) にするための手法の下位概念として理解されている。ところで linking の邦訳については「対応づけ」のほかに，リンキングや連結などと訳されることもあった (柴山・佐藤, 2008; 熊谷・荘島, 2015)。

対応づけに関しては国や時代によって複数の，微妙に異なる定義がなされている (Feuer, Holland, Green, Bertenthal, & Cadelle Hemphill, 1999; Holland & Dorans, 2006; Kolen, 2004a; Kolen, 2004b; M. J. Kolen & Brennan, 2004; Kolen & Brennan, 2014; Linn, 1993; Mislevy, 1992; Newton, 2010a; Newton, 2010b; von Davier, 2010)。たとえば Mislevy (1992) はテストをリンク (link) する方法を Equating, Calibration, Projection, Statistical Moderation, Social Moderation の 5 つに分類し，それぞれの特徴 (descriptiozn) や手続き，実例について検討している。これらの分類は結果として得られる対応づけスコアの結びつきの強さの順に列挙されている。言い換えれば，テスト X のテスト Y における対応づけ得点が，本来のテスト Y についてどれほど確からしい推論をできるかどうかで分類している。この場合，最も結びつきが強いものが等化 (equating) にあたる。Linn (1993) も名称はやや異なるがほぼ同様の分類を提示している。さらに Holland & Dorans (2006) はこれらを対応づけの下位概念として再分類し，細かく定義づけしている。Holland & Dorans (2006) の提案する対応づけの下位概念には垂直尺度化以外にも等化や calibration, concordance など複数の下位分類が存在する。

Feuer et al. (1999) の研究では Mislevy や Linn の研究も踏まえつつ，これまでの NAEP (National Assessment of Educational Progress, 全米学力調査) とそのほかのテスト得点の対応づけについての事例を数多く収集し，結論として NAEP とそのほかの個人スコアを報告できるような商用，あるいは州別のテストを対応づけすることは困難であると述べている。つまり対応づけは統計的に分析できれば何でもありの万能な手法ではなく，意味のある得点の比較のためには，テストの内容や仕様，信頼性などの様々な特徴を考慮する必要があるということである。Feuer et al. (1999) は対応づけ得点そのものというよりも，テスト実施にまつわる様々な特徴に注目して分類していると言える。

今日最も有名な対応づけについての概念の整理をおこなっているのは Kolen & Brennan (2014) であろう。彼らはこれまでの分類とは異なる枠組みとして Degree of Similarity を主張している。母集団やテストの構成概念の類似度によって，対応づけの手法を分類すべきという主張である。もっとも，これまで言われてきた Mislevy (1992)や Linn (1993) ，Holland & Dorans (2006) などの分類と決定的な違いは存在しない。注目する点がテスト得点の連結の強度か，得点そのものではなく構成概念や母集団の類似性に変わっているだけとも言える。

Newton (2010a, 2010b)は Phenomenal, Causal, Predictive という三つの観点で比較し，対応づけを

分類すべきだと述べている。Phenomenal は異なるテストの習熟に関する基準についてであり，従来の構成概念に近い。Causal はそのテスト得点への寄与（例えば生徒が受ける授業やカリキュラム）が同一，あるいは類似しているという観点であり，従来の母集団の等質性に近いアイディアである。Predictive の考え方は対応づけされた得点が予測する将来の達成の程度が類似するかどうかである。Newton のフレームワークはこれまでの対応づけの概念に時間的な広がりを追加し，テスト得点に関して将来や過去の情報も比較することで，対応づけの方法を分類可能なものとしている。

これまでの対応づけの概念分類をまとめると，アメリカで一般的に議論が進んできた等化を頂点とした枠組み (Holland & Dorans, 2006 や Kolen & Brennan, 2014 など) と，イギリスを中心とした比較可能性の 3 次元にグループ化するもの (Newton, 2010a; 2010b) が存在している。ちなみに，わが国では前者の枠組みで理解されることが多く，ほとんどの研究では対応づけを，厳密でない等化，あるいは等化の条件を緩めたものとして考えている (たとえば，日本テスト学会, 2007; 柴山・野口, 2004; 石井・安永, 2011)。近年では対応づけの概念がより明確に認識されるようになってきており，今までは等化の一部として考えられていた研究も，対応づけの下位概念として等化と区別されるようになってきている。しかし，柴山・野口 (2004) が等パーセンタイル等化法を援用することによって異なるテストを concordance する手法について説明しているものの，未だ等化以外の対応づけ（たとえば垂直尺度化）の技術に関しての研究は盛んではない。ところで，特に IRT の項目パラメタと対応づけの一連の過程を尺度調整 (calibration) とまとめて表現することもあるので，対応づけの下位概念としての calibration と，パラメタ推定＋対応づけとしての calibration の表現の重複には注意が必要である。

最後に，これまでに説明した対応づけの下位分類では数値的な指標で対応づけを評価する方法は取り入られてこなかった。その中でも Dorans (2000) や佐藤 & 柴山 (2014), Sato & Shibayama (2018)の研究は，古典的テスト理論の範疇で異なるふたつのテストの対応づけ実行可能性 (linkability) について検討するための指標を提案している数少ない研究例である。これらの指標はテスト間の相関係数や信頼性係数の推定値を用いて計算されるが，使用可能なテストデザインや等化手法が限定される。

3.1.2 等化

対応づけの下位分類の中心的なアイディアが等化 (equating) である。等化とは二つのテストが同一構成概念，難易度で，信頼性が等しく，受検者の母集団が等しい場合のテスト得点の対応づけである (野口・大隅, 2014)。テスト得点の特徴としては左右対称性，交換可能性というものが挙げられる。つまり，テスト X とテスト Y を何らかの手法で等化した場合に，テスト X の 50 点はテスト Y の 60 点相当であるという結果が得られたとすれば，どちらのテストを受けても，必ずもう一方でも等化得点と同等のスコアが保証されるということである。等化をおこなうテストとして有名なものに TOEFL がある。TOEFL では異なる受検時期のテストはすべて等化され，そのスコアは同一尺度上で交換可能 (exchangeable) になる。したがってテスト得点が保証さ

れている期間内であれば、その得点をいつでも最新の TOEFL の得点と等価 (equivalent) なものとして扱うことができる。

先ほど説明した Holland & Dorans (2006) の分類によればテスト等化の条件は，異なるテストが (a) 同一の構成概念を測定していること，(b) 同一の信頼性を持っていること，(c)左右対称であること，(d) 等質であること，(e) 対象とする母集団が不変であることの 5 つである。Kolen & Brennan (2014) の定義に則れば, 推論 (inference) , 構成概念 (constructs) , 母集団 (populations) , 測定の特徴 (measurement characteristic / condition) ， がすべて等しいことである。

IRT でテストを分析するモチベーションのひとつがこの等化を実行することである。等化の手続き自体は，必要な条件を満たせば素点ベースでもおこなうことができる。たとえば平均等化法 (mean equating) や線形等化法 (liner equating) ， 等パーセンタイル等化法 (equipercentile equating) などが素点の等化法の代表的な手段として挙げられる。IRT でおこなう等化のメリットのひとつは，項目のパラメタと受検者の得点である能力パラメタを分離してモデル化されており，互いに依存せずに推定できるため，テスト同士を母集団に依存せずに等化できることである。

3.1.3 尺度の不定性

尺度の不定性 (Scale Invariance) は IRT のパラメタ推定のときにすこし触れたが，尺度の平均と分散を特定の値（たとえば 0 と 1）に固定しなければ解が収束しないという問題に関わっていた。尺度の不定性とはつまり測定したいもの（潜在変数）に対して，観測された得点（顕在変数）があるとして，変数間の関係が尺度によって一様でないことを指す。そもそも個人の能力や特性というものには目に見える大きさがないのだから，尺度か測定対象のどちらかの基準や大きさを仮定しなくては測定値を得られないというのは当然の話である。尺度の不定性の議論については Blanton & Jaccard (2006) が arbitrary metrics (任意尺度) という語で論じているほか，尺度の妥当性と関連して村山 (2012) も議論している。

一般に 2PLM では次のように尺度の単位を定数倍したり，原点に定数を加えたりしても正答確率そのものは変化しない。つまり，

$$P(\theta|a,b) = \frac{1}{1+\exp\big(a(\theta-b)\big)}, \tag{3.1}$$

であるとき，A倍してKを足すという操作をθにおこない，θ^*を得るとすると，

$$\theta^* = A\theta + K, \tag{3.2}$$

となるが，これをθについて解いて，式 (3.1) に代入すると，

$$P(\theta|a,b)=\frac{1}{1+\exp\left(a\left(\frac{1}{A}\theta^{*}-\frac{K}{A}-b\right)\right)}=\frac{1}{1+\exp\left(\frac{a}{A}\left(\theta^{*}-(Ab+K)\right)\right)}=P\left(\theta|\frac{a}{A},Ab+K\right),\quad(3.3)$$

となり，項目パラメタの尺度と一緒に変換されることで，正答確率は同一であることが分かる。このときの係数AとKを，特に等化・対応づけの文脈で等化係数 (equating coefficient) と呼ぶ。

3 PLM における等化係数の推定はやや複雑である。なぜならcパラメタが違うと，aやbパラメタの本質的な意味も変化してしまうためである。3 PLM の等化係数を推定する際には，どちらか一方のcパラメタに固定して推定をおこなうか，両尺度のcの平均をとるなどすることがある (Han et al., 2015)。

等化係数の推定方法には複数の手法がある (Kolen & Brennan, 2014) 。もっとも単純な方法が項目困難度パラメタか能力パラメタの，平均と標準偏差を用いる Mean & Sigma 法 (Marco, 1977) である。いま，等化先のパラメタをT，等化元のパラメタをFの添え字で表すとする。このとき等化先の困難度パラメタの推定値の平均と標準偏差μ_{bT}とσ_{bT}，等化元の困難度パラメタの推定値の平均と標準偏差μ_{bF}とσ_{bF}を用いて，

$$\hat{A}=\frac{\sigma_{bT}}{\sigma_{bF}},\quad(3.4)$$

$$\hat{K}=\mu_{bT}-\hat{A}\mu_{bF},\quad(3.5)$$

という様に等化係数の推定値を求める。この$\hat{A}$の値に識別力パラメタの平均値を用いるのが Loyd & Hoover (1980) の Mean & Mean 法である。この方法による等化係数の$\hat{A}$は，

$$\hat{A}=\frac{\bar{a}_{F}}{\bar{a}_{T}},\quad(3.6)$$

と定義される。さらに，この方法の類似した手法として，識別力パラメタの算術平均ではなく，幾何平均を用いる Mean & Geometric Mean 方法も存在し，この方法による等化係数の$\hat{A}$は，

$$\hat{A}=\frac{\bar{\bar{a}}_{F}}{\bar{\bar{a}}_{T}},\quad(3.7)$$

ただし，

$$\bar{\bar{a}}=\sqrt[J]{\prod_{j=1}^{J}a_{j}},\quad(3.8)$$

で表される幾何平均を用いている。

より数理的に洗練された手法が ICC や TCC を用いて等化係数を推定する手法である (Haebara,

1980; Stocking and Lord, 1983)。Haebara の方法はふたつの尺度の項目反応確率の差を誤差関数と定義し，その誤差関数の全項目の和を最小化するような推定値を等化係数とする手法である。ここでは Haebara (1980) の表記にしたがって項目数を$g = \{1, 2, \ldots, m\}$，受検者数を$a = \{1, 2, \ldots, N\}$とおくこととする。等化先の尺度Tと等化元の尺度Fの誤差と誤差関数は，

$$e_{TF} = P_{g,T}(\theta_{T,a}) - P_{g,F}(\theta_{F,a}), \tag{3.9}$$

$$Q_1 = \int_{-\infty}^{\infty} \sum_{g=1}^{m} \sum_{a=1}^{N} L(e_{TF})\, \mathrm{d}\theta\,, \tag{3.10}$$

である。ただしLは損失 (loss) の頭文字であり，この場合二乗損失関数である。目的関数は適当なθの定義域を設定し，区分求積の要領で計算される。これに加え，逆方向からの等化も考慮した誤差関数Q_2も定義し，これらの和を目的関数とする。この最適化問題を解くためにはガウス・ニュートン法を用いればよいと Haebara (1980) は述べている。ガウス・ニュートン法に関する詳細は割愛するが，ニュートン・ラフソン法におけるヘッセ行列をヤコビ行列とその転置の積で近似した行列を用いる手法である。Stocking-Lord の方法は項目特性曲線ではなくテスト特性曲線を使用して，同様の損失関数を定義する方法である。

結局のところ，能力パラメタか項目パラメタのどちらか一方で等化係数を推定すれば，項目と能力の尺度どちらでも変換できる。しかしこの係数を推定するためには式 (3.2) にあるように，異なる尺度上で等価である得点が必要となる。項目パラメタを推定するときに事前分布の平均を 0，標準偏差 1 に固定すれば，どのような集団の，いかなる尺度も平均 0，標準偏差 1 のスケールになるが，両尺度に共通する情報がなければ式 (3.2) の関係を仮定することができず，等化は成立しない。この共通情報を得るためには，異なるテスト間に共通項目 (common items) を配置するか，共通する母集団 (common subjects) に異なるテストを受検してもらうなどする必要があるが，これらはテストデザインによって決定される。

3.1.4 垂直尺度化の定義

異なる学年に共通の尺度を設けて，学力の伸びや変化を測定しようという試みは 1980 年頃にはすでに始まっていたという(Patz & Yao, 2007)。当時は標準学力テストをサーストンの絶対尺度化法などの方法で垂直尺度化していたが，近年では IRT 研究の発展にともない，後述する IRT にもとづく手法が主流となっている。垂直尺度化は，以前は垂直等化 (vertical equating) という用語で等化の一種として理解されていたものの，最近では信頼性や等質さに厳しい条件を設けている等化と厳密に区別して理解されている (Reckase, 2010)。世界的に垂直尺度化という呼び方が確定したのは約 15 年前である。ERIC(Educational Resources Information Center)での論文検索の結果，vertical equating という単語を含む論文 (たとえば，Camilli, 1999; Lee, 2003 など) が 2003 年以降は確認できないことから，その前後に何らかの決定力のある概念の整理が行われたと推測できる。

わが国では，村木 (2011) と野口・大隅 (2014)が垂直尺度化と呼ぶべきであると提唱する以前に，佐藤・村木 (2008) が等化と区別して垂直尺度化の概念を説明している。しかし，その後の研究である藤森 (2009; 2011) や光永 (2017) などは垂直等化として研究・紹介をおこなっており，未だ国内では垂直尺度化についての理解は十分に深化・統一できていない。今後の研究発展のためにも用語の正確な定着は必須である。

そもそも等化と垂直尺度化が混同，あるいは同一視される問題の根源には，同じ IRT モデル，推定方法，尺度調整方法で実行できるということがある。等化と同様に，垂直尺度の場合であっても異なるテスト間に共通情報を用意することができれば，IRT のパラメタを線形変換することで共通尺度化できるため，基本的に対象とするテストの難易度と受検者のレベルが異なるという点以外で，等化と垂直尺度化に明確な差はないように思われる。しかし尺度得点の解釈において両者には顕著な違いがある。等化後の得点は交換可能で対称性があるのに対し，垂直尺度化された得点は比較可能でしかない。これは、ある学年レベルのテスト得点はその学年の学習内容を強く反映しており、他の学年レベルのテスト得点を厳密に保証するものではないということである。つまり，共通尺度上の異なる難易度のテスト得点が数値上は同じであっても，その得点はわずかに異なる内容を反映しており，測定精度も異なる可能性があるため，完全に等価な得点とは呼べない。しかし，同一個人の異なる時点の得点を比較することは、まさしく学力発達を共通尺度上で表すことに等しい。

最終的に，等化の条件と比較する形で垂直尺度化の条件を定義する。その条件は異なる 2 つのテストが，(a) 類似した構成概念を測定していること，(b) テストの形式や構成が類似していること，(c) 得点は遡及的な比較に限定され，対称ではないこと，(d) 異なる母集団をもつことである，と定義できる。

3.2 垂直尺度化のためのデータ収集デザイン

垂直尺度化では，異なる版のテストのリンキングを行うためのデータ収集デザインをあらかじめ決定しておく必要がある。一般に，垂直尺度化とは単一の年度の複数学年のテストを対応づけして尺度を構成する手法である。しかし，ここでは垂直尺度を構成しながら，同時に対象集団の学力の変化を測定可能なデザインについて説明する。垂直尺度は基本的にひとつの学年の中で完結して作成されるが，垂直尺度間にも共通項目を配置することができれば間接的に特定の集団の学力の変化を共通尺度上で比較することができる。

たとえばある集団 A の学力の変化を縦断的に測定したいときに，中学 1 年生の 12 月にテストを実施し，その一年後の中学 2 年生の 12 月にも同様のテストを実施する状況を考える。共通尺度上で得点を比較するためには，あらかじめ共通尺度化された項目バンクからふたつのテストを作成し，出題するという方法が考えられるが，いまこの項目バンクと共通尺度は得られていないような状況を仮定しているためこの方法は実行不可能である。また同じ集団が一年後に実施するテストに，昨年受けたテストの項目の一部を配置するか，全く同一のテストを受験してもらう方法も考えらえるが，学習効果により正確な学力を測定できないため，この方法も実

行不可能であるとする。そこで，ひとつの年度内の複数学年で垂直尺度構成し，その垂直尺度を維持しながら次年度の測定もおこなうというデザインを考える (O'Neil, 2010; 澁谷・柴山, 2017)。

3.2.1 単一年度の尺度化

主な単一年度向けのデータ収集デザインとしては，隣接する学年のテストに共通した項目を配置する「共通項目デザイン」，1 つの学年をランダムに 2 グループに分け，受験者の学年にあった項目と受験者の 1 つ下の学年の項目をランダムに実施する「等価グループデザイン」，すべての学年で共通した項目を一部に用意した「尺度化テストデザイン」，順序効果を考慮した「均衡型単一グループデザイン (counterbalanced single-group design)」などがある。

「共通項目デザイン (common item design) 」 (図 3.1) は一番実施が容易なデータ収集デザインである。垂直尺度化の文脈では共通項目不等価グループデザイン (common item nonequivalent groups design) などとも呼ばれる。このデザインを用いるならば，隣接した学年の項目の選定には注意しなくてはならない。小学 5 年生と 6 年生の共通項目を用意する際は，未学習の範囲の項目が出題されないように小学 5 年生の範囲から項目を選択するべきである。このデザインはテスト項目の配置によって項目反応が影響される可能性が高いというデメリットを持つ。たとえば一般的なテストでは最初は易しい問題からはじまり，後半の方に難しい問題が配置されるが，テスト後半ともなると児童・生徒の疲労も蓄積されてくるため，前半か後半かといったテストの配置の問題は項目反応や正答率に大きく影響する。そして共通項目はたいていの場合，隣接した学年の上の学年にとっては易しい項目であるため前半に配置され，下の学年では難しい項目のため後半に配置される可能性が高い。そのため，学年によって共通項目の項目反応が異なってくる可能性に注意しなくてはならない。

「等価グループデザイン (equivalent group design) 」 (図 3.2) は 1 つの学年の生徒らを 2 つのグループにランダムに振り分ける必要がある。しかもグループによって実施されるテストの項目が違う上に，片方のグループでは自分よりも 1 つ下の学年の内容のテストが実施されることとなる。そのため生徒間でのテストを受けた際の実感やテスト結果のフィードバックに不平等が生じる可能性があるため，学校カリキュラムに沿った垂直尺度化としてはあまり適切なデザインではない。

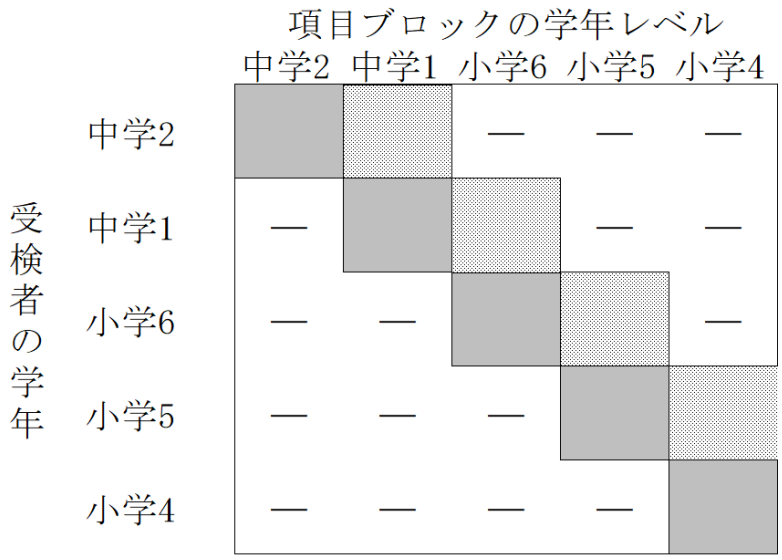

図 3.1 共通項目デザインの図

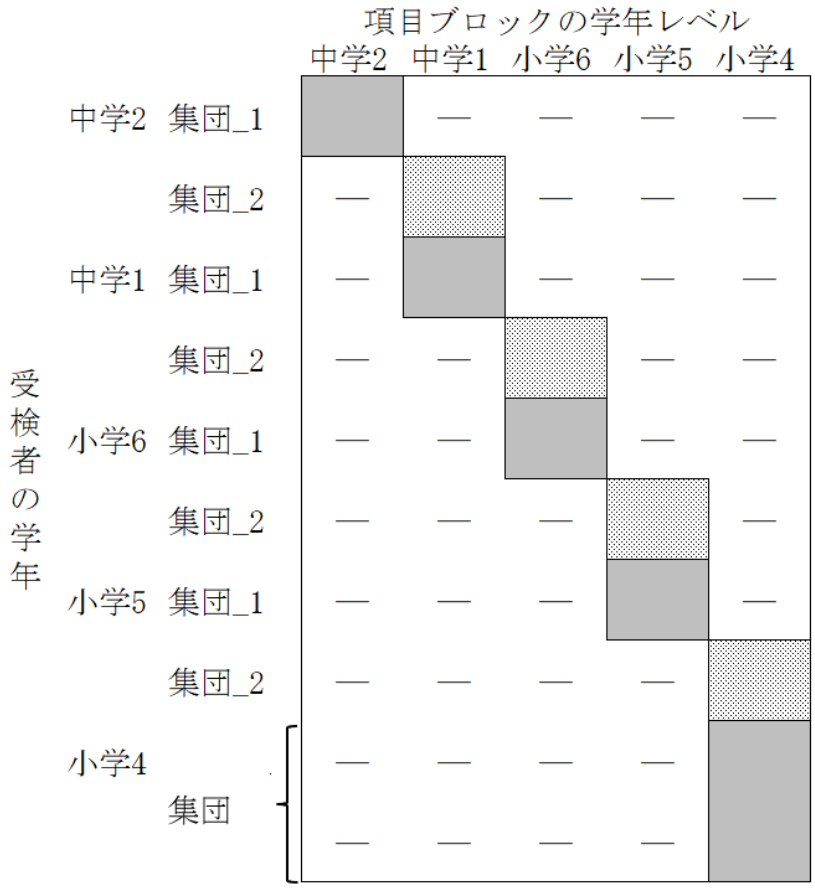

図 3.2 等価グループデザインの図

「尺度化テストデザイン (scaling test design) 」 (図 3.3) は全学年で共通の項目を用意しなくてはならないため，これらのデータ収集デザインのなかで一番実施のためのハードルが高い。たとえば全学年共通の項目 (scaling test item) の作問に関してのノウハウの蓄積が十分でない場合，尺度化テスト項目の作成は困難になると予想される。しかし，ほかの 2 つのデザインとは違い，全学年で一貫した内容の項目があるため，全学年を通した児童・生徒の能力を領域内ではっきりと記すことができるというメリットを持つ。

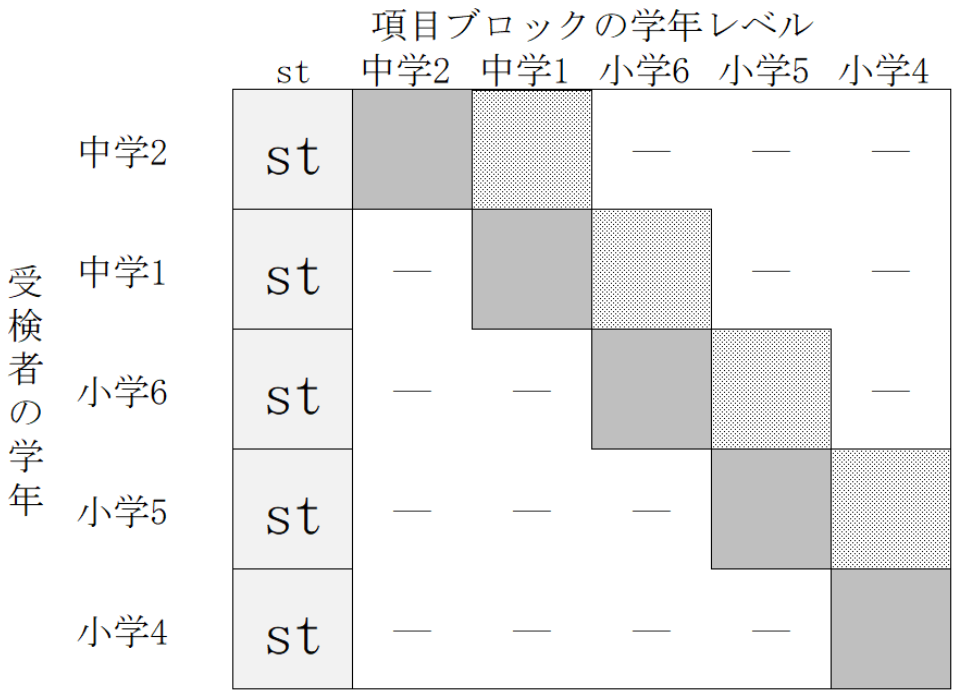

図 3.3 尺度化テストデザインの図

「均衡型単一グループデザイン (counterbalanced single-group design)」 (図 3.4) では，学年をランダムに 2 グループに分けたうえで，片方のグループは学年レベルにあった問題を前半に回答し，下の学年レベルの問題を後半に回答する。もう一方のグループはその逆で，下の学年レベルの問題を前半に，学年レベルにあった問題を後半に回答する。このデザインは受験者の学習や疲労による得点への影響を考慮して実施するために用いられる。テストに含まれる共通項目が特定の部分に偏っている場合などにこのテストデザインは有効である。

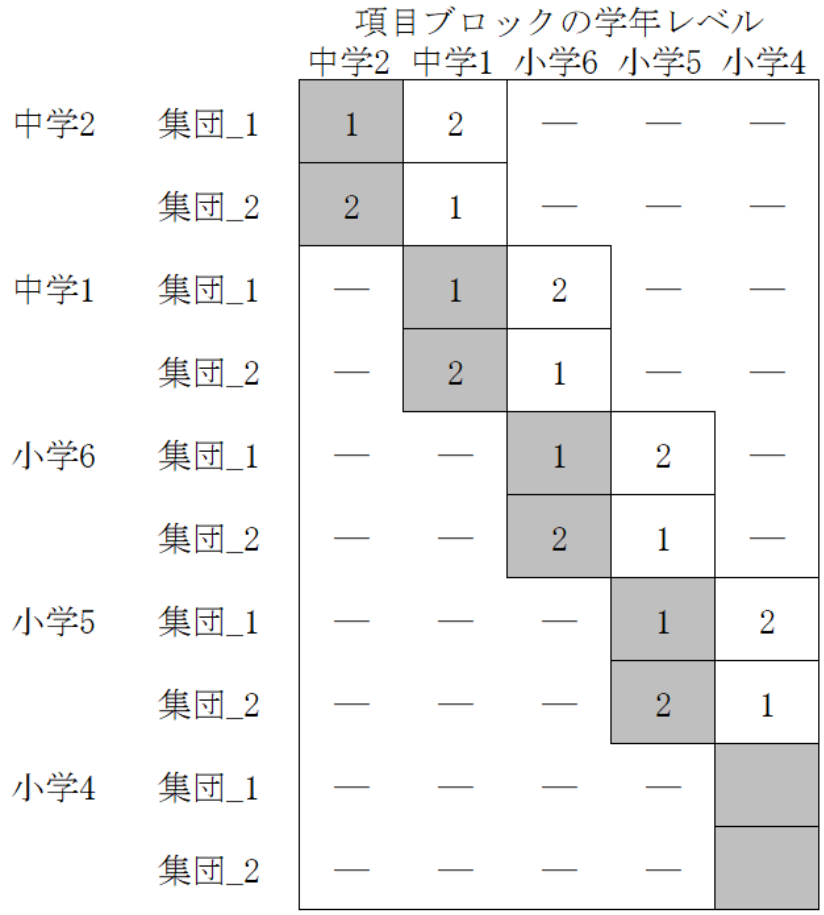

図 3.4 均衡型単一グループデザイン

3.2.2 異なる年度間の垂直尺度化

事前に年度間のテストに共通項目を配置しておけば，事後にアンカーテストを作成する方法（たとえば斉田, 2014; 熊谷ら, 2007 など）を用いなくても年度間の垂直尺度が対応づけ可能である。

重要なことは，年度間の共通項目の配置方法である。O'Neil (2010) によって，学年ごとに異なる年度のテストを対応づけする方法 (horizontal scale maintenance) (図 3.5) と年度ごとに垂直尺度を構成した後で年度間の共通項目全てを使用して対応づけする方法 (vertical scale maintenance) (図 3.6) が提案されている。その際，異なる年度の同一学年のテストには必ず一部に共通項目が配置されているとする。前者は最初の年度のみ垂直尺度化をおこない，年度間の異なるテストは同じ学年同士でのみ等化される。後者の方法では年度ごとに垂直尺度を構成した後に，学年間の共通項目を利用して等化をおこなう。最終的に破線で囲まれた部分のテストは，本来は全く共通項目を含まないのにもかかわらず，共通尺度上でテスト得点を比較することができるようになる。

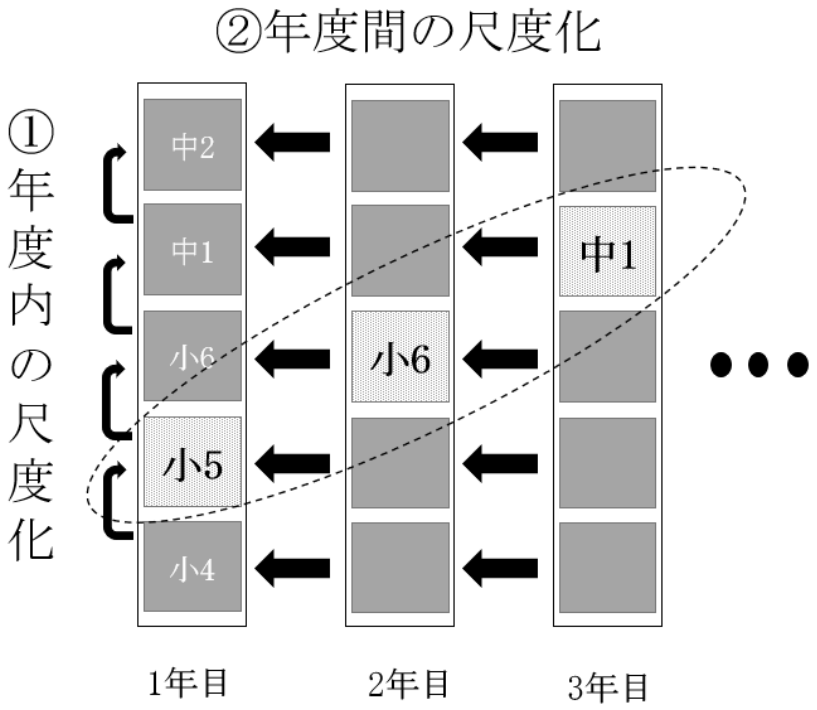

図 3.5 horizontal scale maintenance

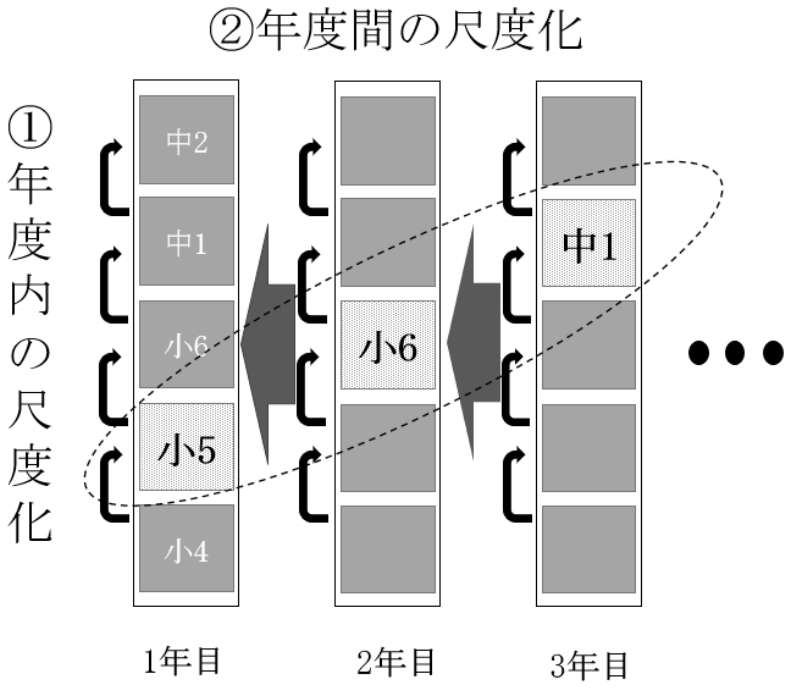

図 3.6 vertical scale maintenance

澁谷・柴山 (2017) の垂直尺度化のデザインは上記ふたつとは異なり，尺度間の共通項目が尺度の最上位，もしくは最下位学年のテスト項目にしか存在しない (図 3.7)。

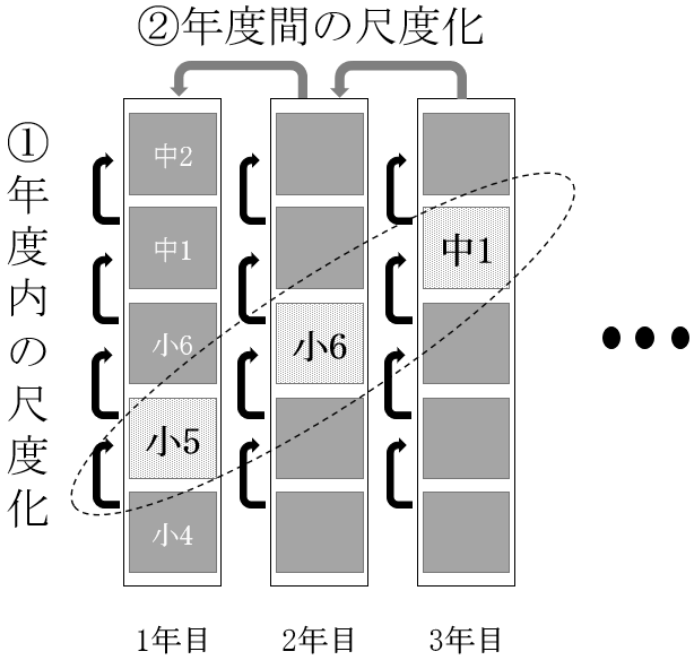

図 3.7 最上位学年のみ年度間の共通項目が配置されるデザイン

異なる年度間の垂直尺度を対応づけすることで得られるメリットは，直接的には全く共通情報がない受検者の 2 時点でのテスト得点を，垂直尺度上で比較することができるようになるということである。

3.3 基本的な尺度化の方法

次に具体的な垂直尺度構成のための計算方法について説明する。この尺度構成の手続きを特に尺度調整 (calibration) と呼び，大きく同時尺度調整法 (concurrent calibration method) と独立尺度調整法 (separate calibration method) の二種類に分けることができる。またそのふたつの折衷案的な手法としてペアワイズ法 (Pair-Wise method, PW) と呼ばれる手法も用いられる(Briggs & Weeks, 2009; Karkee, Lewis, Hoskens, Yao, & Haug, 2003)。

同時尺度調整 (Concurrent Calibration, CC) 法は，共通情報で結びつけられている複数学年のテストの項目反応データをひとつのデータ行列にマージし，そのデータを用いて一度に全項目，全学年，全受検者集団のパラメタを推定する方法である。データをマージする際に当該学年が受検していない項目における反応はすべて欠測値 (missing value) とみなすが，この場合の欠測は通常，多重代入法 (multiple imputation) などで補完しない。一般に，こうした推定方法を完全情報最尤法 (full information maximum likelihood estimation) と呼ぶこともある。CC 法でパラメタを推定するためには，学年ごとに異なる母集団からサンプリングされていることを想

定する多母集団モデルを扱う必要がある。特に，MMLE を用いる場合には集団ごとに事前分布のパラメタを推定する必要があるが，どれかひとつの集団のパラメタを固定しなければ解は収束しない。後述する独立尺度調整法に比べ，CC 法は多くの場面で効率がよく，データの持つすべての情報を用いて推定することができるが，一回の計算時間の増加を招く。

独立尺度調整 (Separate Calibration, SC) 法は各学年で項目パラメタを推定した後に，任意の等化係数推定法により等化係数を推定し，2 学年のペアごとに等化を順次実行し，最終的にひとつの垂直尺度を得る方法である。実用上の点として，基準とする学年をひとつ決定しておき，その学年に向かってすべての学年のデータを等化する方が，パラメタの解釈をする上で都合がよい。等化係数を推定する方法は状況に応じて適切な方法が選択される。たとえば，困難度の外れ値などに左右されやすい Mean & Sigma 法を用いるよりも，Haebara の方法や Stocking & Lord の方法などの特性曲線変換法 (characteristic curve transformation method) を使用する方がよいとされる。また共通受検者法に限って言えば，通常の Mean & Sigma 法を用いるのではなく，EM アルゴリズムにもとづく推定母集団分布の推定値を用いる方法の方がより優れた推定値を与える (熊谷・野口, 2012)。しかし等化係数の推定を重ねるごとに等化誤差が蓄積されていくため，複数回の等化はあまり望ましくないと考えられる。そこで，すべての推定済みの項目パラメタを一度の計算でひとつの共通尺度に落とし込む方法が前川 (1991) によって提案されている。等化誤差の問題があるものの，実用上の観点からは SC 法が好まれる(Briggs and Weeks, 2009)。なぜなら大きな項目バンクを構築する際に後から項目を追加することが容易であり，IRT モデルの一次元性の仮定が守られてない状況でも頑健な手法であると言われている(Beguin, Hanson, & Glas, 2000)からである。

その他の方法としてパラメタ固定 (Fixed Parameter, FP) 法がある。特に，一度推定した項目パラメタの値を固定して，隣接する学年のパラメタを推定していく場合は項目固定 (Fixed Item, FI) 法，共通受検者法を用いる場合に項目ではなく受検者のθの推定値を固定する方法を Fixed Theta (FT) 法と呼ぶ。ただし FP 法では特定の学年の分布の平均と標準偏差を固定することができない。

3.3.1 尺度調整法に関する先行研究

尺度調整法による推定結果の違いについてはいくつかおこなわれている(Briggs and Weeks, 2009; Hanson and Beguin, 2002; Ito et al., 2008; Kim and Cohen, 1998, 2002; Meng, 2007; te Marvelde et al., 2010; Yin, 2013) 。ここで言及する研究のすべては，共通項目デザインのもとでデータが実際に収集されたか，収集されることを想定した研究である。

Hanson & Béguin (2002) の研究では共通項目デザインの等化の文脈においては CC 法の結果が優れているとしている。

Karkee et al. (2003)はコロラド州のアセスメントプログラム CSAP の 2002 年データを用いて，5 学年にまたがる垂直尺度化における 3 つの尺度調整方法を比較している。比較のためのデータセットは全受検者のデータ (54,000 人) から 10,000 人をランダムに抽出し，さらにその

データをふたつの集団にランダムに分割するという方法で得ている。項目パラメタ推定の収束や，適合度，DIF (Differential Item Function) などで評価した結果，ひとつのデータセットではあるものの，モデルデータフィットの観点から SC 法がより好ましい方法であると述べている。

Kim & Cohen (1998; 2002) は二値型の IRT モデルおよび GRM (Graded Response Model, 段階反応モデル) を用いた垂直尺度化における尺度調整法の比較実験をおこなった結果，どちらのモデルにおいてもわずかに CC 法のほうが良い結果を示したが，GRM において共通項目が 5 項目 (テスト全体の約 16 パーセント) の場合は SC 法が良い結果を示すことを報告している。ただし両研究はひとつのテストを能力水準の異なるふたつの集団が受検した場合を想定したシミュレーションである。

Meng (2007) は非常に多くの条件をシミュレーション比較した研究である。尺度調整法はもちろんのこと，共通項目の数，共通項目の形式 (二値のみ，二値と多値の混合)，多値型の項目の数，受検者数において複数の条件を設け，合計で 96 条件の比較をおこなった。すべての条件を総括した結果，PW 法がもっとも良いパフォーマンスを示したと報告している。

Briggs & Weeks (2009) は SC 法と PW 法の比較を，1PLM と PCM，3PLM と GPCM においておこなっている。彼らが使用したデータセットは CSPA の読解力テストであるが，Meng (2007) と同様の結果を報告している。

また，一次元 IRT モデルの仮定を逸脱したデータに対して垂直尺度化の尺度調整法の性能比較をおこなった研究もある。Eastwood (2014) はモデルの一次元性の仮定が守られている場合は CC 法や PW 法が優れているが，そうでない場合には CC 法，PW 法，FT 法のいずれの方法でも推定値にバイアスが生じると結論づけている。Kolen & Brennan (2014) はそのような場合には SC 法を用いるのが安全であるとしている。

3.4 実データの垂直尺度化

実データによる垂直尺度化の研究例および実践例は言語系や計算能力などの一部の能力や学力テストで数多く報告されている。

日本語の例では比較的古い研究ではあるが芝 (1978), 芝・野口・柴山 (1986) 語彙理解尺度構成の研究例がある。近年では(高橋・中村, 2009; 2015) が語彙・漢字に関する適応型テスト ATLAN (Adaptive Tests for Language Abilities) の尺度作成において同時尺度調整法による IRT 垂直尺度化を報告している。

英語の研究例は林 (1996) の英語能力の成長曲線を描写することを目的に，共通項目を利用して垂直尺度を構成した研究や，大規模英語学力テストの同一年度内の異なるテストの共通尺度化をおこなった熊谷ら (2007) の研究などが挙げられる。Kenyon, MacGregor, Li & Cook (2011) は民間企業の作成する K-12 の英語基礎能力試験 (English Language Proficiency Test) の垂直尺度化について研究しており, テストデザインや尺度調整方法, 項目適合度などについて報告している。一次元 IRT モデルを採用する垂直尺度が多い中，Koepfler (2012) の研究は心理計量モデルに

Bifactor IRT モデルを用いている珍しい例である。実務レベルでは公益財団法人日本英語検定協会 (2016) が英検の各級を垂直尺度化し，異なる級のスコアと比較可能な得点として受検者に開示している。しかし，複数技能を測るテストを一次元 IRT モデル当てはめることの根本的問題や影響に関しては分析されておらず，尺度調整法や推定方法の詳細は不明である。

計算能力に関しては, 喜岡 (1991) の小学生の計算能力テストの尺度化の研究, 藤森 (1991) の算数学力尺度の作成を初期のものとしてあげることができる。特に藤森 (1991) は等化係数を推定しない方法 (CC 法) で尺度化している。

比較的規模の大きな学力調査やアセスメントプログラムへの適用例であれば，アメリカの多くの州がテスト専門の業者と提携し，あるいは州独自で垂直尺度を構成し，政策評価などに広く使われている。それぞれのテストがどのようなデータ収集デザインや心理計量モデルを利用しているかどうかは Reckase (2010) や Betebenner & Linn (2009), Patz (2007), (Patz and Yao, 2007), 佐藤･村木 (2008)などに詳しい。それらのうち代表的なものは Harcourt Assessment 社の Metropolitan Achievement Tests と Stanford Achievement Test Series の他，CTB/McGraw-Hill 社の California Achievement Test や TerraNova などである。この他に，埼玉県はさいたま市を除く県内の小学 4 年生から中学 3 年生を対象とした国語と算数，数学の学力に関する悉皆調査を実施し，一次元 IRT モデルによって垂直尺度化している(中室他, 2017; 埼玉県教育委員会, 2016; 2018)。

3.4.1 垂直尺度の評価

垂直尺度化の手法同士を比較したり，その尺度単体の特徴を評価したりするための方法は大きく分けて三つである (Kolen & Brennan, 2014; Young & Tong, 2016) 。一つ目は学年ごとに母集団分布の平均を推定し，学年ごとの変化の推移を確認する方法である。代表値として平均ではなく中央値を用いられることもある。この指標は学年間の成長 (grade-to-grade growth) と呼ばれる。二つ目は平均ではなく分散 (標準偏差) の学年間の変化であり，これは学年間のばらつき (grade-to-grade variability) と呼ばれる。三つ目は Yen (1986) の効果量 (Effect Size) である。効果量は学年間の平均の差を学年内の標準偏差を合わせた値で割った指標であり，

$$Effect\ Size = \frac{\bar{x}_{upper} - \bar{x}_{lower}}{\sqrt{\frac{\left(n_{upper}\sigma_{upper}^2 + n_{lower}\sigma_{lower}^2\right)}{\left(n_{upper} + n_{lower}\right)}}}, \tag{3.11}$$

と計算され，学年分布の分離 (separation of grade distribution) と呼ばれる。

3.4.2 尺度の縮小

Topczewski (2013) によれば，垂直尺度における特有の現象として尺度の縮小 (scale shrinkage) がいくつもの研究で観察されている。歴史的にこの縮小現象をはじめて指摘し，そう名付けたのは Yen (1985) である。もともとは Lord (1975) がいくつかのテストデータを分析したところ

3PLM の項目困難度と項目識別力が有意な正の相関を示したと指摘していることが問題の出発点であり，その後 Yen (1985) が同じく 3PLM で垂直尺度化されたテストバッテリーにおいて，それらのパラメタが有意な相関を示すだけでなく，学力テストのレベルが上がるにつれて困難度の標準偏差が減少することを発見し，これを尺度の縮小と名付けている。この縮小の仮説が正しければ，理論上は等間隔であると想定している垂直尺度上の学力変化が，上級学年では目盛りの間隔が細かくなるため，見かけ上は伸びが現象しているように観測される可能性がある。

この現象は果たして垂直尺度特有のものだろうか。Hoover (1984) は垂直尺度のこの性質に懐疑的であり，一般に受け入れられている考え方 (widely held belief) としては学力の高い集団の方が伸びは大きいはずであると主張しているものの，それに対して Burket (1984, p.16) は basic skills achievement においてはあり得る現象であると反論している。

どちらの主張もやや古いが，近年では尺度の縮小は一次元の垂直尺度に対していくつかの攪乱要因が与える影響の結果として理解されているようである。この後の説明する研究ではテストの回答に必要な能力の多次元性，測定誤差 (推定方法)，項目の局所依存などが，その原因として指摘されている。

Yen (1985; 1986) は仮想的な項目パラメタのもとで，MIRT モデルによって生成した項目反応データなどを一次元 IRT モデルで分析し，上位の学年では項目の正答に必要な能力が複雑 (多次元) になるにつれて尺度は縮小すると結論づけている。最近では単なる測定の多次元性の問題としてではなく，学年が上がるにつれて同じ教科でも測定している能力が微妙に変化している問題として捉え，この現象を construct shift と呼ぶようになっている (Martineau, 2006)。たとえば Wang & Jiao (2009) や Li & Lissitz (2012) は一次元 IRT モデルによる垂直尺度化特有の条件として construct invariance (構成概念の不変) を唱えており，IRT 垂直尺度化は一般的な発達得点尺度の構成概念の変化を無視している，と問題視している。多次元性および構成概念の変化に対処するためには MIRT モデルや双因子 (bifactor) モデルなどを適用することが推奨されている (Eastwood, 2014)。

一方，Camilli (1988) は同じレベルのテストでも一年のうち前半と後半に受けたテストの結果では，後半に受けたテストの能力分布の標準偏差の推定値が小さくなっていることから，テスト項目の多次元性が尺度の縮小の原因ではないと考えた。この研究で指摘されたのは，IRT の初期の研究で使用されていた JMLE の理論的な欠陥による推定誤差と，受検者の能力と項目の困難度のミスマッチのために尺度が縮小していると結論づけた。Camilli, Yamamoto, & Wang (1993) は Camilli (1988) の結果踏まえ，Mislevy & Bock (1982) によるプログラムを使用し，MMLE により NAEP の垂直尺度を構成した。この分析結果は，第 4 学年から第 8 学年にかけては尺度が拡張しているのに対し，第 8 学年から第 12 学年にかけては尺度が縮小しているというものだった。この結果から Camilli らは，必ずしも垂直尺度において縮小の現象は観測されるわけではないが，尺度の縮小が生じる原因として測定誤差やテストの内容，推定方法，テストの多次元性などが考えられると述べている。この場合の推定方法とは単一テストのパラメタ推定

だけにとどまらず，尺度全体のパラメタ推定方法であり，具体的には尺度調整方法の選択に関する問題である。

Yen (1993)は，垂直尺度化に限定せず，局所項目依存 (LID) が IRT 尺度化に及ぼす影響について様々検討している。この研究では LID が直接的に尺度を縮小させると明言していない。その後の Topczewski (2013) が LID の測定精度に及ぼす影響を考慮して，結果的に縮小の原因のひとつである可能性を示唆した。

Topczewski (2013) は多次元性，LID およびテスト間の信頼性が異なる (非類似) 場合の 3 つの条件をシミュレートした結果を報告している。二次元 IRT モデルにより多次元性を，テストレットモデルにより LID を，識別力と当て推量パラメタの分布により信頼性の非類似を再現し，すべてのパラメタは多母集団モデルに拡張された MMLE-EM 法で推定された。シミュレーションの結果から，次元間の相関が低い多次元性と LID は学年間の成長と学年分布の独立にバイアスを生じさせ，尺度を拡張させてしまうが，信頼性の非類似は尺度を拡張もしくは縮小させることが明らかになった。これらの条件の中で識別力パラメタの低下による信頼性の非類似シミュレーションの条件は唯一尺度の縮小の現象を再現しているが，その程度は非常に小さい。

これまでの尺度の縮小に関する結果を総括すると一次元 IRT モデルの仮定を大きく逸脱するような条件が確認されると，尺度の縮小 (拡大) が生じると言える。この尺度の縮小あるいは拡大の問題を統合して尺度の可変性 (scale variability) の問題と呼ぶこととする。現実のデータで問題として指摘されるのはもっぱら尺度の縮小のみであったが，Topczewski (2013) によるシミュレーションからは拡大の可能性も示唆されている。肝心の原因については，測定対象以外の能力の次元，構成概念の変化，LID，テストの信頼性，測定誤差等の，一次元 IRT モデルに対する複数の攪乱要因が影響していると考えられる。さらにこれらの攪乱要因は互いに影響し合うため特定のモデル (MIRT モデル，テストレット IRT モデルなど) を当てはめることだけが最善の方法とは限らない。最後にこれまでのレビューをもとに，尺度に及ぼしうる攪乱要因を図 3.8 にまとめる。

この図においては尺度の可変性の原因を推定誤差や尺度調整法などの測定誤差に起因するものと，テストの内容や IRT モデルの特性に起因するものの二種類に大別している。測定誤差に関しては MMLE-EM やデータに適した尺度調整法を選択することである程度減少させることができる。しかしもう一方の原因については一律に対処するのが困難である。多次元性の一部や construct shift の問題などは MIRT モデルにより対処でき，LID についてはテストレットモデルなどの項目局所依存を認める特殊なモデル (たとえば Adams, Wilson, & Wang, 1997) を使用することで対処できる。しかし，それらが同時に生じているような項目や複数の LID の原因が存在したり，項目反応にそれ以外の要因 (たとえば DIF など) が影響したりする場合などが現実には考えられる。そのためモデリングで対処する際には，考え得る攪乱要因に対して適切なモデルが選択できるように配慮しなくてはならない。

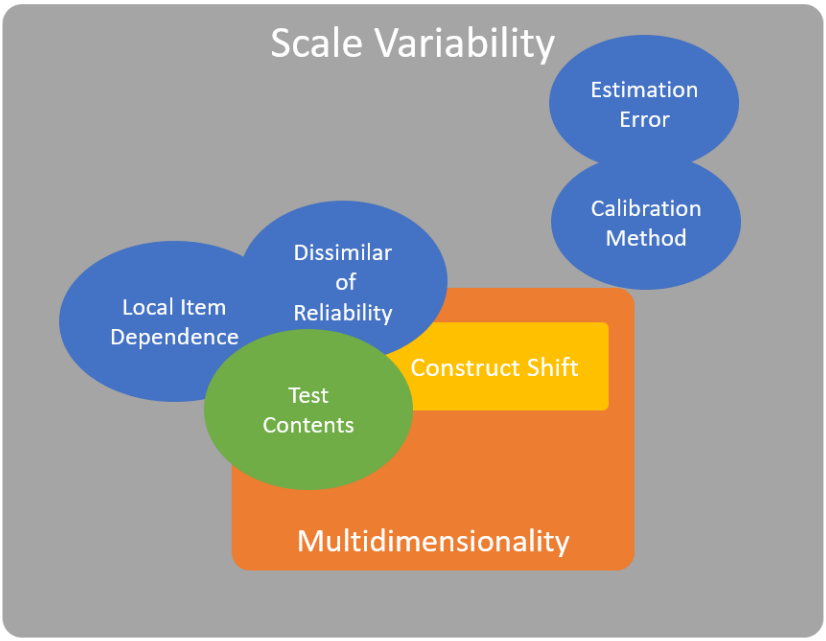

図 3.8 尺度構成における様々な攪乱要因[1]

3.5 垂直尺度化の制約

Reckase (2010) は垂直尺度化のアイディアは新しいアイディアでないにもかかわらず，”it is one of the most challenging areas of applied psychometrics and there are numerous cautions about procedures and interpretations of vertical scales”と述べている。その理由はいくつか考えられる。たとえば得点の解釈において，垂直尺度上の得点は遡及的な比較に限定されるため，一時点での異なる学年の児童生徒らの得点を比較することには向かない。具体的には，垂直尺度上で同じ年度の中学 1 年生の 60 点と中学 2 年生の 60 点を比較することはできないということである。それは垂直尺度上では受検者が学習済みの項目と未学習の項目が混在して位置づけられているが，未学習の項目に対する達成 (achievement) は学習済みの項目の達成からは本質的に測定・予測することができないからである。

また，現状広く用いられている垂直尺度が一次元 IRT を基本とした尺度であるが，一次元 IRT 垂直尺度化は尺度で測定している構成概念がひとつであるという非常に大きな制約がある。しかし実際のところ，学校教育における指導と学習の面では単一の教科であっても下位領域に分けて考えられることが多く，ただ垂直尺度上の数値の変化を追っていくだけでは，学力の発達を可視化することはできても，児童生徒や教師へのフィードバックに支障をきたす恐れがある。

最後に，垂直尺度の可変性の原因が複数指摘されており，それらに対する対処法もそれぞれ異なるという点も，垂直尺度化の実装を困難にする要因だと考えられる。

---

[1] 筆者作成

4　実験

4.1　シミュレーション分析：垂直尺度化に適した標本サイズ

Kolen & Brennan (2014) は，尺度化テストデザインは，垂直尺度上で成長の定義をおこなうときに，垂直尺度の対象とする学習内容全域にわたって成長を定義することができるデザインであると述べている。また，垂直尺度の一次元性を確認する際にも尺度化テスト項目が存在することで，一部の項目ではあるものの，相関行列の固有値を用いる一般的な方法で尺度全体が一次元であるかどうか確認をすることができる。共通項目デザインでは隣接する学年同士でしか確認することができない。しかしながら，これまでの垂直尺度化についての先行研究では，尺度化テストデザインを利用した垂直尺度化に関する知見はほとんど得られていない。そのため，尺度化テストデザインの尺度調整に必要な項目と受検者のサンプルサイズや，尺度調整法が推定値に与える影響について十分に考察できていないのが現状である。

そこで，はじめに特定の項目パラメタ，能力パラメタの条件下のもとで標本サイズと尺度調整法を変化させたときの推定値への影響についてシミュレーション分析をおこなう。

4.1.1　実験デザインとデータ生成方法

適用する IRT モデルは 2PLM である。能力パラメタと項目パラメタの乱数を発生させ，そこから項目反応パタンを生成し，複数の尺度調整法で項目パラメタを推定する推定した項目パラメタ（予測）と項目反応パタン生成に用いた乱数（真値）との誤差をいくつかの指標をもとに確認し，同時に母集団分布の推定精度も確認する。すべての分析には R (R Core Team, 2018) のバージョン 3.5.1 を使用し，項目パラメタの推定と等化係数推定には書き下ろした関数と，lazy.irt パッケージ(Mayekawa, 2016) の calr 関数を使用した。

（1）シミュレーションデータ生成方法

5 つの異なる学力水準の集団を想定し，便宜上 G1～G5 とする。受検者集団の学力分布は正規分布すると仮定し，G1 を$N(-0.8, 1)$とした場合に一学年ごとに平均値を 0.4 ずつ増加させた分布を想定する (図 4.1)。識別力パラメタの事前分布は全学年，全項目共通に対数正規分布$\ln N(0.5, 0.3)$を仮定した (図 4.2)。尺度化テストデザインは学年レベル相当のテスト項目と尺度化テスト項目の 2 種類を含むため，学年レベル相当の項目の困難度パラメタの事前分布は学力分布と同じ正規分布を仮定し，尺度化テスト項目の事前分布は中程度の学力水準である G3 の正規分布の分散をすこし広げた分布$N(0, 1.5)$を仮定した (図 4.3)。

受検者数の設定は一学年あたり 400 人，1,000 人，10,000 人の 3 通りであり，項目数は一学年あたり 15 項目，30 項目，60 項目の 3 通りである。共通項目の割合はすべて固定し，尺度化テスト項目を 1/3，下の学年との共通項目を 1/3，学年レベル相当の項目を 1/3 とした。

上記設定で乱数を発生させ，それをパラメタの真値とし，項目反応パタンを生成する。乱数発

生させた能力・項目パラメタを 2PLM の項目特性関数に代入し，正答確率を得た。同時に，区間[0, 1]の一様乱数をひとつ発生させ，「正答確率≧一様乱数」となった場合には正答反応 (1) を，それ以外は誤答反応 (0) を項目反応データとした。ただし全受検者が正解・不正解となる項目を含むデータセットは，破棄して再度乱数発生からやり直した。

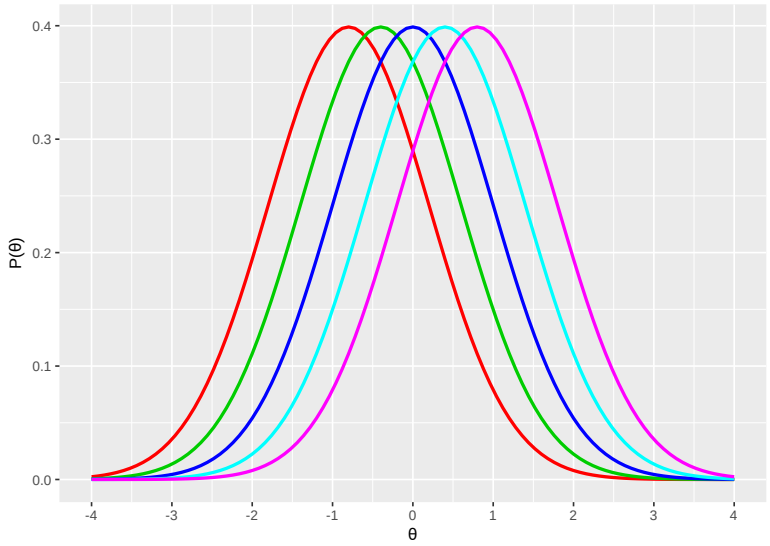

図 4.1 シミュレーション母集団の分布

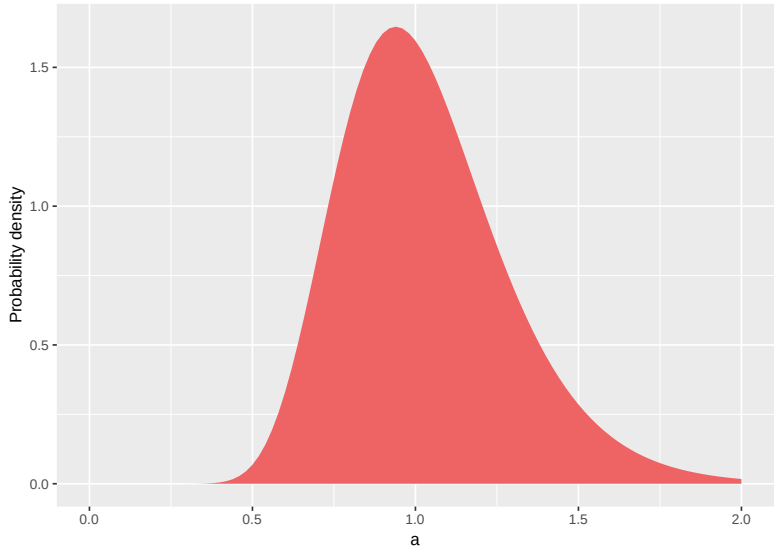

図 4.2 シミュレーション識別力の事前分布

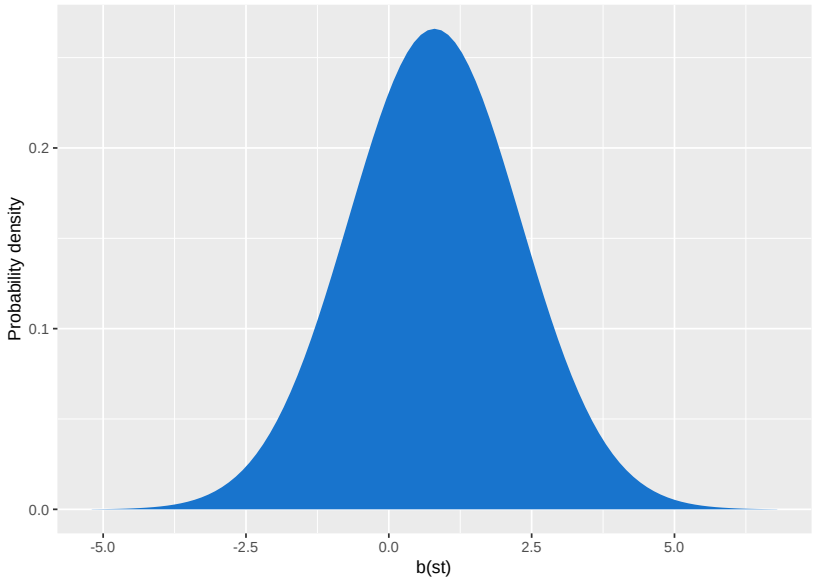

図 4.3 シミュレーション困難度の事前分布

（2）パラメタ推定方法

項目パラメタ推定方法は MMLE-EM 法を採用し，前川 (1991) を参考に多母集団モデルに対応するように推定プログラムを R で作成した。CC では，G3 の事前分布のパラメタを平均 0，標準偏差 1 に固定してパラメタを推定した。

項目パラメタの推定の際は様々なエラーや例外処理の必要性が生じる。MMLE の推定値が通常想定される値よりも大きく外れるか，識別力が負の値をとってしまった場合には，外れる前の段階の値を推定値として扱い，以降の EM サイクルでは更新しないこととした。それ以外の原因で推定が失敗した場合にはデータセットを破棄し，乱数生成からやり直している。

EM サイクルの収束判定は項目パラメタの変化が 1e-4 よりも小さくなるか，$-2\times$周辺対数尤度の変化が 1e-6 よりも小さくなることとし，EM サイクルは最大でも 200 回までとした。

SC では，G3 を平均 0，標準偏差 1 の基準集団とし，G2，G4，G1，G5 の順番で逐次等化をおこなった。SC の等化係数推定には 2 種類の方法を採用した。一般的なテスト等化で用いられ，精度も良いとされる Stocking-Lord の方法 (SL 法，Stocking & Lord, 1983) とすべての等化を一度に実行可能な calr の方法 (Arai and Mayekawa, 2011; 前川, 1991) である。SL 法での逐次等化では共通項目のパラメタの処遇にいくつかの方法が考えられる。今回は，等化先のパラメタと，等化前のパラメタを等化係数で変換したパラメタとの平均を計算して，等化後の共通項目パラメタとした。ただし識別力は幾何平均を計算して，共通項目のパラメタとした。

最後に，3 種類の尺度調整法 (CC，SL，calr) で求めたパラメタと項目反応データから EM アルゴリズムを用いて母集団分布の平均と標準偏差を推定した。

（3）推定値の評価

推定された項目パラメタは RMSE (Root Mean Square Error) と DICC (Difference of ICC) の指標を改良した指標を用いる。DICC は区分求積法と同じ要領でθを適当な区間に限定し，等間隔に分割して得た分点での，真値と推定値における正答確率のズレを平均するものである。また，DIF を判定する指標である指標 K (熊谷, 2012) では DICC おける分点ごとの正答確率を，推定母集団分布で重み付けして評価している。この重み付けにより項目特性曲線のズレを，母集団分布の確率密度が大きい部分を重く評価し，逆に母集団分布の密度が低い部分は軽く評価することができる。この指標 K の計算方法を参考に，今回は DICC の推定方法を改良した指標を DICC-WP (Difference of ICC Weighted by estimated Population distribution) と呼ぶこととする。なお, 推定母集団分布の平均と標準偏差は学年ごとに RMSE を計算した。

4.1.2 実験結果

はじめに項目パラメタの推定結果の比較をおこない，次に母集団分布の推定結果の比較をおこなう。まず各手法 (CC, SL, calr) ごとに 100 回分の識別力と困難度の RMSE のヴァイオリンプロットを描画した (図 4.4~4.9)。ヴァイオリンプロットは，中央値と四分位数，外れ値だけをプロットするボックスプロットとは異なり，データの分布自体を左右対称に表現するグラフである。分布の幅が広くなっている部分ほど，付近の値が多く観測されていることを示している。この場合，縦軸が RMSE であり，横軸は条件が取られている。青色が受験者 10,000 人，赤色が 1,000 人，緑色が 400 人の条件であり，同一受験者内では左から項目数が 15，30，60 の条件である。ただしこれらは軸が揃っていないので注意されたい。次にすべての手法をまとめてプロットしたものを図 4.10 と 4.11 に示す。こちらはすべての結果において目盛りが揃えられている。一部の推定結果が外れ値となっており，RMSE の軸が大きくとられてしまうため，外れ値を除いて軸を小さくとったものを図 4.12 と 4.13 に示した。次に DICC-WP のバープロットを示す (図 4.14)。最後に母集団分布の平均と標準偏差の RMSE のボックスプロットを図 4.15 と 4.16 に示す。図中の破線は各集団の真の平均と標準偏差の値を示している。

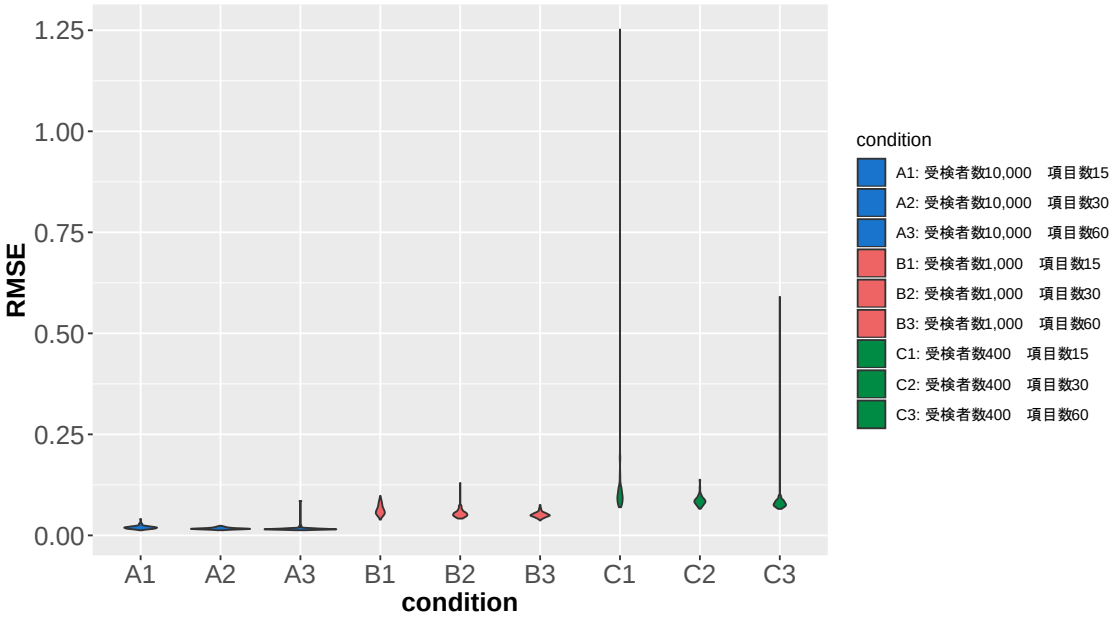

図 4.4 識別力の RMSE（CC）

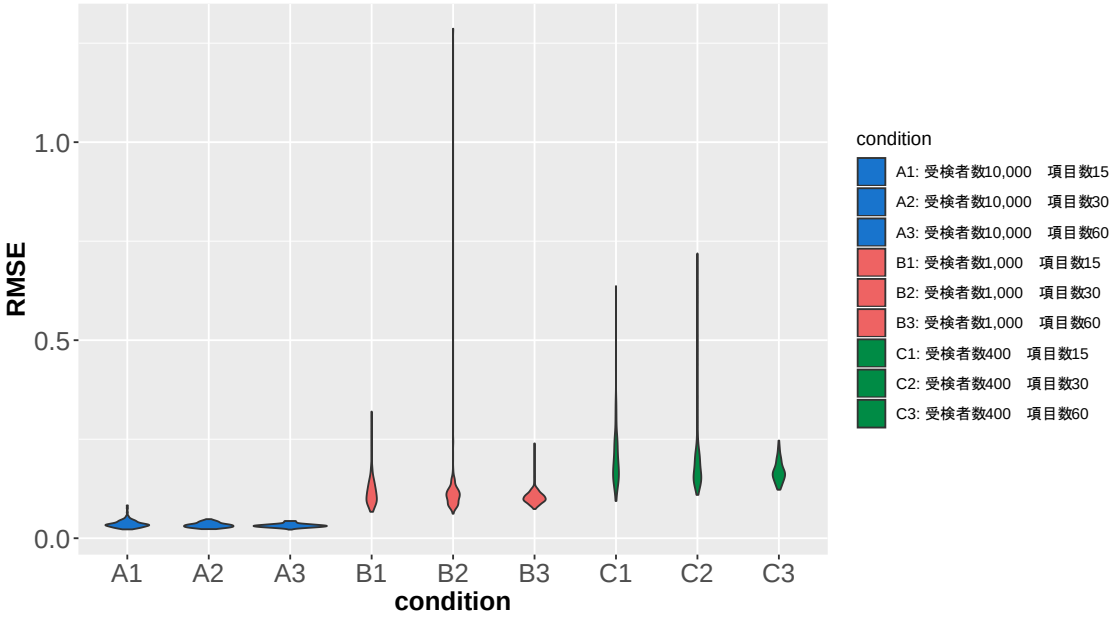

図 4.5 困難度の RMSE（CC）

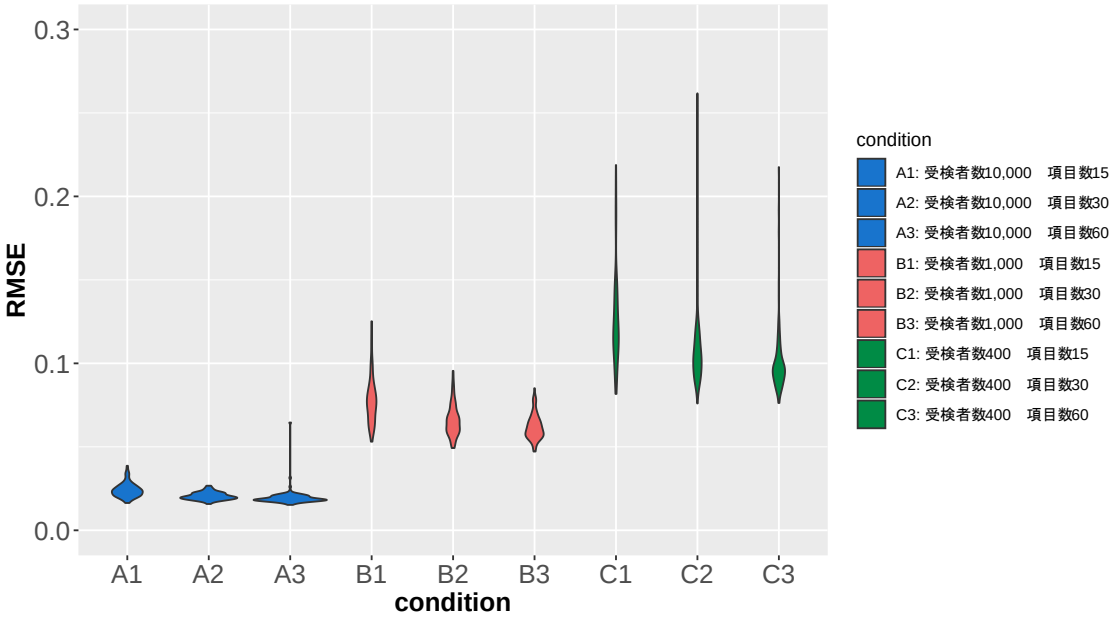

図 4.6 識別力の RMSE（SL）

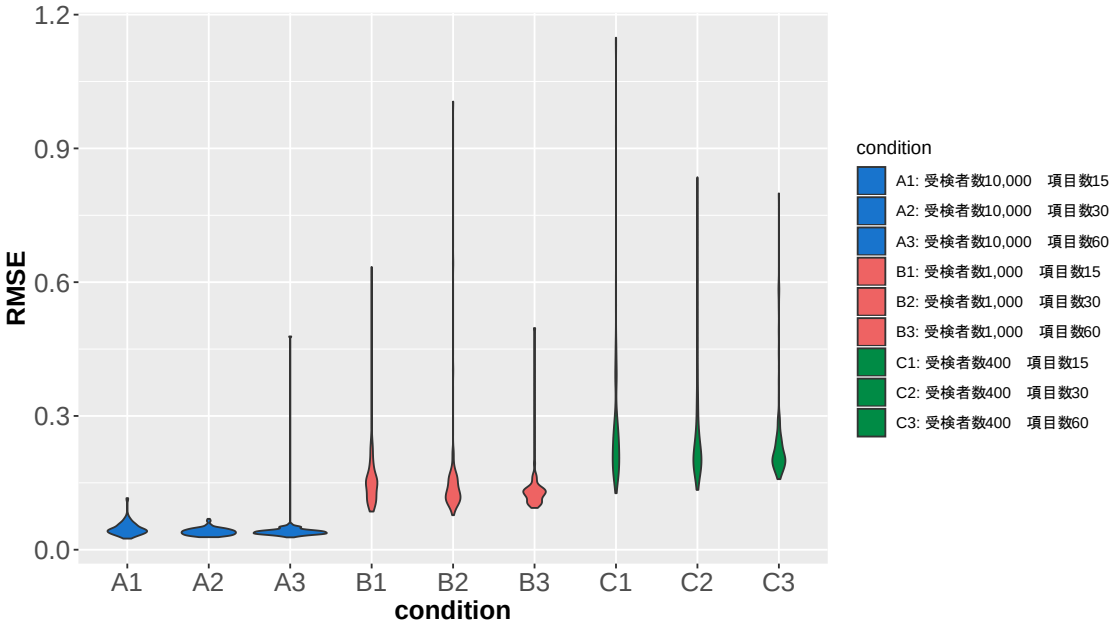

図 4.7 困難度の RMSE（SL）

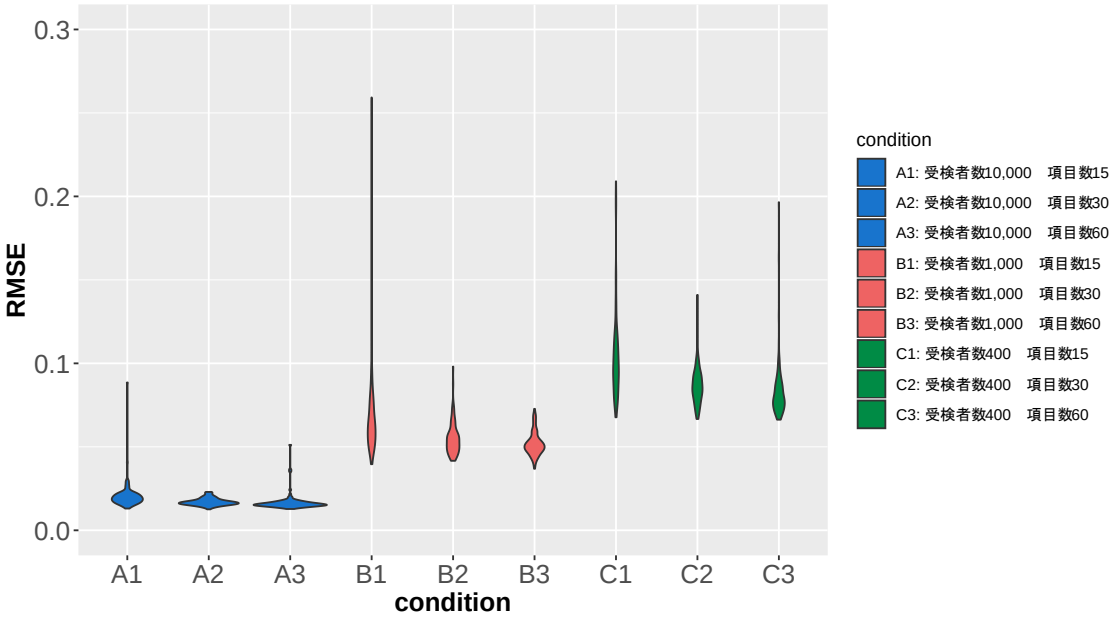

図 4.8 識別力の RMSE（calr）

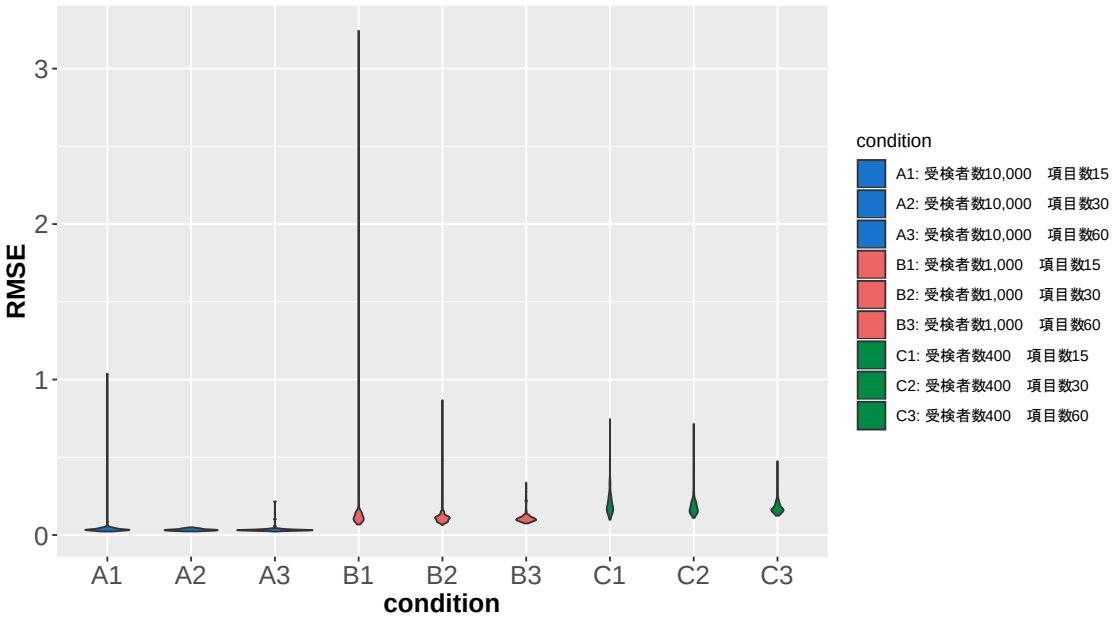

図 4.9 困難度の RMSE（calr）

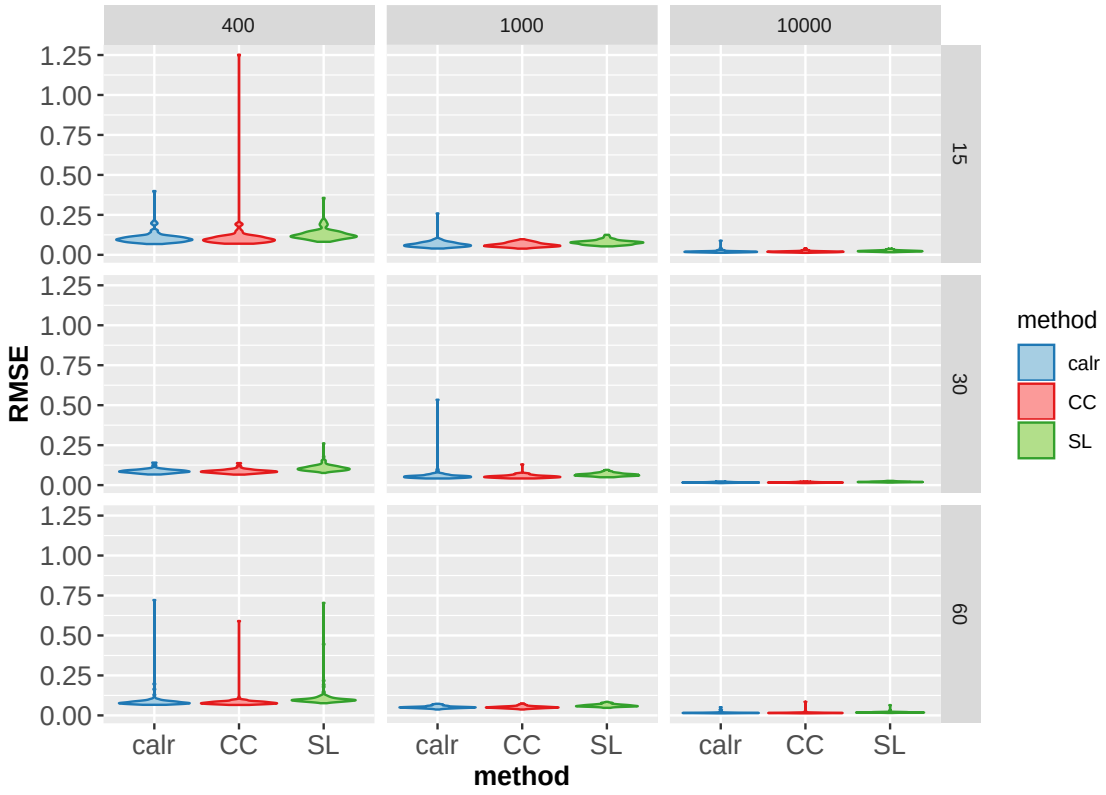

図 4.10 識別力の RMSE（ファセット，外れ値あり）

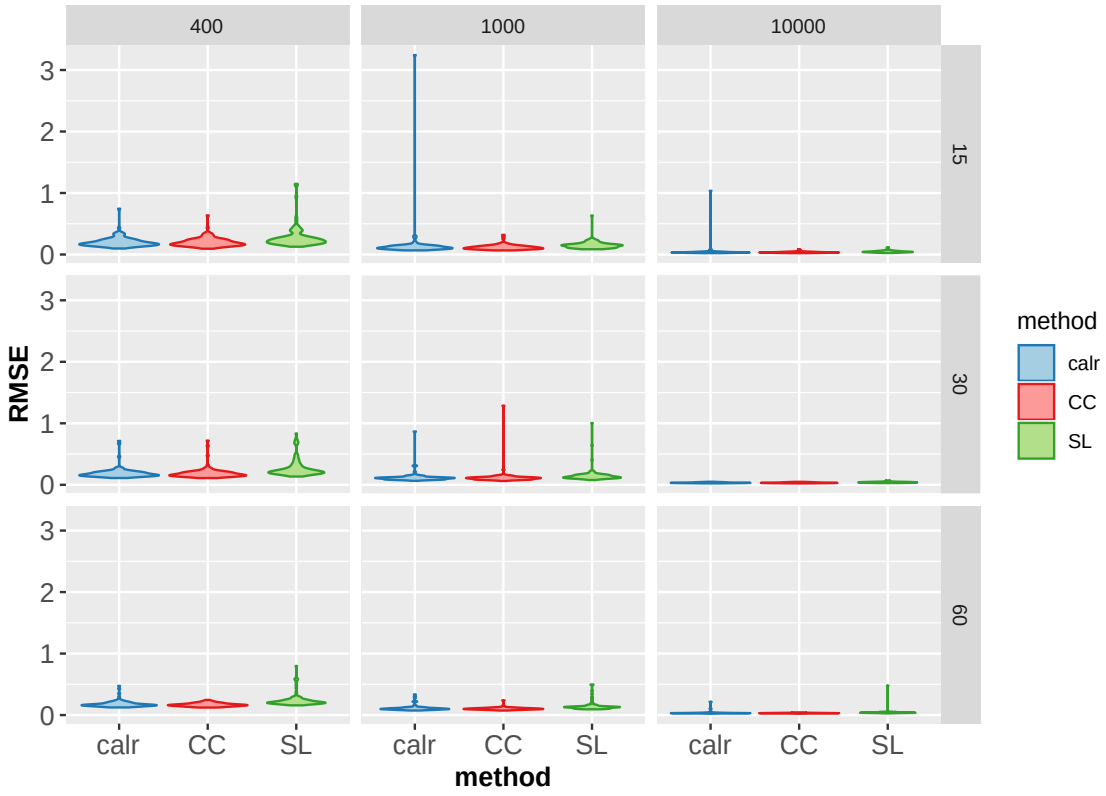

図 4.11 困難度の RMSE（ファセット，外れ値あり）

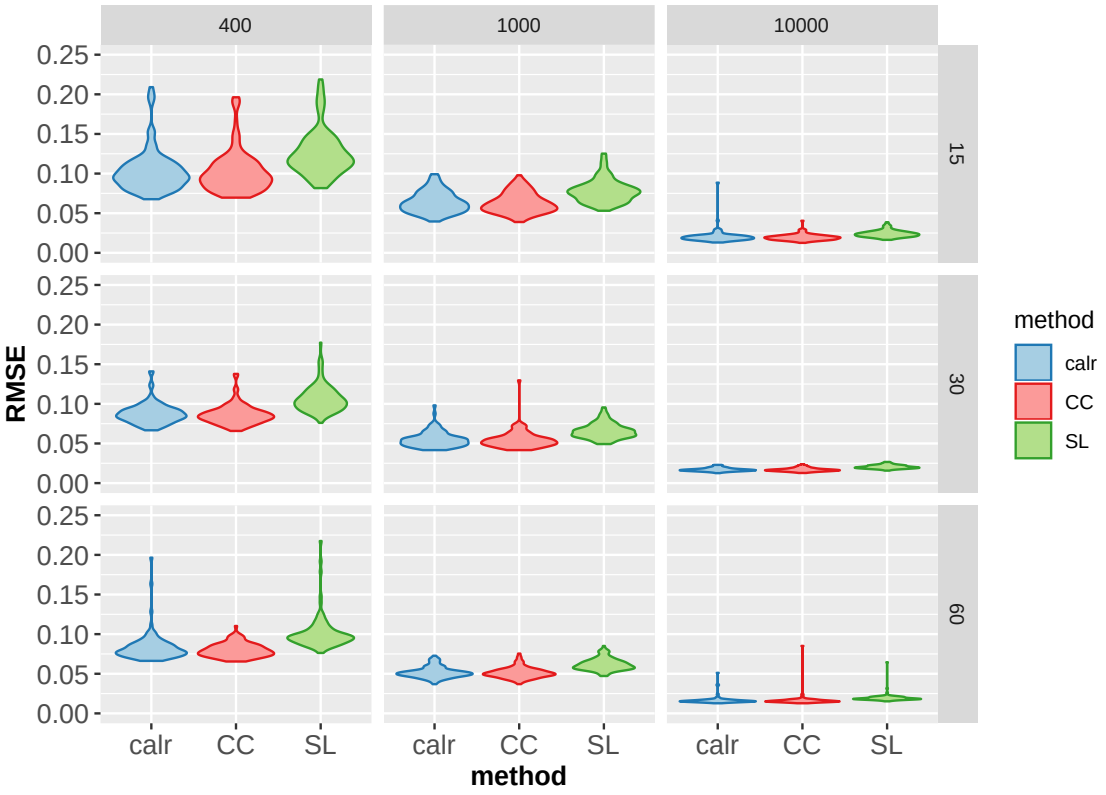

図 4.12 識別力の RMSE（ファセット，外れ値なし）

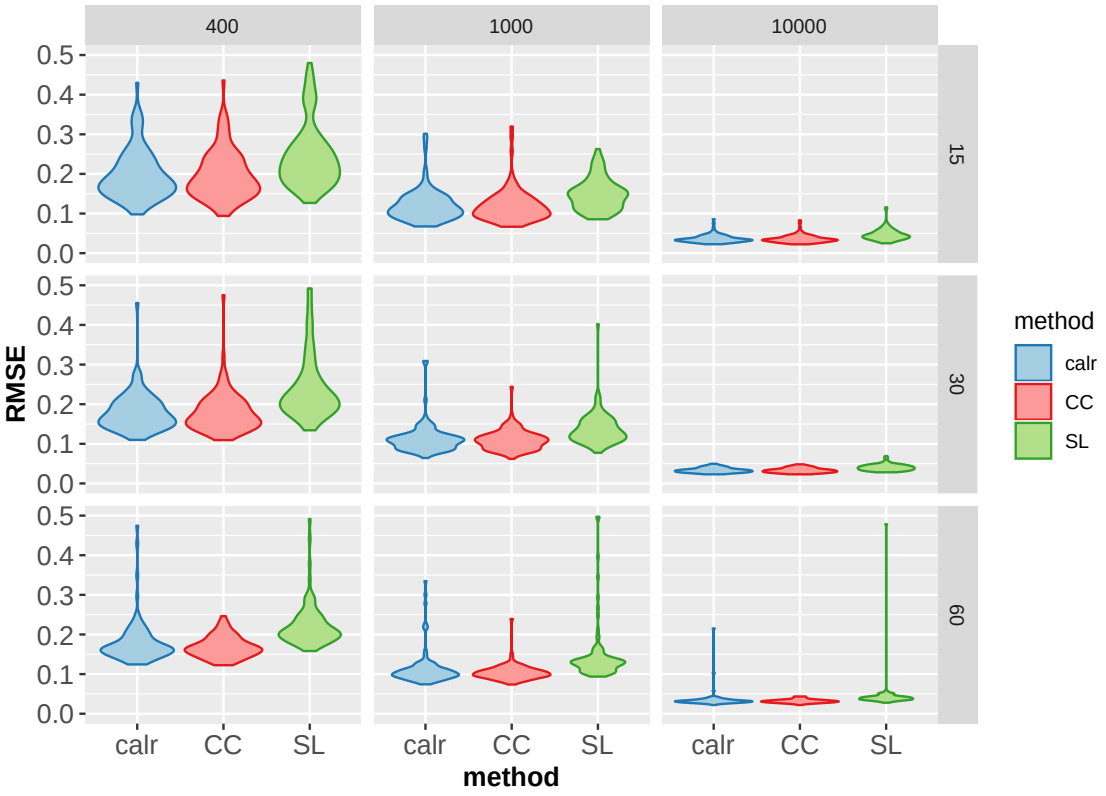

図 4.13 困難度の RMSE（ファセット，外れ値なし）

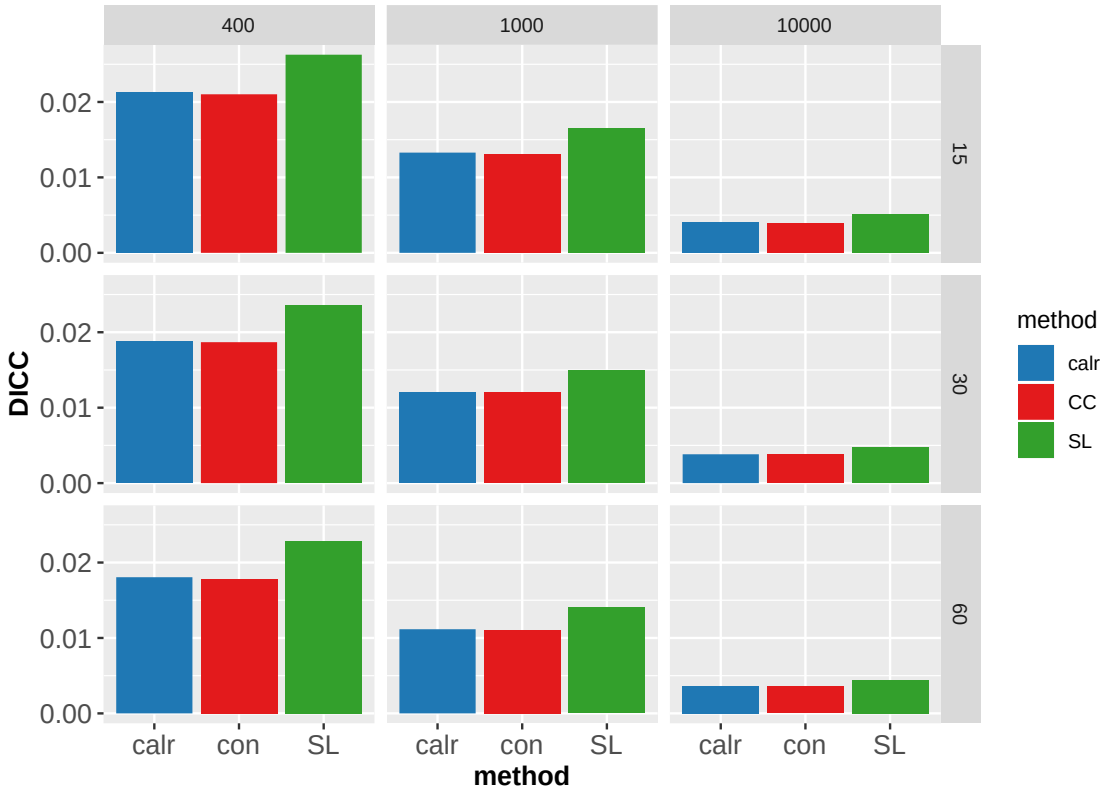

図 **4.14** DICC-WP のバープロット（ファセット）

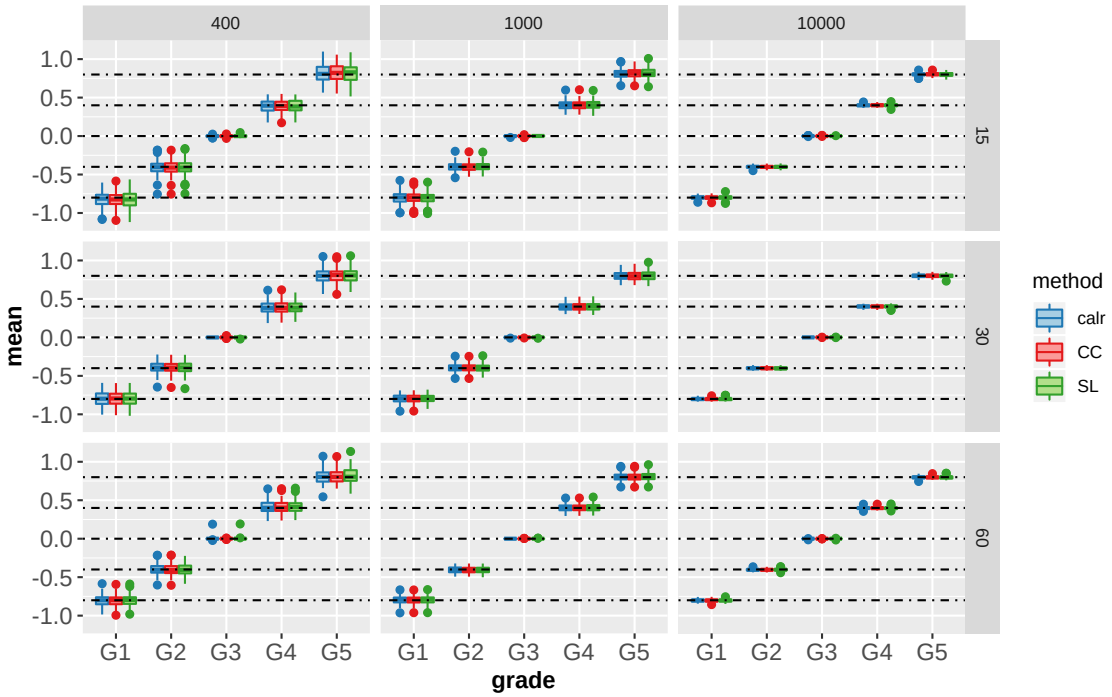

図 4.15 推定母集団分布の平均の RMSE

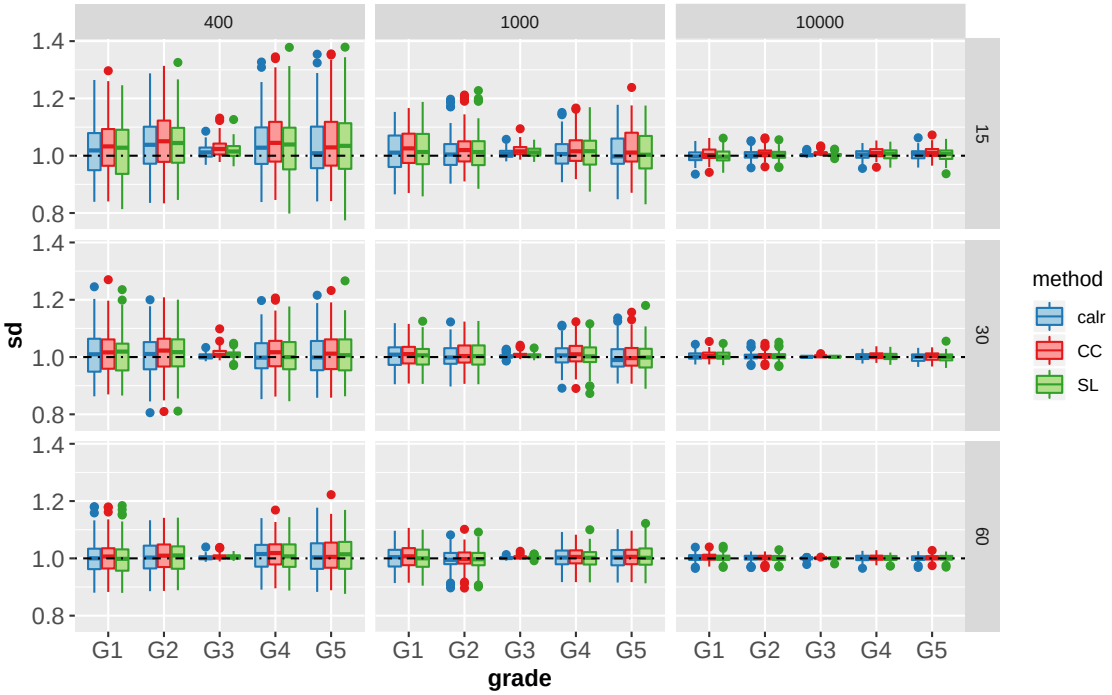

図 4.16 推定母集団分布の標準偏差の RMSE

4.1.3 考察

まず項目パラメタの方から考察すると，CC と calr の一部に非常に大きな RMSE が計算されたデータセットが見受けられる。CC についてはモデルにフィットせず，うまく推定できないパラメタを適当な値で計算を打ち切る設定としたため，おそらくその影響が出ているものと思われる。しかし calr の外れ値の原因は不明である。特に識別力の RMSE で 3 という非常に大きな値を記録しており，明らかに推定が失敗しているようである。

全体的な傾向として受検者の標本数を増やすほど推定結果の RMSE と DICC-WP は減少するが，項目数を増やしてもわずかに推定結果が安定する程度であった。DICC-WP で見ると大きく推定値が安定しているように見えるが，その目盛りについては最大でも 0.02 程度と非常に小さく，意味のある差を示せているのかが不明である。

手法の比較においては，SL が RMSE の平均とばらつき具合において最も性能が悪く，DICC-WP においても同様であった。CC と calr の間では RMSE の平均と DICC-WP ではそれほど大きな差はなかったが，RMSE のばらつき具合においては CC の方が小さく，良い性能を発揮していることが分かる。

次に母集団分布の推定性能について考察する。こちらもやはり項目パラメタと同様で項目数の増加はそれほど結果に影響せず，受検者数を増加させると明らかに推定結果は安定する傾向にある。平均の推定に関しては手法ごとに大きな差は見られなかった。しかし標準偏差の推定において，受検者数が少ない条件では CC が標準偏差をわずかに過大推定している。過大推定のほどはごくわずかであるほか，項目数と受検者数が増加すればほとんどなくなるため，実際の推定では無視できる程度の偏りであるといえる。なお，澁谷・柴山 (2018) の報告では基準となる学年の選択と区分求積の範囲の設定を不適切におこなった場合に，基準から離れた学年ほど標準偏差を過小推定することが指摘されている。

結論として，尺度化テストデザインにおいては従来の等化手法よりも，テスト全体を一度に等化する手法の方が安定した結果を与えることが分かった。今回のシミュレーションで扱った CC と calr では，CC の方が RMSE や DICC-WP において良い結果を与えているが，母集団分布の標準偏差をわずかに過小推定してしまう恐れがある。しかし calr においてもときおり推定結果が大きく不安定になるケースがあるため，総合的に判断すると CC が最も適した手法であると言えるだろう。

4.2 学力テストデータを用いた垂直尺度構成

続いて，実際に小学 4 年生から中学 2 年生までの計 5 学年で，尺度化テストデザインのもとで実施した学力テストデータを用いて垂直尺度構成を試みる。尺度調整方法はシミュレーションの結果を踏まえて CC 法でおこなう。はじめに項目分析と一次元性と LID の確認をおこない，次に項目パラメタを推定し，母集団分布を推定した。

4.2.1 データ収集の手続き

対象学年は小学 4 年生から中学 2 年生であり，対象教科は国語と算数・数学の 2 教科である。この 2 教科を独立に垂直尺度を構成する。テスト項目の作成は外部のテストアセスメント専門業者に委託し，学習指導要領の内容に準拠する形で，尺度化テストデザインのもとで項目を作成した。項目形式は多肢選択式と短答記述式の併用であるため，IRT モデルは 2PLM が適当であると判断した。なお短答記述式は国語の項目の長いもので，おおよそ 120 文字程度である。選択式の項目のみ 3PLM にする方法も考えられるが，その場合 2PLM と識別力や困難度のパラメタを比較することが難しくなるため，今回は全項目を 2PLM で分析する。フィールドテストはおこなっていない。

テスト項目の構成はシミュレーションの条件の一部と同じで，尺度化テスト項目が 10 項目，学年レベル相当のテスト項目が 10 項目，隣接学年との共通項目が 10 項目である。各学年の受検者数はおおよそ 400 人程度で，合計約 2,000 人の受検者が項目に回答している。

作成されたテストは A 県の複数校の児童生徒を対象に，2017 年の 11 月頃に実施された。採点は項目作成と同一の業者で実施され，正答ならば 1，誤答ならば 0 という二値型のデータに変換された。空欄や選択肢以外の記述が見られた項目反応は，誤答や新たなカテゴリとして扱ったり，部分点を代入したりすることなどがあるが，今回は完全に無回答 (欠測) とし，データ上は NA (Not Available) と処理した。同様にそもそも当該学年が受験していないテスト項目についても NA と処理している。

集計された項目は学年レベルごとに項目の頭文字のアルファベットが異なる仕様となっている。すなわち国語の a と b は小学 3・4 年生向け，c は小学 5・6 年生向け d は小学 6・中学 1 年生向け，e と f は中学 1・2 年生向け，st は小学 1 年から小学 3 年生向けの難易度の項目である。数学もほぼ同様の対応になっているが，st だけ小学 3 年から中学 2 年生の難易度で構成されている点が異なる。最後に受検者数と各学年が受検するテスト項目のレベルをまとめた表を表 4.1 に示す。

表 4.1 受検者数と各学年のテストレベル

学年	受検者数（人）		テスト項目のレベル		
	国語	算数・数学			
G1(小4)	411	410	a	b	st
G2(小5)	386	385	b	c	st
G3(小6)	391	391	c	d	st
G4(中1)	415	413	d	e	st
G5(中2)	390	397	e	f	st

4.2.2 項目分析と前処理

IRT のパラメタ推定をおこなう前に国語と算数・数学 (以下，数学) の項目分析をおこなった。確認した事項は項目通過率 (Passing Rate)，項目無回答率 (No Response Rate)，点双列相関係数 (Point Biserial Correlation)，クロンバックの α 係数，である。結果は表 4.2〜4.9 に示す。

表 4.2 項目通過率（国語）

PassingRate

ITEM	G1	G2	G3	G4	G5	ITEM	G1	G2	G3	G4	G5
a101	0.96088	NA	NA	NA	NA	d106	NA	NA	0.61214	0.7385	NA
a102	0.455	NA	NA	NA	NA	d107	NA	NA	0.58311	0.60775	NA
a103	0.58015	NA	NA	NA	NA	d108	NA	NA	0.53639	0.56934	NA
a104	0.88413	NA	NA	NA	NA	d109	NA	NA	0.65426	0.79419	NA
a105	0.67157	NA	NA	NA	NA	d110	NA	NA	0.74271	0.78641	NA
a106	0.62654	NA	NA	NA	NA	e101	NA	NA	NA	0.96117	0.98241
a107	0.31233	NA	NA	NA	NA	e102	NA	NA	NA	0.89806	0.93216
a108	0.75455	NA	NA	NA	NA	e103	NA	NA	NA	0.73366	0.80151
a109	0.63043	NA	NA	NA	NA	e104	NA	NA	NA	0.92494	0.95226
a110	0.52299	NA	NA	NA	NA	e105	NA	NA	NA	0.9322	0.9597
b101	0.72135	0.85753	NA	NA	NA	e106	NA	NA	NA	0.79747	0.85533
b102	0.5578	0.69706	NA	NA	NA	e107	NA	NA	NA	0.6335	0.77078
b103	0.86179	0.93443	NA	NA	NA	e108	NA	NA	NA	0.54412	0.52405
b104	0.88148	0.93158	NA	NA	NA	e109	NA	NA	NA	0.59804	0.60759
b105	0.58421	0.69613	NA	NA	NA	e110	NA	NA	NA	0.94578	0.85417
b106	0.39726	0.55769	NA	NA	NA	f101	NA	NA	NA	NA	0.61631
b107	0.39776	0.48266	NA	NA	NA	f102	NA	NA	NA	NA	0.57592
b108	0.66237	0.78652	NA	NA	NA	f103	NA	NA	NA	NA	0.7832
b109	0.16045	0.25912	NA	NA	NA	f104	NA	NA	NA	NA	0.86683
b110	0.5393	0.58017	NA	NA	NA	f105	NA	NA	NA	NA	0.9194
c101	NA	0.47532	0.59847	NA	NA	f106	NA	NA	NA	NA	0.79088
c102	NA	0.42408	0.55013	NA	NA	f107	NA	NA	NA	NA	0.90452
c103	NA	0.52344	0.63846	NA	NA	f108	NA	NA	NA	NA	0.67506
c104	NA	0.625	0.69231	NA	NA	f109	NA	NA	NA	NA	0.54474
c105	NA	0.79581	0.82776	NA	NA	f110	NA	NA	NA	NA	0.48744
c106	NA	0.75526	0.82902	NA	NA	st101	0.00993	0.01613	0.01681	0.0844	0.08549
c107	NA	0.68966	0.7732	NA	NA	st102	0.58082	0.66205	0.79843	0.82039	0.88665
c108	NA	0.73041	0.79348	NA	NA	st103	0.3454	0.38936	0.39737	0.4733	0.55164
c109	NA	0.73377	0.82834	NA	NA	st104	0.87834	0.86479	0.87701	0.90594	0.87626
c110	NA	0.49789	0.67584	NA	NA	st105	0.38439	0.507	0.52895	0.56311	0.66667
d101	NA	NA	0.59151	0.5495	NA	st106	0.60933	0.77374	0.83465	0.87379	0.92695
d102	NA	NA	0.4511	0.52744	NA	st107	0.35542	0.31143	0.41534	0.33496	0.33924
d103	NA	NA	0.98404	0.96585	NA	st108	0.0607	0.06845	0.13492	0.16302	0.20707
d104	NA	NA	0.81742	0.90571	NA	st109	0.58108	0.75758	0.77867	0.84634	0.89848
d105	NA	NA	0.56782	0.70712	NA	st110	0.61592	0.70909	0.7861	0.74939	0.82025

表 4.3 項目無回答率（国語）

NoResponseRate

ITEM	G1	G2	G3	G4	G5	ITEM	G1	G2	G3	G4	G5
a101	0.00487	NA	NA	NA	NA	d106	NA	NA	0.03069	0.00482	NA
a102	0.02676	NA	NA	NA	NA	d107	NA	NA	0.03069	0.00482	NA
a103	0.0438	NA	NA	NA	NA	d108	NA	NA	0.05115	0.00964	NA
a104	0.03406	NA	NA	NA	NA	d109	NA	NA	0.03836	0.00482	NA
a105	0.0073	NA	NA	NA	NA	d110	NA	NA	0.03581	0.00723	NA
a106	0.00973	NA	NA	NA	NA	e101	NA	NA	NA	0.00723	0
a107	0.11192	NA	NA	NA	NA	e102	NA	NA	NA	0.00723	0
a108	0.46472	NA	NA	NA	NA	e103	NA	NA	NA	0.00482	0
a109	0.44039	NA	NA	NA	NA	e104	NA	NA	NA	0.00482	0
a110	0.57664	NA	NA	NA	NA	e105	NA	NA	NA	0.00482	0.00251
b101	0.06569	0.05685	NA	NA	NA	e106	NA	NA	NA	0.04819	0.01005
b102	0.15815	0.12145	NA	NA	NA	e107	NA	NA	NA	0.00723	0.00251
b103	0.10219	0.05426	NA	NA	NA	e108	NA	NA	NA	0.01687	0.00754
b104	0.0146	0.01809	NA	NA	NA	e109	NA	NA	NA	0.01687	0.00754
b105	0.07543	0.0646	NA	NA	NA	e110	NA	NA	NA	0.2	0.03518
b106	0.11192	0.05943	NA	NA	NA	f101	NA	NA	NA	NA	0.16834
b107	0.13139	0.10594	NA	NA	NA	f102	NA	NA	NA	NA	0.0402
b108	0.05596	0.0801	NA	NA	NA	f103	NA	NA	NA	NA	0.07286
b109	0.34793	0.29199	NA	NA	NA	f104	NA	NA	NA	NA	0
b110	0.10219	0.1137	NA	NA	NA	f105	NA	NA	NA	NA	0.00251
c101	NA	0.00517	0	NA	NA	f106	NA	NA	NA	NA	0.06281
c102	NA	0.01292	0.00512	NA	NA	f107	NA	NA	NA	NA	0
c103	NA	0.00775	0.00256	NA	NA	f108	NA	NA	NA	NA	0.00251
c104	NA	0.00775	0.00256	NA	NA	f109	NA	NA	NA	NA	0.04523
c105	NA	0.01292	0.00512	NA	NA	f110	NA	NA	NA	NA	0
c106	NA	0.01809	0.01279	NA	NA	st101	0.26521	0.19897	0.08696	0.05783	0.03015
c107	NA	0.02584	0.00767	NA	NA	st102	0.11192	0.06718	0.02302	0.00723	0.00251
c108	NA	0.17571	0.05882	NA	NA	st103	0.12652	0.07752	0.02813	0.00723	0.00251
c109	NA	0.20413	0.06138	NA	NA	st104	0.18005	0.08269	0.04348	0.02651	0.00503
c110	NA	0.3876	0.16368	NA	NA	st105	0.15815	0.07752	0.02813	0.00723	0.00503
d101	NA	NA	0.03581	0.02651	NA	st106	0.16545	0.07494	0.02558	0.00723	0.00251
d102	NA	NA	0.18926	0.20964	NA	st107	0.19221	0.09561	0.03325	0.01446	0.00754
d103	NA	NA	0.03836	0.01205	NA	st108	0.23844	0.13178	0.03325	0.00964	0.00503
d104	NA	NA	0.08951	0.02892	NA	st109	0.27981	0.14729	0.04092	0.01205	0.01005
d105	NA	NA	0.18926	0.08675	NA	st110	0.29684	0.14729	0.04348	0.00964	0.00754

表 4.4 点双列相関係数（国語）

P.BIS

ITEM	G1	G2	G3	G4	G5	ITEM	G1	G2	G3	G4	G5
a101	0.24599	NA	NA	NA	NA	d106	NA	NA	0.39726	0.44381	NA
a102	0.34322	NA	NA	NA	NA	d107	NA	NA	0.31405	0.34154	NA
a103	0.47848	NA	NA	NA	NA	d108	NA	NA	0.25101	0.43229	NA
a104	0.35961	NA	NA	NA	NA	d109	NA	NA	0.48097	0.53148	NA
a105	0.42815	NA	NA	NA	NA	d110	NA	NA	0.44362	0.40686	NA
a106	0.24281	NA	NA	NA	NA	e101	NA	NA	NA	0.34157	0.286
a107	0.3038	NA	NA	NA	NA	e102	NA	NA	NA	0.40836	0.43044
a108	0.47307	NA	NA	NA	NA	e103	NA	NA	NA	0.55074	0.43174
a109	0.4392	NA	NA	NA	NA	e104	NA	NA	NA	0.44631	0.34889
a110	0.5117	NA	NA	NA	NA	e105	NA	NA	NA	0.31676	0.18524
b101	0.38947	0.40188	NA	NA	NA	e106	NA	NA	NA	0.42418	0.36405
b102	0.4205	0.44015	NA	NA	NA	e107	NA	NA	NA	0.40829	0.29039
b103	0.32854	0.36668	NA	NA	NA	e108	NA	NA	NA	0.40949	0.41766
b104	0.22591	0.20446	NA	NA	NA	e109	NA	NA	NA	0.52291	0.51297
b105	0.37549	0.33402	NA	NA	NA	e110	NA	NA	NA	0.32831	0.45531
b106	0.4802	0.6002	NA	NA	NA	f101	NA	NA	NA	NA	0.33457
b107	0.32482	0.39099	NA	NA	NA	f102	NA	NA	NA	NA	0.47082
b108	0.40239	0.36792	NA	NA	NA	f103	NA	NA	NA	NA	0.39898
b109	0.32429	0.41444	NA	NA	NA	f104	NA	NA	NA	NA	0.43577
b110	0.16842	0.11704	NA	NA	NA	f105	NA	NA	NA	NA	0.36009
c101	NA	0.31736	0.35334	NA	NA	f106	NA	NA	NA	NA	0.40796
c102	NA	0.30761	0.33818	NA	NA	f107	NA	NA	NA	NA	0.34635
c103	NA	0.31716	0.43369	NA	NA	f108	NA	NA	NA	NA	0.52732
c104	NA	0.35604	0.3669	NA	NA	f109	NA	NA	NA	NA	0.52989
c105	NA	0.42679	0.42513	NA	NA	f110	NA	NA	NA	NA	0.46233
c106	NA	0.40942	0.34335	NA	NA	st101	0.06332	0.10964	0.12593	0.20102	0.19829
c107	NA	0.50769	0.53245	NA	NA	st102	0.49221	0.47583	0.47768	0.50364	0.48924
c108	NA	0.46233	0.35452	NA	NA	st103	0.22397	0.25861	0.22975	0.34	0.36297
c109	NA	0.53236	0.45472	NA	NA	st104	0.32526	0.2908	0.25162	0.19748	0.20698
c110	NA	0.47074	0.37755	NA	NA	st105	0.199	0.2719	0.30047	0.40119	0.30969
d101	NA	NA	0.43968	0.43628	NA	st106	0.43842	0.48346	0.44938	0.46176	0.43552
d102	NA	NA	0.4435	0.49097	NA	st107	0.34292	0.13316	0.33063	0.32434	0.27859
d103	NA	NA	0.17174	0.27808	NA	st108	-0.053	-0.0107	0.06989	0.11272	0.26645
d104	NA	NA	0.44735	0.31123	NA	st109	0.42887	0.48501	0.45755	0.45342	0.34013
d105	NA	NA	0.47206	0.3874	NA	st110	0.4687	0.41614	0.34204	0.43282	0.17775

表 4.5 クロンバックのα係数（国語）

Cronbach_alpha	
G1	0.85895
G2	0.76882
G3	0.77534
G4	0.75167
G5	0.74612

表 4.6 項目通過率（数学）

PassingRate

ITEM	G1	G2	G3	G4	G5	ITEM	G1	G2	G3	G4	G5
a101	0.9656	NA	NA	NA	NA	d106	NA	NA	0.47075	0.35484	NA
a102	0.38987	NA	NA	NA	NA	d107	NA	NA	0.45977	0.58867	NA
a103	0.71814	NA	NA	NA	NA	d108	NA	NA	0.53687	0.66256	NA
a104	0.67901	NA	NA	NA	NA	d109	NA	NA	0.56347	0.55844	NA
a105	0.80732	NA	NA	NA	NA	d110	NA	NA	0.03968	0.08929	NA
a106	0.50129	NA	NA	NA	NA	e101	NA	NA	NA	0.70844	0.78571
a107	0.34383	NA	NA	NA	NA	e102	NA	NA	NA	0.41005	0.39295
a108	0.73935	NA	NA	NA	NA	e103	NA	NA	NA	0.67671	0.77717
a109	0.70171	NA	NA	NA	NA	e104	NA	NA	NA	0.34694	0.42785
a110	0.35509	NA	NA	NA	NA	e105	NA	NA	NA	0.34097	0.43544
b101	0.70106	0.91906	NA	NA	NA	e106	NA	NA	NA	0.6456	0.73315
b102	0.72319	0.60763	NA	NA	NA	e107	NA	NA	NA	0.52865	0.48072
b103	0.28173	0.55556	NA	NA	NA	e108	NA	NA	NA	0.00699	0.04478
b104	0.6409	0.65885	NA	NA	NA	e109	NA	NA	NA	0.19415	0.25195
b105	0.45953	0.55153	NA	NA	NA	e110	NA	NA	NA	0.32609	0.36042
b106	0.29146	0.50394	NA	NA	NA	f101	NA	NA	NA	NA	0.72951
b107	0.32514	0.51862	NA	NA	NA	f102	NA	NA	NA	NA	0.64417
b108	0.36041	0.47644	NA	NA	NA	f103	NA	NA	NA	NA	0.4434
b109	0.65181	0.79944	NA	NA	NA	f104	NA	NA	NA	NA	0.88804
b110	0.30147	0.51242	NA	NA	NA	f105	NA	NA	NA	NA	0.41237
c101	NA	0.83198	0.79301	NA	NA	f106	NA	NA	NA	NA	0.48541
c102	NA	0.71429	0.72987	NA	NA	f107	NA	NA	NA	NA	0.52051
c103	NA	0.53372	0.54645	NA	NA	f108	NA	NA	NA	NA	0.52756
c104	NA	0.59884	0.75668	NA	NA	f109	NA	NA	NA	NA	0.0463
c105	NA	0.016	0.09319	NA	NA	f110	NA	NA	NA	NA	0.02165
c106	NA	0.12647	0.43117	NA	NA	st101	0.79012	0.81016	0.7974	0.89216	0.94177
c107	NA	0.20712	0.34435	NA	NA	st102	0.4625	0.71875	0.77692	0.83495	0.89394
c108	NA	0.38768	0.65761	NA	NA	st103	0.22306	0.23342	0.30435	0.30024	0.32746
c109	NA	0.05426	0.32258	NA	NA	st104	0.06691	0.46875	0.66482	0.70635	0.72877
c110	NA	0.17588	0.33856	NA	NA	st105	0.00692	0.19101	0.21237	0.18342	0.17922
d101	NA	NA	0.92992	0.8963	NA	st106	0.24229	0.18391	0.54266	0.71739	0.70605
d102	NA	NA	0.33333	0.41162	NA	st107	0.22074	0.34375	0.28495	0.25735	0.25
d103	NA	NA	0.38247	0.49363	NA	st108	0.39695	0.41546	0.5784	0.6391	0.60914
d104	NA	NA	0.74788	0.74365	NA	st109	0.03448	0	0.03723	0.08621	0.20977
d105	NA	NA	0.01807	0.04839	NA	st110	0.53571	0.69663	0.69432	0.87069	0.92351

表 4.7 項目無回答率（数学）

NoResponseRate

ITEM	G1	G2	G3	G4	G5	ITEM	G1	G2	G3	G4	G5
a101	0.00732	NA	NA	NA	NA	d106	NA	NA	0.08184	0.02421	NA
a102	0.03659	NA	NA	NA	NA	d107	NA	NA	0.10997	0.01695	NA
a103	0.00488	NA	NA	NA	NA	d108	NA	NA	0.13299	0.01695	NA
a104	0.0122	NA	NA	NA	NA	d109	NA	NA	0.17391	0.0678	NA
a105	0	NA	NA	NA	NA	d110	NA	NA	0.3555	0.18644	NA
a106	0.05122	NA	NA	NA	NA	e101	NA	NA	NA	0.05327	0.01259
a107	0.07073	NA	NA	NA	NA	e102	NA	NA	NA	0.08475	0.07053
a108	0.02683	NA	NA	NA	NA	e103	NA	NA	NA	0.11622	0.07305
a109	0.00244	NA	NA	NA	NA	e104	NA	NA	NA	0.05085	0.00504
a110	0.06585	NA	NA	NA	NA	e105	NA	NA	NA	0.04843	0.00504
b101	0.07805	0.00519	NA	NA	NA	e106	NA	NA	NA	0.11864	0.06549
b102	0.02195	0.04675	NA	NA	NA	e107	NA	NA	NA	0.07022	0.02015
b103	0.03902	0.01818	NA	NA	NA	e108	NA	NA	NA	0.30751	0.15617
b104	0.02195	0.0026	NA	NA	NA	e109	NA	NA	NA	0.08959	0.03023
b105	0.06585	0.06753	NA	NA	NA	e110	NA	NA	NA	0.33172	0.28715
b106	0.02927	0.01039	NA	NA	NA	f101	NA	NA	NA	NA	0.07809
b107	0.10732	0.02338	NA	NA	NA	f102	NA	NA	NA	NA	0.17884
b108	0.03902	0.00779	NA	NA	NA	f103	NA	NA	NA	NA	0.46599
b109	0.12439	0.06753	NA	NA	NA	f104	NA	NA	NA	NA	0.01008
b110	0.33659	0.16364	NA	NA	NA	f105	NA	NA	NA	NA	0.02267
c101	NA	0.04156	0.04859	NA	NA	f106	NA	NA	NA	NA	0.05038
c102	NA	0.03636	0.01535	NA	NA	f107	NA	NA	NA	NA	0.01763
c103	NA	0.11429	0.06394	NA	NA	f108	NA	NA	NA	NA	0.0403
c104	NA	0.10649	0.04348	NA	NA	f109	NA	NA	NA	NA	0.45592
c105	NA	0.35065	0.28645	NA	NA	f110	NA	NA	NA	NA	0.41814
c106	NA	0.11688	0.01535	NA	NA	st101	0.0122	0.02857	0.01535	0.01211	0.00504
c107	NA	0.1974	0.07161	NA	NA	st102	0.02439	0.0026	0.00256	0.00242	0.00252
c108	NA	0.28312	0.05882	NA	NA	st103	0.02683	0.02078	0	0	0
c109	NA	0.32987	0.12788	NA	NA	st104	0.3439	0.16883	0.07673	0.08475	0.0806
c110	NA	0.48312	0.18414	NA	NA	st105	0.29512	0.07532	0.04859	0.03632	0.03023
d101	NA	NA	0.05115	0.01937	NA	st106	0.44634	0.54805	0.25064	0.10896	0.12594
d102	NA	NA	0.08696	0.04116	NA	st107	0.27073	0.33506	0.04859	0.01211	0.00252
d103	NA	NA	0.35806	0.23971	NA	st108	0.36098	0.46234	0.26598	0.0339	0.00756
d104	NA	NA	0.09719	0.046	NA	st109	0.50488	0.5974	0.51918	0.15738	0.12343
d105	NA	NA	0.57545	0.39952	NA	st110	0.45366	0.53766	0.41432	0.15738	0.11083

表 4.8 点双列相関係数（数学）

P.BIS

ITEM	G1	G2	G3	G4	G5	ITEM	G1	G2	G3	G4	G5
a101	0.13023	NA	NA	NA	NA	d106	NA	NA	0.29065	0.1787	NA
a102	0.49835	NA	NA	NA	NA	d107	NA	NA	0.30196	0.35926	NA
a103	0.43231	NA	NA	NA	NA	d108	NA	NA	0.38549	0.44032	NA
a104	0.5468	NA	NA	NA	NA	d109	NA	NA	0.39564	0.47084	NA
a105	0.31672	NA	NA	NA	NA	d110	NA	NA	0.15272	0.28793	NA
a106	0.46308	NA	NA	NA	NA	e101	NA	NA	NA	0.53338	0.48773
a107	0.52772	NA	NA	NA	NA	e102	NA	NA	NA	0.50577	0.50803
a108	0.52249	NA	NA	NA	NA	e103	NA	NA	NA	0.46509	0.41441
a109	0.41989	NA	NA	NA	NA	e104	NA	NA	NA	0.23497	0.18454
a110	0.3963	NA	NA	NA	NA	e105	NA	NA	NA	0.07036	0.365
b101	0.27422	0.26855	NA	NA	NA	e106	NA	NA	NA	0.55069	0.53031
b102	0.37178	0.44658	NA	NA	NA	e107	NA	NA	NA	0.39535	0.54299
b103	0.45208	0.48987	NA	NA	NA	e108	NA	NA	NA	0.01742	0.33627
b104	0.39366	0.40764	NA	NA	NA	e109	NA	NA	NA	0.01501	0.28228
b105	0.50645	0.49729	NA	NA	NA	e110	NA	NA	NA	0.60629	0.5863
b106	0.30987	0.36064	NA	NA	NA	f101	NA	NA	NA	NA	0.50524
b107	0.26508	0.33972	NA	NA	NA	f102	NA	NA	NA	NA	0.49632
b108	0.4049	0.60978	NA	NA	NA	f103	NA	NA	NA	NA	0.60834
b109	0.48333	0.43505	NA	NA	NA	f104	NA	NA	NA	NA	0.14916
b110	0.46415	0.50714	NA	NA	NA	f105	NA	NA	NA	NA	0.39117
c101	NA	0.36585	0.40316	NA	NA	f106	NA	NA	NA	NA	0.34916
c102	NA	0.4905	0.51091	NA	NA	f107	NA	NA	NA	NA	0.49018
c103	NA	0.55899	0.5717	NA	NA	f108	NA	NA	NA	NA	0.4703
c104	NA	0.5186	0.5323	NA	NA	f109	NA	NA	NA	NA	0.35405
c105	NA	0.30832	0.4491	NA	NA	f110	NA	NA	NA	NA	0.2159
c106	NA	0.40158	0.52199	NA	NA	st101	0.41477	0.43338	0.4201	0.29169	0.27704
c107	NA	0.17403	0.18327	NA	NA	st102	0.40249	0.39582	0.45867	0.41176	0.27128
c108	NA	0.42246	0.38362	NA	NA	st103	0.41487	0.39665	0.39022	0.46583	0.53772
c109	NA	0.43162	0.58441	NA	NA	st104	0.39917	0.62509	0.51922	0.57291	0.50121
c110	NA	0.45857	0.53677	NA	NA	st105	0.14335	0.45059	0.51194	0.44487	0.44252
d101	NA	NA	0.21955	0.26721	NA	st106	0.41484	0.55075	0.63142	0.58029	0.56288
d102	NA	NA	0.54201	0.56833	NA	st107	-0.0267	0.20567	0.19339	0.37814	0.39288
d103	NA	NA	0.50338	0.62424	NA	st108	0.47644	0.45893	0.46087	0.44452	0.43245
d104	NA	NA	0.50917	0.60198	NA	st109	0.09373	NA	0.1345	0.30346	0.44888
d105	NA	NA	0.27237	0.29516	NA	st110	0.53413	0.51019	0.43749	0.42157	0.28479

表 4.9 クロンバックのα係数（数学）

Cronbach_alpha	
G1	0.87803
G2	0.90301
G3	0.87384
G4	0.83722
G5	0.83998

項目通過率が極端に低い項目が，国語で 2 項目 (st101=.007~.08, st108=.05~.20)，数学で 7 項目 (c105=.010~.067, d110=.026~.073, e108=.005~.038, f109=.025, f110=.013, st105=.005~.202, st109=0.0~.184) 存在した。なお st109 は小学 5 年生で通過率が 0 となっていた。さらに国語の d101, st104, st107 と数学の st105 はすべて小学 6 年生の学年（G3）でピークを迎えており，通過率が単調増加しなかった。項目通過率が異常に低い項目は，項目自体の難易度設定が難しすぎるか，問題の文章や設定に何らかの問題がある可能性があり，IRT のパラメタ推定を不安定にするひとつの要因である。通過率が単調増加しない項目は特定の学年に有利に働く項目が存在する可能性を示し，IRT の正答確率の単調増加性の仮定を脅かす可能性がある。

点双列相関係数が低い項目は，国語で 2 項目 (st101=.073~.202 , st108=.0003~.269)，数学で 2 項目 (e108=.026~.326, e109=.050~.287) であった。国語の 2 項目と数学の e108 については項目通過率が極端に低いことも影響して，点双列相関係数が不安定になっている可能性がある。

項目無回答率は，国語も数学も 0.5 を超える項目が複数項目存在した。

クロンバックの α 係数はすべての学年のテストで 0.8 付近の値を示しているため，信頼性係数の推定値に大きな問題はないと判断した。

事前の項目分析では，どちらのテストも削除した方が良いと思われる項目が少数個検出された。しかし，尺度全体で 70 項目しか存在せず，項目をひとつ削ることによる尺度全体の情報の損失もあるため，一度 IRT のパラメタ推定をおこない，推定値および推定誤差，適合度などの指標をもとに再検討することとした。

最後に，IRT のパラメタ推定を実行する前にテストの測定の一次元性を，テトラコリック相関係数の固有値の減衰状況より確認した。計画的な欠測を含むデータの場合尺度全体で固有値を計算することはできないため，まずは学年ごとに一次元性の確認をおこない，続いて尺度化テスト項目でも一次元性の確認をおこない，両者の情報から総合的に判断する。計算結果は図 4.17〜4.20 に示した通りである。国語も数学も第一固有値から第二固有値にかけて大きく値が減少しており，概ね測定の一次元性は保証できるものと考える。

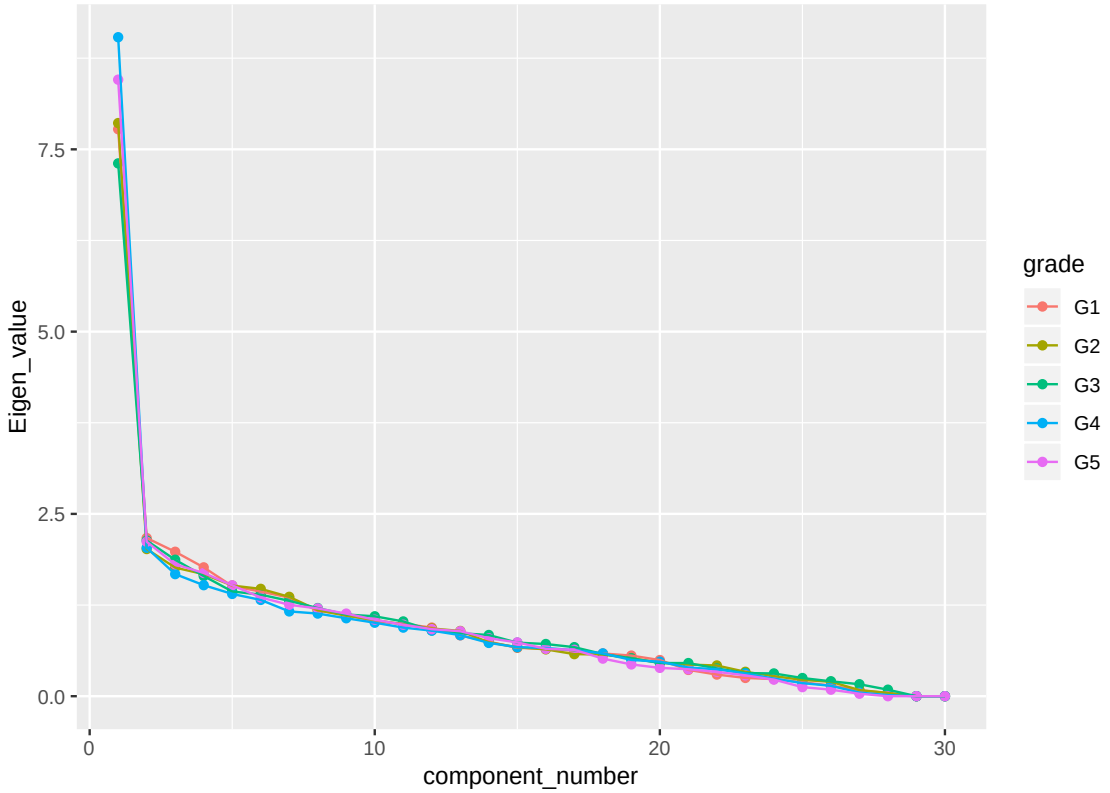

図 4.17 テスト冊子ごとの固有値の減衰状況（国語）

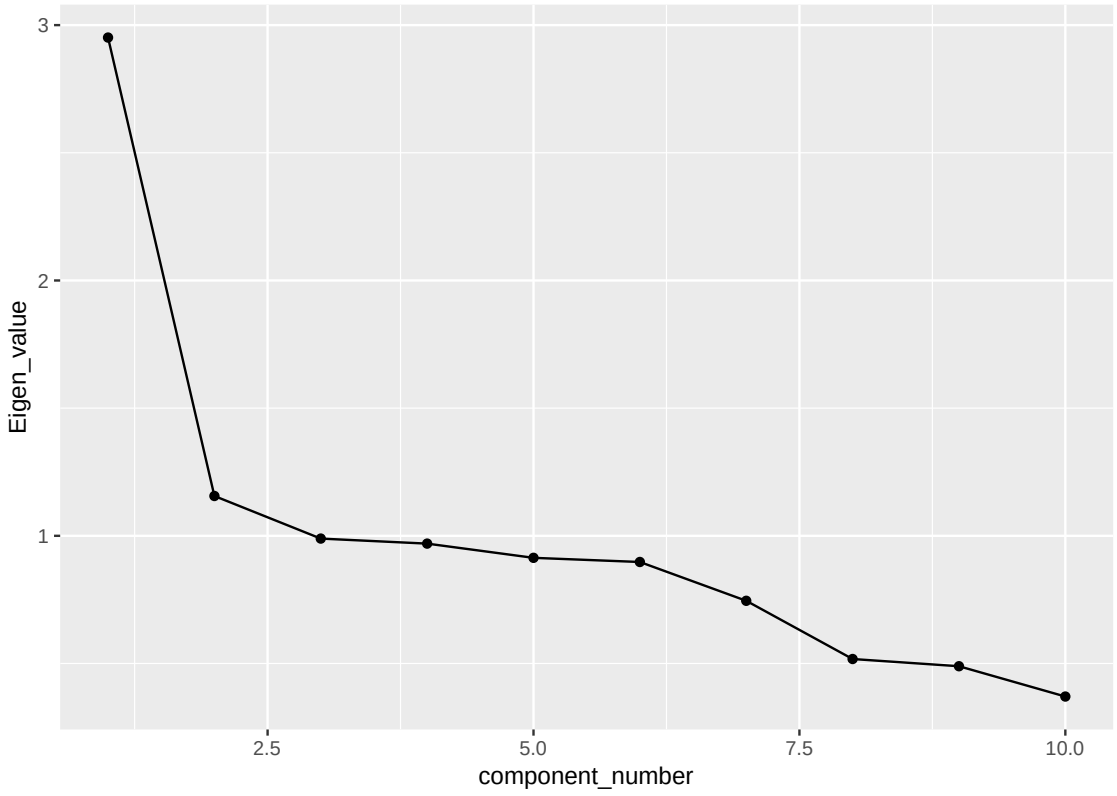

図 4.18 尺度化テストの固有値の減衰状況（国語）

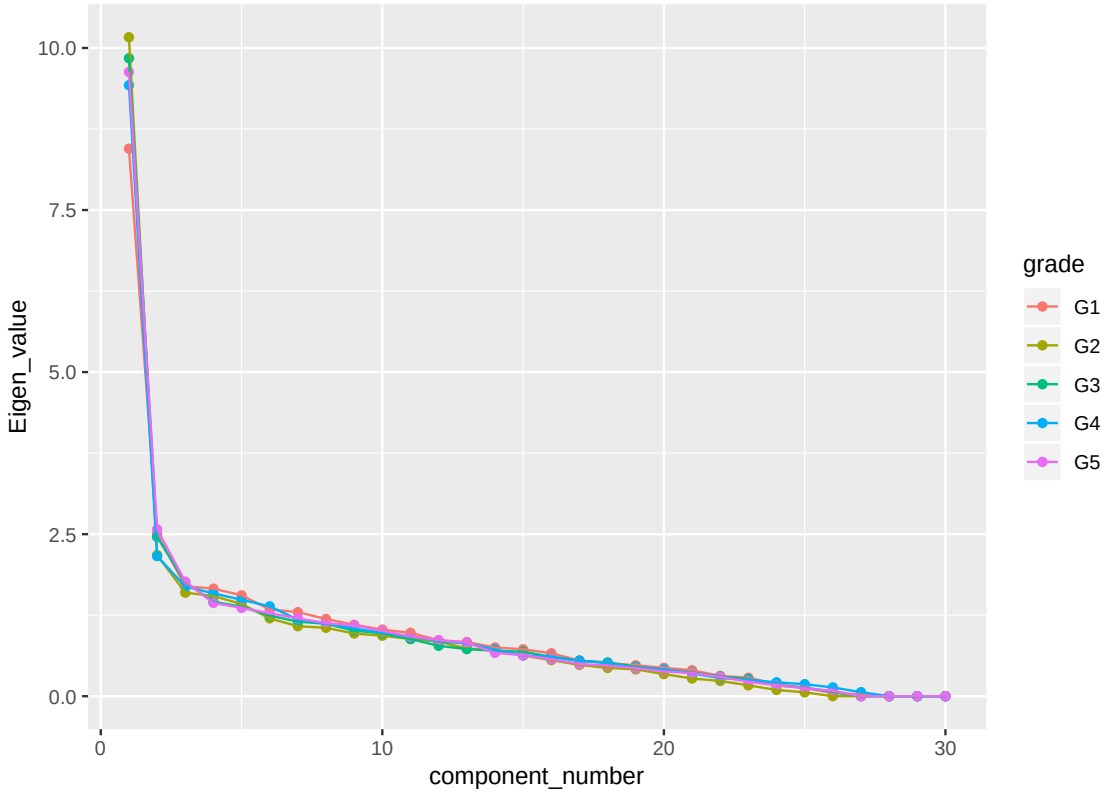

図 4.19 テスト冊子ごとの固有値の減衰状況（数学）

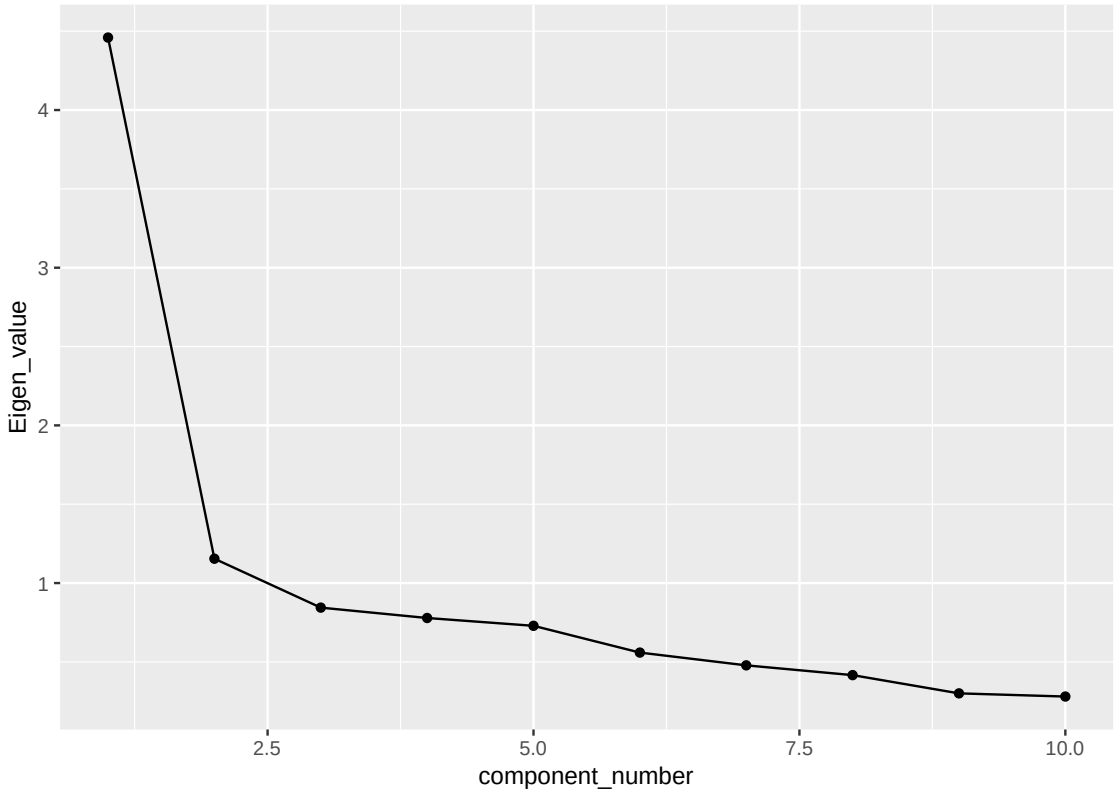

図 4.20 尺度化テストの固有値の減衰状況（数学）

4.2.3 パラメタ推定と局所項目依存および項目適合度統計量の確認

まずはすべての項目を使って MMLE-EM を用いた同時尺度調整を実行した。基準となる集団を G3 と設定し，学年分布が平均 0，標準偏差 1 となるように固定した。項目パラメタと周辺対数尤度の変化率のそれぞれを収束基準として，国語は 111 回，数学は 87 回で収束と判断された。次に，暫定的な項目パラメタから受検者の能力パラメタを推定する。欠測値を除いてすべての項目に正答あるいは誤答している受検者が存在していたため，EAP 推定値を採用したが，事前分布による縮小の影響を抑えるために弱情報事前分布として$N(0,3^2)$を採用した。

推定された項目パラメタ，能力パラメタ，および項目反応データを基にQ_3統計量 (Yen, 1993) を計算した。LID の基準は残差相関の絶対値が 0.2 以上であることとした。どちらの教科においても局所項目依存が確認された項目ペアは存在しなかった。Q_3の数値については欠測値を除いても全部で 2,830 通りの組み合わせがあるため，紙幅の都合で掲載しない。

次に項目適合度を推定した(表 4.10, 4.11)。いくつか指標があるが, OUTFIT 統計量と INFIT 統計量，およびそれらを標準化した統計量の計 4 つを使用した。4 つの指標すべてで大きく逸脱していると判断できそうな項目は国語の a101 と a104, st101 であった。また数学の a101, a104, c105, d105, e108, f109, f110 もモデルから逸脱した項目であると考えられる。

最後に，項目の適合度を視覚的に把握するために観測正答率と予測正答率のプロットを全項目描画した。黒の実線と点が観測正答率であり，同じ枠の中に書かれている色づけされた実線がモデルによる予測正答率である。国語 (図 4.21) は a109 と a110 の項目がモデルのロジスティック曲線から乖離気味である。数学 (図 4.22) は c107, e109 で同じくやや乖離気味である。

国語の a101 は G1 の最初に配置されているテスト項目であり，文章中から擬態語を書き抜くという非常に簡単な項目で，通過率がほぼ 1 に近かったため，適合度が低く計算されたと考えられる。a109, a110, f110 は INFIT と OUTFIT で見るとやや高めの数値ではあるが，トレースプロットではそこまで不自然な上がり方はしていない。

数学の c105 と d105 は非常に OUTFIT 統計量の値が大きいが，推定されたパラメタを見ると識別力も高く，トレースプロットを見てもそこまで不自然な正答率の上がり方をしていないため削除しなくてもよいと考えられる。また，c105 や e108, f109, f110 も同様であると考えられる。これらは通過率が非常に低く，よほど能力の高い受検者しか正答できていない。このことが適合度を下げた要因であると考えられるが，単に識別力が高い項目となるだけなら削除する必要なないと考えられる。逆に e109 は正答確率があまり増加せず，非常に識別力の低い項目である。

項目分析でモデルから逸脱している可能性が示唆された項目の一覧を表 4.12 と 4.13 にまとめた。最終的に複数の指標で IRT モデルにフィットしない可能性が指摘された項目を削除することとした。すなわち国語は st101 と st108 を削除し，数学については f109 を残し，e108 と e109 を削除して最終的な尺度調整をおこなう。

表 4.10 項目適合度（国語）

Item	N	InFit	StdInFit	OutFit	StdOutFit	Item	N	InFit	StdInFit	OutFit	StdOutFit
a101	409	1.297	1.107	1.633	8.026	c106	766	0.991	-0.151	0.973	-0.539
a102	400	0.951	-1.488	0.926	-1.059	c107	765	0.940	-1.117	0.775	-4.695
a103	393	0.974	-0.483	0.901	-1.419	c108	687	1.001	0.030	0.913	-1.652
a104	397	1.177	1.201	1.624	7.805	c109	675	0.983	-0.259	0.849	-2.898
a105	408	1.039	0.697	1.030	0.429	c110	564	0.959	-1.154	0.928	-1.239
a106	407	1.015	0.436	1.024	0.345	d101	781	0.934	-2.905	0.915	-1.707
a107	365	0.925	-1.838	0.894	-1.476	d102	645	0.910	-3.355	0.878	-2.265
b101	749	1.060	0.976	0.991	-0.168	d103	786	1.045	0.279	0.939	-1.236
b102	686	0.994	-0.151	0.955	-0.837	d104	759	0.957	-0.560	0.868	-2.665
b103	735	1.144	1.331	0.957	-0.839	d105	696	0.915	-2.588	0.871	-2.496
b104	785	1.106	1.018	1.093	1.807	d106	792	0.937	-1.832	0.870	-2.670
b105	742	1.001	0.030	0.969	-0.598	d107	792	0.963	-1.834	0.955	-0.902
b106	729	0.877	-3.188	0.773	-4.630	d108	782	0.955	-2.323	0.946	-1.072
b107	703	0.958	-1.942	0.953	-0.898	d109	789	0.907	-2.032	0.754	-5.248
b108	744	1.031	0.669	1.026	0.495	d110	789	0.941	-1.239	0.839	-3.334
b109	542	0.866	-2.519	0.713	-5.140	e101	810	1.010	0.115	0.481	-12.585
b110	712	1.002	0.165	1.003	0.059	e102	810	0.917	-0.838	0.662	-7.543
st101	1746	1.129	1.162	0.887	-3.453	e103	811	0.862	-2.926	0.735	-5.760
st102	1917	0.943	-1.703	0.789	-6.916	e104	811	0.994	-0.006	0.397	-15.356
st103	1905	0.971	-2.530	0.973	-0.823	e105	810	0.957	-0.291	0.818	-3.859
st104	1866	1.000	0.014	0.977	-0.716	e106	789	0.912	-1.514	0.798	-4.245
st105	1891	0.964	-2.991	0.959	-1.270	e107	809	0.932	-2.131	0.906	-1.941
st106	1891	0.981	-0.459	0.792	-6.782	e108	803	0.934	-3.379	0.926	-1.505
st107	1864	0.983	-1.049	0.979	-0.642	e109	803	0.861	-4.652	0.822	-3.740
st108	1834	1.009	0.201	1.020	0.615	e110	716	0.938	-0.671	0.723	-5.688
st109	1805	0.961	-1.067	0.872	-3.967	f101	331	0.942	-1.794	0.933	-0.877
st110	1799	0.959	-1.458	0.958	-1.287	f102	382	0.892	-2.945	0.852	-2.132
a108	220	1.048	0.471	1.200	2.000	f103	369	0.911	-1.314	0.806	-2.773
a109	230	1.023	0.358	0.985	-0.157	f104	398	0.878	-1.236	0.762	-3.592
a110	174	0.980	-0.255	0.912	-0.840	f105	397	0.894	-0.733	0.758	-3.658
c101	776	0.971	-1.414	0.965	-0.686	f106	373	0.905	-1.374	0.839	-2.298
c102	771	0.962	-1.860	0.957	-0.851	f107	398	0.907	-0.721	0.801	-2.971
c103	774	0.962	-1.413	0.943	-1.147	f108	397	0.833	-3.316	0.780	-3.299
c104	774	0.973	-0.788	0.937	-1.257	f109	380	0.840	-3.724	0.789	-3.083
c105	771	0.999	-0.003	0.873	-2.573	f110	398	0.890	-3.005	0.904	-1.382

表 4.11 項目適合度（数学）

Item	N	InFit	StdInFit	OutFit	StdOutFit	Item	N	InFit	StdInFit	OutFit	StdOutFit
st101	1967	0.978	-0.491	0.961	-1.225	c106	725	0.904	-2.080	0.912	-1.709
a101	407	1.084	0.400	1.170	2.335	c107	672	0.989	-0.290	0.991	-0.166
a102	395	0.931	-1.554	0.879	-1.757	c108	644	0.966	-1.131	0.956	-0.797
a103	408	1.019	0.313	1.025	0.354	c109	599	0.853	-2.114	0.666	-6.412
a104	405	0.998	-0.005	0.935	-0.942	c110	518	0.898	-1.797	0.999	-0.010
a105	410	1.045	0.559	1.120	1.670	d101	776	0.987	-0.097	0.945	-1.101
a106	389	0.957	-0.943	0.949	-0.727	d102	753	0.907	-2.350	0.847	-3.098
a107	381	0.864	-2.484	0.748	-3.739	d103	565	0.908	-1.982	0.829	-3.002
a108	399	1.022	0.301	0.900	-1.453	d104	747	0.908	-1.779	0.732	-5.598
a109	409	1.025	0.430	1.010	0.145	d105	414	0.986	0.039	0.122	-21.435
a110	383	0.942	-1.442	0.926	-1.047	d106	762	0.995	-0.314	0.996	-0.086
st102	1982	0.942	-2.007	1.007	0.233	d107	754	0.977	-1.198	0.978	-0.423
st103	1977	0.944	-2.244	0.935	-2.078	d108	745	0.966	-1.212	0.944	-1.090
b101	761	1.053	0.886	1.001	0.018	d109	708	0.956	-1.578	0.945	-1.056
b102	768	1.003	0.110	0.992	-0.152	d110	588	1.039	0.319	0.937	-1.105
b103	772	0.927	-2.093	0.882	-2.392	e101	783	0.865	-2.887	0.938	-1.239
b104	785	0.997	-0.090	1.000	0.003	e102	747	0.933	-1.986	0.908	-1.820
b105	742	0.944	-1.701	0.878	-2.421	e103	733	0.910	-2.104	0.885	-2.259
b106	779	0.960	-1.575	0.959	-0.812	e104	787	0.999	-0.049	0.999	-0.021
b107	742	0.974	-1.158	0.976	-0.474	e105	788	0.996	-0.193	1.004	0.077
b108	776	0.917	-2.521	0.906	-1.894	e106	735	0.885	-2.692	0.810	-3.831
b109	718	0.997	-0.033	0.929	-1.376	e107	773	0.939	-2.271	0.941	-1.175
b110	594	0.926	-1.783	0.830	-3.067	e108	621	1.181	0.756	2.246	18.097
st104	1693	0.850	-4.681	0.778	-6.882	e109	761	1.005	0.128	1.012	0.224
st105	1800	0.954	-1.055	0.857	-4.471	e110	559	0.876	-2.279	0.747	-4.548
st106	1409	0.883	-3.534	0.793	-5.822	f101	366	0.868	-2.051	0.700	-4.430
st107	1731	0.989	-0.424	1.001	0.040	f102	326	0.902	-1.912	0.820	-2.416
st108	1549	0.951	-2.430	0.919	-2.292	f103	212	0.884	-1.483	0.773	-2.492
st109	1242	1.021	0.299	0.914	-2.197	f104	393	0.979	-0.130	0.977	-0.318
st110	1332	0.943	-1.366	0.914	-2.278	f105	388	0.969	-0.854	0.988	-0.164
c101	741	0.986	-0.219	0.999	-0.016	f106	377	0.976	-0.886	0.976	-0.336
c102	756	0.941	-1.261	0.896	-2.073	f107	390	0.936	-1.671	0.917	-1.178
c103	707	0.903	-2.449	0.801	-3.960	f108	381	0.931	-1.690	0.896	-1.469
c104	718	0.929	-1.504	0.846	-3.038	f109	216	1.258	0.864	0.530	-5.732
c105	529	0.799	-1.164	0.203	-19.436	f110	231	1.351	0.804	0.535	-5.845

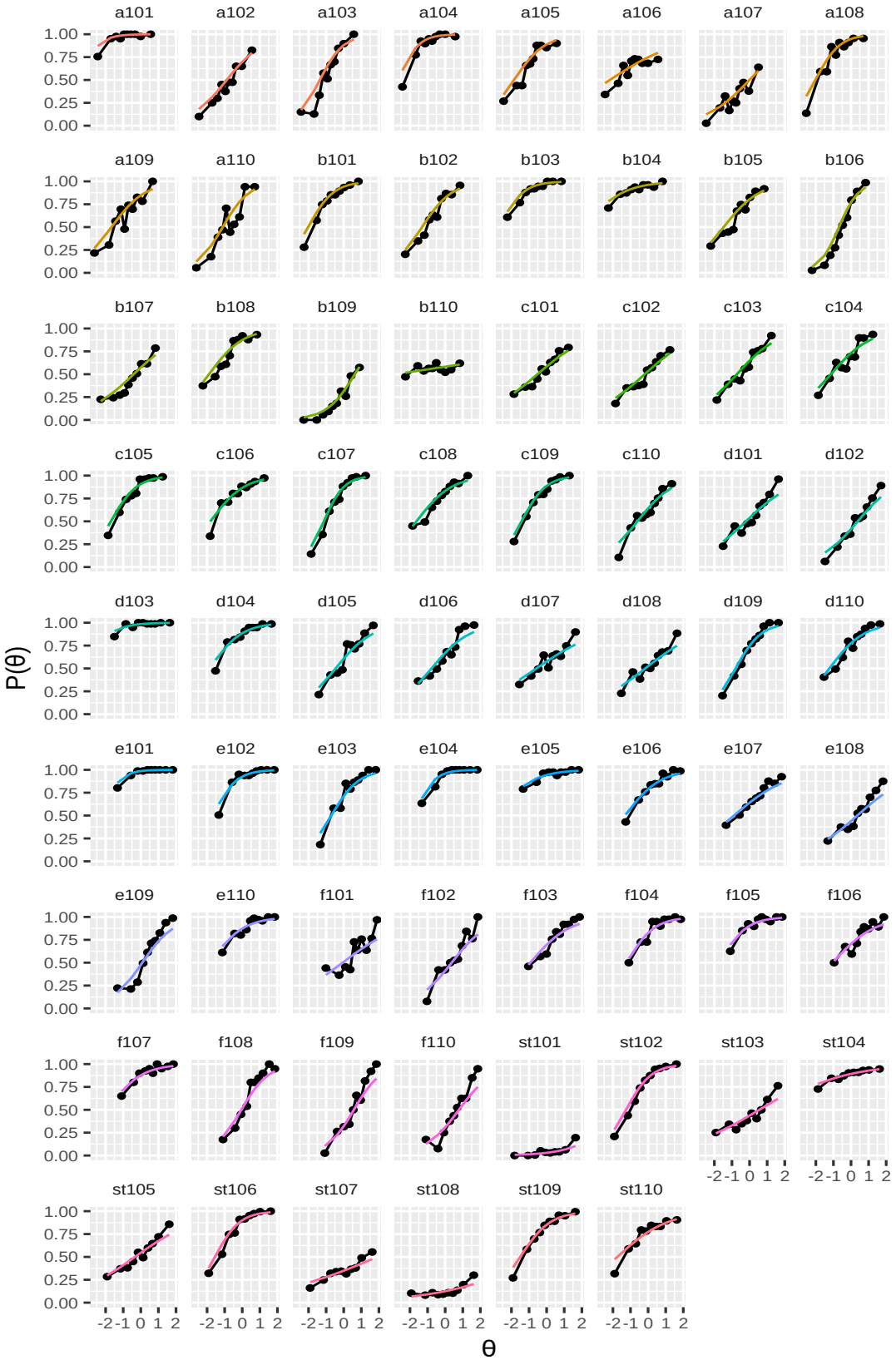

図 4.21 項目適合度プロット（国語）

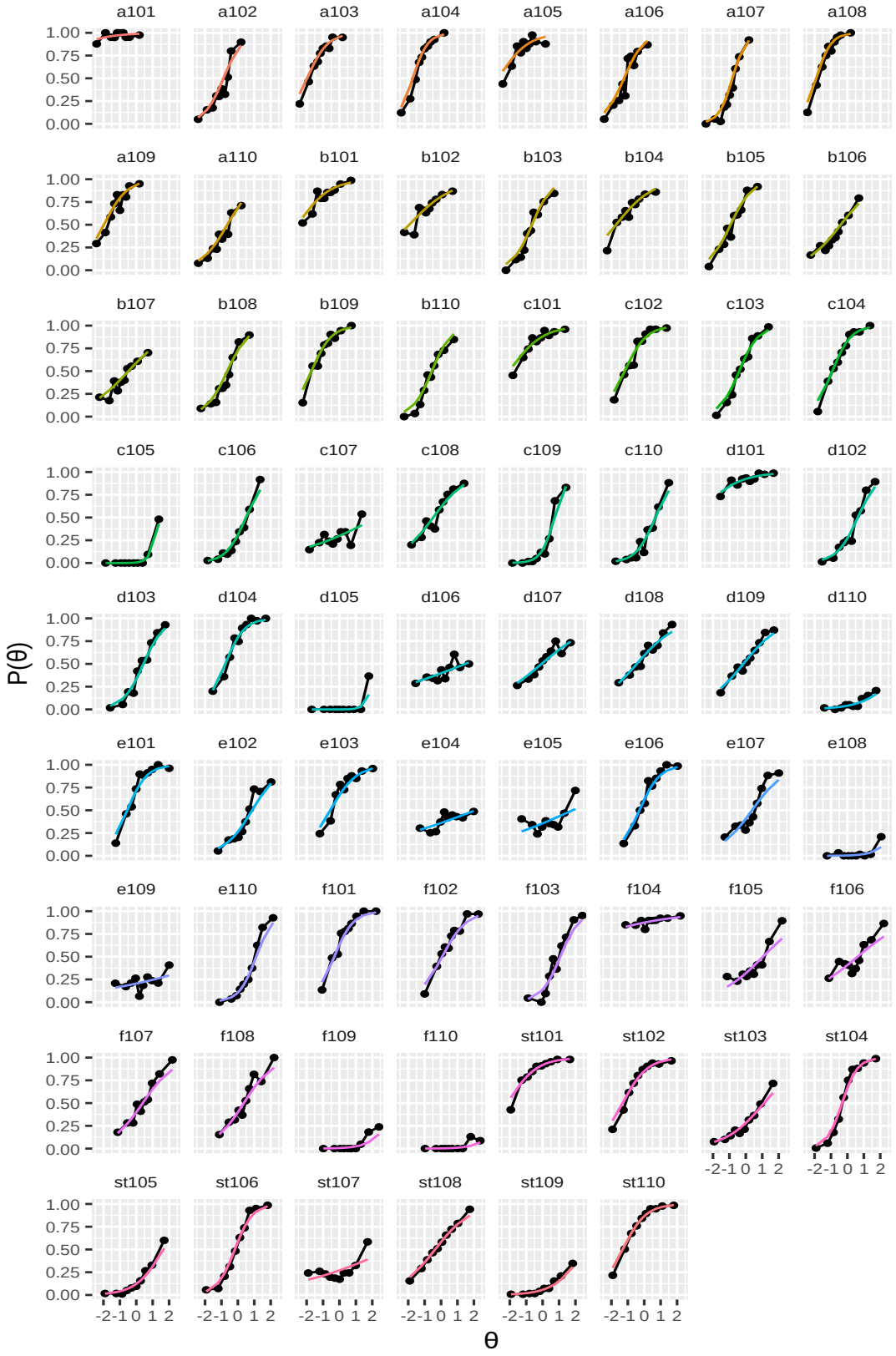

図 4.22 項目適合度プロット（数学）

表 4.12 項目分析結果（国語）

指標	逸脱項目
通過率	st101, st108
単調増加	d101, st104, st107
点双列相関係数	st101, st108
項目局所依存	-
項目適合度	a101, a104
適合度プロット	a109, a110

表 4.13 項目分析結果（数学）

指標	逸脱項目
通過率	c105, d110, c108, f109, f110, st105, st109
単調増加	st105
点双列相関係数	e108, e109
項目局所依存	-
項目適合度	a101, a104, c105, d105, e108, f109, f110
適合度プロット	c107, e109

4.2.4 項目パラメタの推定結果と推定精度および情報量

先ほど決定した削除項目を推定対象から除外して，再推定をおこなった。パラメタの推定値と推定の標準誤差（情報行列の対角要素の平方根）を表 4.14～4.17 に示す。次に項目分析のときの手順と同様に項目適合度を計算し，INFIT,と OUTFIT を表 4.18, 4.19 に示し，Q_3統計量も計算した。ここでもQ_3統計量で絶対値が 0.2 を上回る項目は存在せず，項目の組み合わせも非常に膨大であるため表は省略する。標準誤差が大きい項目が国語の b110 や数学の a101, f104 で確認されたが，適合度はそれほど逸脱していないため削除はしなかった。収束基準は先ほど同様の設定とし，国語は 114 回，数学は 86 回で EM サイクルが終了し，推定中に途中で計算を中止した項目は存在しなかった。

国語と数学の項目特性曲線と項目情報関数をプロットする (図 4.23~図 4.26)。ただし，項目情報関数は国語と数学で縦軸の目盛りを揃えている。併せて識別力と困難度のプロットも図 4.27 と 4.28 に示した。項目情報関数と散布図の軸の目盛りは国語と数学で揃えている。国語の特性曲線は困難度が-4 から 1 の間に集中しており，項目情報関数も 0 以下の低い位置にピークが来ている項目が多い。また識別力が 1 以上の項目が存在せず，項目情報関数も全体的になだらかな項目が多い。一方，数学のテスト項目の特性曲線は困難度が－3 から 3 の間で，識別力が中程度の項目が集中しているため，項目情報関数のピークも尺度全体におおよそ均等に広がっている。

次にテスト情報量を確認する。まずは尺度全体のテスト情報量を図 4.29 と 4.30 に示す。当然だが数学の方が，わずかにピークが 0 に寄っている。しかしどちらの垂直尺度も広い学力水準の児童生徒を対象に作られているため，あまり大きな差はない。

そこでテストのレベルごと（項目の頭文字が a から f と st）に尺度の特徴を捉えるため，レベルごとに項目パラメタの散布図とテスト情報関数を描き，図 4.31～4.34 に示した。ただし, テスト情報関数のθの定義域は, 関数の全体が俯瞰できるように-6 から 6 と広くとった。国語のレベルごとのテスト情報関数は非常に安定した釣り鐘型の分布をしており, st の項目群を除いてレベルが上がるにつれてわずかに右にピークがシフトしているが，やはりすべてのピークが 0 よりも小さい範囲に入っている。数学のレベルごとのテスト特性曲線は上級レベルのテストが多峰型の分布となっているもの，尺度の-3 から 3 という比較的広い範囲をカバーしており，曲線のピークも，低いレベルにおいては，レベルの上昇にしたがって右に推移していっている。しかし，d の項目群だけピークが大幅に右にずれてしまっていたり，e と f の項目群の情報関数が，広い範囲にわたって低い情報量を保持するような曲線になっていたりする。なお，どちらの教科も st の項目はおおよそ中間レベルの位置でピークを迎えている。

表 4.14 項目パラメタの推定結果（国語）

Item	a	b	Item	a	b
a101	1.044	-3.506	c106	0.606	-1.803
a102	0.562	-0.859	c107	1.018	-1.119
a103	0.859	-1.338	c108	0.577	-1.628
a104	1.020	-2.660	c109	0.855	-1.414
a105	0.663	-1.830	c110	0.568	-0.595
a106	0.293	-2.161	d101	0.437	-0.220
a107	0.468	0.020	d102	0.545	0.352
b101	0.807	-2.008	d103	0.752	-3.283
b102	0.675	-1.299	d104	0.662	-1.834
b103	0.852	-2.698	d105	0.558	-0.475
b104	0.507	-3.720	d106	0.545	-0.781
b105	0.560	-1.458	d107	0.314	-0.591
b106	1.035	-0.663	d108	0.365	-0.177
b107	0.422	-0.382	d109	0.821	-0.796
b108	0.601	-1.872	d110	0.608	-1.222
b109	0.719	0.663	e101	1.149	-2.238
b110	0.070	-2.760	e102	0.926	-1.648
st101	0.000	0.000	e103	0.804	-0.707
st102	0.842	-1.238	e104	1.121	-1.722
st103	0.277	0.519	e105	0.606	-2.784
st104	0.274	-4.579	e106	0.624	-1.288
st105	0.333	-0.342	e107	0.409	-0.871
st106	0.854	-1.534	e108	0.423	0.295
st107	0.200	1.806	e109	0.697	0.045
st108	0.000	0.000	e110	0.629	-1.839
st109	0.699	-1.485	f101	0.359	-0.114
st110	0.430	-1.700	f102	0.613	0.298
a108	0.849	-2.001	f103	0.559	-0.919
a109	0.603	-1.644	f104	0.776	-1.223
a110	0.772	-1.084	f105	0.806	-1.677
c101	0.383	-0.513	f106	0.540	-0.993
c102	0.402	-0.181	f107	0.639	-1.861
c103	0.496	-0.710	f108	0.816	-0.096
c104	0.514	-1.135	f109	0.816	0.493
c105	0.790	-1.692	f110	0.650	0.685

表 4.15 項目パラメタの推定の標準誤差（国語）

Item	a	b	Item	a	b
a101	0.199	0.297	c106	0.066	0.153
a102	0.084	0.115	c107	0.083	0.066
a103	0.103	0.083	c108	0.066	0.153
a104	0.146	0.165	c109	0.082	0.099
a105	0.091	0.128	c110	0.067	0.106
a106	0.072	0.324	d101	0.048	0.110
a107	0.089	0.226	d102	0.056	0.094
b101	0.078	0.104	d103	0.149	0.493
b102	0.066	0.085	d104	0.076	0.183
b103	0.099	0.166	d105	0.057	0.106
b104	0.081	0.402	d106	0.053	0.111
b105	0.059	0.103	d107	0.043	0.169
b106	0.082	0.051	d108	0.045	0.128
b107	0.054	0.118	d109	0.069	0.080
b108	0.064	0.122	d110	0.061	0.129
b109	0.088	0.142	e101	0.175	0.216
b110	0.046	1.451	e102	0.102	0.149
st101	NA	NA	e103	0.069	0.090
st102	0.045	0.053	e104	0.131	0.140
st103	0.025	0.112	e105	0.096	0.400
st104	0.038	0.589	e106	0.066	0.161
st105	0.027	0.087	e107	0.048	0.174
st106	0.049	0.063	e108	0.046	0.106
st107	0.025	0.263	e109	0.058	0.074
st108	NA	NA	e110	0.084	0.256
st109	0.043	0.075	f101	0.073	0.243
st110	0.033	0.125	f102	0.080	0.115
a108	0.143	0.164	f103	0.090	0.235
a109	0.109	0.168	f104	0.116	0.207
a110	0.140	0.133	f105	0.140	0.281
c101	0.048	0.119	f106	0.091	0.258
c102	0.049	0.111	f107	0.119	0.362
c103	0.053	0.100	f108	0.095	0.101
c104	0.055	0.117	f109	0.094	0.088
c105	0.076	0.113	f110	0.080	0.102

表 4.16 項目パラメタの推定結果（数学）

Item	a	b	Item	a	b
st101	0.666	-2.144	c106	0.961	0.494
a101	0.450	-5.767	c107	0.220	2.273
a102	1.002	-0.878	c108	0.572	-0.479
a103	0.928	-1.967	c109	1.429	0.731
a104	1.272	-1.720	c110	0.981	0.681
a105	0.640	-2.708	d101	0.478	-3.068
a106	0.970	-1.227	d102	0.911	0.635
a107	1.408	-0.815	d103	0.935	0.442
a108	1.293	-1.895	d104	1.017	-0.768
a109	0.820	-1.967	d105	2.246	2.234
a110	0.728	-0.644	d106	0.165	1.480
st102	0.785	-1.311	d107	0.357	-0.065
st103	0.498	1.105	d108	0.482	-0.454
b101	0.597	-2.551	d109	0.517	-0.152
b102	0.439	-1.931	d110	0.523	3.553
b103	0.987	-0.648	e101	0.986	-0.551
b104	0.524	-1.709	e102	0.702	0.860
b105	0.887	-0.927	e103	0.735	-0.555
b106	0.543	-0.406	e104	0.166	2.035
b107	0.455	-0.476	e105	0.184	1.868
b108	0.904	-0.641	e106	0.955	-0.289
b109	0.966	-1.699	e107	0.585	0.371
b110	1.017	-0.555	e108	0.000	0.000
st104	1.265	-0.321	e109	0.000	0.000
st105	0.717	1.629	e110	1.130	1.108
st106	1.096	-0.186	f101	1.015	-0.315
st107	0.188	3.094	f102	0.810	0.068
st108	0.540	-0.361	f103	1.004	1.049
st109	0.643	2.560	f104	0.212	-5.380
st110	0.842	-1.307	f105	0.441	1.015
c101	0.578	-2.031	f106	0.359	0.607
c102	0.891	-1.186	f107	0.626	0.381
c103	1.002	-0.403	f108	0.678	0.351
c104	1.025	-0.915	f109	0.935	3.338
c105	2.087	1.455	f110	0.995	3.801

表 4.17 項目パラメタの推定の標準誤差（数学）

Item	a	b	Item	a	b
st101	0.046	0.102	c106	0.081	0.069
a101	0.218	2.047	c107	0.053	0.631
a102	0.125	0.073	c108	0.061	0.090
a103	0.120	0.101	c109	0.133	0.063
a104	0.142	0.066	c110	0.098	0.082
a105	0.115	0.244	d101	0.079	0.459
a106	0.120	0.069	d102	0.073	0.061
a107	0.161	0.058	d103	0.083	0.066
a108	0.149	0.075	d104	0.085	0.070
a109	0.113	0.112	d105	0.471	0.117
a110	0.112	0.112	d106	0.044	0.428
st102	0.043	0.054	d107	0.047	0.128
st103	0.032	0.090	d108	0.052	0.110
b101	0.077	0.189	d109	0.055	0.097
b102	0.061	0.161	d110	0.108	0.572
b103	0.081	0.053	e101	0.083	0.070
b104	0.062	0.115	e102	0.062	0.075
b105	0.077	0.056	e103	0.070	0.096
b106	0.061	0.096	e104	0.043	0.492
b107	0.059	0.111	e105	0.044	0.410
b108	0.077	0.056	e106	0.080	0.066
b109	0.089	0.072	e107	0.055	0.080
b110	0.093	0.058	e108	NA	NA
st104	0.061	0.031	e109	NA	NA
st105	0.046	0.089	e110	0.100	0.062
st106	0.058	0.038	f101	0.125	0.095
st107	0.029	0.501	f102	0.109	0.106
st108	0.035	0.063	f103	0.144	0.101
st109	0.062	0.185	f104	0.098	2.625
st110	0.055	0.067	f105	0.071	0.162
c101	0.070	0.186	f106	0.068	0.177
c102	0.077	0.075	f107	0.082	0.106
c103	0.081	0.054	f108	0.086	0.100
c104	0.085	0.061	f109	0.241	0.413
c105	0.317	0.091	f110	0.328	0.592

表 4.18 項目適合度（国語，項目削除後）

Item	N	InFit	StdInFit	OutFit	StdOutFit	Item	N	InFit	StdInFit	OutFit	StdOutFit
a101	409	1.287	1.081	1.646	8.170	c107	765	0.939	-1.130	0.776	-4.667
a102	400	0.951	-1.502	0.925	-1.074	c108	687	1.000	0.015	0.914	-1.638
a103	393	0.972	-0.523	0.902	-1.414	c109	675	0.982	-0.266	0.844	-2.983
a104	397	1.172	1.171	1.624	7.802	c110	564	0.959	-1.171	0.926	-1.267
a105	408	1.037	0.658	1.027	0.385	d101	781	0.936	-2.787	0.918	-1.662
a106	407	1.015	0.419	1.024	0.336	d102	645	0.911	-3.283	0.877	-2.283
a107	365	0.926	-1.809	0.894	-1.467	d103	786	1.045	0.277	0.938	-1.243
b101	749	1.058	0.933	0.988	-0.223	d104	759	0.962	-0.491	0.859	-2.845
b102	686	0.992	-0.188	0.956	-0.831	d105	696	0.918	-2.464	0.877	-2.372
b103	735	1.140	1.301	0.954	-0.889	d106	792	0.940	-1.742	0.874	-2.585
b104	785	1.104	1.001	1.086	1.674	d107	792	0.964	-1.750	0.956	-0.879
b105	742	0.999	-0.004	0.969	-0.599	d108	782	0.956	-2.260	0.946	-1.078
b106	729	0.876	-3.217	0.773	-4.617	d109	789	0.911	-1.930	0.757	-5.174
b107	703	0.958	-1.929	0.953	-0.885	d110	789	0.945	-1.154	0.842	-3.271
b108	744	1.029	0.625	1.026	0.493	e101	810	1.007	0.098	0.494	-12.174
b109	542	0.870	-2.436	0.716	-5.085	e102	810	0.925	-0.749	0.672	-7.287
b110	712	1.002	0.157	1.003	0.057	e103	811	0.872	-2.677	0.740	-5.646
st102	1917	0.946	-1.605	0.795	-6.726	e104	811	1.002	0.054	0.394	-15.462
st103	1905	0.972	-2.499	0.974	-0.809	e105	810	0.965	-0.230	0.821	-3.781
st104	1866	1.003	0.073	0.978	-0.668	e106	789	0.922	-1.326	0.798	-4.238
st105	1891	0.965	-2.843	0.960	-1.236	e107	809	0.937	-1.931	0.906	-1.936
st106	1891	0.983	-0.395	0.792	-6.775	e108	803	0.935	-3.243	0.926	-1.509
st107	1864	0.982	-1.122	0.978	-0.675	e109	803	0.864	-4.394	0.818	-3.825
st109	1805	0.965	-0.960	0.879	-3.755	e110	716	0.948	-0.552	0.744	-5.212
st110	1799	0.962	-1.321	0.958	-1.263	f101	331	0.947	-1.603	0.936	-0.835
a108	220	1.046	0.451	1.200	1.997	f102	382	0.896	-2.737	0.852	-2.132
a109	230	1.021	0.336	0.985	-0.161	f103	369	0.922	-1.115	0.813	-2.667
a110	174	0.979	-0.268	0.910	-0.859	f104	398	0.890	-1.090	0.788	-3.167
c101	776	0.971	-1.416	0.966	-0.673	f105	397	0.905	-0.641	0.773	-3.408
c102	771	0.963	-1.835	0.958	-0.828	f106	373	0.916	-1.189	0.837	-2.326
c103	774	0.962	-1.421	0.942	-1.149	f107	398	0.917	-0.626	0.821	-2.656
c104	774	0.973	-0.795	0.937	-1.268	f108	397	0.844	-2.981	0.780	-3.290
c105	771	0.997	-0.028	0.870	-2.644	f109	380	0.843	-3.555	0.785	-3.154
c106	766	0.990	-0.168	0.970	-0.601	f110	398	0.887	-3.009	0.900	-1.448

表 4.19 項目適合度（数学，項目削除後）

Item	N	InFit	StdInFit	OutFit	StdOutFit	Item	N	InFit	StdInFit	OutFit	StdOutFit
st101	1967	0.978	-0.503	0.963	-1.158	c105	529	0.800	-1.161	0.202	-19.464
a101	407	1.084	0.396	1.170	2.327	c106	725	0.904	-2.074	0.913	-1.701
a102	395	0.931	-1.549	0.879	-1.753	c107	672	0.989	-0.290	0.991	-0.165
a103	408	1.019	0.308	1.025	0.349	c108	644	0.966	-1.128	0.956	-0.794
a104	405	0.998	-0.010	0.934	-0.949	c109	599	0.853	-2.108	0.666	-6.399
a105	410	1.044	0.552	1.120	1.663	c110	518	0.898	-1.791	0.999	-0.018
a106	389	0.957	-0.940	0.949	-0.727	d101	776	0.987	-0.096	0.946	-1.082
a107	381	0.864	-2.480	0.748	-3.738	d102	753	0.907	-2.356	0.849	-3.041
a108	399	1.022	0.296	0.899	-1.463	d103	565	0.909	-1.973	0.829	-3.005
a109	409	1.025	0.425	1.010	0.141	d104	747	0.909	-1.758	0.733	-5.582
a110	383	0.942	-1.435	0.926	-1.042	d105	414	0.987	0.041	0.120	-21.526
st102	1982	0.942	-2.007	1.013	0.394	d106	762	0.995	-0.317	0.996	-0.087
st103	1977	0.944	-2.258	0.935	-2.072	d107	754	0.977	-1.206	0.978	-0.432
b101	761	1.053	0.880	1.000	0.010	d108	745	0.966	-1.217	0.944	-1.094
b102	768	1.003	0.106	0.992	-0.152	d109	708	0.956	-1.588	0.945	-1.048
b103	772	0.927	-2.083	0.882	-2.385	d110	588	1.040	0.327	0.938	-1.078
b104	785	0.997	-0.093	1.000	0.003	e101	783	0.867	-2.843	0.948	-1.052
b105	742	0.944	-1.696	0.878	-2.419	e102	747	0.933	-1.977	0.907	-1.834
b106	779	0.960	-1.568	0.959	-0.808	e103	733	0.911	-2.084	0.887	-2.221
b107	742	0.974	-1.152	0.976	-0.471	e104	787	0.999	-0.060	0.999	-0.023
b108	776	0.918	-2.513	0.907	-1.886	e105	788	0.996	-0.183	1.004	0.069
b109	718	0.997	-0.039	0.928	-1.380	e106	735	0.885	-2.677	0.815	-3.742
b110	594	0.927	-1.775	0.830	-3.063	e107	773	0.939	-2.255	0.943	-1.141
st104	1693	0.851	-4.649	0.775	-6.972	e110	559	0.873	-2.337	0.748	-4.527
st105	1800	0.951	-1.142	0.860	-4.355	f101	366	0.872	-1.981	0.704	-4.364
st106	1409	0.883	-3.511	0.795	-5.771	f102	326	0.903	-1.880	0.822	-2.384
st107	1731	0.989	-0.426	1.001	0.035	f103	212	0.882	-1.527	0.803	-2.136
st108	1549	0.951	-2.411	0.920	-2.259	f104	393	0.979	-0.136	0.972	-0.395
st109	1242	1.018	0.260	0.913	-2.219	f105	388	0.969	-0.861	0.989	-0.159
st110	1332	0.944	-1.353	0.914	-2.273	f106	377	0.977	-0.875	0.977	-0.317
c101	741	0.986	-0.221	0.999	-0.010	f107	390	0.937	-1.627	0.919	-1.148
c102	756	0.941	-1.261	0.896	-2.068	f108	381	0.931	-1.658	0.900	-1.413
c103	707	0.903	-2.443	0.801	-3.957	f109	216	1.260	0.867	0.485	-6.419
c104	718	0.929	-1.500	0.847	-3.027	f110	231	1.380	0.841	0.469	-6.912

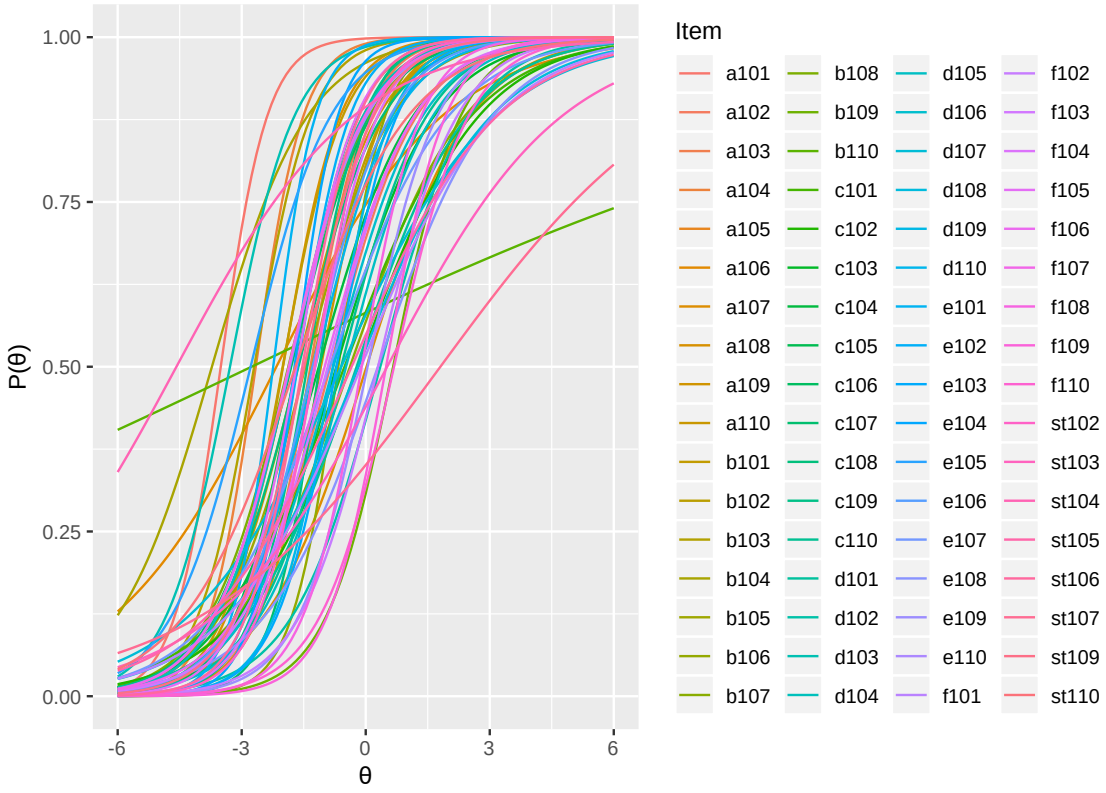

図 4.23 項目特性曲線（国語）

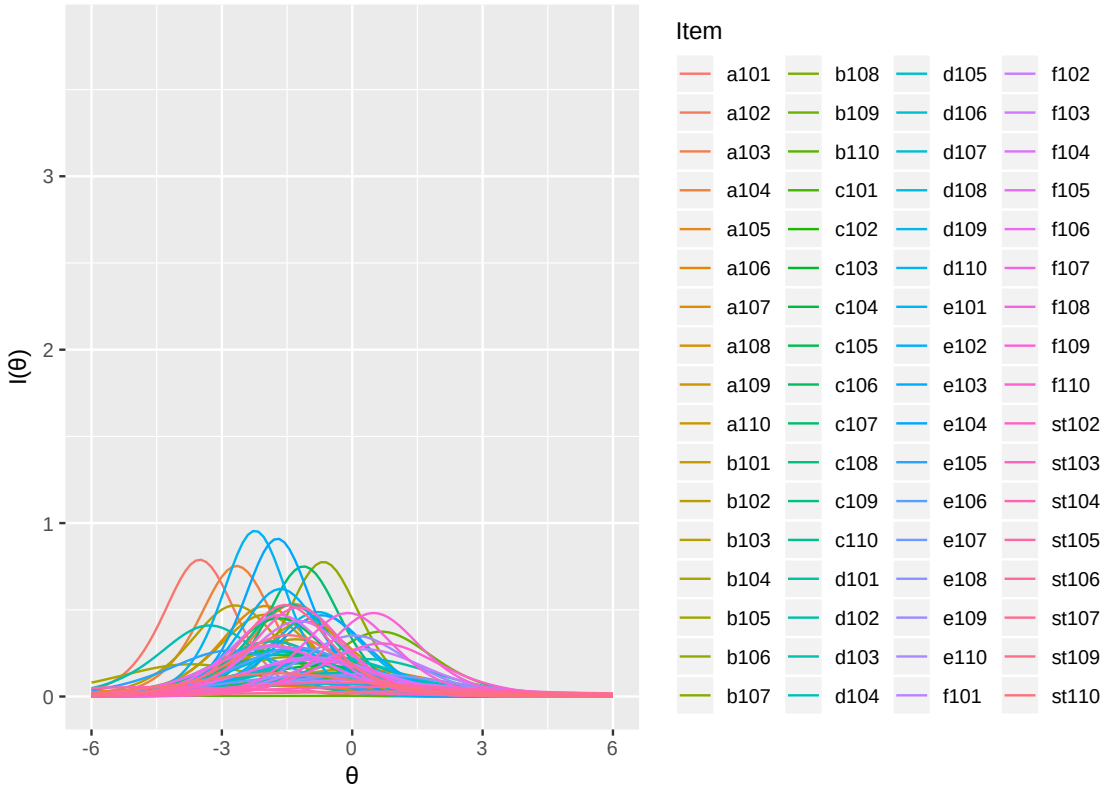

図 4.24 項目情報関数（国語）

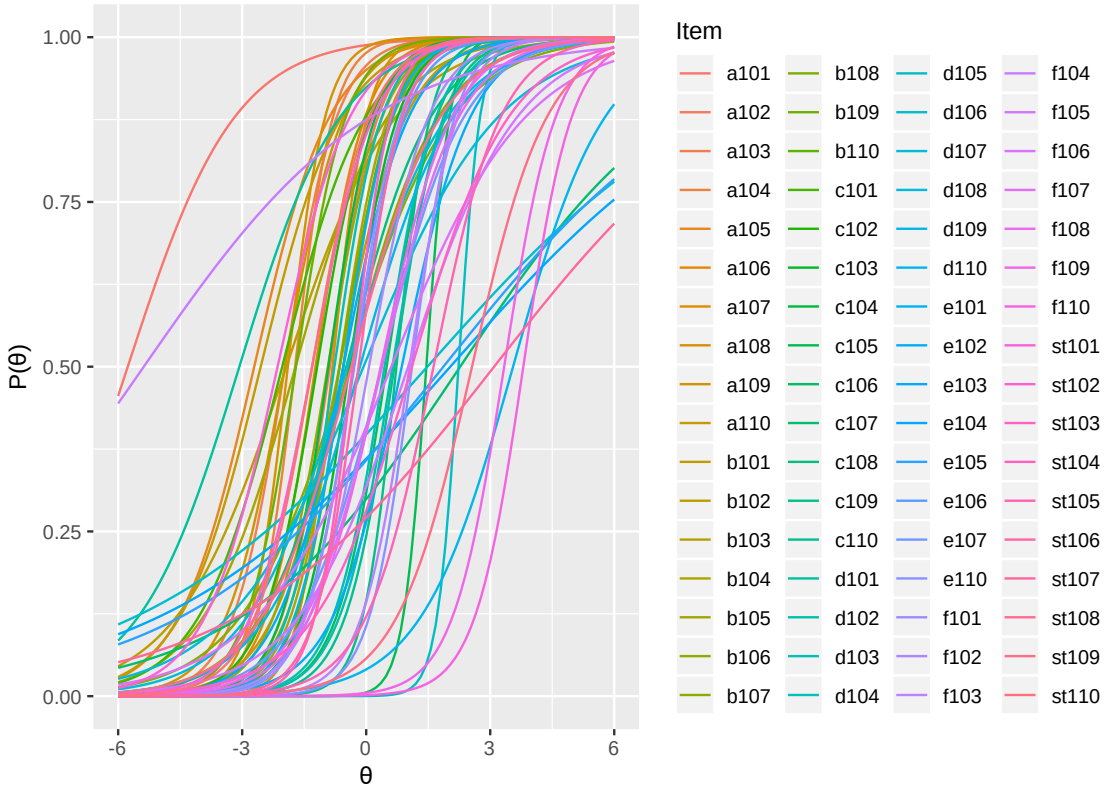

図 4.25 項目特性曲線（数学）

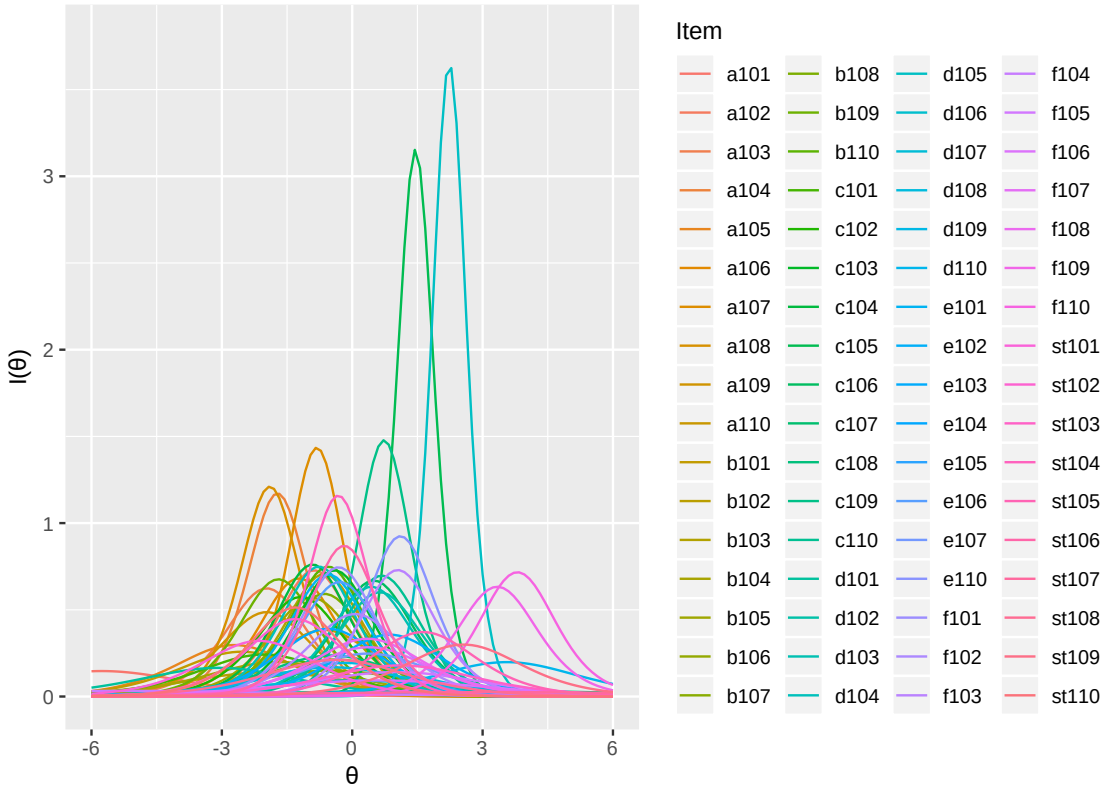

図 4.26 項目情報関数（数学）

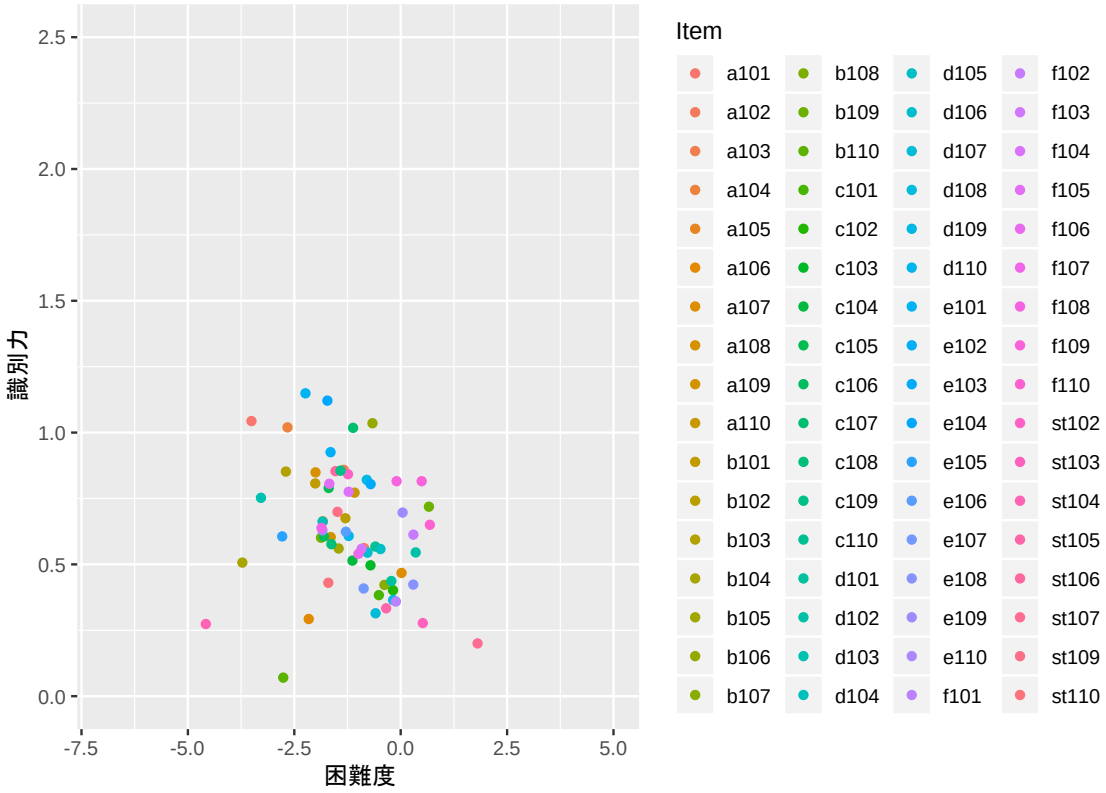

図 4.27 項目パラメタの散布図（国語）

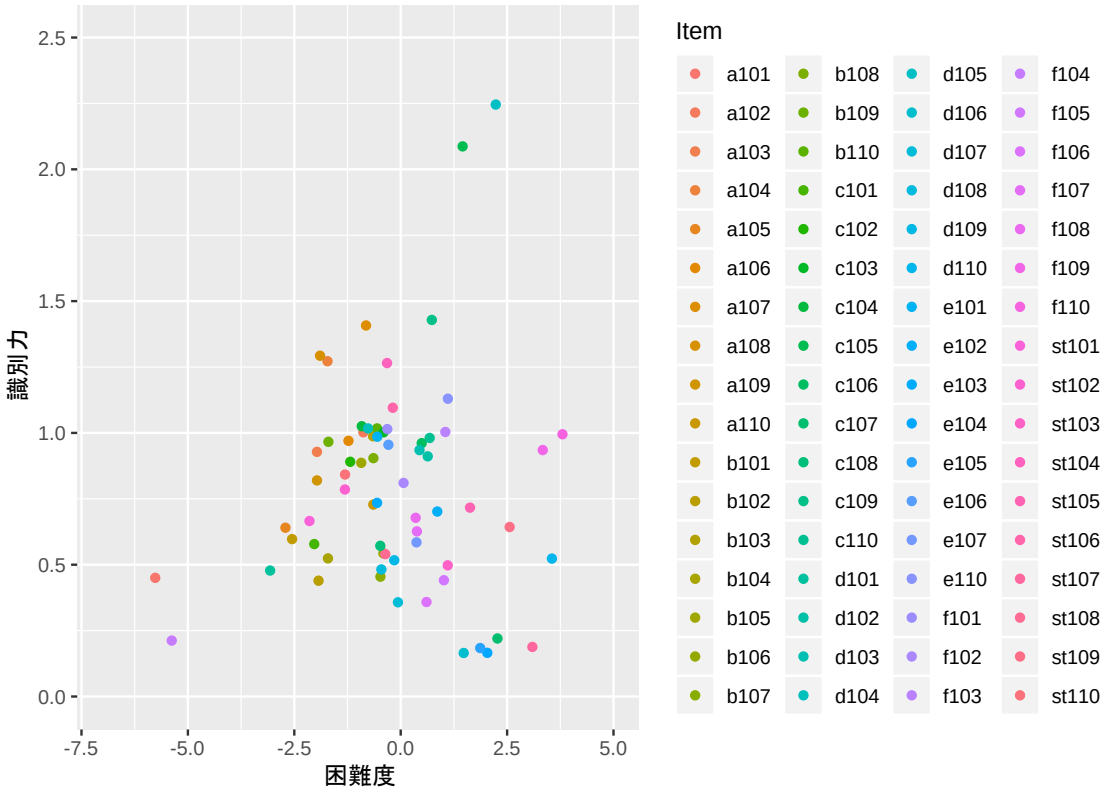

図 4.28 項目パラメタの散布図（数学）

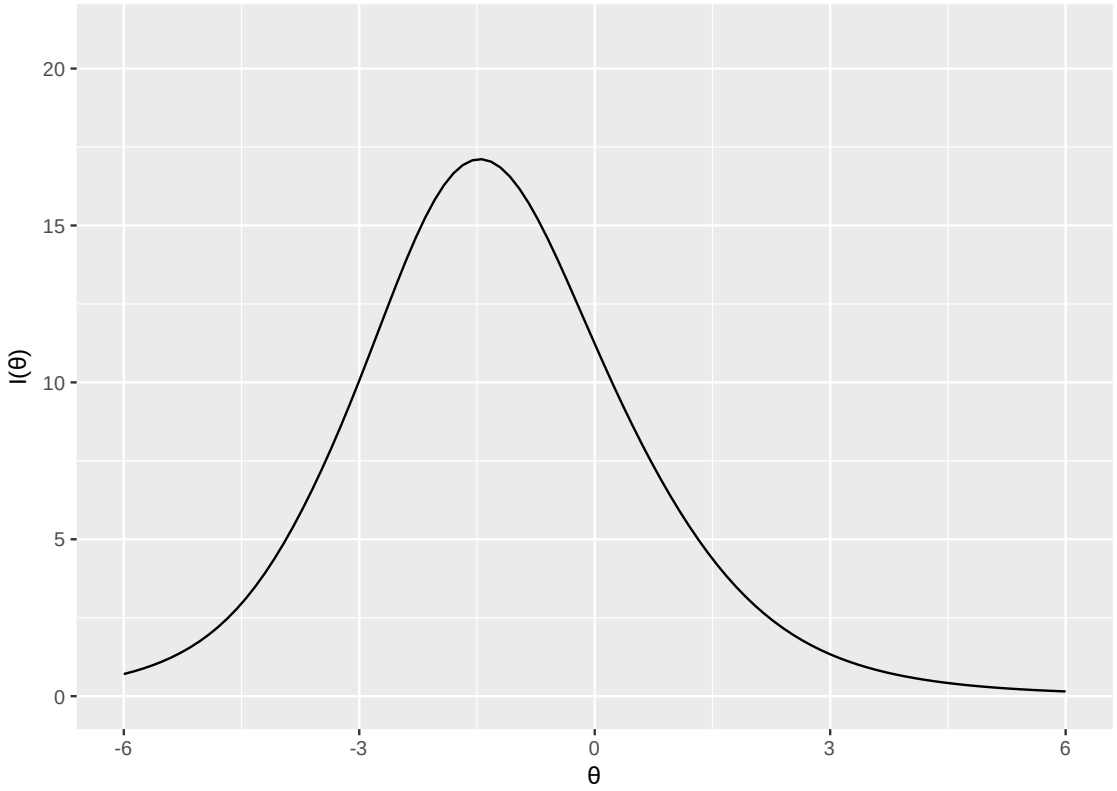

図 4.29 テスト情報関数（国語）

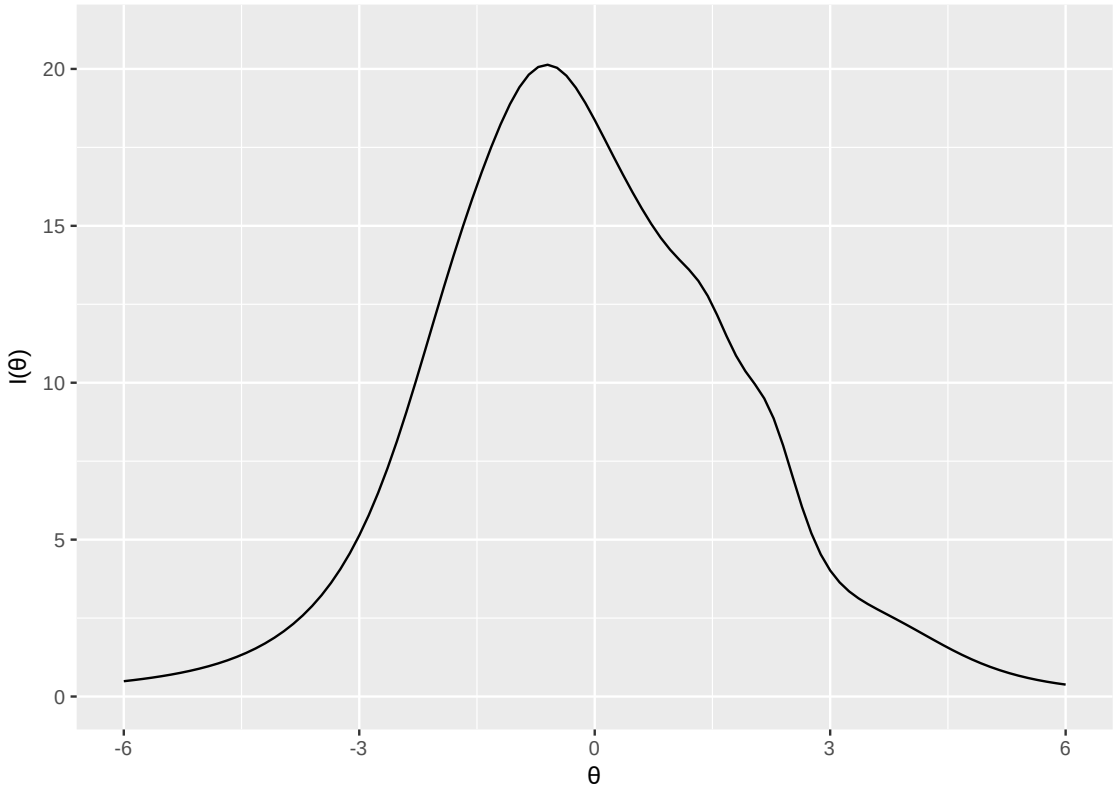

図 4.30 テスト情報関数（数学）

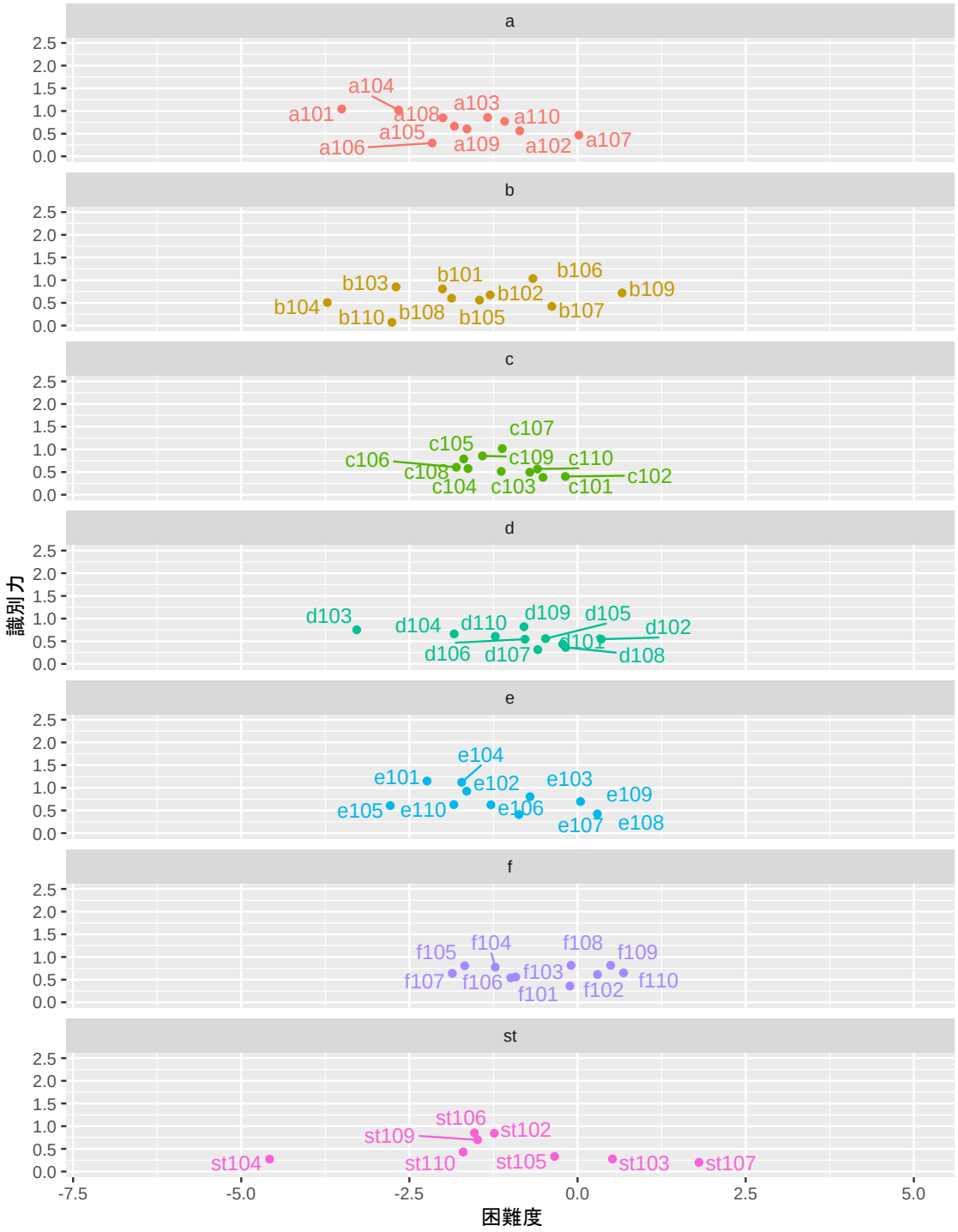

図 4.31 レベルごとの項目パラメタの散布図（国語）

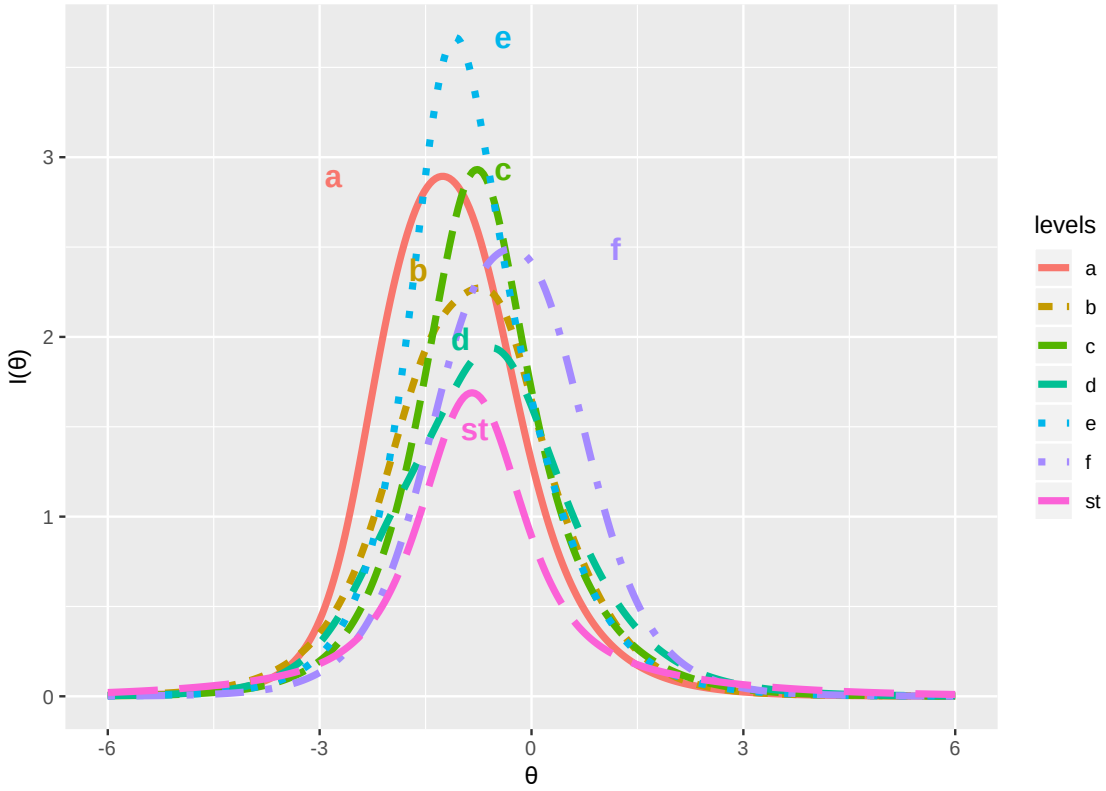

図 4.32 レベルごとのテスト情報関数（国語）

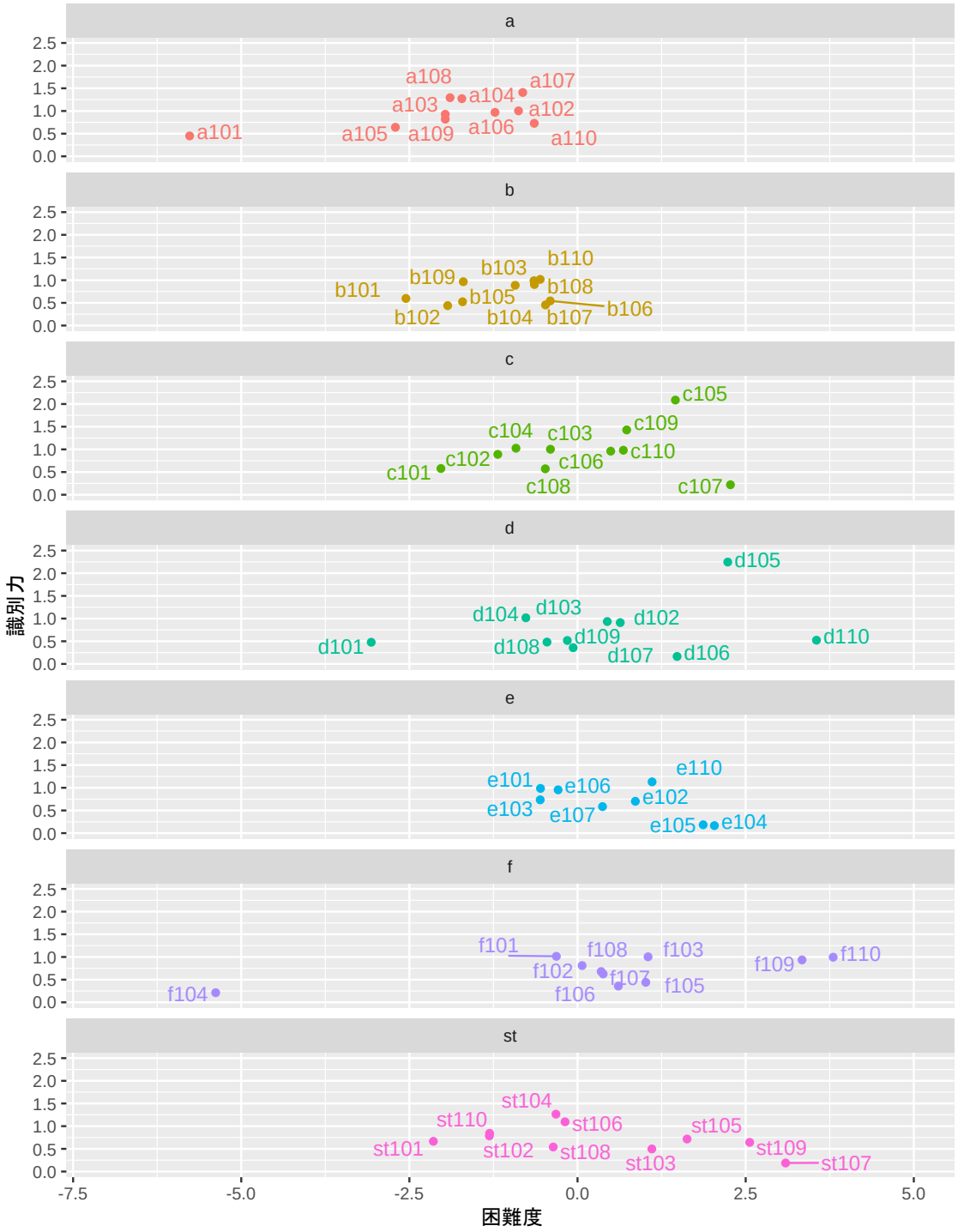

図 4.33 レベルごとの項目パラメタの散布図（数学）

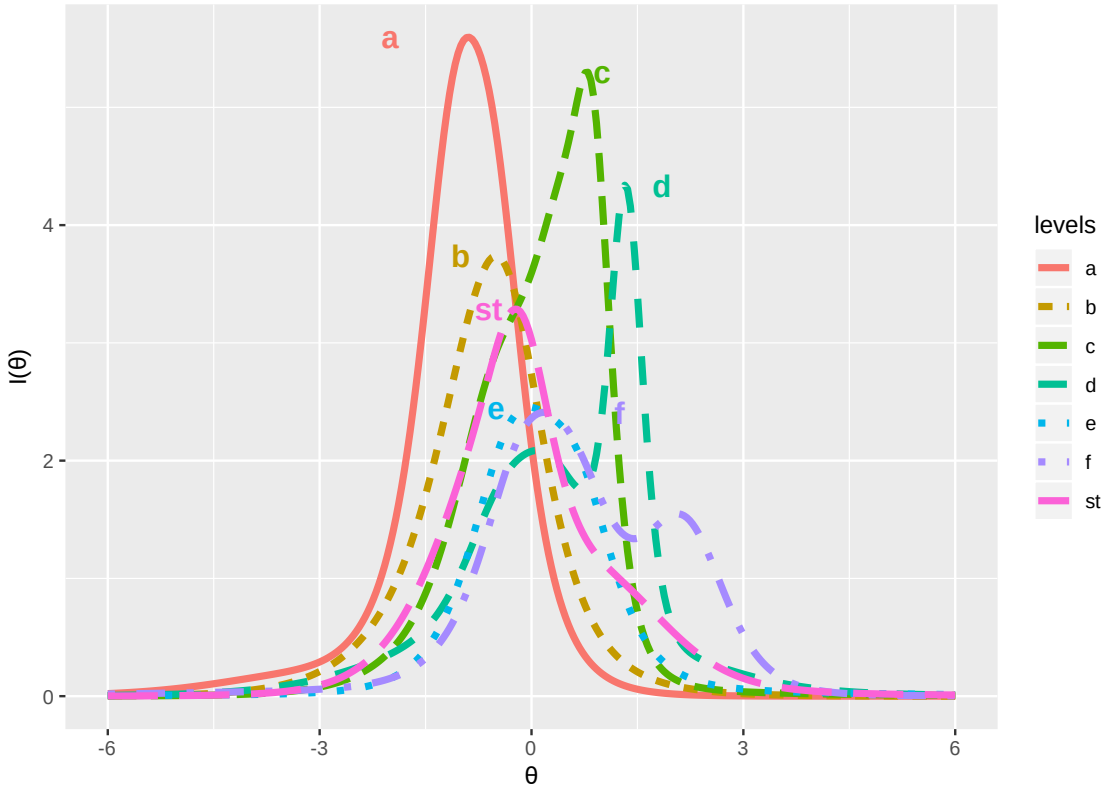

図 4.34 レベルごとのテスト情報関数（数学）

4.2.5 学力分布の推定

同時尺度調整法では，学年ごとに異なる平均と標準偏差の分布を仮定する多母集団モデルに基づいて項目パラメタの推定をおこなった。ここではそれを応用して EM アルゴリズムによって母集団ごとに学力分布を推定した結果を示す。母集団分布の推定では EM サイクルが収束した時点の項目パラメタを固定し，母集団分布の平均と標準偏差をパラメタとして推定する。先述した項目パラメタ推定の EM サイクル時にも平均と標準偏差を更新しながらサイクルを進めたが，そのときはどこかひとつの集団の平均と標準偏差を 0 と 1 に固定した。しかし，今回のように母集団分布のみに関心がある場合には項目パラメタの値が固定されているため，特定の集団のパラメタを固定せずとも適当な値に収束する。また，母集団分布推定の際には正規分布などの特定の分布族を用いず, E ステップの区分求積時のノードにおける受検者の期待度数によって分布を近似している。分点数は 31 であり，積分区間は-6 から 6 で固定している。2 教科の母集団分布を図 4.35 と 4.36 に示し，平均と標準偏差および，それらによって計算された効果量 (式 3.11 を参照) の一覧を表 4.20 に示す。

まず母集団分布のパラメタと効果量の表を見るとどちらの教科でも学年分布が上がるにつれて標準偏差は拡大していっている。どちらの教科も標準偏差 2 程度がピークであった。この現象は母集団分布の山が徐々に平たくなっていることからも分かる。学年間の分布の独立，すなわち効果量を見ると，学年が上がるにつれて減少していっている。グラフの形状から分かることは，国語は全学年で，やや学力が低い層の小さな山と平均的な層の大きな山の二山の形状をしており，おなじ形状のまま右に推移していっているが，数学は逆で，平均的な層の大きな山と学力の高いそうの小さな山の二山の形状であり，中学生 (G4, G5) 以降で大きく二山の形状にシフトしている。

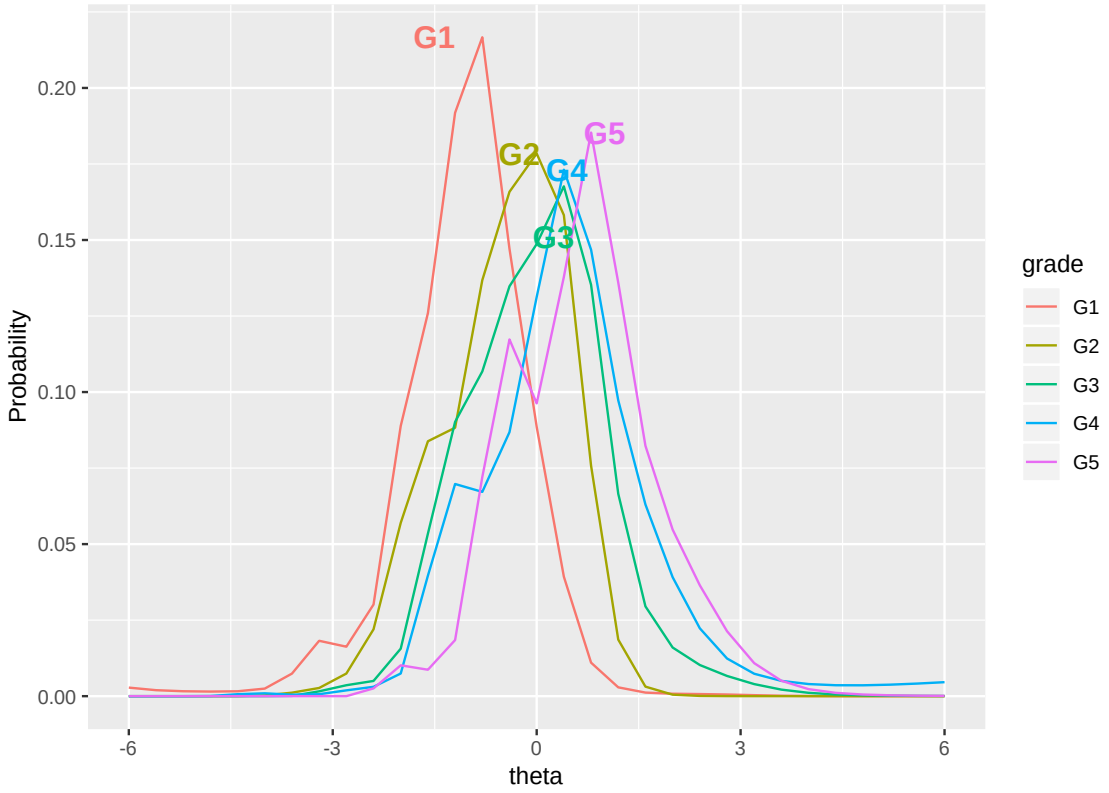

図 4.35 推定母集団分布（国語）

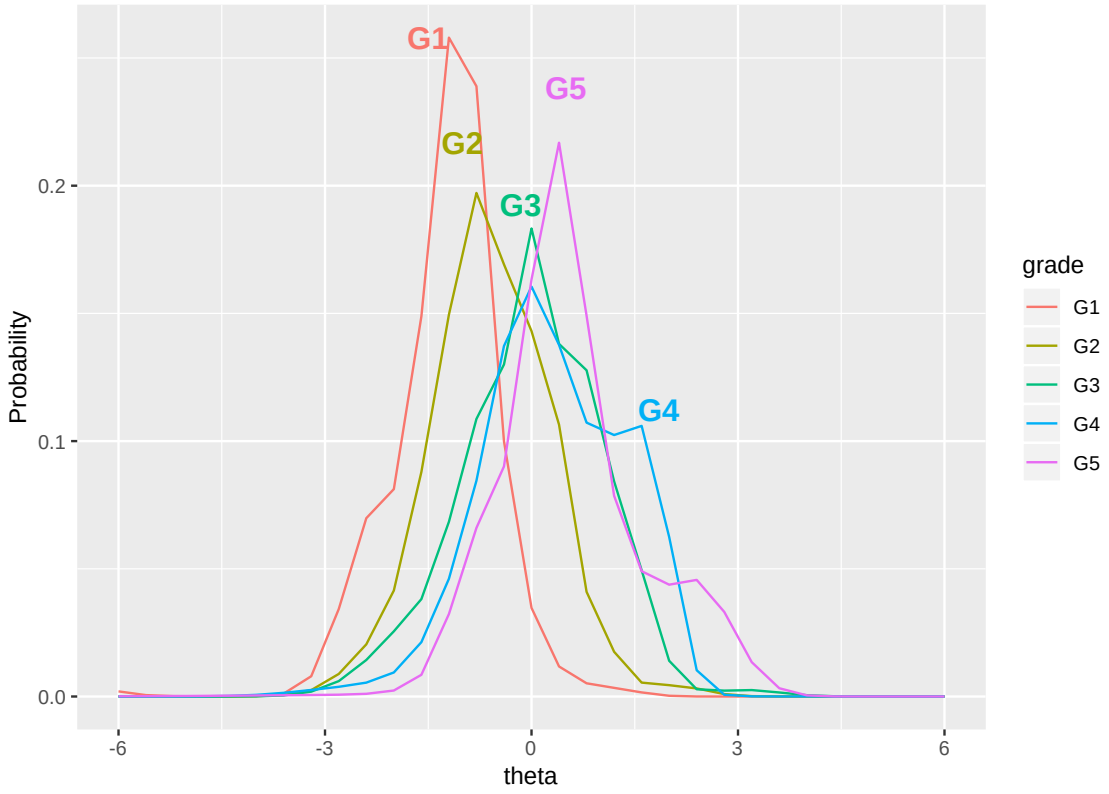

図 4.36 推定母集団分布（数学）

表 4.20 母集団分布のパラメタと効果量

	国語				算数・数学			
学年	N	平均	SD	効果量	N	平均	SD	効果量
G1	411	-1.069	0.953	—	410	-1.235	0.775	—
G2	386	-0.465	0.900	0.651	385	-0.588	0.874	0.785
G3	391	0.016	1.032	0.497	391	0.003	1.02	0.622
G4	415	0.458	1.308	0.374	413	0.313	1.04	0.301
G5	390	0.661	1.067	0.170	397	0.565	1.055	0.241

4.2.6 考察

今回の垂直尺度化は事前のシミュレーション結果から最も性能が高いと予想される CC 法によって尺度調整を実行した。しかし，受検者の標本数が一学年あたり 400 人程度であったことと，項目自体の質を十分に検討できなかったことも相俟って，安定した推定値を得ることができない項目がいくつか見受けられた。特に困難度の絶対値がパラメタ推定の際に基準とした分布$N(0,1)$から推測して非常に大きな項目 (たとえば 3 以上) を削除することでテスト情報関数や母集団分布はより安定した形状になると考えられる。しかし，実際に非常に困難度の高い項目も尺度に必要ではあるので，今回は特に適合度等の指標で検出されない限り項目を残したまま推定している。

また，別の安定した推定手法としては比較的制限を加えた事前分布をもちいてベイズ推定を実行する方法が考えられる。その場合は Stan (Stan Development Team, 2018)や greta (Golding, 2018) 等の確率的プログラミング言語でコードを書いて，階層ベイズ推定法を実行することになる。しかし階層ベイズ推定は同時事後分布を最大化するため，理論的には一致推定量が得られないことや，確率的プログラミング言語による推定アルゴリズムは確率的に発生させた乱数によりパラメタを推定するものであるため，推定値のランダムネスには注意が必要である。

今回垂直尺度化された項目のパラメタを見てみると，やはり先行研究同様に尺度の縮小が起こっている。特に縮小の様子は数学において顕著であったが，国語ではそもそも学年のテスト間の難易度がほとんど動かないという様子も確認された。先行研究で指摘されていたいくつかの原因から，今回の縮小の原因を推測するのであれば，推定誤差が最も大きな原因を占めていると考えられる。次に尺度化テストという特殊な形式の項目も含んでいるためテストの内容も影響している可能性がある。また，固有値の減衰状況を見ると尺度化テストなどは 2 因子の仮説を支持する可能性も考えられるため，わずかではあるが測定の多次元性が考えられる。

母集団分布から推定された学年分布は学年が上がるにつれてばらつきが大きくなり，学力の伸びも鈍くなっていくという結果になった。ばらつきが徐々に大きくなっていく現象については，学年が上がるにつれて個人差が大きくなっていくことが一因ではないかと推測できる。また，特に国語では少数の学力下位層とそれ以外の平均的な層に分断されていき，数学では逆に平均的な層と少数の学力上位層に分断されている様子が確認された。一般的に分布が二峰 (bimodal) 型になると平均や標準偏差は単峰型の分布と比較することは困難になる。また，本来単峰型でない集団に対して正規分布のような単峰型の事前分布を仮定して推定することも，当然，推定誤差を大きくすることにつながる。これまでの尺度の縮小に関する研究では母集団分布の計算をおこなったものは存在しないが，尺度が縮小する，つまり項目パラメタの困難度の標準偏差が減少し，見かけ上の能力の伸びが徐々に小さくなっていく現象の原因は，学力分布が二峰型であることもひとつの要因かもしれない。

4.3 周辺ベイズ推定法による垂直尺度構成

4.3.1 方法

周辺ベイズ推定法 (MBE) は局外母数である受検者の能力について積分消去した周辺事後分布を最大化するように項目パラメタを推定する手法である。ベイズ推定の一種であるため，仮に全受検者が正答あるいは誤答している項目でもパラメタを推定することができるほか，事前分布によってパラメタに制約をかけるため，困難度における絶対値の大きな推定値や，異常に識別力が高く推定されてしまう問題などにも対処できる。この実験で用いるMBE は M ステップにおける目的関数に事前分布を加えて，それを最大化するため，いわゆる MAP 推定値を求めていることに他ならない。

4.3.2 結果

通常の MMLE による垂直尺度構築と同様の収束基準を用いて MBE を実行した。困難度の事前分布は正規分布$N(0,1)$を，識別力の事前分布は$\ln N(0,1)$を採用した。今回は項目を削除せずに，推定できるパラメタをすべて使用することとした。EM サイクルは国語が 100 回で収束と判断された。数学は 251 回で収束したものの a101 の項目が 1 回目の推定時点で非負の識別力が推定されてしまったため，その項目を削除して再推定したところ，86 回で収束と判断された。

推定された項目パラメタとその推定の標準誤差を表 4.21〜4.24 に示す。項目パラメタの推定値はどちらの項目も安定している。標準誤差はわずかではあるが全体的に減少した。

項目の適合度 (INFIT と OUTFIT) を表 4.25 と 4.26 に示し，適合度プロットを図 3.37 と 3.38 に示す。INFIT 統計量はどちらの教科もすべての項目で 0.7 から 1.3 の間に収まっている。適合度のプロットは MMLE のときとほぼ全く同じであり，しっかりとデータにフィットした推定値が得られている。

次に項目特性曲線とレベルごとの項目パラメタの散布図，テスト情報関数を図 3.39〜3.44 に図示する。全体的に推定値が事前分布の平均の方向に縮小したため，散布図やテスト情報関数の軸の目盛りの範囲が縮まっている。このことからも，安定した推定値が得られていることがうかがえる。しかしテスト情報関数を見ると，全体的に縮小したものの，関数の形自体は MMLE のときとほとんど変わらない。また尺度の縮小についても同様に生じている。

最後に推定母集団分布について図 4.45 と 4.46 に示し，パラメタと効果量を表 4.21 に示す。分布の形状を見ると全体的に平均 0 の方向に縮小しているが，二山の形状がやや緩和されおおむね単峰型の分布になっていることが分かる。分布のパラメタを見ると全体的に平均 0，標準偏差 1 の方向に縮小しているが，効果量を見ると MMLE に比べて尺度の縮小が緩和されていることが分かる。

表 4.21 周辺ベイズ推定法による推定項目パラメタ（国語）

Item	a	b	Item	a	b
a101	1.048	-3.502	c106	0.603	-1.814
a102	0.546	-0.870	c107	1.007	-1.132
a103	0.844	-1.358	c108	0.577	-1.633
a104	1.009	-2.684	c109	0.843	-1.430
a105	0.649	-1.857	c110	0.560	-0.607
a106	0.308	-2.110	d101	0.431	-0.204
a107	0.458	0.020	d102	0.530	0.378
b101	0.798	-2.033	d103	0.773	-3.197
b102	0.673	-1.316	d104	0.648	-1.849
b103	0.849	-2.716	d105	0.553	-0.460
b104	0.530	-3.618	d106	0.535	-0.775
b105	0.560	-1.473	d107	0.312	-0.574
b106	1.026	-0.679	d108	0.355	-0.164
b107	0.425	-0.400	d109	0.802	-0.795
b108	0.602	-1.886	d110	0.594	-1.228
b109	0.717	0.647	e101	1.099	-2.284
b110	0.112	-1.964	e102	0.883	-1.678
st101	0.499	4.128	e103	0.756	-0.707
st102	0.823	-1.251	e104	1.061	-1.762
st103	0.273	0.548	e105	0.597	-2.776
st104	0.287	-4.359	e106	0.591	-1.312
st105	0.322	-0.331	e107	0.386	-0.878
st106	0.831	-1.555	e108	0.403	0.350
st107	0.197	1.853	e109	0.651	0.083
st108	0.242	4.737	e110	0.620	-1.820
st109	0.682	-1.503	f101	0.340	-0.062
st110	0.412	-1.743	f102	0.567	0.362
a108	0.826	-2.032	f103	0.523	-0.928
a109	0.596	-1.660	f104	0.734	-1.238
a110	0.744	-1.101	f105	0.754	-1.729
c101	0.385	-0.519	f106	0.492	-1.041
c102	0.406	-0.190	f107	0.610	-1.885
c103	0.493	-0.719	f108	0.753	-0.060
c104	0.511	-1.145	f109	0.754	0.572
c105	0.782	-1.708	f110	0.599	0.778

表 4.22 周辺ベイズ推定法による推定の標準誤差（国語）

Item	a	b	Item	a	b
a101	0.198	0.291	c106	0.066	0.154
a102	0.083	0.118	c107	0.082	0.067
a103	0.102	0.085	c108	0.066	0.153
a104	0.145	0.167	c109	0.081	0.101
a105	0.090	0.131	c110	0.066	0.108
a106	0.072	0.297	d101	0.047	0.112
a107	0.089	0.233	d102	0.055	0.096
b101	0.077	0.106	d103	0.148	0.456
b102	0.066	0.086	d104	0.074	0.187
b103	0.099	0.166	d105	0.056	0.107
b104	0.081	0.367	d106	0.053	0.113
b105	0.059	0.103	d107	0.043	0.170
b106	0.081	0.052	d108	0.044	0.131
b107	0.055	0.117	d109	0.068	0.082
b108	0.064	0.122	d110	0.060	0.133
b109	0.088	0.142	e101	0.168	0.226
b110	0.047	0.626	e102	0.097	0.156
st101	0.059	0.389	e103	0.065	0.096
st102	0.044	0.055	e104	0.124	0.149
st103	0.025	0.114	e105	0.092	0.394
st104	0.037	0.523	e106	0.063	0.170
st105	0.026	0.090	e107	0.046	0.185
st106	0.048	0.065	e108	0.044	0.112
st107	0.025	0.264	e109	0.054	0.079
st108	0.034	0.631	e110	0.082	0.254
st109	0.042	0.076	f101	0.069	0.255
st110	0.032	0.132	f102	0.075	0.124
a108	0.140	0.169	f103	0.084	0.251
a109	0.108	0.170	f104	0.110	0.218
a110	0.137	0.137	f105	0.131	0.299
c101	0.049	0.118	f106	0.084	0.287
c102	0.049	0.110	f107	0.112	0.372
c103	0.053	0.100	f108	0.088	0.110
c104	0.055	0.118	f109	0.087	0.095
c105	0.076	0.114	f110	0.074	0.110

表 4.23 周辺ベイズ推定法による推定項目パラメタ（数学）

Item	a	b	Item	a	b
st101	0.663	-2.151	c106	0.950	0.500
a101	0.000	0.000	c107	0.240	2.047
a102	0.958	-0.868	c108	0.567	-0.479
a103	0.887	-1.998	c109	1.404	0.738
a104	1.201	-1.745	c110	0.969	0.687
a105	0.635	-2.721	d101	0.501	-2.930
a106	0.928	-1.231	d102	0.920	0.624
a107	1.343	-0.804	d103	0.939	0.433
a108	1.232	-1.924	d104	1.020	-0.767
a109	0.795	-1.990	d105	1.886	2.268
a110	0.691	-0.620	d106	0.183	1.332
st102	0.780	-1.316	d107	0.359	-0.069
st103	0.501	1.096	d108	0.485	-0.454
b101	0.590	-2.569	d109	0.522	-0.155
b102	0.434	-1.939	d110	0.569	3.300
b103	0.965	-0.642	e101	0.991	-0.552
b104	0.515	-1.722	e102	0.702	0.851
b105	0.869	-0.927	e103	0.734	-0.559
b106	0.531	-0.395	e104	0.181	1.873
b107	0.446	-0.469	e105	0.205	1.696
b108	0.880	-0.635	e106	0.951	-0.295
b109	0.940	-1.717	e107	0.590	0.365
b110	0.992	-0.548	e108	0.990	3.325
st104	1.258	-0.321	e109	0.204	4.007
st105	0.713	1.629	e110	1.116	1.102
st106	1.089	-0.188	f101	0.998	-0.323
st107	0.200	2.889	f102	0.783	0.056
st108	0.541	-0.363	f103	0.997	1.054
st109	0.650	2.534	f104	0.319	-3.453
st110	0.836	-1.313	f105	0.437	1.017
c101	0.578	-2.028	f106	0.360	0.605
c102	0.881	-1.190	f107	0.616	0.380
c103	0.992	-0.402	f108	0.667	0.350
c104	1.013	-0.919	f109	0.905	3.378
c105	1.953	1.482	f110	0.96469	3.83302

表 4.24 周辺ベイズ推定法による推定の標準誤差（数学）

Item	a	b	Item	a	b
st101	0.045	0.102	c106	0.080	0.069
a101	NA	NA	c107	0.053	0.531
a102	0.120	0.076	c108	0.061	0.091
a103	0.116	0.106	c109	0.131	0.063
a104	0.135	0.070	c110	0.097	0.083
a105	0.112	0.242	d101	0.080	0.419
a106	0.116	0.072	d102	0.074	0.060
a107	0.155	0.061	d103	0.083	0.065
a108	0.143	0.078	d104	0.085	0.069
a109	0.109	0.115	d105	0.383	0.141
a110	0.108	0.119	d106	0.045	0.365
st102	0.043	0.054	d107	0.048	0.127
st103	0.032	0.089	d108	0.053	0.110
b101	0.075	0.190	d109	0.055	0.096
b102	0.060	0.161	d110	0.111	0.489
b103	0.079	0.054	e101	0.084	0.070
b104	0.061	0.117	e102	0.062	0.075
b105	0.075	0.057	e103	0.070	0.096
b106	0.060	0.099	e104	0.044	0.421
b107	0.058	0.114	e105	0.044	0.340
b108	0.075	0.058	e106	0.080	0.066
b109	0.087	0.074	e107	0.056	0.079
b110	0.091	0.060	e108	0.177	0.310
st104	0.061	0.031	e109	0.051	0.895
st105	0.046	0.090	e110	0.099	0.063
st106	0.058	0.038	f101	0.124	0.097
st107	0.029	0.443	f102	0.107	0.109
st108	0.035	0.063	f103	0.143	0.102
st109	0.062	0.181	f104	0.100	1.153
st110	0.054	0.068	f105	0.070	0.163
c101	0.070	0.185	f106	0.068	0.176
c102	0.076	0.075	f107	0.081	0.107
c103	0.080	0.054	f108	0.085	0.102
c104	0.084	0.062	f109	0.226	0.408
c105	0.292	0.097	f110	0.300	0.560

表 4.25　周辺ベイズ推定法による項目適合度（国語）

Item	N	InFit	StdInFit	OutFit	StdOutFit	Item	N	InFit	StdInFit	OutFit	StdOutFit
a101	409	1.270	1.030	1.621	7.890	c106	766	0.989	-0.174	0.971	-0.577
a102	400	0.949	-1.579	0.925	-1.073	c107	765	0.937	-1.171	0.773	-4.734
a103	393	0.970	-0.563	0.899	-1.452	c108	687	0.999	-0.002	0.911	-1.679
a104	397	1.169	1.155	1.569	7.167	c109	675	0.980	-0.304	0.846	-2.946
a105	408	1.035	0.640	1.026	0.374	c110	564	0.958	-1.208	0.926	-1.260
a106	407	1.015	0.438	1.025	0.357	d101	781	0.934	-2.911	0.915	-1.713
a107	365	0.924	-1.878	0.893	-1.482	d102	645	0.909	-3.386	0.877	-2.273
b101	749	1.059	0.957	0.989	-0.213	d103	786	1.026	0.192	0.954	-0.920
b102	686	0.992	-0.205	0.953	-0.889	d104	759	0.955	-0.589	0.867	-2.676
b103	735	1.142	1.313	0.954	-0.887	d105	696	0.914	-2.601	0.871	-2.499
b104	785	1.106	1.017	1.105	2.022	d106	792	0.936	-1.847	0.870	-2.681
b105	742	0.999	-0.010	0.968	-0.622	d107	792	0.963	-1.815	0.955	-0.897
b106	729	0.874	-3.295	0.771	-4.673	d108	782	0.956	-2.310	0.947	-1.066
b107	703	0.956	-2.002	0.951	-0.923	d109	789	0.905	-2.070	0.754	-5.238
b108	744	1.030	0.645	1.023	0.432	d110	789	0.940	-1.256	0.839	-3.336
b109	542	0.863	-2.580	0.711	-5.179	e101	810	0.995	0.037	0.469	-12.942
b110	712	1.005	0.431	1.006	0.121	e102	810	0.913	-0.879	0.659	-7.622
st101	1746	1.126	1.141	0.886	-3.471	e103	811	0.862	-2.942	0.735	-5.754
st102	1917	0.943	-1.715	0.789	-6.934	e104	811	0.987	-0.067	0.397	-15.369
st103	1905	0.971	-2.550	0.973	-0.831	e105	810	0.949	-0.356	0.817	-3.865
st104	1866	0.998	-0.025	0.982	-0.558	e106	789	0.911	-1.530	0.798	-4.245
st105	1891	0.964	-2.980	0.959	-1.262	e107	809	0.932	-2.129	0.906	-1.935
st106	1891	0.981	-0.464	0.792	-6.771	e108	803	0.934	-3.386	0.926	-1.510
st107	1864	0.983	-1.100	0.978	-0.664	e109	803	0.861	-4.672	0.822	-3.744
st108	1834	1.005	0.113	1.033	0.991	e110	716	0.935	-0.705	0.719	-5.786
st109	1805	0.961	-1.075	0.872	-3.992	f101	331	0.943	-1.781	0.933	-0.874
st110	1799	0.959	-1.460	0.958	-1.277	f102	382	0.891	-2.989	0.852	-2.130
a108	220	1.038	0.381	1.174	1.746	f103	369	0.909	-1.346	0.805	-2.788
a109	230	1.017	0.271	0.979	-0.229	f104	398	0.874	-1.280	0.756	-3.681
a110	174	0.970	-0.410	0.902	-0.936	f105	397	0.886	-0.799	0.751	-3.771
c101	776	0.971	-1.433	0.965	-0.691	f106	373	0.903	-1.404	0.837	-2.319
c102	771	0.962	-1.893	0.957	-0.860	f107	398	0.900	-0.784	0.794	-3.083
c103	774	0.961	-1.450	0.942	-1.150	f108	397	0.831	-3.367	0.779	-3.313
c104	774	0.972	-0.816	0.936	-1.269	f109	380	0.839	-3.787	0.789	-3.081
c105	771	0.997	-0.033	0.872	-2.604	f110	398	0.889	-3.052	0.902	-1.417

表 4.26 周辺ベイズ推定法による項目適合度（数学）

Item	N	InFit	StdInFit	OutFit	StdOutFit	Item	N	InFit	StdInFit	OutFit	StdOutFit
st101	1967	0.978	-0.492	0.954	-1.446	c107	672	0.988	-0.314	0.992	-0.147
a102	395	0.924	-1.728	0.874	-1.836	c108	644	0.965	-1.186	0.954	-0.828
a103	408	1.014	0.232	1.022	0.315	c109	599	0.849	-2.195	0.655	-6.648
a104	405	0.990	-0.119	0.944	-0.807	c110	518	0.893	-1.885	0.988	-0.195
a105	410	1.041	0.518	1.119	1.655	d101	776	0.980	-0.172	0.951	-0.983
a106	389	0.951	-1.097	0.939	-0.856	d102	753	0.904	-2.417	0.842	-3.190
a107	381	0.855	-2.676	0.740	-3.870	d103	565	0.905	-2.059	0.826	-3.068
a108	399	1.014	0.207	0.894	-1.540	d104	747	0.905	-1.840	0.729	-5.678
a109	409	1.020	0.351	1.003	0.044	d105	414	0.984	0.023	0.153	-19.590
a110	383	0.937	-1.577	0.922	-1.103	d106	762	0.996	-0.260	0.997	-0.065
st102	1982	0.941	-2.012	1.000	-0.007	d107	754	0.977	-1.194	0.979	-0.419
st103	1977	0.942	-2.317	0.933	-2.128	d108	745	0.966	-1.229	0.944	-1.096
b101	761	1.052	0.861	1.002	0.047	d109	708	0.956	-1.601	0.943	-1.086
b102	768	1.002	0.085	0.991	-0.179	d110	588	1.026	0.235	0.941	-1.020
b103	772	0.923	-2.225	0.877	-2.506	e101	783	0.864	-2.916	0.928	-1.455
b104	785	0.996	-0.130	0.998	-0.032	e102	747	0.932	-2.027	0.906	-1.863
b105	742	0.940	-1.823	0.873	-2.527	e103	733	0.910	-2.113	0.883	-2.302
b106	779	0.958	-1.656	0.957	-0.848	e104	787	1.000	-0.010	1.000	-0.003
b107	742	0.973	-1.221	0.974	-0.510	e105	788	0.996	-0.179	1.005	0.103
b108	776	0.915	-2.637	0.902	-1.973	e106	735	0.884	-2.717	0.806	-3.932
b109	718	0.995	-0.089	0.925	-1.458	e107	773	0.938	-2.285	0.941	-1.178
b110	594	0.922	-1.923	0.826	-3.142	e108	621	1.128	0.578	2.215	17.703
st104	1693	0.850	-4.724	0.776	-6.948	e109	761	1.003	0.087	1.023	0.446
st105	1800	0.951	-1.132	0.858	-4.439	e110	559	0.872	-2.369	0.744	-4.603
st106	1409	0.881	-3.581	0.793	-5.819	f101	366	0.866	-2.110	0.699	-4.456
st107	1731	0.987	-0.489	1.003	0.080	f102	326	0.901	-1.941	0.821	-2.404
st108	1549	0.950	-2.450	0.919	-2.303	f103	212	0.877	-1.605	0.768	-2.544
st109	1242	1.015	0.218	0.911	-2.278	f104	393	0.957	-0.337	1.000	-0.003
st110	1332	0.942	-1.392	0.905	-2.509	f105	388	0.968	-0.880	0.987	-0.179
c101	741	0.984	-0.257	0.997	-0.063	f106	377	0.976	-0.869	0.976	-0.334
c102	756	0.938	-1.320	0.889	-2.220	f107	390	0.935	-1.703	0.918	-1.172
c103	707	0.899	-2.545	0.798	-4.028	f108	381	0.929	-1.723	0.896	-1.478
c104	718	0.926	-1.581	0.842	-3.120	f109	216	1.185	0.682	0.477	-6.544
c105	529	0.787	-1.276	0.208	-19.213	f110	231	1.208	0.577	0.469	-6.915
c106	725	0.901	-2.153	0.906	-1.831						

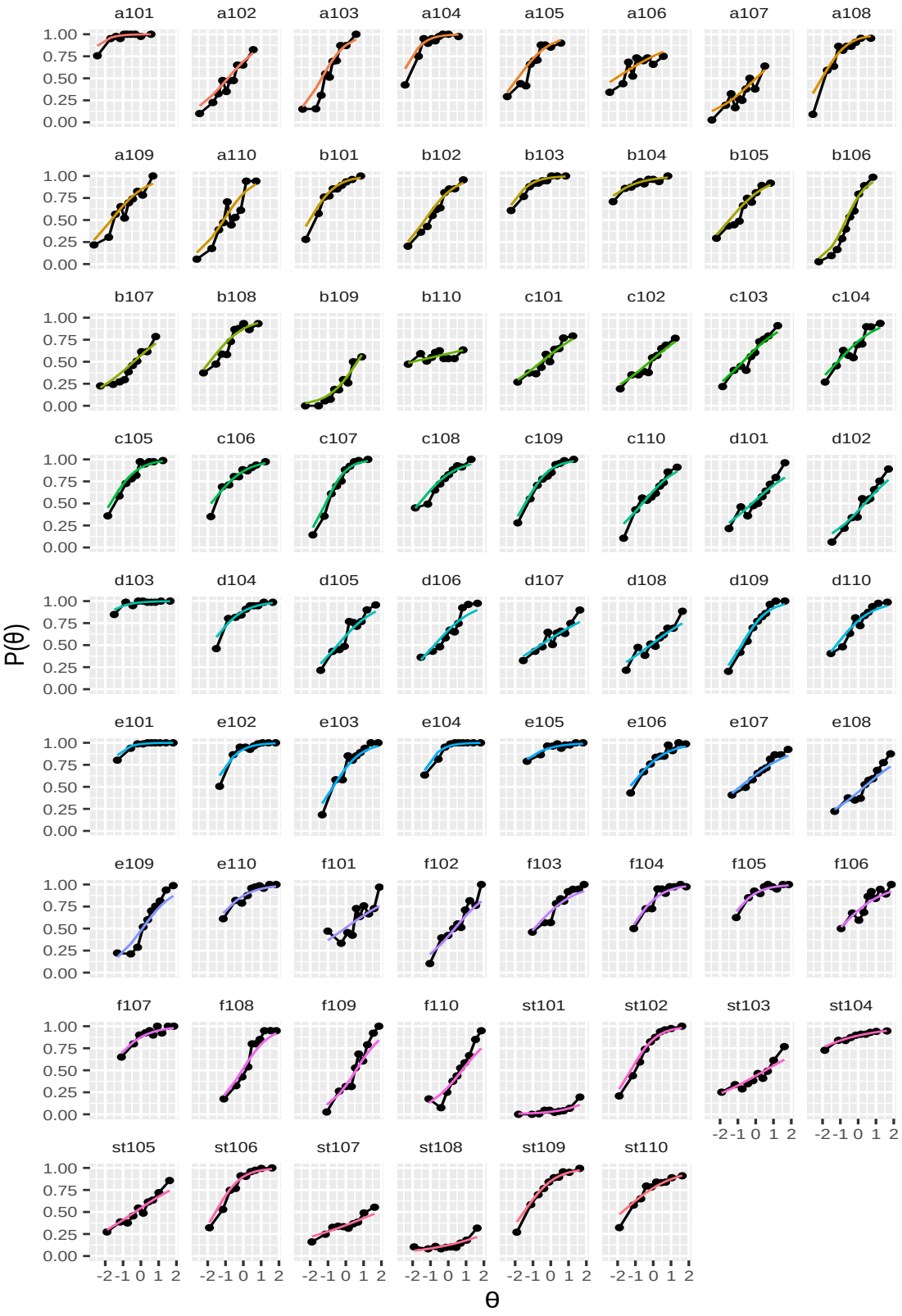

図 4.37 周辺ベイズ推定法による推定値の適合度（国語）

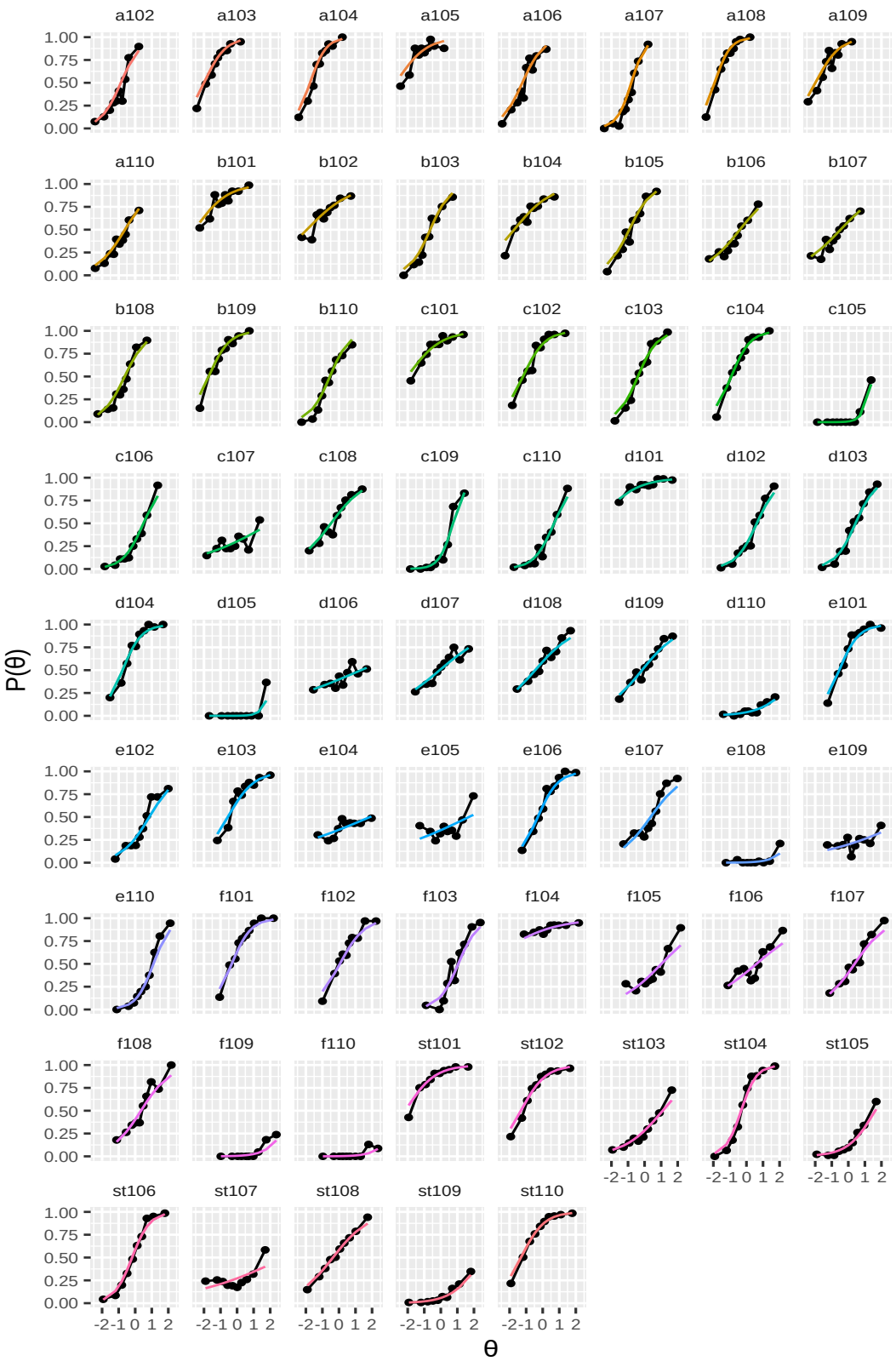

図 4.38 周辺ベイズ推定法による推定値の適合度（数学）

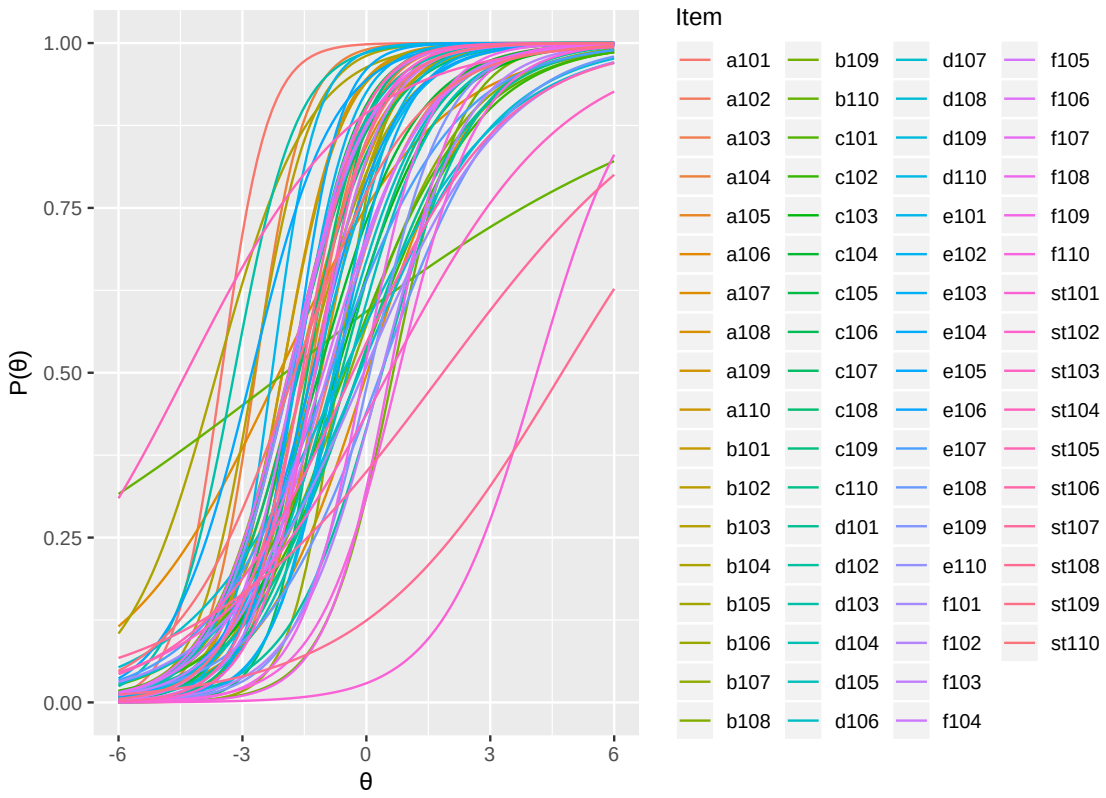

図 4.39 周辺ベイズ推定法による項目特性曲線（国語）

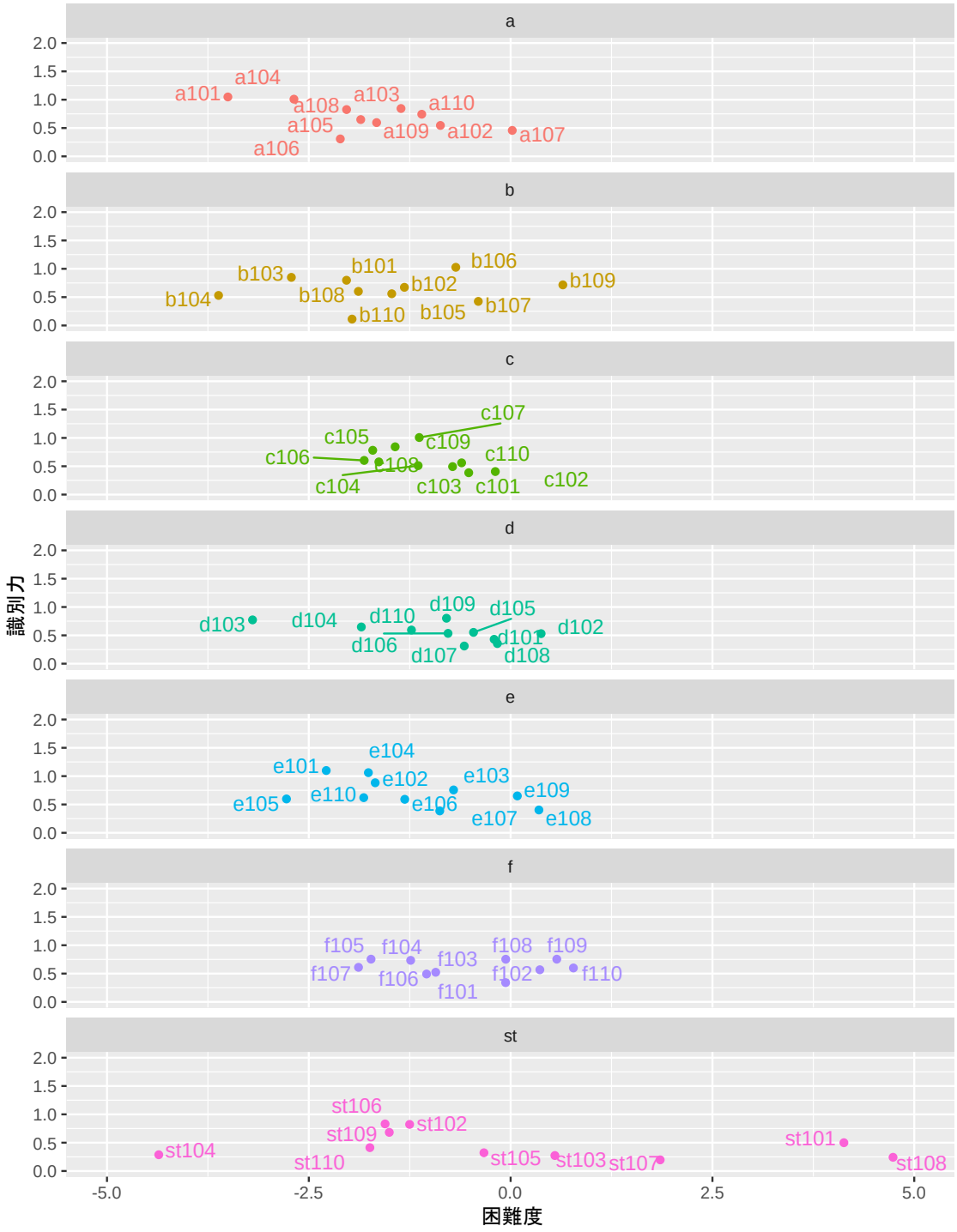

図 4.40 周辺ベイズ推定法による推定パラメタの散布図（国語）

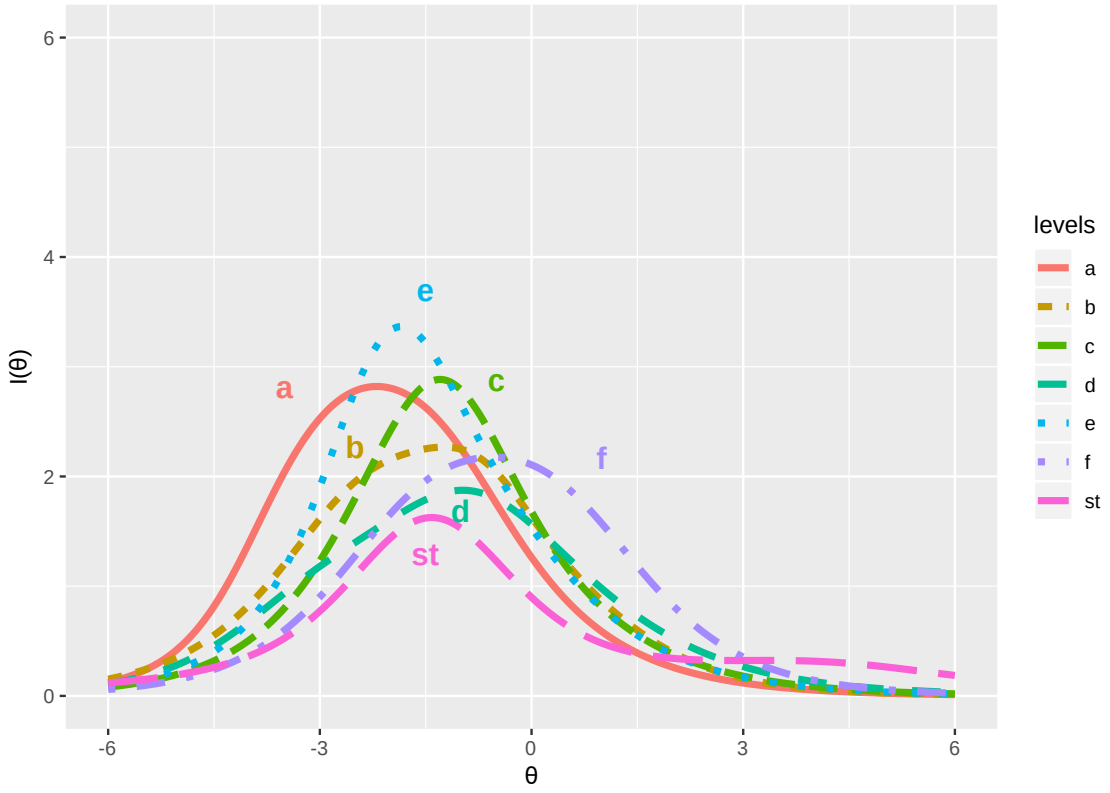

図 4.41 周辺ベイズ推定法によるレベルごとのテスト情報関数（国語）

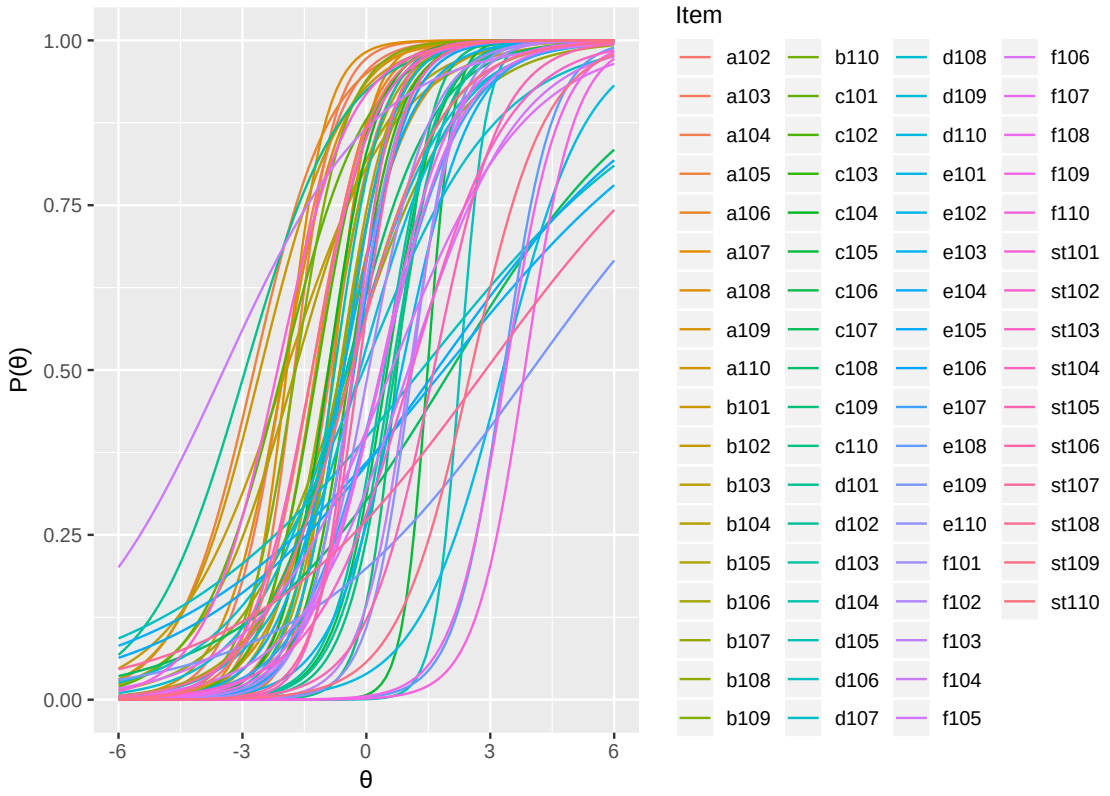

図 4.42 周辺ベイズ推定法による項目特性曲線（数学）

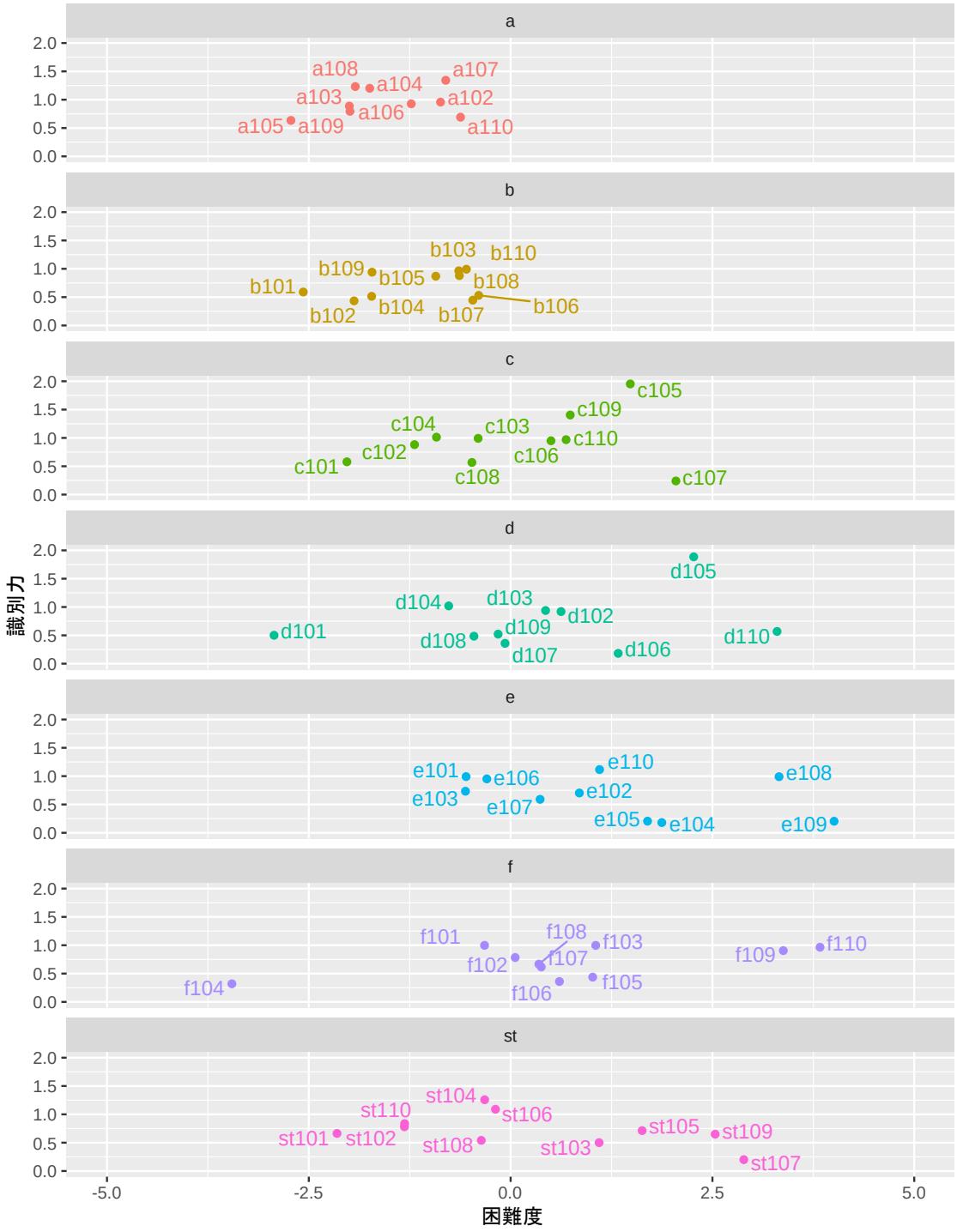

図 4.43 周辺ベイズ推定法による推定パラメタの散布図（数学）

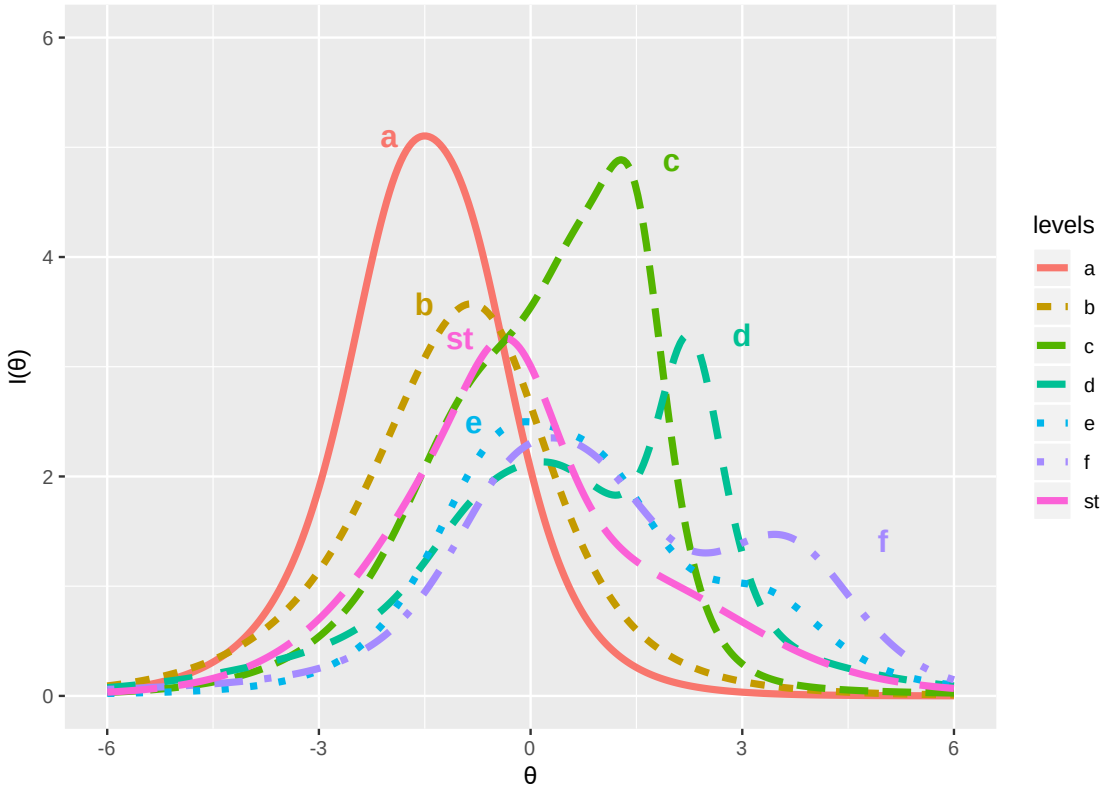

図 4.44 周辺ベイズ推定法によるレベルごとのテスト情報関数（数学）

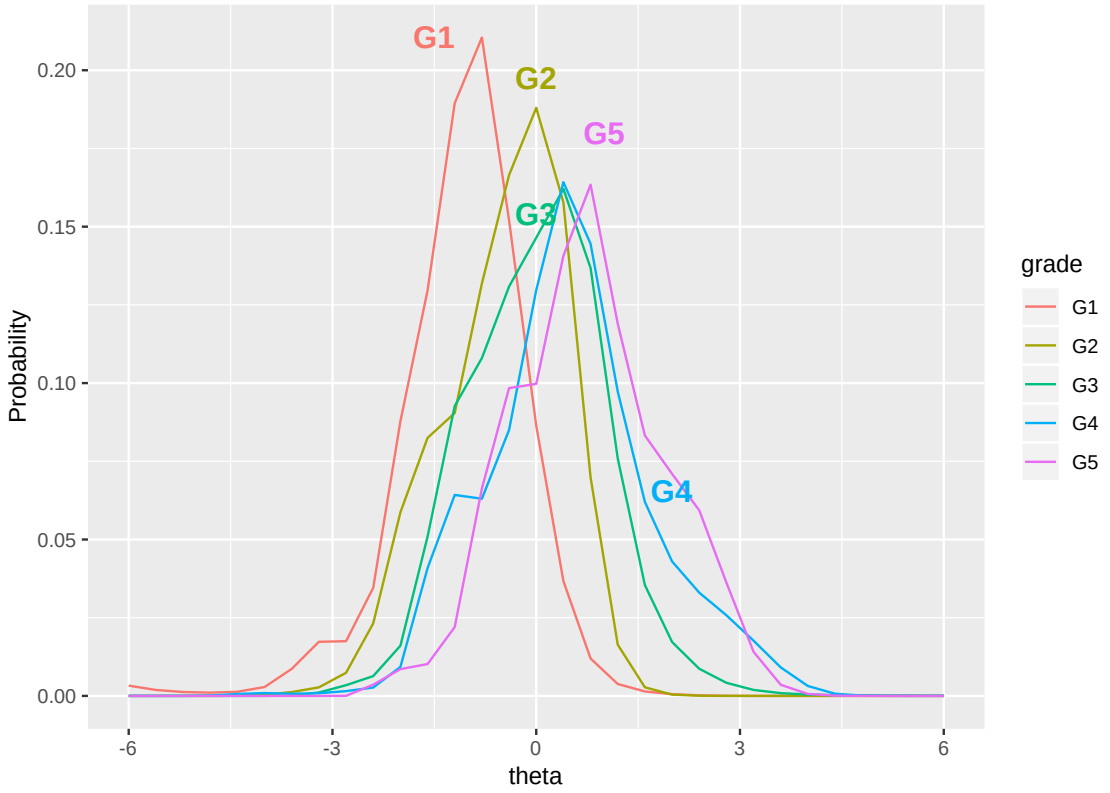

図 4.45 周辺ベイズ推定法による推定母集団分布（国語）

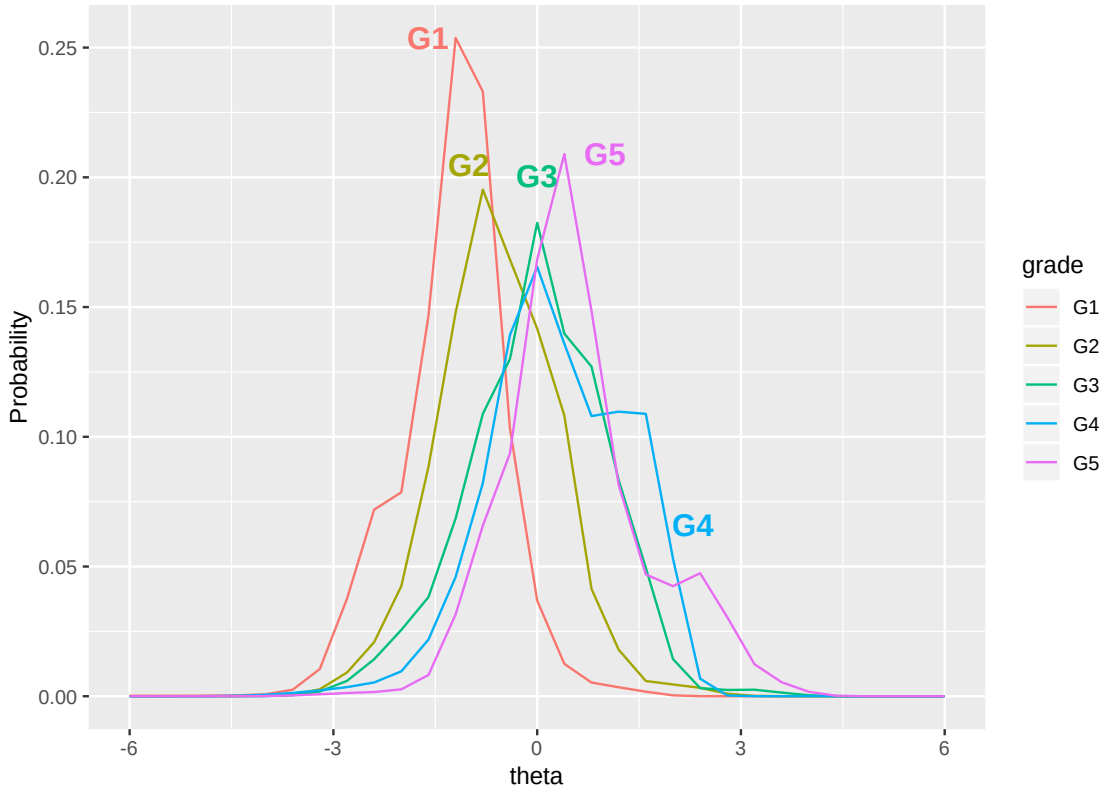

図 4.46 周辺ベイズ推定法による推定母集団分布（数学）

表 4.27 周辺ベイズ推定法による推定母集団分布のパラメタと効果量

	国語				算数・数学			
学年	N	平均	SD	効果量	N	平均	SD	効果量
G1	411	-1.084	0.947	—	410	-1.237	0.774	—
G2	386	-0.477	0.896	0.658	385	-0.588	0.882	0.784
G3	391	0.011	1.005	0.512	391	0.003	1.018	0.620
G4	415	0.472	1.218	0.412	413	0.301	1.018	0.293
G5	390	0.749	1.111	0.237	397	0.568	1.054	0.258

4.3.3 考察

MBE を用いることで推定値の標準誤差を低下させることができ，項目適合度も全体的に向上した推定値を得ることができた。特に，絶対値の大きな困難度の項目や低い識別力の項目のパラメタの推定値をより安定させることができており，シミュレーションにもとづく結果ではないが，少数標本の推定時には MBE が有効な手法であることが分かった。

しかし，推定誤差を低下させ，適合度を向上させてもなお，尺度の縮小は確認された。また，分布が多峰から単峰に近づいても同様に尺度の縮小が生じている。したがって分布の多峰性は尺度の縮小の直接の要因でないと結論づけることができる。

5　結論

IRT 垂直尺度化は IRT モデルに基づいて推定された異なる難易度のテスト項目のパラメタを共通尺度化する手法である。IRT モデルには正答と誤答という二値型の反応データをモデリングしたもの (1PLM〜3PLM) 以外に部分点を認めるモデル（GPCM）やひとつのテストが複数の構成概念を測定していることを許容するモデル (MIRT モデル), 項目反応に関わる攪乱要因を個人の誤差分散として定義するモデル (GIRT モデル) などの様々な拡張が存在する。モデル選択には適合度などの数値的な指標を用いることができるが，何よりもモデルの持つ仮定がデータに当てはまっていることと，安定的に実行するための条件が揃っているかどうかを慎重に検討しなくてはならない。

IRT 垂直尺度化についてのシミュレーション分析と実データの分析から以下のことが明らかになった。まず，垂直尺度化特有のデータ収集デザインである多母集団の尺度化テストデザインにおいては，従来の等化係数推定方法よりも，一度にすべての等化係数を推定する手法や，等化係数を推定せず，多母集団モデルを仮定する推定方法の方が性能がよい。また，母集団分布の推定においてはごくわずかだが，同時尺度調整法では標準偏差を過大推定する傾向が見られた。また calr の方法は比較的マイナーなプログラムであり，一般にそれほど広くは普及していない方法である。そして今回使用した lazy.irt の calr 関数では希に推定値が大きく真値から逸脱するケースがあった。これらを踏まえると尺度化テストデザインにおいては同時尺度調整法にてパラメタを推定する方法が最も良い方法であると結論づけた。しかし同時尺度調整法を実行する際は基準となる学年を適切に選択肢，区分求積の分点の範囲を十分広くとる必要があるだろう。

実データの分析では大きくふたつの知見が得られた。ひとつは尺度の縮小に関して先行研究と類似した結果が得られたが，原因は項目局所依存と測定誤差以外の要因である可能性が大きいということである。尺度の縮小の原因として残された可能性は多次元性と construct shift であるが，標本数が少ないことを加味すると，多次元 IRT モデルのような複雑なモデルを用いてある程度精度の高いパラメタ推定することは非常に困難であると予想される。もうひとつの知見は少数標本の推定において MBE が有効であるということである。MMLE-EM では許容し得ない値をとっていた項目であっても，適切と思われる推定値を得ることができた。しかし事前分布がやや制約の強い分布であったため，ほとんどのパラメタが常識的な範囲の値に収束したとも考えられる。もちろん本来であれば弱情報事前分布を適用し，徐々に条件を変えながら，得られた推定値の頑健性を検討すべき (松浦, 2016) であるが，少なくとも適合度と標準誤差の観点からはこの設定でも問題のない推定値が得られていると言えるだろう。事前分布の設定を変えることで，たとえ同じデータからであってもより適切な推定値を得ることができることも，ベイズ推定法のひとつのメリットである。

今後の課題はモデルを改良することである。モデルについては，GIRT モデルが有効である可能性が高い。測定の多次元性や construct shift を考慮するのであれば MIRT モデルやその一種で

ある双因子モデルも適当かもしれないが，孫 (1997) が指摘したように，次元がはっきりと分けられない場合には MIRT モデルは適切ではない。たとえば坂本 (2015) は数学の下位領域を知識，推論，応用として TIMSS データに対していくつかの MIRT モデルを適用しているが，本来これらの次元はお互いに密接に関連し合い，はっきりと別の次元であることを想定することは不適切かもしれない。GIRT モデルは 2PLM に潜在変数をひとつだけ加えたモデルであるため，はっきりと次元を分けられない場合に有効であることに加えて，MIRT モデルよりも安定した推定に必要な標本数も少なくてすむと考えられる。

付記

本研究は科研費 (16H03731) の助成を受けたものである。

謝辞

本稿を書き上げるにあたって，指導教員である柴山直教授には大変お世話になりました。心より感謝申し上げます。研究指導は勿論のこと，筆者の卒業論文および本稿のテーマである IRT と垂直尺度化という研究テーマのきっかけをいただき，さらに研究に対する姿勢やプログラミングについて非常に多くのことを学ばせていただきました。

また，副指導教員である熊谷龍一准教授には，本稿の実験部分の核となる推定プログラムの作成とシミュレーションに関して，大いにご助言いただきました。筆者の唐突な質問に対しても，いつも快くお答えいただきありがとうございました。

さらに，教育設計評価専攻と教育情報アセスメントコース教育評価測定論領域の先輩，同級生，学部生，事務補佐の方々にも，本稿を書き上げるにあたって暖かい励ましをいただきました。ありがとうございました。

最後に，実家から離れた仙台での一人暮らしを支えていただいた両親と，何かと心配をかけた姉にも心より感謝申し上げます。

参考文献

Adams, R. J., Wilson, M., Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, *21*(1), 1-23.

Arai, S., Mayekawa, S. (2011). A comparison of equating methods and linking designs for developing an item pool under item response theory. *Behaviormetrika*, *38*(1), 1-16.

Baker, F. B., Kim, S. H. (2004). *Item Response Theory: Parameter Estimation Techniques* (2nd ed.). Boca Raton: CRC Press.

Beguin, A. A., Hanson, B. A., Glas, C. A. W. (2000). Effect of Multidimensionality on Separate and Concurrent Estimation in IRT Equating. *Paper presented at the Annual Meeting of the National Council on Measurement in Education*.

Betebenner, D. W., Linn, R. L. (2009). Measurement Challenges Within the Race to the Top Agenda Growth in Student Achievement: Issues of Measurement, Longitudinal Data Analysis, and Accountability. Retrieved from http://www.k12center.org/publications.html.

Blanton, H., Jaccard, J. (2006). Arbitrary metrics in psychology. *American Psychologist*, *61*(1), 27-41.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*(1), 29-51.

Bock, R. D., Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika*, *35*(2), 179-197.

Bock, R. D., Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443-459.

Bock, R. D., Gibbons, R., Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement*, *12*(3), 261-280.

Briggs, D. C., Weeks, J. P. (2009). The impact of vertical scaling decisions on growth interpretations. *Educational Measurement: Issues and Practice*, *28*(4), 3-14.

Burket, G. R. (1984). Response to Hoover. *Educational Measurement: Issues and Practice*, *3*(4), 15-16.

Cai, L., du Toit, S. H. C., Thissen, D. (2011). IRTPRO. Skokie, IL: Scientific Software International, Inc.

Camilli, G. (1988). Scale shrinkage and the estimation of latent distribution parameters. *Journal of Educational Statistics*, *13*(3), 227-241.

Camilli, G. (1994). Origin of the scaling constant d = 1.7 in item response theory. *Journal of Educational and Behavioral Statistics*, *19*(3), 293-295.

Camilli, G. (1999). Measurement error, multidimensionality, and scale shrinkage: A reply to Yen and Burket. *Journal of Educational Measurement Spring*, *36*(1), 73-78.

Camilli, G., Yamamoto, K., Wang, M. M. (1993). Scale shrinkage in vertical equating. *Applied Psychological Measurement*, *17*(4), 379-388.

Chen, W., Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, *22*(3), 265-289.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297-334.

de Ayala, R. J. (2009). *The Theory And Practice Of Item Response Theory*. New York: The Guilford Press.

Dempster, A. P., Laird, N. M., Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, *39*(1), 1-38.

Dorans, N. J. (2000). Distinctions among Classes of Linkages. *Research Notes. RN-11*. New York: College Entrance Examination Board. Retrieved from https://eric.ed.gov/?id=ED562636

Eastwood, M. (2014). The Effects of Construct Shift and Model-Data Misfit on Estimates of Growth Using Vertical Scales (doctoral dissertation). University of Connecticut, Storrs, Connecticut. Retrieved from http://digitalcommons.uconn.edu/dissertations/544

Feuer, M. J., Holland, P. W., Green, B. F., Bertenthal, M. W., Cadelle Hemphill, F. (1999). Uncommon Measure. Retrieved from http://www.nap.edu/catalog/6332.html

Fox, J. P. (2010). *Bayesian item response modeling: Theory and applications*. New York: Springer.

藤森進 (1991). 小学校3年生から5年生の算数学力尺度の作成. 心理学研究, *62*(2), 82-87.

藤森進 (2009). 部分得点モデルにおける同時尺度調整法による垂直的等化の研究 人間科学研究, *31*, 95-102.

藤森進 (2011). 部分得点モデルにおける同時尺度調整法による垂直的等化の改訂報告 人間科学研究, *32*, 21-29.

Golding, N. (2018). greta: Simple and Scalable Statistical Modelling in R. Retrieved from https://github.com/greta-dev/greta

南風原朝和 (1991). 項目反応理論概説 芝祐順 (編) 項目反応理論 (pp. 9-30). 東京大学出版会

南風原朝和 (2000). 個人正答確率に基づく局所独立性の概念の明確化 －－実験的独立性および一次元性との関係を中心に－－ Retrieved December 21, 2018, from http://www.p.u-tokyo.ac.jp/~haebara/local_ind/

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research*, *22*(3), 144-149.

Haley. D.C. (1952). Estimation of the Dosage Mortality Relationship When the Dose is Subject to Error. *Technical Report, 15*. Retrieved from https://statistics.stanford.edu/research/estimation-dosage-mortality-relationship-when-dose-subject-error

Hambleton, R. K., Jones, R. W. (2005). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, *12*(3), 38-47.

Han, K. T., Wells, C. S., Hambleton, R. K. (2015). Effect of adjusting pseudo-guessing parameter estimates on test scaling when item parameter drift is present. *Practical Assessment, Research & Evaluation*, *20*(16). Retrieved from https://pareonline.net/getvn.asp?v=20&n=16

Hanson, B. A., Beguin, A. A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement*, *26*(1), 3-24.

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement*, *9*(2), 139-164.

林規生 (1996). 生涯学習の観点から日本人の英語能力発達過程を探る――項目反応理論の応用―― JACET全国大会要綱, *35*, 152-155.

Holland, P. W., Dorans, N. J. (2006). Linking and Equating. In L. R. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 187-220). Westport, CT: Praeger Publishers.

Hoover, H. D. (1984). The most appropriate scores for measuring educational development in the elementary schools: GE's. *Educational Measurement: Issues and Practice*, *3*(4), 8-14. Wiley/Blackwell (10.1111).

市川伸一 (1991). 心理測定法への招待――測定からみた心理学入門―― サイエンス社

印東太郎 (1995). 尺度化の意義. 行動計量学, *22*(2), 135-154.

石井秀宗・安永和央 (2011). 全項目が開示されるテスト文化のもとでの得点分布の経年比較--全国テストと自治体テストのリンキング 日本テスト学会誌, *7*(1), 24-35.

Ito, K., Sykes, R. C., Yao, L. (2008). Concurrent and separate grade-groups linking procedures for vertical scaling. *Applied Measurement in Education*.

樺島祥介・上田修功 (2003). 平均場近似・EM法・変分ベイズ法 甘利俊一・竹内啓・竹村彰通・伊庭幸人 (編), 計算統計I――確率計算の新しい手法―― (pp. 121-191). 岩波書店

Karkee, T., Lewis, D. M., Hoskens, M., Yao, L., Haug, C. (2003). Separate versus concurrent calibration methods in vertical scaling. *the Annual Meeting of the National Council on Measurement in Education*. Chicago, IL. Retrieved from https://eric.ed.gov/?id=ED478167

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

Kenyon, D. M., MacGregor, D., Li, D., Cook, H. G. (2011). Issues in vertical scaling of a K-12 English language proficiency test. *Language Testing*, *28*(3), 383-400.

Kim, S., Cohen, A. S. (2002). A comparison of linking and concurrent calibration under the graded response model. *Applied Psychological Measurement*, *26*(1), 25-41.

Kim, S., Cohen, A. S. (1998). A comparison of linking and concurrent calibration under item response theory. *Applied Psychological Measurement*, *22*(2), 131-143.

喜岡恵子 (1991). 計算能力の尺度化 芝祐順 (編) 項目反応理論 (pp. 163-174). 東京大学出版会

Koepfler, J. (2012). Examining the Bifactor IRT Model for Vertical Scaling in K-12 Assessment (doctoral dissertation). James Madison University, Harrisonburg, VA. Retrieved from http://commons.lib.jmu.edu/diss201019

Kolen, M. J., Brennan, R. L. (2004). *Test Equating, Scaling, and Linking: Methods and Practices (2nd ed.)* (2nd ed.). New York: Springer.

Kolen, M. J. (2004). Population invariance in equating and linking: Concept and history. *Journal of Educational Measurement*, *41*(I), 3-14.

Kolen, M. J. (2004). Linking assessments: Concept and history. *Applied Psychological Measurement*, *28*(4), 219-226.

Kolen, M. J., Brennan, R. L. (2016). *Test Equating, Scaling, and Linking Methods and Practices.* (3rd ed.). New York: Springer Verlag.

熊谷龍一・山口大輔・小林万里子・別府正彦・脇田貴文・野口裕之 (2007). 大規模英語学力テストにおける年度間・年度内比較――大学受験生の英語学力の推移―― *Japanese journal for research on testing*, *3*(1), 83-90.

熊谷龍一 (2009). 初学者向けの項目反応理論分析プログラムEasyEstimationシリーズの開発. 日本テスト学会誌, *5*(1), 107-118.

熊谷龍一・野口裕之 (2012). 推定母集団分布を利用した共通受験者法による等化係数の推定. 日本テスト学会誌, *8*(1), 9-18. Retrieved from https://ci.nii.ac.jp/naid/40019469809

熊谷龍一 (2012). 統合的DIF検出方法の提案―“EasyDIF”の開発― 心理学研究, *83*(1), 35-43.

熊谷龍一・荘島宏二朗 (2015). 教育心理学のための統計学 誠信書房

Lee, O. (2003). Rasch simultaneous vertical equating for measuring reading growth. *Journal of Applied Measurement*, *4*(1), 10-23.

Li, Y., Lissitz, R. W. (2012). Exploring the full-information bifactor model in vertical scaling with construct shift. *Applied Psychological Measurement*, *36*(1), 3-20.

Linn, R. L. (1993). Linking results of distinct assessments. *Applied Measurement in Education*, *6*(1), 83-102.

Liu, Y., Maydeu-Olivares, A. (2012). Local Dependence Diagnostics in IRT Modeling of Binary Data. *Educational and Psychological Measurement*, *73*(2), 254–274.

Lord, F. M., Novick, M. R., Birnbaum, A. (1968). *Statistical theories of mental test scores.* Oxford, England: Addison-Wesley.

Lord, F. M., Novick, M. R., Birnbaum, A. (2008). *Statistical theories of mental test scores*. Information Age Pub.

Lord, F. M. (1952). *A Theory of Test Scores*. *Psychometric monographs* (Vol. 7). Retrieved from https://www.psychometricsociety.org/sites/default/files/pdf/MN07.pdf

Lord, F. M. (1953). The relation of test score to the trait underlying the test. *Educational and Psychological Measurement*, *13*(4), 517-549.

Lord, F. M. (1975). The “ability” scale in item characteristic curve theory. *Psychometrika*, *40*(2), 205-217.

Lord, F. M. (1980). *Applications of Item Response Theory To Practical Testing Problems*. New York: Routledge. Retrieved from https://www.amazon.co.jp/dp/B00ABL6D40/ref=dp-kindle-redirect?_encoding=UTF8&btkr=1

Lord, F. M., Wingersky, M. S. (1984). Comparison of IRT True-Score and Equipercentile Observed-Score “Equatings.” *Applied Psychological Measurement*, *8*(4), 453-461.

Loyd, B. H., Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement*, *17*(3), 179-193.

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement*, *14*(2), 139-160.

Martineau, J. A. (2006). Distorting value added: The use of longitudinal, vertically scaled student achievement data for growth-based, value-added accountability. *Journal of Educational and Behavioral Statistics*, *31*(1), 35-62.

松浦健太郎 (2016). StanとRでベイズ統計モデリング　共立出版

前川眞一 (1991). 項目パラメタの推定　芝祐順 (編) 項目反応理論 (pp. 87-129). 東京大学出版会

Mayekawa, S. (2016). lazy.irt: Some IRT functions for lazy boys and girls.

Mckinley, R. L., Mills, C. N. (1985). A comparison of several goodness-of-fit statistics. *Applied Psychological Measurement*, *9*(1), 49-57.

McKinley, R. L., Reckase, M. D. (1982). The Use of the General Rasch Model with Multidimensional Item Response Data. Iowa City, Iowa. Retrieved from http://www.dtic.mil/docs/citations/ADA125099

Meng, H. (2007). A comparison study of IRT calibration methods for mixed-format tests in vertical scaling (doctoral dissertation). University of Iowa, Iowa. Retrieved from http://ir.uiowa.edu/etd/338http://ir.uiowa.edu/etdhttp://ir.uiowa.edu/etd/338.

Mislevy, R. J., Bock, R. D. (1982). BILOG: Maximum likelihood item analysis and test scoring with logistic models. Mooresville IN: Scientific Software.

Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika*, *51*(2), 177-195.

Mislevy, R. J. (1992). Linking Educational Assessments: Concepts , Issues, Methods, and Prospects. Princeton, NJ.

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

村木英治 (2011). 項目反応理論　朝倉書店

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *ETS Research Report Series*. Princeton, NJ.

村山功 (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.

Muthén, L. K., Muthén, B. O. (2006). Mplus User’s Guide. Los Angeles, CA.

永野重史 (2001). 発達とはなにか　東京大学出版会

中村知靖・豊田秀樹 (1991). 比較判断の法則と項目反応理論　芝祐順 (編) 項目反応理論 (pp. 201-209). 東京大学出版会.

中村知靖・前川眞一 (1993). 一般項目反応モデルにおける項目パラメタの周辺最尤推定法　教育心理学研究, *41*(1), p22-30.

中室牧子・星野崇宏・松岡亮二・益川弘如・二宮裕之・本橋幸康・及川賢 (2017). 埼玉県学力・学習状況調査のデータを活用した効果的な指導方法に関する分析研究.

Nelder, J. A., Mead, R. (1965). A Simplex Method for Function Minimization. *The Computer Journal*, *7*(4), 308-313.

Newton, P. (2010). Thinking about linking. *Measurement*, *8*(1), 38-56.

Newton, P. E. (2010). Conceptualizing comparability. *Measurement*, *8*(4), 172-179.

Neyman, J., Scott, E. L. (1948). Consistent Estimates Based on Partially Consistent Observations. *Econometrica*, *16*(1), 1-32.

日本テスト学会 (2007). テスト・スタンダード――日本のテストの将来に向けて――　金子書房

野口裕之・大隅敦子 (2014). テスティングの基礎理論　研究社

O'Neil, T. P. (2010). Maintenance of Vertical Scales Under Conditions of Item Parameter Drift and Rasch Model-data Misfit (doctoral Dissertation). University of Massachusetts - Amherst, Amherst, MA. Retrieved from http://scholarworks.umass.edu/open_access_dissertations/239

岡田謙介 (2015). 心理学と心理測定における信頼性について ――Cronbachのα係数とは何なのか, 何でないのか――　教育心理学年報, *54*, 71-83.

Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, *44*(4), 443-460.

Orlando, M., Thissen, D. (2000). Likelihood-Based Item-Fit Indices for Dichotomous Item Response Theory Models. *Applied Psychological Measurement*, *24*(1), 50-64.

大友賢二 (1996). 項目応答理論入門ー言語テスト・データの新しい分析法ー　大修館書店

Patz, R. J. (2007). Vertical Scaling in Standards-Based Educational Assessment and Accountability Systems. Retrieved from www.ccsso.org

Patz, R. J., Yao, L. (2007). Vertical Scaling: Statistical Models for Measuring Growth and Achievement. In C. R. Rao and S. Sinharay (Eds.), *Handbook of Statistics* (pp. 955-975). Elsevier.

R Core Team (2018). R: A Language and Environment for Statistical Computing. Vienna, Austria. Retrieved from https://www.r-project.org/

Ramsay, J. O. (1975). Solving implicit equations in psychometric data analysis. *Psychometrika*, *40*(3), 337-360.

Reckase, M. (2009). *Multidimensional item response theory*. New York: Springer.

Reckase, M. D. (2010). Study of Best Practices for Vertical Scaling and Standard Setting with Recommendations for FCAT 2.0 The Requirements for State Assessment Programs. Retrieved from http://www.fldoe.org/core/fileparse.php/5663/urlt/0086369-studybestpracticesverticalscalingstandardsetting.pdf

Rizopoulos, D. (2006). ltm: An R package for Latent Variable Modelling and Item Response Theory Analyses. *Journal of Statistical Software*, *17*(5), 1-25. Retrieved from http://www.jstatsoft.org/v17/i05/

斉田智里 (2014). 英語学力の経年変化に関する研究――項目応答理論を用いた事後的等化法による共通尺度化―― 風間書房

埼玉県教育委員会 (2016). 平成28年度埼玉県学力・学習状況調査報告書

埼玉県教育委員会 (2018). 埼玉県学力・学習状況調査 Retrieved January 4, 2019, from https://www.pref.saitama.lg.jp/f2214/gakutyou/20150605.html

坂本佑太朗 (2015). わが国のTIMSS2011数学データにおける多次元IRTを使った妥当性の検証について 日本テスト学会誌, *12*(1), 37-53.

佐藤喜一・村木英治 (2008). 垂直尺度. 池田央 (編) テスト作成ハンドブック (pp. 494-512). 株式会社教育測定研究所

佐藤喜一・柴山直 (2014). 対応づけ得点のための信頼性指標の提案 : 対応づけ可能性分析への応用 日本テスト学会誌, *10*(1), 69-80. Retrieved from https://ci.nii.ac.jp/naid/40020167548

Sato, Y., Shibayama, T. (2018). Linkability analysis focused on reliability of linked scores. *Paper presented at the annual meeting of the National Council on Measurement in Education*. New York, NY.

Savalei, V. (2006). Logistic Approximation to the Normal: The KL Rationale. *Psychometrika*, *71*(4), 763-767.

芝祐順 (1978). 語彙理解尺度作成の試み 東京大学教育学部紀要, *17*, 47-58.

芝祐順 (1981). 因子分析法 (第二版) 東京大学出版会

芝祐順・渡部洋・石塚智一 (1984). フィッシャー情報量. 統計用語辞典 新曜社

芝祐順・野口裕之・柴山直 (1986). 語彙理解力の発達に関する追跡的研究. 東京大学教育学部紀要, *25*, 27-40.

柴山直・繁桝算男 (1994). 一般項目反応モデルにおけるパラメタの推定方法. 行動計量学, *21*(1), 57-65. 日本行動計量学会.

柴山直・野口裕之 (2004). 「法科大学院統一適性試験」と大学入試センター「法科大学院適性試験」の得点対応づけ.日弁連法務研究財団 (編), 法科大学院ー適性試験テクニカル・レポート (p. 139). 商事法務

柴山直・佐藤喜一 (2008). 等パーセンタイル等化法における誤差評価の試み 東北大学大学院教育学研究科年報, *57*(1), 395-409.

柴山直・佐藤喜一・熊谷龍一・澁谷拓巳・板宮千尋・江尻大亮 (2018). 経年変化分析調査と対応づけによる本体調査の年度間比較の試み. Retrieved from http://www.mext.go.jp/a_menu/shotou/gakuryoku-chousa/1406895.htm

澁谷拓巳・柴山直 (2017). 学力の発達を追跡するための垂直尺度化について. 日本教育心理学会第59回総会発表論文集 (p. 751). Retrieved from https://confit.atlas.jp/guide/event-

img/edupsych2017/PH76/public/pdf?type=in
澁谷拓巳・柴山直 (2018). 学力テストの IRT 垂直尺度化に適したサンプル数と尺度調整法の検討. 日本テスト学会第16回大会発表論文抄録集 (pp. 94-95).
靜哲人 (2007). 基礎から深く理解するラッシュモデリング――項目応答理論とは似て非なる測定のパラダイム―― 関西大学出版部
孫媛・芝祐順 (1990). 特異な反応パタンを示す被験者の能力推定――一般項目反応理論の適用―― 教育心理学研究, *38*(4), 360-368.
孫媛 (1997). 多次元データに対する項目反応モデル 学術情報センター紀要, *9*, 103-111.
Stan Development Team (2018). Stan Modeling Language Users Guide and Reference Manual. Retrieved from http://mc-stan.org
Stevens, S. S. (1946). On the Theory of Scales of Measurement. *Science*, *103*, 677-680.
Stocking, M. L., Lord, F. M. (1983). Developing a common metric in Item Response Theory. *Applied Psychological Measurement*, *7*(3), 201-210.
Stout, W., Nandakumar, R., Habing, B. (1996). Analysis of latent dimensionality of dichotomously and polytomously scored test data. *Behaviormetrika*, *23*(1), 37-65.
Swaminathan, H., Gifford, J. A. (1986). Bayesian estimation in the three-parameter logistic model. *Psychometrika*, *51*(4), 589-601.
高橋登, 中村知靖 (2009). 適応型言語能力検査（ATLAN）の作成とその評価. 教育心理学研究, *57*(2), 201-211.
高橋登, 中村知靖 (2015). 漢字の書字に必要な能力──ATLAN書取り検査の開発から──. 心理学研究, *86*(3), 258-268.
Takane, Y., De Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, *52*(3), 393-408.
te Marvelde, J. M., Glas, C. A. W., van Landeghem, G., van Damme, J. (2010). Application of Multidimensional Item Response Theory Models to Longitudinal Data. *Applied Psychological Measurement*, *8*(2), 5-7.
Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, *34*(4), 273-286.
登藤直弥 (2012). 項目反応間の局所依存性が項目母数の推定に与える影響――項目母数の比較可能性を確保した上での検討―― 行動計量学, *39*(2), 81-91.
Topczewski, A. M. (2013). Effect of Violating Unidimensional Item Response Theory Vertical Scaling Assumptions on Developmental Score Scales (doctoral disserttion). University of Iowa, Iowa City, Iowa. Retrieved from http://ir.uiowa.edu/etd/4921http://ir.uiowa.edu/etdhttp://ir.uiowa.edu/etd/4921.
Torgerson, W. S. (1958). *Theory and methods of scaling*. New York: Wiley.
豊田秀樹 (1998). 共分散構造分析［入門編］――構造方程式モデリング―― 朝倉書店
豊田秀樹 (2005). 項目反応理論［理論編］ 朝倉書店

豊田秀樹 (2012). 項目反応理論［入門編］ 朝倉書店

豊田秀樹・岩間徳兼・中村彩子・齋藤康寛 (2015). 項目反応理論を用いたテスト運用への切り替えコスト軽減の試み――多数の潜在特性尺度の同時等化法を利用して―― 日本オペレーションズ・リサーチ学会和文論文誌, *58*, 122-147.

von Davier, A. A. (2010). What dictates the meaning of test linking? A reaction to “Thinking About Linking.” *Measurement*, *8*(4), 161-167.

Wang, S., Jiao, H. (2009). Construct equivalence across grades in a vertical scale for a K-12 large-scale reading assessment. *Educational and Psychological Measurement*, *69*(5), 760-777.

Wingersky, M. S., Barton, M. A., Lord, F. M. (1982). LOGIST user’s guide. Princeton, NJ.

Wright, B. D., Masters, D. N. (1982). *Rating Scale Analysis*. Chicago, Il: MESA Press.

Wright, B. D., Stone, M. H. (1979). *Best Test Design*. Chicago, Il: MESA Press.

柳井晴夫・前川眞一・繁桝算男・市川雅教 (1990). 因子分析――その理論と方法―― 朝倉書店

Yen, W. M. (1985). Increasing item complexity: A possible cause of scale shrinkage for unidimensional item response theory. *Psychometrika*, *50*(4), 399-410.

Yen, W. M. (1986). The choice of scale for educational measurement: An IRT parspective. *Journal of Educational Measurement*, *23*(4), 299-325.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, *30*(3), 187-213.

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement*, *5*(2), 245-262.

Yin, L. (2013). The Robustness of IRT-Based Vertical Scaling Method to Violation Unidimensionality (doctoral dissertation). University of Pittsburgh, Pittsburgh, PA.

Young, M. J., Tong, Y. (2016). Vertical Scales. In S. Lane, M. D. Raymond, and T. M. Haladyna (Eds.), *Handbook of Test Development* (2md ed., pp. 450-466). New York: Routledge.

Zimowski, M. F., Muraki, E., Mislevy, R. J., Bock, R. D. (2003). BILOG-MG 3. Skokie, IL: Scientific Software International, Inc.

付録 A　周辺最尤推定法のプログラムの妥当性検証

本研究で使用した項目パラメタ推定プログラムは筆者作成によるものである。ここではそのプログラム (**irtfun2**) によって得られた推定値の妥当性を検討するために，同種のプログラム **lazy.irtx** と Easy Estimation による分析結果との比較をおこなう。

1. プログラムの使用方法

はじめに本研究で用いたプログラムの概要と使用方法について説明する。プログラムの推定アルゴリズムは一般的な周辺最尤推定法であり，その推定アルゴリズムを C++で記述し，**Rcpp** パッケージによってそのアルゴリズムを R 上で実行している[2]。C++で記述している理由は R に比べて計算速度が圧倒的に速いためである。一般にベクトル処理においては R と C++の実行速度に大差はないが, for ループをともなうような計算を実行する場合には, その差は顕著になる。EM アルゴリズムの E ステップでは 2 重，3 重の for ループをともなう計算が必要になるため，C++で記述している。

R 上でこの関数を使用するためには，まず Rtools[3]と呼ばれる C++コンパイラと **Rcpp** という R パッケージをインストールする必要がある。詳しいインストール方法などは「みんなの Rcpp」などの Web ページ[4]の解説を参照されたい。**Rcpp** パッケージについては通常の R によるパッケージインストール方法と同じで CRAN から入手できる。C++コンパイラと **Rcpp** の他に，あと 2 つのパッケージが必要となる。一つ目は **devtools** と呼ばれる R パッケージ作成支援パッケージである。さらに，このパッケージが提供する **install_github** 関数を通して GitHub 上にある筆者自作のパッケージ **irtfun2** をインストールすることで，本研究で使用したプログラムが実行可能になる。インストールの際には，必要な依存パッケージ（たとえば **stringr** や **magrittr** など）も自動的にダウンロードされるほか，C++のコードをコンパイルするため，インストールには数分程度の時間がかかる。

R シンタックス

```
install.packages(“Rcpp”) # Rcpp のインストール
install.packages(“devtools”) # devtools のインストール
devtools::install_github(“takuizum/irtfun2”) # irtfun2 のインストール
```

irtfun2 パッケージには項目パラメタ推定関数 **estip** と **estip2** のほか，能力パラメタ推定関数 **estheta** や項目適合度計算関数 **ifind**, **ifind3** や一般項目反応モデルのパラメタ推定関数

[2] バージョン 3.4.3 以上が必要である。

[3] Mac の場合には Xcode command line tools である。

[4] https://teuder.github.io/rcpp4everyone_ja/

estGip, 復元得点分布 (observed score distribution) 推定関数 **obscore_dist** などが含まれている。IRT の項目パラメタ推定関数は **estip** と **estip2** である。**estip2** は通常の周辺最尤推定法のほかに，周辺ベイズ推定法や正則化周辺最尤推定法 (Regularized MML) が実装されており，さらに M ステップの最適化手法として **optim** 関数により提供されている準ニュートン法などを使用できるが，計算速度は **estip** にやや劣る。

本研究で使用した関数は **estip** の方である。**estip** では項目パラメタと－2×周辺対数尤度による収束基準判定をおこなっているが，収束基準はユーザーの任意の値を設定できる。また，項目反応パタンを含む R 上のデータフレームの列名を指定することで，任意の項目を推定から除外することも可能である。以下に関数の使用方法の例を示す。いま，一行目に受検者の ID，二行目に受検者の集団についての番号，三行目以降に項目反応データが含まれているデータフレーム **sim_dat_st** を使用する。このデータは乱数によって発生させた項目反応データであり，**irtfun2** パッケージに含まれている。

```
library(irtfun2)
head(sim_dat_st)
# item response data
# must be data.frame
res <- estip(sim_dat_st,
             fc = 3, # 項目反応データの始まりの列数
             gc = 2, # 母集団変数の列数
             ng = 5, # 母集団数
             eEM = 1e-4, # 項目パラメタの変化率による収束基準の値
             eMLL = 1e-6 # 周辺対数尤度の変化率による収束基準の値
             )
```

2. estip の推定結果の妥当性検証

estip の妥当性検証のために乱数により発生させた項目反応データを **lazy.irtx** の **uIRT** 関数と Easy Estimation でも推定し，その推定結果を比較する。シミュレーションデータの発生方法と Easy Estimation で推定するためのデータの書き出し方法を以下に示す。シミュレーションデータは，垂直尺度構成で使用したデータ構造を参考に，5 学年の異なる能力水準の集団を対象とした尺度化テストデザイン構造としている。

```
library(tidyverse) # dplyr と magrittr を使うためのパッケージ

# ICC of 2PLM
ptheta2 <- function(theta, a, b){
  # IRT 2 PLM response probability
  D <- 1.702
  1/(1+exp(-D*a*(theta-b)))
}

# 反応確率と一様乱数から 01 データを発生させる関数。
resfunc2 <- function(prob){
  prob <- matrix(prob, ncol = 1)
  subfunc <- function(prob){
    if(prob < runif(1)) res <- 0
    else res <- 1
```

```
    return(res)
  }
  res <- apply(prob, 1, subfunc)
  return(res)
}

# 母集団のパラメタ（平均, SD)
thetaMS <- matrix(c(0.0,1.0,
                    0.4,0.9,
                    0.8,0.8,
                    1.2,0.7,
                    1.6,0.6),
                  ncol = 2, byrow = T)

# grade ID of items
gradeID <- c("A","B","C","D","E")

# true item parameter matrix
true_para <- matrix(nrow = 30*7/3, ncol = 2)

jj <- 1

# fix the seed
set.seed(0204)
# grade item parameter
for(g in 1:5){

  if(g == 1){ # 第一学年だけ，項目数が異なる（基準となるので）
    # item parameters(exept scaling test item)
    for(j in 1:(30*2/3)){
      true_para[jj,1] <- rlnorm(1, -0.5, 0.3)
      true_para[jj,2] <- rnorm(1, thetaMS[g,1], thetaMS[g,2])
      jj <- jj + 1 # 挿入行数カウント
    }
  } else {
    for(j in 1:(30*1/3)){
      true_para[jj,1] <- rlnorm(1, -0.5, 0.3)
      true_para[jj,2] <- rnorm(1, thetaMS[g,1], thetaMS[g,2])
      jj <- jj + 1 # 挿入行数カウント
    }
  }
}

# scaling test item parameter
for(g in 1:(30*1/3)){
  true_para[jj,1] <- rlnorm(1, -1, 0.5)
  true_para[jj,2] <- runif(1, -3, 3)
  jj <- jj + 1
}
true_para <- data.frame(V3 = true_para[,1], V4 = true_para[,2])

rm(jj)

al <- c("a","b","c","d","e","f","st")
```

```
c <- 2
for(i in c("a","b","c","d","e")){

  num <- formatC(c(1:(30*1/3)), width = 3, flag = 0)

  if(i =="a"){
    itemn1 <- apply(matrix(i, ncol = 1), 1, paste0, num)
    itemn2 <- apply(matrix(al[c], ncol = 1), 1, paste0, num)
    itemn3 <- apply(matrix(al[7], ncol = 1), 1, paste0, num)
    itemID <- rbind(itemn1,itemn2,itemn3)
  }

  if(i !="a") {
    itemn1 <- apply(matrix(i, ncol = 1), 1, paste0, num)
    itemn2 <- apply(matrix(al[c], ncol = 1), 1, paste0, num)
    itemn3 <- apply(matrix(al[7], ncol = 1), 1, paste0, num)
    itemn <- rbind(itemn1,itemn2,itemn3)
    itemID <- cbind(itemID,itemn)
    rm(itemn1,itemn2,itemn3,itemn,num)
  }
  c <- c + 1
} # end of i

rm(al,c)


for(g in 1:5){

  cat("grade ",g,".¥n")
  # item parameters(exept scaling test item)

  if( g == 1){
    gradeitem <- c(seq.int(1,length.out = 30*2/3), seq.int(30*2 + 1, length.out = 30*1/3))
  } else {
    gradeitem <- c(seq.int(g*30*1/3+1,length.out = 30*2/3), seq.int(30*2 + 1, length.out =
30*1/3))
  }

  a <-  true_para[gradeitem, 1]
  b <-  true_para[gradeitem, 2]

  # ability parameter
  theta <- rnorm(1000, thetaMS[g,1], thetaMS[g,2])

  # generate response patterns exept scaling test item
  resp <- theta %>%
    matrix(ncol = 1) %>%
    apply(1,ptheta2, a = a, b = b) %>%
    apply(2,resfunc2) %>%
    t() %>%
    as.data.frame()

  colnames(resp) <- itemID[,g]
```

```
  grade <- rep(g, 1000) %>% as.numeric() # グループIDはnumeric型であること。

  ID <- apply(matrix(gradeID[g], ncol = 1), 1, paste0,
              formatC(c(1:1000), width = 5, flag = 0)) %>%
    as.character()

  resp <- cbind(ID, grade, resp)

  if(g == 1) {
    RESP <- resp
  } else {
    # combine response data for concurrent calibration
    suppressMessages(
      suppressWarnings(
        RESP <- RESP %>% dplyr::full_join(resp)
      )
    )
  }

} # end of one grade
rm(a,b,g,grade,gradeID,gradeitem,i,ID,j,theta,itemID,resp)

write.table(RESP, file="vald_data.dat", quote = F, sep = "", col.names = F, row.names = F,
na="N")
true_para <- data.frame(a=true_para$V3, b=true_para$V4)
```

次に項目反応データを使って項目パラメタを推定し，それらの結果を比較する。**lazy.irtx** のインストール方法と使用方法は光永（2017）[5]などを，Easy Estimation のインストール方法と使用方法は作者である東北大学准教授熊谷龍一氏のサイト[6]を参照されたい。

得られた推定値の表とプロットを次のページ以降に示す。

[5] 光永悠彦 (2017). テストは何を測るのか　ナカニシヤ出版

[6] http://irtanalysis.main.jp/

表 A.1 識別力の真値と推定値の表

Item	TRUE	irtfun2	lazy	Easy	Item	TRUE	irtfun2	lazy	Easy
a001	0.57110	0.43136	0.43186	0.43078	c006	0.68637	0.73476	0.73533	0.73465
a002	0.75217	0.80731	0.80828	0.82369	c007	0.59359	0.50958	0.50996	0.51056
a003	0.50608	0.61615	0.61688	0.60287	c008	0.79283	0.78567	0.78627	0.79082
a004	0.30042	0.33877	0.33917	0.34236	c009	0.39928	0.51928	0.51966	0.52195
a005	0.59215	0.63663	0.63738	0.64400	c010	0.38294	0.40695	0.40724	0.40527
a006	0.67182	0.60914	0.60987	0.60552	d001	0.76029	0.68154	0.68165	0.67555
a007	0.41616	0.40074	0.40123	0.39850	d002	0.51440	0.47222	0.47229	0.47095
a008	0.54550	0.60524	0.60596	0.61312	d003	0.60799	0.55946	0.55957	0.55571
a009	0.75604	0.80889	0.80986	0.83718	d004	0.53879	0.48299	0.48305	0.48082
a010	0.53350	0.48348	0.48405	0.49593	d005	0.40833	0.41348	0.41355	0.41275
b001	0.47937	0.27146	0.27170	0.26650	d006	0.46037	0.40069	0.40076	0.40232
b002	0.62499	0.40424	0.40463	0.39966	d007	0.54778	0.55898	0.55909	0.55468
b003	1.19762	1.25095	1.25249	1.31232	d008	0.57737	0.59311	0.59324	0.59296
b004	0.61544	0.62253	0.62319	0.62324	d009	0.49237	0.52373	0.52380	0.51974
b005	0.54452	0.49592	0.49646	0.49686	d010	0.66815	0.70798	0.70813	0.70401
b006	0.77039	0.77629	0.77707	0.77405	e001	0.79879	0.75954	0.75922	0.74650
b007	0.76118	0.32793	0.32823	0.32404	e002	0.95506	0.89667	0.89620	0.88645
b008	0.67947	0.36391	0.36423	0.36478	e003	0.45755	0.36621	0.36606	0.36275
b009	0.67313	0.58114	0.58184	0.59354	e004	0.97031	0.80809	0.80774	0.78526
b010	0.61856	0.53851	0.53907	0.53914	e005	0.45229	0.36691	0.36679	0.36272
st001	0.64509	0.23342	0.23357	0.23530	e006	0.46203	0.42675	0.42658	0.42540
st002	0.50889	0.37609	0.37639	0.37465	e007	0.60234	0.55629	0.55609	0.55579
st003	1.17466	0.30497	0.30517	0.30352	e008	1.00740	0.90718	0.90680	0.89846
st004	0.72128	0.27555	0.27571	0.27640	e009	0.90365	0.80219	0.80200	0.80533
st005	0.66328	0.40820	0.40845	0.40609	e010	0.62134	0.55106	0.55081	0.55333
st006	0.84430	0.14236	0.14247	0.14152	f001	0.23494	0.18317	0.18301	0.17519
st007	0.43605	0.19257	0.19267	0.19252	f002	0.38535	0.37246	0.37213	0.36943
st008	0.43117	0.69491	0.69545	0.69457	f003	0.30306	0.21335	0.21314	0.20404
st009	0.53589	0.50508	0.50523	0.50538	f004	0.32985	0.40939	0.40902	0.40210
st010	0.49508	0.29922	0.29938	0.29649	f005	0.44316	0.39649	0.39618	0.38566
c001	0.68096	0.65020	0.65070	0.65367	f006	0.14956	0.29215	0.29194	0.28194
c002	0.58132	0.57761	0.57805	0.57601	f007	0.19215	0.20902	0.20885	0.20254
c003	0.48104	0.46438	0.46475	0.46902	f008	0.68532	0.57812	0.57762	0.55152
c004	0.38624	0.36947	0.36976	0.36989	f009	0.52815	0.47904	0.47863	0.48551
c005	0.41987	0.38697	0.38726	0.38539	f010	0.32488	0.30938	0.30910	0.29360

表 A.2 困難度の真値と推定値の表

Item	TRUE	irtfun2	lazy	Easy	Item	TRUE	irtfun2	lazy	Easy
a001	-3.31339	-3.87102	-3.87078	-3.66743	c006	0.05057	0.00023	0.00058	0.2068
a002	0.27653	0.27540	0.27571	0.46729	c007	0.20895	0.18012	0.18055	0.38563
a003	0.82248	0.79145	0.79176	1.00152	c008	0.12196	0.15117	0.15158	0.35997
a004	-0.24053	-0.35262	-0.35230	-0.14631	c009	1.91981	1.69371	1.69483	1.89295
a005	0.02341	0.01215	0.01247	0.21120	c010	0.11816	0.11764	0.11804	0.32006
a006	0.68592	0.66424	0.66455	0.86490	d001	1.71959	1.74918	1.75073	1.95995
a007	1.32931	1.40500	1.40527	1.61402	d002	1.06067	1.05032	1.05116	1.25482
a008	-0.91125	-0.94721	-0.94688	-0.73358	d003	1.76431	1.7606	1.76215	1.97011
a009	-1.08191	-1.12912	-1.12879	-0.89648	d004	2.6388	2.61635	2.61884	2.82824
a010	-1.12014	-1.09183	-1.09151	-0.86610	d005	1.16226	1.11233	1.11324	1.31697
b001	-1.45763	-0.39020	-0.38994	-0.19748	d006	1.87195	1.87843	1.88011	2.08005
b002	-0.83324	-0.41596	-0.41566	-0.21868	d007	0.5066	0.49572	0.49601	0.69644
b003	0.68942	0.21839	0.21882	0.42455	d008	0.80076	0.82021	0.82083	1.02395
b004	0.72446	0.73562	0.73612	0.93735	d009	1.24296	1.1543	1.15525	1.35982
b005	-0.85790	-0.98876	-0.98845	-0.78310	d010	1.44742	1.41826	1.41945	1.62481
b006	0.78362	0.88921	0.88975	1.09112	e001	1.40895	1.36343	1.36447	1.56842
b007	-1.58082	-1.58790	-1.58792	-1.40410	e002	2.23768	2.24474	2.24725	2.4638
b008	-0.17087	0.47697	0.47748	0.67942	e003	1.53508	1.53455	1.53588	1.7413
b009	1.40282	0.49236	0.49280	0.69089	e004	2.01177	2.10402	2.10625	2.32882
b010	0.67293	0.75854	0.75904	0.96059	e005	2.35503	2.45463	2.45733	2.67193
st001	0.96816	0.09752	0.09787	0.30665	e006	1.22265	1.20509	1.20589	1.4098
st002	0.26064	-2.06797	-2.06842	-1.87407	e007	1.4067	1.4721	1.47332	1.67805
st003	-0.13012	-0.65190	-0.65197	-0.45440	e008	2.18931	2.24647	2.24893	2.4647
st004	0.77642	1.37237	1.37343	1.57493	e009	0.14343	-0.01241	-0.01333	0.19169
st005	-0.71288	0.29390	0.29431	0.49539	e010	2.91187	3.14128	3.14523	3.34292
st006	1.03416	-2.13281	-2.13340	-1.94589	f001	0.22183	-0.28593	-0.2885	-0.16568
st007	-0.20217	2.70044	2.70243	2.90502	f002	-1.99142	-2.07696	-2.08321	-1.90207
st008	1.34673	-0.59814	-0.59809	-0.39467	f003	-0.63466	-1.45747	-1.4627	-1.38944
st009	-0.26865	2.73150	2.73385	2.93626	f004	1.27129	1.39682	1.3977	1.59936
st010	0.83316	1.64839	1.64967	1.85952	f005	0.35481	0.21458	0.21315	0.38278
c001	1.23376	1.21116	1.21202	1.41319	f006	-2.16638	-0.29316	-0.29545	-0.15443
c002	-0.42279	-0.34889	-0.34869	-0.14582	f007	2.5924	2.73798	2.74157	2.97887
c003	1.23665	1.18544	1.18628	1.38477	f008	-0.5779	-0.89118	-0.89492	-0.7957
c004	1.17268	1.18667	1.18753	1.38983	f009	2.72081	2.8603	2.86417	3.0538
c005	0.68932	0.56362	0.56422	0.76751	f010	1.48851	1.42428	1.42522	1.62026

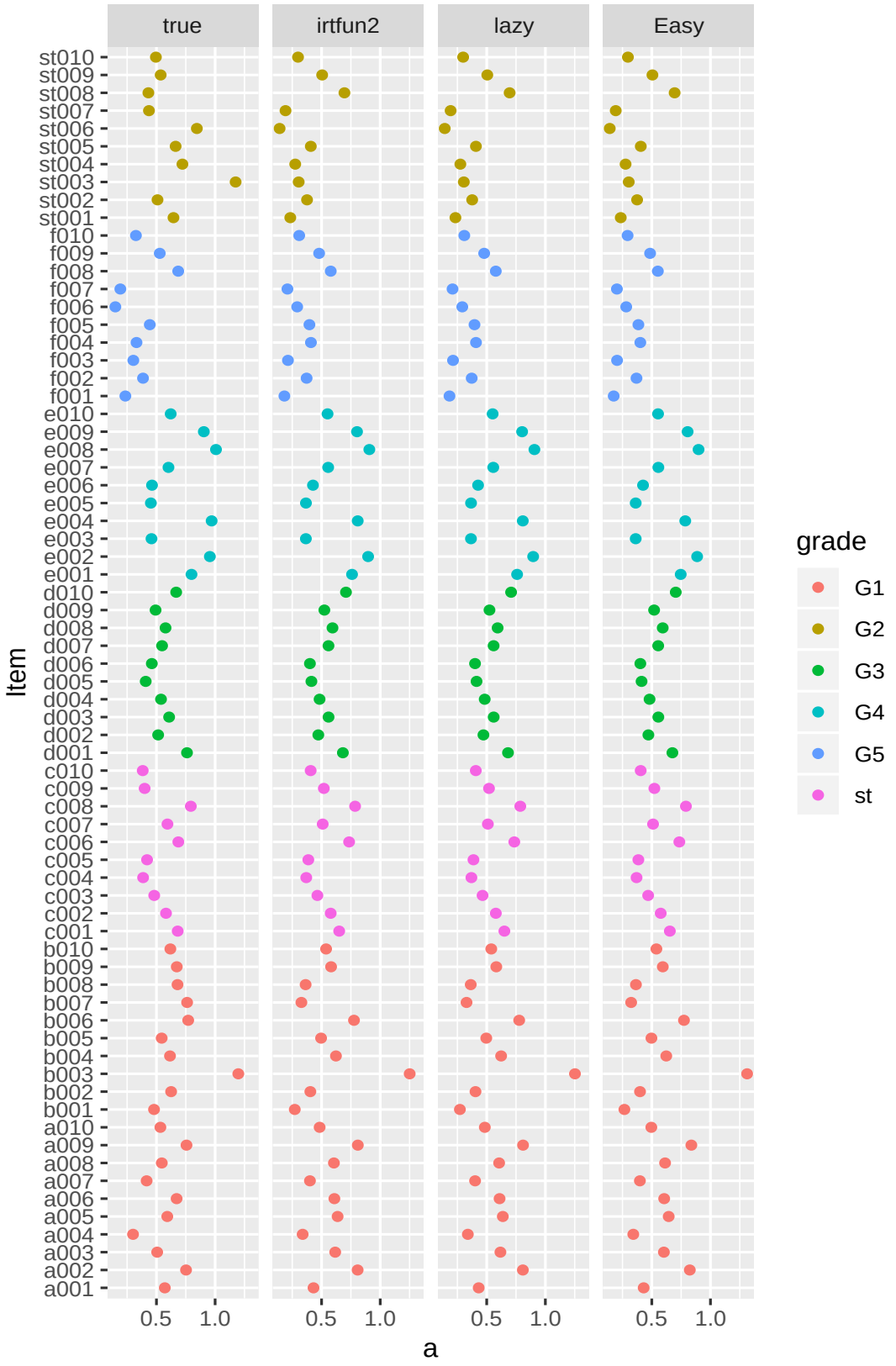

図 A. 1 識別力の真値と推定値のプロット

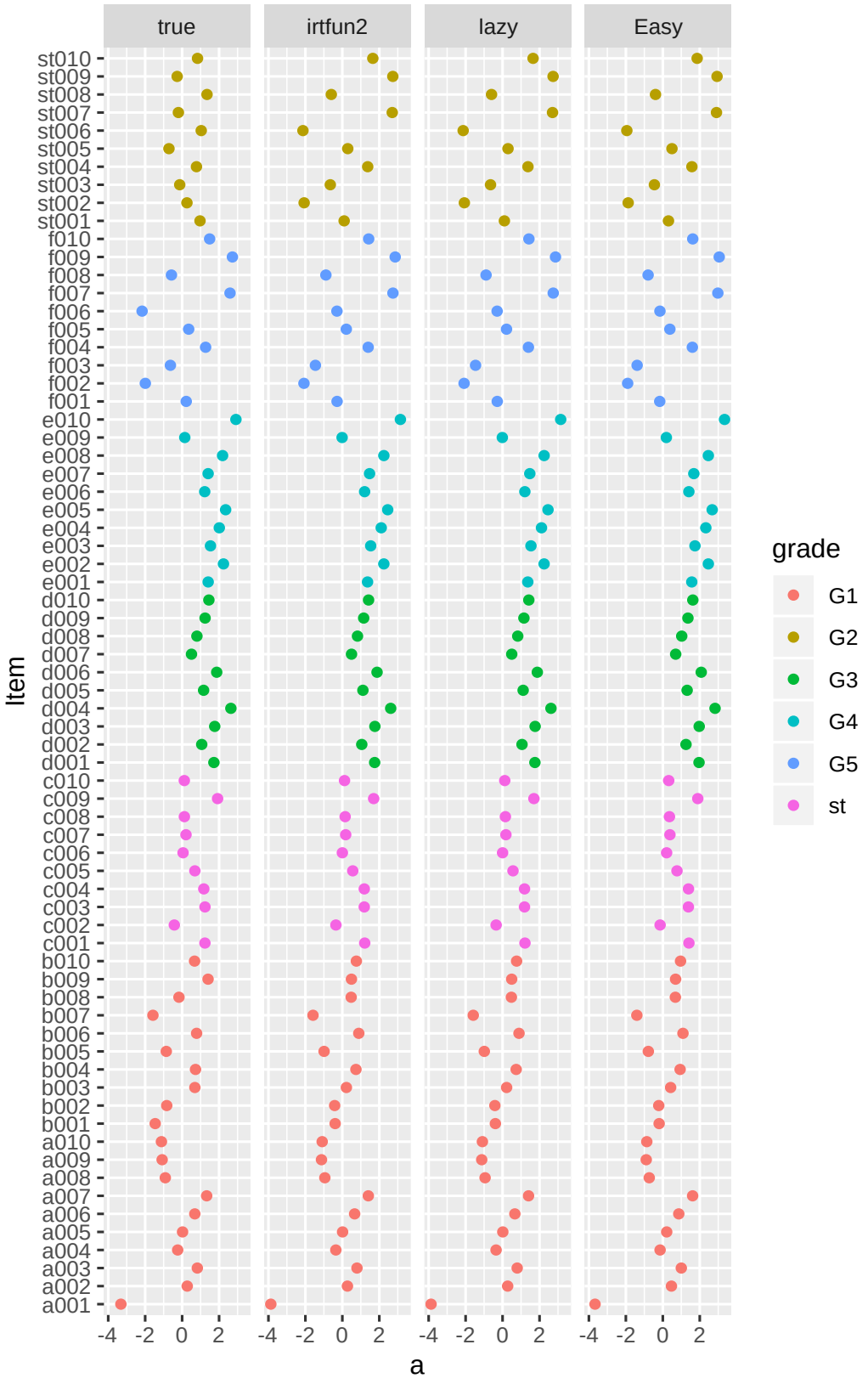

図 A.2 困難度の真値と推定値のプロット

三者の推定値は小数点第 3 位あたりまで概ね一致しているため，本研究で使用した **estip** 関数の推定結果に大きな問題はないと言える。いくつかの項目で Eazy Estimation の推定結果だけ **irtfun2** と **lazy.irtx** の推定結果とわずかにずれることがあるが，これは E ステップのθの事前分布に使用している分布が受検者の期待度数に基づく多項分布か，正規分布かの違いによるものであると考えられる。**irtfun2** の推定関数である **estip** 関数のデフォルト設定と **lazy.irtx** の **uIRT** 関数は正規分布を使用しているため，推定結果がかなり似ている。真値との関係を見る限りでは分布の違いに優劣はないようである。ちなみに **estip** 関数では引数に **thdist=”empirical”** を与えることで，多項分布を使用した推定をおこなうこともできる。

付録Ｂ　シミュレーション研究のＲシンタックス

```
#  尺度化テストデザイン

install.packages(c("pracma","dplyr"))
# lazy.irtx の簡易インストール用 URL
install.packages("http://mayekawa.in.coocan.jp/Rpackages/lazy.irtx_1.0.1.tar.gz",
repos=NULL)

library(pracma)
library(tidyverse)
library(lazy.irtx)
library(irtfun2)

simN <- 100

gradeN <- 5

# 母集団のパラメタ（平均, SD)
thetaMS <- matrix(c(-0.8,1.0,
                    -0.4,1.0,
                    0,1.0,
                    0.4,1.0,
                    0.8,1.0),
                  ncol = 2, byrow = T)

# set output directory
setwd("YOUR OWN DIR")

# starts simulation
# 計算過程のほとんどは sink 関数で txt 形式で出力されます。

simN <- 100
for(sampleN in c(400,1000,10000)){

  for(itemN in c(15,30,60)){
    sink(file = paste0("sampleN",sampleN,"_itemN",itemN,"_SimulationResult.txt"))

    message("sample number is ",sampleN)
    message("item number is ", itemN)

    # まずは項目パラメタの用意
    # prepare item names
    al <- c("a","b","c","d","e","f","st")
    c <- 2
    for(i in c("a","b","c","d","e")){

      num <- formatC(c(1:(itemN*1/3)), width = 3, flag = 0)
      if(i =="a"){
        itemn1 <- apply(matrix(i, ncol = 1), 1, paste0, num)
        itemn2 <- apply(matrix(al[c], ncol = 1), 1, paste0, num)
        itemn3 <- apply(matrix(al[7], ncol = 1), 1, paste0, num)
```

```
      itemID <- rbind(itemn1,itemn2,itemn3)
    }
    if(i !="a") {
      itemn1 <- apply(matrix(i, ncol = 1), 1, paste0, num)
      itemn2 <- apply(matrix(al[c], ncol = 1), 1, paste0, num)
      itemn3 <- apply(matrix(al[7], ncol = 1), 1, paste0, num)
      itemn <- rbind(itemn1,itemn2,itemn3)
      itemID <- cbind(itemID,itemn)
      rm(itemn1,itemn2,itemn3,itemn,num)
    }
    c <- c + 1
  } # end of i

  rm(al,c)

  # true equating coefficient
  true_ec <- matrix(NA,ncol=2, nrow=4)

  for(i in 1:4){
    true_ec[i,1] <- thetaMS[i+1,2]/thetaMS[i,2] # A
    true_ec[i,2] <- thetaMS[i+1,1] - true_ec[i,1]*thetaMS[i,1] # K
  } # end of i


  # 誤差推定値結果代入用
  ParaError_sep <- ParaError_calr<- matrix(nrow = simN, ncol = 12)
  ParaError_con <- matrix(nrow = simN, ncol = 12)

  colnames(ParaError_sep) <- c("MEA-a", "MAE-b", "MSE-a", "MSE-b", "RMSE-a", "RMSE-b",
"MAPE-a", "MAPE-b", "NMSE-a", "NMSE-b", "rSTD-a", "rSTD-b")
  colnames(ParaError_con) <- c("MEA-a", "MAE-b", "MSE-a", "MSE-b", "RMSE-a", "RMSE-b",
"MAPE-a", "MAPE-b", "NMSE-a", "NMSE-b", "rSTD-a", "rSTD-b")
  colnames(ParaError_calr) <- c("MEA-a", "MAE-b", "MSE-a", "MSE-b", "RMSE-a", "RMSE-b",
"MAPE-a", "MAPE-b", "NMSE-a", "NMSE-b", "rSTD-a", "rSTD-b")

  # DICC計算結果代入用
  DICC_sep  <- matrix(ncol = 1,nrow = 100)
  DICC_con  <- matrix(ncol = 1,nrow = 100)
  DICC_calr <- matrix(ncol = 1,nrow = 100)

  # 母集団平均・標準偏差代入用
  pdist_mean_sep <- matrix(nrow=100,ncol=5)
  pdist_mean_con <- matrix(nrow=100,ncol=5)
  pdist_mean_calr <- matrix(nrow=100,ncol=5)
  pdist_sd_sep <- matrix(nrow=100,ncol=5)
  pdist_sd_con <- matrix(nrow=100,ncol=5)
  pdist_sd_calr <- matrix(nrow=100,ncol=5)

  t <- 0
  system.time(

    while(t != simN){

      t <- t + 1
```

```
cat("starts ",t," times simulation¥n")
message("starts ",t," times simulation")
cat("generate dichotomous response data.¥n")

# grade ID
gradeID <- c("A","B","C","D","E")
# the matrix to insert simulation equating coefficients
sim_ec <- matrix(NA, ncol = 2, nrow = 4)
# true item parameter matrix
true_para <- matrix(nrow = itemN*7/3, ncol = 2)

jj <- 1
# grade item parameter
for(g in 1:gradeN){
  if(g == 1){ # 第一学年だけ，項目数が異なる（基準となるので）
    # item parameters(exept scaling test item)
    for(j in 1:(itemN*2/3)){
      true_para[jj,1] <- rlnorm(1, -0.5, 0.3)
      true_para[jj,2] <- rnorm(1, thetaMS[g,1], thetaMS[g,2])
      jj <- jj + 1 # 挿入行数カウント
    }
  } else {
    for(j in 1:(itemN*1/3)){
      true_para[jj,1] <- rlnorm(1, -0.5, 0.3)
      true_para[jj,2] <- rnorm(1, thetaMS[g,1], thetaMS[g,2])
      jj <- jj + 1 # 挿入行数カウント
    }
  }
}

# scaling test item parameter
for(g in 1:(itemN*1/3)){
  true_para[jj,1] <- rlnorm(1, -0.5, 0.3) # 学年独自項目と同じ値を設定
  true_para[jj,2] <- rnorm(1, 0, 1.5)
  jj <- jj + 1
}

true_para <- data.frame(V3 = true_para[,1], V4 = true_para[,2])

rm(jj)

#-----------------------------------------------------------------------------#
# generate simulation data & separate calibration
#-----------------------------------------------------------------------------#

cat("separate parameter estimation.¥n")
gd <- 0
grade_dummy <- c(3,2,4,1,5)
while(gd < gradeN){

  gd <- gd + 1
  g <- grade_dummy[gd]
  cat("grade ",g,".¥n")
  # item parameters(exept scaling test item)
```

```
          # エラーチェック用
          res_sep <- NULL

          if( g == 1){
            gradeitem <- c(seq.int(1,length.out = itemN*2/3), seq.int(itemN*2 + 1, length.out
= itemN*1/3))
          } else {
            gradeitem <- c(seq.int((g-1)*itemN*1/3+1,length.out = itemN*2/3), seq.int(itemN*2
+ 1, length.out = itemN*1/3))
          }

          a <-  true_para[gradeitem, 1]
          b <-  true_para[gradeitem, 2]

          # ability parameter
          theta <- rnorm(sampleN, thetaMS[g,1], thetaMS[g,2])

          # generate response patterns exept scaling test item
          resp <- sim_gen(theta = theta, a = a, b = b)[,-1]

          if(any(colMeans(resp,na.rm=T) == 1)){ # もしも全問正解の項目が含まれていた場合
            message("detected item that all subject response correctly in grade", g)
            gd <- gd -1
            next
          }

          if(any(colMeans(resp,na.rm=T) == 0)){ # もしも全問不正解の項目が含まれていた場合
            message("detected item that all subject response incorrectly in grade", g)
            gd <- gd -1
            next
          }

          colnames(resp) <- itemID[,g]
          grade <- rep(g, sampleN) %>% as.numeric() # グループ ID は numeric 型であること。
          ID <- apply(matrix(gradeID[g], ncol = 1), 1, paste0,
                      formatC(c(1:sampleN), width = 5, flag = 0)) %>%
            as.character()
          resp <- cbind(ID, grade, resp)

          try(res_sep <- estip(resp, fc = 3,max = 4, min = -4 , maxiter_em = 100,
                               min_a = 0, maxabs_b = 20, EM_dist = 0, Bayes = 0, print = 0))

          #if(res_sep$converged == 0 || is.null(res_sep) || class(res_sep) == "try-error"){
          if(is.null(res_sep) || class(res_sep) == "try-error"){

            message("Error in ",t," times ",g," grade item parameter estimation")
            t <- t -1 # 途中で計算に失敗した場合には, カウントを減らす。
            break
          }

          sim_a <- res_sep$para$a
          sim_b <- res_sep$para$b
```

```
          cat("separate calibration: method SL.¥n")
          if(g == 3) {
            RESP <- resp
            paraT <- data.frame(V1 = itemID[,g], V3 = sim_a, V4 = sim_b, stringsAsFactors =
F)
            # for calr
            para1 <- data.frame(giid = itemID[,g], gfid = rep(g, itemN), type = rep("B2",
itemN), ncat = rep(2, itemN),
                                p1 = sim_a, p2 = sim_b, p3 = rep(0, itemN))
          } else {
            # combine response data for concurrent calibration
            suppressMessages(
              suppressWarnings(
                RESP <- RESP %>% dplyr::full_join(resp)
              )
            )

            # equating
            paraF <- data.frame(V1 = itemID[,g], V3 = sim_a, V4 = sim_b, stringsAsFactors =
F)

            # for calr
            para2 <- data.frame(giid = itemID[,g], gfid = rep(g, itemN), type = rep("B2",
itemN), ncat = rep(2, itemN),
                                p1 = sim_a, p2 = sim_b, p3 = rep(0, itemN))

            res_eq <- CEquating(paraT, paraF, method = "SL", output = F, Change = 1, Easy =
T)
            #  共通項目は等化後と等化先の項目パラメタの平均に置き換え（識別力は幾何平均）

            para1 <- rbind(para1,para2)

            if(g == 2){
              para_sep <- paraT[!paraT$V1 %in% res_eq$para$V1,]
              para_sep <- rbind(para_sep,res_eq$para)
            } else {
              paraT <- res_eq$para[!res_eq$para$V1 %in% para_sep$V1,]
              para_sep <- rbind(para_sep,paraT)
            }

            paraT <- res_eq$para

            sim_ec[g-1,] <- res_eq$EquatingCoefficient_A_K

          } # end of one grade


        } # end of separate calibration

        if( is.null(res_sep) || class(res_sep) == "try-error"){
          next
        }
```

```
        cat("separate calibration; method calr.¥n")
        try(res_calr <- calr(para1, baseform = 1, nsubj = rep(sampleN,gradeN), maxiter = 200,
npoints = 31,
                            thmin = -4, thmax = 4, print = 0))


        if(class(res_calr) == "try-error"){
          t <- t -1 # 途中で計算に失敗した場合には，カウントを減らす。
          break
        }

        para_sep <- dplyr::arrange(para_sep,V1)

        rm(para1,para2)
        calr_ec <- res_calr$qr[-1,]
        para_calr <- dplyr::arrange(res_calr$param,giid)

        # 独立尺度調整法の推定誤差の確認
        # pracma の関数で，RMSE などの値が計算できる。詳しくは?rmserr

        # a para
        er <- rmserr(true_para$V3, para_sep$V3)
        ParaError_sep[t,1] <- er$mae
        ParaError_sep[t,3] <- er$mse
        ParaError_sep[t,5] <- er$rmse
        ParaError_sep[t,7] <- er$mape
        ParaError_sep[t,9] <- er$nmse
        ParaError_sep[t,11] <- er$rstd

        # b para
        er <- rmserr(true_para$V4, para_sep$V4)
        ParaError_sep[t,2] <- er$mae
        ParaError_sep[t,4] <- er$mse
        ParaError_sep[t,6] <- er$rmse
        ParaError_sep[t,8] <- er$mape
        ParaError_sep[t,10] <- er$nmse
        ParaError_sep[t,12] <- er$rstd

        # a para
        er <- rmserr(true_para$V3, para_calr$p1)
        ParaError_calr[t,1] <- er$mae
        ParaError_calr[t,3] <- er$mse
        ParaError_calr[t,5] <- er$rmse
        ParaError_calr[t,7] <- er$mape
        ParaError_calr[t,9] <- er$nmse
        ParaError_calr[t,11] <- er$rstd

        # b para
        er <- rmserr(true_para$V4, para_calr$p2)
        ParaError_calr[t,2] <- er$mae
        ParaError_calr[t,4] <- er$mse
        ParaError_calr[t,6] <- er$rmse
        ParaError_calr[t,8] <- er$mape
        ParaError_calr[t,10] <- er$nmse
```

```
        ParaError_calr[t,12] <- er$rstd

        # starts IRT analysis

        cat("concurrent calibraiton.¥n")

        # エラーチェック用
        res_con <- NULL

        try(res_con <- estip(RESP, ng = 5, gc = 2, fc = 3, bg=3, maxiter_em = 100,
                             min_a = 0, maxabs_b = 20, EM_dist = 1, Bayes = 0, print = 0))

        if(class(res_con) == "try-error" || is.null(res_con)) {
          message("Error in ",t," times concurrent caribration")
          t <- t - 1
          next
        }

        #MML_EM(RESP)
        res_con$para <- res_con$para %>% dplyr::arrange(Item)

        # a para
        er <- rmserr(true_para$V3, res_con$para$a)
        ParaError_con[t,1] <- er$mae
        ParaError_con[t,3] <- er$mse
        ParaError_con[t,5] <- er$rmse
        ParaError_con[t,7] <- er$mape
        ParaError_con[t,9] <- er$nmse
        ParaError_con[t,11] <- er$rstd

        # a para
        er <- rmserr(true_para$V4, res_con$para$b)
        ParaError_con[t,2] <- er$mae
        ParaError_con[t,4] <- er$mse
        ParaError_con[t,6] <- er$rmse
        ParaError_con[t,8] <- er$mape
        ParaError_con[t,10] <- er$nmse
        ParaError_con[t,12] <- er$rstd

        #DICC
        dicc_f <- function(a1,b1,a2,b2,theta,w){
          sum(abs(1/(1+exp(-1.702*a2*(theta-b2)))*w - 1/(1+exp(-1.702*a1*(theta-b1)))*w))
        }

        cat("calculation estimated theta population distribution")
        para_sep$c <- 0

        dist_sep      <-      MML_EM_dist(RESP,data.frame(a=para_sep$V3,       b=para_sep$V4,
c=rep(0,nrow(para_sep)))[match(colnames(RESP)[c(-1,-2)],para_sep$V1),],fc=3,gc=2,ng=5)
        dist_calr     <-     MML_EM_dist(RESP,data.frame(a=para_calr$p1,      b=para_calr$p2,
c=para_calr$p3)[match(colnames(RESP)[c(-1,-2)],para_calr$giid),],fc=3,gc=2,ng=5)
        dist_con <- res_con$population_dist

        cat("calculation DICC_est")
```

```
        true_para$V1 <- para_sep$V1
        temp1 <- temp2 <- temp3 <- 0
        for(g in 1:5){
          d1 <- true_para[match(itemID[,g],true_para$V1),c("V3","V4")]
          d2 <- para_calr[match(itemID[,g],para_calr$giid),c("p1","p2")]
          d3 <- para_sep[match(itemID[,g],para_sep$V1),c("V3","V4")]
          d4 <- res_con$para[match(itemID[,g],res_con$para$Item),c("a","b")]
          temp1                          <-                          temp1                          +
mean(mapply(FUN=dicc_f,a1=d1$V3,b1=d1$V4,a2=d2$p1,b2=d2$p2,MoreArgs                                    =
list(dist_calr$population_dist[,1],dist_calr$population_dist[,1+g]), SIMPLIFY = T), na.rm=T)
          temp2                          <-                          temp2                          +
mean(mapply(FUN=dicc_f,a1=d1$V3,b1=d1$V4,a2=d3$V3,b2=d3$V4,MoreArgs                                    =
list(dist_sep$population_dist[,1],dist_sep$population_dist[,1+g]), SIMPLIFY = T), na.rm=T)
          temp3 <- temp3 + mean(mapply(FUN=dicc_f,a1=d1$V3,b1=d1$V4,a2=d4$a,b2=d4$b,MoreArgs
= list(dist_con[,1],dist_con[,1+g]), SIMPLIFY = T), na.rm=T)
        }
        DICC_calr[t,] <- temp1/5
        DICC_sep[t,] <- temp2/5
        DICC_con[t,] <- temp3/5

        write.table(true_para,                                  file                                =
paste0("sampleN",sampleN,"_itemN",itemN,"_simN",t,"para_true.csv"),  sep  =  ",",  quote  =  F,
row.names = F)
        write.table(para_sep,                                 file                                =
paste0("sampleN",sampleN,"_itemN",itemN,"_simN",t,"para_sep.csv"),  sep  =  ",",  quote  =  F,
row.names = F)
        write.table(para_calr,                                  file                                =
paste0("sampleN",sampleN,"_itemN",itemN,"_simN",t,"para_calr.csv"),  sep  =  ",",  quote  =  F,
row.names = F)
        write.table(res_con$para,                                   file                                =
paste0("sampleN",sampleN,"_itemN",itemN,"_simN",t,"para_con.csv"),  sep  =  ",",  quote  =  F,
row.names = F)

        pdist_mean_calr[t,] <- dist_calr$population_mean
        pdist_mean_con[t,] <- res_con$population_mean
        pdist_mean_sep[t,] <- dist_sep$population_mean
        pdist_sd_calr[t,] <- dist_calr$population_sd
        pdist_sd_con[t,] <- res_con$population_sd
        pdist_sd_sep[t,] <- dist_sep$population_sd

      } # end of while

    )
    write.table(ParaError_sep,                                  file                                =
paste0("sampleN",sampleN,"_itemN",itemN,"_ParaError_sep.csv"),  sep  =  ",",  quote  =  F,
row.names = F)
    write.table(ParaError_calr,                                   file                                =
paste0("sampleN",sampleN,"_itemN",itemN,"_ParaError_calr.csv"),  sep  =  ",",  quote  =  F,
row.names = F)
    write.table(ParaError_con,                                  file                                =
paste0("sampleN",sampleN,"_itemN",itemN,"_ParaError_con.csv"),  sep  =  ",",  quote  =  F,
row.names = F)
    write.table(DICC_calr, file = paste0("sampleN",sampleN,"_itemN",itemN,"_DICC_calr.csv"),
sep = ",", quote = F, row.names = F)
```

```
    write.table(DICC_sep,  file  =  paste0("sampleN",sampleN,"_itemN",itemN,"_DICC_sep.csv"),
sep = ",",  quote = F,  row.names = F)
    write.table(DICC_con,  file  =  paste0("sampleN",sampleN,"_itemN",itemN,"_DICC_con.csv"),
sep = ",",  quote = F,  row.names = F)
    write.table(pdist_mean_calr,                                file                                =
paste0("sampleN",sampleN,"_itemN",itemN,"_pdist_mean_calr.csv"),   sep   =   ",",   quote   =   F,
row.names = F)
    write.table(pdist_mean_sep,                                 file                                =
paste0("sampleN",sampleN,"_itemN",itemN,"_pdist_mean_sep.csv"),   sep   =   ",",   quote   =   F,
row.names = F)
    write.table(pdist_mean_con,                                 file                                =
paste0("sampleN",sampleN,"_itemN",itemN,"_pdist_mean_con.csv"),   sep   =   ",",   quote   =   F,
row.names = F)
    write.table(pdist_sd_calr,                                  file                                =
paste0("sampleN",sampleN,"_itemN",itemN,"_pdist_sd_calr.csv"),   sep   =   ",",   quote   =   F,
row.names = F)
    write.table(pdist_sd_sep,                                  file                                 =
paste0("sampleN",sampleN,"_itemN",itemN,"_pdist_sd_sep.csv"), sep = ",", quote = F, row.names
= F)
    write.table(pdist_sd_con,                                  file                                 =
paste0("sampleN",sampleN,"_itemN",itemN,"_pdist_sd_con.csv"), sep = ",", quote = F, row.names
= F)

    sink() #  アウトプットの終了。

  } # end of itemN


} # end of samoleN



#-------------------------------------------------------------------------------------------#


# シミュレーションプロット用
library(tidyverse)  # ggplot & tidyr data
library(RColorBrewer)  # usefull colour chart
library(devEMF)  # for output emf, only for windows


cols <- brewer.pal(6, "Paired")



setwd("YOUR OWN DIR")
OUTDIR <- "SET OUTPUT DIR"


# 項目パラメタ，母集団平均，標準偏差ファイルの読み込み
cc <- 0
```

```
calr_dicc_pdist <- numeric(9)
sep_dicc_pdist <- numeric(9)
con_dicc_pdist <- numeric(9)
calr_dicc <- numeric(9)
sep_dicc <- numeric(9)
con_dicc <- numeric(9)
itemsize <- numeric(9)
subjectsize <- numeric(9)
condition <- c("A1","A2","A3","B1","B2","B3","C1","C2","C3")
for(sampleN in c(400,1000,10000)){
  for(itemN in c(15,30,60)){
    cc <- cc + 1
    # item parameter error
    cone_para <- read.csv(paste0("sampleN",sampleN,"_itemN",itemN,"_ParaError_con.csv"))
    sepe_para <- read.csv(paste0("sampleN",sampleN,"_itemN",itemN,"_ParaError_sep.csv"))
    calre_para <- read.csv(paste0("sampleN",sampleN,"_itemN",itemN,"_ParaError_calr.csv"))
    # pdist
    con_mean <- read.csv(paste0("sampleN",sampleN,"_itemN",itemN,"_pdist_mean_con.csv"))
    con_sd <- read.csv(paste0("sampleN",sampleN,"_itemN",itemN,"_pdist_sd_con.csv"))
    sep_mean <- read.csv(paste0("sampleN",sampleN,"_itemN",itemN,"_pdist_mean_sep.csv"))
    sep_sd <- read.csv(paste0("sampleN",sampleN,"_itemN",itemN,"_pdist_sd_sep.csv"))
    calr_mean <- read.csv(paste0("sampleN",sampleN,"_itemN",itemN,"_pdist_mean_calr.csv"))
    calr_sd <- read.csv(paste0("sampleN",sampleN,"_itemN",itemN,"_pdist_sd_calr.csv"))
    # DiCC(pdist)
    con_dicc_temp <- read.csv(paste0("sampleN",sampleN,"_itemN",itemN,"_DICC_con.csv"))
    sep_dicc_temp <- read.csv(paste0("sampleN",sampleN,"_itemN",itemN,"_DICC_sep.csv"))
    calr_dicc_temp <- read.csv(paste0("sampleN",sampleN,"_itemN",itemN,"_DICC_calr.csv"))
    # assign
    assign(sprintf("con_para_%d_%d",sampleN,itemN), cone_para)
    assign(sprintf("sep_para_%d_%d",sampleN,itemN), sepe_para)
    assign(sprintf("calr_para_%d_%d",sampleN,itemN), calre_para)

    assign(sprintf("con_mean_%d_%d",sampleN,itemN), con_mean)
    assign(sprintf("sep_mean_%d_%d",sampleN,itemN), sep_mean)
    assign(sprintf("calr_mean_%d_%d",sampleN,itemN), calr_mean)
```

```
    assign(sprintf("con_sd_%d_%d",sampleN,itemN), con_sd)
    assign(sprintf("sep_sd_%d_%d",sampleN,itemN), sep_sd)
    assign(sprintf("calr_sd_%d_%d",sampleN,itemN), calr_sd)

    calr_dicc_pdist[cc] <- mean(calr_dicc_temp$V1)
    sep_dicc_pdist[cc]  <- mean(sep_dicc_temp$V1)
    con_dicc_pdist[cc]  <- mean(con_dicc_temp$V1)

    # DICC
    calr_t <- sep_t <- con_t <- 0 # initialization
    for(t in 1:100){
      true  <-  read.csv(paste0("sampleN",sampleN,"_itemN",itemN,"_simN",t,"para_true.csv"),
header = T)
      calr  <-  read.csv(paste0("sampleN",sampleN,"_itemN",itemN,"_simN",t,"para_calr.csv"),
header = T)
      sep   <-  read.csv(paste0("sampleN",sampleN,"_itemN",itemN,"_simN",t,"para_sep.csv"),
header = T)
      con   <-  read.csv(paste0("sampleN",sampleN,"_itemN",itemN,"_simN",t,"para_con.csv"),
header = T)
      # theta interval and point
      N <- 31
      theta <- seq(-4,4, length.out = N) # Arai & Mayekawa (2011) では-3~3の区間だった。
      # 全項目，全分点の平均確率を計算。
      calr_t <- calr_t + mean(mapply(FUN = function(a1,a2,b1,b2,theta) mean(abs(1/(1+exp(-
1.702*a2*(theta-b2))) - 1/(1+exp(-1.702*a1*(theta-b1))))),
                                     a1=true$V3,   a2=calr$p1,   b1=true$V4,   b2=calr$p2,
MoreArgs = list(theta=theta), SIMPLIFY = T))

      sep_t  <- sep_t + mean(mapply(FUN = function(a1,a2,b1,b2,theta) mean(abs(1/(1+exp(-
1.702*a2*(theta-b2))) - 1/(1+exp(-1.702*a1*(theta-b1))))),
                                    a1=true$V3, a2=sep$V3, b1=true$V4, b2=sep$V4, MoreArgs =
list(theta=theta), SIMPLIFY = T))

      con_t  <- con_t + mean(mapply(FUN = function(a1,a2,b1,b2,theta) mean(abs(1/(1+exp(-
1.702*a2*(theta-b2))) - 1/(1+exp(-1.702*a1*(theta-b1))))),
                                    a1=true$V3, a2=con$a, b1=true$V4, b2=con$b, MoreArgs =
```

```
list(theta=theta), SIMPLIFY = T))
    } # end of t
    calr_dicc[cc] <- calr_t/100
    sep_dicc[cc]  <- sep_t/100
    con_dicc[cc]  <- con_t/100
    itemsize[cc]  <- itemN
    subjectsize[cc] <- sampleN
  }
}

rm(true,calr,sep,con,calr_t,sep_t,con_t,con_dicc_temp,sep_dicc_temp,calr_dicc_temp)
rm(cone_para,sepe_para,calre_para,con_mean,sep_mean,calr_mean,con_sd,sep_sd,calr_sd)

cat("create data.frame¥r")
# para
# a
RMSEa_sep <- data.frame(i400j15 = sep_para_400_15$RMSE.a, i400s30 = sep_para_400_30$RMSE.a,
i400s60 = sep_para_400_60$RMSE.a,
                        i1000j15      =      sep_para_1000_15$RMSE.a,      i1000j30      =
sep_para_1000_30$RMSE.a, i1000j60 = sep_para_1000_60$RMSE.a,
                        i10000j15     =     sep_para_10000_15$RMSE.a,     i10000j30     =
sep_para_10000_30$RMSE.a, i10000j60 = sep_para_10000_60$RMSE.a)

RMSEa_calr    <-    data.frame(i400j15    =    calr_para_400_15$RMSE.a,    i400s30    =
calr_para_400_30$RMSE.a, i400s60 = calr_para_400_60$RMSE.a,
                         i1000j15      =      calr_para_1000_15$RMSE.a,      i1000j30      =
calr_para_1000_30$RMSE.a, i1000j60 = calr_para_1000_60$RMSE.a,
                         i10000j15     =     calr_para_10000_15$RMSE.a,     i10000j30     =
calr_para_10000_30$RMSE.a, i10000j60 = calr_para_10000_60$RMSE.a)

RMSEa_con <- data.frame(i400j15 = con_para_400_15$RMSE.a, i400s30 = con_para_400_30$RMSE.a,
i400s60 = con_para_400_60$RMSE.a,
                        i1000j15      =      con_para_1000_15$RMSE.a,      i1000j30      =
con_para_1000_30$RMSE.a, i1000j60 = con_para_1000_60$RMSE.a,
                        i10000j15     =     con_para_10000_15$RMSE.a,     i10000j30     =
con_para_10000_30$RMSE.a, i10000j60 = con_para_10000_60$RMSE.a)
```

```
#b
RMSEb_sep <- data.frame(i400j15 = sep_para_400_15$RMSE.b, i400s30 = sep_para_400_30$RMSE.b,
i400s60 = sep_para_400_60$RMSE.b,
                        i1000j15 = sep_para_1000_15$RMSE.b, i1000j30 =
sep_para_1000_30$RMSE.b, i1000j60 = sep_para_1000_60$RMSE.b,
                        i10000j15 = sep_para_10000_15$RMSE.b, i10000j30 =
sep_para_10000_30$RMSE.b, i10000j60 = sep_para_10000_60$RMSE.b)

RMSEb_calr <- data.frame(i400j15 = calr_para_400_15$RMSE.b, i400s30 =
calr_para_400_30$RMSE.b, i400s60 = calr_para_400_60$RMSE.b,
                         i1000j15 = calr_para_1000_15$RMSE.b, i1000j30 =
calr_para_1000_30$RMSE.b, i1000j60 = calr_para_1000_60$RMSE.b,
                         i10000j15 = calr_para_10000_15$RMSE.b, i10000j30 =
calr_para_10000_30$RMSE.b, i10000j60 = calr_para_10000_60$RMSE.b)

RMSEb_con <- data.frame(i400j15 = con_para_400_15$RMSE.b, i400s30 = con_para_400_30$RMSE.b,
i400s60 = con_para_400_60$RMSE.b,
                        i1000j15 = con_para_1000_15$RMSE.b, i1000j30 =
con_para_1000_30$RMSE.b, i1000j60 = con_para_1000_60$RMSE.b,
                        i10000j15 = con_para_10000_15$RMSE.b, i10000j30 =
con_para_10000_30$RMSE.b, i10000j60 = con_para_10000_60$RMSE.b)
# mean
#sep
sep_mean_g1 <- data.frame(i400j15=sep_mean_400_15$V1, i400s30 = sep_mean_400_30$V1, i400s60
= sep_mean_400_60$V1,
                          i1000j15 = sep_mean_1000_15$V1, i1000j30 = sep_mean_1000_30$V1,
i1000j60 = sep_mean_1000_60$V1,
                          i10000j15 = sep_mean_10000_15$V1, i10000j30 = sep_mean_10000_30$V1,
i10000j60 = sep_mean_10000_60$V1)
sep_mean_g2 <- data.frame(i400j15=sep_mean_400_15$V2, i400s30 = sep_mean_400_30$V2, i400s60
= sep_mean_400_60$V2,
                          i1000j15 = sep_mean_1000_15$V2, i1000j30 = sep_mean_1000_30$V2,
i1000j60 = sep_mean_1000_60$V2,
                          i10000j15 = sep_mean_10000_15$V2, i10000j30 = sep_mean_10000_30$V2,
i10000j60 = sep_mean_10000_60$V2)
```

```
sep_mean_g3 <- data.frame(i400j15=sep_mean_400_15$V3, i400s30 = sep_mean_400_30$V3, i400s60
= sep_mean_400_60$V3,
                          i1000j15 = sep_mean_1000_15$V3, i1000j30 = sep_mean_1000_30$V3,
i1000j60 = sep_mean_1000_60$V3,
                          i10000j15 = sep_mean_10000_15$V3, i10000j30 = sep_mean_10000_30$V3,
i10000j60 = sep_mean_10000_60$V3)
sep_mean_g4 <- data.frame(i400j15=sep_mean_400_15$V4, i400s30 = sep_mean_400_30$V4, i400s60
= sep_mean_400_60$V4,
                          i1000j15 = sep_mean_1000_15$V4, i1000j30 = sep_mean_1000_30$V4,
i1000j60 = sep_mean_1000_60$V4,
                          i10000j15 = sep_mean_10000_15$V4, i10000j30 = sep_mean_10000_30$V4,
i10000j60 = sep_mean_10000_60$V4)
sep_mean_g5 <- data.frame(i400j15=sep_mean_400_15$V5, i400s30 = sep_mean_400_30$V5, i400s60
= sep_mean_400_60$V5,
                          i1000j15 = sep_mean_1000_15$V5, i1000j30 = sep_mean_1000_30$V5,
i1000j60 = sep_mean_1000_60$V5,
                          i10000j15 = sep_mean_10000_15$V5, i10000j30 = sep_mean_10000_30$V5,
i10000j60 = sep_mean_10000_60$V5)
sep_mean_all                                                                          <-
dplyr::bind_rows(sep_mean_g1,sep_mean_g2,sep_mean_g3,sep_mean_g4,sep_mean_g5)
#calr
calr_mean_g1 <- data.frame(i400j15=calr_mean_400_15$V1, i400s30 = calr_mean_400_30$V1,
i400s60 = calr_mean_400_60$V1,
                           i1000j15 = calr_mean_1000_15$V1, i1000j30 = calr_mean_1000_30$V1,
i1000j60 = calr_mean_1000_60$V1,
                           i10000j15 = calr_mean_10000_15$V1, i10000j30 =
calr_mean_10000_30$V1, i10000j60 = calr_mean_10000_60$V1)
calr_mean_g2 <- data.frame(i400j15=calr_mean_400_15$V2, i400s30 = calr_mean_400_30$V2,
i400s60 = calr_mean_400_60$V2,
                           i1000j15 = calr_mean_1000_15$V2, i1000j30 = calr_mean_1000_30$V2,
i1000j60 = calr_mean_1000_60$V2,
                           i10000j15 = calr_mean_10000_15$V2, i10000j30 =
calr_mean_10000_30$V2, i10000j60 = calr_mean_10000_60$V2)
calr_mean_g3 <- data.frame(i400j15=calr_mean_400_15$V3, i400s30 = calr_mean_400_30$V3,
i400s60 = calr_mean_400_60$V3,
                           i1000j15 = calr_mean_1000_15$V3, i1000j30 = calr_mean_1000_30$V3,
```

```
i1000j60 = calr_mean_1000_60$V3,
                          i10000j15 = calr_mean_10000_15$V3, i10000j30 =
calr_mean_10000_30$V3, i10000j60 = calr_mean_10000_60$V3)
calr_mean_g4 <- data.frame(i400j15=calr_mean_400_15$V4, i400s30 = calr_mean_400_30$V4,
i400s60 = calr_mean_400_60$V4,
                          i1000j15 = calr_mean_1000_15$V4, i1000j30 = calr_mean_1000_30$V4,
i1000j60 = calr_mean_1000_60$V4,
                          i10000j15 = calr_mean_10000_15$V4, i10000j30 =
calr_mean_10000_30$V4, i10000j60 = calr_mean_10000_60$V4)
calr_mean_g5 <- data.frame(i400j15=calr_mean_400_15$V5, i400s30 = calr_mean_400_30$V5,
i400s60 = calr_mean_400_60$V5,
                          i1000j15 = calr_mean_1000_15$V5, i1000j30 = calr_mean_1000_30$V5,
i1000j60 = calr_mean_1000_60$V5,
                          i10000j15 = calr_mean_10000_15$V5, i10000j30 =
calr_mean_10000_30$V5, i10000j60 = calr_mean_10000_60$V5)
calr_mean_all                                                                    <-
dplyr::bind_rows(calr_mean_g1,calr_mean_g2,calr_mean_g3,calr_mean_g4,calr_mean_g5)
#con
con_mean_g1 <- data.frame(i400j15=con_mean_400_15$V1, i400s30 = con_mean_400_30$V1, i400s60
= con_mean_400_60$V1,
                          i1000j15 = con_mean_1000_15$V1, i1000j30 = con_mean_1000_30$V1,
i1000j60 = con_mean_1000_60$V1,
                          i10000j15 = con_mean_10000_15$V1, i10000j30 = con_mean_10000_30$V1,
i10000j60 = con_mean_10000_60$V1)
con_mean_g2 <- data.frame(i400j15=con_mean_400_15$V2, i400s30 = con_mean_400_30$V2, i400s60
= con_mean_400_60$V2,
                          i1000j15 = con_mean_1000_15$V2, i1000j30 = con_mean_1000_30$V2,
i1000j60 = con_mean_1000_60$V2,
                          i10000j15 = con_mean_10000_15$V2, i10000j30 = con_mean_10000_30$V2,
i10000j60 = con_mean_10000_60$V2)
con_mean_g3 <- data.frame(i400j15=con_mean_400_15$V3, i400s30 = con_mean_400_30$V3, i400s60
= con_mean_400_60$V3,
                          i1000j15 = con_mean_1000_15$V3, i1000j30 = con_mean_1000_30$V3,
i1000j60 = con_mean_1000_60$V3,
                          i10000j15 = con_mean_10000_15$V3, i10000j30 = con_mean_10000_30$V3,
i10000j60 = con_mean_10000_60$V3)
```

```
con_mean_g4 <- data.frame(i400j15=con_mean_400_15$V4, i400s30 = con_mean_400_30$V4, i400s60
= con_mean_400_60$V4,
                          i1000j15 = con_mean_1000_15$V4, i1000j30 = con_mean_1000_30$V4,
i1000j60 = con_mean_1000_60$V4,
                          i10000j15 = con_mean_10000_15$V4, i10000j30 = con_mean_10000_30$V4,
i10000j60 = con_mean_10000_60$V4)
con_mean_g5 <- data.frame(i400j15=con_mean_400_15$V5, i400s30 = con_mean_400_30$V5, i400s60
= con_mean_400_60$V5,
                          i1000j15 = con_mean_1000_15$V5, i1000j30 = con_mean_1000_30$V5,
i1000j60 = con_mean_1000_60$V5,
                          i10000j15 = con_mean_10000_15$V5, i10000j30 = con_mean_10000_30$V5,
i10000j60 = con_mean_10000_60$V5)
con_mean_all                                                                          <-
dplyr::bind_rows(con_mean_g1,con_mean_g2,con_mean_g3,con_mean_g4,con_mean_g5)

# sd
#sep
sep_sd_g1 <- data.frame(i400j15=sep_sd_400_15$V1, i400s30 = sep_sd_400_30$V1, i400s60 =
sep_sd_400_60$V1,
                        i1000j15 = sep_sd_1000_15$V1, i1000j30 = sep_sd_1000_30$V1, i1000j60
= sep_sd_1000_60$V1,
                        i10000j15 = sep_sd_10000_15$V1, i10000j30 = sep_sd_10000_30$V1,
i10000j60 = sep_sd_10000_60$V1)
sep_sd_g2 <- data.frame(i400j15=sep_sd_400_15$V2, i400s30 = sep_sd_400_30$V2, i400s60 =
sep_sd_400_60$V2,
                        i1000j15 = sep_sd_1000_15$V2, i1000j30 = sep_sd_1000_30$V2, i1000j60
= sep_sd_1000_60$V2,
                        i10000j15 = sep_sd_10000_15$V2, i10000j30 = sep_sd_10000_30$V2,
i10000j60 = sep_sd_10000_60$V2)
sep_sd_g3 <- data.frame(i400j15=sep_sd_400_15$V3, i400s30 = sep_sd_400_30$V3, i400s60 =
sep_sd_400_60$V3,
                        i1000j15 = sep_sd_1000_15$V3, i1000j30 = sep_sd_1000_30$V3, i1000j60
= sep_sd_1000_60$V3,
                        i10000j15 = sep_sd_10000_15$V3, i10000j30 = sep_sd_10000_30$V3,
i10000j60 = sep_sd_10000_60$V3)
sep_sd_g4 <- data.frame(i400j15=sep_sd_400_15$V4, i400s30 = sep_sd_400_30$V4, i400s60 =
```

```
sep_sd_400_60$V4,
                        i1000j15 = sep_sd_1000_15$V4, i1000j30 = sep_sd_1000_30$V4, i1000j60
= sep_sd_1000_60$V4,
                        i10000j15 = sep_sd_10000_15$V4, i10000j30 = sep_sd_10000_30$V4,
i10000j60 = sep_sd_10000_60$V4)
sep_sd_g5 <- data.frame(i400j15=sep_sd_400_15$V5, i400s30 = sep_sd_400_30$V5, i400s60 =
sep_sd_400_60$V5,
                        i1000j15 = sep_sd_1000_15$V5, i1000j30 = sep_sd_1000_30$V5, i1000j60
= sep_sd_1000_60$V5,
                        i10000j15 = sep_sd_10000_15$V5, i10000j30 = sep_sd_10000_30$V5,
i10000j60 = sep_sd_10000_60$V5)
sep_sd_all <- dplyr::bind_rows(sep_sd_g1,sep_sd_g2,sep_sd_g3,sep_sd_g4,sep_sd_g5)
#calr
calr_sd_g1 <- data.frame(i400j15=calr_sd_400_15$V1, i400s30 = calr_sd_400_30$V1, i400s60 =
calr_sd_400_60$V1,
                         i1000j15 = calr_sd_1000_15$V1, i1000j30 = calr_sd_1000_30$V1,
i1000j60 = calr_sd_1000_60$V1,
                         i10000j15 = calr_sd_10000_15$V1, i10000j30 = calr_sd_10000_30$V1,
i10000j60 = calr_sd_10000_60$V1)
calr_sd_g2 <- data.frame(i400j15=calr_sd_400_15$V2, i400s30 = calr_sd_400_30$V2, i400s60 =
calr_sd_400_60$V2,
                         i1000j15 = calr_sd_1000_15$V2, i1000j30 = calr_sd_1000_30$V2,
i1000j60 = calr_sd_1000_60$V2,
                         i10000j15 = calr_sd_10000_15$V2, i10000j30 = calr_sd_10000_30$V2,
i10000j60 = calr_sd_10000_60$V2)
calr_sd_g3 <- data.frame(i400j15=calr_sd_400_15$V3, i400s30 = calr_sd_400_30$V3, i400s60 =
calr_sd_400_60$V3,
                         i1000j15 = calr_sd_1000_15$V3, i1000j30 = calr_sd_1000_30$V3,
i1000j60 = calr_sd_1000_60$V3,
                         i10000j15 = calr_sd_10000_15$V3, i10000j30 = calr_sd_10000_30$V3,
i10000j60 = calr_sd_10000_60$V3)
calr_sd_g4 <- data.frame(i400j15=calr_sd_400_15$V4, i400s30 = calr_sd_400_30$V4, i400s60 =
calr_sd_400_60$V4,
                         i1000j15 = calr_sd_1000_15$V4, i1000j30 = calr_sd_1000_30$V4,
i1000j60 = calr_sd_1000_60$V4,
                         i10000j15 = calr_sd_10000_15$V4, i10000j30 = calr_sd_10000_30$V4,
```

```
i10000j60 = calr_sd_10000_60$V4)
calr_sd_g5 <- data.frame(i400j15=calr_sd_400_15$V5, i400s30 = calr_sd_400_30$V5, i400s60 = calr_sd_400_60$V5,
                         i1000j15 = calr_sd_1000_15$V5, i1000j30 = calr_sd_1000_30$V5, i1000j60 = calr_sd_1000_60$V5,
                         i10000j15 = calr_sd_10000_15$V5, i10000j30 = calr_sd_10000_30$V5, i10000j60 = calr_sd_10000_60$V5)
calr_sd_all <- dplyr::bind_rows(calr_sd_g1,calr_sd_g2,calr_sd_g3,calr_sd_g4,calr_sd_g5)
#con
con_sd_g1 <- data.frame(i400j15=con_sd_400_15$V1, i400s30 = con_sd_400_30$V1, i400s60 = con_sd_400_60$V1,
                        i1000j15 = con_sd_1000_15$V1, i1000j30 = con_sd_1000_30$V1, i1000j60 = con_sd_1000_60$V1,
                        i10000j15 = con_sd_10000_15$V1, i10000j30 = con_sd_10000_30$V1, i10000j60 = con_sd_10000_60$V1)
con_sd_g2 <- data.frame(i400j15=con_sd_400_15$V2, i400s30 = con_sd_400_30$V2, i400s60 = con_sd_400_60$V2,
                        i1000j15 = con_sd_1000_15$V2, i1000j30 = con_sd_1000_30$V2, i1000j60 = con_sd_1000_60$V2,
                        i10000j15 = con_sd_10000_15$V2, i10000j30 = con_sd_10000_30$V2, i10000j60 = con_sd_10000_60$V2)
con_sd_g3 <- data.frame(i400j15=con_sd_400_15$V3, i400s30 = con_sd_400_30$V3, i400s60 = con_sd_400_60$V3,
                        i1000j15 = con_sd_1000_15$V3, i1000j30 = con_sd_1000_30$V3, i1000j60 = con_sd_1000_60$V3,
                        i10000j15 = con_sd_10000_15$V3, i10000j30 = con_sd_10000_30$V3, i10000j60 = con_sd_10000_60$V3)
con_sd_g4 <- data.frame(i400j15=con_sd_400_15$V4, i400s30 = con_sd_400_30$V4, i400s60 = con_sd_400_60$V4,
                        i1000j15 = con_sd_1000_15$V4, i1000j30 = con_sd_1000_30$V4, i1000j60 = con_sd_1000_60$V4,
                        i10000j15 = con_sd_10000_15$V4, i10000j30 = con_sd_10000_30$V4, i10000j60 = con_sd_10000_60$V4)
con_sd_g5 <- data.frame(i400j15=con_sd_400_15$V5, i400s30 = con_sd_400_30$V5, i400s60 = con_sd_400_60$V5,
                        i1000j15 = con_sd_1000_15$V5, i1000j30 = con_sd_1000_30$V5, i1000j60
```

```
= con_sd_1000_60$V5,
                          i10000j15 = con_sd_10000_15$V5, i10000j30 = con_sd_10000_30$V5,
i10000j60 = con_sd_10000_60$V5)
con_sd_all <- dplyr::bind_rows(con_sd_g1,con_sd_g2,con_sd_g3,con_sd_g4,con_sd_g5)
# rm
rm(sep_mean_g1,sep_mean_g2,sep_mean_g3,sep_mean_g4,sep_mean_g5,calr_mean_g1,calr_mean_g2,ca
lr_mean_g3,calr_mean_g4,calr_mean_g5,

con_mean_g1,con_mean_g2,con_mean_g3,con_mean_g4,con_mean_g5,sep_sd_g1,sep_sd_g2,sep_sd_g3,s
ep_sd_g4,sep_sd_g5,

calr_sd_g1,calr_sd_g2,calr_sd_g3,calr_sd_g4,calr_sd_g5,con_sd_g1,con_sd_g2,con_sd_g3,con_sd
_g4,con_sd_g5,

con_mean_400_15,con_mean_400_30,con_mean_400_60,con_mean_1000_15,con_mean_1000_30,con_mean_
1000_60,con_mean_10000_15,con_mean_10000_30,con_mean_10000_60,

calr_mean_400_15,calr_mean_400_30,calr_mean_400_60,calr_mean_1000_15,calr_mean_1000_30,calr
_mean_1000_60,calr_mean_10000_15,calr_mean_10000_30,calr_mean_10000_60,

sep_mean_400_15,sep_mean_400_30,sep_mean_400_60,sep_mean_1000_15,sep_mean_1000_30,sep_mean_
1000_60,sep_mean_10000_15,sep_mean_10000_30,sep_mean_10000_60,

con_sd_400_15,con_sd_400_30,con_sd_400_60,con_sd_1000_15,con_sd_1000_30,con_sd_1000_60,con_
sd_10000_15,con_sd_10000_30,con_sd_10000_60,

calr_sd_400_15,calr_sd_400_30,calr_sd_400_60,calr_sd_1000_15,calr_sd_1000_30,calr_sd_1000_6
0,calr_sd_10000_15,calr_sd_10000_30,calr_sd_10000_60,

sep_sd_400_15,sep_sd_400_30,sep_sd_400_60,sep_sd_1000_15,sep_sd_1000_30,sep_sd_1000_60,sep_
sd_10000_15,sep_sd_10000_30,sep_sd_10000_60)

cat("tidyr for facet¥r")
# item para
# すべてをまとめてファセットする
RMSE_a_res      <-      data.frame(subject=c(rep(400,300),rep(1000,300),rep(10000,300))      ,
```

```
item=rep(c(rep(15,100),rep(30,100),rep(60,100)),3),
                          SL=tidyr::gather(RMSEa_sep)$value,
CC=tidyr::gather(RMSEa_con)$value, calr=tidyr::gather(RMSEa_calr)$value) %>%
  tidyr::gather(key=method, value=RMSE, -subject, -item)

RMSE_b_res      <-      data.frame(subject=c(rep(400,300),rep(1000,300),rep(10000,300))      ,
item=rep(c(rep(15,100),rep(30,100),rep(60,100)),3),
                          SL=tidyr::gather(RMSEb_sep)$value,
CC=tidyr::gather(RMSEb_con)$value, calr=tidyr::gather(RMSEb_calr)$value) %>%
  tidyr::gather(key=method, value=RMSE, -subject, -item)
# mean
mean_res      <-      data.frame(subject=c(rep(400,1500),rep(1000,1500),rep(10000,1500))      ,
item=rep(c(rep(15,500),rep(30,500),rep(60,500)),3),

grade=rep(c(rep("G1",100),rep("G2",100),rep("G3",100),rep("G4",100),rep("G5",100)),3),
                        SL=tidyr::gather(sep_mean_all)$value,
                        CC=tidyr::gather(con_mean_all)$value,
                        calr=tidyr::gather(calr_mean_all)$value) %>%
  tidyr::gather(key=method, value=mean, -subject, -item, -grade)
# sd
sd_res      <-      data.frame(subject=c(rep(400,1500),rep(1000,1500),rep(10000,1500))      ,
item=rep(c(rep(15,500),rep(30,500),rep(60,500)),3),

grade=rep(c(rep("G1",100),rep("G2",100),rep("G3",100),rep("G4",100),rep("G5",100)),3),
                      SL=tidyr::gather(sep_sd_all)$value,
                      CC=tidyr::gather(con_sd_all)$value,
                      calr=tidyr::gather(calr_sd_all)$value) %>%
  tidyr::gather(key=method, value=sd, -subject, -item, -grade)
# DICC
dicc_res             <-             data.frame(condition=condition,             item=itemsize,
subject=subjectsize,calr=calr_dicc, SL=sep_dicc, con=con_dicc) %>%
  tidyr::gather(key=method, value=DICC, -item, -subject, -condition)
# DICC_pdist
dicc_pdist_res          <-          data.frame(condition=condition,          item=itemsize,
subject=subjectsize,calr=calr_dicc_pdist, SL=sep_dicc_pdist, con=con_dicc_pdist) %>%
  tidyr::gather(key=method, value=DICC, -item, -subject, -condition)
```

```
# まずは全実験結果を一覧でプロット

leg_lab <- c("A1","A2","A3","B1","B2","B3","C1","C2","C3") # legend
leg <- c("A1: 受検者数 10,000　項目数 15","A2: 受検者数 10,000　項目数 30","A3: 受検者数 10,000　項目数 60",
         "B1: 受検者数 1,000　項目数 15","B2: 受検者数 1,000　項目数 30","B3: 受検者数 1,000　項目数 60",
         "C1: 受検者数 400　項目数 15","C2: 受検者数 400　項目数 30","C3: 受検者数 400　項目数 60") # legend

calrplot_a <- RMSEa_calr %>% tidyr::gather(key = condition, value = RMSE) %>%
  ggplot(aes(x=condition,y=RMSE,fill=condition))+
  ylim(0,0.3) +
  #geom_boxplot()+
  geom_violin()+
  theme(axis.text=element_text(size=15),axis.title=element_text(size=15,face="bold")) +
  scale_fill_manual(values =
c("#1874CD","#1874CD","#1874CD","#EE6363","#EE6363","#EE6363","#008B45","#008B45","#008B45"
),
                    labels = leg) +
  scale_x_discrete(labels = leg_lab)

calrplot_b <- RMSEb_calr %>% tidyr::gather(key = condition, value = RMSE) %>%
  ggplot(aes(x=condition,y=RMSE,fill=condition))+
  #ylim(0,2) +
  #geom_boxplot()+
  geom_violin()+
  theme(axis.text=element_text(size=15),axis.title=element_text(size=15,face="bold")) +
  scale_fill_manual(values =
c("#1874CD","#1874CD","#1874CD","#EE6363","#EE6363","#EE6363","#008B45","#008B45","#008B45"
),
                    labels = leg) +
  scale_x_discrete(labels = leg_lab)

sepplot_a <- RMSEa_sep %>% tidyr::gather(key = condition, value = RMSE) %>%
```

```
  ggplot(aes(x=condition,y=RMSE,fill=condition))+
  ylim(0,0.3) +
  #geom_boxplot()+
  geom_violin()+
  theme(axis.text=element_text(size=15),axis.title=element_text(size=15,face="bold")) +
  scale_fill_manual(values                                                    =
c("#1874CD","#1874CD","#1874CD","#EE6363","#EE6363","#EE6363","#008B45","#008B45","#008B45"
),
                    labels = leg) +
  scale_x_discrete(labels = leg_lab)

sepplot_b <- RMSEb_sep %>% tidyr::gather(key = condition, value = RMSE) %>%
  ggplot(aes(x=condition,y=RMSE,fill=condition))+
  #ylim(0,1) +
  #geom_boxplot()+
  geom_violin()+
  theme(axis.text=element_text(size=15),axis.title=element_text(size=15,face="bold")) +
  scale_fill_manual(values                                                    =
c("#1874CD","#1874CD","#1874CD","#EE6363","#EE6363","#EE6363","#008B45","#008B45","#008B45"
),
                    labels = leg) +
  scale_x_discrete(labels = leg_lab)

conplot_a <- RMSEa_con %>% tidyr::gather(key = condition, value = RMSE) %>%
  ggplot(aes(x=condition,y=RMSE,fill=condition))+
  #ylim(0,0.3) +
  #geom_boxplot()+
  geom_violin()+
  theme(axis.text=element_text(size=15),axis.title=element_text(size=15,face="bold")) +
  scale_fill_manual(values                                                    =
c("#1874CD","#1874CD","#1874CD","#EE6363","#EE6363","#EE6363","#008B45","#008B45","#008B45"
),
                    labels = leg) +
  scale_x_discrete(labels = leg_lab)

conplot_b <- RMSEb_con %>% tidyr::gather(key = condition, value = RMSE) %>%
```

```
  ggplot(aes(x=condition,y=RMSE,fill=condition))+
  #ylim(0,2) +
  #geom_boxplot()+
  geom_violin()+
  theme(axis.text=element_text(size=15),axis.title=element_text(size=15,face="bold")) +
  scale_fill_manual(values                                                    =
c("#1874CD","#1874CD","#1874CD","#EE6363","#EE6363","#EE6363","#008B45","#008B45","#008B45"
),
                    labels = leg) +
  scale_x_discrete(labels = leg_lab)

# ggplot_facet
RMSE_a_plot <- ggplot(RMSE_a_res, aes(y=RMSE, x=method, fill=method, colour=method)) +
  #geom_boxplot()+
  geom_violin()+
  #ylim(0,0.4) +
  facet_grid(item~subject) +
  theme(axis.text=element_text(size=12),axis.title=element_text(size=12,face="bold")) +
  scale_fill_manual(values = c(cols[1],cols[5],cols[3])) +# 色分けを独自定義
  scale_colour_manual(values = c(cols[2],cols[6],cols[4])) # 色分けを独自定義

RMSE_b_plot <- ggplot(RMSE_b_res, aes(y=RMSE, x=method, fill=method, colour=method)) +
  #geom_boxplot()+
  geom_violin()+
  #ylim(0,2) + # 外れ値がいくつか消される。
  facet_grid(item~subject) +
  theme(axis.text=element_text(size=12),axis.title=element_text(size=12,face="bold")) +
  scale_fill_manual(values = c(cols[1],cols[5],cols[3])) +# 色分けを独自定義
  scale_colour_manual(values = c(cols[2],cols[6],cols[4])) # 色分けを独自定義

# DICC
# DICC_nomal
dicc_plot <- ggplot(dicc_res, aes(y=DICC, x=method, fill=method)) +
  geom_bar(stat = "identity") +
  facet_grid(item~subject) +
  theme(axis.text=element_text(size=12),axis.title=element_text(size=12,face="bold")) +
```

```
  scale_fill_manual(values = c(cols[2],cols[6],cols[4]), lab=c("calr","CC","SL"))
#scale_fill_manual(values = c("#1874CD","#EE6363","#008B45")) # 色分けを独自定義(カラフル)
#scale_fill_manual(values = c("#030303","#454545","#ADADAD")) # モノクロ（論文，白黒用）

# DICC_pdist
dicc_pdist_plot <- ggplot(dicc_pdist_res, aes(y=DICC, x=method, fill=method)) +
  geom_bar(stat = "identity") +
  facet_grid(item~subject) +
  theme(axis.text=element_text(size=12),axis.title=element_text(size=12,face="bold")) +
  scale_fill_manual(values = c(cols[2],cols[6],cols[4]), lab=c("calr","CC","SL")) # 色分けを
独自定義(カラフル)

mean_res_plot <- mean_res %>% ggplot(aes(x=grade,y=mean,fill=method,colour=method)) +
  geom_boxplot()+
  #geom_violin()+
  #ylim(0,2)+
  facet_grid(item~subject) +
  #geom_hline(yintercept = c(0,0.4,0.8,1.2,1.6), colour="blue", linetype="dashed") +
  theme(axis.text=element_text(size=12),axis.title=element_text(size=12,face="bold")) +
  scale_fill_manual(values = c(cols[1],cols[5],cols[3]), lab=c("calr","CC","SL")) +# 色分け
を独自定義
  scale_colour_manual(values = c(cols[2],cols[6],cols[4])) # 色分けを独自定義

sd_res_plot <- sd_res %>% ggplot(aes(x=grade,y=sd,fill=method,colour=method)) +
  geom_boxplot()+
  #geom_violin()+
  facet_grid(item~subject) +
  theme(axis.text=element_text(size=12),axis.title=element_text(size=12,face="bold")) +
  scale_fill_manual(values = c(cols[1],cols[5],cols[3]), lab=c("calr","CC","SL")) +# 色分け
を独自定義
  scale_colour_manual(values = c(cols[2],cols[6],cols[4])) # 色分けを独自定義

cat("output plot as meta file¥r")

setwd(OUTDIR)
```

```
# グラフ出力の前に，記述統計量を確認
sink(file="tapply_stat.txt")
cat("RMSE_a¥n")
with(RMSE_a_res,tapply(RMSE, list(method,subject,item), mean, simplify=F))
cat("RMSE_b¥n")
with(RMSE_b_res,tapply(RMSE, list(method,subject,item), mean))
cat("pdist_mean¥n")
with(mean_res, tapply(mean, list(method,grade,subject,item), mean))
cat("pdist_sd¥n")
with(sd_res, tapply(sd, list(method,grade,subject,item), mean))
sink()

emf(paste0(sim,"_sepplot_a.emf"), width = 9, height = 5)
sepplot_a
dev.off()

emf(paste0(sim,"_sepplot_b.emf"), width = 9, height = 5)
sepplot_b
dev.off()

emf(paste0(sim,"_calrplot_a.emf"), width = 9, height = 5)
calrplot_a
dev.off()

emf(paste0(sim,"_calrplot_b.emf"), width = 9, height = 5)
calrplot_b
dev.off()

emf(paste0(sim,"_conplot_a.emf"), width = 9, height = 5)
conplot_a
dev.off()

emf(paste0(sim,"_conplot_b.emf"), width = 9, height = 5)
conplot_b
dev.off()
```

```
emf(paste0(sim,"_RMSE_a_facet.emf"), width=7, height = 5)
RMSE_a_plot
dev.off()

emf(paste0(sim,"_RMSE_a_facet_zoom.emf"), width=7, height = 5)
RMSE_a_plot+ylim(0,0.25)
dev.off()

emf(paste0(sim,"_RMSE_b_facet.emf"), width=7, height = 5)
RMSE_b_plot
dev.off()

emf(paste0(sim,"_RMSE_b_facet_zoom.emf"), width=7, height = 5)
RMSE_b_plot+ylim(0,0.5)
dev.off()

emf(paste0(sim,"_DICC_facet_plot.emf"), width = 7,height = 5)
dicc_plot
dev.off()

emf(paste0(sim,"_DICC_pdist_plot.emf"), width=7, height=5)
dicc_pdist_plot
dev.off()

emf(paste0(sim,"_mean_plot.emf"), width=8, height=5)
mean_res_plot+
  geom_hline(yintercept = 0, linetype="dotdash")+
  geom_hline(yintercept = 0.4, linetype="dotdash")+
  geom_hline(yintercept = 0.8, linetype="dotdash")+
  geom_hline(yintercept = -0.4, linetype="dotdash")+
  geom_hline(yintercept = -0.8, linetype="dotdash")

dev.off()

emf(paste0(sim,"_sd_plot.emf"), width=8, height=5)
sd_res_plot+
```

```
  geom_hline(yintercept = 1, linetype="dotdash")
dev.off()
```