論文 / 著書情報
Article / Book Information

題目(和文)	項目反応理論を用いた大規模試験の運用に関する諸問題の検討とその解決：予備的試験の規模に制約がある場合を通じて
Title(English)	
著者(和文)	光永悠彦
Author(English)	Haruhiko Mitunaga
出典(和文)	学位:博士(学術), 学位授与機関:東京工業大学, 報告番号:甲第9219号, 授与年月日:2013年3月26日, 学位の種別:課程博士, 審査員:前川 眞一
Citation(English)	Degree:Doctor (Academic), Conferring organization: Tokyo Institute of Technology, Report number:甲第9219号, Conferred date:2013/3/26, Degree Type:Course doctor, Examiner:
学位種別(和文)	博士論文
Type(English)	Doctoral Thesis

項目反応理論を用いた大規模試験の運用に関する諸問題の検討とその解決

— 予備的試験の規模に制約がある場合を通じて —

光永悠彦

和文要約

項目反応理論 (IRT) に基づく等化を行うことによって、複数回のテストの結果を相互に比較可能にする標準化テストが、多くのテスト実施機関によって近年行われるようになってきた。本論では、本試験の前に大規模な予備試験を行えず、十分な大きさの項目バンクが用意できない場合のためのテストデザインを考え、本デザインで構築される項目バンクの中に入れるべき各項目の特性（識別力・困難度）の推定値が等化方法によってどのように影響を受けるかを探り、実際の試験でこのデザインが適用可能な形で、最も推定誤差が小さくなる等化方法を提案することを目的とした。同時に、等化の前提として「尺度の一次元性」が満たされる場合と不十分な場合において、最適な等化方法を検討した。

本論では、第 1 章において、IRT に基づく等化の考え方を概説し、次いで等化の要件、とりわけテストの一次元性について、IRT と因子分析との関係を述べた。さらに、テストの等化方法として個別推定と同時推定があり、これらの方法にはそれぞれ複数の方法が提案されていることを述べた。その上で、本論で取り上げる「予備試験を十分な規模で行えない場合のテストデザイン」を提示し、このデザインにおける等化上の注意点について述べた。

続いて第 2 章において、個別推定の場合に考えなければならない順序性の問題を取り上げ、ペアワイズな等化方法を用いた場合に、規準集団への等化の順序によって等化後の項目バンクにおけるパラメタの推定値がどのように影響を受けるかを検討した。同時に、順序性を考えなくてよい等化法として calr 法 (Mayekawa,1991 他) を取り上げ、結果を比較した。シミュレーション研究の結果、calr 法が最も真値に近い推定値を示した。また、ICC 法を適用すれば順序の効果は小さいものの、Mean/Sigma 法を用いた場合は項目パラメタの推定において順序性の影響を大きく受けることが分かった。

また、第 3 章において、本論で提示したテストデザインを用いて項目バンクを構築する場面を考え、calr 法を用いた個別推定と、同時推定のいずれを用いた方が真値に近い推定結果となるかを、シミュレーション研究により検討した。結果、予備試験の真の識別力が小さく、本試験の真の識別力が大きい場合、識別力の推定値が過小に推定されることが分かった。また、同時推定を行う際、累積分の反応データから得られた項目パラメタのセットを、あらためて規準集団で得られた項目パラメタに等化する方法をとることにより、多群 IRT モデルを適用するのみの場合に比べ、より真値に近い推定結果となることが分かった。

第 4 章においては、同時推定を行う際に、多群 IRT モデルによる項目パラメタの推定を行うソフトウェアとして BILOG-MG を取り上げ、群の数や新作項目数などを操作したときにメモリの消費量がどのように変化するか検討した。その結果、多群 IRT モデルにおいて、群の数がメモリ消費量に大きく影響することが示された。本論で取り上げるテストデザインは、実施のたびにテスト冊子のバリエーションが増えることから、群の数を減らした場合に等化済み項目パラメタの推定値がどのようになるかを検討する必要が生じた。

第 5 章においては、同時推定場面で多群 IRT による推定を、群の数を減らして行った場合、項目パラメタの推定値がどのように影響を受けるかを、シミュレーションにより検討した。結果、互いに似た真の能力値の平均を持つような複数の群を 1 群にまとめることで、1 冊子 1 群に対応した多群 IRT モデルと同じ程度の真値からのずれを持つ項目パラメタの推定値が得られた。このモデルを用いることで、群の数を減らして同時推定を行い続けることが可能であることが示唆された。

第 6 章では、本論で提案されたテストデザインにおいて、個別推定で calr 法を用いる方法が、最も真値に近い項目パラメタとなる可能性に触れた。しかし、個別推定は、毎回の試験が測定している構成概念がいずれの回においても同一で、複数の回をまたいで 1 因子構造である場合には妥当な結果を返すが、そうでない場合は同時推定を行った方が、結果として妥当な推定値を返すことを指摘した。いずれの方法をとるかは、テストの一次元性の仮定をどのように検証するかに依存するものの、それぞれの方法で実際に実行可能な等化方法が提案されたことにより、テストデータの分析を事後的に行うことで、最適な等化方法を見出すことが可能であることを指摘した。最後に、今後の等化研究の展望を述べ、日本における等化方法の研究のあり方について議論した。

Abstract

Although the common-item-nonequivalent-groups design has been widely used for the purpose of equating large-scale test items, the performance of this design is not fully investigated in such situations where the number of items in the item bank is small and the number of linking items between forms and item banks are small. In this thesis, a method to administer the test when the size of the item bank is not large enough is proposed.

The purpose of the thesis is threefold: (1) to explore the effect of the order or the equating when the items are separately calibrated, (2) to find the best method to calibrate/equate item parameters which best recovers the true value, (3) to show the computational limit of the concurrent calibration/equating methods and the way to avoid limitation by reducing the number of groups under the multi-group IRT model.

Simulation studies show the following result: (a) separate calibration method produces the estimate close to the true value than concurrent calibration, whereas concurrent calibration must be chosen in case where the assumption of unidimensionality is not supported, (b) it is impossible to estimate parameters using BILOG-MG in multi-group IRT model when the number of groups is increased to the magnitude of the typical large-scale test, (c) in order to use concurrent calibration method in such a large-scale dataset, it is useful to reduce the number of groups in such a way that the mean values of the latent ability are similar among the merged groups in a multi-group IRT model.

Finally, the effectiveness of the proposed calibration/equating procedure in a practical standardized test context is discussed, as well as the need for the future research.

目次

表目次

図目次

第 1 章

序論および目的

1.1　試験における等化の役割とその位置づけ

1.1.1　試験の制度設計と項目バンク

我々の身の回りには、多くの試験が存在する。これらの試験では、受験者の持つ特性を測定し、結果を数字で表すことで、その特性があらかじめ定められた基準を満たしているか、あるいは背後に仮定される母集団上、いいかえれば社会の上でどれほどのレベルに位置づけられるかを判断する。たとえば、英語能力試験では、まず「英語能力」なるものとは何かを専門家が吟味し、その社会的位置づけ、あるいは一般性について一定の合意を得たうえで、それを的確に問うための試験問題（以下、試験問題の意味で「試験項目」「項目（item)」と表記する）を受験者に提示し、得られた正誤反応（以下、データとしてみる場合は「0-1 データ」と表記する）から各受験者について成績を推定する。ここで推定された成績は、一般化された英語能力が、母集団上においてどのレベルにあるかを各受験者について表しているので、社会一般において試験成績として有効であるとみなされ、以て「英語能力を測定する公的試験」の意味を成している。

近年、実施のたびに異なる項目が受験者に提示されるのにもかかわらず、テスト得点が異なる回をまたいで比較可能である、という利点を持つテストが出現し、実際の成績判断に利用されている。たとえば OECD が 3 年に一度行っている PISA(Programme for International Student Assessment) では、複数の実施地域（主に国単位）の中等教育課程に属する生徒の学力を、複数の会場をまたいで共通な尺度上の点数で表現することで、実施地域ごと、あるいは実施年ごとに異なる項目を提示しながら、それらの間で学力がどの程度異なるかを把握することができる。また、国際的でないテストであっても、公的機関や企業が行うテストで、実施回をまたいで比較可能な成績を返す目的のテストが存在する。このようなテストにおいては、一定期間、比較可能な形の得点を受験者に示し続けることが求められる。一方、日本においては、これまで素点に基づく成績の表示が一般的に行われてきた。素点とは「回答を直接点数化した数値、得点」(日本テスト学会、2007、p.23) と定義される値である。受験者の正答の数を数えることによって得られた値を、そのまま受験者の能力を反映した値とみなすことは、素点を用いた成績表示を行っていることに等しい。しか

し、素点による成績が実施地域をまたいで同一であると言える場面は、受験者に提示した項目が同一であり、また、採点の方法もすべて同一である場合に限られる。したがって、受験者の能力の経時的変化をとらえるようなテストを行う場合は、同一の項目を複数回同一の受験者に提示することを避けなければならない以上、素点以外の方法による共通尺度の構成を行う必要がある。

このような比較可能な尺度を構成する要求は、日本でも多くなりつつある。たとえば、日本の文部科学省による小中学生を対象とした「全国学力調査」を実施するにあたり、倉元 (2008, pp. 219-220) は「共通尺度化の方法」に関する議論を行っている。この中で、日本のテストにおいて素点を成績表示に用い続けているという現状を「理由のない素点主義」と指摘し、「全国学力調査の議論を通じて尺度化の重要性が理解されれば、さまざまな場面で利用されているテストの品質向上につなげていくことができるかもしれない」(p.219) と述べている。また、Arai and Mayekawa (2005) は、2005 年時点において、すでに尺度得点を用いた成績表示をしているテストが日本において存在している（IT パスポート試験、日本留学試験、共用試験（医療系大学間共用試験実施評価機構、2005 年時点においては予備試験段階））ことをテスト関係者へのアンケート調査によって示している。従来、日本において「素点による成績表示」に代表されるような「日本的試験文化」(Arai and Mayekawa, 2005) が存在する中で、このような共通尺度に基づく成績表示を行う試験の事例が増えることによって、どのようにして共通尺度を構成するか、実践的な見地からの研究が必要となってきている。

一般的に、受験者が数千人から数万人に及ぶテストにおいて、新たにテストを立ち上げる団体・機関（以下、「テスト実施機関」と呼ぶ）は、図 1.1 のような段階を経てテストを実施することが必要とされる (日本テスト学会、2010)。図 1.1 にある「予備テスト」（予備試験）は、本試験で出題される前に、項目の特性を把握し、特性が好ましくない項目を排除する目的で実施されるテストを指す。

本論では、同じ内容のテストが毎回繰り返されるようなテストを考える。すなわち、図 1.1 で記した「全体目標と計画作り」から「資料の保存管理」までを 1 回のテスト実施で行い、そのたびに「報告書の作成」を実施し、技術報告の形でテストの実施履歴を記録していく。図 1.1 で重要なのは、毎回の試験で出題された項目のみならず、出題前に実施した予備テストで出題の対象外となった項目も「項目バンク」に入れ、毎回のテスト実施のたびに項目バンクに項目が追加されていく、という点である。「項目バンク」は一種のデータベースであり、項目を一意に識別できるような ID、教科名、領域（単元）名、出題形式、問題文（図表、音声、動画等を含む）の内容、正答、互いにヒントになるため一緒に出してはいけない項目（敵対項目と呼ばれる）、項目統計量、使用履歴など、項目に関する属性を出題するたびに改訂し、項目を新規に作成する際には既存項目一覧として参照する（日本テスト学会、2010、p.75）。項目バンクの存在は、複数の回の成績が比較可能となるテストに必要な条件の一つであるといえる。項目バンクの構築例としては、大学入試センターにおける事例が吉村 (2009,pp.167-189) にある。

しかし、各々のテスト実施回における受験者の得点は、特別な工夫をしなければ比較可能と解釈することは無理がある。ある項目について、正答率が高いからといって、必ずしもその項目が易しかったからであるとは限らない。その回の受験者が総じて成績が良かったために、比較的難易度が

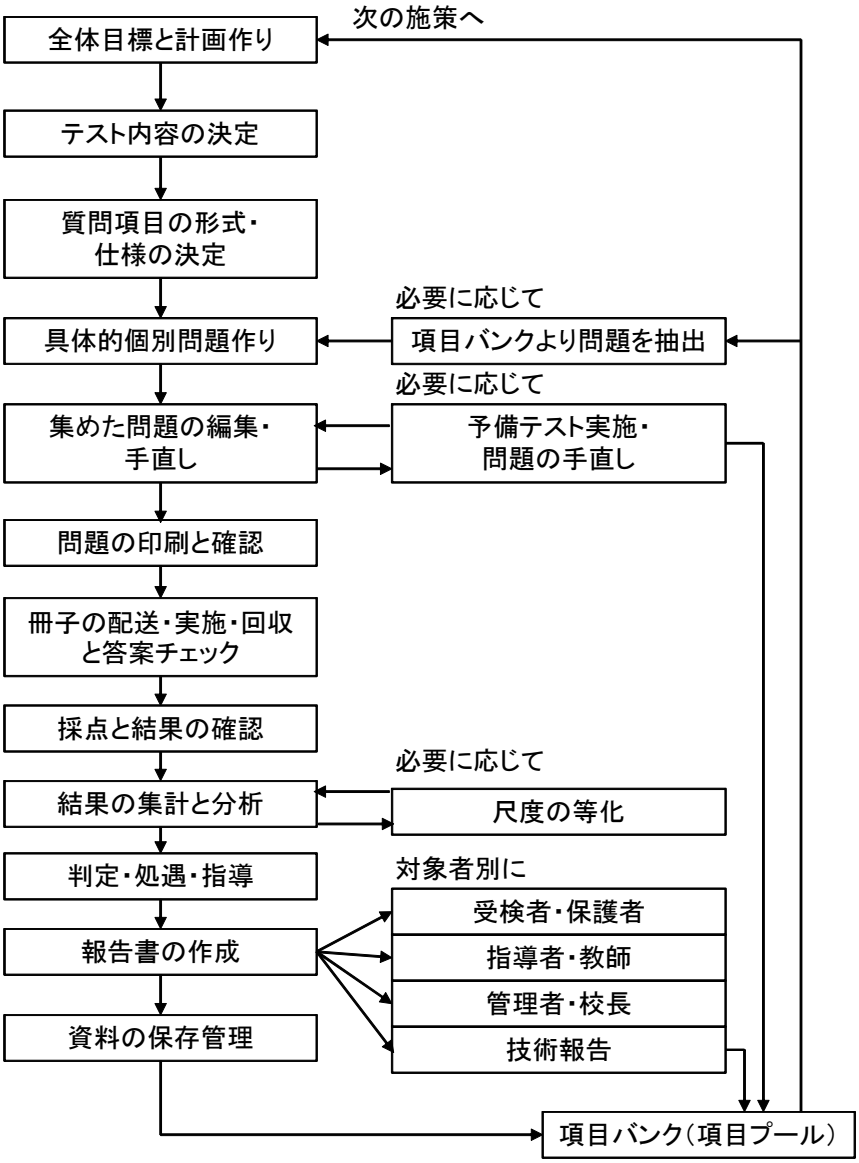

図 1.1　大規模テストの作成・実施における作業手順（日本テスト学会、2010、p.6 を一部改変）

高めであったとしても、正答率が高まる場合がある。したがって、正答率のみを根拠にした形で得点を調整することで、各回の受験者の能力を比較可能な形にすることは、たとえ同一項目を複数の実施回にわたって出題した場合であっても、きわめて強い仮定をおく必要がある。たとえば、「項目の難しさが同等である」ということをプレテストなどの手段で事前に明らかにしておかなければならない。しかし、このようなプレテストを本試験実施前に行い続けることは、現実的ではない。また、同一項目を複数回提示する場合、受験者がテストに対する「対策」をとることが予想され、テスト得点が「どれだけ対策をしたか」を反映した値になってしまうことも予想される。これは、試験の公平性を危うくする要因であるといえる。

また、項目の問う概念が広範囲である場合、その概念のすべての範囲から出題することが理想である。しかし、多数の項目を提示できたとしても、その多数の項目が限られた下位概念を問う項目のみであった場合、その下位概念のみを偶然勉強した者の成績が高くなり、試験の公平さを欠く、といった問題点がある。事前にプレテストを行えば、この問題も克服できるものの、限られた下位概念が共通して測定している「因子」を 1 つ抽出し、その因子の軸、すなわち受験者の「能力」という潜在変数 (latent variable) で受験者がどれだけ勉強したかを位置づけることができれば、教育測定の方法としては妥当性が高いだろう。たとえば、英語の複数の単元を問う項目から、「英語読解能力」という 1 つの因子を抽出し、受験者の「英語読解能力」を問う、といった場面が想定できる。

これらの問題点をまとめると、既存の正答率、あるいは正答の数（素点）からでは、(1)「項目の特性（難易度など)」と「受験者の能力」を分離できない、(2) 「受験者の能力」に関して、測定される概念の抽象化ができない、といった問題点が指摘できる。このような問題点を克服するために、項目反応理論（item response theory; IRT; Lord,1980; 芝、1991; 池田、1994）を応用した方法が用いられるようになった。以下に、IRT に基づくテストの特性について述べる。

1.1.2　2 値データに対する IRT モデル

IRT においては、項目 j $(j = 1, 2, ..., J)$ について、正答確率 P_j と受験者の「能力」という潜在変数 θ との関係を、ある関数 $P_j(\theta)$ によってあらわす。この関数を項目特性関数 (item characteristic function, ICF)、あるいは項目特性曲線 (item characteristic curve, ICC) と呼ぶ。各項目に対し、「正答・非正答」のいずれかの値しかとらないような 2 値データ (dichotomous data) に対する ICC としては、$0 \leq P_j \leq 1$ となる単調増加関数として、ロジスティック曲線が用いられる (Birnbaum, 1968)。ロジスティック曲線の形を決めるモデルとして、主に以下の 3 種が知られている。

■1 パラメタ・ロジスティックモデル (1PL)　1PL は式 1.1 で表される ICF を用いる。

$$P_j(\theta|b_j) = \frac{1}{1 + \exp(-D(\theta - b_j))} \tag{1.1}$$

ここで $D = 1.7$ のとき、θ が全域にわたって正規累積曲線と良い近似になることが知られている。正規累積曲線は、ロジスティック曲線と同様、単調に増加する ICF の一種として、ロジスティック曲線が用いられるようになる前には使われていた。

b_j は項目 j の特性を表すパラメタで、「困難度」「b パラメタ」と称する。困難度は、大きければ大きいほど、ICC の形で見ると変曲点（$\theta = b_j$ となる点）が右、すなわち θ の大きな方にシフトし、小さければ小さいほど変曲点が左にシフトする。このことは、困難度の大きい項目に関しては、より大きな θ を持った受験者でなければ、正答確率が高くならないことを意味する。

■2 パラメタ・ロジスティックモデル (2PL)　2PL では、式 1.2 で表される ICF を用いる。

$$P_j(\theta|a_j, b_j) = \frac{1}{1 + \exp(-Da_j(\theta - b_j))} \tag{1.2}$$

1PL と比較して、a_j なるパラメタが付加されている。これは「識別力」「a パラメタ」と呼ばれるもので、$a_j > 0$ である。識別力が大きければ大きいほど、ICC の変曲点の近傍で、接線の傾きが大きくなり、小さければ小さいほど接線の傾きが小さくなる。よって、識別力の大きな項目は、ICC の変曲点付近の θ を持つ受験者において、少しの θ の増加によって正答確率が大きく高まるのに対し、識別力の小さな項目は、正答確率の変化が小さい。言い換えれば、識別力の大きな項目は、$\theta = b_j$ の点において受験者の能力の違いをよく識別できる項目である、といえる。

1PL、2PL では、$\theta = b_j$ となる点は、式 1.1 および 1.2 から、どちらも $P_j = 0.5$ となることがわかる。したがって、b_j は、正答確率が 0.5 となる θ の値であるとも表現できる。

■3 パラメタ・ロジスティックモデル (3PL)　3PL では、式 1.3 で表される ICF を用いる。

$$P_j(\theta|a_j, b_j, c_j) = c + \frac{1-c}{1+\exp(-Da_j(\theta - b_j))} \tag{1.3}$$

2PL と比較して、3PL では c_j なるパラメタが付加されている。c_j は「あて推量」を表すパラメタで、多肢選択項目などにおいて偶然に正答できる水準を表す。たとえば、4 つの選択肢がある項目に対し、受験者がランダムに解答する場面では、偶然の要素で 4 分の 1 の確率で正答する場合がある。あて推量パラメタは、そのような場合をモデル化したものである。c_j は、確率を表すパラメタであるから、$0 \leq c_j \leq 1$ である。

以上述べた 3 種のうち、テストの測定すべき内容、また受験者の特性を勘案し、最も適切なモデルを用いて、予備試験および本試験の 0-1 データに対して項目パラメタを推定し、また θ を推定する。

1.1.3　IRT モデルと因子分析モデルとの関係

一般的に、IRT モデルでは θ が各受験者における「一般性を持った潜在的な因子」の大きさを表す「因子得点」、項目パラメタが「因子負荷」を構造化した形のパラメタであると見ることもできる。モデルとして、IRT モデルはカテゴリカルなデータにおける因子分析と等価である (Takane and de Leeuw, 1987; 柳井・繁桝・前川・市川、1990、p.145-146)。

2PL の IRT モデルにおいては、項目パラメタは識別力 a_j および困難度 b_j が用いられる。このモデルにおいては、ロジスティック曲線を ICC に用いる前は、正規累積曲線

$$P_j = \Phi(Z_j) = \int_{-\infty}^{Z_j} \phi(t)dt \tag{1.4}$$

ただし

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right) \tag{1.5}$$

を用いたモデル (normal ogive item response model) であり、$Z_j = (\theta - \mu_j)/\sigma_j$ と表される。

Lord(1980) の定式化は、項目 j には「能力」を表す連続的な潜在変数 Γ_j が仮定され、ある受験者にとって閾値定数 γ_j が $\Gamma_j > \gamma_j$ であれば、「正答」という反応が得られ、$\Gamma_j < \gamma_j$ であれば「非

正答」という反応が得られている、とするモデルである。このモデルは、正誤反応の背景として、ある閾値を超えるだけの能力があるかどうかを仮定している。以上の前提のもと、Lord は Γ_j を θ の関数であり、その関数形は 1 因子の因子分析モデル

$$\Gamma_j = \rho_j \theta + \epsilon_j \tag{1.6}$$

であると仮定した。ただし、ρ_j は回帰係数を示す。ここで、Γ_j の θ への回帰直線 $\mu_{j|\theta}$ は、

$$\mu_{j|\theta} = \rho_j \theta \tag{1.7}$$

と書ける。因子分析モデルで考えると、因子パタン係数が、因子スコアを説明変数、データを基準変数とした場合の重回帰分析における偏回帰係数として与えられる (柳井ほか、1990、p.31) ことから、$\mu_{j|\theta}$ をデータ、ρ_j を因子パタン係数（因子負荷)、θ を因子スコアとおきかえれば、式 1.6 は 1 因子の因子分析モデルと等価である。図 1.2 に、以上の関係を示した。

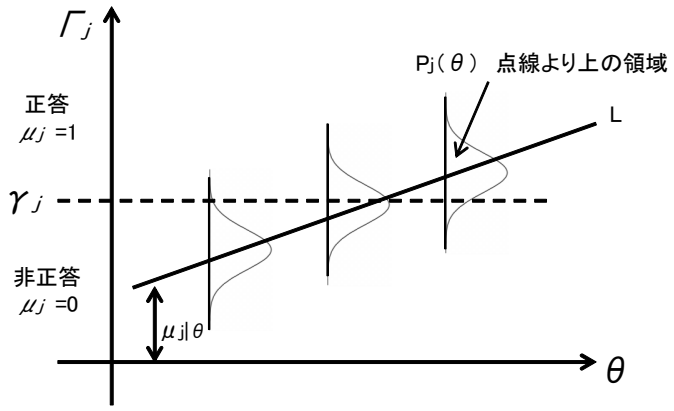

図 1.2　θ の 3 つのレベルにおける Γ_j の条件付き分布。直線 L は Γ_j の回帰直線を表す。村木 (2011,p.43) を一部改変

ここで、Γ_j について、Lord(1980) では (1)Γ_j の θ への回帰関数 $\mu_{j|\theta}$ は線形である、(2)Γ_j の分散 $\sigma^2_{j|\theta}$ は θ の全域にわたって等しい、(3)Γ_j の任意の θ のレベルにおける条件付き分布は正規分布である、という 3 つの仮定をおいた。これらの仮定と、θ の分散を 1 とする仮定（Thurstone,1947）を用いると、Z_j は

$$Z_j = -\frac{\gamma_j - \mu_{j|\theta}}{\sigma_{j|\theta}} \tag{1.8}$$

$$= -\frac{\gamma_j - \rho_j \theta}{\sqrt{1-\rho_j^2}} \tag{1.9}$$

$$= \frac{\rho_j}{\sqrt{1-\rho_j^2}} \left(\theta - \frac{\gamma_j}{\rho_j} \right) \tag{1.10}$$

と表すことができる。したがって、2PL の IRT モデルにおいて、

$$a_j = \frac{\rho_j}{\sqrt{1-\rho_j^2}} \tag{1.11}$$

$$b_j = \frac{\gamma_j}{\rho_j} \tag{1.12}$$

と置き換えれば、式 1.10 は $Z_j = a_j(\theta - b_j)$ と表すことができる。

以上の議論より、因子分析モデルの表現でいうところの因子負荷は、式 1.11 にしたがって IRT における識別力を変換した値と考えることができる。その意味で、識別力パラメタは、因子分析の 1 因子モデルにおける因子負荷とみなすことができる。

1.1.4 IRT に基づくテストの特長

IRT を用いたテストは、「各受験者について、テスト得点 X が真の得点 T と受験者ごとに独立な誤差 E との線形和によって表される」という古典的テスト理論 (classical test theory; CTT) にはない特徴を備えている。IRT に基づくテストでは、各テスト項目に対して得られた各受験者の反応 (順序尺度、あるいは名義尺度に対するカテゴリデータ) より、その共通な変動を代表するような因子を抽出し、その因子の特徴を表すような「項目パラメタ」を各項目に推定、また各受験者に対してはその因子の特徴の大小を反映した「能力値 θ」を推定する。

IRT に基づくテストでは、(1)CTT に基づくテストでは不可能だった、「正誤反応が得られた要因が、受験者の能力に依拠するのか、項目の特性に依拠するのか」を分離した形でテスト結果を解釈することができる、(2)CTT に比べ、正誤反応が得られた背景を数理モデルで表現するうえでの柔軟性が高い、(3)CTT では得点の範囲が 0 から任意の値までという定まった値で与えられるのに対し、IRT では θ の基準を任意に与えることが可能である、といった利点がある。テストの実践場面においては、(1) は「項目パラメタ」と「θ」の解釈を独立して行うことができる、(2) は段階反応モデル (graded response model; Samejima, 1969) や一般化部分採点モデル (generalized partial credit model; Muraki, 1992) などという形でモデルを拡張することで、「正答－非正答」以外の正誤反応データ、たとえば「完全正答－部分正答－非正答」といった順序データなどであっても IRT の枠組みで扱うことが可能となる、(3) は成績判断のための規準となるような「規準集団」(norm group) を設け、規準集団からの相対的な能力の差異を各回のテストを受験した集団において比較することが可能となる、という形で、CTT に基づくテストにはない特長を生かしたテストを実施するためのツールとして IRT が用いられる。

このうち、IRT に基づくテストにおいては、(3) の点で述べた θ の尺度の不定性により、任意の規準集団を定め、そこからの θ のずれを集団ごとに比較するといったことが可能となる。また、IRT に基づくテストの場合、受験者の能力を判断する材料として、CTT に比べて「モデルで説明できる部分」が大きいという特徴がある。CTT、すなわち素点に基づくテストにおいては、尺度化において、受験者の能力を素点の範囲内でしか表現できない。たとえば、100 点満点のテストであれば、0 点から 100 点までの 101 通りの得点という形でしか受験者の能力を表現できない。それに

対し、IRT に基づくテストでは、理論上は「受験者の反応パタン」の数だけ、とりうる θ の値のバリエーションが存在する。たとえば、10 項目からなるテストにおいては、$2^{10} = 1024$ 通りの θ の値が想定できる。このため、IRT を用いたテストは、CTT のテストに比べて、より現実に即した形の得点を返すことが期待できる。

1.1.5 用語の定義

以下の記述において、「フォーム」「アンカー項目」「グループ」「受験者集団」「規準集団」の各用語について、定義を述べる。

■フォーム（テストフォーム）、アンカー項目 1 種類の問題冊子を「フォーム」と呼ぶ。各々のフォームは複数の項目から構成され、それらは部分的に同一である場合もある。複数のフォームにまたがって出題されている項目を「アンカー項目」と呼ぶ。

■グループ（受験者グループ、受験者集団） あるテストにおいて、それぞれ互いに異なる特徴を持つ受験者の群が想定できるとき、それらを「グループ」の違いと定義する。互いに異なる特徴としては、受験者の属性や受験時期などが挙げられる場合もあれば、それぞれ異なるフォームを提示されているという意味で、各フォームに対して各グループを仮定する場合もある。「受験者集団」と表記した場合、その意味としてはほぼ「グループ」と等しい。ただし、「グループ」は、多群 IRT モデルの中で外的基準として与える群の違いに関するモデルの情報を意味するのに対し、「受験者集団」と表記した場合は必ずしもモデル上に限定せず、質的に異なる受験者の一群を指す。

■規準集団 IRT に基づくテストにおいては、θ の不定性により、θ のスケールは任意に決められる。そのため、θ の尺度に何らかの意味付けをしなければ、テストの結果表示としては不適切な場合がある。そこで、ある時点における θ の尺度を「規準集団」と定め、その θ の尺度において、$\theta \sim N(0, 1)$ などという形で規準集団上の θ の平均と標準偏差を固定した上での項目パラメタを求めておき、他の受験者集団における θ は、規準集団上の θ と相対的な値の比較が可能となる形で項目パラメタを求める。この手続きにより、全ての受験者集団の θ の平均および標準偏差は、規準集団を介して比較可能となる。このように、集団同士の θ の尺度を比較可能とするための規準となる集団を「規準集団」と呼ぶ。

1.1.6 IRT に基づくテストのもつ問題点 — テストのリンクと等化

しかしながら、これらの特長を生かしたテストを実施し、実際の成績判断に使用するためには、事前にいくつかの問題を解決する必要がある。特に、資格試験や語学試験において、受験者の技能や能力水準が一定の基準を超えているかどうかを測定する必要がある場合、その基準がテストの実施回をまたいで一定であることが保証されていることが必要となる。そこで、前述した IRT に基づくテストが持つ特長のうち、(3) の特長を生かしたテストが行われている。ただし、規準集団と比較可能な成績を毎回のテストにおいて算出するためには、異なるテスト冊子 (フォーム) のテス

トにおける成績を、同一尺度上の成績に変換する統計的な操作が行われることが前提となる。　この操作のことをリンク (linking) と呼ぶ (日本テスト学会、2007、p.224)。さらに、毎回のテストにおいて、同じ特性を測定している場合のリンク操作を等化 (equating) と呼ぶ (日本テスト学会、2007、p.220)。実際の試験においては、テストで測定される特性は同一であることが一般的であるため、毎回のテストの結果得られた尺度を規準集団上の尺度に等化する操作を行えば、毎回のテストにおける受験者集団の θ の分布が異なっても、共通の尺度上で表現された等化済み θ を用いて、共通尺度上での成績評価を行うことが可能である。

1.1.7　等化の前提

もちろん、等化はいかなるテストに対しても行える手続きではない。また、無制約のもとに行えるものではない。「等化先」（たとえば、規準集団上の尺度）の尺度に「等化元」（たとえば、毎回のテスト）の尺度を何らかの操作によって比較可能であると主張するためには、いくつかの制約が必要であることは自明である。Lord(1980) は、等化の前提として、(1) 等化元と等化先で同じ能力を測定していなければならない、(2) 能力が同じ集団において、等化処理を施した等化元の条件付き得点分布と、等化先の条件付き得点分布が同一でなければならない（公平性の条件)、(3) テスト得点の変換の形が、そのテスト得点が得られた母集団に依存しない、(4) 等化元から等化先への得点の変換は、等化先から等化元への得点の変換に等しい（対称性の条件)、といった点を挙げている (Petersen, Kolen, & Hoover, 1989(前川訳 (1992)、p.340))。これらの点は、特に (2) や (3) については、等化が必要ない場面、すなわち「等化元と等化先が平行テストである」場合が現実には困難であるという点から導かれている。

また、Dorans and Holland (2000) で指摘されている等化の条件としては、(a) 測定対象となる構成概念が同一である、(b) 信頼性が等しい、(c) 対称性が保たれている、(d) どちらのテストを受験しても同等である、(e) 母集団不変、の 5 つが挙げられている (佐藤・柴山、2010; von Davier et al., 2004; 倉元、2011; 条件の表記は佐藤・柴山 (2010) に基づく)。これらを見ると、等化の前提としては「構成概念の同一性」「等信頼性」「公平性」「対称性」「母集団不変」であることがわかる。実際にテストを行った場合、これらの条件を満たしていない場合がほとんどであるものの、たとえば「構成概念の同一性」や「等信頼性」は、あらかじめ項目の測定している構成概念が経験的に同一とされる内容の項目を出題することや、プレテスト（本試験を実施する前に、受験者とは別の集団に本試験用の項目を提示するようなテスト）を実施し、その反応から項目特性や信頼性の推定を行うことで、等化の前提を満たそうとする努力をすることは可能である。また「対称性」のように、等化の技術的な方法を工夫することで達成可能な条件もある。しかしながら、「公平性」や「母集団不変」に関しては、テストを実施する前に予測することがきわめて困難な条件も含まれている。これらの前提が満たされているかどうかの判断は、テスト得点をどのような目的で使用するかに依存する。テストの目的に関わる事項であるので、本論では詳しく立ち入らない。

本論では、テストを実施しようとする機関（以下、「テスト実施機関」と表記）は、Dorans and Holland の等化 5 条件を、できるかぎり満たそうと努力する場合を考える。その上で、「リンク」

ではなく、「等化」を前提としたテストデザインを提案し、その方法の評価を行う。ただし、等化 5 条件のうち、「構成概念の同一性」に関しては、予備試験段階で企図された構成概念と、本試験で意図した構成概念が互いに異なる場合について、等化の条件が満たされない場合に、等化を行った場合について取り上げる。

1.1.8　大規模テスト

本論では、IRT に基づくテストで、複数の回にわたって共通の尺度を構成するようなテストの場面を想定し、各々の回に 1000 名規模の受験者が存在し、各回の受験者に対応したフォームが複数回出題されるようなテスト場面を想定し、複数の回のテストが一定の期間ごとに繰り返されるようなテストにおける共通尺度の構成をとりあげる。このようなテストを、本論では大規模テストと呼ぶことにする。

1.2　テストデザイン

等化を行うテストにおいては、事前にどのような計画（デザイン）で受験者の得点を相互に比較可能とするかを決める必要がある。大規模テストにおいては、その目的から、採用できるデザインに制約が課せられる。本節では、等化を行うテストにおける一般的なテストデザインを示す。

1.2.1　共通項目デザイン

いま、受験者集団 1 と受験者集団 2 に対して、内容が異なるテストフォーム X および Y を出題し、集団 1 と 2 の成績を相互に比較可能な尺度に乗せる場面を考える。この場合、図 1.3 で示すように、フォーム X と Y の一部を共通な項目とすれば、共通項目に対する正誤反応を手掛かりに等化することが可能である。ここで IRT を用いることとすれば、項目分析の仮定で、項目特性と受験者集団の特徴を別個に推定することが可能である。よって、共通項目によって、集団 1 と集団 2 との能力差が推定でき、同様に、共通項目について、集団 1 と集団 2 の両方における項目特性が推定できることとなる。さらに、フォーム X の集団 1 において出題された部分の項目特性、およびフォーム Y の集団 2 において出題された部分の項目特性が、共通項目上での項目特性と比較可能な形で推定される。これによって、両集団を またいで比較可能な形での尺度化ができることとなる。この方法を「共通項目デザイン」と呼ぶ。共通受験者デザインにおいては、各フォームに対応する受験者グループは、互いに θ の分布が異なるという仮定をおく。

1.2.2　共通受験者デザイン

共通項目デザインとは異なり、同一の受験者に異なる複数の種類のテストフォームを同時に受けさせることで、異なる種類のフォームを受験した複数の集団間で共通の尺度を構成することが可能である。図 1.4 に示すこの方法を「共通受験者デザイン」と呼ぶ。共通項目デザインとは異

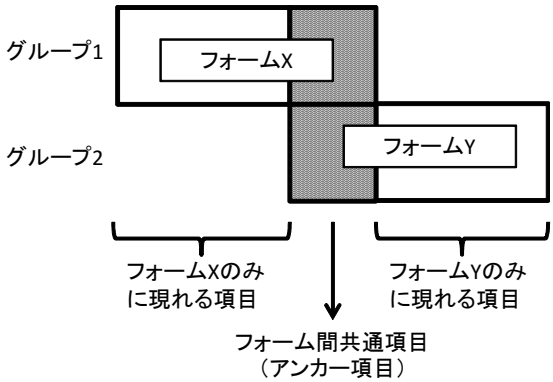

図 1.3　共通項目デザインの例。異なる受験者集団 1 と 2 の間で共通な尺度を構成するために、2 種のテストフォーム X と Y に共通項目を設け、IRT を用いた尺度化を行う

なり、IRT を用いなくとも、たとえば素点を偏差値やパーセンタイル順位に変換することで、異なる集団間で相互に比較可能な得点を推定することができる (素点を用いた等化、リンクに関する議論、および IRT を用いた等化との比較は Muraki, Hombo and Lee (2000) を参照) が、IRT を用いても等化が可能である。IRT を用いた共通受験者デザインの方法に関する研究は、たとえば Ogasawasa(2001) がある。

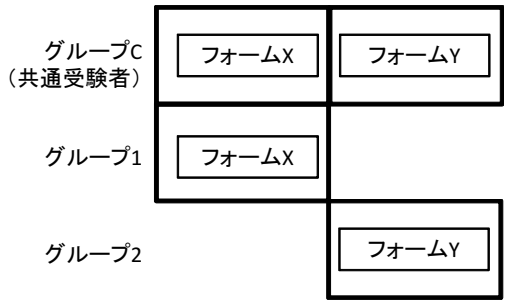

図 1.4　共通受験者デザインの例。異なる受験者集団 1 と 2 の間で共通な尺度を構成するために、2 種のテストフォーム X と Y を同時に解答する集団 C を設ける

1.2.3　利点と欠点

共通項目デザインには、共通受験者デザインで必要な「複数あるフォームにある全項目を受験する」受験者集団（図 1.4 の「受験者集団 C」）がいらない、という利点がある。仮に、1 フォームに 30 項目が含まれ、等化したいフォームの数が 10 である場合、受験者集団 C は 300 項目を解答しなければならない。人間の注意力や体力を考えると、300 項目をすべて解答するような作業は、時間をかけて行った場合であっても、非現実的であるといわなければならない。共通項目デザインの場合、各フォームに、部分的に共通項目を含ませることによって、等化すべきフォーム（集団）の

数が増えたとしても、現実的に実施可能なテストの範疇でおさめることができる。

一方で、共通項目デザインにおいては、IRT による等化が必要である。IRT に基づくテストにおける等化は、これまで多くの方法が提案されてきている（後の節で詳述する）ものの、いずれの方法をとるかによって等化の結果が変わってくる。結果が等化の方法に依存して変わる、という点は、「等化の前提」の上で、「対称性」を乱す場合には好ましくない傾向といえる。しかし、対称性を満たすような等化であったとしても、等化の方法によって少しずつ結果が異なる。したがって、等化の方法を吟味する上で、新たな等化方法の評価尺度が必要となる。

1.3 IRT に基づく尺度化、および等化

本節では、IRT に基づくテストにおいて、複数のテストフォームに対し、θ の分布が異なると仮定される複数の受験者グループが解答している場面において、どのように等化後の項目パラメタを求めるかについて、「尺度化」と「等化」の 2 つのステップについて述べる。

1.3.1 尺度化

一般に、IRT に基づく試験においては、まず正誤反応から項目パラメタを推定する。この過程で、受験者をまたいだ一般的な θ の尺度上での、項目の困難度や識別力、あて推量パラメタといった「項目パラメタ」が推定される。これらの値を推定する過程は、受験者の能力を判断するための材料として各項目のパラメタを推定することから、受験者の能力に関する尺度を構成していることに他ならない。その意味で、「尺度化」と呼ぶ。

IRT における尺度化は、ICF にどのモデルを使用するかによって異なる意味付けとなる。すなわち、1PL においては、項目は「困難度」という特徴において記述されるというモデルであるのに対し、2PL では「困難度」に加えて「識別力」というパラメタが加わる。仮定するモデルは、これから尺度化したい尺度の特徴や、尺度の使用目的などによって決定される。当然、パラメタの種類が多ければ多いほど、データを豊かに表現できる。その意味では、3PL は最も優れたモデルであるといえるかもしれない。しかし、3PL の安定した推定においては、多くの人数が必要 (豊田 (2012, p.73) によると、経験的目安として、1PL で 100 人、2PL で 300 人、3PL では 1000 人以上の受験者が必要という。ただし、項目パラメタを求めるためには、受験者の数ではなく、回答パタンのバリエーションの数が増えることが必要である。第 4 章で詳述) であることから、テスト実施機関は毎回の受験者数がどの程度になるかを予測したうえで、モデルを選択することが求められる。

1.3.2 等化の際のパラメタ推定方法

前述のように、等化とは、受験者について同一の内容を問うテストにおいて、異なる尺度を相互に比較可能な尺度の上に乗せるという操作であった。この場合、「フォームごとに異なる θ の分布を仮定する」という前提を置くと、g 番目の受験者グループにおける θ の分布に関して、(1) フォームをまたいだ受験者全体において θ の分布を正規分布 $N(0,1)$ とおき、その中で各グループ

の θ について $N(\mu_g, \sigma_g)$ とおく、(2) あるフォームにおいて、θ の分布を $N(0,1)$ とおき、その他のフォームに対応する受験者グループ g に関して $N(\mu_g, \sigma_g)$ とおく、という 2 通りの仮定ができる。(2) の場合を用いると、θ $N(0,1)$ とおいたグループを「規準集団」とみなせば、他のグループにおいて、相対的な規準集団との $theta$ の分布の比較ができることになる。このように、グループごとに異なる θ の分布を仮定する共通項目デザインを、「共通項目非等価集団法」(common-item nonequivalent groups design) と呼ぶ。

共通項目非等価集団法に基づく等化方法はこれまでさまざまな種類が提案されているが、それらは大きく分けて「同時推定」(concurrent calibration) と「個別推定」(separate calibration)、それに「項目パラメタ固定法」(fixed common item parameters; FCIP; Li, Griffith, and Tam, 1997) に大別される (Hanson and Béguin, 2002; Hu, Rogers, & Vukmirovic, 2008; Arai and Mayekawa, 2011)。

以下に、図 1.3 の場合を用いて、それぞれのパラメタ推定方法について述べる。いずれの場合も、2 つのグループがあり、グループ 1 が規準集団、グループ 2 が等化されるべき受験者集団で、それぞれのグループに対応するフォーム X・フォーム Y を提示している。ただし、フォーム X とフォーム Y の間には、アンカー項目が含まれている。なお、本節以降において、項目 j の項目パラメタをベクトル $\boldsymbol{\xi}_j$ で表す。モデルが 2PL の場合、$\boldsymbol{\xi}_j = (a_j, b_j)$ であり、3PL の場合は $\boldsymbol{\xi}_j = (a_j, b_j, c_j)$ である。また、行列 $\boldsymbol{\xi}$ は、$\boldsymbol{\xi}_j$ を $j = 1, 2, ..., J$ 項目分まとめたものを示す。

■同時推定　まず、フォーム X とフォーム Y のそれぞれについて、「1 行が 1 人分の正誤反応データ」となるように並べる。このとき、同じ列に並んだ 0-1 データは、同一の項目に対する正誤反応となるように並べる。アンカー項目の部分はグループ 1 および 2 の両方において正誤反応が得られているが、それ以外の項目に関しては、フォーム X のみに出現する項目に関してはグループ 2 の受験者において欠損となっており、フォーム Y のみに出現している項目に関してはグループ 1 の受験者において欠損となっている。図 1.5 に、データの概略を示す。

次に、先に用意したデータに対し、多群 (Multi group) を仮定した IRT モデル (multiple group IRT; Bock and Zimowski, 1996) を適用し、項目パラメタの推定を行う。多群 IRT モデルについては、Mislevy(1984) や前川 (1991, pp.107-114,117-118) などが方法を提案しており、計算を行うためのソフトウェアも用意されている（たとえば、`BILOG-MG`(Zimowski, Muraki, Mislevy and Bock, 2003) や `ICL`(Hanson, 2002)、`PARSCALE`(Muraki and Bock, 2003) など)。多群 IRT モデルを適用した場合、受験者が母集団上で互いに異なる θ の分布をもち、それらが同じテストを受験したと仮定される。したがって、データとして、各受験者がどのフォームを受験したかという情報が必要である。同時推定の場合は、そのデータは、各受験者がどのフォームを受験したかという情報から得る。多群 IRT モデルの場合、得られるパラメタは、各グループをまたいだ共通の項目パラメタ、すなわち等化済みの項目パラメタのセットと、各グループにおける θ の分布に関する情報である。θ の分布に関する情報は、連続量の θ をいくつかの求積点に区切って離散化し、各求積点における確率の値を多項分布の形で表すこともできるが、実際には分布の形の解釈のしやすさから、グループごとに単一の正規分布を仮定し、グループごとに平均と標準偏差を求める。

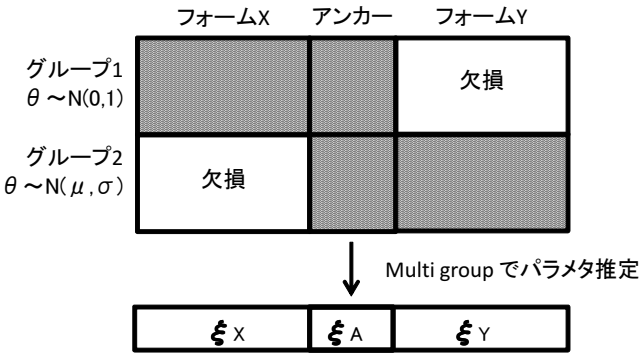

図 1.5　共通項目デザインの場合の、同時推定による等化

■個別推定　フォーム X とフォーム Y でそれぞれについて項目パラメタを推定し、$\boldsymbol{\xi}_X$ および $\boldsymbol{\xi}_Y$ を得る。次に、$\boldsymbol{\xi}_X$ と $\boldsymbol{\xi}_Y$ の中に含まれるアンカー項目のパラメタ $\boldsymbol{\xi}_{XA}$ と $\boldsymbol{\xi}_{YA}$ を手掛かりに、$\boldsymbol{\xi}_X$ を規準として $\boldsymbol{\xi}_Y$ を等化する。この際、項目パラメタの等化の方法として、いくつかの方法が提案されている（後述）。図 1.6 に、方法の概略を示す。

この方法は、項目パラメタの等化を伴う方法で、仮に理論通りであれば $\boldsymbol{\xi}_{XA} = \boldsymbol{\xi}_{YA}$ であるところ、実際には異なる項目パラメタが得られたので、$\boldsymbol{\xi}_{XA}$ の尺度に（$\boldsymbol{\xi}_{YA}$ を含む）$\boldsymbol{\xi}_Y$ を乗せる、という操作を行う必要がある。すると、等化後の $\boldsymbol{\xi}_{YA}$ が完全に $\boldsymbol{\xi}_{XA}$ と一致しない、という場合が生じる。したがって、$\boldsymbol{\xi}_{XA}$ と「等化後の $\boldsymbol{\xi}_{YA}$」のいずれを「等化後のパラメタ」と扱うか、という問題が生じうる。規準集団への等化、という観点で見れば、規準集団がフォーム X を解いたグループ 1 である以上、$\boldsymbol{\xi}_{XA}$ をもって等化後のパラメタとするのが自然である。しかしこの場合、アンカー項目においては、等化後の項目パラメタの情報にグループ 2 の正誤反応情報が反映されていないという問題も指摘できる。

■項目パラメタ固定法 (FCIP 法)　図 1.7 に、方法の概略を示す。まず、フォーム X において項目パラメタを推定し、$\boldsymbol{\xi}_X$ を得る。次に、フォーム Y において項目パラメタを推定する。この際に、フォーム X におけるアンカー項目のパラメタ $\boldsymbol{\xi}_{XA}$ が、あらかじめ既知であるというモデルのもとで、$\boldsymbol{\xi}_Y$ を推定する。実際には、$\boldsymbol{\xi}_Y$ の推定場面で、アンカー項目 $\boldsymbol{\xi}_{YA}$ は $\boldsymbol{\xi}_{XA}$ であると固定して推定するので、個別推定で見られるような、$\boldsymbol{\xi}_{XA}$ と等化後の $\boldsymbol{\xi}_{YA}$ が異なるという問題は生じない。しかし、$\boldsymbol{\xi}_Y$ の推定で、$\boldsymbol{\xi}_{XA}$ があまりにもグループ 2 の受験者の尺度からかけ離れている場合、推定が不安定となり、項目パラメタの推定ができない、といった事態が考えうる。BILOG-MG による FCIP 推定は DeMars and Jurich (2012) を参照のこと。

1.3.3　共通項目のパラメタを用いた等化の方法

個別推定を行う際に、複数の異なる項目パラメタのセットを規準集団上の項目パラメタと比較可能にする操作が必要である。この方法においては、「困難度等化法」(equated *b*s method)「特性曲

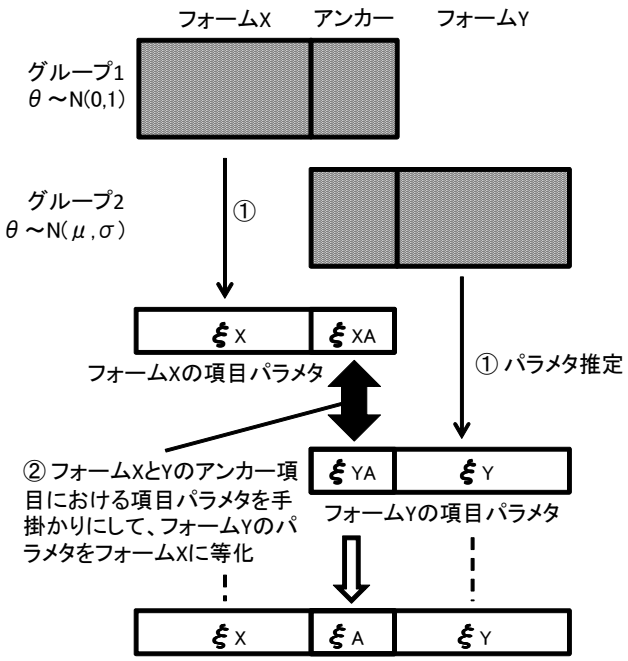

図 1.6　共通項目デザインの場合の、個別推定による等化

線変換法」(characteristic curve transformation method) に大別することができ、それぞれにいくつかの方法が提案されている（村木 (2010, p.110) や Petersen et al(1989); 前川訳 (1991,p.359) においては「同時尺度調整法」として「同時推定」が、また「困難度固定法」として「FCIP 法」が紹介されている)。さらに、特性曲線変換法の一種で、複数のフォームについてまとめて等化を行う方法として calr 法 (前川、1991、p.119-122; Arai and Mayekawa, 2011, Appendix) がある。

■困難度等化法　「困難度等化法」においては、等化前と等化後において、線形等化法に基づく項目パラメタの変換を行う。線形等化法とは、テスト X とテスト Y という 2 つのテストを同一受験者が受験した場合を考え、2PL または 3PL において、困難度の平均が 0、標準偏差 1 とする。ここで、テスト X における θ_x の平均を μ_{θ_x}、標準偏差を σ_{θ_x}、テスト Y における θ_y の平均を μ_{θ_y}、標準偏差を σ_{θ_y} とおくと

$$\frac{\theta_x - \mu_{\theta_x}}{\sigma_{\theta_x}} = \frac{\theta_y - \mu_{\theta_y}}{\sigma_{\theta_y}} \tag{1.13}$$

との関係から、

$$\theta_y = \frac{\sigma_{\theta_y}}{\sigma_{\theta_x}}\theta_x + \left(\mu_{\theta_y} - \frac{\sigma_{\theta_y}}{\sigma_{\theta_x}}\mu_{\theta_x}\right) \tag{1.14}$$

$$= k\theta_x + l \tag{1.15}$$

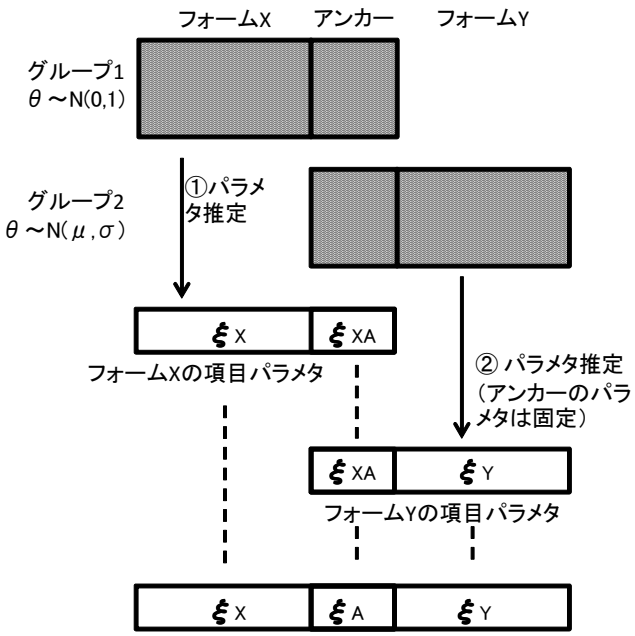

図 1.7　共通項目デザインの場合の、FCIP による等化

と書くことができ、等化係数 (k, l) を用いて θ_y の尺度上に θ_x を乗せることができるとの考えに基づいた方法である。この関係を項目パラメタの場合に当てはめると、

$$b_y = kb_x + l \tag{1.16}$$

$$a_y = \frac{a_x}{k} \tag{1.17}$$

$$c_y = c_x \tag{1.18}$$

となる。以上の関係を用いて、テスト X における困難度の平均を μ_{b_x}、テスト Y における困難度の平均を μ_{b_y}、テスト X における困難度の標準偏差を σ_{b_x}、テスト Y における困難度の標準偏差を σ_{b_y} とおくと、式 1.16 から 1.18 において、

$$k = \frac{\sigma_{b_y}}{\sigma_{b_x}} \tag{1.19}$$

$$l = \mu_{b_y} - k\mu_{b_x} \tag{1.20}$$

なる (k, l) を用いる方法を Marco(1977) は mean/sigma 法と呼んだ。それに対して、テスト X における識別力の平均を μ_{a_x}、テスト Y における識別力の平均を μ_{a_y} とおき、

$$k = \frac{\mu_{a_x}}{\mu_{a_y}} \tag{1.21}$$

$$l = \mu_{b_y} - k\mu_{b_x} \tag{1.22}$$

とする方法が Loyd and Hoover (1980) によって提案されており、これは mean/mean 法と呼ばれている。一般的に識別力パラメタは推定が不安定なため、μ_{a_x} や μ_{a_y} を用いない mean/sigma 法

がよく使われている (村木、2011、p.111)。

■特性曲線変換法　困難度等化法においては、主として困難度の平均や標準偏差といった情報のみが等化の際に用いられる。しかし、仮に困難度の推定値に極端な値が少数含まれていた場合、それに伴う等化係数 (k, l) の値、および等化後の項目パラメタの値が、その「はずれ値」に引きずられてしまう恐れがある。それに対し、特性曲線変換法においては、アンカー項目において 2 つのグループにおいて推定された項目パラメタの値のセットを用いて、規準となるグループにおけるアンカー項目の ICC に近くなるように、もう一方のグループの ICC を変換するような等化係数を求める方法である。

テスト X のみで得られた項目パラメタを $P_j^X(\theta; a_j^X, b_j^X, c_j^X)$ とおき、テスト Y のみで得られた項目パラメタを $P_j^Y(\theta; a_j^Y, b_j^Y, c_j^Y)$ とおく。X と Y のアンカー項目群 A において、ICC の「差」の値 $DICC$ を定義することができる。以下の 2 つの場合

$$DICC_H = \sum_{j \subset A} \left(P_j^X(\theta; a_j^X, b_j^X, c_j^X) - P_j^Y(\theta; a_j^Y, b_j^Y, c_j^Y) \right)^2 \tag{1.23}$$

$$DICC_{SL} = \left(\sum_{j \subset A} P_j^X(\theta; a_j^X, b_j^X, c_j^X) - \sum_{j \subset A} P_j^Y(\theta; a_j^Y, b_j^Y, c_j^Y) \right)^2 \tag{1.24}$$

$$\tag{1.25}$$

において、式 1.23 のように $DICC$ を定義した場合を Haebara 法 (Haebara,1980)、式 1.24 と定義した場合を Stocking-Lord 法 (Stocking and Lord, 1983) と呼ぶ。ただし、$j \subset A$ は、項目 j がアンカー項目群 A に含まれている場合、を示す。

■calr 法　以下の説明は Arai and Mayekawa(2011,Appendix) による。

等化済みの項目パラメタのセットと、グループごとに別々の等化係数を同時に推定する方法である。g 番目のグループ $(g = 1, 2, ...G)$ における θ の分布 $\theta^{(g)}$ および項目パラメタ $(a_j^{(g)}, b_j^{(g)}, c_j^{(g)})$ について、共通尺度上での θ および項目パラメタ (a_j, b_j, c_j) への変換を以下のように施すことを考える。

$$\theta^{(g)} = u_g + v_g\theta, \;\; a_j^{(g)} = \frac{1}{v_g} a_j, \;\; b_j^{(g)} = u_g + v_g b_j, \;\; c_j^{(g)} = c_j \tag{1.26}$$

ただし、u_g および v_g はグループ g に対する等化係数を表す。また、各々の $\theta^{(g)}$ から共通尺度上の θ への変換、すなわち

$$\theta = q_g + r_g\theta^{(g)}, \;\; a_j = v_g a_j^{(g)}, \;\; b_j = q_g + r_g b_j^{(g)}, \;\; c_j = c_j^{(g)} \tag{1.27}$$

に関しては、

$$q_g = -\frac{u_g}{v_g}, \;\; r_g = \frac{1}{v_g} \tag{1.28}$$

という逆変換用の等化係数を用いる。

ここで、以下の最小二乗基準 RSS_{prob} を定める。

$$RSS_{prob} = \sum_{g=1}^{G} \sum_{j \subset g} \int_{\Theta_g} \left(P(\theta|\hat{a}_j^{(g)}, \hat{b}_j^{(g)}, \hat{c}_j^{(g)}) - P(q_g + r_g\theta|a_j, b_j.c_j) \right)^2 h_g(\theta) d\theta \quad (1.29)$$

ここで $P(\cdot|\cdot)$ はモデルの式であり、1PL の場合は式 1.1、2PL の場合は式 1.2、3PL の場合は式 1.3 を用いる。w_A, w_B, w_C はパラメタに対する重み、$h_g(\theta)$ は g グループにおける能力値の分布関数で、各グループについて N(0,1) に比例するとおく。また、$j \subset g$ は、項目 j がグループ g に提示されている場合のみ、ということを表す。

式 1.29 を最小にする項目パラメタの推定値 $(\hat{a}_j^{(g)}, \hat{b}_j^{(g)}, \hat{c}_j^{(g)})$ および等化係数の推定値 $(\hat{q}_g, \hat{r}_g)$ を求めるためには、交互最小二乗法 (alternative least square method; ALS) を用いればよい。すなわち、まず項目パラメタに関する最適化を行い、その最適化時点での $(\hat{a}_j^{(g)}, \hat{b}_j^{(g)}, \hat{c}_j^{(g)})$ を所与として RSS_{prob} の評価を行い、次に等化係数に関する最適化を行い、その最適化時点での $(\hat{q}_g, \hat{r}_g)$ を用いて RSS_{prob} を評価し、前回の反復における RSS_{prob} とほとんど変わらなければ収束したとみなし、そうでなければ再び項目パラメタのの最適化を行う、という手順を繰り返す。また、数値計算の上では、式 1.29 中の θ に関する積分について、θ を適当な区間で離散化し、和で置き換えることが必要である。

1.3.4 垂直等化と水平等化

IRT を用いたテストにおいては、等化の目的、関心が「異なるテストを用いて、あらかじめレベルが異なるとわかっている学習者間の成績を比較する」場合と、「毎回のテストで異なるフォームを提示し、それらの間で成績を比較可能にする」場合に分かれる。前者を「垂直等化」(vertical equating)、後者を「水平等化」(horizontal equating) と呼ぶ (渡辺・野口、1999)。

垂直等化の例としては、芝・野口 (1982) の日本語語彙理解力に関する 10 年にわたる経年調査結果を、IRT を用いて単一の尺度に等化した例が挙げられる（芝、1991、p.131-140)。この事例では、小学校 1 年から高校 2 年までの 10 の学力集団を仮定し、それらの受験者に共通項目デザインによる等化デザインを適用し、語彙理解力を単一の尺度で表すことが可能となっている。垂直等化の場合、異なるテストに現れる項目が受験者のどのレベルに相当するのかといった「項目の特徴」（主として「困難度」）に関心がある場合や、「学年が上がった」「学習の成果があった」といったような、受験者の能力が学習や履修といった「処置」によってどの程度「効果」があったということを議論する場合に用いられる。

一方、水平等化の例としては、先に述べた規準集団への等化が挙げられる。また、吉村・荘島・杉野・野澤・清水・齋藤・根岸・岡部・フレイザー (2005) は、1990 年より 2004 年までに出題された大学入試センター試験本試験の「英語」の項目を大学 1 年生 424 名に提示した正誤反応を用いて、IRT に基づく共通受験者デザインにて等化を行っている。等化された尺度は、1990 年で出題された項目の特性値を規準とする形で表されており、1997 年以後で英語学力の特性値が大きく低下していることが見出された。そのほか、斉田 (2003) などが、水平等化の実践例として挙げられる。

水平等化と、垂直等化の両方を同時に行う場合もある。熊谷・山口・小林・別府・脇田・野口 (2007) は、1995 年度から 2005 年度にわたり、年間に複数回行われた英語のテストの正誤反応から、受験者の英語学力の年度間比較、および年度内の「学習前–学習後」の比較を行っている。前者は水平等化、後者は垂直等化を、いずれも IRT に基づく形で行っている。また、Nakamura and Mitsunaga (2011) は、2006 年度から 2010 年度に行われた大学 1 年次に対して学期初めに行われるプレースメントテスト（習熟度別クラスに振り分けるためのテスト）および学期末に行われる確認テスト（次年度にどのクラスに進むべきかを指示するためのテスト）の正誤反応データを IRT に基づく共通項目デザインにより等化を行っている。このテストにおいても、プレースメントテストと確認テストの間で学力がどのように変化したかを垂直等化の結果から、また 2006 年より 2010 年の学力の変化、特に入学時点でのプレースメントテストにおける学力の変化を水平等化の結果として利用している。このほか、斉田・柳川 (2011) も、水平等化と垂直等化を同時に行うデザインである。

1.4　等化を伴う大規模テストとその制約

以上述べたように、単に「等化」を行う、と言っても、その方法には多数のオプションがある。また、等化の方法は、テストのデザインと関連して決定されるべきものであることが分かる。

IRT に基づくテストのデザインは、テストの仕様によって柔軟に設定される。たとえば、1 回に複数のフォームを異なるグループに提示し、1 回の実施の中で共通項目非等価集団デザインによるテストを実施しつつ、複数のフォームで得られた θ の尺度を規準集団に等化するテストもあるし、毎年 2 回のテストの間だけで共通な尺度を得られさえすればよいようなテストもある。

大規模試験においては、資格試験や語学試験など、受験者の能力の水準において、主に水平等化を行い、規準集団と比較可能な形の尺度にあわせる、という操作を、試験実施のたびに行い続けるようなテスト実施法をとれば、一定数の期間にわたって、規準集団と比較可能な尺度を受験者に提供し続けることが可能である。そのためには、図 1.1 の流れで、「予備試験」を実施し、あらかじめ規準集団上で項目パラメタ既知の項目を項目バンクに入れておくことが必要である。毎回の試験においては、項目バンク上の項目を受験者に提示し、規準集団上での尺度の等化の手がかりとする。

しかし、実際のテスト場面を考えると、以下のような制約があることがわかる。

1.4.1　作題体制上の制約

テスト実施機関がテストの全体計画を立てる段階では、テストに関する作題体制をどのように整えるかを計画することが必要である。しかし、作題の体制が整ったとしても、安定的に多くの項目を作り続けることは困難を伴う作業になりうる。毎回実施するテストが 1 つの因子を測定しているという前提があるため、実施のたびに 1 因子の内容が当初の計画通りか、言い換えれば因子妥当性があるかをチェックする必要がある。そのためには、問題文を作成し、問題冊子を編集する役割以外に、問題文が出題者の意図通りに機能するかを事前にチェックする役割を設ける必要がある。さ

らに、実際に受験者に問題冊子を提示する前に、少数の「受験者と似た属性をもつ人」（たとえば、大学入試のテストにおいては、大学 1 年次生）に問題冊子を提示し、意見を求めるなどの試みが必要である。このように、問題冊子を作成するうえでは、多くの人的リソースが必要であるので、予備試験の前に現実的な問題作成量を見積もっておくことが必要である。

1.4.2　テストデザイン上の制約

規準集団を仮定し、その規準集団に毎回の尺度を乗せるテストの場合、第 1 回試験実施前に、規準集団上で項目パラメタが既知の項目を多く項目バンクに入れておく必要がある。テスト実施機関にとっては、予備試験を行う労力や、多数の項目を第 1 回試験前に準備しなければならない労力など、事前準備に多くの作業が発生し、しかもそれらの作業の質が、本試験の質（信頼性や妥当性）に影響するという点で、負担が大きい作業となることが予想される。毎回のテストにおいては、規準集団と共通の項目をアンカー項目として提示しつつ、規準集団のアンカー項目も提示する。仮に、同一受験者が連続した複数回を受験することを許せば、隣接した実施回の相互間で同一の項目を提示することは、テストの公平性の観点から、避けなければならない。よって、テストの実施のたびに、規準集団上でのアンカー項目が「枯渇」（受験者にとって提示済みとなり、再出題に適さない項目となる）していくことになるので、第 1 回試験前に十分な数の項目を項目バンクに入れる必要がある。

「枯渇」を防ぐためには、何らかの形で「規準集団上で項目パラメタ既知の項目」を補充しなければならない。予備試験を本試験と並行して行うことが一つの方法であるが、本試験と別のテストを実施し続けるような手間をかけることは現実的ではない。

また、等化を行う際、共通受験者デザインは、受験者に多くの負担をかける原因となりうる。多数のフォームの間で共通受験者デザインを用いることは、現実的ではない。したがって、テストデザインの工夫をするうえでは、共通項目非等価集団デザインに関する等化が主となる。しかしながら、共通項目非等価集団デザインの等化については、「規準集団にいくつかのフォームの項目パラメタを等化する」という場面の研究が多く、前述のようなテストデザイン上の問題点を克服するようなデザインにおいて、その結果が援用できるかどうかは不明である。

1.4.3　テストの本文、および出題ルールを秘匿する必要性

項目バンクを設け、そこから同一の項目を複数フォームで再出題するような場合、テストで受験者に提示する場面以外において、項目バンクの中身、特に項目の文面を秘匿しておく必要がある。試験対策と称して過去問（項目バンクの中身）を練習問題として使われると、項目バンクの内容を知り得た受験者とそうでない受験者の間で、公平性がなくなるという事態になりかねない。そうした公平性の問題を考え、項目バンクの管理には特別な注意を要する。

また、同一受験者が異なる実施回を受験可能としたときに、異なる実施回で出題されたフォーム間で共通の項目が同一受験者で 2 度提示される可能性がある。この可能性を排除するには、将来行

われる試験に用いるフォームについて、同一の項目をアンカーに使わないようにするような計画を立てた上で項目を選定する必要がある。また、定められた期間（6 か月や 1 年)、同一受験者の再受験を禁止し、その間に出されるフォームの間では共通の項目を含んでよい、とするルールを設ける場合もある。いずれにしても、これらの計画によって、項目バンクの中にある項目をいつ提示するか、という点について制約が生じる。また、公平性の観点から、これらのルールを公表しないという方針をとることが望ましいと考える。

1.4.4　計算の可能性に関する制約

テスト実施機関が大規模試験を実施する場面で IRT モデルを使用する場合、E-M アルゴリズムによる最尤法を用いて項目パラメタを推定するのが、現実的な選択肢であろう。なぜなら、テスト実施機関においては、項目パラメタの推定は「ブラックボックス」となっている単一のプログラムによって行い、毎回同一のプログラムを使用することで、試験の公平性を確保しようと考えるためである。大学入試センター試験のように、毎年同じルールに基づいて採点するのと同じ考えで、採点後の尺度化・等化・尺度得点の算出処理までもが毎回の試験で同一であることが、テスト実施機関にとって自然な考えである。20 年ほど前より、MCMC(Markov chain Monte Carlo) 法、特に Gibbs sampler や Metropolis-Hastings within Gibbs 法によるサンプリングを用いた項目パラメタのベイズ推定を行う方法が紹介され (たとえば、Baker and Kim, 2004, pp.295-312 や Shigemasu and Nakamura, 1996、Patz and Junker, 1999 など) モデルの拡張の柔軟性から研究者の間では普及しつつある。しかし、テスト実施機関にとっては、乱数発生に依存した推定を伴う手続きは、受け入れがたいと考えるのも無理はないだろう。

また、E-M アルゴリズムを行うソフトウェアは数種知られているが、本研究においては BILOG-MG 3 を取り上げる。BILOG-MG 3 は多群 IRT モデルの推定が可能で、推定に関するオプションが豊富で取り扱いやすい、といった利点に加え、多くの研究で利用されているという点が挙げられる (Arai and Mayekawa (2011) や Hanson and Béguin(2002)、張 (2009) などで使用されている)。しかしながら、どのようなソフトウェアを利用する場合であっても、多数の項目が複雑なフォーム（グループ）に分かれて配置されているというデータに対しては、メモリの制約などにより、推定できなくなる場合も考えうる。本研究においては、経年的に本試験を繰り返すテスト場面を取り扱うため、実際に等化が可能であるようなデザインが提案されたとしても、1 回あたり数十のフォームであった時、数回のテスト実施で計算不可能になるとすれば、現実のテストでは適用できないテストデザインであるといわざるを得ないだろう。

IRT モデルの分析ソフトウェアに関しては、テストの公平性の観点から、予備試験及び毎回の本試験において同一の手続きが行われることがテスト実施団体にとって求められることが制約となる。その場合、BILOG-MG を用いた分析を続けているテスト実施団体においては、常に別のソフトウェアによる分析の代替案を用意する意味で、BILOG-MG ではない別のソフトウェア（例えば、ICL や PARSCALE）による分析を並行して行うことが望ましい。多群 IRT におけるパラメタ推定は 4.1.1 節にて述べる方法などが提案されており、これらの方法を適用する数値計算のため

のプログラムを用意することによって、理論通りの推定結果が得られているかを確かめながらテストを実施することが可能である。しかし、テスト実施機関において、そのようなプログラムを用意し、維持することが可能な人材を長期にわたって確保することは、テストの実施を目的としている実施機関から見れば、本来の目的とは異なる人材を求めることでもあり、容易ではないことは確かであろう。したがって、本研究で扱うように、分析のためのソフトウェアを固定し、そのソフトウェアの特性をテスト実施前に知ることは、本研究で取り上げるテスト実施法（次節にて詳述）を長期にわたって安定して実施し続けるための最も現実的な選択肢のひとつであるといえる。

1.5　本研究で扱うテスト実施法

1.5.1　テスト実施法の概要

以上の問題点および制約を克服し、かつ、安定的に長期にわたって実施可能なテストの実施が可能な方法として、本研究では、「本試験に新作項目をアンカー項目よりも多く提示し、新作項目を項目バンクに入れ続ける」テスト実施場面を取り上げる。

大規模試験においては、規準集団に対して行う「予備試験」によって、毎回の試験結果を比較可能としたときの規準を定め、規準集団上での項目特性を知る必要があることは先に述べた。しかし、大規模試験に先立ち、多くの項目を提示するような大規模な予備試験を行うことは、実践的には困難である。テスト実施機関にとって、このような大規模予備試験が行えないのであれば、最初に行う予備試験の規模を小さくすることが必要である。すると、第 1 回目の本試験において、予備試験で得られた「項目パラメタ既知」の項目と、それより多い「新作項目」とで構成されたフォームを受験者に提示し、本試験を予備試験に等化することで、予備試験と比較可能な尺度での受験者の能力を表示できる。さらに、ここで等化された「第 1 回目の本試験における新作項目」が、第 2 回試験においては項目バンクの中身の一部として取り扱うことができる。第 2 回試験の実施においては、項目バンクとして「予備試験の項目」に「第 1 回試験で等化した項目」が加わることとなり、これを本試験実施のたびに繰り返すことにより、テスト実施のたびに項目バンクの中身が増えることになる。以上述べたテスト実施法を図 1.8 に示した。ただし、図 1.8 においては、第 3 回試験のアンカー項目は第 2 回試験と試行試験と共通に出題されているが、必ずしも過去行われたすべての回からアンカー項目を出題しなければならないわけではない。

このテスト実施法で「等化」を行うべき場面は、本試験を行うたびに現れる。本研究では、これらの等化方法は互いに同一である場合を考える。なぜなら、テスト実施機関は事前に試験の仕様を決定する際、事前に採点から等化までのルールを決めることで、実施回をまたいだテストの公平性を確保することができるからである。言い換えれば、テストの実施回によって、等化方法を変えることにより、特定の実施回の受験者に有利・不利が生じる事態を防ぐことができるからである。

本研究で扱うテスト実施法は、すでにいくつかの試験で実際に採用されている可能性がある。しかし、先に述べたように、項目バンクを用いた試験においては、項目の内容や、どのフォームにどのような項目を出題するかに関するルールを秘匿する必要がある。等化の方法を公表することは、

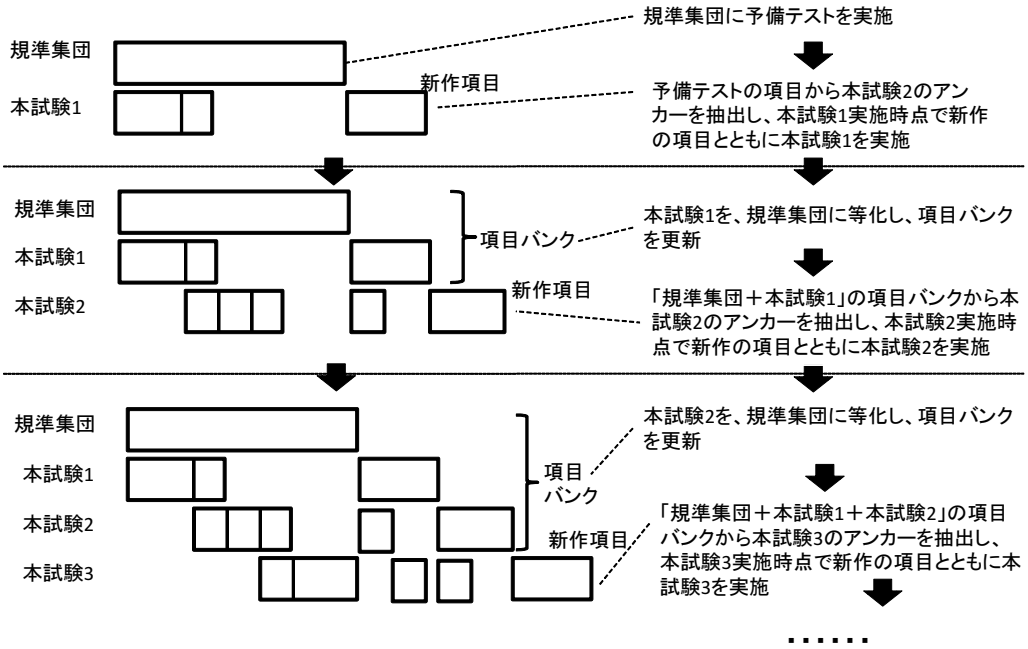

図 1.8　項目バンクのサイズがテスト実施のたびに増えるようなテスト実施法の概要

一定の法則性を持って定期的に本試験上にアンカー項目として同一項目が再出題されるというルールを公表することにつながりかねないため、等化ルールは通常、秘匿されるか、テストの研究課題としてテスト実施とは独立した形で行うという条件がついた上で、限定的に開示されることがほとんどである。したがって、このテスト実施法に関する学術的に一般性を持った先行研究は、特に日本においては、きわめて限られた形にならざるを得ない。そこで、本研究では、IRT に基づいて生成したシミュレーションデータを用い、そのデータを理論通りに分析した場合、それでもなお等化方法の違いによって生じる影響を探ることにより、実践場面においても適用可能な等化方法に関する知見を得ることを目的とする。

1.5.2　1 回の試験に複数のフォームが提示されている場合

本研究においては、予備試験・本試験を問わず、1 回の試験について一つのフォームが提示されている場面を考えている。しかし、1 回の本試験において、互いに異なるアンカー項目をもつ複数種類のフォームを用いてテストを行うデザインも考えることができる。たとえば、同一試験会場で隣同士に座った受験者に異なる種類のフォームを提示したり、時差のある試験会場で同時に試験を行うなどの場合、複数フォームのデザインが有効である。

このような複数フォームを等化する場合は、本研究で「本試験 1」「本試験 2」・・・としていた枠組みを、「フォーム通し番号 1」「フォーム通し番号 2」・・・と考え、それぞれの通し番号のフォームを、複数フォームが提示される 1 回の本試験に割り当てるようにすれば、本研究のテスト実施法

と同様の等化処理で項目バンクを構成することが可能である。表 1.1 に、本試験に複数フォームを提示した場合の、各テストフォームの割り当てを示す。

フォーム通し番号	1 フォーム/回	2 フォーム/回	3 フォーム/回
フォーム 1	本試験 1-1	本試験 1-1	本試験 1-1
フォーム 2	本試験 2-1	本試験 1-2	本試験 1-2
フォーム 3	本試験 3-1	本試験 2-1	本試験 1-3
フォーム 4	本試験 4-1	本試験 2-2	本試験 2-1
フォーム 5	本試験 5-1	本試験 3-1	本試験 2-2
フォーム 6	本試験 6-1	本試験 3-2	本試験 2-3
フォーム 7	本試験 7-1	本試験 4-1	本試験 3-1
フォーム 8	本試験 8-1	本試験 4-2	本試験 3-2
フォーム 9	本試験 9-1	本試験 5-1	本試験 3-3
フォーム 10	本試験 10-1	本試験 5-2	本試験 4-1
フォーム 11	本試験 11-1	本試験 6-1	本試験 4-2
フォーム 12	本試験 12-1	本試験 6-2	本試験 4-3

表 1.1　本試験 1 回あたり複数フォームを提示するデザインにおける、各フォームの割り当て。「本試験 1-1」は、「本試験 1 回目における第 1 フォーム」を表す。全体でフォーム数 12 のデザインで、本試験 1 回あたりのフォーム数が 1 から 3 の場合について示した

1.6　本研究で扱うテスト実施法における問題点

テスト実践場面において、前項で提案したテスト実施法を実際に行う際、以下のような問題点が生じることが予測される。

1.6.1　項目バンクのパラメタ更新の方法について

本研究で扱うテスト実施法において、毎回の試験において受験者の成績を推定する際、成績を出す根拠になる要素は、毎回試験における問題項目の等化済みの項目パラメタと、受験者の正誤反応データである。後者はデータに基づいているので、その値の解釈に疑義が生じることはない。しかし、前者に関しては、毎回の試験において「正しい」と信じられる方法によって等化され、確実に規準集団のスケールで比較可能であると解釈できる項目パラメタが求められる。したがって、より理論通りの結果となるような等化方法の検討が必要であった。

本研究で取り上げるテスト実施法においては、特に毎回の本試験において、項目全体に占める新作項目数の割合（以下、「新作項目割合」と呼ぶ）が大きいときに、項目バンク上のパラメタ推定値がどうなるかを検討する必要がある。本研究で取り上げるテスト実施法において、毎回の本試験で新作項目割合が大きければ、毎回の試験においてアンカー項目よりも多くの新作項目を出題する

こととなり、項目バンクのサイズは実施のたびにより大きくなっていくことになる。しかし、そのような場合においては、本試験においてアンカー項目数が相対的に少ないこととなり、等化が不安定になる可能性がある。このような場合の検討がこれまで必要であった。

1.6.2　同時推定時における計算上の限界

同時推定を行う場合、データセットの大きさはテストの実施とともに大きくなる。テストの実施上、1 回あたりに用意するフォームの種類を増やす場合も考えうる。その場合、計算上の限界が出来する可能性がある。理論上は計算可能であっても、メモリの消費量や計算時間が莫大である場合、実践場面で同時推定を行うことは不可能である。しかし、理論上は推定すべきパラメタであるものの、実践上はあまり必要とされないパラメタとして、各フォームに対応する受験者グループの能力値の分布に関するパラメタが挙げられる。このパラメタを推定しなければ、特にメモリの消費量を減らすことができる可能性がある。しかし、グループごとの能力値に関するパラメタを推定する場合と推定しない場合で、項目パラメタの推定値に影響があるかどうかはこれまで不明であった。

1.6.3　テストの一次元性が等化済み項目パラメタに及ぼす影響

毎回の試験で、項目パラメタの推定を行う際、規準集団に対して行う試験と、それ以降に行われる本試験とで、同一の概念が測定されている必要がある。とりわけ、本研究で提案するテスト実施法においては、本試験の受験者の成績を推定するために、規準集団上でのスケールに等化した項目パラメタを用いる。そのため、識別力の推定値が測定している概念に対する因子負荷と等価であるという解釈に則れば、特に予備試験と本試験との間で、識別力が同等であることが最低限求められる。

ここで、一次元性が満たされない場合について、図 1.9 に示す 2 種を考えることができる。図 1.9 の (a) に示すような一次元性の乱され方は「予備試験と本試験との間で測定したい概念が変化した場合」を示し、図 1.9 の (b) に示すような場合は「そもそも一次元性に乏しい構成概念に関するテスト」であるといえる。

等化の前提の中に「一次元性」が含まれている以上、。図 1.9(b) のようなテストにおいては、等化の前提が満たされているとは言えない。一方、図 1.9(a) のように一次元性が満たされていない場合においては、等化の前提は本試験のみで見た場合には満たされている。すなわち、後者の場合は前者の場合に比べて、等化に関する前提を局所的に満たしているという点で、より等化の結果が妥当なものになる可能性がある。しかし、後者の場合であっても、予備試験と本試験との間で比較可能なスケールに乗せる操作の際に、項目バンク上の推定結果が理論とかけ離れたものになる可能性は避けられない。Karkee, Lewis, Hoskens, Yao and Haug (2003) では、図 1.9(b) のような場合においては個別推定を行った方が良い、としている。しかし、図 1.9(a) の場合、どのような等化法をとるべきかに関する議論は、ほとんど行われていない。

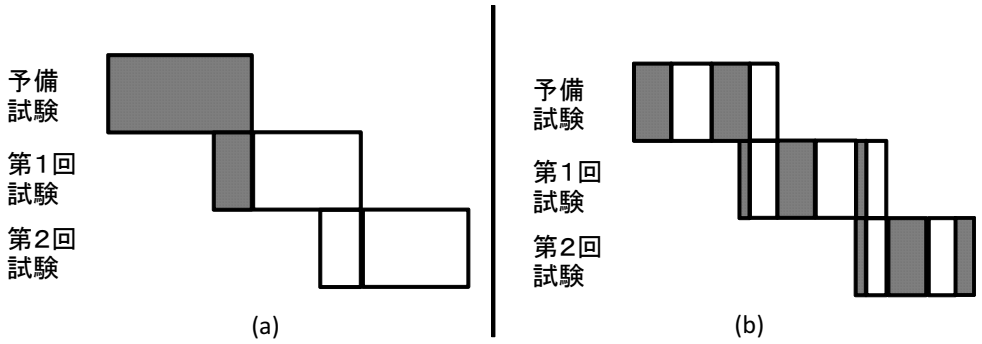

図 1.9　テストの一次元性が満たされない場合 2 種。(a) は「予備試験と本試験間で一次元性は満たされないが、予備試験と本試験の内では一次元性が満たされる場合」、(b) は「予備試験と本試験の間および内のどちらも一次元性が満たされない場合」を示す。同じ色は当該項目が同一の構成概念を測定していることを表す

テスト実施機関にとっては、これら 2 通りの「テストの一次元性が満たされない場合」のうち、「予備試験と本試験で一次元性が満たされないが、それぞれの中で一次元性が満たされる場合」の方が実際に起こりやすい場合であると考える。たとえば、水平等化を行うテストを毎年繰り返し、学力の経年変化を共通項目デザインでとらえるテストにおいて、各テストの中で同時に複数のフォームを用いて垂直等化を行うようなデザインを考える。この場合、垂直等化を行うフォーム間においては、一次元性の問題は大きくないことが予想される。なぜなら、成績上位群向けと下位群向けの項目は、同一の概念を測定していることが前提だからである。しかし、水平等化の部分においては、本研究で記したテスト実施法を採用する背景として、あらかじめ複数年度の項目をまとめて作成し、年度ごとのフォームを事前に用意するだけの作題能力に乏しい場合が多い。この場合、各年度に出題される項目を個別に毎年作成することになり、年度内の項目作成場面においては一次元性に留意した項目作成およびフォーム編集作業が行われるものの、「年度内では一次元だが年度間で一次元性が保証されない」場合が生じうることが予想される。

1.7　目的、および制限事項

1.7.1　本研究の目的

以上の背景を踏まえ、本論では以下の 4 つの手順に従い、大規模テストで規準集団上に等化する試験場面において、実現可能性の高い等化方法をテストデザインを含めたテストの仕組みとともに提案し、その仕組みによって得られた項目バンクの中身について、理論にかなった結果を返しているか、パラメタの推定値にバイアスがかかる要因がどのようなものか、といった点についてシミュレーション研究を行う。

■第 2 章・個別推定における等化の順序性の検討　1.3.2 節の「個別推定」の項で述べたとおり、個別推定においては等化において項目バンク上の項目パラメタと等化後の項目パラメタとの間で値に違いが生じうる。項目バンク上の値をそのまま使い続ける場合と、等化後の項目パラメタを使い続ける場合、また両者の平均をとる場合で、個別推定の方法別に項目バンク上のパラメタ推定値にどのような影響が及ぶか、シミュレーションによる検討を行う。

■第 3 章・IRT を用いた大規模テストにおける個別推定と同時推定とのパラメタ比較　1.4 節で挙げた制約を解決するようなテスト実施法の一案として、項目バンクのサイズが本試験実施のたびに増えるような実施法を取り上げる。この方法において、項目バンク上での項目パラメタの推定値が等化方法（同時推定、個別推定）によってどのように変わるかを、シミュレーション研究によって検討する。同時推定に関しては、多群 IRT モデルを用いる方法と、多群 IRT モデルを用いたのちにあらためて規準集団に等化する方法とを比較する。

■第 4 章・多群 IRT モデルのパラメタ推定における計算機資源の限界　同時推定を行う場合、テスト実施のたびにデータのサイズが大きくなっていく。このような大きなデータを継続して分析し続けるためには、実際にそれが計算可能かどうかを確かめておく必要がある。この点を、シミュレーションによって検討する。

■第 5 章・多群 IRT モデルにおけるモデル簡素化の評価　1.4.4 節で示した通り、項目バンクのサイズがテスト実施のたびに増えるテスト実施法を、同時推定により継続して等化することができるように、複雑なフォーム・グループ対応を簡素化した場合を考える。その上で、項目バンク上における項目パラメタ推定値にどのような影響が及ぶかを、等化方法間、およびグループのまとめ方を変えた場合で比較する。

最後に、「第 6 章：考察・総論」にて、本研究で扱うテスト実施法における最適な等化方法に関する議論を行い、本テストデザインによる大規模テストの実施について、実践場面における問題点を指摘する。

以上の章節構成を図示すると、図 1.10 のようになる。

1.7.2　本研究の制限事項

本研究においては、全ての章において、シミュレーション研究を行っている。シミュレーション研究の利点は、研究者が検討したい事項を事前に統制したうえで、検討したい作業仮説を検証するような実験的な方法論に基づいて研究を行うことが容易であることが挙げられる。一方、テストの研究としては、実際の受験者から得られたデータを用いて、そのデータを推定した項目パラメタを「真値」とし、「真値」を持った項目バンクから一定数の項目を抽出したうえで、それらの項目を用いて等化のためのテスト実施法を構築した場合を仮定して研究を行うことも可能である。あるいは、あらかじめ共通項目デザインに基づくテストを実施し、その結果を複数の等化方法で等化した結果を比較するような研究も考えることができる。

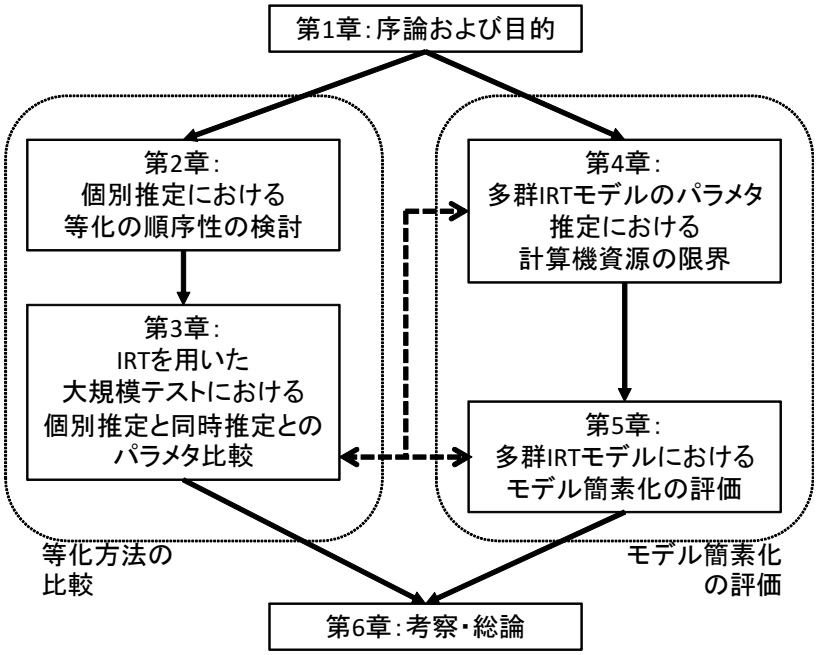

図 1.10　本研究の構成。実線は論の順序、点線は相互に関連あるトピックを示す

テスト実践場面で得られた正誤反応を用いて等化の研究を行うためには、本試験を実施する前にテスト実施機関が受験者に対し、「本テストの答案データは、将来のテスト研究のために使用する場合がある」旨の同意を得る必要がある。この手続きは、テストの仕組みや提示する問題冊子を心理学における実験データを収集するための材料であると考えた場合は倫理上欠かすことのできない手続きであり、そのような配慮がなされないまま受験者の同意なく研究に用いることは推奨されない。

本研究の実施に当たり、全ての研究でシミュレーションによる研究にとどまったのは、実データを用いた研究を行うための同意を得ることができなかったためである。よって、これらの研究成果は、実データによるさらなる検証の余地があるだろう。このことは、6.5.1 節においても触れることとする。

また、本研究においては、項目バンクにおけるパラメタを毎回の本試験において抽出し、毎回の本試験の受験者における θ の真値と比較することは、行っていない。理由として、θ の真値の比較を行う場合、各回の本試験における受験者数の効果を検討する必要があり、「受験者数」という従属変数を本試験および予備試験について設定する必要があり、条件数を増やすことがシミュレーション計画を複雑にすることに直結することが挙げられる。また、本研究で取り上げるテストの実施法では、項目バンク上のパラメタの値を手掛かりに次の試験におけるアンカー項目を抽出する必要があり、常に試験実施後の項目バンクにおける予備試験に等化済みの項目パラメタの値が参照されるので、等化済み θ を理論通りに推定するためには最低限、等化済み項目パラメタが正しく推定されていなければならないことが挙げられる。理論的には、毎回の試験の後に得られた項目バンク

の中にある等化済みの項目パラメタを θ の推定に用いることにすれば、その値は (任意の平均と標準偏差を持つ) 予備試験の θ の尺度と直接比較可能である。なぜなら、項目パラメタの値が規準集団上の尺度に等化されているということは、式 1.16、式 1.17、式 1.18 に従って適切な (k, l) が推定されているということを意味し、その (k, l) の値を用いて式 1.14 を適用すれば、等化済み θ_y を等化前 θ_x から変換する操作と等価だからである。以上より、本研究においては、項目バンクに入れるべき等化済み項目パラメタの値に関する議論を行うこととする。

第 2 章

個別推定における等化の順序性の影響

2.1　序論

2.1.1　等化の順序性の問題

前章において、あらかじめ本試験実施前に定義した規準集団上の尺度に本試験を等化し続けるような試験場面を考え、最適な等化方法の検討や実際に実施可能なテストデザインの検討の必要性について述べた。本研究では、個別推定において、等化方法の違いのみならず、どのフォームから規準集団に等化するかによって、等化後の項目パラメタの値がどのようにばらつくかについて検討する。

規準集団への等化を繰り返す試験場面で、個別推定を毎回行う場合、本試験を実施するたびに当該本試験のみの受験者の 0-1 データを用いて、本試験 1 回に出題された項目に関して項目パラメタを推定する。その後、本試験中に含まれる「項目バンク上での規準集団に等化された項目パラメタ」の尺度に、本試験の項目パラメタを合わせる。この作業が「等化」である。個別推定においては、毎回の試験で測定されている概念は、同一であることが前提となる。

個別推定場面における等化の方法として、前章においては「Mean/Mean 法」「Mean/Sigma 法」「Haebara 法」「Stocking-Lord 法」「calr 法」をとりあげた。これらの方法は、いずれも「等化元」の尺度を「等化先」の尺度に等化する。ここで、「等化先」は「規準集団」であり、その数はただ 1 つである。その一方、「等化元」は複数の尺度、たとえば 5 回の本試験があった場合、5 つの異なるテストフォームから得られた項目パラメタがすべて「等化元」となる。

等化方法のうち、「Mean/Mean 法」「Mean/Sigma 法」「Haebara 法」「Stocking-Lord 法」は、いずれも「等化元」に一つのフォームしか取れない。その意味で、「ペアワイズな等化法」であるといえる。この方法の場合、ペアとなる 1 つの「等化元」を「等化先」に等化するたびに、等化処理によって「等化元において等化先と共通項目である項目で、等化処理によって変換された項目パラメタの値」が算出される。この値を「等化済みの等化元パラメタ」と呼ぶ。「等化済みの等化元パラメタ」は、理論通りであれば、等化先と大きく傾向が異ならない値となるはずである。しかし、実際には、個別推定によって得られる「等化済み項目パラメタ」は「等化元パラメタ」の線形変換

であり、一部の項目において大きく値が異なる可能性が捨てきれない。等化先のパラメタと大きく異なる「等化済みの等化元パラメタ」を項目バンクに入れるかどうかで、(1) 「等化済みの等化元パラメタ」を項目バンクに入れ、元あった等化先のパラメタは破棄する、(2) 元の等化先のパラメタを入れ、「等化済みの等化元パラメタ」は破棄する、(3)「等化済みの等化元パラメタ」と元あった等化先のパラメタの平均を、項目パラメタの種類（識別力、困難度）についてとる、の 3 通りのオプションが考えうる。このような「元の等化先のパラメタ」と「等化済みの等化元パラメタ」という 2 種類の項目パラメタが現れるのは、アンカー項目についてのみであり、その他の「本試験のみに提示されている新作項目」においては、このような問題は生じない。しかし、アンカー項目における項目パラメタの扱いを変えたことによって、項目バンク上の項目パラメタの推定値が今後参照され、新たな「等化先」となることを考えると、2 種類の項目パラメタの扱いについては検討しなければならない実践上の問題であるといえる。

一方、個別推定による等化方法のうち calr 法は、一つの「等化先」に、複数の「等化元」を同時に等化することが可能である。calr 法では、各フォームに対応する受験者集団ごとに異なる等化係数を推定し、同時に、全項目における等化された項目パラメタを推定する。したがって、このような等化の順序性に関する問題は生じない。

2.1.2 個別推定における方法の比較研究

個別推定の方法間の比較については、Arai and Mayekawa (2011) では Mean/sigma 法と calr 法を比較し、calr 法の方が推定の誤差が少ないという結果を得ている。また Hanson and Béguin (2002) や Lei and Zhao (2012) では Mean/Mean 法、Mean/sigma 法、Haebara 法、Stocking-Lord 法を比較している。Hanson and Béguin (2002) では実データからサンプリングする方法でデータセットを生成し、2 つのフォームを共通項目デザインで等化する場面のシミュレーションを行い、Mean/Mean 法、および Mean/Sigma 法において、等化先のスケールにおける等化元の項目パラメタの値の等化後推定値が理論値よりも大きくずれるという結果を得た。原因として、Mean/Sigma 法では困難度の平均、また Mean/Mean 法では識別力の平均を等化係数の推定に用いるため、極端な困難度の影響を受けやすいことを指摘している。また Lei and Zhao (2012) では、垂直等化を行う場合において受験者の能力が上昇する場面をシミュレーションにより検討し、項目数が少ない場合や受験者人数が少ない場合に Mean/Sigma 法において推定誤差が大きいという結果を得た。

これらの研究においては、Arai and Mayekawa(2011) や Lei and Zhao (2012) は複数のフォーム（前者は規準集団 +2 フォーム、後者は規準集団 +3 フォーム）を仮定したテストデザインを用いているが、Hanson and Béguin(2002) は 2 つのフォームに関する等化のみを扱っている。また、いずれの研究も、本研究のような「等化の順序性」を考えた計画とはなっていない。しかし、これらの先行研究の知見から、Mean/Sigma 法においては、項目パラメタの推定値に誤差が大きいことが予想され、等化の順序によって推定結果に大きくばらつきが生じることが予想される。

2.1.3　本研究の目的

本研究においては、このような等化の順序性に関する問題点を明らかにするため、規準集団への等化を繰り返すようなテスト場面を模したシミュレーションを行う。同時に、calr 法で得られた結果と比較し、前節で述べた (1) から (3) のオプションのうち、いずれの方法が真値、あるいは calr 法による推定値に近くなるかを検討する。その際、本研究の目的の一つである「新作項目割合」を操作することで、項目パラメタの推定結果にどのような影響があるかを検討する。

2.2　シミュレーション方法

2.2.1　想定されたテストデザインおよび受験者

はじめに、能力値の規準となる「規準集団」を定義するための「予備試験」を N_{trial} 人の受験者を対象に行ったと想定した。ここでの能力を $\theta \sim N(0,1)$ となるように項目パラメタを推定した。ここで、$N_{trial} = 4000$ とした。

以後、予備試験で出題された項目のうち J_{anchor} 項目を「アンカー項目」として、J_{new} 項目の「新作項目」（その実施フォームで初出の項目）とともに提示する場面を想定した。ここで、全ての実施フォームについて、アンカー項目は予備試験時点でパラメタが推定された項目のみを用いることとした。本試験の受験者 N_{exam} 人が $(J_{anchor} + J_{new})$ 項目を解き、0-1 データを得たと想定した。ここで、$N_{exam} = 8000$ とした。

本試験の 0-1 データより、そのフォームの受験者における項目パラメタ（等化前）を推定し、アンカー項目の予備試験時点のパラメタを手掛かりに予備試験のパラメタに等化した。

本試験のフォーム数は受験者集団の数と同じとし、受験者集団 g は、$(3, 4, 5)$ の 3 通りを考えた。また、試験の項目数に応じ、「小アンカー条件」「大アンカー条件」「多項目条件」の 3 通りを考えた。「小アンカー条件」では $(J_{anchor}, J_{new}) = (6, 24)$、「大アンカー条件」では $(J_{anchor}, J_{new}) = (12, 18)$、「多項目条件」では $(J_{anchor}, J_{new}) = (12, 48)$ とした。「小アンカー条件」および「多項目条件」では $(J_{anchor} : J_{new}) = (1 : 4)$、「大アンカー条件」では $(J_{anchor} : J_{new}) = (1 : 1.5)$ となるようにした。また、アンカー項目のうち $(J_{anchor}/3)$ 項目を、隣り合ったフォームと共通のアンカー項目とした（第 1 回目を除く）。以上より、予備試験における項目数（本試験におけるアンカー項目数）は、本試験の回数および項目数条件に応じ、表 2.1 のとおりとした。「小アンカー条件」および「多項目条件」は、新作項目割合が大きく、「大アンカー条件」は新作項目割合が小さい場合に相当する。また、本研究で用いたテストデザインを、小アンカー条件の場合について、図 2.1 （$g = 3$）、図 2.2($g = 4$)、図 2.3($g = 5$) にそれぞれ示した。

予備試験の項目パラメタの真の値については、識別力 $a_{trial} = 0.75$ とした。また、困難度の真値 b_{trial} は $N(0, 1)$ に従うものとし、これをもとに、本試験のアンカー項目における困難度の真値も生成した。すなわち、予備試験における困難度の真値のリストにおいて、小さい方から J_{anchor} 項目をフォーム 1 のアンカー項目に、その次に小さい方から J_{anchor} 項目をフォーム 2 のアンカー

表 2.1　本研究で用いた予備試験の項目数

	$g=3$	$g=4$	$g=5$
小アンカー	14	18	22
大アンカー	28	36	44
多項目	28	36	44

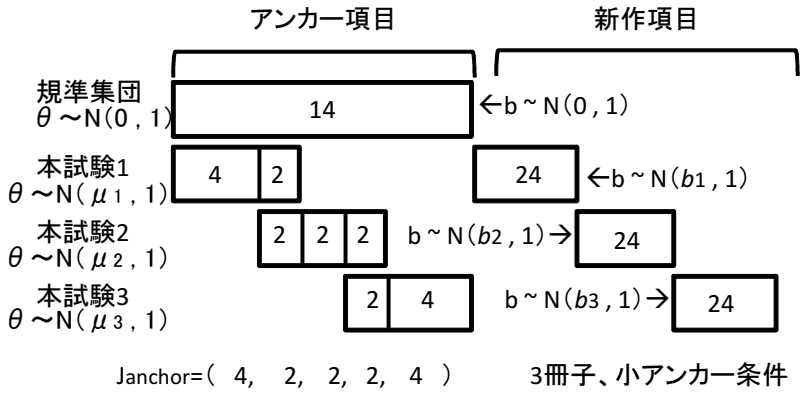

図 2.1　本研究で用いたテストデザイン (本試験 3 回の場合)。N(0.00,1.00) などの表記は、該当する受験者グループの真の能力値分布を示す

項目に、という形で、それぞれのフォームが「フォームの番号が小さくなればなるほど易しい項目が提示される」という真値の困難度を持つようにした。

一方、本試験の新作項目の真値は識別力 $a_{exam}=0.5$ といた。本試験の新作項目における困難度 b_{exam}、および、各フォームに対応する受験者の能力の真の平均 θ_{exam} は同じ値とした。この値を、以下の 3 通りについて考えた。(1) アンカー項目に対応した、すなわちフォームの番号が小さいほど、困難度の真値も小さい条件、(2) どのフォームにおいても、新作項目の困難度の真値が 0 となる条件、(3) 新作項目の困難度の真値が、アンカー項目の困難度の真値とマッチしていない条件、の 3 つの条件をそれぞれ「H1」「H2」「H3」の条件とし、それぞれの g の条件別に設定した困難度、および θ の真値を表 2.2 に示した。

表 2.2　本試験における困難度の真値。各々のセルの条件で、それぞれ (フォーム 1),(フォーム 2),……における困難度の真値の (平均, 標準偏差) をもつ正規分布に従うことを示す。g の条件、および、H の条件ごとに示した

	$g=3$	$g=4$	$g=5$
H1	$(-1,1),(0,1),(1,1)$	$(-1,1),(-0.5,1),(0.5,1),(1,1)$	$(-1,1),(-0.5,1),(0,1),(0.5,1),(1,1)$
H2	$(0,1),(0,1),(0,1)$	$(0,1),(0,1),(0,1),(0,1)$	$(0,1),(0,1),(0,1),(0,1),(0,1)$
H3	$(1,1),(0,1),(-1,1)$	$(1,1),(0.5,1),(-0.5,1),(-1,1)$	$(1,1),(0.5,1),(0,1),(-0.5,1),(-1,1)$

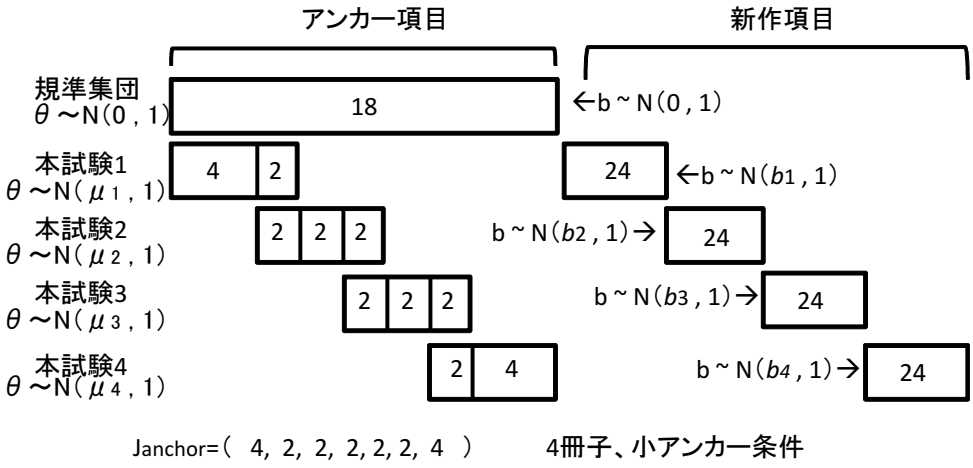

図 2.2　本研究で用いたテストデザイン (本試験 4 回の場合)。N(0.00,1.00) などの表記は、該当する受験者グループの真の能力値分布を示す

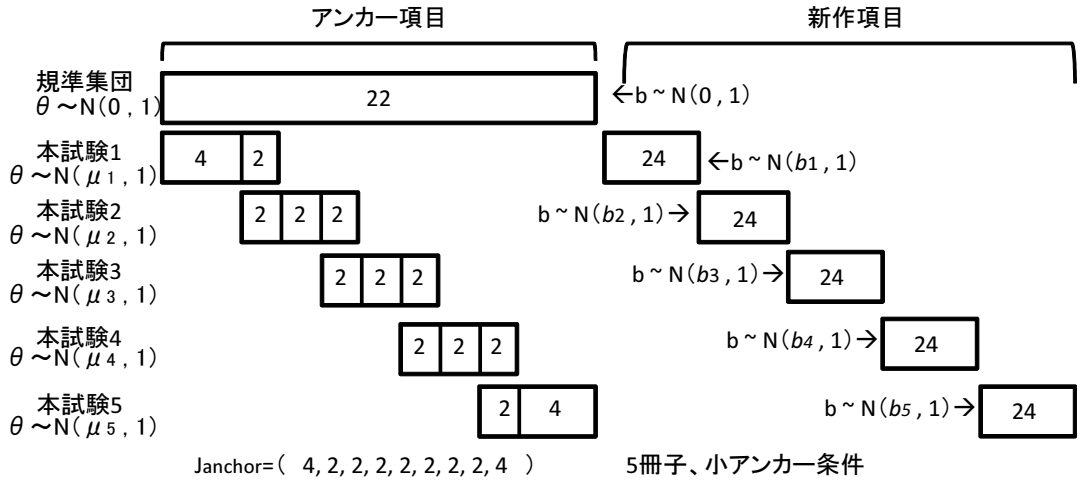

図 2.3　本研究で用いたテストデザイン (本試験 5 回の場合)。N(0.00,1.00) などの表記は、該当する受験者グループの真の能力値分布を示す

2.2.2　シミュレーション手続き

テストデザインに従って、アンカー項目および毎回の新作項目の真の値を定めた。初めに予備試験の 0-1 データのみを用いて項目パラメタを推定し、のちのアンカー項目における項目パラメタの推定値 $\boldsymbol{\xi}_{norm}$ とした。この手続きののち、以下に示す「ペアワイズの方法」および「calr 法」それぞれにおいて等化手続きを行った。以下の手続きをそれぞれの条件において 100 回繰り返した。0-1 データの生成には RESGEN4(Muraki,2000)、項目パラメタ推定には BILOG-MG 3、ま

た HB、SL、MS の等化には R 言語の plink パッケージ[*1] を用いた。また、全ての項目パラメタの推定では、2 パラメタ・ロジスティックモデル (2PL) を用いた。

ペアワイズの方法に関する手続き

g 個ある本試験のフォームについて、等化順を定めた。定めた等化順の通りに該当するフォームについてアンカー項目および本試験の項目パラメタの真値を定め、そこから 0-1 データを生成した。アンカー項目の真値は、予備試験における真値と同一の値とした。次にその 0-1 データより項目パラメタ $\boldsymbol{\xi}_{ancnew}$ を求め、$\boldsymbol{\xi}_{ancnew}$ 内にあるアンカー項目の項目パラメタ $\boldsymbol{\xi}_{anc}$ を等化元、$\boldsymbol{\xi}_{norm}$ を等化先とするペアワイズの等化を行った。続いて、「項目パラメタ更新ルール」(後述) に基づき、$\boldsymbol{\xi}_{norm}$ の値を更新した。以上の 0-1 データ生成から項目パラメタ推定、等化、項目バンクの更新の手続きまでを、等化順にしたがって全フォームについて行った。

等化の方法については、Haebara 法 (HB)、Mean/sigma 法 (MS)、および Stocking-Lord 法 (SL) の 3 通りを行った。さらに、「項目パラメタ更新ルール」として、(1)$\boldsymbol{\xi}_{norm}$ をそのまま使い続ける (equate to Same norm group; S)、(2)$\boldsymbol{\xi}_{norm}$ の値を $\boldsymbol{\xi}_{anc}$ に置き換える (equate using each Examination; E)、(3)$\boldsymbol{\xi}_{norm}$ と $\boldsymbol{\xi}_{anc}$ の平均を困難度、識別力それぞれについてとる、すなわち $(\boldsymbol{\xi}_{norm} + \boldsymbol{\xi}_{anc})/2$ を計算して $\boldsymbol{\xi}_{norm}$ とする (calculate Mean of parameter; M)、の 3 通りを考えた。これらをクロスさせた 9 通り (HBS,HBE,HBM,MSS,MSE,MSM,SLS,SLE,SLM) において等化手続きを行った。

以上の方法のうち、常に規準集団のパラメタを等化に用いる (1) の「項目パラメタ更新ルール」の場合は、等化順をどのようにしても、それらの間で推定される項目バンクの推定値は同一である。しかし、(2) および (3) については、等化順によって項目バンクの推定値に違いが出る。そこで、(2) および (3) の方法については、$g!$ 通りの等化順のすべてについて上記の等化手順を行い、それぞれの等化順における従属変数について、$g!$ 通りの試行で最も大きな従属変数を示した試行 (最も真値とかけ離れた等化順)、最も小さな従属変数を示した試行 (最も真値に近い等化順)、全試行における従属変数の平均値、を記録した。

calr 法における手続き

g 個あるフォームについて、フォーム 1、フォーム 2、の順に、それぞれアンカー項目の真値を予備試験の項目パラメタの真値から引用し、さらに新作項目のパラメタ真値とともに定め、そこから 0-1 データを生成した。さらに、それぞれのフォームについて個別に項目パラメタを推定した。それらの値をもとに、calr 法に基づいて複数フォームの尺度を規準集団の尺度に等化した。最後に、従属変数を記録した。

[*1] http://cran.r-project.org/web/packages/plink/index.html

2.2.3　従属変数

各試行について、DICC(Arai and Mayekawa,2011) を式 2.1 に基づいて算出した。DICC は、求められたすべての項目パラメタの数値（新作項目かアンカー項目かは問わない）を対象にして、その ICC の差を表す値として算出された。値が小さいほど、推定された項目パラメタが真値に近いことを示す。

$$DICC = \frac{1}{J}\sum_{j=1}^{J}\frac{1}{Q}\sum_{q=1}^{Q}\left|P_j(\theta_q|\hat{a}_j^l,\hat{b}_j^l) - P_j(\theta_q|a_{jT},b_{jT})\right| \tag{2.1}$$

ここで J は $1,2,...,j,...J$ 番目の項目、Q は $1,2,...,q,...Q$ 番目の θ の求積点を表し、θ_q は -3 から 3 までのスケールを Q 等分した各値とし、$Q=31$ とした。また、a_{jT} および b_{jT} はそれぞれ j 番目の項目における識別力と困難度の真値を表し、$\hat{a}_j^l$ および $\hat{b}_j^l$ はそれぞれ j 番目の l 回目の推定における識別力と困難度の推定値を表す。

2.3　結果

困難度条件、フォーム数条件別の推定結果における DICC の平均値を小アンカー条件においては図 2.4 に、大アンカー条件においては図 2.4 に、多項目条件においては図 2.6 に、それぞれ示した。また、困難度条件、フォーム数条件別の、100 回の推定値を推定方法別にプロットした結果（散布図）を、小アンカー条件では図 2.7、大アンカー条件では図 2.8、多項目条件では図 2.9 に、それぞれ示した。いずれの図においても、各列（横）にフォーム数 $g=3,4,5$ の場合を、各行（縦）には上から H1（アンカー項目と新作項目の真値の困難度の傾向が一致)、H2（アンカー項目の困難度に関わらず真値の困難度の平均が 0)、H3（アンカー項目の困難度と新作項目の困難度の傾向が不一致）の場合をそれぞれ示した。

平均のグラフにおいては、順序性が影響する推定方法（HBM,HBS,MSM,MSS,SLM,SLS）について、「100 回の試行の、$g!$ 通りの試行における平均 DICC」、「$g!$ 通りの組み合わせのもとでの最大 DICC を、100 回の試行で平均した値」、「$g!$ 通りの組み合わせのもとでの最小 DICC を、100 回の試行で平均した値」の 3 つの点をプロットした。この 3 点の間隔が大きい条件は、順序性が推定結果に影響すると解釈した。また、同様に、散布図において、順序性が影響する推定方法については、平均のグラフで示した 3 つの要素それぞれについて、100 試行の DICC を別の系列で示した。「最小 DICC」の最小値は、「平均 DICC」の最小値よりも小さく、「最大 DICC」の最小値より小さくなる。また、「最小 DICC」の最大値は、「平均 DICC」の最大値よりも小さく、「最大 DICC」の最大値より小さくなる。

2.3.1　フォーム数の違いによる DICC の傾向の違い

図 2.4、図 2.5、および図 2.6 において、各行に並んだグラフを比較すると、フォーム数 g が 3、4、5 の場合でも推定方法によって DICC の平均値に大きな値の傾向の違いがないことが分かった。

この傾向は、困難度条件、およびアンカー項目の大小とは関係がない。また、同様に、DICC の推定値の標準偏差も、フォーム数の違いによる値の傾向の差は見られなかった。

2.3.2 等化方法の間における DICC の傾向の違い

図 2.4、図 2.5、および図 2.6 を見ると、いずれの困難度およびフォーム数条件においても、Mean/sigma 法 (MS) が大きな DICC を示した。それ以外の Haebara 法 (HB)、Stocking-Lord 法 (SL)、および calr 法は、MS よりも小さな DICC であった。

また、等化の順序性が影響する等化方法のうち、MS、HB および SL においては、項目バンク更新時に平均をとる方法 (MSM、HBM、SLM) が、更新時に常に各本試験の等化済みパラメタを記録する方法 (MSE、HBE、SLE) よりも値の最大値、最小値がより平均値に集まる傾向にあることが分かった。また、項目バンク更新時に常に規準集団上の値を記録する方法 (HBS、MSS、SLS) の DICC は、平均をとる方法よりも大きな DICC となるが、常に本試験のパラメタの値を記録する方法に比べると小さな DICC となることがわかった。さらに、DICC の平均値で見ると、大アンカー条件、および多項目条件において、HBS、MSS、SLS は他の 2 つの方法よりも大きな DICC を示し、その差が小アンカー条件よりも大きな傾向が見られた。

DICC の散布図 (図 2.7、図 2.8、図 2.9) より、特に MS の方法において、条件によっては 0.03 を超える DICC となり、これは他の方法には見られない大きな値であるといえる。さらに、MS においては、項目バンクに値を入れる際の方法によって、DICC に大きな差が見られた一方で、HB、SL においてはその差が小さいという結果となった。

2.3.3 等化順の違いによる DICC の傾向

DICC の散布図 (図 2.7、図 2.8、図 2.9) より、HBM、HBE、MSM、MSE、SLM、SLE のそれぞれにおいて、等化順の違いが推定結果にどのように影響するかを検討したところ、HBM、HBE、SLM、SLE の方法においては、1 つの試行において、等化順の違いの結果得られた DICC の最小値と平均値、それに最大値が各試行内で接近した値になっている傾向が見られた。この傾向は、項目数、テストの長さ、困難度の違いによらず、ほぼ一定の傾向であった。それに対し、MSM、MSE の場合は、等化順の違いの結果得られた DICC の最小、平均および最大が同一試行内でかけ離れた値になる傾向が見られた（たとえば、図 figdiccplotminor において、MSM および MSS の 100 試行で、DICC の最小値・平均値・最大値が同一試行内で点線でつないて示しているが、これらの値が試行間で大きさが逆転しているのに対し、HBM、HBE や SLM、SLE では試行間で大小関係がほぼ一定であることがわかる）。この結果から、MS の方法は、等化順の影響で、得られた推定値が真値とかけ離れる場合が見られる可能性が、他の方法よりも大きいことが分かった。

2.3.4　アンカー項目の大小による DICC の傾向の違い

テストの長さ（項目数）が同じ場合において、アンカー項目が少ない場合（平均：図 2.4、散布図：2.7）とアンカー項目が多い場合（平均：図 2.5、散布図：図 2.8）を比較すると、アンカー項目が多い場合、いずれの等化方法においてもより真値に近い推定値となる傾向を示した。また、アンカー項目の多い場合には、散布図より、DICC の値の範囲がより狭くなる傾向が見られた。この傾向は、MS において顕著で、その他の方法においてもわずかながら見られた。

2.3.5　テスト全体の長さによる DICC の傾向の違い

アンカー項目の比率が同じ場合において、テストの長さが短い場合（平均：図 2.4、散布図：2.7）と長い場合（平均：図 2.6、散布図：2.9）とで DICC を比較すると、テストの長さが長い場合に、いずれの等化方法においてもより真値に近い推定結果となる傾向を示した。また、テストの長さが長い場合には、アンカー項目の多い場合と同程度に、DICC の値の範囲が縮小する傾向が見られた。結果的に、テスト全体の長さを長くした場合でも、アンカー項目の割合を大きくするのと同程度に DICC を小さくする効果が見られた。

2.3.6　困難度条件による DICC の傾向の違い

図 2.4、図 2.5、および図 2.6 において、各行に並んだ 3 つのグラフを比較すると、主に MS の方法において、アンカー項目と新作項目の困難度の傾向が一致している場合（図の一番上の行 3 枚）と比較して下の 2 枚（困難度の値がアンカー項目によらず一定の場合、および、困難度の傾向がアンカー項目と不一致の場合）の場合に大きな DICC を示すという結果となった。さらに、小アンカー条件 (図 2.4) において、他の 2 つの項目条件 (図 2.5 および図 2.6) に比べて困難度条件の違いにより DICC の範囲が大きくなる傾向が見られた（この傾向は、MS の方法において最も顕著であったが、他の方法においてもわずかながら見られた)。

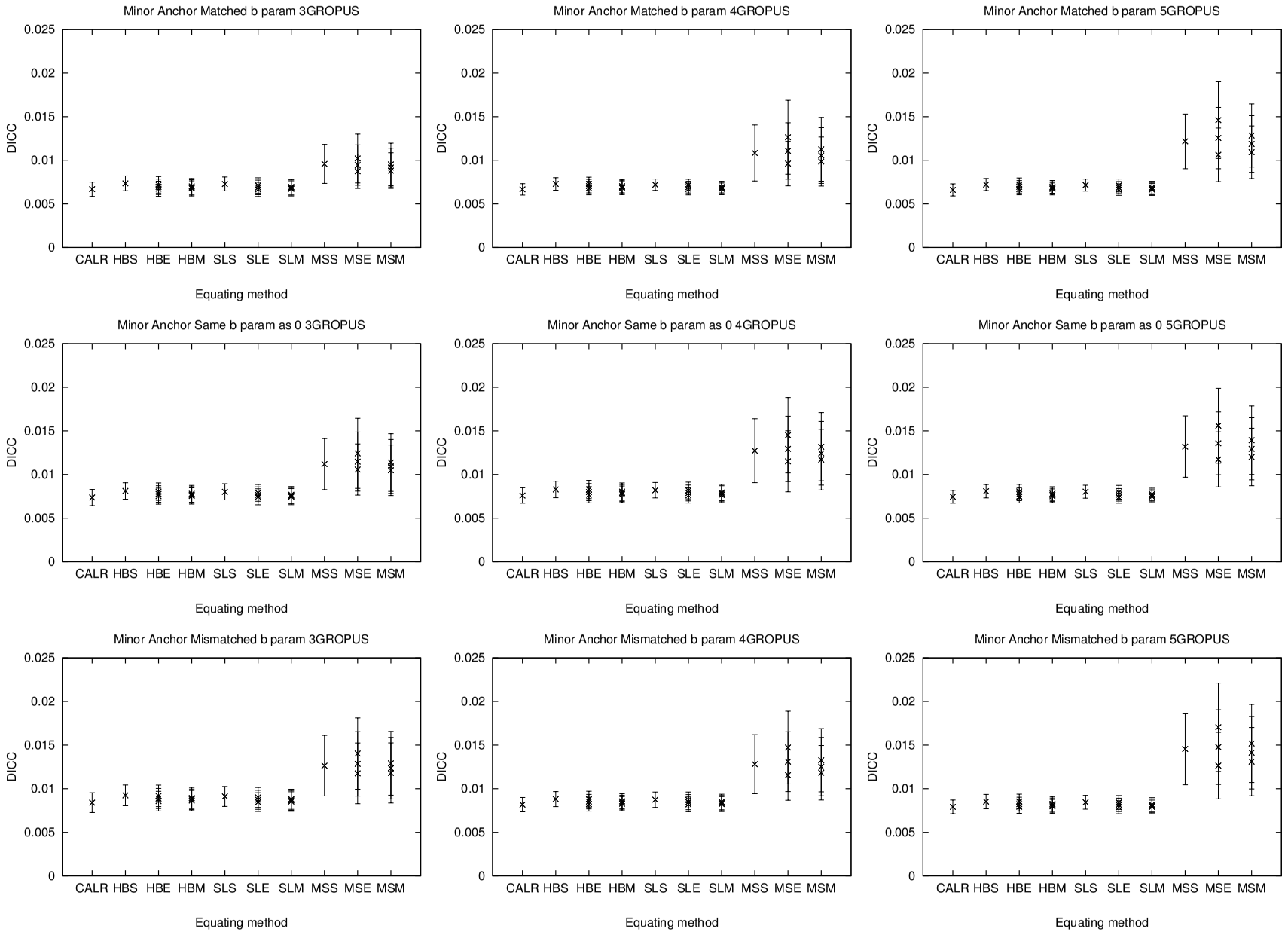

図 2.4　条件ごとの DICC の平均値（小アンカー条件）。各行（縦）は上から H1、H2、H3、各列は左から順に $g = 3, 4, 5$ の場合。HBM,HBE,SLM,SLE,MSM,MSE では、上から最大値、平均値、最小値の 100 回における平均を示す

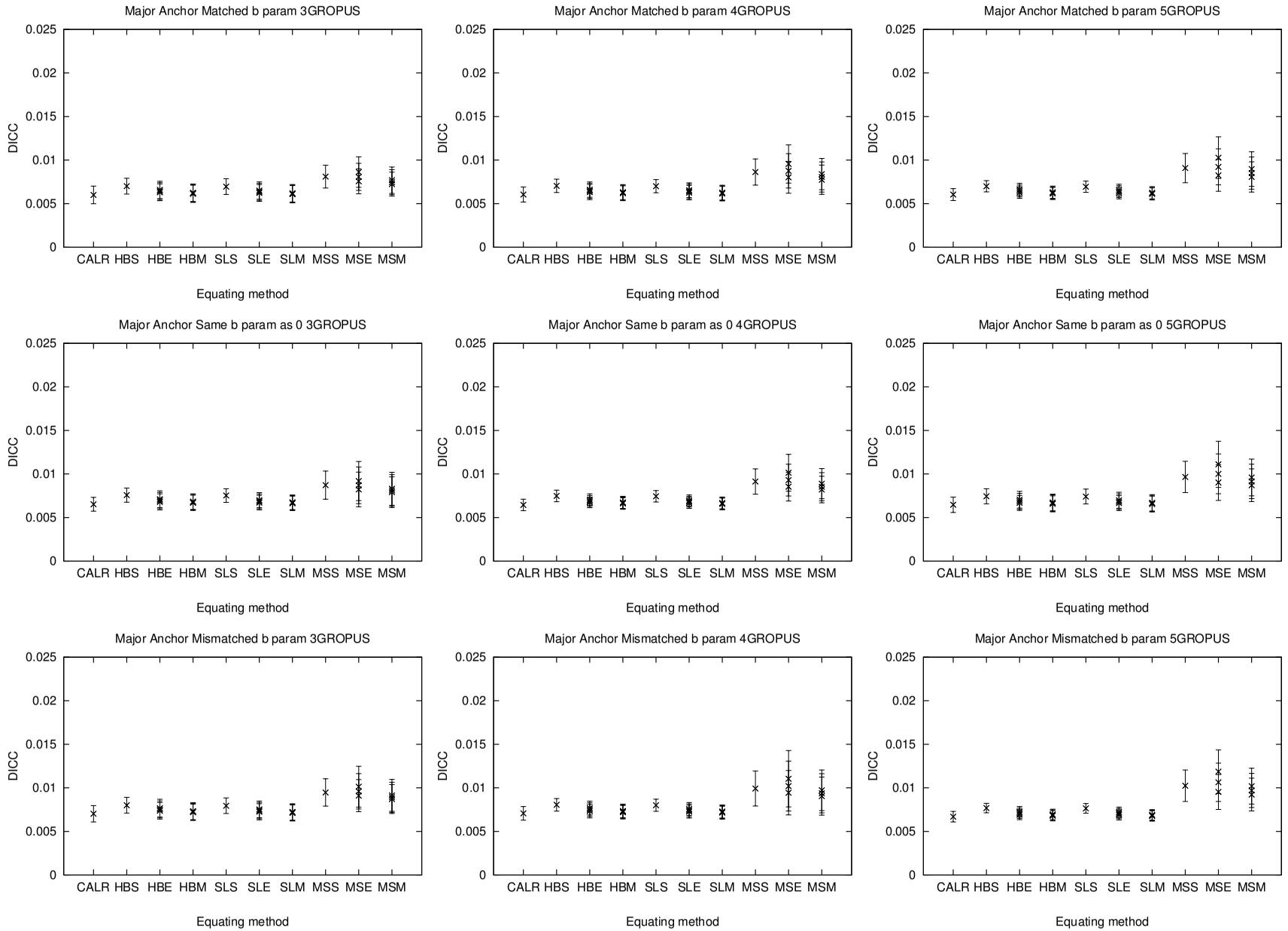

図 2.5　条件ごとの DICC の平均値（大アンカー条件）。 各行（縦）は上から H1、H2、H3、各列は左から順に $g = 3, 4, 5$ の場合。HBM,HBE,SLM,SLE,MSM,MSE では、上から最大値、平均値、最小値の 100 回における平均を示す

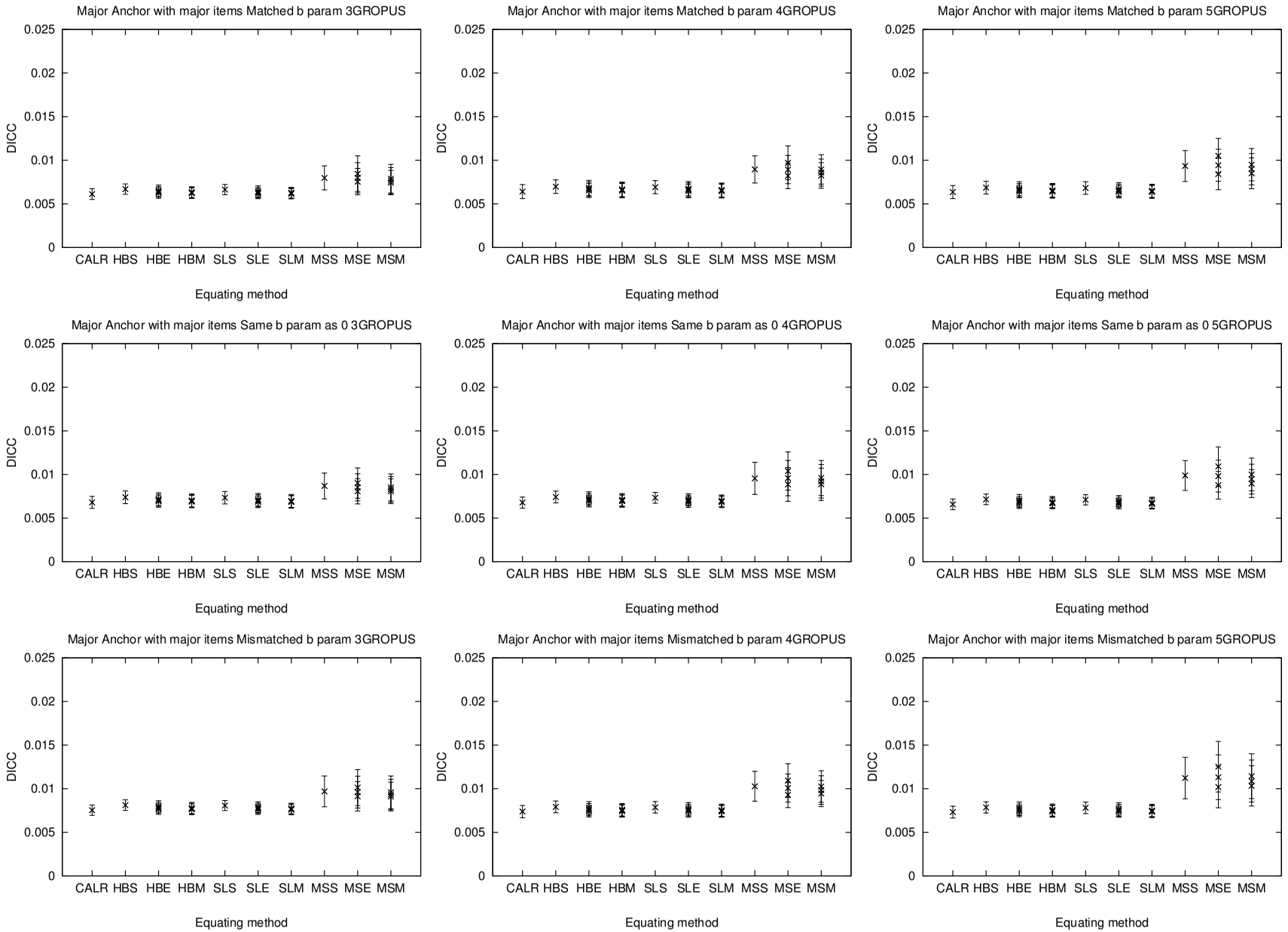

図 2.6　条件ごとの DICC の平均値（多項目条件）。 各行（縦）は上から H1、H2、H3、各列は左から順に $g = 3, 4, 5$ の場合。HBM,HBE,SLM,SLE,MSM,MSE では、上から最大値、平均値、最小値の 100 回における平均を示す

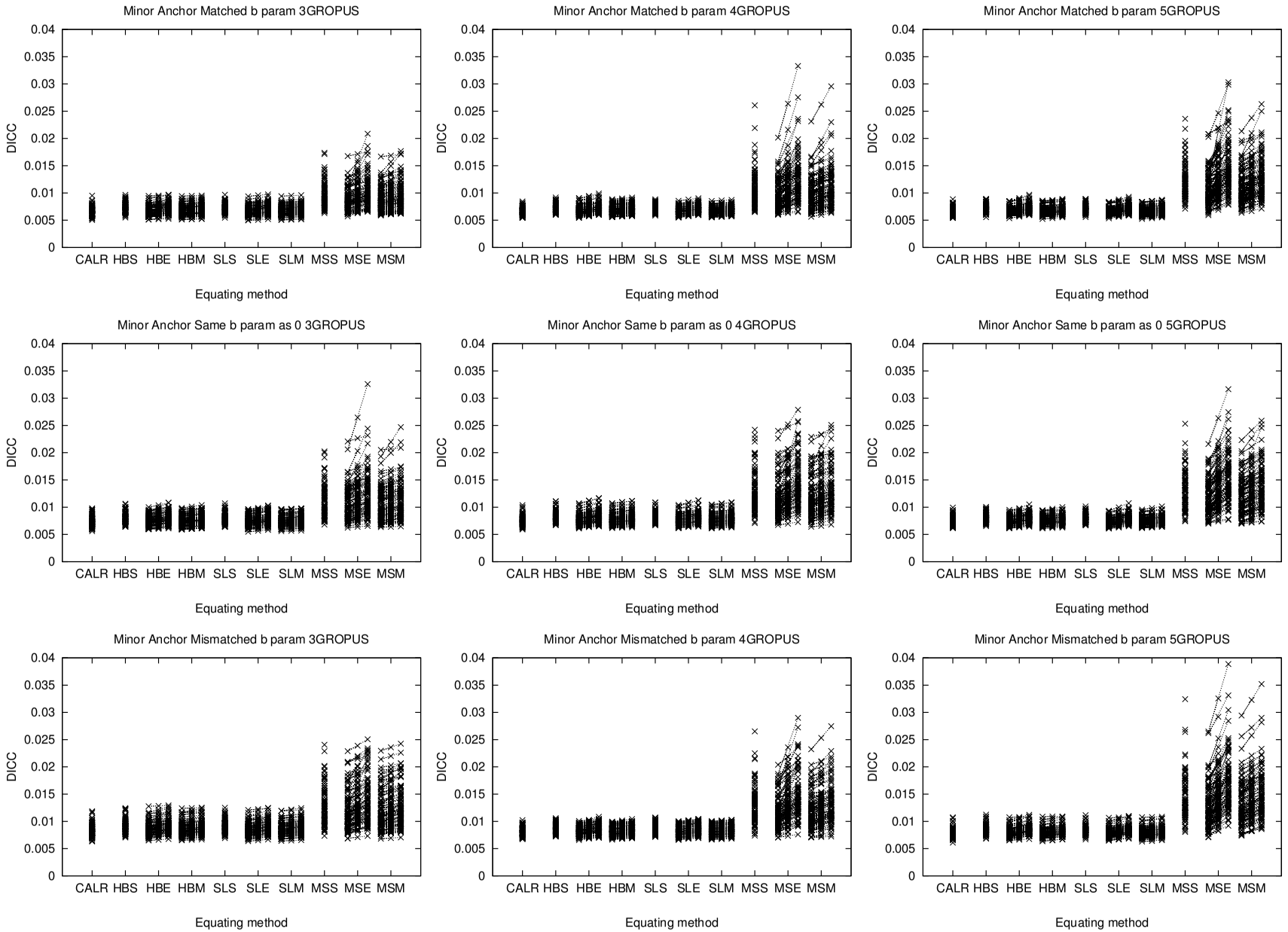

図 2.7　100 回の試行における、条件ごとの DICC の散布図（小アンカー条件）。各行（縦）は上から H1、H2、H3、各列は左から順に $g = 3, 4, 5$ の場合。HBM,HBE,SLM,SLE,MSM,MSE で 3 系列あるうちの左は最小値、中央は平均値、右は最大値を示す

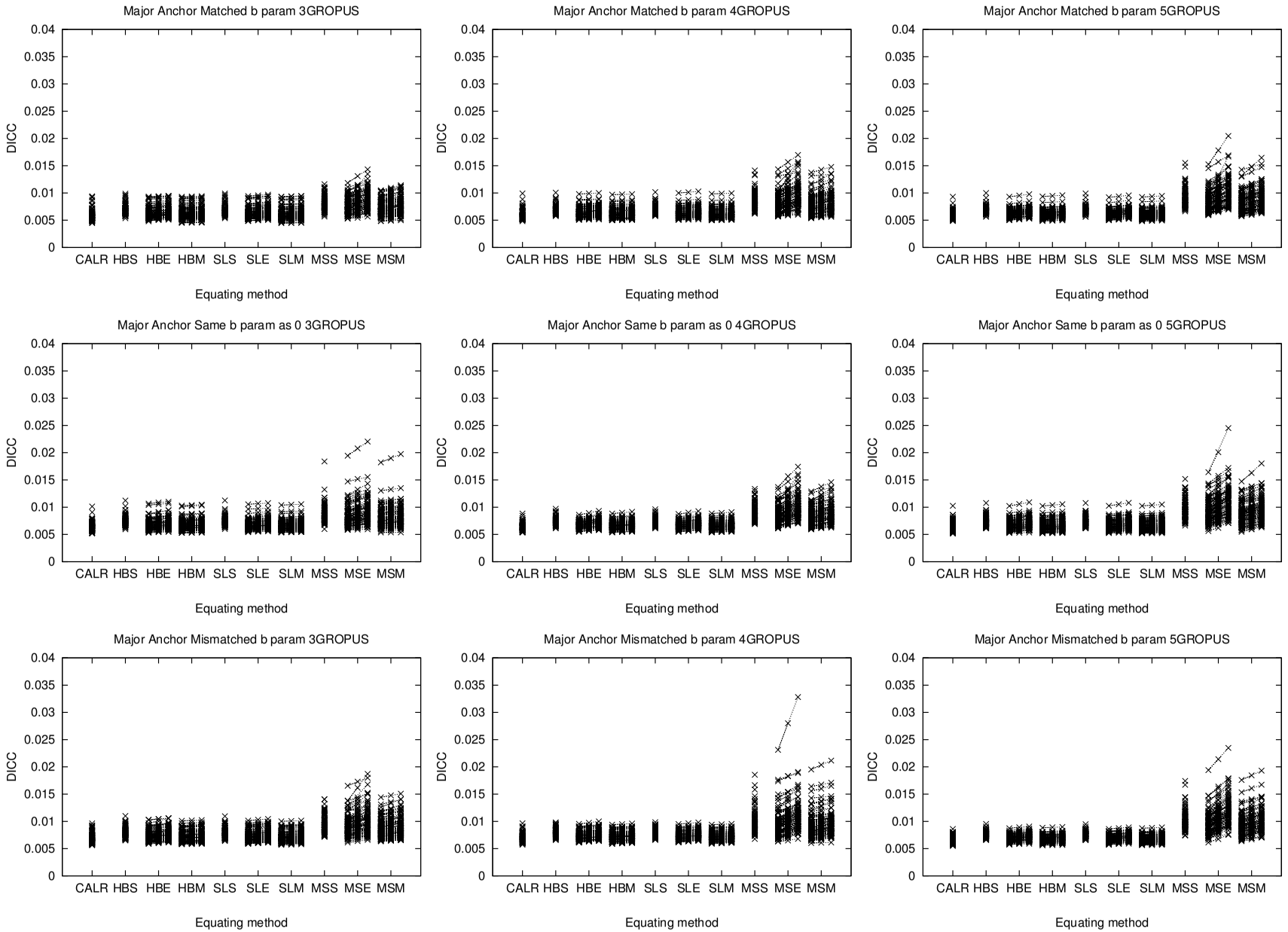

図 2.8　100 回の試行における、条件ごとの DICC の散布図（大アンカー条件）。各行（縦）は上から H1、H2、H3、各列は左から順に $g = 3, 4, 5$ の場合。HBM,HBE,SLM,SLE,MSM,MSE で 3 系列あるうちの左は最小値、中央は平均値、右は最大値を示す

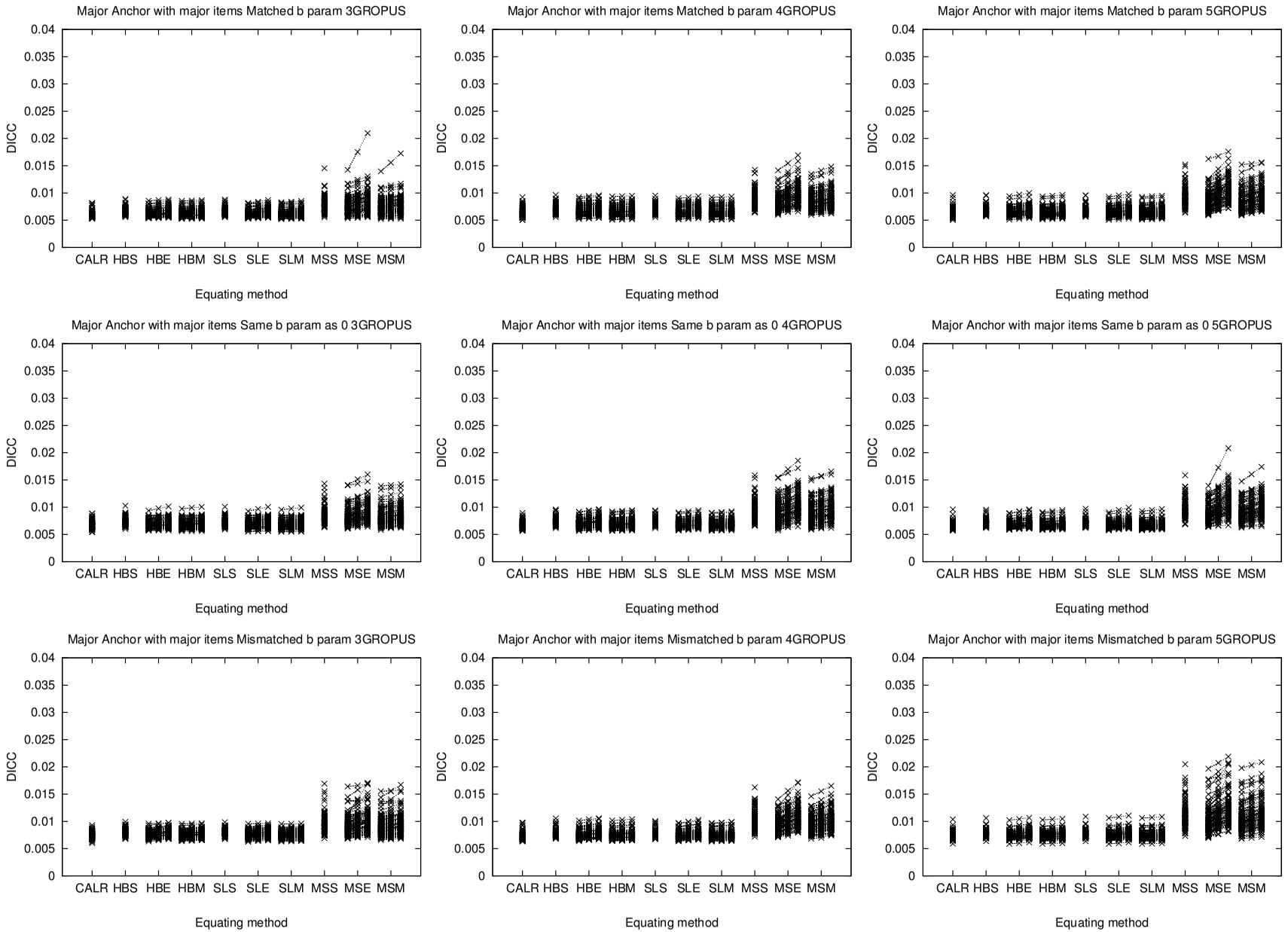

図 2.9　100 回の試行における、条件ごとの DICC の散布図（多項目条件）。各行（縦）は上から H1、H2、H3、各列は左から順に $g = 3, 4, 5$ の場合。HBM,HBE,SLM,SLE,MSM,MSE で 3 系列あるうちの左は最小値、中央は平均値、右は最大値を示す

2.4 考察

2.4.1 等化方法の違いについて

先行研究 (Arai and Mayekawa, 2011、Hanson and Béguin, 2002、Lei and Zhao, 2012) においては、個別推定時の項目パラメタ推定の精度に関する多くの比較が行われ、個別推定についてはMean/Sigma 法 (MS) の推定において真値と異なる結果となることが報告されていた。本研究においても、この結果と同様の傾向が見られた。また、Stocking-Lord 法 (SL) や Haebara 法 (HB)、calr 法は、MS に比べてより真値に近い推定結果を返している。これらはいずれも ICC 法であり、テストフォームに含まれる共通項目において、1 項目程度大きく等化元と等化先における項目パラメタの困難度が異なる場合であっても、その 1 項目に引きずられて等化後の項目パラメタが決まってしまうことがない。すなわち、ICC 法においては全ての項目における項目特性情報を「総合的に」用いて等化を行っていると言え、テストの運用をより安定的に行うことが可能であると期待される。

また、calr 法は、HB や MS において、「項目パラメタの平均をとる項目バンク更新法」の結果に近い DICC となった。このことは、HB や MS の方法しかとれない場合、たとえば、これまで HB や MS の方法でテストが等化され続けてきた場面などでは、項目パラメタの平均をとる項目バンク更新法を用いれば、ほぼ calr 法に近い結果を得ることができることを示している。ただし、項目バンク更新法の間の比較で、ICC 法を用いた場合、DICC の値に大きな差が見られていないことを考えると、実践的な意味での「等化の方法の妥当性」には、大きな影響はないといえる。

MS において、毎回の試験において等化元の困難度の推定値に極端な値が存在する場合、その推定値に引きずられた形の等化係数となってしまうという問題点（Hanson and Béguin,2002）が、MS の推定に誤差が大きかった原因ではないかと考える。たとえば、DICC の散布図において、多くの条件で、MS の方法のみで極端に大きな DICC となった試行が数件発生していることがわかる。これらの試行において DICC が大きくなった原因として、極端な困難度の値にひきずられた可能性が指摘できる。したがって、本研究の結果から、ICC 法が使用できる場合は、なるべく ICC 法を用いることが、テストを安定して運用するために必須であることが示唆される。

2.4.2 新作項目割合、および、項目数全体の影響

新作項目割合が大きい「小アンカー条件」においては、新作項目割合が小さい「大アンカー条件」に比べ、MS の方法をとった場合、DICC の値が大きくばらつく傾向が見られた。一方、他の方法をとった場合、等化の順序性の影響を含め、推定結果にばらつきが見られなかった。以上の結果より、MS の方法は、新作項目割合が大きい場合には、安定した結果を返さない方法であることが分かる。確かに、テストの現場では、簡易に実施可能な MS の方法を用いるという選択が行われる可能性がある。しかし、本研究の結果からは、テスト実施のたびに等化を行うようなデザインの場合、MS の方法を用いるべきではないことが指摘できる。

また、項目数の大きな「多項目条件」と「小アンカー条件」を比較した場合、項目数の大きさがDICC の大きさに影響を及ぼすという結果であった一方で、フォーム数はほとんど影響を及ぼさなかった。このことは、項目バンクのサイズを大きくする際、試験の回数を多くするのと、1 フォームあたりの項目数を増やす場合では、後者の方がやや推定が不安定になる傾向があることを示している。しかし、実際に試験の長さを決めるのは、推定の安定性だけではなく、受験者にかける労力や時間の制約など、多岐にわたる。本研究の結果からは、項目数を増やすことによって、これらの制約を覆すほど、重大な推定の不安定さをもたらすという証拠が見出されたとは言えないと考える。

2.4.3 項目バンク更新法について

ICC 法、また MS 法のいずれにおいても、テストのフォーム数が増えても、推定結果の DICC に大きな影響が見られなかった。このことは、特に「常に規準集団の項目パラメタを記録 (S)」や「常に本試験の項目パラメタを記録 (E)」した場合に、フォームの数が増える、すなわち、等化の回数が増えた場合であっても、推定に影響がないことを意味する。したがって、ペアワイズに等化する場合は、等化の回数を考えなくてもよいことや、DICC に影響する因子は等化の回数以外の要因、たとえば等化方法や項目バンクの更新法、元の困難度の条件に依存することが示唆される。等化を繰り返す場合に、それぞれの等化の前提となる等化元の項目パラメタが正しく（真値に近い形で）推定され、それが正しく（理論通りに）等化される場合は、フォームの数が増えても、それにより項目バンク全体の推定結果に及ぼす影響は小さいと考える。

項目バンク更新法のうち、「規準集団の値を使い続ける方法」や「本試験の値を使い続ける方法」に比べて「平均をとる方法」がより真値に近い結果となった。このことは、項目バンクを構築するうえで、平均をとる方法が、等化元と等化先のパラメタそれぞれを平等に扱っているためであると考える。また、「平均をとる方法」に近い結果を返す calr 法は、項目バンクを構築するような目的の等化の場合、等化順を考えなくともよい方法であるので、項目パラメタが一意に定まり、しかも DICC が小さい結果となるという点で、より好ましい方法であると考える。

もっとも、先に述べたように、ICC 法による等化の順序性の影響は、あまり大きいとは言えない。しかし、テストの実践的な場面においては、「ルールとして定められた方法による運用」が求められるという側面がある。この場合、「各フォームを個別に項目パラメタ推定し、それぞれのフォームに含まれる共通項目を用いて calr 法で等化する」という単一のルールさえ用意できれば、等化の順序性による（小さいとはいえ）差が生じることなく等化を行うことができる calr 法が、このようなテストデザインの場合、最も好ましい方法であるといえるだろう。

一方、MS においては、等化の順序によって大きく結果が異なっている。このことは、DICC が極端に大きくなるような等化順が存在する場合、言い換えれば少なくとも一つのフォームに極端な困難度を持つ項目が存在し、そのフォームが等化元となって等化が行われ、等化結果の値が項目バンクに反映された場合、MS の方法は不適切な結果を返す可能性があることを示している。本研究では、極端な困難度を持つ項目を考えていないが、実際のテスト場面においては、「困難度の絶対

値が大きい」項目が等化元となる場合が少なくないことが予想される。このような場合は、等化順の影響を避ける意味でも、ICC 法をとることが好ましいといえる。

2.4.4　今後の課題

本研究では、人工的に生成したシミュレーションデータに基づき、等化を繰り返すという方法をとった。したがって、実際のデータを用いた際に現れるような、極端な困難度の値を持つような場合の検証が不十分であると考える。特に、ICC 法の結果に関しては、極端な項目パラメタが等化元あるいは等化先となった場合の推定結果の検証が不可欠であると考える。困難度に外れ値を持つ項目パラメタによる等化方法の検討については、Hu, Rogers and Vukmirovic (2008) にあるように、人工的に項目パラメタの値を生成する研究もあるものの、実データに基づく方法がより実践場面では有効であると考える。

また、本研究においては、FCIP 法との比較を行っていない。FCIP 法は、本研究におけるペアワイズ等化法と同等の手続きを、IRT の項目パラメタ推定の手続きで行うことができる方法である。本研究のテストデザインにおいて FCIP 法を用いた場合と比較することも、今後の課題である。

第 3 章

IRT を用いた大規模テストにおける個別推定と同時推定とのパラメタ比較

3.1 序論

前章では、個別推定を行う場面において、等化の順序性が calr 法を用いることで回避できることを示した。一方で、同時推定においては、等化の順序性に関する問題は発生せず、また、規準集団を IRT のモデル上で表現することができるという利点がある。この 2 つの方法を、1.5.1 節で述べたテストデザインに適用した場合に、等化後の項目バンクにおける項目パラメタがどのような値を示すかを、シミュレーションによる研究を通じて検討する。特に、新作項目割合が項目パラメタの推定に及ぼす影響について検討する。

3.1.1 先行研究、その問題点

1.5.1 節によって提案されたテスト実施法においては、(1) 等化の方法をどうするか、特に個別推定と同時推定のいずれをとるか、(2) 予備試験と本試験で同様な項目特性となるような項目が作成できない場合、等化後の項目特性の推定値にどのような影響がみられるか、といった点が未検証であった。Kolen and Brennan (2004) には、本研究で扱うテスト実施法に類似する方法の紹介がある (pp. 201-207) ものの、等化の詳細な手続に関する統計学的検討はわずかである。

検証すべき点のうち (1) は、異なる θ の分布をもつような複数のグループがアンカー項目を含む複数のテストフォームを受験している場合の共通項目デザインによる等化を行う研究（common-item nonequivalent groups design）において、等化に使用するソフトウェアの種類と同時に比較されるトピックである（Kang and Petersen (2009)、Hanson and Béguin (2002)、Li, Tam and Tompkins (2004) を参照。また実践例では Briggs and Weeks (2009) や Pang, Madera, Radwan and Zhang(2010) などが挙げられる）。また、Lee and Ban (2010) は、実際の試験で得られたデータを用いて、randomly equivalent とされる 2 グループ間の等化を個別推定と同時推定で行い、個別推定がより誤差の少ない等化が行えるとした。ただし、本研究で述べるような共通項目に

よる等化デザインをとっていない。また、Jodoin, Keller and Swaminathan (2003) は、同一の実データに個別推定と同時推定、および FCIP 法による分析を行い、推定された等化後の項目パラメタが異なることを指摘している。また、Arai and Mayekawa (2011) は、「項目パラメタ既知」のアンカー項目を手掛かりに予備試験の受験者集団へ等化する試験デザインに関するシミュレーション研究を行っている。この研究では予備試験実施後に 2 種の本試験を実施した場合を想定し、(a) 同じアンカー項目を 2 種の本試験フォームに共通して用いるデザイン、(b) 2 種の本試験フォームでそれぞれ別のアンカー項目を用いるデザイン、それに (c) 第 1 種目の本試験フォームのみ予備試験から選抜し、第 2 種目は 1 種目の本試験フォームからアンカー項目を選抜するデザイン、の 3 種の条件で個別推定と同時推定、それに FCIP の推定法の間で比較を行った結果、(b) のデザインがどちらの推定方法においても安定して真の項目パラメタ値に近いパラメタの推定値を得た。しかしながら、実際の大規模試験を運営するに当たり、試験を重ねていく間を通して毎回のアンカー項目の選定方法を固定することは現実問題として難しいといえる。

また、検証すべき問題 (2) は、本研究で扱うテストデザインが、予備試験から一貫して一定の品質の項目を本試験で作成することを前提としているため、予備試験及び本試験の項目特性が項目パラメタ推定に影響をおよぼすかを見極める必要がある（IRT に基づくテストにおける項目特性に関しては、2.1 節で詳細を述べる）。項目パラメタのうち困難度については、項目の作成者、テストフォームの編集者が経験的に特定の項目の難易度をテスト実施前に判断しうるが、識別力に関しては実施前の判断が難しい。したがって、識別力に関して、予備試験と本試験で異なる特性であった場合に、特定の等化方法を用いることによって項目バンク上の推定に影響が及ぶことは、項目特性の評価上、好ましいことではない。実際の試験場面において、テストを立ち上げる段階（予備試験実施の段階）では、テストで問うべき観点が多様な項目によってカバーできているが、試験実施につれて既出項目と類似の項目が過剰に出題されることを避けるあまり、テストで問うべき概念を直截的に問うのではなく、遠回しに問うことが多くなっていき、結果として本試験の識別力が予備試験に比べて低くなることが想定できる。逆に、予備試験の段階で、測定すべき概念の統一がとれておらず、本試験の段階においてはある程度の水準の識別力が得られる、という場面も想定できる。この場合、予備試験のみで、識別力が低い項目が多くみられることになる。項目の識別力に関するこれら 2 つの場合は、いずれもテストを実施してみないとわからないので、等化方法をテスト実施前に決定する必要がある以上、それぞれの等化の方法に関する性質を前もって明らかにしておく必要があるといえる。

3.1.2 研究の目的

本研究では、「予備試験ののち、新作項目を多く含んだ本試験を 2 種のテストフォームを用いて行い、それぞれのテストフォームを予備試験に等化して項目バンクに登録する」ような大規模テスト場面を模したシミュレーションを行い、個別推定と同時推定のいずれが真値に近い値となるかを、項目パラメタの種類（識別力および困難度）ごとに検証する。同時に、先の (2) で示した要因の効果を検証する。さらに、受験者数の規模の違いによって、また、アンカー項目の数、あるいは

毎回の試験で出される新作項目数の違いによって、項目パラメタの推定値に違いがみられるかも検証する。

Hanson and Béguin (2002) や Arai and Mayekawa (2011) において、受験者数の効果を検証しているように、テスト実施機関の立場であっても、受験者数が増えれば項目パラメタの推定精度がどの程度向上するかという点に関心を持つことが多い。たとえば Arai and Mayekawa (2011) では、受験者数を 4 倍（500 名から 2000 名）に増やした条件では、等化の方法にかかわらず、フォーム内に含まれる項目の ICC の差異の和に関して、真値との誤差がほぼ半減するという結果が得られた。このような関係が、本試験を複数回実施した場合にもあてはまるかどうかを検討する。また、本試験におけるアンカー項目の数は、同様にテスト実施機関が操作しうる事項の一つであり、テスト実施上の制約を受けやすいことから、Arai and Mayekawa (2011) においても独立変数の一つとしてその効果が検討されている。とりわけ本研究のテスト実施法の場合、アンカー項目数を減らすことは、本試験の新作項目を増やすことにつながるため、アンカー項目を少なくした場合の検討を行う。

さらに、テスト実施機関としては、短期間になるべく大規模の項目バンクを整備することを志向し、1 回の本試験でなるべく多くの新作項目を出題する方略をとることが想定できる。そのため、新作項目数が増えた場合の効果を検証する。Lee and Ban (2010) や Arai and Mayekawa (2011) においては、フォーム全体の TCC (Test Characteristic Curve) またはフォーム内に含まれる項目の ICC の和を比較していたが、本研究では、項目パラメタの値に関する直接的な真値と推定値のずれを評価指標とする。これは、(1)TCC や、テスト情報量曲線を用いた項目選択は、テスト実施機関が編集した単一フォームの特性を記述するうえで重要であるものの、テスト実施機関が項目バンクの中に含まれる全項目にわたる大まかな傾向を知る際に、項目バンク内に含まれる個々の項目のパラメタ値に着目した議論を行うことが多い点、(2) 特定の θ の分布をもった受験者グループに対して情報量が多くなるようなフォームを TCC やテスト情報量曲線に基づいて作成する項目選択方法は、本試験においてアンカー項目よりも項目パラメタ未知の新作項目が多い本研究の実施法では不可能である点、(3) 本研究のテスト実施法では、テスト実施機関は本試験実施後に次回本試験のアンカー項目を少数（10 項目程度）選択するが、その際に参考となる統計指標は TCC よりむしろ項目パラメタそのものの値であること、を考慮してのことである。なお、本研究では、1 回の実施で 1 つのフォームを提示することは必ずしも前提とはしていない。複数のテストフォームを同一時点に別々の受験者グループに提示する場面も想定している。したがって、Arai and Mayekawa (2011) とは異なり、常にその時点で識別力が良好な項目をアンカー項目として抽出することを想定した。

ただし、本研究では、FCIP 法を扱っていない。これは、項目数やフォームの数、受験者のグループの数がテスト実施のたびに増えるようなテスト実施法においては、FCIP を用いた分析を行うことで推定が不安定となる可能性があり、実践場面において適用できるかどうかが不明なためである。個別推定、および同時推定は、いずれも推定方法の過程で安定して推定が可能であり、これらの方法を本研究では比較することにした。

3.2　方法

3.2.1　シミュレーションで仮定したテストフォーム、受験者集団、モデル

毎回の試験場面は、規準集団（等化先）を定義するための「予備試験」およびそれに続いて受験者にスコアを返す「本試験」（第 1 回試験、第 2 回試験）の計 3 種のテストフォームから構成した。これらの項目は、毎試験後に 2PL を当てはめて IRT に基づく分析を行うものとした。

本試験に先立ち、予備試験を実施したと仮定した。予備試験は、基準集団における項目特性が既知である項目を作成して項目バンクに入れるために実施された。予備試験実施後、2 度の本試験を実施した。本試験にはそれぞれ項目バンクに入っている項目を「アンカー項目」として含むものとした。本試験には、さらに、その試験実施時点では項目特性が未知である「非アンカー項目」を含むものとし、毎回の試験においてはアンカー項目と非アンカー項目を同時に受験者に提示し、反応を得たと想定した。

予備試験のフォームおよび受験者集団

予備試験として、30 項目を想定した。30 項目の真の識別力を平均 $\log(\alpha)$ で標準偏差 0.2 の対数正規分布に従う乱数から発生させ、真の困難度を平均 0、標準偏差 1 の正規分布に従う乱数から発生させた。これらの項目に対し、真の $\theta \sim N(0,1)$ をとる受験者 N_{trial} 人からの反応を得たものとした。

第 1 回・第 2 回試験のフォームおよび受験者集団

第 1 回試験、および、第 2 回試験においては、項目バンクから「項目パラメタ既知の項目」を J_{anc} 項目、3.2.4 節に述べる「アンカー項目抽出ルール」にしたがって抽出し、それに第 1 回試験実施時点の「新作項目」を J_{new} 項目追加した（$J_{anc} + J_{new}$）項目を用いた。新作項目の真の識別力を平均 $\log(\delta)$ で標準偏差 0.2 の対数正規分布に従う乱数から発生させ、真の困難度を平均 0、標準偏差 1 の正規分布に従う乱数から発生させた。これらの項目に対し、第 1 回試験では真の $\theta \sim N(0.2,1)$ をとる受験者 N_{exam} 人からの反応を得たものとした。また、第 2 回試験では真の $\theta \sim N(0.4,1)$ をとる受験者 N_{exam} 人からの反応を得たものとした。いずれも、試験実施とともに受験者の能力が向上する場面を想定した。

3.2.2　1 回のテスト場面におけるシミュレーション手続き

予備試験の実施

予備試験実施によって、予備試験のテストフォームが予備試験の受験者グループに出題され、反応パタンから各項目の項目特性が推定された。次に、推定された項目特性の値を項目バンクに入れた。

第 1 回試験の実施

第 1 回試験実施に先立ち、「アンカー項目抽出ルール」に基づき、アンカー項目が選抜された。アンカー項目と新作項目からなる第 1 回試験フォームが、第 1 回試験の受験者に提示され、得られた反応パタンから「第 1 回試験単独の項目パラメタ」が推定された。次に「第 1 回試験単独の項目パラメタ」に含まれるアンカー項目の項目パラメタと、アンカー項目の項目バンク上での項目パラメタを用いて、第 1 回試験を予備試験に等化した。最後に、等化済みの項目パラメタを「項目パラメタ既知の項目」として項目バンクに入れた。その際、アンカー項目において、予備試験における項目パラメタと等化済みの項目パラメタが異なる場合、両者の平均値を入れた。Kolen and Brennan (2004) は、項目バンク内に既に存在する項目パラメタの値と異なる値が存在する場合には、テスト実施前にルールを考えておく必要がある (p.204) としているが、具体的な手続きには触れていない。また、前章にて議論した通り、項目バンク更新の方法の違いは、等化方法によっては大きな違いを及ぼさないことがわかっている。そこで、実際のテスト場面において、最も公平に項目パラメタの値を推定しているであろう値として、識別力、困難度ともに、平均値を項目バンクに入れることとした。

第 2 回試験の実施

第 2 回試験の前に「アンカー抽出ルール」に基づき、アンカー項目が選抜された。アンカー項目は、第 1 回試験が初出となる項目を含む、第 1 回試験実施後時点での項目バンクから抽出するものとした。第 1 回試験と同様、第 2 回試験においてもアンカー項目と新作項目からなるテストフォームが第 2 回試験受験者に提示され、反応を得た。次に、反応パタンから「第 2 回試験単独の項目パラメタ」が推定された。さらに「第 2 回試験単独の項目パラメタ」に含まれるアンカー項目の項目パラメタと、同一項目の項目バンク上の項目パラメタを用いて、第 2 回試験を試行試験に等化した。最後に、等化済み項目パラメタを項目バンクに入れた。アンカー項目において、項目バンク上の値と異なる等化済みパラメタの値が得られた場合は、両者の平均値を入れた。

3.2.3 等化手順

等化手順として、以下に示す「個別推定」「同時推定＋規準集団への等化」「多群 IRT モデルを用いた同時推定」をそれぞれ行った。

個別推定 (separate calibration)

3.2.2 節の 1 回のテスト場面における手続きをすべて終えたのち、項目バンク上にある項目パラメタは、すべて予備試験での θ のスケールで比較可能になっているので、これをもって、個別推定の結果とし、記録した。予備試験から第 2 回試験までの、個別推定の手続きを図 3.1 に示した。

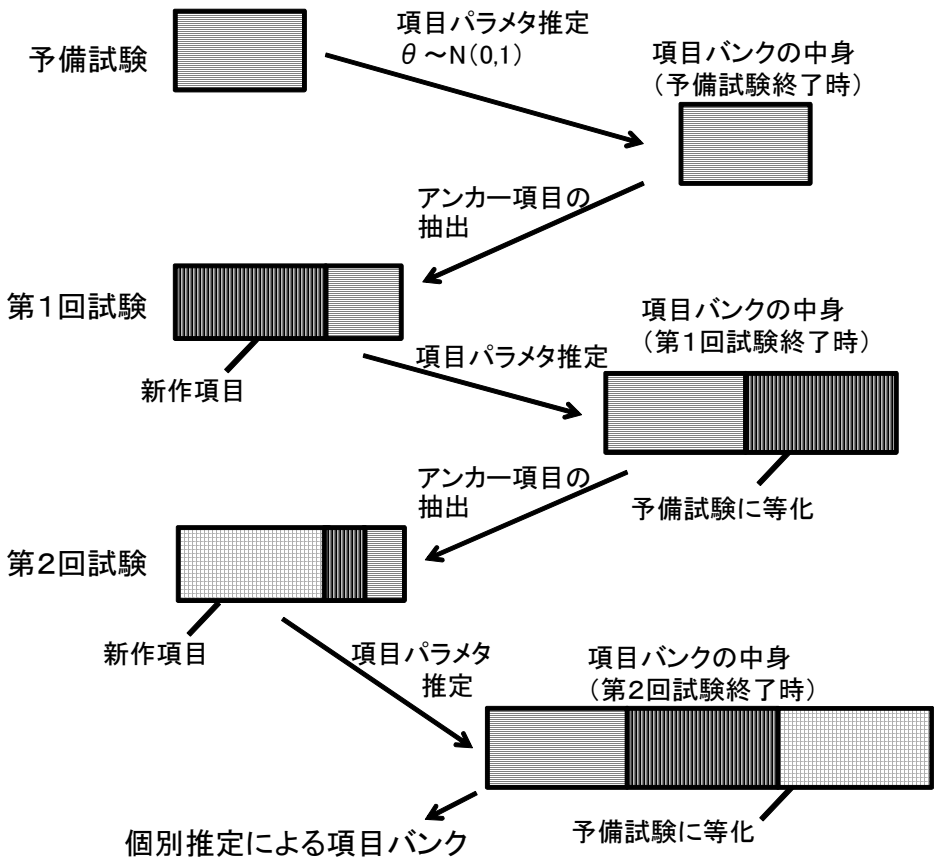

図 3.1　個別推定における項目バンク構築手順。同じパタンの枠は同一の項目であることを示す

同時推定＋規準集団への等化 (concurrent calibration with equating into norm group)

第 2 回試験終了後、予備試験、第 1 回試験、第 2 回試験のすべての受験者で記録された反応パタンを、同一項目が縦に並ぶように連結した 0-1 データセット（以下、「同時推定用データセット」と呼ぶ）を作成した。この 0-1 データセットに対して 3 つの試験の受験者をそれぞれ別々のグループであるとみなした multi-group model を用いて項目パラメタを一括して推定した。その際、θ の分布は、全受験者において平均が 0、標準偏差 1 になるように推定した。これにより得られた項目パラメタは、θ の分布が規準集団のスケールに載っていないので、規準集団のスケールで比較可能な項目パラメタに変換する必要がある。そこで、同時推定用データセットを用いて推定した項目パラメタのセットに含まれる予備試験 30 項目の項目パラメタを、予備試験単独で実施した際に推定された項目パラメタに等化する処理を行った。この処理によって、同時推定用データセット内にあるすべての項目、言い換えれば項目バンク内のすべての項目のパラメタが、予備試験の項目パラメタに等化されることとなった。このようにして算出した等化済みの項目パラメタを「同時推定＋規準集団への等化」によるパラメタ推定値と定義し、記録した。

なお、同時推定に関しては、第 2 回試験までの手続き終了後の 0-1 データセットに対して行った。すなわち、アンカー項目の抽出ルールは個別推定の項目パラメタに対して適用されている。これは、Arai and Mayekawa (2011) などにおいて個別推定の方が真値に近い項目パラメタを得たとする先行研究から、同時推定において何らかのバイアスがかかった項目バンクが第 1 回試験において得られた場合、その効果が第 2 回試験のアンカー項目抽出において影響することを考慮しての措置であった。以上の手続きを図 3.2 に記した。

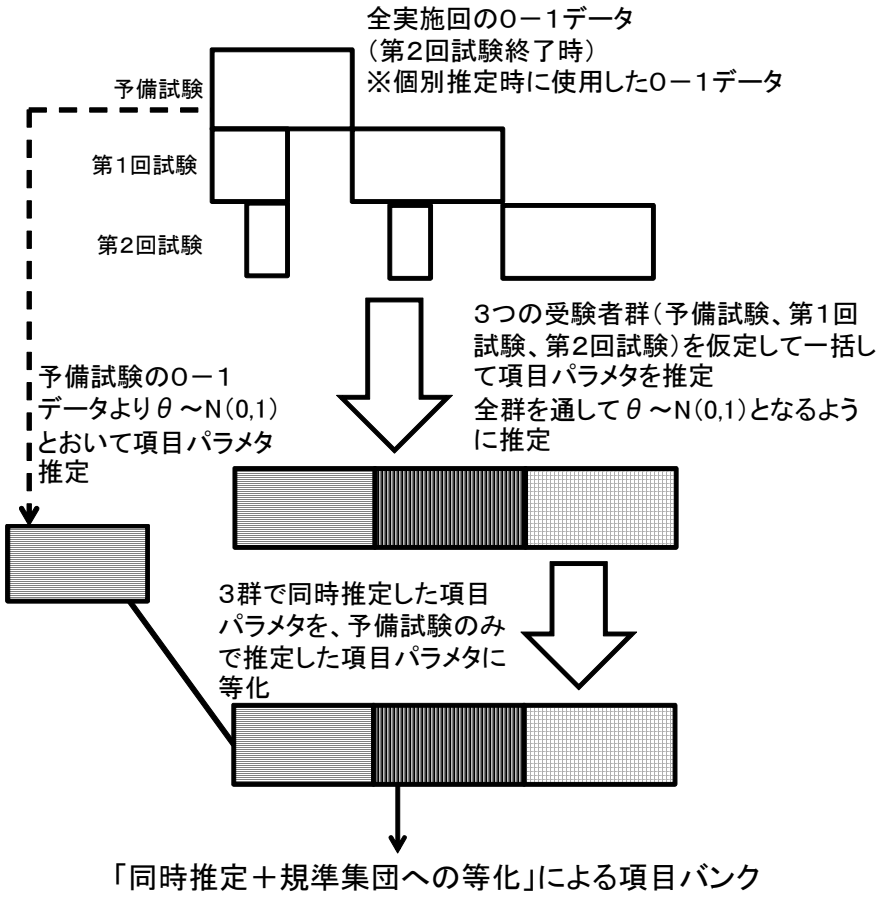

図 3.2　「同時推定＋規準集団への等化」による項目バンク構築手順。白枠は 0-1 データの並びであることを示し、同じパタンの枠は同じ項目であることを示す

多群 IRT モデルを用いた同時推定 (concurrent calibration using multi-group IRT model)

第 2 回試験終了後、同時推定用データセットに対して、multi-group model を用いて項目パラメタを一括して推定した。その際、予備試験に該当するグループのの分布の平均を 0、標準偏差を 1 とするように項目パラメタを推定した。本研究の場合、予備試験における θ の分布に関しては、真の平均が 0、標準偏差が 1 とおいたため、この推定の結果得られた第 1 回試験、第 2 回試験の θ の分布は、予備試験の受験者を基準として相互に比較可能になっていると解釈できる。この方法で推

定した項目パラメタを「同時推定 MG」と定義した。この方法においても、「同時推定＋規準集団への等化」と同様、項目パラメタの一括推定は個別推定時の 0-1 データに対して行った。以上の手続きを図 3.3 に記した。

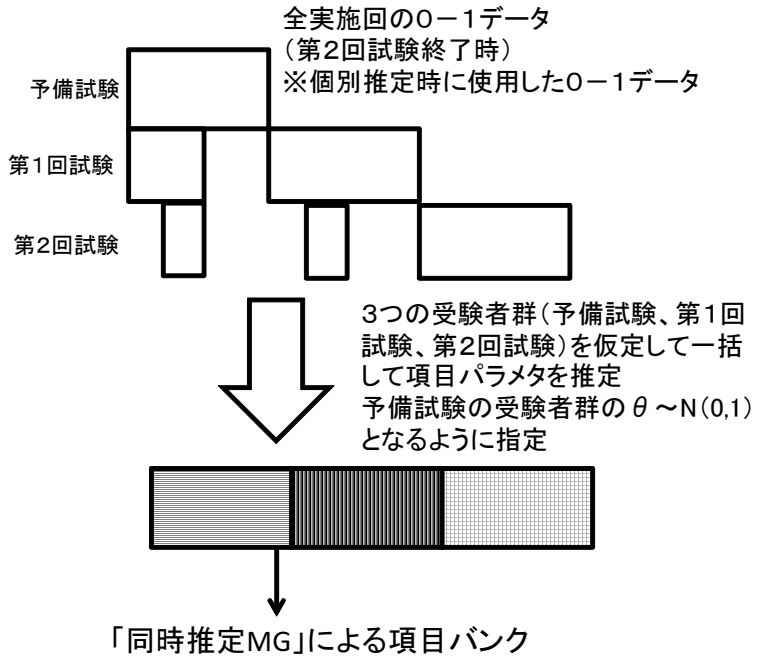

図 3.3　「同時推定 MG」による項目バンク構築手順。白枠は 0-1 データの並びであることを示し、同じパタンの枠は同じ項目であることを示す

3.2.4　アンカー項目抽出ルール

本研究の目的にしたがい、項目バンクに登録されている全項目から、識別力が高く、また項目バンク全体の縮図となるような項目をアンカー項目として抽出するような以下の方法が用いられた。まず、項目バンクの中身を困難度の値の大きい順に並べ、次に、困難度の値の順に（項目バンクのサイズ ÷ アンカー項目数）項目ずつ区切り、それぞれの区切りの中から最も識別力の大きな項目をアンカー項目として抽出した。

3.2.5　項目パラメタ推定、および等化方法

項目パラメタの推定にはすべて BILOG-MG 3 を用いた。すべての項目パラメタ推定において、CALIB コマンドの NOSPRIOR、NOTPRIOR を指定した。これにより、事前分布は適用されなかった。個別推定の際、等化方法として calr 法を用いた。この方法は、Arai and Mayekawa (2011) において、個別推定時に最も真値に近い項目パラメタを得たとされた方法であり、前章の研究においても真値に近い等化結果を得ることができ、さらに複数のフォームを同時に等化するという場面である

ことから、本研究で用いられた。また、「同時推定＋規準集団への等化」における項目パラメタを推定する際、同時推定用データセットに対し、BILOG-MG 3 の CALIB コマンドの REF=オプションで値に 0 を指定し、受験者全体で θ の平均が 0、標準偏差が 1 になるように推定したうえで、calr 法により予備試験の受験者のみで得られた項目パラメタに等化した。一方、「同時推定 MG」で項目パラメタを推定する場合は、REF=の値に予備試験の受験者集団を示す番号を指定して推定したうえで、得られた推定値をそのまま項目バンク上での等化済み項目パラメタとした。

3.2.6 シミュレーションデザイン

アンカー項目数 J_{anc} は（5,10）の 2 条件とした。一方、2 回の本試験の新作項目数 J_{new} を（10,20,40）の 3 条件とした。これにより 2 度の本試験終了後の項目バンクのサイズは (50,70,110) 項目となった。また、予備試験での識別力 α の値が (0.5, 1.0) の 2 条件、また本試験での識別力 δ の値が (0.5, 1.0) の 2 条件をそれぞれ考え、これらをクロスさせた 4 条件を実行した。真値において θ の平均が 0、標準偏差が 1 なので、識別力 0.5 は中程度の識別力の項目として、また識別力 1.0 は高い識別力を持つ項目としてこれらの値を設定した。予備試験の受験者数 N_{trial}、および本試験の受験者数 N_{exam}（2 回とも同じ受験者数）に関して、小集団条件として $(N_{trial}, N_{exam}) = (1000, 2000)$、大集団条件として $(N_{trial}, N_{exam}) = (4000, 8000)$ の 2 条件を考えた。全部で $(2 \times 3 \times 2 \times 2 \times 2)$ の 48 条件について、それぞれの条件で 100 回ずつの試験実施場面を実行し、それぞれについて 3 種類の等化済み項目パラメタを項目バンク内にあるすべての項目について求めた。1 回の試験場面ごとに、予備試験、第 1 回試験、第 2 回試験それぞれで独立した受験者の θ の分布が仮定された。また、ランダムな項目パラメタの真値、およびランダムな 0-1 データの生成には RESGEN4 を用いた。

3.2.7 従属変数

J 項目からなる項目バンクにおいて、識別力、困難度それぞれの項目パラメタの j 番目の推定値 $\hat{\xi}_j$ に関して、その真値 ξ_j との差を評価するために、RMSE（Root mean squared error）および真値との差の値（平均差 Mean Difference; MD）として定義した。各条件において l 回目のテスト場面での RMSE の値を $RMSE_l$、真値との平均差を MD_l とし、テスト場面ごとに式 3.1、式 3.2 に基づいて差の指標を計算した。RMSE は真値と推定値のずれの大きさを評価する値として、また真値との差の値は真値と推定値のずれの方向性を検討する値として算出した。

$$RMSE_l = \sqrt{\frac{\sum_{j=1}^{J}(\hat{\xi}_j - \xi_j)^2}{J}} \tag{3.1}$$

$$MD_l = \frac{1}{J}\sum_{j=1}^{J}(\hat{\xi}_j - \xi_j) \tag{3.2}$$

これらの値については、「$\hat{\xi}_j$ を a_j の推定値および ξ_j を a_j の真値と考えた」場合を「識別力の推定結果」として、また「$\hat{\xi}_j$ を b_j の推定値および ξ_j を b_j の真値と考えた」場合を「困難度の推定結

表 3.1　識別力パラメタまたは困難度パラメタの項目バンク上の推定結果が、真値から 2 以上隔たったために、結果の算出から除外された試行数。推定方法ごとに示した。表中「同時規準」は「同時推定＋規準集団への等化」を、「同時 MG」は「同時推定 MG」を、それぞれ表す

J_{anc}	J_{new}	(α, δ)	個別推定	同時規準	同時 MG
5	20	(1.0,1.0)	1	1	1
5	40	(1.0,1.0)	2	2	2
5	10	(1.0,0.5)	3	2	2
5	20	(1.0,0.5)	2	1	1
5	40	(1.0,0.5)	2	2	2
5	10	(0.5,1.0)	1	1	1
5	40	(0.5,1.0)	1	1	1
5	10	(0.5,0.5)	1	1	1
5	20	(0.5,0.5)	1	1	1
5	40	(0.5,0.5)	1	0	0
10	20	(1.0,1.0)	2	2	2
10	40	(1.0,1.0)	3	3	3
10	10	(1.0,0.5)	1	1	1
10	40	(1.0,0.5)	2	1	2
10	10	(0.5,1.0)	2	2	2

果」として算出した。

これらの値に関して、条件ごと、および 3 種の推定方法の違い別に l に関して平均および標準偏差を算出し、記録した。

3.3　結果

すべての条件で、100 回の項目パラメタの推定に成功した。いずれの推定方法でも、値が求まらなかった試行はなかった。ただし、小集団条件で、真の項目パラメタから著しくかけ離れた推定結果となる場合が見られた。このような推定結果は、項目パラメタの種類別に RMSE や平均差を評価するうえで、標準偏差が過大に表示されることから、項目バンク上において、真の項目パラメタからの絶対差が 2 を超える推定結果が（識別力パラメタと困難度パラメタのいずれか一方でも）得られた試行に関しては、平均および標準偏差の算出から除外した。表 3.1 に除外した回数を推定方法別に示した。表 3.1 に示されている以外の小集団条件の試行および大集団条件での試行においては、除外された試行はなかった。

3.3.1 $\alpha = 1.0$ の場合の RMSE

はじめに、$\alpha = 1.0$ で小集団、アンカー項目数 10 の条件での RMSE を図 3.4 に示した。$\delta = 1.0$ の場合においても、また $\delta = 0.5$ の場合でも、いずれのパラメタでも個別推定と同時推定で顕著な推定誤差の違いは見られなかった。また、大集団条件での RMSE を図 3.5（$\alpha = 1.0$、大集団、アンカー項目数 10 の条件）に示した。大集団条件の場合、小集団条件の場合と比べていずれのパラメタにおいても全体的に真値からの誤差が少なくなったが、個別推定と同時推定の間で推定誤差の違いは見られなかった。

次に、アンカー項目数を 5 に減らした小集団条件の結果を図 3.6（$\alpha = 1.0$、小集団、アンカー項目数 5 の条件）に記した。アンカー項目数 10 の場合（図 2）と比較しても、いずれの方法においても RMSE の値に大きな違いは見られなかった。大集団条件の場合を図 3.7（$\alpha = 1.0$、大集団、アンカー項目数 5 の条件）に記した。図 3.5 と図 3.7 を比較しても、RMSE の推定値の平均に大きな傾向の違いは見られなかった。Arai and Mayekawa (2011) では、受験者数が倍加すると RMSE がほぼ半減する関係がみられた。図 3.4 および図 3.5 を比較すると、すべての条件で RMSE がほぼ半減しており、Arai and Mayekawa (2011) の結果と一致した。また、図 3.6 と図 3.7 の比較からも、同様な人数の効果が見て取れる結果となった。さらに、アンカー項目数を減らした場合との比較（図 3.4 と図 3.6、図 3.5 と図 3.7）においては、両者の間にはほとんど違いがみられなかった。

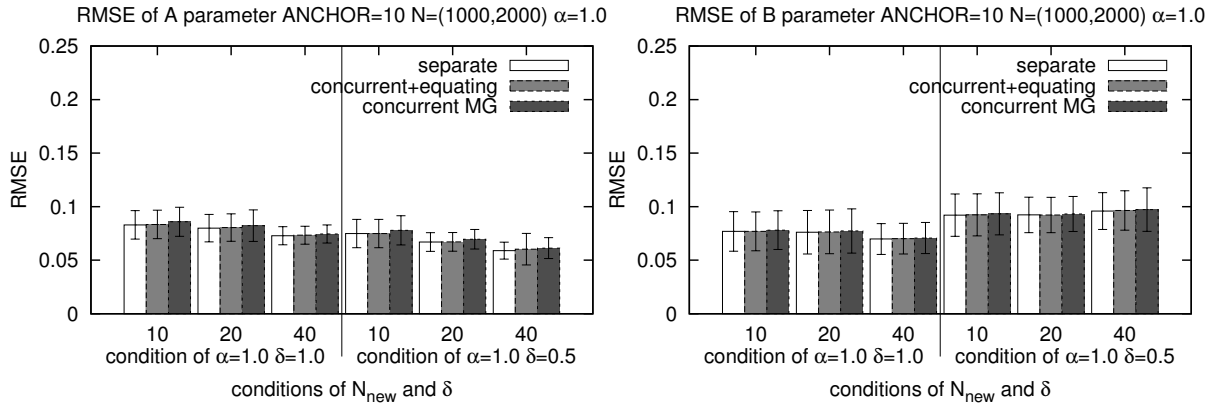

図 3.4　$\alpha = 1.0$、アンカー項目数 10 で小集団条件における、識別力の推定値の RMSE（左）および困難度の推定値の RMSE（右）。それぞれのグラフにおいて、左側は $\delta = 1.0$、右側は $\delta = 0.5$ で、左から新作項目数が 10,20,40 の条件を示す

3.3.2 $\alpha = 0.5$ の場合の RMSE

次に、$\alpha = 0.5$ の条件で、小集団条件、アンカー項目数 10 の場合の RMSE を図 3.8 （$\alpha = 0.5$、小集団、アンカー項目数 10 の条件）に示した。また、同条件で大集団の場合を図 3.9（$\alpha = 0.5$、大集団、アンカー項目数 10 の条件）に示した。これらの条件では、$\delta = 1.0$ の条件において、新作

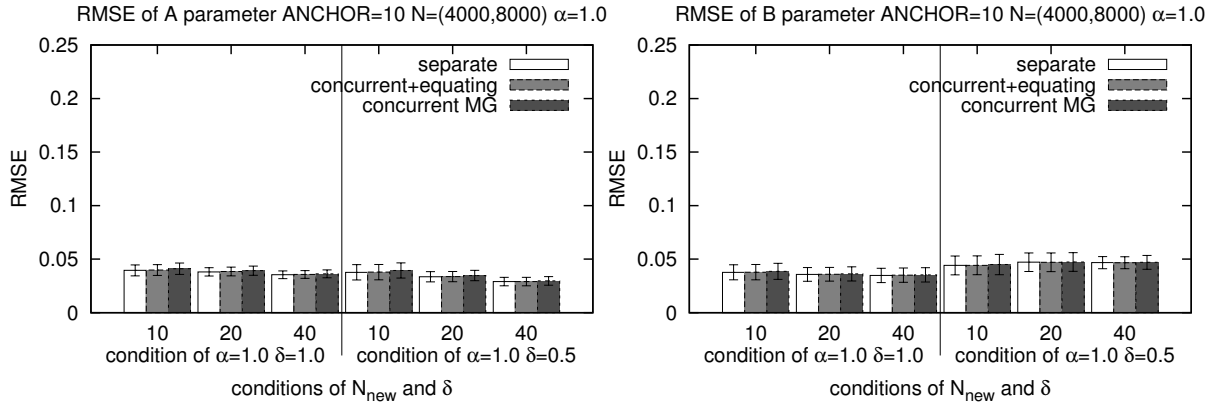

図 3.5　$\alpha = 1.0$、アンカー項目数 10 で大集団条件における、識別力の推定値の RMSE（左）および困難度の推定値の RMSE（右）。それぞれのグラフにおいて、左側は $\delta = 1.0$、右側は $\delta = 0.5$ で、左から新作項目数が 10,20,40 の条件を示す

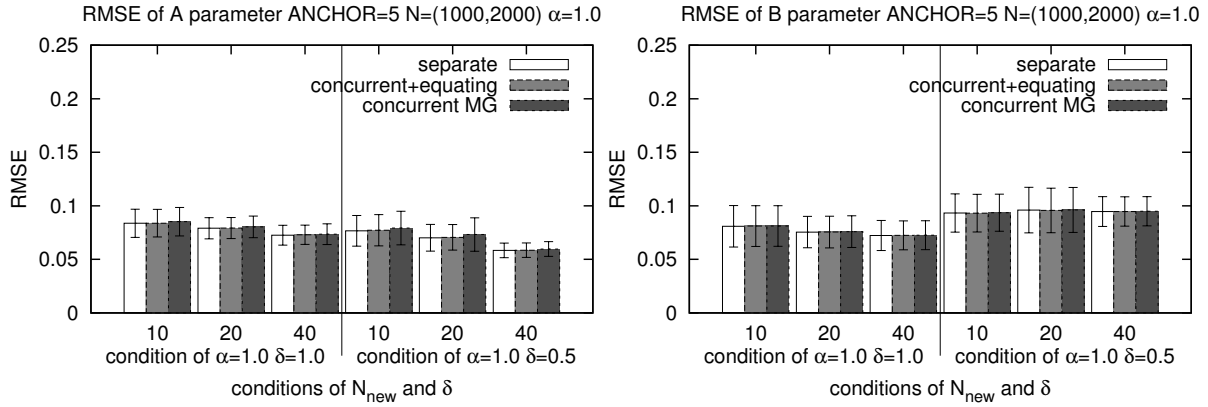

図 3.6　$\alpha = 1.0$、アンカー項目数 5 で小集団条件における、識別力の推定値の RMSE（左）および困難度の推定値の RMSE（右）。それぞれのグラフにおいて、左側は $\delta = 1.0$、右側は $\delta = 0.5$ で、左から新作項目数が 10,20,40 の条件を示す

項目数が 20 項目、40 項目の場合で推定方法の間に傾向の違いがみられた一方、$\delta = 0.5$ の条件においては、方法の間で推定値に大きな傾向の差は見られず、人数が増えることによって推定誤差が縮小する傾向のみ見られた。

$\delta = 1.0$ の場合、同時推定の場合で、識別力の推定値に大きな真値からの差異がみられた。同時推定の 2 種の方法の違いでみると、「同時推定＋規準集団への等化」の方法のほうが、「同時推定 MG」の方法よりもより真の値に近い結果となった。また、その差異の大きさは、本試験における新作項目数が多くなるにしたがって増大する傾向が顕著であった。それに対し、個別推定では、いずれの新作項目数の条件においても差異はほぼ一定であった。また同時推定の場合、大集団条件であっても、小集団条件に比べて小さいものの、真値からのずれは解消していないことがわかった。また、$\delta = 1.0$ の条件で困難度の推定値は、識別力とは違った傾向を示した。新作項目数を増やす

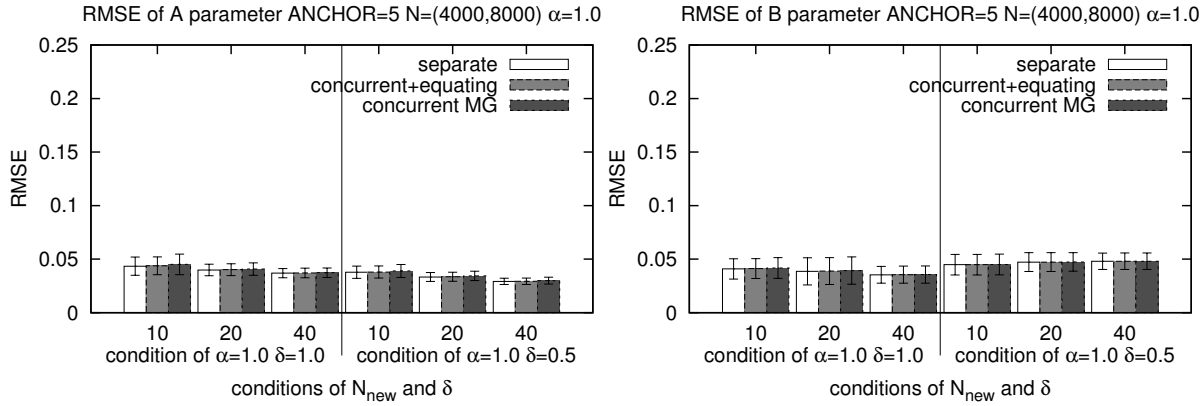

図 3.7　$\alpha = 1.0$、アンカー項目数 5 で大集団条件における、識別力の推定値の RMSE（左）および困難度の推定値の RMSE（右）。それぞれのグラフにおいて、左側は $\delta = 1.0$、右側は $\delta = 0.5$ で、左から新作項目数が 10,20,40 の条件を示す

ことによって、個別推定の条件で RMSE が減少した一方、同時推定においては RMSE の値が小集団条件においてはやや大きくなる傾向に、大集団条件においては大きくなる傾向を示した。全体的に「同時推定＋規準集団への等化」の方法よりも「同時推定 MG」の方が、真値からかけ離れた推定値を示した。

さらに、$\alpha = 0.5$ の条件で、小集団条件、アンカー項目数 5 の場合の RMSE を図 3.10（$\alpha = 0.5$、小集団、アンカー項目数 5 の条件）に示した。また、同条件で大集団の場合を図 3.11（$\alpha = 0.5$、大集団、アンカー項目数 5 の条件）に示した。これらを図 3.8、図 3.9 と比較すると、特に大集団条件において、個別推定と同時推定とで項目パラメタの推定値に大きな違いがみられた。図 3.9 と図 3.11 を比較すると、アンカー項目を減らした場合、新作項目数 20 と 40 の条件において、同時推定で生じる RMSE の大きさが顕著に増大する一方、個別推定においては RMSE の平均がほぼ同一であることが分かった。

Arai and Mayekawa (2011) でみられたような、受験者数が増大すると RMSE がほぼ半減する関係は、$\alpha = 0.5$ の条件においても、個別推定で見られた。ただし、同時推定で $\delta = 1.0$ の場合、すなわち、RMSE に違いがみられた条件に関しては、識別力の推定値に関して半減する傾向は見られなかった（図 3.8 と図 3.9、および図 3.10 と図 3.11 の比較）。さらに、アンカー項目を減らした場合について比較すると、同時推定で $\delta = 1.0$ の条件の場合、アンカー項目 10 の条件の方が、アンカー項目 5 の条件より小さな RMSE の平均を示した（図 3.8 と図 3.10、および図 3.9 と図 3.11 との比較）。一方、それ以外の条件では、アンカー項目数の違いで RMSE の平均に大きな差異は見られなかった。

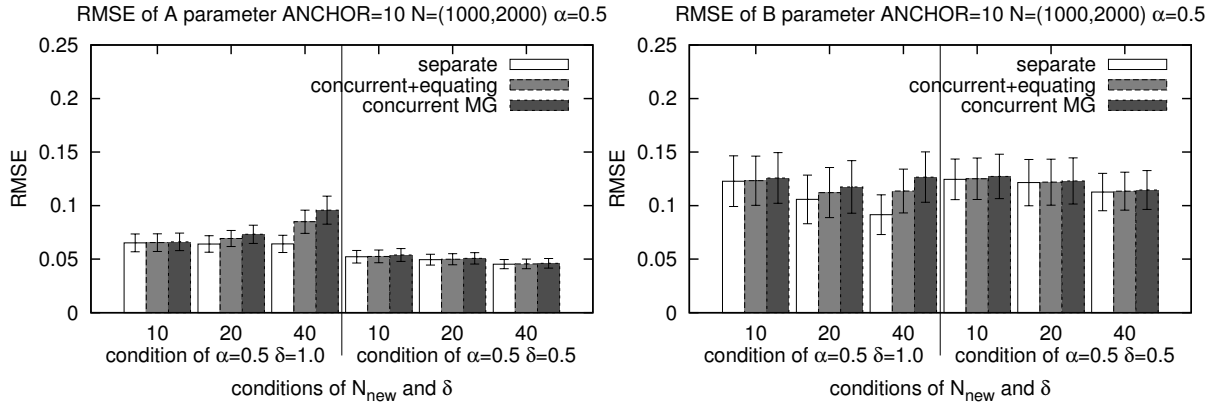

図 3.8　$\alpha = 0.5$、アンカー項目数 10 で小集団条件における、識別力の推定値の RMSE（左）および困難度の推定値の RMSE（右）。それぞれのグラフにおいて、左側は $\delta = 1.0$、右側は $\delta = 0.5$ で、左から新作項目数が 10,20,40 の条件を示す

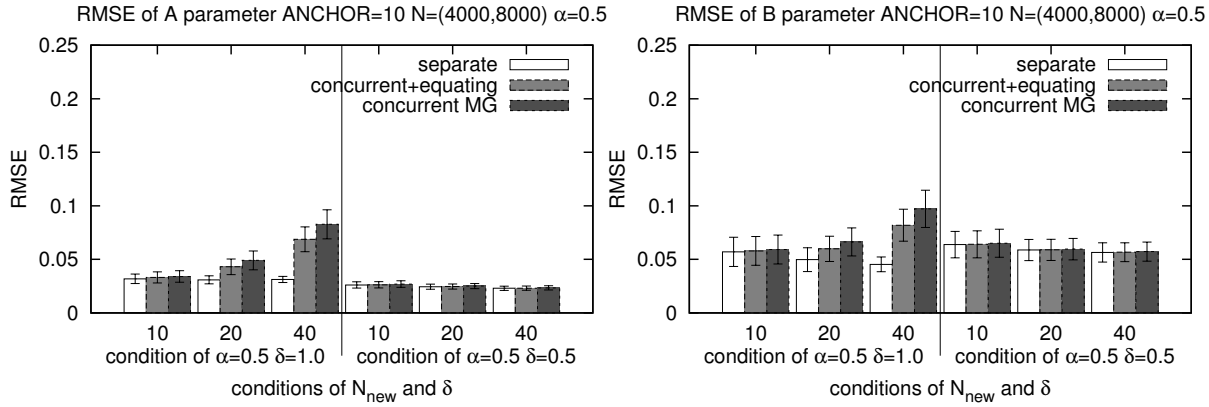

図 3.9　$\alpha = 0.5$、アンカー項目数 10 で大集団条件における、識別力の推定値の RMSE（左）および困難度の推定値の RMSE（右）。それぞれのグラフにおいて、左側は $\delta = 1.0$、右側は $\delta = 0.5$ で、左から新作項目数が 10,20,40 の条件を示す

3.3.3　真値との差の方向性

差の方向性を見るために、識別力、困難度別の真値との差の値を $\alpha = 1.0$ の場合については図 3.12（$\alpha = 1.0$、小集団、アンカー項目数 10）、図 3.13（$\alpha = 1.0$、大集団、アンカー項目数 10）、図 3.14（$\alpha = 1.0$、小集団、アンカー項目数 5）、図 3.15 （$\alpha = 1.0$、大集団、アンカー項目数 5）に、$\alpha = 0.5$ の場合は図 3.16（$\alpha = 0.5$、小集団、アンカー項目数 10）、図 3.17（$\alpha = 0.5$、大集団、アンカー項目数 10）、図 3.18 （$\alpha = 0.5$、小集団、アンカー項目数 5）、図 3.19（$\alpha = 0.5$、大集団、アンカー項目数 5）にそれぞれ示した。推定方法の間で RMSE の値に差が出た $\alpha = 0.5$ かつ $\delta = 1.0$（図 3.16、図 3.17、図 3.18 および図 3.19）の場合でみると、同時推定の場合に、いずれの受験者

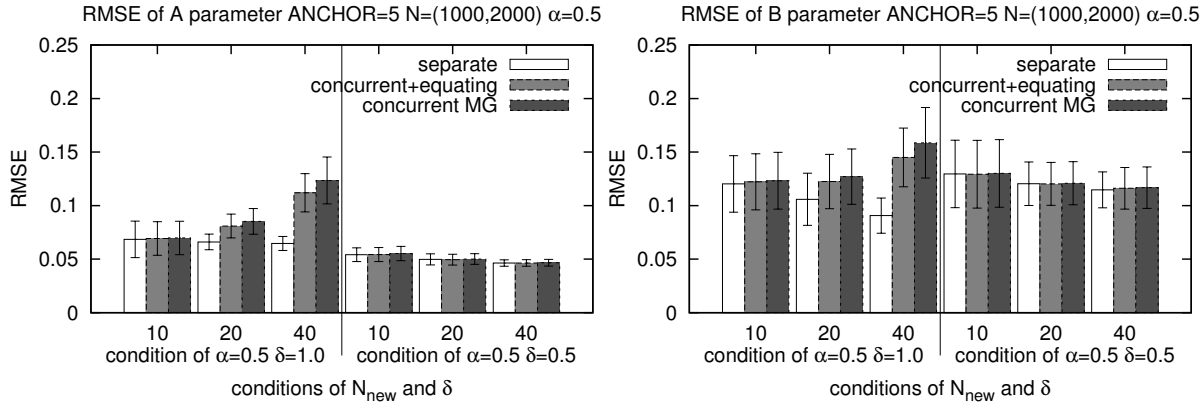

図 3.10　$\alpha = 0.5$、アンカー項目数 5 で小集団条件における、識別力の推定値の RMSE（左）および困難度の推定値の RMSE（右）。それぞれのグラフにおいて、左側は $\delta = 1.0$、右側は $\delta = 0.5$ で、左から新作項目数が 10,20,40 の条件を示す

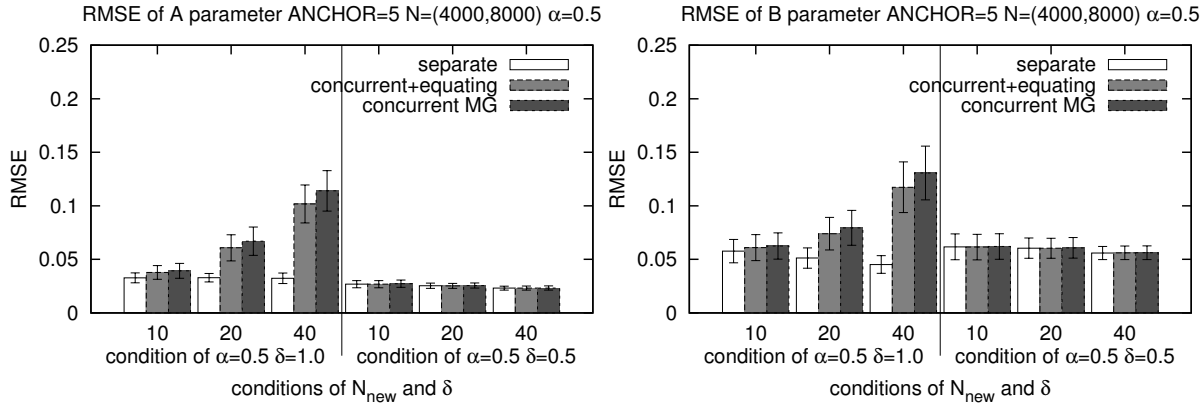

図 3.11　$\alpha = 0.5$、アンカー項目数 5 で大集団条件における、識別力の推定値の RMSE（左）および困難度の推定値の RMSE（右）。それぞれのグラフにおいて、左側は $\delta = 1.0$、右側は $\delta = 0.5$ で、左から新作項目数が 10,20,40 の条件を示す

数条件でも、またいずれのアンカー項目数の条件でも、新作項目数が 20 または 40 の場合に、識別力パラメタを真値より過小に推定していることが分かった。さらに、新作項目数が増加するにしたがって、過小推定の度合いも増大することがわかった。また小集団条件と大集団条件の結果を比較すると、受験者の人数が増えても、真値との差の値そのものはあまり減少しない半面、標準偏差が小さくなる傾向が見られた。人数を増やすことによって、過小推定される傾向は変わらず、過小推定された推定値の信頼区間が狭まるという結果であった。一方、個別推定では、このような傾向は見られなかった。さらに、困難度の真値との差は、識別力の場合と異なり、差の方向に一貫性は見られなかった。

真値との差が過小推定された条件で、アンカー項目数の違いでみると、図 3.16 と図 3.18、および図 3.17 と図 3.19 との比較から、いずれの受験者数条件においても、アンカー項目数が減少する

と過小推定の度合いが大きくなることがわかった。

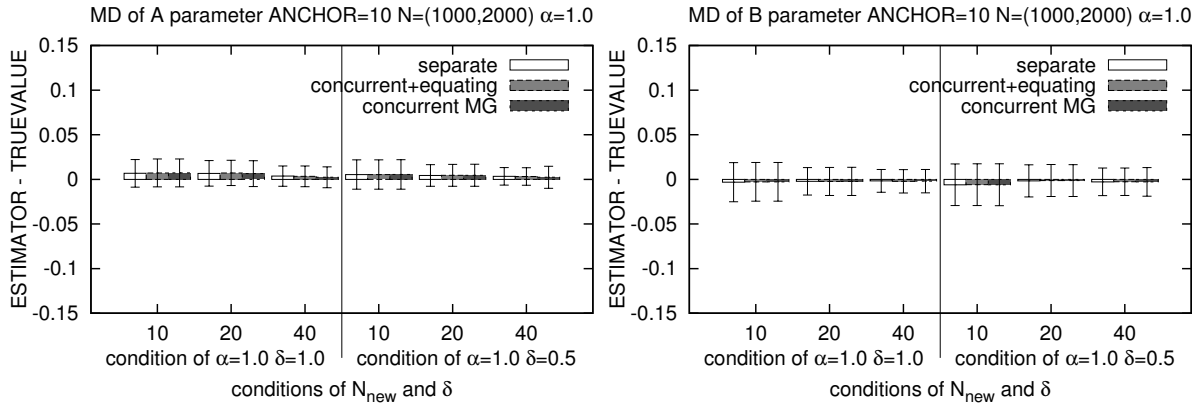

図 3.12　$\alpha = 1.0$、アンカー項目数 10 で小集団条件における、識別力の推定値の MD（左）および困難度の推定値の MD（右）。それぞれのグラフにおいて、左側は $\delta = 1.0$、右側は $\delta = 0.5$ で、左から新作項目数が 10,20,40 の条件を示す

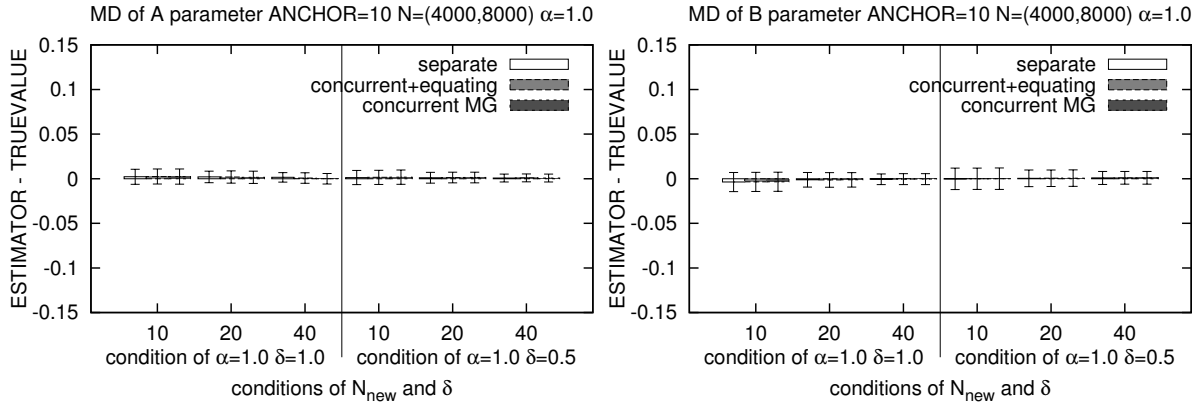

図 3.13　$\alpha = 1.0$、アンカー項目数 10 で大集団条件における、識別力の推定値の MD（左）および困難度の推定値の MD（右）。それぞれのグラフにおいて、左側は $\delta = 1.0$、右側は $\delta = 0.5$ で、左から新作項目数が 10,20,40 の条件を示す

3.4　考察

3.4.1　等化方法間における RMSE の差異

実際に試験を実施しようとしているテスト実施機関が、十分にサイズの大きい項目バンクを事前に作成することが困難な場合に、毎回の試験実施に際して新作項目を含め、試験終了後に等化後の項目パラメタを項目バンクに追加する方法で標準化テストを実施するシミュレーションを行った。その結果、同時推定を行った場合、予備試験での識別力が本試験の推定値に比べて低い値の場合

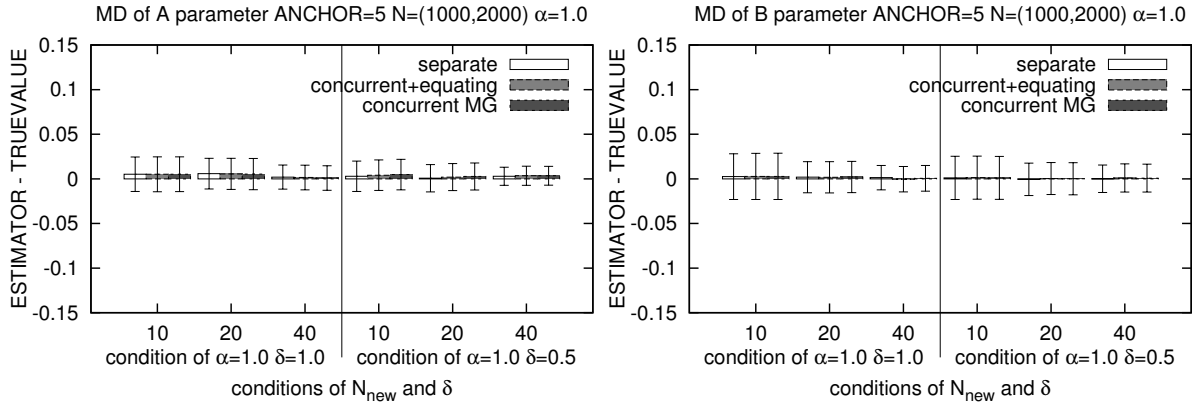

図 3.14　$\alpha = 1.0$、アンカー項目数 5 で小集団条件における、識別力の推定値の MD（左）および困難度の推定値の MD（右）。それぞれのグラフにおいて、左側は $\delta = 1.0$、右側は $\delta = 0.5$ で、左から新作項目数が 10,20,40 の条件を示す

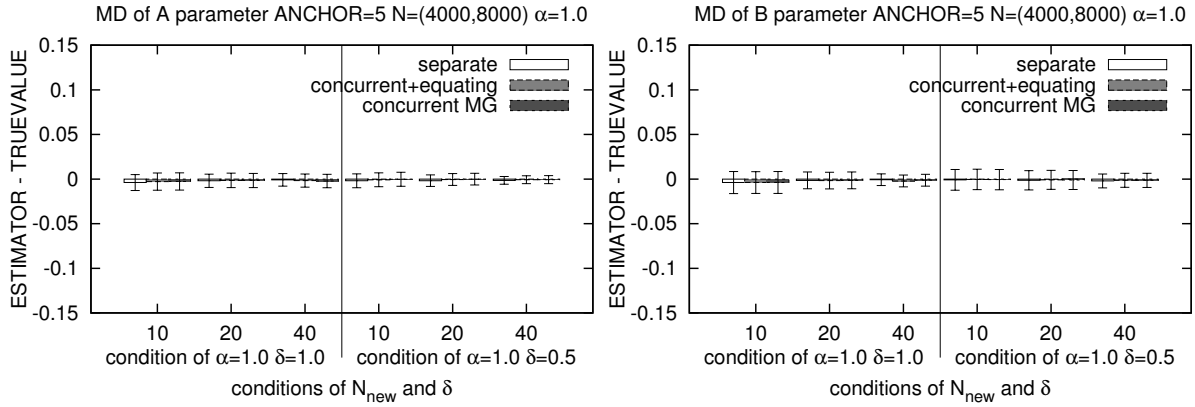

図 3.15　$\alpha = 1.0$、アンカー項目数 5 で大集団条件における、識別力の推定値の MD（左）および困難度の推定値の MD（右）。それぞれのグラフにおいて、左側は $\delta = 1.0$、右側は $\delta = 0.5$ で、左から新作項目数が 10,20,40 の条件を示す

に、項目パラメタの推定値が真値から大きくずれるという結果であった。この傾向は、Arai and Mayekawa (2011) の結果を支持する結果であるといえる。本研究のアンカー項目抽出ルールによって採用されるリンキングのデザインは、Arai and Mayekawa (2011) で検討された 3 種のリンキングのデザインを平均的に含む場合が多く、本研究のデザインにおいても、同時推定の推定結果の偏りが現れたものと推測できる。識別力の値が予備試験において低い場面は、規準となる予備試験において、因子分析の「因子負荷」に該当する値が低い場合に相当する。具体的には、ある単一のテストフォームにおいて、j 番目の項目の識別力 は、2 値（正誤）反応とテスト通過率 x との双列相関係数 ρ_{jx} を用いると、式 1.11 の $\rho_j = \rho_{jx}$ とおいた場合に相当する。これは、BILOG-MG での計算において、識別力の初期値として用いられる（村木、2011、p.44）。本研究のデザインで同時推定時に RMSE が大きくなったのは、項目バンク上で「識別力の真値が小さい」項目に「識

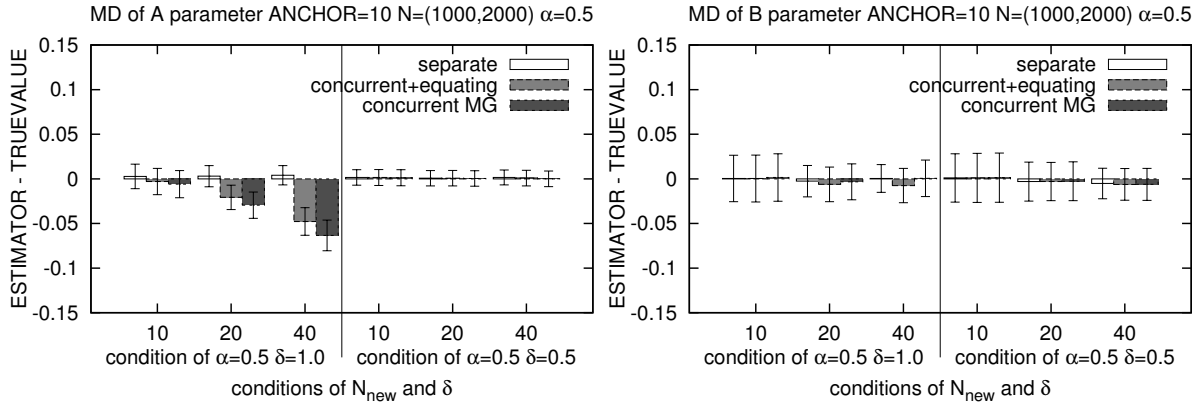

図 3.16　$\alpha = 0.5$、アンカー項目数 10 で小集団条件における、識別力の推定値の MD（左）および困難度の推定値の MD（右）。それぞれのグラフにおいて、左側は $\delta = 1.0$、右側は $\delta = 0.5$ で、左から新作項目数が 10,20,40 の条件を示す

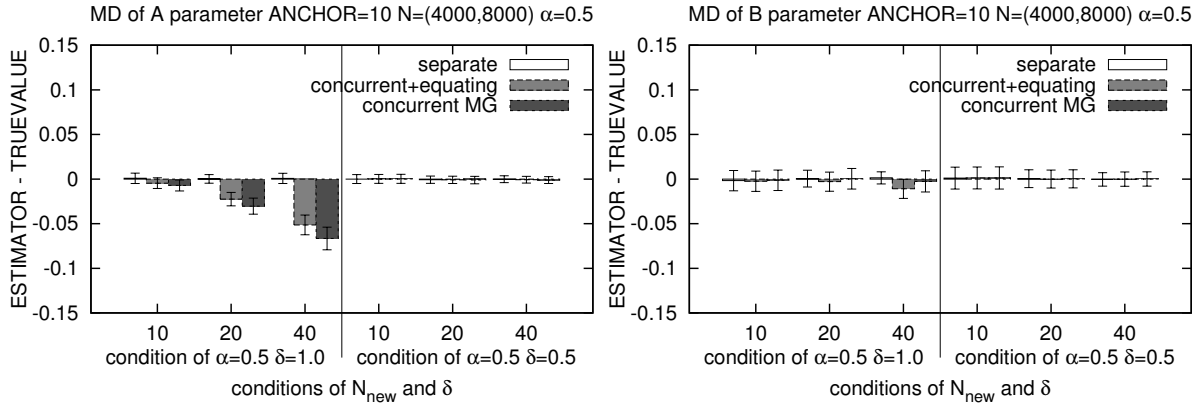

図 3.17　$\alpha = 0.5$、アンカー項目数 10 で大集団条件における、識別力の推定値の MD（左）および困難度の推定値の MD（右）。それぞれのグラフにおいて、左側は $\delta = 1.0$、右側は $\delta = 0.5$ で、左から新作項目数が 10,20,40 の条件を示す

別力の真値が大きい」多数の項目を等化する場合であった。同時推定では、真の識別力が高い項目も低い項目も同一のモデル上で一度に分析される。そのため、それぞれのフォームをまたいで共通の因子が一つ抽出される過程で、予備試験における「受験者の能力と個々の項目の正誤との相関が低いような正誤反応データセット」に尺度を合わせるという影響を受け、項目バンク上のパラメタ推定値が真値から大きくずれたと考える。ただし、同時推定の中でも、一括して項目パラメタを推定したのち、予備試験の項目パラメタに等化する処理を行った「同時推定＋規準集団に等化」条件では、規準集団の項目パラメタという手掛かりに等化しているため、項目バンクにおいて「同時推定 MG」に比べて真の識別力に近い値が得られたと考える。一方、個別推定では、真の識別力の低いフォームと高いフォームを別個に推定したうえで、等化を行っているので、識別力の低いフォームの推定結果が他のフォームに直接影響しない。そのため、項目バンク上で真値からのずれが 3 種

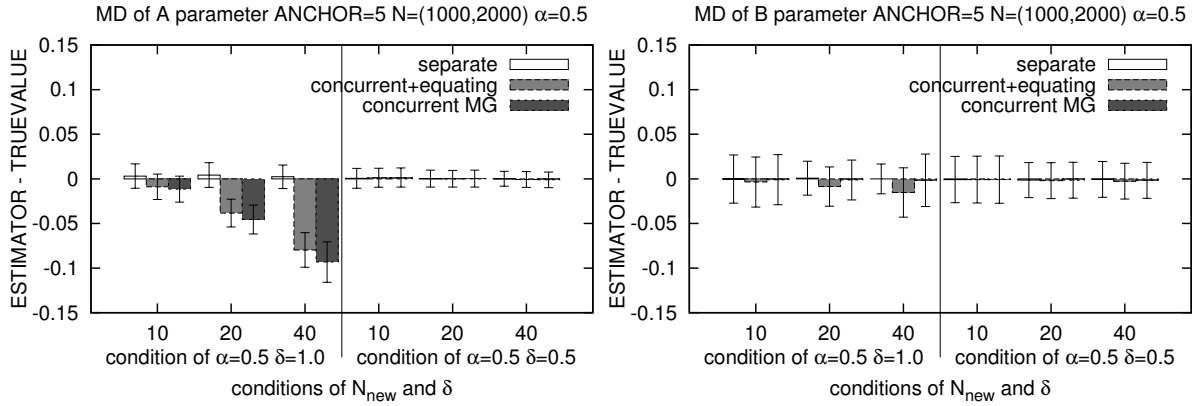

図 3.18　$\alpha = 0.5$、アンカー項目数 5 で小集団条件における、識別力の推定値の MD（左）および困難度の推定値の MD（右）。それぞれのグラフにおいて、左側は $\delta = 1.0$、右側は $\delta = 0.5$ で、左から新作項目数が 10,20,40 の条件を示す

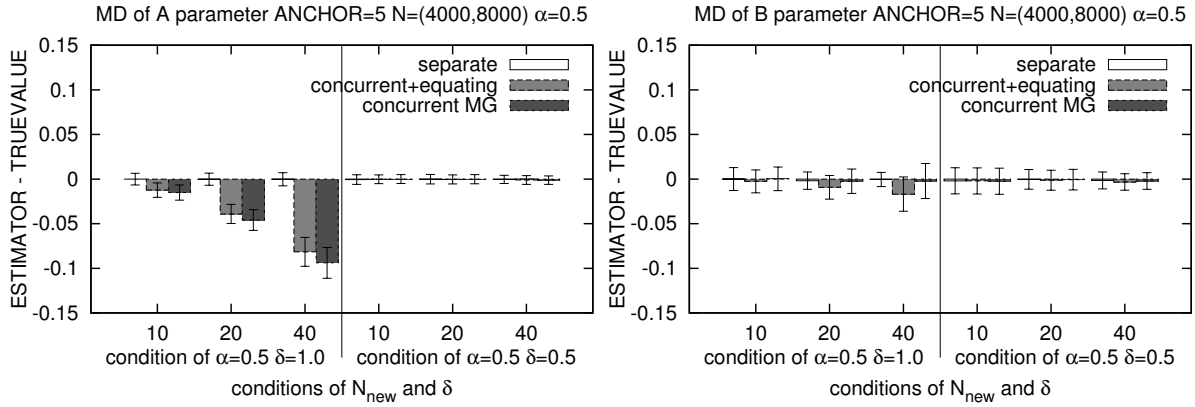

図 3.19　$\alpha = 0.5$、アンカー項目数 5 で大集団条件における、識別力の推定値の MD（左）および困難度の推定値の MD（右）。それぞれのグラフにおいて、左側は $\delta = 1.0$、右側は $\delta = 0.5$ で、左から新作項目数が 10,20,40 の条件を示す

の方法の中で最も少なかったと考える。

3.4.2　等化方法間における MD の差異

真値との差の方向性に関しては、識別力パラメタの推定値において、個別推定と同時推定とで RMSE に違いがみられた条件で過小に推定されるという結果を得た。過小推定の度合いは、アンカー項目数が 5 の条件の方が 10 の条件に比べて大きかった。仮にアンカー項目として選ばれた項目のみが特異的に過小推定されている場合、真値との平均差はアンカー項目数が多い条件においてより大きくなるはずである。よって、アンカー項目のみならず、項目バンク全体で見たときに識別力が過小に推定されているといえる。また、過小推定は、アンカー項目数と同程度の新作項目の

条件においては顕著でなかったのに対し、新作項目を増やすとその傾向が大きくみられた。さらに、本研究においては、アンカー項目として選ばれ、繰り返し提示される項目は、項目バンク内で識別力の高い項目とした。このルールに従うと、識別力の推定値が過小推定された「$\alpha = 0.5$ かつ $\delta = 1.0$」の条件において、「本試験 1 の項目のみ本試験 2 のアンカー項目として選ばれる」場合が多いことが指摘できる（他の条件においては、本試験 2 のアンカー項目は、本試験 1 の項目と予備試験の項目とが両方含まれうる）。これは、Arai and Mayekawa (2011) において Type 3 とされているデザインに相当し、等化方法の比較において同時推定が最も真値から異なるという結果を示していた。以上をまとめると、アンカー項目数が少なく、新作項目数が多く、かつ、同時推定によって推定結果にバイアスがかかるようなテストデザインにおいて、識別力の過小推定が起こっているといえる。また、識別力の推定値が過小に推定された原因として、同時推定時における 1 因子抽出の影響が指摘できる（後述）ものの、同時推定時における過小推定の理論的説明に関して、今後さらに検討が必要であるといえる。

3.4.3　アンカー項目数の効果

また、項目パラメタ推定の RMSE および真値との平均差について、同時推定を行う場合であっても、アンカー項目数が 5 の場合も 10 の場合でも、新作項目の数が 10 の条件では、個別推定と同様の RMSE を示した。この点は、新作項目数を大きく増やす必要がない場合、個別推定と同時推定のいずれをとっても同様の結果であることを示している。しかし、本研究で述べたテスト実施法が採用される場合、項目バンクのサイズを短期間のうちに大きくしたいという意図のもと、新作項目数を毎回の試験で多く出題する場合が多いことが予想される。したがって、実際にこの試験デザインを適用する場面においては、RMSE および真値との平均差が大きくなる場合に該当する場合が多いことが指摘できる。

3.4.4　項目パラメタ更新の問題、とるべき等化方法

ここで、実際の試験場面で想定される、項目バンク上での項目パラメタ更新の問題について述べる。本研究で述べた試験デザインを適用する場面において、同時推定により項目バンクの内容を更新し続ける方針をとったならば、毎回の試験を行うたびに同時推定による等化済み項目パラメタを推定し、その値を項目バンクに登録する際、項目バンク内のすべての項目のパラメタが一斉に変化する。一方で、個別推定を行うことにすれば、項目バンク内にて変更が必要な部分は、直近の試験でアンカー項目として出題した項目に限られる。テスト実施機関から見れば、同時推定によって、アンカー項目として出題した項目以外の項目の特性まで変わってしまう点は、値の解釈として不自然と考える可能性があり、その意味でも本研究のデザインで等化を行い続けるならば、個別推定を行った方がより実践的ではないかと考える。

しかしながら、心理測定の場面としてテストを見たときに、個別推定の方法が最もパフォーマンスが良かったという本研究の結果から、「個別推定を行い続けてさえいれば常に最良の等化が可能

である」とは限らないことに注意しなければならない。Lord(1980) は、等化を行う上での条件の一つとして、等化元と等化先で同一の特性（潜在特性）を測定していることを挙げている。すなわち、予備試験、毎回の本試験において、テストフォームはすべて同一の構成概念を測定していることが、等化の前提となっている。IRT のパラメタ推定を、因子分析の因子負荷の推定と等価であると考えた場合、個別推定では各々のテストフォームごとに抽出される因子の意味が異なるにもかかわらず、等化係数を算出する手続きによってあたかも適切な等化が行えたかのように見える一方、同時推定ではモデルの適合度はともかく、推定の結果として常に「テストフォームをまたいだ」1 因子が抽出される。したがって、等化元と等化先で共通の因子が仮定できない場合に、本研究の結論がそのまま当てはまるとは限らない。

また、同時推定時に識別力の推定値に過小推定が生じた要因は、同時推定により 1 因子が抽出される過程で、モデルでは表されない第 2 因子以降の因子負荷の部分が、項目パラメタの推定値に反映されないため、識別力の推定値が過小に推定されたのではないかと考える。個別推定においては、同時推定に比べて全体の項目数が少ない、いいかえればより 1 因子性が高いような個々のテストフォームごとに 1 因子を抽出した結果を用いて、アンカー項目を手掛かりに規準集団に等化するため、識別力の過小推定は起こりにくいと考える。

3.4.5　今後の課題

本研究では、項目パラメタの困難度の真値を、受験者の能力値分布とほぼ包含するような仮定をおいた。もちろん、実際の試験では、このような理想的な場面が毎回の試験にわたって続くとは考えにくい。その意味で、困難度が実際の受験者のレベルに合致しない場合の検討をすべきであろう。また、アンカー項目抽出のルールを変えた場合、とりわけ直近の試験で使用（受験者へ提示）した項目を次の回の試験で使用しない場合と、使用する場合でどのように項目バンク中の項目パラメタの推定値が影響を受けるか、あるいは項目バンクに存在する項目に対し、複数回の項目パラメタの推定値がある場合、平均をとる方法が最適なのかどうかについて、今後さらなる検討を要すると考える。

第 4 章

多群 IRT モデルのパラメタ推定における計算機資源の限界

4.1 序論

4.1.1 同時推定における計算の限界

同時推定においては、前章で述べたとおり、実施回を重ねるごとに大きくなっていく反応パタン行列を用いて、多群 IRT モデルを用いて分析を行わなければならない。特に、新作項目割合が大きい場合、その影響は顕著となる。一方で、個別推定の場合は、毎回の試験について個別に反応パタン行列から項目パラメタを推定すればよい。したがって、計算可能性の観点から見れば、同時推定において計算上の問題が生じる可能性がある。本節では、この計算場面における限界がどのような要因で起こるかを指摘し、シミュレーションによって実際の計算の限界を明らかにする。

周辺最大尤度法を用いた項目パラメタ推定

一般に、多群 IRT に限らず、複数のフォームから得られた 0-1 データから項目パラメタを推定する際、EM アルゴリズムによる最尤法を用いる。この方法は、項目パラメタと θ の推定を同時に行う同時最尤推定法 (joint maximum likelihood estimation method; JML) に端を発し、Bock and Lieberman (1970) の周辺最尤推定法 (marginal maximum likelihood estimation method, MML) によって一致推定量 (受験者の数が増えれば増えるほど、計算されたパラメタの推定値が母数値に近づくという性質をもった推定量) の問題を統計学的に解決したモデルに発展した (村木、2011、p.75)。さらに、計算機資源を消費しない MML の推定法として、Bock and Aitkin (1981) は「EM アルゴリズムを用いた MML 推定法 (marginal maximum likelihood estimation method with EM algorithm; MML-EM) を提案した。MML-EM は、IRT パラメタ推定法の標準であり、BILOG-MG は、MML-EM を用いて計算を行っている (村木、2011、p.77)。本節では、MML-EM の推定法について述べる。

N 人の被験者 i が全 J 項目からなるテストの j 番目の項目に正答したかどうかを表すデータを

$\boldsymbol{u} = u_{ij}$ とすると、その確率は各被験者について

$$P(\boldsymbol{u}_i|\theta_i, \boldsymbol{\xi}) = \prod_{j=1}^{J} P_j(\theta_i)^{u_{ij}} Q_j(\theta_i)^{1-u_{ij}} \tag{4.1}$$

という尤度関数で表せる。ただし、$\boldsymbol{u}_i$ は i 番目の受験者の正誤反応データを表すベクトル、$P_j(\theta_i)$ は項目反応モデル、$Q_j(\theta_i) = 1 - P_j(\theta_i)$ であり、$\boldsymbol{\xi}$ は項目反応モデルに含まれるパラメタを表す。

ここで、θ_i に関する事前分布を、分布の特徴を示すパラメタ $\boldsymbol{\tau}$ を用いて $g(\theta|\boldsymbol{\tau})$ と表す。通常、θ_i の事前分布には正規分布 $N(\mu, \sigma^2)$ を仮定する。この場合、$\boldsymbol{\tau} = (\mu, \sigma^2)'$ である。

$\boldsymbol{u}$ の周辺確率は $\boldsymbol{u}_i, i = 1, 2, ..., N$ に基づくパラメタの周辺尤度で、

$$P(\boldsymbol{u}) = \prod_{i=1}^{N} \int P(\boldsymbol{u}|\theta_i, \boldsymbol{\xi}) g(\theta_i|\boldsymbol{\tau}) d\theta_i \tag{4.2}$$

と書ける。4.2 式は実際のデータから観測されない θ_i を、θ_i の事前分布をかけたうえで積分により消去したものを N 人分かけ合わせたもので、周辺尤度関数と呼ぶ（4.2 式中、積分の中に現れる事前分布 $g(\theta_i|\boldsymbol{\tau})$ には、i という添字がつくことに注意）。周辺尤度関数と、データ $\boldsymbol{u}$ が得られたら、ベイズ理論に基づき事後確率分布

$$h(\theta_i|\boldsymbol{u}, \boldsymbol{\tau}, \boldsymbol{\xi}) = \frac{P(\boldsymbol{u}|\theta_i, \boldsymbol{\xi}) g(\theta_i|\boldsymbol{\tau})}{\int P(\boldsymbol{u}|\theta_i, \boldsymbol{\xi}) g(\theta|\boldsymbol{\tau}) d\theta_i} \tag{4.3}$$

が得られる。

Bock and Aitkin (1981) では、以下の E ステップと M ステップをそれぞれ反復して行い、項目パラメタを推定している。まず、E(expectancy) ステップとして、4.3 式の事後確率分布において、$g(\theta_i|\boldsymbol{\tau})$ を Q 個の求積点 $h_q (q = 1, 2, ..., q, ...Q)$ 上における重み $A(h_q)$ で置き換える。これにより、4.3 式は

$$h(h_q|\boldsymbol{u}, \boldsymbol{\tau}, \boldsymbol{\xi}) = \frac{L_j(h_q) A(h_q)}{\sum_{q=1}^{Q} L_j(h_q) A(h_q)} \tag{4.4}$$

$$L_j(h_q) = \prod_{j=1}^{J} P_j(h_q)^{u_{ij}} Q_j(h_q)^{1-u_{ij}} \tag{4.5}$$

というように、分母における積分を和で置き換えた形で表すことができる。さらに、式 4.4 において、求積点上の期待値

$$\bar{N}_q = \sum_{i=1}^{N} \left[\frac{L_j(h_q) A(h_q)}{\sum_{q=1}^{Q} L_j(h_q) A(h_q)} \right] \tag{4.6}$$

$$\bar{r}_{jq} = \sum_{i=1}^{N} \left[\frac{u_{ij} L_j(h_q) A(h_q)}{\sum_{q=1}^{Q} L_j(h_q) A(h_q)} \right] \tag{4.7}$$

を求める。$\bar{N}_q$ は、全受験者の中で、h_q となる能力値を持つとされる期待度数であり、$\bar{r}_{jq}$ は、h_q という能力値の受験者における項目 j に正答する受験者の期待度数である。

次に、M(maximization) ステップとして、E ステップで求まった $\bar{N}_q$ および $\bar{r}_{jq}$ を、あたかもデータとして観測されているかのように扱い、以下に示す期待対数完全データ尤度の最大値を見出す。

$$\ln L(\boldsymbol{\xi}, \boldsymbol{\tau}) = \sum_{j=1}^{J}\sum_{q=1}^{Q}\left[\bar{r}_{jq}\ln P_j(A(h_q)|\boldsymbol{\xi}_j) + (\bar{N}_q - \bar{r}_{jq})\ln(1 - P_j(A(h_q)|\boldsymbol{\xi}_j))\right] + \sum_{q=1}^{Q}\bar{N}_q \ln g(\theta_i|\boldsymbol{\tau}) \tag{4.8}$$

ただし $\boldsymbol{\xi}_j$ は j 番目の項目のパラメタを表す。最大値を見出す方法としては、4.8 式中に含まれている $\boldsymbol{\xi}$ の各要素について偏微分をとり、Newton-Raphson 法、あるいは Fisher の scoring 法により最適解を求める。詳細な式の形は Baker and Kim(2004)、尾崎 (2005)、前川 (1991) などを参照のこと。

ところで、MML-EM による推定においては、E ステップにおける式 4.6 および式 4.7 の $\bar{N}_q$ および $\bar{r}_{jq}$ を求める必要がある。しかし、実際の計算においては、同一の u_{kj} という項目反応パタンを持つ受験者が F_k 人存在し、u_{kj} のバリエーションが K 個（$k = (1, 2, ..., k, ...K)$）存在するというデータ (u_{kj}, F_k) があれば、

$$\bar{r}_{jq} = \sum_{i=1}^{N} u_{ij}h_{iq} = \sum_{k=1}^{K} u_{kj}F_k h_{kq} \tag{4.9}$$

$$\bar{N}_q = \sum_{i=1}^{N} h_{iq} = \sum_{k=1}^{K} F_k h_{kq} \tag{4.10}$$

と考えることにより、$\bar{N}_q$ および $\bar{r}_{jq}$ を求めることができる。ここで h_{iq} は、i 番目の受験者の q 番目の求積点における事前分布の離散値であり、h_{iq} は、k 番目のバリエーションの反応パタンにおける q 番目の求積点における事前分布の離散値である。したがって、MML-EM における数値計算は、受験者の人数ではなく、受験者が反応した正誤反応パタンのバリエーションの数に依存した形で、メモリが消費されることがわかる。また、バリエーションの数は $K \leq 2^J$ であるので、たとえば全体で 15 項目であるなら、K は最大でも $2^{15} = 32768$ である。したがって、15 項目からなるテストの場合、これ以上の人数からなるテストデータを扱う場合は、ローデータを (u_{kj}, F_k) を用いた形にあらかじめ変換することにより、メモリの消費量を確実に減らすことが可能である。

MML-EM における多群 IRT モデルのパラメタ推定

一方、Mislevy(1984) の多群 IRT モデルにおいては、受験者グループが G 個の母集団から別個にサンプリングされた場面を想定し、g 番目の受験者グループにおいて、能力値 θ の分布が $\theta_{i \in g} \sim h_g(\boldsymbol{\theta}|\boldsymbol{\tau}_g)$ に従っていると仮定する。ただし、$i \in g$ は受験者 i がグループ g に属することを示し、$\boldsymbol{\tau}_g$ は、g 番目のグループにおける分布の特徴を示すパラメタである。以下、Mislevy(1984) および室橋 (2005) に基づき、グループをまたいだ項目パラメタの推定値 $\boldsymbol{\xi}$ および各グループにおける分布のパラメタ $\boldsymbol{\tau}_g$ の推定法を記す。

多群 IRT モデルにおける周辺確率は

$$P_g(\boldsymbol{u}|\boldsymbol{\xi},\boldsymbol{\tau}_g) = \int P(\boldsymbol{u}|\theta_i,\boldsymbol{\xi})h(\theta_i|\boldsymbol{\tau}_g)d\theta_i \tag{4.11}$$

と書くことができる。ここで、前節で導入した u_{kj} と F_k を用いた記法を用いると、グループをまたいだ全体の正誤反応パタン $\boldsymbol{U}$ を得る確率は、

$$P_j(\boldsymbol{U}|\boldsymbol{\xi},\boldsymbol{\tau}) = \prod_{g=1}^{G} \frac{N_g!}{\prod_{k=1}^{K} F_{kg}!} \prod_{k=1}^{K_g} P_g(u_{kj}|\boldsymbol{\xi},\boldsymbol{\tau}_g)^{F_{kg}} \tag{4.12}$$

となる。ただし、N_g は g 番目のグループに属する受験者人数、K_g は g 番目のグループにおける反応パタンのバリエーションの数、F_{kg} は g 番目のグループの中での、k 番目のバリエーションとなった受験者の人数を表す。

数値計算を行う際の対数尤度は、4.12 式を尤度関数とみなしたものから、定数項を除いた部分の対数、すなわち

$$\log L(\boldsymbol{U}|\boldsymbol{\xi},\boldsymbol{\tau}) = \sum_{g=1}^{G}\sum_{k=1}^{K_g} F_{kg} \log P_g(u_{kj}|\boldsymbol{\xi},\boldsymbol{\tau}_g) \tag{4.13}$$

となる。

テストの実践場面においては、各グループごとに $\theta_{i\in g} \sim N(\mu_g, \sigma_g^2)$ とおくことが多い。これは、結果の解釈が容易なためと、もともと受験者の能力が正規分布に従うことが仮定されているような場合に対応させるためである。この場合、4.13 式において、$\boldsymbol{\tau}_g = (\mu_g, \sigma_g^2)$ の E ステップにおける μ_g の推定値 $\hat{\mu}_g$ を、

$$\hat{\mu}_g = \frac{1}{N_g} \sum_{k=1}^{F_k} \frac{F_{kg}}{P_g(u_{kj}|\boldsymbol{\xi},\mu_g,\sigma_g^2)} \sum_{q=1}^{Q} h_q P_j(u_{kj}|h_q,\mu_g,\sigma_g^2) h_g(h_q|\mu_g,\sigma_g^2) \tag{4.14}$$

で得ることができる。ただし、前節と同様、θ の積分を h_q による和で置き換えた形で示した。また、σ_g の推定値 $\hat{\sigma}_g$ を、

$$\hat{\sigma}_g = \frac{1}{N_g} \sum_{k=1}^{F_k} F_{kg} \left[(\bar{\theta}_{kg} - \hat{\mu}_g)^2 + \frac{1}{P_g(u_{kj}|\boldsymbol{\xi},\mu_g,\sigma_g^2)} \sum_{q=1}^{Q} (h_q - \hat{\theta}_{kg}) h_q P_j(u_{kj}|h_q,\mu_g,\sigma_g^2) h_g(h_q|\mu_g,\sigma_g^2) \right] \tag{4.15}$$

と導くことができ、これらの値を E ステップの期待対数尤度に用いることで、EM アルゴリズムによる推定を行うことができる。ただし

$$\bar{\theta}_{kg} = \frac{1}{P_g(u_{kj}|\boldsymbol{\xi},\mu_g,\sigma_g^2)} \sum_{q=1}^{Q} (h_q - \hat{\theta}_{kg}) h_q P_j(u_{kj}|h_q,\mu_g,\sigma_g^2) h_g(h_q|\mu_g,\sigma_g^2) \tag{4.16}$$

である。

以上の計算を行う場合、データ u_{kj} および F_k、それに受験者のグループに関する情報が必要である。多群 IRT の期待対数尤度の式は、各々のグループにおける母集団分布 h_g を、項目パラメタ

を 1 グループで推定する場合の期待度数の計算式にかけ合わせる形をとっている。ここで推定される項目パラメタは、グループをまたいで共通であるから、h_g をまたいで共通のパラメタである。よって、期待対数尤度の計算の過程で、g 個の別々な行列が必要となり、1 グループの場合に比べてメモリの使用量が増えることは避けられない。このことは、グループの数を増やすことによって、同時推定が推定不可能になる可能性を示している。

4.2 メモリ不足となる要因

4.2.1 モデルの上での要因

前節で述べたように、多群 IRT の計算場面において、計算に要するリソースは、グループの数 g に比例する。また、求積点 q の数にも依存して増えることがわかる。ただし、q を増やすことは、数値積分によって消去される θ の項について、その精度を高めるという効果はあるものの、大きくすればするほど実践上のメリットが増すというわけではない。むしろ、一定の水準以上があればよい、という性質の値であろう。

また、受験者に提示する項目数 j や受験者数 i も、メモリの消費量を決める要素になりうる。しかし、前節で述べたように、得られた反応パタンのバリエーションが多いほど、メモリの使用量も増大する。「1 フォームに 1 グループが対応するテストデザイン」をとることによって、1 フォーム当たりの項目数が多ければ大きいほど、とりうる 0-1 パタンのバリエーションも増える。一方、フォームの数が増えた場合、仮に 2 フォームの間に共通項目が多く存在する場合は、その部分に関して 0-1 パタンのバリエーションが増える要因となりうる。したがって、項目数、とくにフォーム間に共通の項目数が全項目の中でどれだけ含まれているかが、メモリ消費量の増大の要因として挙げられる。

4.2.2 計算機環境に起因する要因

近年では、特別に設計された大型コンピュータでなくても、パソコンを使用することによって多群 IRT モデルによる項目パラメタ推定が可能となってきた。しかし、特に 32bit 環境の OS(operating system) を導入したパソコンを使用する場合、その OS の上では、すべてのプログラムが 32bit のメモリ空間を用いる、という前提で計算機環境が設計されている。32bit の環境におけるメモリ番地指定のバリエーションは、一般的に $2^{32} = 4294967296$bit である。言い換えれば、32bit 環境においては、2^{32} 個のメモリ番地しか管理することができない。ただし、特別な仕組みによって、これを超えるだけのメモリ空間を管理する OS もある。しかし、そのような仕組みを持った OS であったとしても、1 つのプロセスが使用するメモリ空間は有限であり、大きなリソースを消費するようなプログラムの実行をテストのたびごとに同時推定場面で行うことは、テスト実践場面では現実的でない。

4.2.3 研究の目的

以上を踏まえ、テスト実践場面で一般的である代表的な多群 IRT モデルに基づくパラメタ推定ソフトウェアである BILOG-MG をとりあげ、実際にグループの数や項目数を増やした場合、あるいは求積点を増やした場合に、メモリの使用量がどのように変化するか調べる。また、計算が不可能になる限界となる点を調べることを目的とする。

4.3 方法

4.3.1 想定されたテストデザイン

テストデザインとしては、次章で扱う多群 IRT モデルと同様とした。すなわち、20 項目を N_{trial} 人からなる受験者に予備テストとして提示し、規準集団上での θ を定義した上で、それに続く第 n 回の本試験に N_{exam} 人が解答する場面を想定した。ただし、本試験では予備テストで提示された項目のうち 10 項目と、K 項目の「第 n 回試験実施時点における新作項目」の、$(10+K)$ 項目が提示されるものとした。このテストデザインでは、n 回目の試験実施場面において、G 回分のテストフォームが同時推定される $(G=n+1)$。デザインを図 5.1 に示した。

受験者の人数については、現実のテスト場面を想定し、またモデルから考えうる反応パタンのバリエーションをなるべく尽くすように、N_{trial} および N_{exam} をどちらも 1000 とした。また、新作項目数 K は 10 から 100 までの 10 通りを想定した。この場合、新作項目割合 P は、$K/10$、すなわち 1 から 10 までの値に対応する。

4.3.2 実験材料および仮定されたモデル

項目パラメタの推定には BILOG-MG 3 を用いた。また、前項のテストデザインに従うデータは、RESGEN4 で生成した。これらのプログラムを、Windows 7 Professional 上の cygwin bash shell でバッチ実行した。マシンのメモリは 8GB 搭載されているものの、実際に OS が管理しているメモリ空間は 2.99GB となっている状態でバッチ実行した。

BILOG-MG におけるメモリ管理は、バッチ実行の場合、コマンドライン上から `BLM1 *.blm NUM=90000` などとする (ただし`*.blm` は任意の BILOG-MG のコマンドファイル名を表す) と、使用可能なメモリの Numerical space の上限が変更できる（character space の上限は、`NUM=`のかわりに `CHAR=` を指定すると変更できる)。EM アルゴリズムによる項目パラメタの推定は Phase 2 で行われるので、Phase 2 の実行プログラムである `BLM2` に、`NUM=95000` を指定して推定を行った。BILOG-MG の出力によると、この状態で使用可能なメモリ量は 389120000byte となった (おそらく $95000 \times 1024 \times 4 = 389120000$ を示すと考える)。`NUM=95000` を超えると、「メモリが allocate できない」旨のエラーとなり、推定できなくなったので、この状態がメモリ使用量の上限であると判断した。

モデルは 2PL とした。ここで、求積点 q の個数について、BILOG-MG の CALIB コマンドで NQPT=オプションを指定し、求積点の数を 15 から 45 まで 5 刻みで操作した。多群の推定については、1 フォームに対して 1 グループを指定した。

4.4 結果

4.4.1 推定時のメモリ使用量

図 4.1 から図 4.7 に、順に $q = 15$、$q = 20$、$q = 25$、$q = 30$、$q = 35$、$q = 40$、$q = 45$ の場合の、フォーム数 g と新作項目数 K 別のメモリ消費量を示した。なお、図 4.1 から図 4.7 にある「P」に続く数字は、新作項目割合、すなわち $K/10$ を示す。

図 4.1 によると、メモリの消費量は、K が大きい場合、小さい場合ともにグループ数のべき乗に比例するので、K が小さい場合にはメモリ不足になるまでの g がより多くなるということがわかる。また、図 4.1 と図 4.2 を比較すると、$q = 15$ と $q = 20$ の場合で、どの g や K の場合でもメモリの消費量がほぼ 1.33 倍、すなわち (20/15) 倍になることが見て取れる。同様に、図 4.1 と図 4.4 を比較すると、q の値が倍になると、どの g や K の場合でもメモリの消費量がほぼ 2 倍になることが分かる。このように、q の値は、ほぼ線形にメモリ使用量を増加させることが分かった。

4.4.2 メモリ使用量の予測式の導出

以上の結果から、g からメモリ消費量 M を予測するためのべき乗回帰式

$$M = ag^b + e \tag{4.17}$$

を各 q および K の条件について当てはめたところ、表 4.1 のようになった。

各 q の条件における表 4.1 の回帰係数の推定結果を見ると、パラメタ b における結果は、K が大きくなればおおむね 2.927 前後の値となることが分かった。また、パラメタ a に関しては、a と K の関係を知るために、式 4.18 に示す単回帰モデルのパラメタ b_0 および b_1 をそれぞれの q ごとに推定した。結果を表 4.2 に示した。

$$a = b_1 K + b_0 + e,\ e \sim N(0, 1) \tag{4.18}$$

表 4.2 の結果より、さらに q から b_0 および b_1 を予測する式を考えることで、M を q、K、g で表すような推定式を求めた。表 4.2 より、q の増加に伴い、b_0、b_1 ともに線形に増加していると考え、単回帰モデルを想定した。q から b_0 および b_1 を予測する単回帰式は、以下のようになった。

$$b_1 = 19.729q + 31.447 \tag{4.19}$$
$$b_0 = 72.481q - 855.910 \tag{4.20}$$

ただし、b_1 の予測式に関しては、回帰係数の標準誤差が 0.026、切片の標準誤差が 0.807 であった。また、ただし、b_0 の予測式に関しては、回帰係数の標準誤差が 1.803、切片の標準誤差が 57.03 と、比較的大きな値となった。

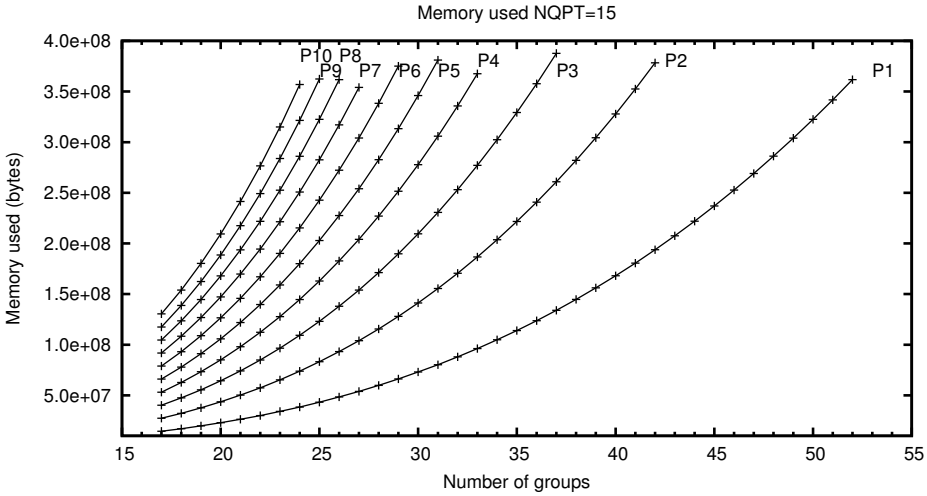

図 4.1　メモリ消費量 ($q = 15$)。P に続く数字は、新作項目割合を表す

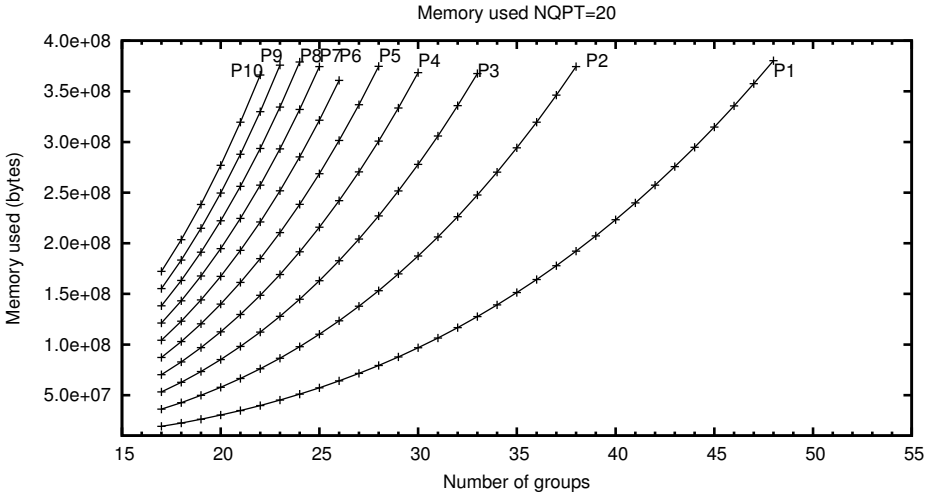

図 4.2　メモリ消費量 ($q = 20$)。P に続く数字は、新作項目割合を表す

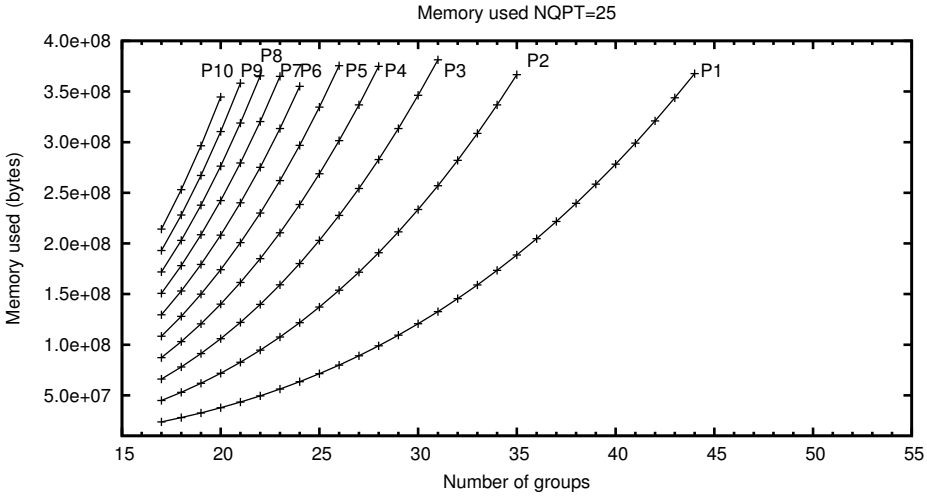

図 4.3　メモリ消費量 ($q = 25$)。P に続く数字は、新作項目割合を表す

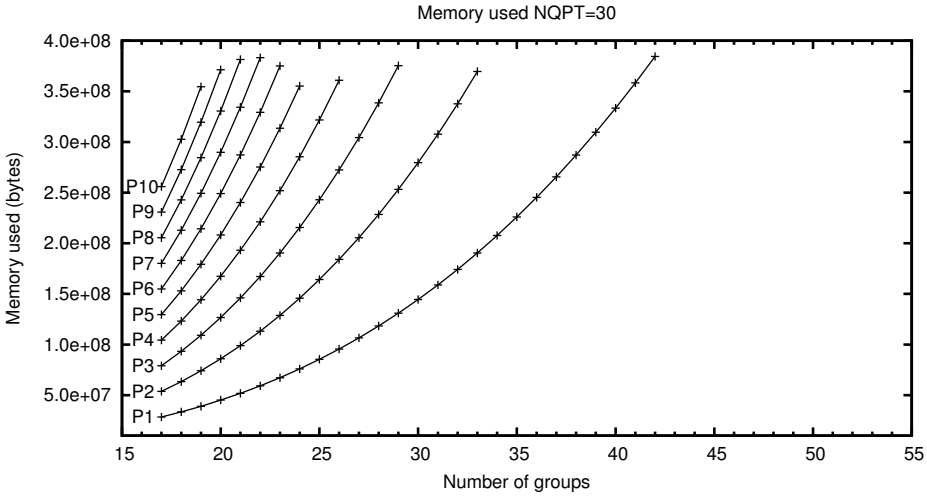

図 4.4　メモリ消費量 ($q = 30$)。P に続く数字は、新作項目割合を表す

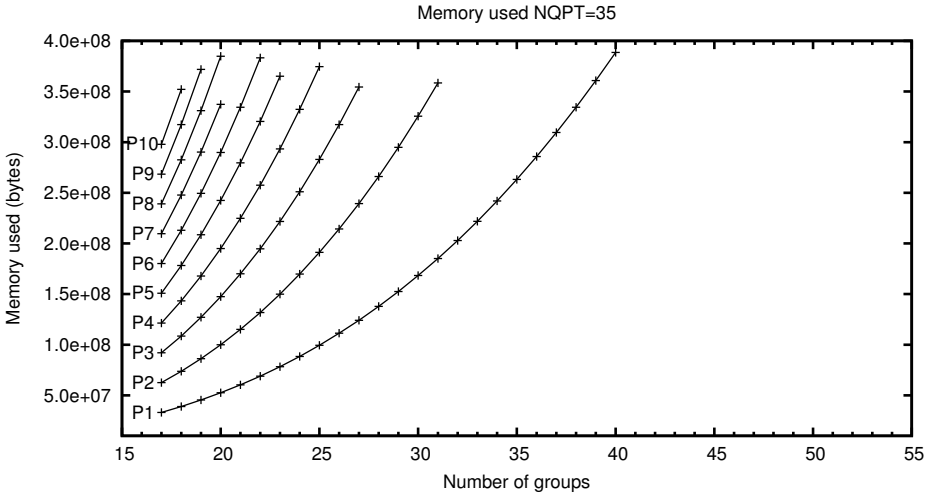

図 4.5　メモリ消費量 ($q = 35$)。P に続く数字は、新作項目割合を表す

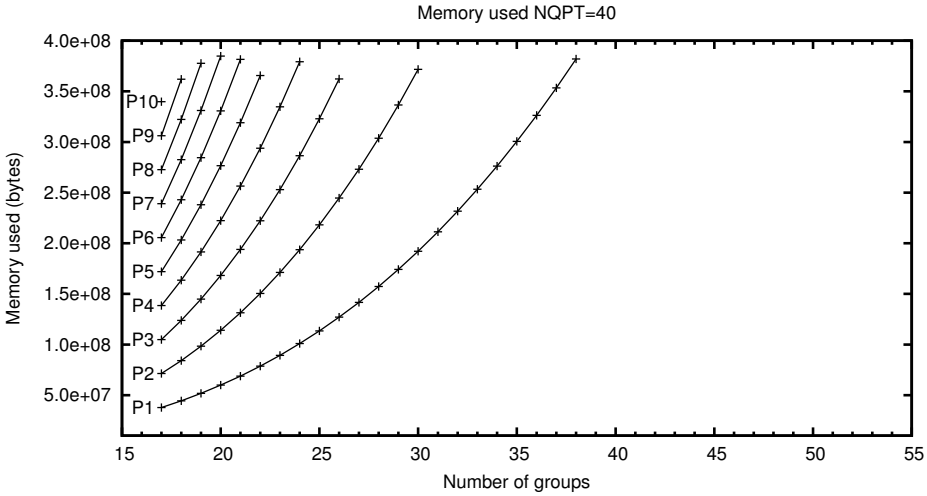

図 4.6　メモリ消費量 ($q = 40$)。P に続く数字は、新作項目割合を表す

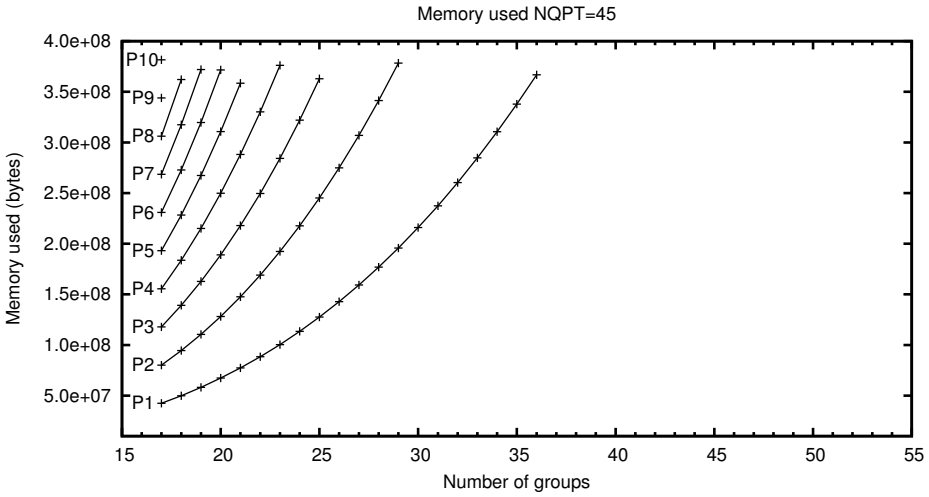

図 4.7　メモリ消費量 ($q = 45$)。P に続く数字は、新作項目割合を表す

以上より、本研究で取り上げたテストデザインにおける M の推定値 $\hat{M}$ は、q、K、g、および $b = 2.927$ を用いて、

$$\hat{M} = ag^b \tag{4.21}$$
$$= (b_1K + b_0)g^b \tag{4.22}$$
$$= ((19.729q + 31.447)K + (72.481q - 855.910)) \times g^{2.927} \tag{4.23}$$

と書き表すことができた。

4.4.3　テスト場面における推定の可否

先の結果において、図 4.1 から図 4.7 に示したメモリ使用量は、一定の上限に達するとそれ以上の q や g、K の場合において推定ができなくなる。そこで、テスト実践場面において、実際に推定ができるかどうかを明らかにするため、表 4.3 に、$q = 15$ の場合の、各 K とグループ数 G に対応する推定の可否を示した。また、表 4.4 から表 4.9 に、順に $q = 20$、$q = 25$、$q = 30$、$q = 35$、$q = 40$、$q = 45$ の場合の推定の可否を示した。

表 4.3 によると、$q = 15$、$K = 10$ の場合に、48 グループ（本試験 47 回）条件でもメモリ不足に陥らずに推定できている。しかし、53 グループを超える条件の場合、メモリ不足に陥るという結果となった。同様に、表 4.4 によると、$q = 20$、$K = 10$ の場合も、48 グループを超えるとメモリが不足するという結果となった。

	K=10	K=20	K=30	K=40	K=50	K=60	K=70	K=80	K=90	K=100
q=15										
a	3682.3	6855.9	10053.1	13328.2	16548.9	19831.1	23179.0	26448.1	29749.9	33085.5
se(a)	30.4	46.9	62.0	76.6	89.4	101.7	112.9	123.7	133.8	142.9
b	2.90885	2.92091	2.92423	2.92400	2.92478	2.92429	2.92297	2.92292	2.92251	2.92182
se(b)	0.00215	0.00189	0.00176	0.00170	0.00162	0.00157	0.00152	0.00148	0.00144	0.00140
q=20										
a	4930.3	9138.0	13367.3	17597.2	21824.5	26128.5	30350.5	34610.8	38910.3	43250.3
se(a)	41.5	61.4	78.3	93.1	106.4	118.4	129.9	140.6	150.1	158.1
b	2.90649	2.91945	2.92325	2.92503	2.92609	2.92580	2.92645	2.92656	2.92629	2.92575
se(b)	0.00224	0.00190	0.00173	0.00160	0.00151	0.00143	0.00137	0.00131	0.00126	0.00121
q=25										
a	6244.7	11457.3	16632.8	21871.1	27102.4	32425.2	37645.5	42911.1	48222.9	53582.3
se(a)	54.1	75.8	93.1	108.4	121.8	132.7	143.8	153.3	161.0	166.2
b	2.90198	2.91750	2.92346	2.92552	2.92678	2.92666	2.92744	2.92766	2.92750	2.92705
se(b)	0.00236	0.00192	0.00168	0.00153	0.00142	0.00132	0.00125	0.00118	0.00112	0.00105
q=30										
a	7529.8	13775.7	19966.8	26229.5	32479.7	38633.7	44838.3	51094.4	57402.9	63765.3
se(a)	65.7	89.5	107.4	122.3	134.7	147.1	158.0	166.8	173.0	176.0
b	2.9000	2.9161	2.9225	2.9248	2.9262	2.9280	2.9289	2.9292	2.9291	2.9287
se(b)	0.00241	0.00192	0.00165	0.00147	0.00134	0.00125	0.00117	0.00109	0.00102	0.00095
q=35										
a	8852.5	16151.9	23375.1	30508.0	37759.1	44897.3	52357.9	59346.0	66658.1	-
se(a)	78.0	103.1	120.8	135.9	147.8	160.0	162.7	177.4	181.3	
b	2.89735	2.91402	2.92071	2.92503	2.92660	2.92848	2.92770	2.92983	2.92981	-
se(b)	0.00246	0.00192	0.00162	0.00142	0.00128	0.00118	0.00105	0.00101	0.00093	
q=40										
a	10216.7	18469.3	26703.1	34829.5	43088.7	51217.5	59410.1	67667.2	-	-
se(a)	91.0	115.9	133.7	148.6	159.3	170.7	179.3	184.3		
b	2.89411	2.91339	2.92032	2.92482	2.92651	2.92848	2.92952	2.92996	-	-
se(b)	0.00253	0.00190	0.00158	0.00138	0.00122	0.00112	0.00102	0.00093		
q=45										
a	11626.8	20816.3	30068.7	39195.5	48470.3	57595.9	66792.3	-	-	-
se(a)	104.6	128.6	145.9	160.1	168.9	178.8	185.0			
b	2.89038	2.91244	2.91959	2.92426	2.92605	2.92811	2.92922	-	-	-
se(b)	0.00259	0.00189	0.00155	0.00134	0.00117	0.00105	0.00095			

表 4.1　各 q における、項目数 K=10 から 100 の場合のべき乗回帰結果。se(a) は係数 a の、se(b) は係数 b の、それぞれ漸近標準誤差を表す。「-」は、当該条件でデータ数が 3 未満のため、推定できなかったことを示す

4.5　考察

これらの結果から、BILOG-MG においては、フォームごとに 1 グループを仮定するモデルにおいては、特に新作項目割合が大きい場合、メモリ不足により推定できなくなる場合があることが分かった。これに類似するたのプログラム (たとえば、ICL など) においても、メモリの管理手法による違いはあるものの、メモリ不足に陥る可能性があることが予想される。また、BILOG-MG は標準的なテスト場面や研究においても用いられていることもあり、BILOG-MG 以外の計算方法に関しては、十分な計算結果の検証を経たうえでなければテストの実践場面に使用するのは難しいと考える。

BILOG-MG を用いた場合、$q = 15$、$K = 10$ の場合であっても、4.23 式から、$g = 121$ 以上の場合、メモリの上限が 32bit 環境の一般的上限である 2^{32}byte を超えると予想される。特に、4.23 式より、g に関しては約 2.927 乗のオーダーで $\hat{M}$ が増加するということを意味し、グループの数はメモリ消費量に大きな影響を与えていることが見て取れる。g が多くなる場合は、毎回行われる

q	b_1	se(b_1)	b_0	se(b_0)
15	327.089	0.817	286.32	50.68
20	425.495	0.567	608.54	35.16
25	525.718	0.767	895.02	47.57
30	623.714	0.727	1267.32	45.11
35	721.933	1.030	1670.68	57.97
40	819.833	0.599	2057.77	30.25
45	919.490	0.659	2444.07	29.49

表 4.2　各 q 条件における単回帰係数の推定結果。se(b_0) および se(b_1) はそれぞれ b_0、b_1 の漸近標準誤差を表す

試験に複数のフォームを提示する場合 (1.5.2 節参照) であり、たとえば毎回 10 フォーム提示されていれば、12 回の試験で $g = 120$ となる。したがって、テスト実施のたびにフォーム数が増え、それに伴ってグループ数が増加するテスト実施法で多群 IRT モデルによる同時推定を行うためには、グループの数 g を減らすことが必須であるといえる。いくつかのフォームにおける受験者が、同一の母集団からサンプリングされていると考えて、グループをまとめることにより、フォームの数が増えても g が増えないような多群 IRT モデルを考えればよい。

本研究で提案されたメモリ消費量の予測式は、あくまで $q = 15$ から 45 までの場合におけるデータからの予測値である。q の値が小さな場合、あるいは、K の値が小さい場合には、グループの数が少ない場合に、当てはまりが悪い。これは、主として 4.18 式の g の係数 b が、K の小さな条件で低い値になっていることに起因する。ただし、この部分の精度を高めることは、q の値を小さくしすぎることにより、項目パラメタの推定に影響を及ぼすことが懸念される。特に多群 IRT モデルを用いた垂直等化場面では、それぞれのグループに特定の範囲の学力を持つ集団が割り当てられる場合があり、複数の学力群をまたいだ等化を行う必要があることから、θ の求積点の範囲を広げた方が良いという場合があり、q の値を多くとる必要があることが指摘できる。

また、q を減らすことによるメモリ節約の効果は、新作項目数 K が小さい場合にはある程度見られることがわかる。BILOG-MG のマニュアル (du Toit, 2003) には、デフォルトの `NQPT=`の値として、1 グループの場合は 10、多群の場合は 20 とある。しかし、この点の個数を減らすことは、特にモデルとかけ離れたパラメタの推定値となるようなデータの分析においては、E ステップの期待度数の推定に影響する恐れがあるので、あまり小さくするのは好ましくない。ただし、受験者数が少ない場合（正確には、とりうる u_{kj} のバリエーションが少ない場合）は、q を多くとると、一つ一つの求積点に入る受験者の期待度数が少なくなるので、q はデフォルトの 10 で十分であり、それ以上の値を用いても、項目パラメタ推定の安定性や精度に影響はないように見える (村木、2011、p.121)。いずれにしても、q を極端に減らすことは好ましくないといえ、この方法によるメモリの節約も限界があることが示唆される。

なお、グループの数を減らすと、メモリの使用量は劇的に減少する。たとえば、$q = 30$(表 4.6)、

33 グループ（本試験 32 回）で $K = 20$ の条件において「推定可能」という結果となった。この試行において、使用されたメモリは 369581152 byte であった。同じ 0-1 データに対し、グループを「規準集団」と「実受験者」の 2 グループとして、多群 IRT モデルを適用したところ、メモリの使用量は 4295228 byte であった。メモリの使用量は、グループを減らしたことにより、およそ 86 分の 1 で済んだことになる。しかし、実受験者グループとしてグループをひとまとめにした場合、実際には提示されたフォームの違い (すなわちグループの違い) によって母集団における θ の分布が異なるモデルを当てはめた方が、当てはまりが良くなるにもかかわらず、グループの違いを無視してしまっていることになる。また、項目パラメタの推定値にどのような影響が出るかは不明である。次章では、多群 IRT モデルにおいて、グループの数を減らしてモデルを簡素化した場合における、項目パラメタの推定について述べる。

表 4.3　NQPT=15 の場合の、BILOG-MG による項目パラメタ推定の可否。新作項目数 K（横列）とグループ数（縦行）別に示した

	10	20	30	40	50	60	70	80	90	100
17	○	○	○	○	○	○	○	○	○	○
18	○	○	○	○	○	○	○	○	○	○
19	○	○	○	○	○	○	○	○	○	○
20	○	○	○	○	○	○	○	○	○	○
21	○	○	○	○	○	○	○	○	○	○
22	○	○	○	○	○	○	○	○	○	○
23	○	○	○	○	○	○	○	○	○	○
24	○	○	○	○	○	○	○	○	○	○
25	○	○	○	○	○	○	○	○	○	×
26	○	○	○	○	○	○	○	○	×	×
27	○	○	○	○	○	○	○	×	×	×
28	○	○	○	○	○	○	×	×	×	×
29	○	○	○	○	○	○	×	×	×	×
30	○	○	○	○	○	×	×	×	×	×
31	○	○	○	○	○	×	×	×	×	×
32	○	○	○	○	×	×	×	×	×	×
33	○	○	○	○	×	×	×	×	×	×
34	○	○	○	×	×	×	×	×	×	×
35	○	○	○	×	×	×	×	×	×	×
36	○	○	○	×	×	×	×	×	×	×
37	○	○	○	×	×	×	×	×	×	×
38	○	○	×	×	×	×	×	×	×	×
39	○	○	×	×	×	×	×	×	×	×
40	○	○	×	×	×	×	×	×	×	×
41	○	○	×	×	×	×	×	×	×	×
42	○	○	×	×	×	×	×	×	×	×
43	○	×	×	×	×	×	×	×	×	×
44	○	×	×	×	×	×	×	×	×	×
45	○	×	×	×	×	×	×	×	×	×
46	○	×	×	×	×	×	×	×	×	×
47	○	×	×	×	×	×	×	×	×	×
48	○	×	×	×	×	×	×	×	×	×

表 4.4　NQPT=20 の場合の、BILOG-MG による項目パラメタ推定の可否。新作項目数 K（横列）とグループ数（縦行）別に示した

	10	20	30	40	50	60	70	80	90	100
17	○	○	○	○	○	○	○	○	○	○
18	○	○	○	○	○	○	○	○	○	○
19	○	○	○	○	○	○	○	○	○	○
20	○	○	○	○	○	○	○	○	○	○
21	○	○	○	○	○	○	○	○	○	○
22	○	○	○	○	○	○	○	○	○	○
23	○	○	○	○	○	○	○	○	○	×
24	○	○	○	○	○	○	○	○	×	×
25	○	○	○	○	○	○	○	×	×	×
26	○	○	○	○	○	○	×	×	×	×
27	○	○	○	○	○	×	×	×	×	×
28	○	○	○	○	○	×	×	×	×	×
29	○	○	○	○	×	×	×	×	×	×
30	○	○	○	○	×	×	×	×	×	×
31	○	○	○	×	×	×	×	×	×	×
32	○	○	○	×	×	×	×	×	×	×
33	○	○	○	×	×	×	×	×	×	×
34	○	○	×	×	×	×	×	×	×	×
35	○	○	×	×	×	×	×	×	×	×
36	○	○	×	×	×	×	×	×	×	×
37	○	○	×	×	×	×	×	×	×	×
38	○	○	×	×	×	×	×	×	×	×
39	○	×	×	×	×	×	×	×	×	×
40	○	×	×	×	×	×	×	×	×	×
41	○	×	×	×	×	×	×	×	×	×
42	○	×	×	×	×	×	×	×	×	×
43	○	×	×	×	×	×	×	×	×	×
44	○	×	×	×	×	×	×	×	×	×
45	○	×	×	×	×	×	×	×	×	×
46	○	×	×	×	×	×	×	×	×	×
47	○	×	×	×	×	×	×	×	×	×
48	○	×	×	×	×	×	×	×	×	×

表 4.5　NQPT=25 の場合の、BILOG-MG による項目パラメタ推定の可否。新作項目数 K（横列）とグループ数（縦行）別に示した

	10	20	30	40	50	60	70	80	90	100
17	○	○	○	○	○	○	○	○	○	○
18	○	○	○	○	○	○	○	○	○	○
19	○	○	○	○	○	○	○	○	○	○
20	○	○	○	○	○	○	○	○	○	○
21	○	○	○	○	○	○	○	○	○	×
22	○	○	○	○	○	○	○	○	×	×
23	○	○	○	○	○	○	○	×	×	×
24	○	○	○	○	○	○	×	×	×	×
25	○	○	○	○	○	×	×	×	×	×
26	○	○	○	○	○	×	×	×	×	×
27	○	○	○	○	×	×	×	×	×	×
28	○	○	○	○	×	×	×	×	×	×
29	○	○	○	×	×	×	×	×	×	×
30	○	○	○	×	×	×	×	×	×	×
31	○	○	○	×	×	×	×	×	×	×
32	○	○	×	×	×	×	×	×	×	×
33	○	○	×	×	×	×	×	×	×	×
34	○	○	×	×	×	×	×	×	×	×
35	○	○	×	×	×	×	×	×	×	×
36	○	×	×	×	×	×	×	×	×	×
37	○	×	×	×	×	×	×	×	×	×
38	○	×	×	×	×	×	×	×	×	×
39	○	×	×	×	×	×	×	×	×	×
40	○	×	×	×	×	×	×	×	×	×
41	○	×	×	×	×	×	×	×	×	×
42	○	×	×	×	×	×	×	×	×	×
43	○	×	×	×	×	×	×	×	×	×
44	○	×	×	×	×	×	×	×	×	×
45	○	×	×	×	×	×	×	×	×	×
46	×	×	×	×	×	×	×	×	×	×
47	×	×	×	×	×	×	×	×	×	×
48	×	×	×	×	×	×	×	×	×	×

表 4.6　NQPT=30 の場合の、BILOG-MG による項目パラメタ推定の可否。新作項目数 K（横列）とグループ数（縦行）別に示した

	10	20	30	40	50	60	70	80	90	100
17	○	○	○	○	○	○	○	○	○	○
18	○	○	○	○	○	○	○	○	○	○
19	○	○	○	○	○	○	○	○	○	○
20	○	○	○	○	○	○	○	○	○	×
21	○	○	○	○	○	○	○	○	×	×
22	○	○	○	○	○	○	○	×	×	×
23	○	○	○	○	○	○	×	×	×	×
24	○	○	○	○	○	×	×	×	×	×
25	○	○	○	○	×	×	×	×	×	×
26	○	○	○	○	×	×	×	×	×	×
27	○	○	○	×	×	×	×	×	×	×
28	○	○	○	×	×	×	×	×	×	×
29	○	○	○	×	×	×	×	×	×	×
30	○	○	×	×	×	×	×	×	×	×
31	○	○	×	×	×	×	×	×	×	×
32	○	○	×	×	×	×	×	×	×	×
33	○	○	×	×	×	×	×	×	×	×
34	○	×	×	×	×	×	×	×	×	×
35	○	×	×	×	×	×	×	×	×	×
36	○	×	×	×	×	×	×	×	×	×
37	○	×	×	×	×	×	×	×	×	×
38	○	×	×	×	×	×	×	×	×	×
39	○	×	×	×	×	×	×	×	×	×
40	○	×	×	×	×	×	×	×	×	×
41	○	×	×	×	×	×	×	×	×	×
42	○	×	×	×	×	×	×	×	×	×
43	×	×	×	×	×	×	×	×	×	×
44	×	×	×	×	×	×	×	×	×	×
45	×	×	×	×	×	×	×	×	×	×
46	×	×	×	×	×	×	×	×	×	×
47	×	×	×	×	×	×	×	×	×	×
48	×	×	×	×	×	×	×	×	×	×

表 4.7　NQPT=35 の場合の、BILOG-MG による項目パラメタ推定の可否。新作項目数 K（横列）とグループ数（縦行）別に示した

	10	20	30	40	50	60	70	80	90	100
17	○	○	○	○	○	○	○	○	○	○
18	○	○	○	○	○	○	○	○	○	○
19	○	○	○	○	○	○	○	○	○	×
20	○	○	○	○	○	○	○	○	×	×
21	○	○	○	○	○	○	×	×	×	×
22	○	○	○	○	○	○	×	×	×	×
23	○	○	○	○	○	×	×	×	×	×
24	○	○	○	○	×	×	×	×	×	×
25	○	○	○	○	×	×	×	×	×	×
26	○	○	○	×	×	×	×	×	×	×
27	○	○	○	×	×	×	×	×	×	×
28	○	○	×	×	×	×	×	×	×	×
29	○	○	×	×	×	×	×	×	×	×
30	○	○	×	×	×	×	×	×	×	×
31	○	○	×	×	×	×	×	×	×	×
32	○	×	×	×	×	×	×	×	×	×
33	○	×	×	×	×	×	×	×	×	×
34	○	×	×	×	×	×	×	×	×	×
35	○	×	×	×	×	×	×	×	×	×
36	○	×	×	×	×	×	×	×	×	×
37	○	×	×	×	×	×	×	×	×	×
38	○	×	×	×	×	×	×	×	×	×
39	○	×	×	×	×	×	×	×	×	×
40	○	×	×	×	×	×	×	×	×	×
41	×	×	×	×	×	×	×	×	×	×
42	×	×	×	×	×	×	×	×	×	×
43	×	×	×	×	×	×	×	×	×	×
44	×	×	×	×	×	×	×	×	×	×
45	×	×	×	×	×	×	×	×	×	×
46	×	×	×	×	×	×	×	×	×	×
47	×	×	×	×	×	×	×	×	×	×
48	×	×	×	×	×	×	×	×	×	×

表 4.8　NQPT=40 の場合の、BILOG-MG による項目パラメタ推定の可否。新作項目数 K（横列）とグループ数（縦行）別に示した

	10	20	30	40	50	60	70	80	90	100
17	○	○	○	○	○	○	○	○	○	○
18	○	○	○	○	○	○	○	○	○	×
19	○	○	○	○	○	○	○	○	×	×
20	○	○	○	○	○	○	○	×	×	×
21	○	○	○	○	○	○	×	×	×	×
22	○	○	○	○	○	×	×	×	×	×
23	○	○	○	○	×	×	×	×	×	×
24	○	○	○	○	×	×	×	×	×	×
25	○	○	○	×	×	×	×	×	×	×
26	○	○	○	×	×	×	×	×	×	×
27	○	○	×	×	×	×	×	×	×	×
28	○	○	×	×	×	×	×	×	×	×
29	○	○	×	×	×	×	×	×	×	×
30	○	○	×	×	×	×	×	×	×	×
31	○	×	×	×	×	×	×	×	×	×
32	○	×	×	×	×	×	×	×	×	×
33	○	×	×	×	×	×	×	×	×	×
34	○	×	×	×	×	×	×	×	×	×
35	○	×	×	×	×	×	×	×	×	×
36	○	×	×	×	×	×	×	×	×	×
37	○	×	×	×	×	×	×	×	×	×
38	○	×	×	×	×	×	×	×	×	×
39	×	×	×	×	×	×	×	×	×	×
40	×	×	×	×	×	×	×	×	×	×
41	×	×	×	×	×	×	×	×	×	×
42	×	×	×	×	×	×	×	×	×	×
43	×	×	×	×	×	×	×	×	×	×
44	×	×	×	×	×	×	×	×	×	×
45	×	×	×	×	×	×	×	×	×	×
46	×	×	×	×	×	×	×	×	×	×
47	×	×	×	×	×	×	×	×	×	×
48	×	×	×	×	×	×	×	×	×	×

表 4.9　NQPT=45 の場合の、BILOG-MG による項目パラメタ推定の可否。新作項目数 K（横列）とグループ数（縦行）別に示した

	10	20	30	40	50	60	70	80	90	100
17	○	○	○	○	○	○	○	○	○	○
18	○	○	○	○	○	○	○	○	×	×
19	○	○	○	○	○	○	○	×	×	×
20	○	○	○	○	○	○	×	×	×	×
21	○	○	○	○	○	×	×	×	×	×
22	○	○	○	○	×	×	×	×	×	×
23	○	○	○	○	×	×	×	×	×	×
24	○	○	○	×	×	×	×	×	×	×
25	○	○	○	×	×	×	×	×	×	×
26	○	○	×	×	×	×	×	×	×	×
27	○	○	×	×	×	×	×	×	×	×
28	○	○	×	×	×	×	×	×	×	×
29	○	○	×	×	×	×	×	×	×	×
30	○	×	×	×	×	×	×	×	×	×
31	○	×	×	×	×	×	×	×	×	×
32	○	×	×	×	×	×	×	×	×	×
33	○	×	×	×	×	×	×	×	×	×
34	○	×	×	×	×	×	×	×	×	×
35	○	×	×	×	×	×	×	×	×	×
36	○	×	×	×	×	×	×	×	×	×
37	×	×	×	×	×	×	×	×	×	×
38	×	×	×	×	×	×	×	×	×	×
39	×	×	×	×	×	×	×	×	×	×
40	×	×	×	×	×	×	×	×	×	×
41	×	×	×	×	×	×	×	×	×	×
42	×	×	×	×	×	×	×	×	×	×
43	×	×	×	×	×	×	×	×	×	×
44	×	×	×	×	×	×	×	×	×	×
45	×	×	×	×	×	×	×	×	×	×
46	×	×	×	×	×	×	×	×	×	×
47	×	×	×	×	×	×	×	×	×	×
48	×	×	×	×	×	×	×	×	×	×

第 5 章

多群 IRT モデルにおけるモデル簡素化

5.1 序論・目的

5.1.1 大きなデータ行列に対する多群 IRT モデル

前章では、BILOG-MG の項目パラメタ推定において、一つの問題冊子に一つのグループを仮定すると、グループの数が多数になった場合、メモリの不足により項目パラメタが推定できないといった現象が生じることを述べた。この問題は、BILOG-MG 特有ではなく、計算機の記憶容量や CPU が管理できるメモリ空間の制限など、計算機環境に起因することが指摘できる。グループの数を減らすことにより、また項目パラメタ推定時に積分消去される、「グループごとの θ の分布」の求積点を減らすことにより、メモリ不足の問題は回避可能であるが、求積点を減らすことによるメモリ節約効果は限定的である。

問題は、このようなグループの簡素化を行った場合と、すべての冊子に対応するグループを仮定した場合とで、推定され等化される項目パラメタが異なるか否かである。項目パラメタ推定時に同時に最大化されるモデルの最大尤度は、1 冊子 1 グループの場合が最大であるだろうが、尤度が高まるのはグループごとの θ の分布を別個に推定しているためである。学力の比較、すなわち、各グループの能力値の比較においては、等化された項目パラメタを用いてグループごとに θ を推定し、その値をグループごとに比較すればよい。したがって、グループ数を減らしたモデルであっても、等化の結果得られた項目パラメタの推定値が「1 冊子 1 グループ」モデルの推定値に近ければ、グループ数を減らした場合であっても有効な推定結果であるといえるだろう。

5.1.2 水平等化を連続して行うための試験デザイン

個別推定と同時推定のいずれが良いかという議論は、試験のデザインに依存する。たとえば、Arai and Mayekawa (2011) では比較的大きな項目バンクがあらかじめ備わっている場合を想定し、項目バンクから項目パラメタ既知の項目を共通項目として出題する場合を扱っている。また、Hanson and Béguin (2002) では、共通な項目をもつ二つの問題冊子を別々の受験者グループが解答した場面を想定している。いずれの研究においても、個別推定の結果がより真値に近いパラメタ

推定値を得た。しかし、Arai and Mayekawa (2011) で扱われているような、項目パラメタ既知の項目を予備試験で用意するテスト実施法が必ずしも実施可能であるとは限らず、この場合に適用可能なテストデザインの提案、および、そのデザインを適用した際の項目パラメタの推定値に関する研究に乏しい、といった問題点があった。本章で取り上げるテストデザインは、(1) 規準集団に予備試験を実施、(2) 予備試験を特性が似ている 2 種類に分割し、それぞれ別個に新作項目を追加した項目群から構成される問題冊子を作成し、各々を本試験の 2 グループに提示、(3) 以降、(2) で予備試験を折半した共通項目と、毎回新規に作成した新作項目とをあわせて毎年提示、というもので、(2) の段階は、経年変化を見たい期間にわたって繰り返される。その繰り返し間の θ の尺度は毎年出題され続ける共通項目によって等化され、さらに学力の基準となる集団を、実際の本試験以外の集団に置くことにより、規準集団との相対的な比較という形で毎年の受験者の学力が比較可能になる。また、同様に、毎年の 2 グループ間の学力差は、試行試験で共通な受験者が解いた共通項目によって等化されることとなり、毎年において相互に比較可能となる。このデザインは、Kolen and Brennan (2004) において、すべての受験者グループの θ の尺度が直接「単一の規準集団が受験した等化先」に等化され、他のグループを介して間接的に等化する場面がない、という特徴を持ち、等化の精度が比較的良好なことが指摘されている (pp.282-283)。また、毎回のどの本試験においても受験者にとってユニークな項目が出題される。したがって、等化のためのデザイン上、また実践上好ましいデザインであるといえる。

5.1.3 目的

以上の背景を踏まえ、本章における研究では、前節で述べたような規準集団に等化する試験実施法において、多群 IRT モデルを用いた同時推定と個別推定の間でいずれが真の項目パラメタの推定値となるかをシミュレーションを通じて検証する。その際、同時推定で仮定されるグループの数を減らした場合に、項目パラメタの推定についてどのような影響があるか検討する。さらに真の θ の分布、とりわけ θ の平均がグループ間で異なる場合、等化方法間で項目パラメタの推定結果のずれの傾向が異なるか検討する。また、新作項目割合を変えた場合に、真の θ の平均がグループ間で異なる場面のシミュレーションを通じて、新作項目割合がモデル簡素化に及ぼす影響を検討する。

5.2 シミュレーション方法

5.2.1 想定された試験場面およびモデル

学力の規準となる集団として「規準集団」を設定し，すべての試験における受験者の θ は規準集団と比較可能になるように，毎回の試験で推定された項目パラメタを等化するデザインを用いた．まず、規準集団の $\theta \sim N(0,1)$ という仮定の下で項目パラメタを推定し、あらかじめ「項目パラメタが既知」である項目を用意するための「予備試験」を実施した。次に、予備試験のうち半数の項目と、項目パラメタが未知である「新作項目」とで構成される冊子を用いて、第 1 年度第 1 回本試験として実施した。その後、予備試験のうち、第 1 回本試験で出題されなかった項目（アンカー項

目）と、第 1 回本試験とは別に用意した新作項目とを第 1 年度第 2 回本試験として実施した。以降、毎年の本試験について第 1 回と第 2 回の 2 種類の冊子を新たに作成し、アンカー項目とともに提示し続けたと想定した。また、IRT モデルとしては、個別推定、同時推定場面ともに、2PL を仮定した。

5.2.2 シミュレーションで使用したデータ、およびデザイン

まず、予備試験として、$\theta \sim N(0.00, 1.00)$ という真値を持つ N_t 人の受験者集団が、識別力の真値が平均 $\log(\alpha)$、標準偏差 0.20 の対数正規分布、困難度の真値が $N(0.00, 1.00)$ からそれぞれランダムに発生させた値をもつ 10 項目のセット二つに解答したと想定した 0-1 データをランダムに生成した。次に、本試験 1 として、予備試験のうち一つ目のセットと同じ真値を持つ項目パラメタのセットと、新作項目として識別力の真値が平均 $\log(\delta)$、標準偏差 0.20 の対数正規分布、困難度の真値が $N(0.00, 1.00)$ からそれぞれランダムに発生させた値を持つ K 項目の項目特性の情報を付加して、$(10 + K)$ 項目に N_k 人が反応したとする 0-1 反応パタンを生成した。N_k 人の受験者集団は真の $\theta \sim N(0.00, 1.00)$ とした。続いて、本試験 2 として、本試験 1 と同様な手続きでランダムに設定した真値の項目パラメタを持つ K 項目（新作項目）と、予備試験で本試験 1 に提示しなかった方のアンカー 10 項目（予備試験における真値をそのまま引き写した値を本試験 1 の冊子での真値とする。以下の本試験に関しても同様）、合わせて $(10 + K)$ 項目に N_k 人が反応したとする 0-1 反応パタンをランダムに生成した。本試験 2 では、N_k 人の受験者集団は真の $\theta \sim N(0.25, 1.00)$ とした。ここで平均 0.25 は、1 年間を通じて勉学を行ったために、受験者に平均的に起こりうる学力の上昇幅として設定した値である。続いて、本試験 3 として、本試験 1 と同様な手続きでランダムに設定した真値の項目パラメタを持つ K 項目と、予備試験で本試験 1 に提示したアンカー 10 項目、合わせて $(10 + K)$ 項目に N_k 人が反応したとする 0-1 反応パタンを作成した。本試験 3 では、本試験 2 と同様の θ の分布とした。さらに、本試験 4 として、本試験 1 と同様な手続きで設定した真値の項目パラメタのセットを持つ K 項目と、予備試験で本試験 2 に提示したアンカー 10 項目の、合わせて $(10 + K)$ 項目に N_k 人が反応したとする 0-1 データを作成した。また、本試験 4 では、Nk 人の受験者集団は真の $\theta \sim N(0.50, 1.00)$ とした。

以降、4 回の試験を 1 ブロックとして、5 回目の試験からは項目パラメタ、θ の真値の設定を本試験 1 からと同様に繰り返し適用した。したがって、奇数回目の本試験においては予備試験での一つ目のセットが共通項目として提示され、偶数回目の本試験においては二つ目のセットが共通項目となる。また、本試験の回数を 4 で割った時の余りが 1 になる回の受験者集団の真の $\theta \sim N(0.00, 1.00)$、余りが 2 または 3 になる回では真の $\theta \sim N(0.25, 1.00)$、余りが 0 になる回では真の $\theta \sim N(0.50, 1.00)$ となる。以上のデザインを図 5.1 に示した。

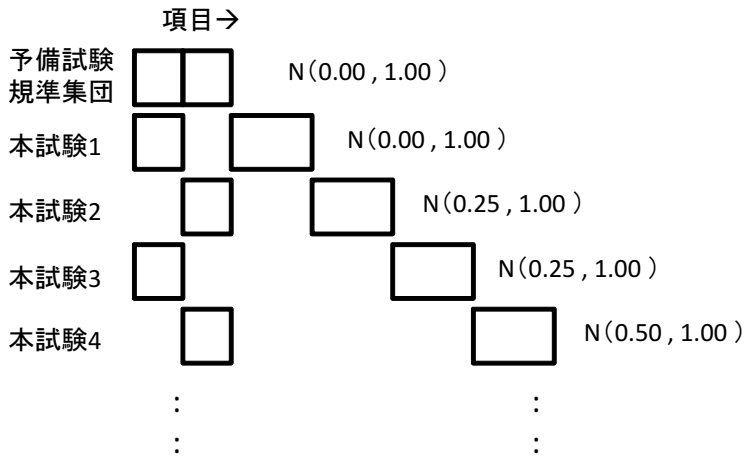

図 5.1　本研究で用いたテストデザイン。N(0.00,1.00) などの表記は、該当する受験者グループの真の能力値分布を示す。「本試験 1」～「本試験 4」までの 4 グループを「1 ブロック」と定義した

5.2.3　シミュレーションによる等化手続き

個別推定 separate calibration (SC)

個別推定では、次の 2 段階で項目パラメタを規準集団（予備試験）上のスケールに等化した。まず (1) 予備試験、および本試験のすべての回ごとに独立に項目パラメタを推定し、(2) それぞれの冊子に含まれている予備試験と共通な項目の項目パラメタを手掛かりに、予備試験の受験者集団の $\theta \sim N(0.00, 1.00)$ となるようにそれぞれの試験の項目パラメタを予備試験の項目パラメタに等化した。等化には calr 法 (Arai and Mayekawa, 2011, Appendix; 前川、1991) を使用した。この方法は、一つの冊子を規準集団 (等化先) としたとき、他の冊子 (等化元) が多数ある場合でも、等化元の冊子ごとにそれぞれ等化係数を推定することが可能である。他の方法 (たとえば、Marco (1977) の Mean-Sigma 法) においては、このような一つの規準に多数の等化元が存在する場合、それぞれの等化元ひとつひとつを個別に等化先に等化する操作を繰り返すことになり、等化先の項目パラメタが等化のたびにわずかずつ更新されていくことになる。一方、calr 法は、モデル中のパラメタにグループごとの等化係数を置いているため、等化元が複数ある場合でも等化先の項目パラメタの更新は一度で済み、解釈が容易になるというメリットがある。

同時推定（冊子ごとに 1 グループ）Concurrent Calibration (CC)

前章で行った「同時推定＋規準集団に等化」と同様の手続きを行い、「同時推定」の推定結果とした。すなわち、次の 3 段階で項目パラメタを規準集団上のスケールに等化した。(1) 規準集団（予備試験）の 20 項目のみで項目パラメタを推定した、(2) 予備試験、本試験（すべての回）の全反応データを、共通な項目が同じカラム（列）に並ぶように並べた大きな累積データセットを作成し、

これに対して全体の $\theta \sim N(0.00, 1.00)$ を仮定して項目パラメタを推定した、(3) このままでは (2) の推定結果は全体として規準集団上のスケールに乗っていないので、(2) の推定の結果得られた予備試験の項目パラメタを共通項目とし、(2) の累積分の全項目のパラメタを等化元、(1) で推定された項目パラメタを等化先とする等化を calr 法で行い、累積分の全項目の等化済み項目パラメタを得た。(2) で、累積分の大きな 0-1 データに対して項目パラメタを推定する際、一つの問題冊子（1 回の試験）に対して一つのグループを仮定した。したがって、この推定法では実施回数分のグループの定義が必要になる。 概要を図 5.2 に示した。

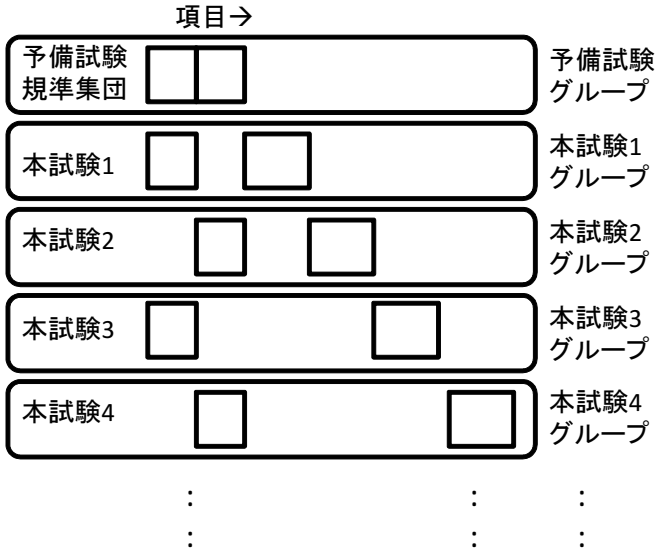

図 5.2　CC によるモデル。簡素化する前の同時推定における受験者グループの指定

同時推定（真の θ の平均値差を忠実に表現したモデル）Concurrent Calibration, Fit to True theta (CCFT)

同時推定の手続きは「同時推定＋規準集団に等化」と同様とした。ただし、もとの「真の能力値分布の平均」が同一である実施回は、同一のグループであるとし、「本試験 1 グループ」、「本試験 2 および 3 グループ」、「本試験 4 グループ」「規準集団（予備試験の冊子を解いたグループ）」の 4 グループを定義した。この場合、ブロック数が増えても、グループ数は 4 のままである。このモデルは、真の θ の平均の違いを、忠実に表現したモデルである。 概要を図 5.3 に示した。

同時推定（偶数回グループ、奇数回グループ）Concurrent Calibration, Even and Odd (CCEO)

同時推定の手続きは「同時推定＋規準集団に等化」と同様とした。ただし、「学習前」と「学習後」の間、すなわち、偶数回に行われたテストと奇数回に行われたテストでは受験者の能力が異なる場面を想定し、「偶数回目の本試験での問題冊子を解いたグループ」、「奇数回目の本試験での冊子を解いたグループ」、「規準集団」の 3 グループを定義した。この方法では、定義されたグループ

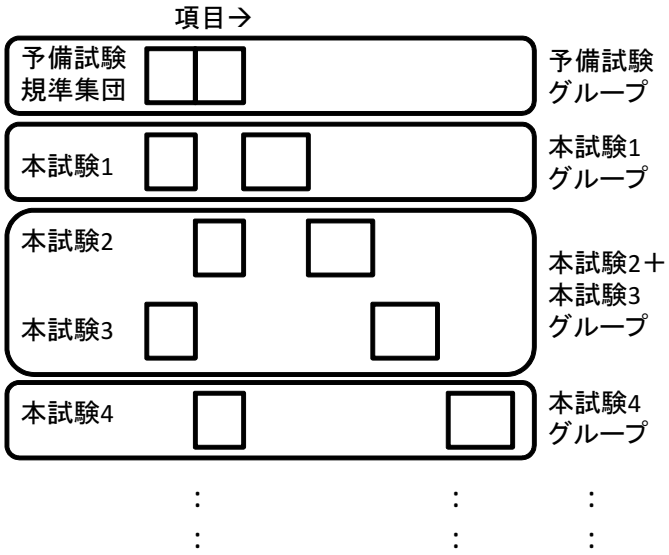

図 5.3　CCFT によるモデル簡素化

数は、実施回が多くなっても 3 のままである。 概要を図 5.4 に示した。

同時推定（近接した 2 回をまとめて 1 グループ）Concurrent Calibration, Adjacent two Forms (CCAF)

同時推定の手続きは「同時推定＋規準集団に等化」と同様とした。ただし、近接した回（本試験 1 と本試験 2、本試験 3 と本試験 4、以下同様）は一つのグループとして定義した。これとは別に、「規準集団」グループを定義した。この方法では、定義されたグループ数は (本試験回数 ÷2)+1 である。このモデルでは、学力が上昇する前と後の受験者を一つのグループ内に込みにしているものの、それぞれの年度ごとに学力差が見られるかに注目したモデルである。 概要を図 5.5 に示した。

同時推定（本試験をすべてまとめた 1 グループ） Concurrent Calibration, One Group (CCOG)

同時推定の手続きは「同時推定＋規準集団に等化」と同様とし、グループの定義を「本試験全体」「規準集団（予備調査の冊子を解いたグループ）」の 2 グループとした。この方法では、定義されたグループの数は、実施回が多くなっても 2 のままである。 概要を図 5.6 に示した。

5.2.4　シミュレーションの手続き

項目パラメタの真値を変化させた場合と、グループ間の真の θ の平均値差を変化させた場合との 2 通りについて、シミュレーションを行った。

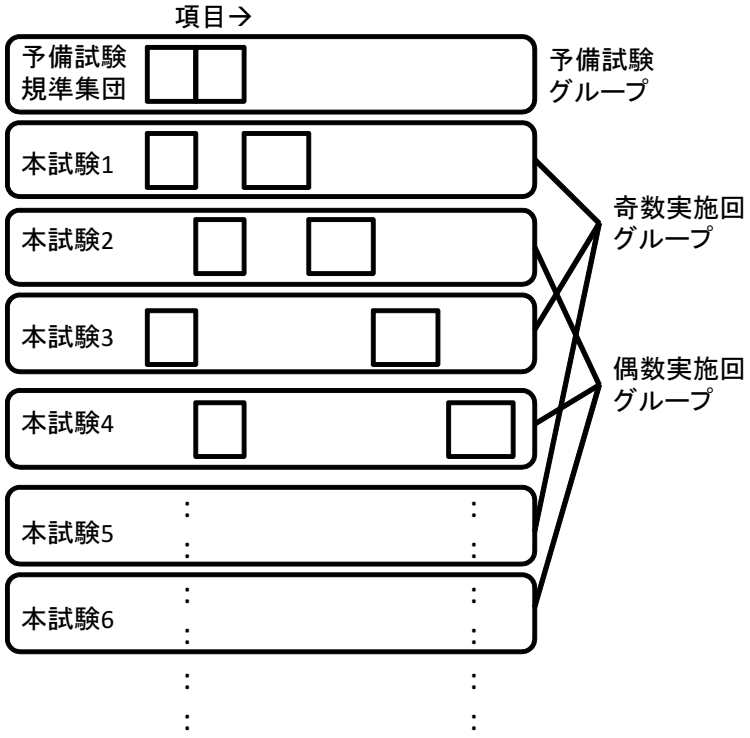

図 5.4　CCEO によるモデル簡素化

真の識別力と困難度を変化させた場合の検討（シミュレーション 1）

5.2.2 節に記したデータについて、予備試験の受験者数 4000 人、本試験 1 回あたり 8000 人として 0-1 データを発生させた。その際、予備試験において真の識別力 $\alpha = (0.5, 1.0)$ の 2 条件、本試験において真の識別力 $\delta = (0.5, 1.0)$ の 2 条件について発生させた。さらに、本試験において、冊子をまたいでユニークに付与される「新作項目」の項目数を $(10, 20, 40)$ の 3 条件を仮定し、本試験のブロック数を $(1, 2)$ の 2 条件（ブロック数 1 では本試験の冊子数 4、ブロック数 2 では冊子数 8）について仮定した。以上をクロスさせた 24 条件に関して、RESGEN4 を用いて 0-1 データを生成した。次に、個別推定、同時推定 4 通りのそれぞれのパラメタ推定手続きを行い、本試験の全項目について規準集団に等化した項目パラメタの推定値を得た。最後に、5.2.5 節に述べる従属変数を、それぞれの推定結果に対して計算し、記録した。以上の手続きを、各条件について 100 回繰り返した。すべての推定において、項目パラメタの推定には BILOG-MG 3 を用いた。

真のグループ間 θ を変化させた場合の検討（シミュレーション 2）

節で示した真の θ の平均値差 (0.25) を 0.0 から 1.0 まで 0.1 刻みで変化させた 11 通りについて、規準集団に等化した項目パラメタの推定値を得る手続きを前節と同様に 100 回行い、それぞれの推定方法について 5.2.5 節に述べる従属変数のうち「DICC」および「パラメタの種類ごとの RMSE」

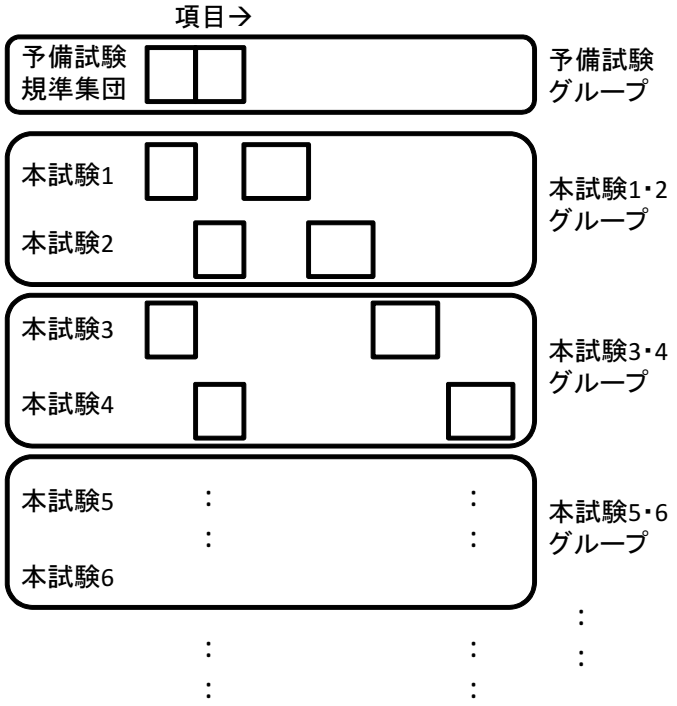

図 5.5　CCAF によるモデル簡素化

を計算し、記録した。ただし、α と δ、およびブロック数の条件については、 シミュレーション 1 の結果、推定方法間で最も従属変数の値が隔たった条件について行い、新作項目数は $(10, 20, 40)$ の 3 条件について行うこととした。また、各試験の受験者数は前節と同様とし、使用ソフトウェアも前節で述べたものと同様とした。

5.2.5　従属変数

等化された項目パラメタの真値からのずれ

それぞれの条件について、等化された項目パラメタの推定値が真の値からどの程度かけ離れているかの指標として、Arai and Mayekawa(2011) に従い、推定された全項目について ICC を算出し、その差を式 5.1 で定義される DICC(5.1 式) として定義した。この値が小さければ推定された項目パラメタと真の項目パラメタとの差が小さいことを意味する (2.2.3 節で述べた DICC は、試行の繰り返しを含まない形であったが、本研究の DICC は試行の繰り返しで平均をとっている点に注意)。

$$DICC = \frac{1}{L}\sum_{l=1}^{L}\frac{1}{J}\sum_{j=1}^{J}\frac{1}{Q}\sum_{q=1}^{Q}\left|P_j(\theta_q|\hat{a}_j^l, \hat{b}_j^l) - P_j(\theta_q|a_{jT}, b_{jT})\right| \tag{5.1}$$

ここで L は $1, 2, ..., l, ...L$ 回目の繰り返し試行を表し、J は $1, 2, ..., j, ...J$ 番目の項目、Q は $1, 2, ..., q, ...Q$ 番目の θ の求積点を表し、θ_q は -3 から 3 までのスケールを Q 等分した各値とし、

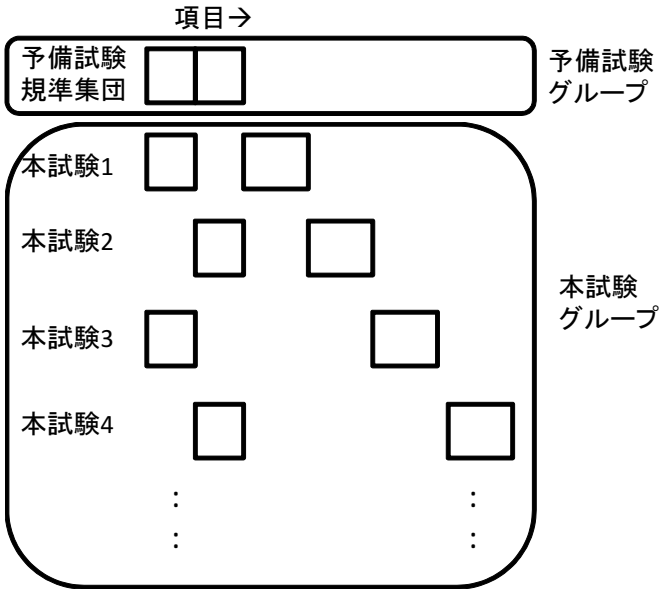

図 5.6　CCOG によるモデル簡素化

$Q = 31$ とした。また、a_{jT} および b_{jT} はそれぞれ j 番目の項目における識別力と困難度の真値を表し、$\hat{a}_j^l$ および $\hat{b}_j^l$ はそれぞれ j 番目の l 回目の推定における識別力と困難度の推定値を表す。

また、各項目に関して、識別力と困難度の真値と推定値との差を検討するため、式 5.2 の RMSE を識別力パラメタ、困難度パラメタそれぞれについて算出した。さらに、識別力と困難度のそれぞれについて、真値との差の方向性を検討するため、真値からの平均差（バイアス）を式 5.3 に基づいて算出した。

$$RMSE = \frac{1}{L}\sum_{l=1}^{L}\sqrt{\frac{1}{J}\sum_{j=1}^{J}(\hat{\xi}_j^l - \xi_{jT})^2} \tag{5.2}$$

$$BIAS = \frac{1}{L}\sum_{l=1}^{L}\left[\frac{1}{J}\sum_{j=1}^{J}(\hat{\xi}_j^l - \xi_{jT})\right] \tag{5.3}$$

ここで $\hat{\xi}_j^l = \hat{a}_j^l$ および $\xi_{jT} = a_{jT}$ とおいた場合を識別力の RMSE および BIAS、$\hat{\xi}_j^l = \hat{b}_j^l$ および $\xi_{jT} = b_{jT}$ とおいた場合を困難度の RMSE および BIAS と定義した。

同時推定の場合の尤度相対低下率

モデルのあてはまりがグループを減らした場合にどの程度変化するかを見るために、各条件の 100 回の試行に関し、式 5.4 に示す尤度相対低下率（R_Δ^2, de Ayala,2009,p.141）を算出した。

$$R_\Delta^2 = \sum_{l=1}^{L}\frac{-2\ln(L_R^l) - (-2\ln(L_F^l))}{-2\ln L_R^l} \tag{5.4}$$

ここで L_R^l は 1 冊子 1 グループ (CC) の l 回目の同時推定における最大化された尤度を表し、L_F^l はグループ数を減らした場合の l 回目の同時推定時における最大化された尤度を表す。本研究の場合、L_F^l は CCEO、CCFT、CCAF、CCOG の 4 通りに算出される。$L_R^l > L_F^l$ の場合、すなわち CC に比べて比較対象のモデルのあてはまりが悪い場合、R_Δ^2 は負の値をとり、 $L_R^l < L_F^l$ の場合は正の値をとる。

5.3 結果

シミュレーション 1 および 2 の両方で、すべての条件において、BILOG-MG による推定は収束基準 (0.005) を満たして収束した。また、真値から 2.0 以上かけ離れた推定結果を識別力または困難度のいずれかで示す項目もなかった。

5.3.1 シミュレーション 1 の結果

DICC に関して

条件ごとに DICC を算出したものを図 5.7（ブロック数 1)、図 5.8（ブロック数 2）にそれぞれ記した。いずれの条件においても、SC、CCFT、CC、CCEO、CCAF、CCOG の順に DICC が低い、すなわち真の項目パラメタに近い推定値が得られたという結果となった。ただし、$\alpha = 1.0$ の条件（図 5.7 および図 5.8 の左側 6 条件）においては、SC と CCFT が同程度で最も小さな DICC、次いで SC および CCFT に近い値で CC および CCEO が同じ程度の DICC となり、CCAF および CCOG の方法が同程度に大きい DICC となった。一方、$\alpha = 0.5$ の条件 (図 5.7 および図 5.8 の右側 6 条件) は、$\alpha = 1.0$ 条件で見られたような「SC および CCFT」「CC および CCEO」「CCAF および CCOG」が互いに近い値をとるという結果とはならず、SC、CCFT、CC、CCEO、CCAF、CCOG の順で真値に近いという結果が得られた。

新作項目数の大小でみると、全体的に新作項目数が増加することによって、SC および CCFT と「CCFT 以外の CC4 条件」(CC、CCEO、CCAF、CCOG の総称) との間に大きな DICC の差が見られるようになった。また、$\alpha = 1.0$ の場合よりも $\alpha = 0.5$ の場合が SC および CCFT と CCFT 以外の CC4 条件との間で DICC の差がより大きかった。さらに、$\alpha = 1.0$ の場合に見られた (CC、CCEO) が (CCAF、CCOG) よりも小さな DICC となったという傾向は、新作項目数が多くなると顕著ではなくなった。また、$\alpha = 0.5$ かつ $\delta = 1.0$ の条件においては、SC が CCFT よりも小さな DICC となる傾向が、特に新作項目数が多い条件において見られた。ブロック数の大小で比較すると、1 ブロック条件の方が 2 ブロック条件よりも DICC がわずかに減少した。α および δ の条件間でブロック数の効果による顕著な DICC の差は見られなかったものの、SC、CCFT、CCAF、CCOG においては DICC の減少の幅が小さく、CC および CCEO においては減少の幅が大きかった。

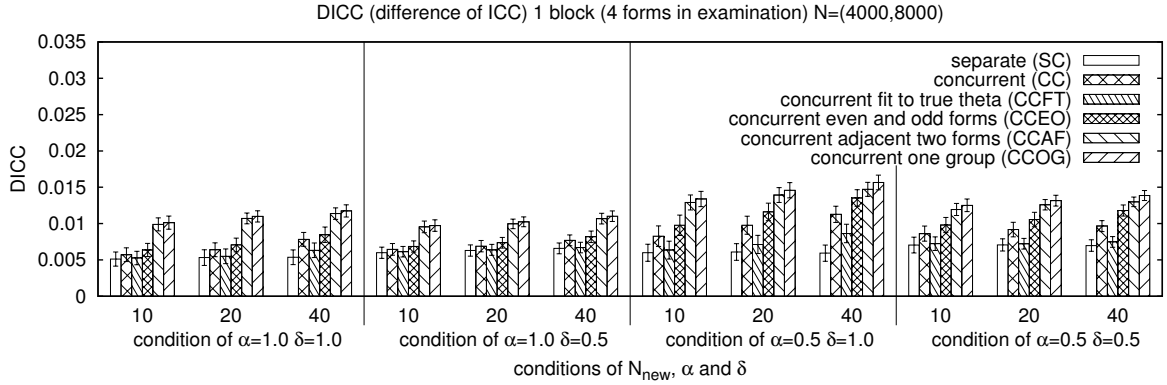

図 5.7　DICC（1 ブロック条件）。新作項目数 (10,20,40)、α(1.0,0.5)、δ(1.0,0.5) の違いごとに記した

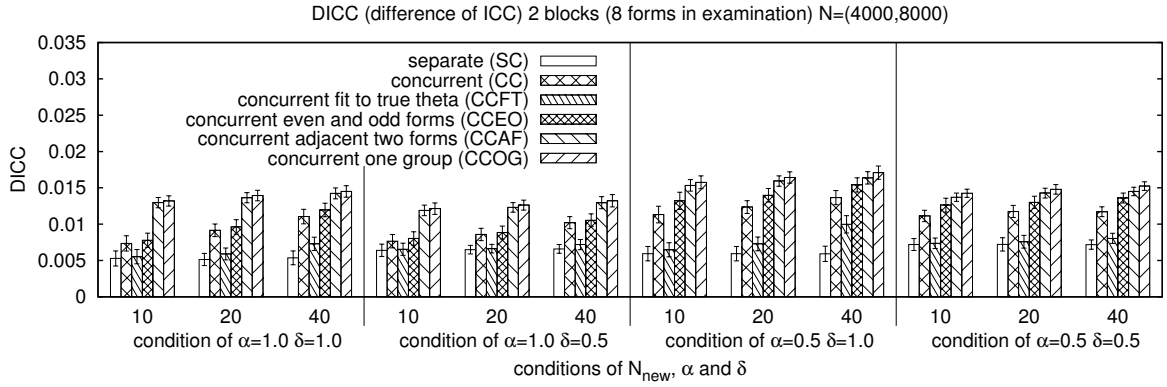

図 5.8　DICC（2 ブロック条件）。新作項目数 (10,20,40)、α(1.0,0.5)、δ(1.0,0.5) の違いごとに記した

項目パラメタの RMSE に関して

項目パラメタ（識別力、困難度）に関して、RMSE を図 5.11（1 ブロック条件の困難度について）、図 5.12（2 ブロック条件の困難度について）に記した。識別力の RMSE （1 ブロック条件を図 5.9、2 ブロック条件を図 5.10 に記した）を見ると、1 ブロック条件、2 ブロック条件のいずれでも、RMSE の値は 0.04 以下となり、いずれも小さな値になった。特に、$\delta = 0.5$ の条件において、$\delta = 1.0$ の条件の約半分の RMSE となった。ただし、新作項目数の多寡は、RMSE に影響を及ぼさなかった。RMSE が最大の条件は「新作項目数 10、ブロック数 2、$\alpha = 0.5$、$\delta = 1.0$」で、SC は 0.039、 CC は 0.039、 CCFT は 0.038、CCEO は 0.040、 CCAF は 0.039、 CCOG は 0.040 であった。

一方、困難度の RMSE については、ほぼ一貫して DICC の値と同様な傾向を得た。すなわち、全体的に SC および CCFT が最も RMSE の小さな結果であり、$\alpha = 1.0$ の条件においては CC と

CCEO が CCAF と CCOG よりも真値に近い困難度の推定値であった。SC や CCFT の RMSE が 0.05 前後に対して、CCFT 以外の CC4 条件においては RMSE が 0.1 前後で、ほぼ 2 倍の真値からの誤差が生じたことになる。

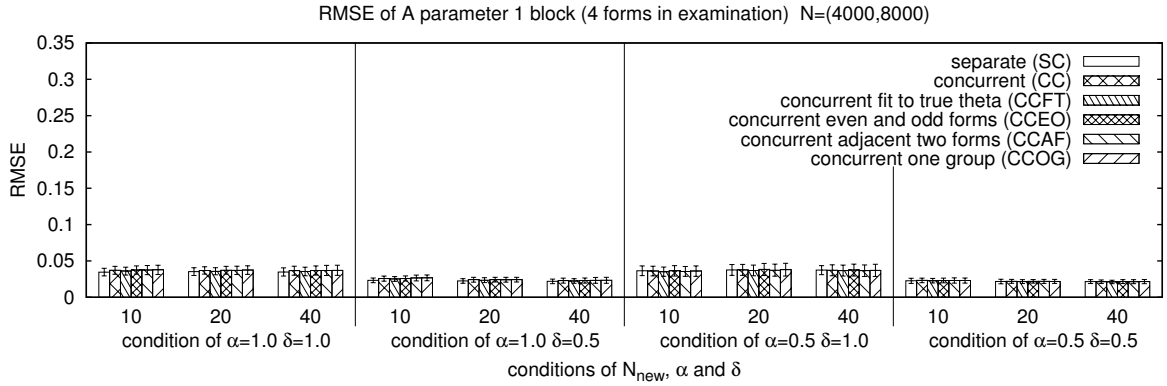

図 5.9　識別力の RMSE（1 ブロック条件）。新作項目数 (10,20,40)、α(1.0,0.5)、δ(1.0,0.5) の違いごとに記した

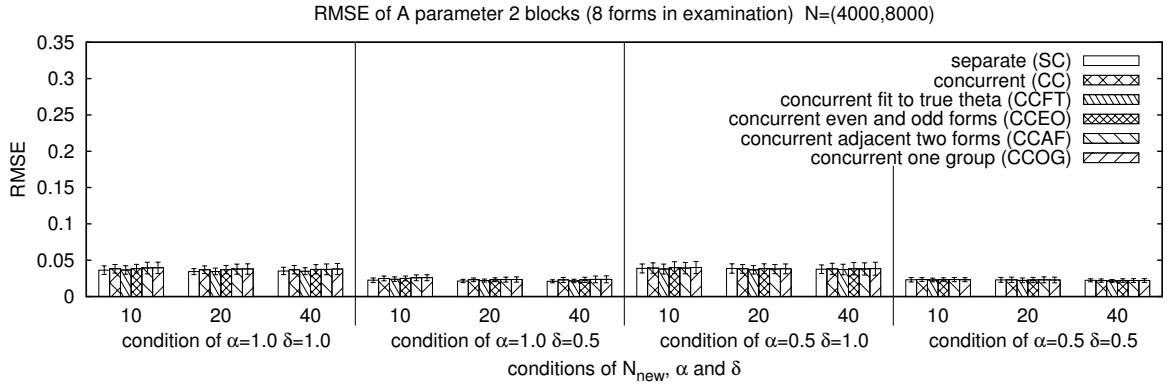

図 5.10　識別力の RMSE（2 ブロック条件）。新作項目数 (10,20,40)、α(1.0,0.5)、δ(1.0,0.5) の違いごとに記した

項目パラメタの BIAS に関して

項目パラメタの真値からの平均差について値を算出したところ、$\alpha = 0.5$ でかつ $\delta = 1.0$ の条件において、識別力の推定値で、いずれの新作項目数、ブロック数においても CC4 条件のみわずかながら一貫して過小推定をしていることがわかった（過小推定が最大の条件は「新作項目数 40、1 ブロック」で、SC は 0.000、 CC は -0.012、 CCFT は -0.013、CCEO は -0.014、 CCAF は -0.012、CCOG は -0.014）。ほかの条件においては、識別力の推定値に差の方向性は見られな

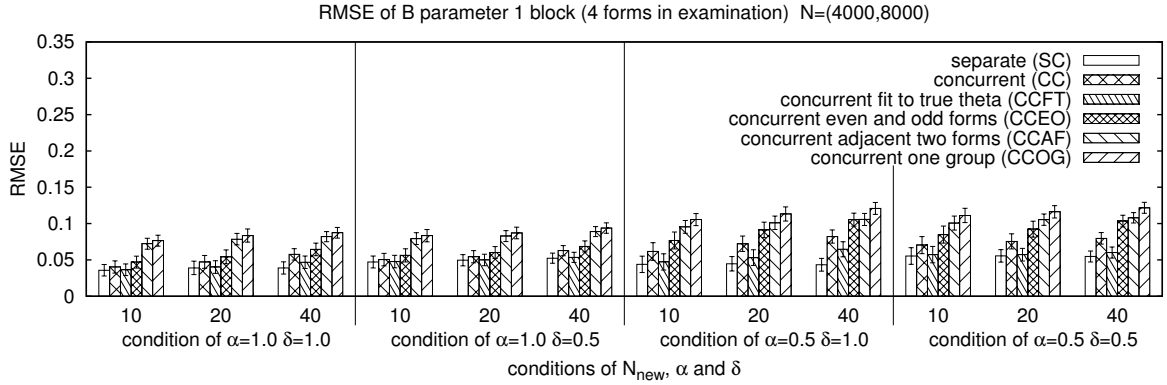

図 5.11　困難度の RMSE（1 ブロック条件）。新作項目数 (10,20,40)、α(1.0,0.5)、δ(1.0,0.5) の違いごとに記した

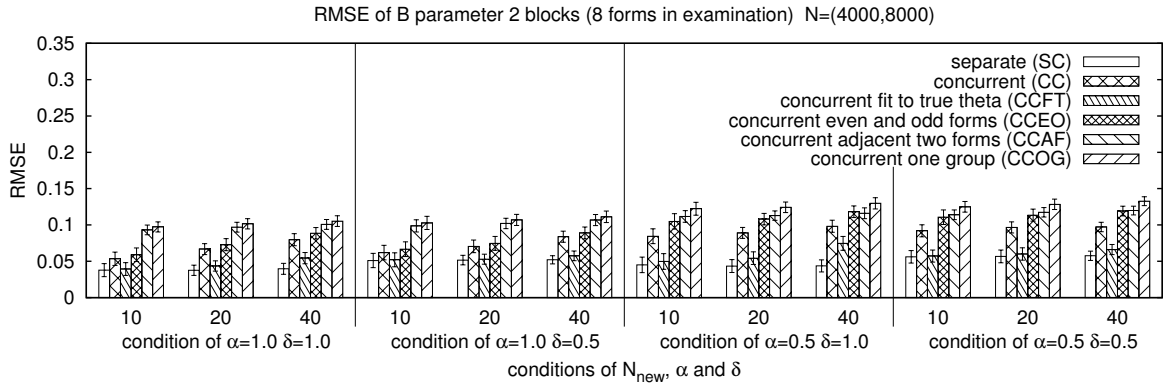

図 5.12　困難度の RMSE（2 ブロック条件）。新作項目数 (10,20,40)、α(1.0,0.5)、δ(1.0,0.5) の違いごとに記した

かった。また、困難度の推定値においては、すべての条件において真値からの平均差は ± 0.007 の範囲に収まっており、特定の条件において特異的な差の方向性は見られなかった。

尤度相対低下率 R^2_Δ に関して

各条件における R^2_Δ の値を、1 ブロック条件については図 5.13、2 ブロック条件の場合については図 5.14 に記した。1 ブロック条件、2 ブロック条件ともに値の大きさに大きな傾向の差は見られなかった。2 ブロック条件でみると、CCFT において、CC よりもわずかに尤度が高い条件が見られた。CC よりも尤度が低下した条件の場合、CCEO がいずれの α、δ、新作項目数の条件においても CC からの尤度の低下が最も少なく、以下 CCAF、CCOG の順であった。また、新作項目数が少ない条件ほど、CCAF および CCOG の推定法で CC からの尤度の低下が著しいという結果となった。さらに、全体的に見て $\alpha = 1.0$ 条件の方が、$\alpha = 0.5$ 条件に比べて尤度の低下が大き

かった。

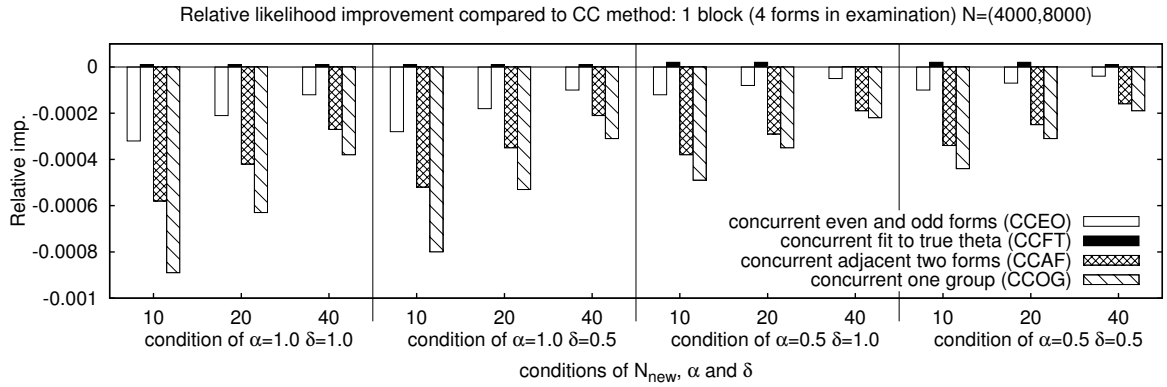

図 5.13　1 ブロック条件における尤度相対低下率。CC を基準として、CCFT、CCEO、CCAF、CCOG の 4 種の推定法でどの程度尤度が低下するかを示した。新作項目数 (10,20,40)、α(1.0,0.5)、δ(1.0,0.5) の違いごとに記した

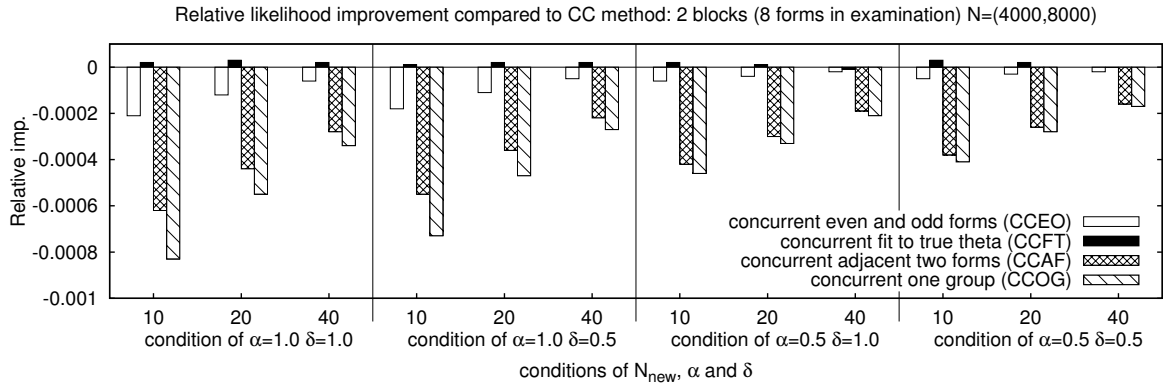

図 5.14　2 ブロック条件における尤度相対低下率。CC を基準として、CCFT、CCEO、CCAF、CCOG の 4 種の推定法でどの程度尤度が低下するかを示した。新作項目数 (10,20,40)、α(1.0,0.5)、δ(1.0,0.5) の違いごとに記した

5.3.2　シミュレーション 2 の結果

5.3.1 節の結果より、CCFT および CC4 条件の間でグループ数の差による違いが大きかった「2 ブロック条件」で、予備試験において識別力が高く、本試験において予備試験を上回る識別力となる項目が少なくなったテスト場面を想定し「$\alpha = 1.0$ かつ $\delta = 0.5$」条件について、シミュレーション 2 を行った。結果を以下に示す。

DICC に関して

新作項目数が 10 の場合の DICC を図 5.15 に、同じく 20 の場合を図 5.16 に、40 の場合を図 5.17 に示した。θ の上昇幅が 0.2 以下の条件に関しては推定方法の間で大きな差異が見られないという結果であった。しかし、θ の上昇幅が 0.3 を超える条件においては、SC が最も小さな DICC を示し、わずかな差で CCFT がそれに次いで小さな DICC を示した。また、CCFT 以外の CC4 条件でみると CC および CCEO が「SC および CCFT」に次いで小さな DICC となった。しかし、θ の上昇の度合いが大きくなるにつれて、わずかではあるが、CCEO の方が CC よりも大きな DICC となる傾向が見られた。一方、θ の上昇幅が 0.3 を超える場合、CCAF および CCOG において θ の上昇幅が大きくなるほど DICC が線形に増加する傾向が見られた。CCAF と CCOG の間で DICC の大きさに違いは見られなかった。新作項目数の違いでみると、SC、CCFT および CCAF、CCOG の場合は新作項目数の違いで大きな傾向の違いは見られなかったものの、CC および CCEO の場合は新作項目数が大きな条件でより大きな DICC の値となる傾向が見られた。たとえば、θ の上昇幅が 0.4 の場合でみると、新作項目数 10 の場合は SC と大きく変わらない値であるのに対し、新作項目数 40 の場合は大きく離れた DICC の値を示した。

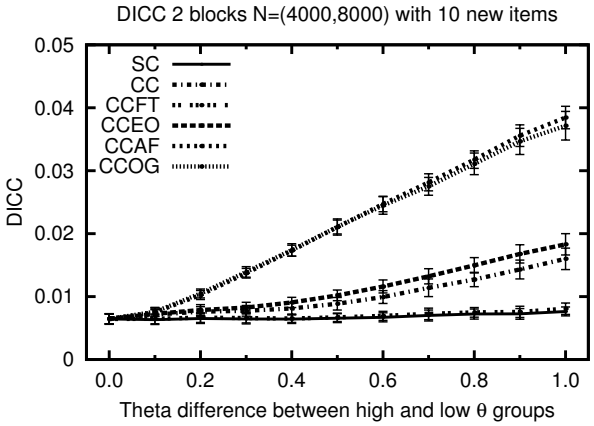

図 5.15　グループ間の真の平均 を 0.0 から 1.0 まで変化させた場合の DICC。新作項目数 10、2 ブロック条件で、$(\alpha, \delta) = (1.0, 0.5)$ の場合を記した

項目パラメタの RMSE に関して

項目パラメタの種類別に真値との差を示した図を図 5.18（新作項目数 10）、図 5.19（新作項目数 20）、図 5.20（新作項目数 40）にそれぞれ示した。識別力の推定値に関しては、SC および CCFT においてはどの θ の上昇幅においてもほぼ一定の RMSE であったのに対し、新作項目数が 10 および 20 の条件において、CCFT 以外の CC4 条件では RMSE は θ の上昇幅が大きくなるにつれて増大する傾向が見られた。ただし、CCFT 以外の CC4 条件間で増大の傾向に差は見られなかった（新作項目数 20 の条件において θ の上昇幅が 0.7 を上回る場合に、CCAF および CCOG がより

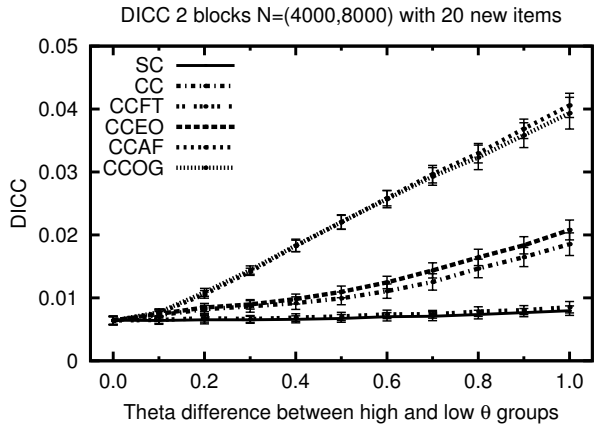

図 5.16　グループ間の真の平均 を 0.0 から 1.0 まで変化させた場合の DICC。新作項目数 20、2 ブロック条件で、$(\alpha, \delta) = (1.0, 0.5)$ の場合を記した

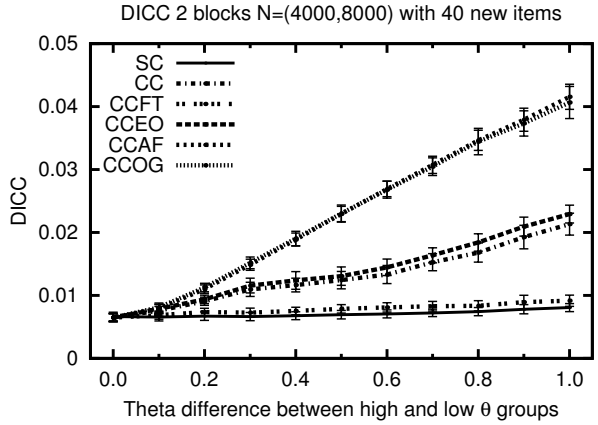

図 5.17　グループ間の真の平均 を 0.0 から 1.0 まで変化させた場合の DICC。新作項目数 40、2 ブロック条件で、$(\alpha, \delta) = (1.0, 0.5)$ の場合を記した

真値に近い結果を示した)。それに対し、新作項目数 40 の条件においては、CCAF および CCOG が、CC および CCEO に比べて RMSE が小さくなる傾向が見られたが、θ の上昇幅が 0.5 以下の場合は、CCFT 以外の CC4 条件の間で識別力の RMSE に違いは見られなかった。ただし、いずれの推定結果においても、CCFT 以外の CC4 条件においては、最大の θ の上昇幅である 1.0 の場合においても、0.08 を下回る RMSE であった。

困難度の推定値に関する RMSE についてみると、いずれの新作項目数の場合においても SC および CCFT が最も小さな RMSE となり、以下、CC、CCEO、CCAF、CCOG の順となった。新作項目数 10、および 20 の条件では θ の上昇幅が 0.3 以上 0.7 以下の場合、SC、CCFT、CC および CCEO が RMSE の大きさで 0.1 以内に収まる一方で、CCAF および CCOG では RMSE の大きさが 0.1 を上回る結果となった。一方、新作項目数 40 の場合、SC および CCFT ではどの θ の

上昇幅においても RMSE の値は 0.1 を下回るのに対し、CC および CCEO においては θ の上昇幅が 0.3 から 0.7 であっても RMSE の値が 0.1 に近い値を示した。しかし、CCAF および CCOG の結果は、それらを上回る RMSE であった。

これらの結果をまとめると、困難度の推定値において、新作項目数が 10 または 20 の場合で、かつ、θ の上昇幅が小さい場合は、SC のかわりに CCFT や CC、CCEO を用いても、困難度の推定値が真値と大きくかけ離れることはないものの、CCAF や CCOG の方法を用いると真値と大きくかけ離れること、また、新作項目数が 40 の場合や、少ない新作項目数であっても θ の上昇幅が大きい場合は、「SC および CCFT」の結果が最も真値に近い推定値となり、次いで「CC および CCEO」の結果となることがわかった。また、識別力の推定に際しては、SC が最も小さい RMSE であったものの、CC5 条件で最大でも 0.07 前後の RMSE と、いずれの方法でも推定の結果としては十分小さな RMSE となることがわかった。

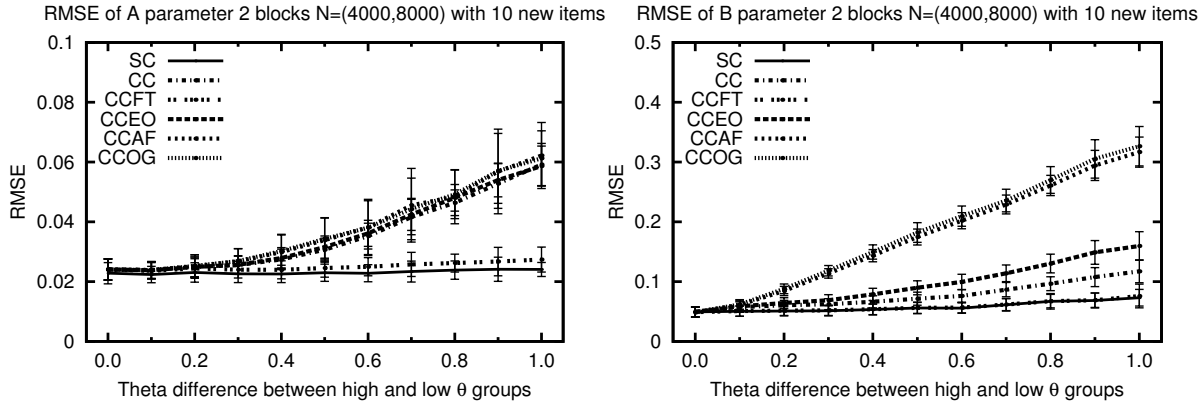

図 5.18　グループ間の真の平均 θ を 0.0 から 1.0 まで変化させた場合の識別力（左）、困難度（右）の推定値の RMSE。2 ブロック条件で、$(\alpha, \delta) = (1.0, 0.5)$、新作項目数 10 の条件において、推定方法別に示した

5.4　考察

5.4.1　モデル簡略化の効果とテストの一次元性

本研究における「予備試験で項目パラメタ既知の項目グループを用意し、本試験において予備試験と共通のアンカー項目を出題する」という等化場面で、(CC および CCFT) と (CCEO、CCAF、CCOG) の間で DICC および項目パラメタの推定値の RMSE に差が見られた。SC の方が「CCFT 以外の CC4 条件」、特に CC よりも DICC が小さかったという結果は、Arai and Mayekawa (2011)、Hanson and Béguin (2002) で見られた結果と同様であったと考える。項目パラメタ別に見ると、識別力の推定値には新作項目数の影響が少なく、困難度の推定値にはその影響が表れている。また、識別力の推定値の BIAS が、α=0.5 で δ=1.0 の場合にわずかながら過小に推定されるという結果となった。理由として、本研究の手続きで、RESGEN4 に「すべての冊子

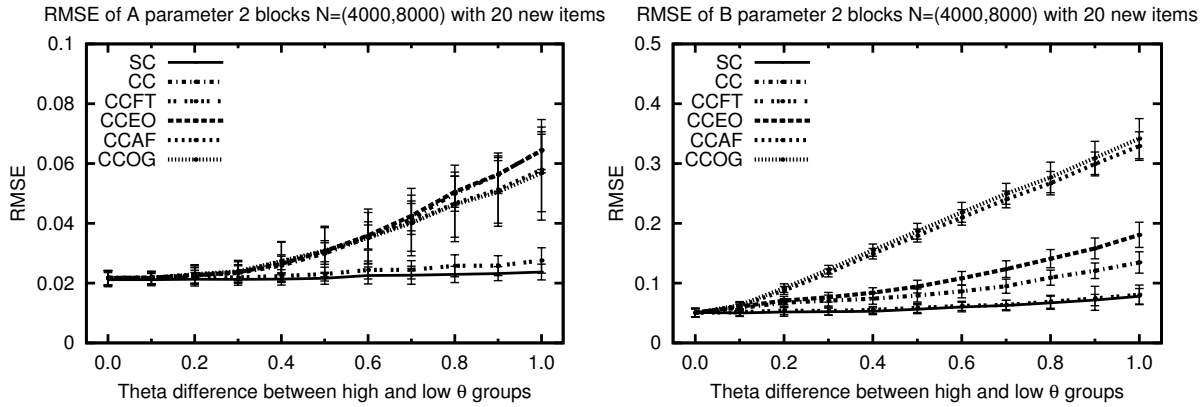

図 5.19　グループ間の真の平均 θ を 0.0 から 1.0 まで変化させた場合の識別力（左）、困難度（右）の推定値の RMSE。2 ブロック条件で、$(\alpha, \delta) = (1.0, 0.5)$、新作項目数 20 の条件において、推定方法別に示した

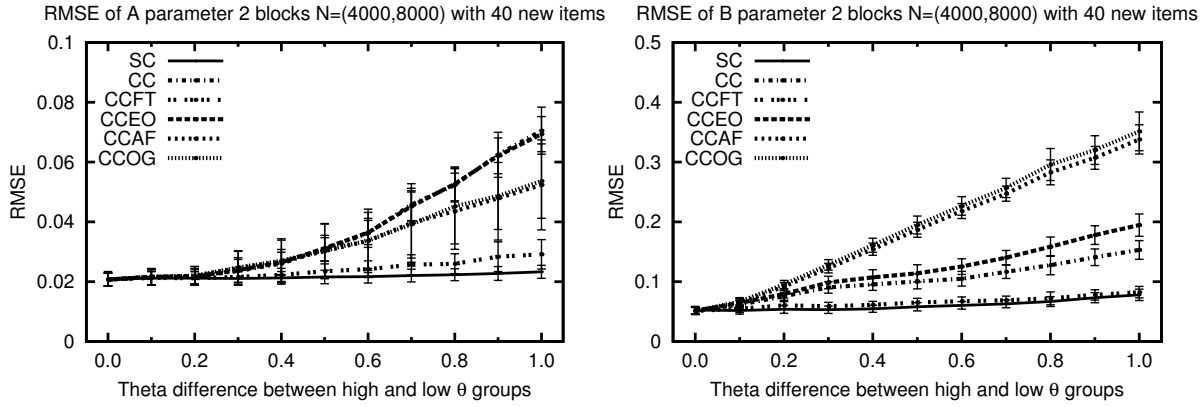

図 5.20　グループ間の真の平均 θ を 0.0 から 1.0 まで変化させた場合の識別力（左）、困難度（右）の推定値の RMSE。2 ブロック条件で、$(\alpha, \delta) = (1.0, 0.5)$、新作項目数 40 の条件において、推定方法別に示した

をまたいで一次元になるような」乱数を発生させた点が挙げられる。項目パラメタの推定を因子分析の一種と考えると、識別力の推定は因子負荷の推定と等価 (村木、2011、p.44；尾崎、2003、p.212) であり、元の 0-1 データにおいて点双列相関係数が高い項目グループにおいては識別力が高い。したがって、識別力の推定に影響するのは、元の 0-1 データにおける因子の一次元性が主であり、本研究ではそれを手続き上で統制しているので、識別力の RMSE の推定結果に条件間の差が少なかったと考える。また、「予備試験の真の識別力が低く、かつ、本試験の真の識別力が高い」条件において、CC の条件（CCFT を含む 5 条件）で識別力が過小に推定されたのは、SC において「冊子ごとに 1 因子を抽出する」場合は、「全体をまたいで 1 因子を抽出する」CC5 条件に比べ、よりモデルのあてはまりの良い推定結果を返したからであると考える。ただし、SC は、各回個別に項目パラメタを推定し、共通項目のパラメタを手掛かりに等化するという 2 段階推定を行っているため、毎回の試験をまたいだ因子の一次元性が仮定できない場合、モデル上で一次元性を表現す

ることが極めて困難である。それに対し、CC5 条件は、各回の試験で一貫した共通の因子を仮定しない場合であっても、すべての回のデータを同時に推定した結果がパラメタ推定に反映される。したがって、一次元性が保証できない場合の推定結果の妥当性は、SC に比べて高いといえる。仮に、ある単一次元の能力を問うテストを本研究で取り上げたデザインで 10 回にわたって実施した場合、そのうちの 1 回（冊子 X と表記）において、出題者の意図と異なり、新作項目が他の 9 回とは異なる次元を測定していたとする。CC5 条件では、冊子 X において識別力が低く推定されるはずである。なぜなら、他の 9 回の試験で問われている単一次元で見た場合、冊子 X に含まれる新作項目は因子負荷が低いといえるからである。一方 SC では、冊子 X の正誤反応のみで識別力を推定し共通尺度上に等化した結果は、CC5 条件に比べて高い値となることが予測できる。なぜなら、冊子 X の新作項目のみで見た場合、他の 9 回とは異なる別の 1 因子を測定しているということで識別力が高く推定され、この推定された識別力が他の 9 回の試験における共通尺度上に等化されるためである。この場合、実際のデータの特徴をより正確にとらえるという意味で妥当な推定結果は、「冊子 X が他の 9 回の試験と比べて識別力が低く、一次元性が満たされていない可能性が示される」というものであり、CC5 条件の方法によるものであると指摘できる。このような場合、本質的には多次元 IRT の適用が望ましいものの、実践的なテスト場面での多次元 IRT の適用は、結果の解釈や推定方法の確立などの諸問題の解決を待たなければならないだろう。

5.4.2　簡略化によるモデルのあてはまりの良さへの影響

また、尤度相対低下率の比較から、グループ数を減らした 4 方法の中では、CCFT の方法において尤度が CC とほぼ同等であるという結果であった。この点から、CC はそれぞれのグループにおける θ の分布の特徴を詳しく記述できるモデルであるが、本研究の目的である「より真値に近い項目パラメタの推定値」を必ずしも返さないモデルであることが指摘できる。CCFT 以外の場合でみると、減少幅は、新作項目数を増やすことによってより小さくなったことから、モデル上の尤度は新作項目数に依存する関係が見て取れる。しかし、項目パラメタの推定値、特に困難度の推定結果は、新作項目数を増やしても CC と CCEO の間で RMSE の大きさにおいて差は縮小しなかった。このことは、モデルのあてはまりと推定結果は必ずしも対応しないことが指摘できる。

5.4.3　実際のテスト場面における簡略化の方針

DICC は、推定方法、α、δ、新作項目数の違いによって異なる結果となった。この違いは、主に困難度の RMSE によって生じたものと推測できる。特に、CC と、それよりグループ数を減らした CCFT、CCEO、CCAF、CCOG の推定結果を比較すると、グループ数が 3 しかない CCEO が、CCAF よりも CC に近い尤度を得た。それに伴い、CCEO が CCAF や CCOG に比較してより真値に近い困難度の推定値を返した。さらに、グループ数が 4 の CCFT を適用すると、CC とほとんど変わらない尤度となり、項目パラメタの推定結果も CC より真値に近くなった。このことは、テストを実施する側の立場に立てば、仮にグループ数を減らす必要に迫られた場合、似たよう

な特性の受験者が含まれるグループをひとまとめにし、グループ間で θ の分布の違いが明確になるようなグループ分けにするという方法が望ましいことを示している。実際、2 ブロック条件において、CC ではグループの数が 9 であるのに対し、CCFT は 4、CCEO は 3、CCAF は 5 であり、CCAF は CCEO よりも真値からはずれた推定結果を返した。また、CCFT および CCEO では、ブロック数が増加してもグループ数は増えない。また、テストデザイン上、CCEO においては同一のグループ内に同一のアンカー項目が含まれるという関係にあった一方で、CCFT は同一のグループ内に異なるアンカー項目を含むデザインであった。両者を比較すると、真値に近い項目パラメタを得るためには、必ずしも同一の共通項目を同一のグループに提示している必要はなく、むしろ θ の分布が似ている集団を同一グループと定義した場合に真値に近い項目パラメタの推定値が得られる、ということが指摘できる。

項目パラメタの種類別に見ると、識別力の推定値はグループ数を変えたことにより大きな影響を受けなかった一方で、困難度の推定値は推定方法によって RMSE の値に差が生じた。仮に、本研究で得られた全項目を「項目パラメタ既知の項目」として項目バンクに入れて CBT (computer based test) などで学力の判断に使用する場合、等化元の項目パラメタに誤差が生じたことによる影響は、実践上の意味においては識別力よりも困難度の方が大きいと考える。なぜなら、困難度の推定値は等化元の受験者集団を規準集団としたときの値であり、等化先の受験者の能力を直接定義するのは困難度の値だからである。ただし、4 章での結果より、新作項目数が 10 から 20 前後で、真の θ の上昇幅が 0.7 以内に収まっている場合は、CC や CCFT、CCEO の方法を用いても、得られた推定値に大きな影響がないことが予測できる。仮に、SC の方法が何らかの理由により適用することが不適切な場合（たとえば、毎回の試験で同一の一次元を測定している前提の試験において、測定内容が実施回によって異なり、一次元性が保証されない場合など）は、CC の方法を適用する代わりに、本研究で取り上げた CCFT や CCEO の方法のように、互いに似た θ の分布、もしくは共通の背景を持つ受験者グループをひとまとめにして新たにグループを定義し、同時推定を行うことで、「1 冊子 1 グループ」の代替とすることができる、というのが本論の結論である。また、テスト場面において真の θ の上昇幅がどれほどかが不明な場合は、SC の方法を適用して各問題冊子に対応するグループごとに等化後の θ の分布を推定し、それを参考に θ の平均が似ているグループをまとめるようにすれば良い。

5.4.4 新作項目割合が大きな場合のモデル簡素化の影響

モデル簡素化の影響は、新作項目数が大きくなる場合、すなわち、新作項目割合が大きな場合に、より大きく表れるという結果が得られた。しかし、グループ間の真の θ を変えた場合のシミュレーション結果より、比較的小さな θ の差の場合においては、新作項目割合が大きい（アンカー項目数 10 に対して新作項目数 40 の条件）であっても、CC、CCFT と CCEO においては項目パラメタの推定結果に差は見られなかった。したがって、これら 3 種のモデル簡素化は、新作項目割合が大きな場面においても有効であると考える。このことは、個別推定しか不可能であったような多群の場合であっても、同時推定が可能となることを示している。

5.4.5　今後の課題

本研究においては、同じ 0-1 データに同じ 2PL モデルを適用した場合であっても、グループの定義を変えたことによって項目パラメタの推定値に影響が及んでいる。理由としては、以下の点が挙げられる。多群 IRT モデルのパラメタ推定においては「項目パラメタ」と「各グループにおける θ の分布を規定するパラメタ」の 2 種のパラメタに分けて考えることができ、その尤度関数は、項目パラメタの尤度と θ の尤度の積となっている（前川、1991、p.113)。したがって、「互いに異なる θ の分布を持つグループ」を同一グループとして定義した場合、θ の分布の推定が真値からかけ離れた結果となり、そのかけ離れた推定結果をもとにして項目パラメタを推定したため、項目パラメタの推定に影響が及んだものと考える。一方で、SEM(Structural Equation Model) において 2 値のカテゴリカル順序尺度による因子分析モデルと同等であることを利用した等化方法が尾崎 (2003, pp.210-220) で提案されている。この方法は、SEM のソフトウエアを用いた方法で、大規模データセットの場合であっても適用可能であるならば、実践的である可能性がある。この方法との比較は、今後の課題である。

本研究においては、現実のテスト場面で起こりうるデザインとして、予備試験であらかじめ項目パラメタ既知の項目を用意し、本試験でそれらの項目を手掛かりに予備試験の規準集団に等化する場面を取り上げた。もちろん、現実のテストのすべてが、本研究と同様のデザインを適用するわけではない。たとえば、項目パラメタ既知の項目と新作項目を毎試験において受験者に提示して新作項目のパラメタを推定し、試験のつど新作項目を規準集団に等化していくデザインなども考えうる。このような場合に対しても、本研究で見出された同時推定のグループ分けとパラメタの推定精度との関係を検討する必要があるであろう。

第 6 章

考察・総論

前章までにおいて、本試験実施前に大きな項目バンクを用意できない場合のテスト実施法においてどのような方法で等化を行えばよいかを、シミュレーション研究を通じて検討してきた。本章では、これまで検討された結果から、実際にテストを実施する際にどのような等化方法をとればよいか、その手続きを述べ、さらにテスト実施上の問題点、特に尺度の一次元性の問題について述べる。最後に、本研究から見える等化の研究上の課題について述べ、今後の研究がどのように発展するか、その展望を述べる。

6.1　本研究で取り上げたテスト実施法における等化

本論で提案した項目バンクの中身が増大するようなテスト実施法においては、毎回の試験終了後、受験者に成績を返すまでの間に、当該実施回のフォームの項目パラメタより各受験者の能力を返す必要がある。また、項目バンクの中から、将来のアンカー項目が提示される。そのため、毎回の試験の後に当該実施回のフォームのアンカー項目および新作項目のパラメタを規準集団に等化する。その際、本研究の各章で見出された知見から、より真値に近いパラメタを返す等化方法を提案する。

6.1.1　個別推定の方法

第 2 章では、個別推定の順序性が等化後の項目パラメタの推定値に及ぼす影響を検討した。この結果、Mean/Sigma 法は ICC 法 (Stocking-Lord 法、Haebara 法、calr 法) に比べて等化の順序によって等化後のパラメタの推定値が真値と異なる傾向が示された。また、calr 法に関しては、他の ICC 法に比較し、一度に複数のフォームに共通して現れるアンカー項目の項目パラメタを手掛かりに、複数のフォームを規準集団上の尺度に乗せることが可能である。また calr 法は、真値からのずれに関しても、ICC 法と同等か、より真値に近い結果を示す方法であった。したがって、個別推定の場合は、calr 法が使えるのであれば calr 法を、使用できない場合は ICC 法を使用するのがよいというのが結論である。ただし、ICC 法でペアワイズに等化する場合、項目バンクにアンカー項目

のパラメタを入れる際のルールは、より真値に近い傾向を示したことから、等化の結果得られた等化後のパラメタと、既存の項目バンクのパラメタの平均をとるのがよいと考える。

calr 法は、基本原理は前川 (1991) に記述されており、詳細な計算方法は Arai and Mayekawa (2011) の Appendix に記されている。しかし、実際の計算は、代表的な等化係数推定プログラムである R の `plink` パッケージや Kolen and Brennan(2004, Appendix B) に記載されている `POLYST` や `POLYEQUATE` などのソフトウェア[*1]には導入されていない。そのため、いまだ一般的ではないというのが現状である。また、calr 法に類似するような方法により、グループごとの等化係数やグループに共通の等化済み項目パラメタを推定するモデル研究は、あまり注目されているとは言えず、今後さらなる発展が望まれる。

また、1.5.2 節で述べたような「1 回の試験において複数フォームを提示する」場合、個別推定において、同一の本試験に出題された複数のフォームについて等化の順序性の影響が生じる。このため、等化の原理的に順序性の影響が生じない、calr 法を用いるのが望ましいと考える。

6.1.2 同時推定の方法

第 3 章では、同時推定の方法として、「同時推定＋規準集団に等化」と「同時推定 MG」の 2 種を比較した。この 2 種の間において、より真値に近い結果は「同時推定＋規準集団に等化」であった。本研究で示したテストデザインにおいては、全受験者における 0-1 データに対して一括して項目パラメタを推定し、規準集団にあらためて等化する方法が有効であることが示唆される。

6.1.3 大規模な 0-1 データの同時推定

同時推定（「同時推定＋規準集団に等化」）において、BILOG-MG を使用する場合、グループの数が増えると、メモリが不足してパラメタ推定が不可能になる。一方、グループの数を減らせば、メモリの使用量は少なくて済み、現実に計算可能な計算量となる（第 4 章）。グループの数を減らした場合、パラメタ推定値にどのような影響が見られるかを第 5 章で検討した。結果、各フォームに対応する θ の分布の形状（平均、標準偏差）が互いに似ているグループを一つにまとめるという操作をした場合、「1 フォームが 1 グループに対応する」と仮定した分析結果と比べて、推定結果に大きな差は見られないことがわかった。実際には、個別推定によって各グループの θ の平均を推定し、互いに似た θ の平均を持つグループをまとめるようにしたり、あらかじめ事前に母集団における θ の平均が似ている（同一の母集団からサンプリングされたと考える）と仮定されるグループをまとめる、などの方法で、グループをまとめたモデルに対して同時推定を行えばよいと考える。

[*1] `http://www.education.uiowa.edu/centers/casma/computer-programs.aspx` で入手可能

6.1.4　新作項目割合が大きな場合における推定

第 2 章の結果より、新作項目割合の大きな個別推定場面では、Mean/Sigma 法以外の方法を用いることにより、実用上問題のない推定の安定性を得ることができることが分かった。また、第 3 章の結果より、同時推定場面においては、新作項目割合が大きい場合に、「本試験の識別力が高く、かつ、予備試験の識別力が低い」状況で等化を行うと、項目バンク上で識別力の過小推定が起こることが分かった。しかし、第 5 章の結果は、新作項目割合が大きい場合、同時推定場面でグループの数が多くなったとしても、グループのまとめ方によっては「1 フォーム 1 グループ」の場合と同等な項目パラメタを得ることが可能であることを示している。以上の結果をまとめると、新作項目割合が大きい場合は、同時推定場面における過小推定の可能性について、留意しなければならないことが分かる。過小推定の条件は、予備試験と本試験において異なる次元の構成概念を測定しているが、各試験のフォーム内では単一の次元を測定している場合であることが示唆される。

6.1.5　同時推定か個別推定かの選択における一次元性の検討の必要性

第 3 章では、等化を繰り返すテストデザインにおいて、個別推定と同時推定のいずれが真値に近いパラメタとなるかを、等化後の項目バンクにおける項目パラメタの数値について比較した。結果、同時推定においては、予備試験の識別力が低く、本試験の識別力が高い条件において、識別力の推定値が真値より低く推定される傾向が見られた。一方、個別推定においては、そのようなパラメタの推定値の偏りは見られなかった。予備試験と本試験で識別力の推定値が変わらない、もしくは予備試験の識別力より本試験の識別力が高い傾向が一貫してみられるならば、個別推定の方がより真値に近い推定結果を返すので、より好ましい推定方法といえる。しかし、「予備試験の識別力が本試験の識別力よりも小さくなる」ということは、予備試験と本試験との間で一次元性が保証されない場合といえ、等化の前提を満たしていない可能性がある。したがって、常に個別推定の方が真値に近いからと言って、個別推定が最良の方法であるとは言い切れない側面があるといえ、このような場合は同時推定による項目パラメタの等化が適しているといえる。特に、一次元性に関しては、後に節を改めて議論する。

実践場面における手続きとして、一次元性が毎回の試験において満たされているかを調べるためには、本試験実施後に行われる等化に先立ち、一次元のカテゴリカル因子分析モデルによる一因子性の検討を行うようにすれば良い。一因子性の検討の結果、一次元性が仮定できるうちは、個別推定を行い、一次元性が仮定できなくなるような本試験の新作問題を出題したことがテスト実施後に判明した場合は、同時推定に切り替える、というような操作が必要となる。ただし、今後、アンカー項目を出題する際、一次元が仮定できなかった回の新作項目は、再出題せず、個別推定を続けられるようなテストデザインとすることが望まれる。

6.2　テスト実践場面での応用可能性

6.2.1　個別推定における項目バンクの更新

本研究のテスト実施法に限らず、個別推定の場合の項目バンク更新法は、実際のテスト場面において大きな問題点とされてこなかった場合が多い。なぜなら、等化方法の検討を行うまでもなく、「もっともらしい結果」を返すことができれば、その方法を使い続けるのがテスト実施機関にとって好都合だからである。言い換えれば、テスト実施機関は実務的な観点から「等化後の項目バンクが何らかの事情で得られない」場合を最も恐れている。したがって、それらしい結果が毎回の試験の中で得られさえすれば、それが良い方法であると判断され続ける傾向にあるといえる。

本研究の第 2 章の知見より、本研究で取り上げたテスト実施法で見られるような、複数のフォームを規準集団に等化する際には、Mean/Sigma 法は不適であることが分かった。しかしながら、多くのテスト場面において、Mean/Sigma 法は用いられることが多い。理由としては、(1) 反復計算を伴わず、Excel 等の表計算ソフトウェアでも簡単に計算が可能なこと、(2) 反復計算がないため、解の発散などにより解が求まらない場合がありえないこと、(3)ICC 法で設定される θ の求積点の数、および θ の上限と下限の値を恣意的に決めることが、テスト実施機関によっては「外部に説明できない事項である」等の理由で許されないこと、などが挙げられる。ここで (3) は、経験的な知見の集積によって、テストの分析に携わる者の間で求積点などの値に一定のコンセンサスが得られることや、シミュレーション研究などで、これらの値を操作することによって等化後の項目パラメタの推定値がどの程度変わるかが分かれば、Mean/Sigma 法を採用せず ICC 法を採用する根拠になりうる。しかし (1) や (2) に関しては、代替案を提示することが難しい場面が想定できる。

しかしながら、本研究の知見より、多少の困難を伴ったとしても、IRT に基づく等化を行いながら項目バンクを構築するようなテストにおいては、ICC 法を用いるべきであることは論を俟たないであろう。特に calr 法によって、最も理論的にかなった等化が行われることが示されたことから、今後 calr 法がテスト実務家にとって手軽に実行可能になることが必要であると指摘できる。

6.2.2　項目バンクをより急速に増やすための新作項目数の検討

本研究で検討したテスト実施法において、新作項目の数がアンカー項目と大きく異ならない場合は、同時推定でも個別推定でもほぼ同様な結果となった。しかし、本研究の目的は、毎回のテスト実施後に、項目パラメタが既知の項目が実施前よりも多くなっていることが前提であり、その度合いが大きいほど、テスト実施場面においてより望ましいといえる。一般的にアンカー項目として抽出され、再利用される項目は、項目パラメタの値、特に識別力が良い項目であることが多い。しかし、識別力の大きな項目が安定して多く得られる場合はテスト実践場面においては少ない。よって、識別力が高い項目が多く現れることを期待し、本試験において多くの新作項目を出題する試みがなされることが合理的である。しかしながら、本試験の新作項目数を多くすることや、アンカー項目数を多くすることによって、受験者に過度の負担をかけるような項目数となることは避けなけ

ればならない。また、新作項目数は、作題および問題冊子の編集作業に割くことのできる労力の大きさによっても、制約を受けるだろう。

また、実際のテスト場面では、出題分野、あるいはトピックが問題冊子間で均一になるように工夫することが多い。この場合、識別力が高いためにアンカー項目に選ばれた項目に特定の分野を問う内容が集中すると、それ以外の分野が新作項目として出題されることとなる。このようなアンバランスが重なると、識別力が全体的に良好な分野とそうでない分野が出てくる。テスト実施機関としては、どの分野においても識別力の高い項目が、どの困難度の値においても複数問そろっているのが理想であろう。したがって、受験者の負担にならない範囲、あるいは作題体制の制約の範囲の中で、テスト実施機関が本試験の新作項目数を大きく設定する場面が多いことが予想される。

本研究からの知見より、項目バンクのサイズが大きくなった場合に、個別推定の方が項目パラメタの値が真値に近い結果を得たということは、本研究のテスト実施法には個別推定が適している可能性を示唆した。しかし、前提として、測定している概念が一次元であることが、冊子間で求められていることも指摘した。これらの結論を実際のテスト場面に今すぐに当てはめることが難しい場合（たとえば、既存の等化方法を変えられない場合、あるいは等化作業を第三者に委託してきたために、テスト実施機関にとって過去に行った等化法が不明なまま項目バンクを更新してきた場合などが考えうる）もあるものの、今後本研究で取り上げたテスト実施法による項目バンクの構築においては、本研究の知見を反映した等化法とすることが望まれる。

6.2.3　同時推定における群の併合

実際に行われるテストでは、複数の異なる背景を持つ受験者が受験していると仮定するのが自然であることがある。たとえば、日本国内における受験者と海外における受験者、といった場合である。また、大学 1 年次向けのクラス分けテストの場合、推薦入試に合格した受験者と一般入試に合格した受験者とで、勉学の背景が異なる、と仮定することにより、受験者の特徴を加味した θ の尺度を構成することが可能となる。

このような、毎回の試験において複数の背景を持つ受験者が混在するようなテストの場合は、それぞれの集団に別々の問題冊子を提示し、それらにアンカー項目を混ぜた共通受験者デザインを考え、本研究で示したテスト実施法に基づく試験を行い続けることができる。たとえば大学 1 年次向けテストであれば、予備試験からのアンカー項目を適切に各年度用の問題冊子に振り分けることにより、同一の問題項目を全く含まない複数の問題冊子を用いることが実現できる。ただし、少数の項目について、予備試験を行い、最初の数年度用のアンカー項目を項目バンクに入れることが前提である。

本研究から得られた知見によれば、各年度のテスト実施後に行われる等化に際し、似た θ の平均を示す群をひとまとめに定義して多群 IRT モデルを適用し、「同時推定＋規準集団への等化」の方法を用いることにすれば、BILOG-MG を用いた場合であっても、メモリ不足に陥ることなく等化を行うことが可能である。たとえば大学 1 年次向けテストを「推薦入試群」「AO 入試群」「一般入試群」「その他群」それぞれに 1 フォームずつ作成するような場合、1 年あたり 4 フォームが出題

される。この場合であっても、それぞれの選抜方法によって入学を許可された受験者の能力の平均が、選抜方法ごとに異ならないという傾向が続くのであれば、群の違いを「規準集団」「推薦入試群」「AO 入試群」「一般入試群」「その他群」の 5 群と考え、同時推定においては常に 5 群を仮定した IRT モデルを適用すればよいことになる。

6.3 尺度の一次元性と等化方法

6.3.1 個別推定と同時推定

一次元性があらかじめデータからわかっている場合、もしくは他の外的基準によって一次元であるということがもっともらしいと判断される場合は、個別推定が有効な手段である。そうでない場合、すなわち、一次元性がテスト実施前に確認できない場合は、モデルのあてはまりの観点から考えると、個別推定はバイアスのかかった推定結果を返すことが示唆され、同時推定を行うことが有用であると考える。Karkee, Lewis, Hoskens, Yao and Haug (2003) は、垂直等化の場合、「成績上位群向けと下位群向けのフォーム間で同時推定を行い、そののち、規準集団への等化を個別推定で行う」という方法を提案している。このような方法の妥当性を検証することも、今後必要であると考える。Karkee et al. (2003) は同時に、「個別推定の方が多次元データに対してより適した方法である」と指摘している。しかし、ここでいう「多次元データ」とは、毎回の試験の間で一次元性が保証できない場合を指しており、1.6.3 節で述べた「各実施回の間で一次元性が満たされない場合」に相当する。一方、本研究においては、「等化の前提が満たされているとテスト実施機関が考えうる場合」、すなわち、予備試験及び各実施回において一次元性が仮定されるものの、予備試験と本試験の間で異なる次元の構成概念が測定されているテスト場面を想定した。このような場合は、5.4.1 節で述べたように、個別推定は妥当な結果とならない可能性が指摘できる。

いずれの場合においても、テストのデータが事後的に採集されなければ、このような議論をするのは難しい。しかし、事後的であっても、最低限、予備試験と本試験との間で一次元性が見られるかどうか、因子分析などの方法により、検討することが望ましいと考える。その上で、個別推定（テストをまたいでの一次元性が見られる場合）と同時推定（一次元性が見られない場合）のいずれをとるかを考えることも必要であると考える。

6.3.2 共通受験者デザインでの多次元 IRT モデルにおける等化の可能性

共通受験者デザインの場合に限定されるが、毎回の試験において多次元性を持つようなことが分かっている尺度に多次元 IRT モデルを適用し、解釈可能な因子が毎回得られていることが分かっている場合の等化法として、2 段階推定による分析を応用することができると考える。この方法は、個別推定を多次元 IRT モデルの「困難度等化法」の考え方で行うことを目的としている。以下の説明では、規準となるテスト X（等化先）とテスト Y（等化元）があり、それらをすべての受験者が受験しているものとする。ここで、それぞれのテストで見出された F 個の因子が同一の構造を持つであると認められる場合、多次元 IRT モデルによる解をまず求め、次にテスト X のそれぞれ

の因子 f $(f = 1, 2, ..., f, ..., F)$ における $\theta_f^{(X)}$ の平均と分散をある値（たとえば、0 と 1）に固定した上でテスト Y の $\theta_f^{(Y)}$ からテスト X の $\theta_f^{(X)}$ に対する回帰（多次元 IRT モデルの場合は、パス解析）を行う。このような「2 段階推定」を行うと、その回帰係数および切片を用いて等化係数を推定することができる。その等化係数を用いて、テスト Y に存在する各因子において、項目パラメタを変換することにより、テスト Y の「テスト X の尺度上」における項目パラメタが得られる。

このような 2 段階推定を行う場合、回帰係数を過小推定することが知られており、光永・星野・繁桝・前川 (2005) の方法でその偏りを修正することができる。しかし、パス解析モデルを用いた場合、たとえばテスト Y における第 2 因子の能力値 $\theta_2^{(Y)}$ からテスト X の第 1 因子の $\theta_1^{(X)}$ への回帰係数をどのように等化係数に反映させるかを考える必要がある。多次元 IRT モデルにおいての等化及びリンクに関する議論 (Reckase, 2009, chapter 8) と整合性のとれる等化係数となるような工夫が必要である。また、本研究で示したような複数回にわたって試験を実施し続ける場合に、どれほど安定した結果となるかについては、今後の実践研究を待つしかないだろう。

6.4　モデル研究として考えた場合の個別推定と同時推定

これまで述べてきた個別推定および同時推定について、それぞれ多変量解析のモデルとしてみた場合、個別推定は以下のように記述できる。すなわち、「各々の回のデータから得られた項目パラメタの推定値」と「規準集団のデータから得られた項目パラメタの推定値」を別個に推定し、それらの間に共通して現れる項目パラメタの推定値を基に、等化元の項目パラメタのセットを線形変換するための係数を推定する。このように記述すると、本質的には個別推定は「2 段階推定」を行っていることになる。このような 2 段階推定は、光永ほか (2005) で指摘したような「回帰係数の推定値の過小推定」や「推定値の一致性 (consistency) の欠如」といった数理的な欠点を持つ。一方、同時推定の場合は、そのような欠点はない。

しかし、本研究の多群 IRT モデルに関するパラメタ推定においては、「求められたパラメタが真値と異なる度合い」と、「推定された際の尤度」に顕著な関連が見られなかった。多群 IRT モデルにおいては、グループの数を増やした方がより適合度の高い結果を返す。しかし、グループの数を増やして高い適合度を得た条件において、真値からのずれが大きい項目パラメタの推定結果となったことから、グループの特徴を示すパラメタを増やしたモデルにおいては、それらのパラメタに関して過剰に適合する結果を返しやすいという傾向があることが指摘できる。この傾向に注意すれば、モデルのパラメタ推定という枠組みにおいては、個別推定よりも同時推定がより好ましい結果を返すことが期待される。

ただし、第 3 章の結果において考えた 2 種の同時推定で、規準集団の θ の平均が 0、標準偏差が 1 とおいただけの「同時推定 MG」のモデルと比較して、「同時推定＋規準集団への等化」の方が、項目パラメタ推定値はより真値に近い結果となった。この結果は、数理的に好ましい推定法だけでは、必ずしも目的にかなった推定値を返さないことを示している。このような点に、テストの研究における「より数理的に好ましいモデルを考える」モデル研究と「よりテスト場面で実効性のある結果が得られる方法を考える」実践研究の違いが見える。また、モデル研究と実践研究のいずれか

一方だけをとりあげる方法では、テストの尺度化、あるいは等化の研究では不足であることも指摘できる。

6.5 本研究の限界、また将来の等化研究に向けて

本研究では、テスト実施とともに項目バンクのサイズが増えるようなテストデザインを取り上げた。一般に、「すべての場合に共通してあてはまる、最良の等化方法というものは存在しない」(Skaggs and Lissitz, 1986, p.516) ことが、等化研究において知られており、本研究のように、個々のテストデザインに関して、実践的であるかどうかといった観点からの考察を交えて「最も良い」等化方法を提案する、というのが、等化研究のスタンダードであるといえるだろう。本節では、今後の研究の指針、また今後の課題について述べる。

6.5.1 テスト研究に実データを用いることの意義 – テスト結果による政策決定–

本研究に類する等化に関する研究は、テスト機関に心理統計学者 (psychometrician) が従事し、論文誌に掲載される場合と、そのテスト機関が研究ノート (research note) の形で公刊する場合がある。後者の場合は、たとえば Hanson and Béguin (1999) や Pang et al(2010) などがある。これらの研究では、実際の受験者から得られたテストのデータが用いられている。これらの研究成果は web site に公開されることが一般的になってきている（前述の 2 種の研究ノートはいずれも web site からだれでも入手可能である)。これらの研究においては、(1) 等化結果を得るための手続き、(2) 従属変数の定義の明確化、の 2 つが記載されていることが望まれる。特に (1) は、どのようなテストデザインを用いたか、どのようなソフトウェアを用いたか、といった点に関して、再現可能なように記述する必要があろう。

翻って、日本においては、このような研究が進んでいるとは言えない。等化やリンクといった方法で得点の尺度を調整するテストが、日本では一般的ではないという側面も指摘できるかもしれない。しかし、日本のテスト機関がテストの研究をし、研究者が実データを分析した結果を公刊することが難しい現状が指摘できる。特に、実際に得点を返す場面に用いた 0-1 データを用いた研究の場合、たとえ項目困難度の分析のためにのみ用いる場合であっても、その項目分析の結果を公表することが難しい。受験者に返した得点の根拠として 0-1 データをとっている以上、後から研究目的に転用するためには、仮にその結果を公表することで生じる不都合に関し、テスト機関が責任を負わなければならず、やむを得ない部分もあるだろう。しかし、テストの実施回をまたいで比較可能なスケールを提供し続けることのできるテストの実践研究は、本研究で行ったようなシミュレーション研究のみならず、実データを用いた研究が行われなければならないと考える。そのような研究成果が多く集積できれば、テスト実施機関が行いたいテストの仕様に合わせたテストデザインを考えるだけで、適切な等化方法を先行研究の知見から選択できるようになるだろう。しかしそうなるためには、研究の数のみならず、テスト実施団体における学術研究への理解と協力姿勢がこれまで以上に深まらなければならないだろう。一例として、アメリカにおいて「全米学力調査」を教

育省の責任において実施するにあたり、常勤および非常勤合わせて 5000 名規模のスタッフが働いており、その中には連邦政府及び教育省から委託を受けた大手のテスト機関に属するスタッフも含まれる（荒井、2008、p.29）。大手のテスト機関が行政および立法といった国の機関とともにテストを実施し、それらを通じて分析結果を国民の間で共有するという「アメリカにおけるテストの文化」を、日本においても導入することが必要と考える。また、全米学力調査においては、木村 (2008) で記述されているように、政府の「落ちこぼれ防止法」(The No Child Left Behind Act of 2001) 制定に伴う制度変更といった履歴が見られ、その過程で常に「学力調査の結果何が分かったのか」に関する説明責任がテスト実施者側のみならず政府（連邦政府、州政府）に問われていることがわかる。この種の説明責任は、国が違えども、普遍的にテスト実施団体が問われるべき性質のものであり、説明責任の重大性を何らかの形でテスト実施団体自らが認識しなければ、日本におけるテストを取り巻く環境は現状のままである可能性が高いといえる。

なお、実データを用いたシミュレーションを行った場合であっても、本研究で見られた知見は有効であると考える。項目パラメタの大小について、識別力の高低と困難度の絶対値の高低（すなわち「極端に困難度が高いか低い」か「困難度が中庸である」か）をクロスさせた 4 パタンを考えると、「識別力が高く困難度が極端」な項目は実際にはほとんど出現しない。そのような項目は、学力レベルの極端に高い、もしくは低い受験者を良く識別できる項目であるものの、極端な学力レベルを持った受験者が大多数を占める受験者集団があって初めてそのような推定結果が得られることから、現実にはほとんど出現しないことがわかる。そのほかの 3 パタン、すなわち「識別力が高く、困難度が中庸」「識別力が低く、困難度が極端」「識別力が低く、困難度が中庸」の項目は、現実のテスト場面でも頻出する。本研究のシミュレーションにおいては、このような「現実のテスト場面で見られる項目パラメタの偏り」を、忠実にシミュレートしていないことは確かであり、その影響の効果を検証する研究を、実データから推定した項目パラメタのセットを用いて行うことが必要であるといえる。しかし、本研究で発生させたランダムな項目パラメタの真値は、前述の 4 パタンを尽くすものであると考えれば、本研究において検討してきた「等化後の項目バンク上における項目パラメタの妥当性」は、等化の理論的枠組みの範疇において検証できている部分が大きいと考える。

6.5.2　多次元 IRT における等化方法の検討

本研究では、テストの一次元性に関して、項目パラメタの識別力の大きさによって 1 因子が仮定できるかどうかを議論している。しかし、実際の試験においては、むしろ一次元性が満たされていない場合が多く、そのためには多次元 IRT モデルを適用すべきであるという提案がなされている。多次元 IRT における等化の研究に関しては、多次元 IRT のパラメタの解釈など、多次元 IRT の方法論に関するテスト機関におけるコンセンサスや、実践例が豊富に存在することが前提となっていると考える。しかし、多次元 IRT の実践例が増えるためには、多次元 IRT の方法論に関する議論が多くなることが前提であるという側面もあり、現状ではあまり進んでいないといえるだろう。

本研究では、人工的に生成したデータに関して、多次元性の操作をおこなっているものの、極めて限定的な操作にとどまっている。多次元 IRT モデルによくフィットするようなデータに関して

一次元の IRT を仮定した場合のリンク方法の検討は、多次元 IRT におけるパラメタ推定の研究ともども、テストの実践場面においても有益な手掛かりを与えるものといえよう。今後は、これらの研究において新たな知見が蓄積され、本研究で取り上げたテストデザインにおけるテスト実践場面における応用がなされるものと考える。

6.5.3 多値型データにおける等化方法の検討

本研究では、0-1 データのような、2 値型のデータ (dichotomous data) に対する 2PL をとりあげた。データの型としては、他に多値型データ (polytomous data) がある。これはたとえば、小論文の評価を 3 段階の順序尺度上で行う場合に相当する。この場合、ある θ を持った受験者が「0」「1」「2」の 3 種の評価になる確率をそれぞれ ICC で表現することで、IRT の枠組みで順序尺度を扱うことが可能である。具体的なモデルとしては、1.1.4 節で述べた段階反応モデルや一般化部分採点モデルなどが挙げられる。

多値型データで得られた項目パラメタを用いた、共通項目デザインによる等化方法は Kolen and Brennan (2004, pp. 208-230) 等で扱われている。しかし、本研究で取り上げたデザインにおける等化については、実践例を含め、今後研究される必要があると考える。

謝辞

本研究は、小生がこれまでかかわってきたいくつかの試験分析の実務場面で、疑問であったトピックに関し、一つの結論を見出したいと考え、行われたものです。したがって、小生が試験の分析の仕事に関わることがなければ、編み出されることはありませんでした。

本論をまとめるにあたり、まず、指導教授である前川眞一先生に、多大なる感謝の念を表します。小生が様々な意味で未熟であったころからさまざまな場面で奮励を促し、時に厳しく、時にあたたかい指導をいただきました。また、試験の分析を行うポジションの中で、実データに触れる経験を小生に与えてくださったのも先生のおかげであり、このような研究が面白いと思えるようになったのも、先生のおかげであります。また、本論文の審査にあたりまして、多忙な中特に貴重な時間を頂戴いたしまして、有益なコメントをいただきました中川正宣教授、石井源信教授、中山実教授、室田真男教授に、心より御礼申し上げます。

また、新潟大学の杉澤武俊先生、株式会社教育測定研究所の野上康子様、山口大学の大澤公一先生には、テストの分析の現場において、小生が分析のノウハウを学ぶ上で多くの良い影響を受けました。本論文の内容にも、お三方から学んだ部分が多く反映されています。一方、大学入試センターの荒井清佳先生、東海大学の藤田智子先生には、多様なテストの実践例をお教えいただき、本論の執筆に際しましても多くの有益な示唆をいただきました。さらに、テストの妥当性、とりわけ一次元性に関する議論につきましては、慶應義塾大学の中村優治先生とのディスカッションが有益であったと考えております。このほか、ここに記すには紙幅の都合上かないませんが、多くの先生方にご指導いただきましたこと、ここに記して感謝申し上げます。

最後になりましたが、小生を支え、また小生がここまで至る道をずっと見守ってくれた、父・崇彦、母・トク子に、感謝の念をここに表します。

業績一覧 (平成 24 年 12 月 6 日現在)

本研究に関連する研究発表

査読付き論文

光永悠彦・前川眞一 (2013). 多群 IRT モデルにおける簡素化の評価 —— 水平等化場面のシミュレーョンを通じて ——　行動計量学 (印刷中) （第 5 章）

光永悠彦・前川眞一 (2012). 項目反応理論に基づくテストにおける項目バンク構築時の等化方法の比較　日本テスト学会誌, **8**, 31-48.　（第 3 章）

学会発表

光永悠彦・前川眞一 (2012). 共通項目デザインを用いた大規模テストにおける等化順序の効果の検討　日本教育心理学会第 54 回大会発表論文集（琉球大学）（ポスター発表）（第 2 章）

光永悠彦・前川眞一 (2011). 項目プールのサイズが実施ごとに増大するような大規模テストにおける項目パラメタの更新法　日本テスト学会第 9 回大会（口頭発表、日本テスト学会大会発表賞）（第 3 章）

国内講演

光永悠彦 (2010). わが国における言語テストの社会的位置づけ ～テスト分析者から見た現状と課題～　日本言語テスト学会（JLTA）研究例会（第 1 章、第 6 章）

その他の研究発表

査読付き論文

光永悠彦・星野崇宏・繁桝算男・前川眞一　(2005). 因子スコアや潜在変数得点を用いた構造方程式モデルの母数推定の偏りの解決　行動計量学, **32(1)**, 21-33.

国際学会発表（審査あり）

Mitsunaga,H. & Mayekawa, S. (2005). The Development of an Item Parameter Calibration Program for IRT models with the Improved Initial Estimates and Distribution Estimation Method. *International Meeting of the Psychometric Society 2005,* Tilburg, the Netherlands.(ポスター発表)

Mitsunaga,H. & Mayekawa, S. (2009). Item parameter calibration with non normal ability distribution. *International Meeting of the Psychometric Society 2009,* Cambridge, UK.(口頭

発表)

Nakamura, Y., & Mitsunaga, H. (2011). Five-consecutive-year analysis of Japanese students' English proficiency using a large-scale in-house placement test. *The 16th World Congress of Applied Linguistics (AILA2011)*, Beijing, China. (口頭発表)

Nakamura, Y., Murray, A., & Mitsunaga, H. (2012). Second Language Vocabulary Assessment. *The 17th Conference of Pan-Pacific Association of Applied Linguistics (PAAL2012)*, Beijing, China. (口頭発表)

国内学会発表

星野崇宏・光永悠彦・繁桝算男・前川眞一 (2003). 潜在変数得点の推定値を用いた構造方程式の母数の推定について　日本行動計量学会第 31 回大会発表論文集 pp.134-137(口頭発表)

光永悠彦・前川眞一 (2005). 正規分布の確率密度に比例した離散能力分布の推定法　日本教育心理学会第 47 回大会（ポスター発表）

光永悠彦・前川眞一 (2008). テスト得点への数値カテゴリを持つ多項分布の当てはめ　日本テスト学会第 6 回大会（口頭発表）

中村優治・光永悠彦 (2009). 英語プレイスメントテスト開発と英語読解能力の経年的変化に関する分析の試み　日本テスト学会第 7 回大会（口頭発表）

Nakamura,Y. & Mitsunaga, H. (2010). Constructing a large-scale placement test for measuring students' English proficiency. *2010 JACET (The Japan Association of College English Teachers) 49th Convention,* Miyagi, Japan. （口頭発表、英語）

Nakamura,Y. & Mitsunaga, H. (2010). Constructing a Large-scale English Placement Test. *JALT 2010: 36th Annual International Conference on Language Teaching and Learning & Educational Materials Exhibition,* Nagoya, Japan. (口頭発表、英語)

引用文献

[1] 荒井克弘 (2008) 全米学力調査の概要とその背景　荒井克弘・倉元直樹 (編著) 全国学力調査　日米比較研究　金子書房

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

[3] Arai, S., & Mayekawa, S. (2005). The characteristics of large-scale examinations administered by public institutions in Japan: From the viewpoint of standardization. 日本テスト学会誌, **1**, 82-92.

[4] Baker,F.B., & Kim, S-H. (2004). *Item response theory: Parameter estimation techniques (2nd ed., revised and expanded).* NY: Marcel Dekker.

[5] Birnbaum, A. (1968). Some latent trait models and their uses in inferring an examinee's ability. In Lord, F.M.. & Novick, M.R., *Statistical theories of mental test scores (pp. 395-479).* Reading, MA: Addison-Wesley.

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

[7] Bock, R.D., & Lieberman, L. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika*, **35**, 179-197.

[8] Bock,R.D., & Zimowski, M.F. (1996). Multiple group IRT. in van der Linden, W. & Hambleton,R.K. (eds) . *Handbook of modern item response theory.* NY: Splinger-Verlag.

[9] 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.

[10] DeMars, C.E., & Jurich, D.P. (2012). Software note: using BILOG for fixed-anchor item calibration. *Applied psychological measurement*, **36(3)**, 232-236.

[11] Dorans, N.J. & Holland, P.W. (2000). Population invariance and the equatability of tests: Basic theory and the linear case. *Journal of Educational Measurement, 37,* 281-306

[12] du Toit, M. (Ed.) (2003) *IRT from SSI.* Scientific Software International.

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

[14] Hanson, B.A. (2002). IRT command language Version 0.020301 [Computer Program].

Monterey, CA: Author (http://www.b-a-h.com/software/irt/icl/index.html)

[15] Hanson, B.A., & Béguin, 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**, 3-24.

[16] Hanson, B.A., & Béguin, A.A. (1999) *Separate versus concurrent estimation of IRT item parameters in the common item equating design.* ACT research report series 99-8. Iowa City, IA: ACT, Inc.

[17] Hu, H., Rogers, W.T. & Vukmirovic, Z. (2008). Investigation of IRT-based equating methods in the presence of outlier common items. *Applied Psychological Measurement,* **32**, 311-333.

[18] 池田央 (1994). 現代テスト理論　行動計量学シリーズ 7　朝倉書店

[19] Jodoin, M.G., Keller, L.A. & Swaminathan, H. (2003). A comparison of linear, fixed common item, and concurrent parameter estimation equating procedures in capturing academic growth. *The journal of experimental education*,**71(3)**,229-250.

[20] Kang, T. & Petersen, N.S. (2009). Linking item parameters to a base scale. *Paper presented at the National Council on Measurement in Education,* San Diego, CA.

[21] Karkee, T., Lewis, D.M., Hoskens, M., Yao, L., & Haug, C. (2003) Separate versus concurrent calibration methods in vertical scaling. *Paper presented at the Annual Meeting of the National Council on Measurement in Education,* Chicago, IL.

[22] 木村拓也 (2008) 2003 年以降の全米学力調査の変質　荒井克弘・倉元直樹 (編著) 全国学力調査　日米比較研究　金子書房

[23] 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 –　日本テスト学会誌, **3**,84-90.

[24] 倉元直樹 (2011). 個別大学の追試験における得点調整方法に関する一提案 – タッカーの線形等化法を用いて – 日本テスト学会誌, **7**, 68-83.

[25] 倉元直樹 (2008). テスト・スタンダードからみたわが国の全国学力調査の条件 荒井克弘・倉元直樹 (編著) 全国学力調査　日米比較研究　金子書房

[26] Lee, W.-C., & Ban, J.-C. (2010). A comparison of IRT linking procedures. *Applied Measurement in Education*, **23**, 23-48.

[27] Lei, P-W., & Zhao, Y. (2012). Effects of vertical scaling methods on linear growth estimation. *Applied psychological measurement* , **36(1)**, 21-39.

[28] Li, Y-H., Griffith, W.D., & Tam, H.P. (1997) Equating multiple tests via an IRT linking design: utilizing a single set of anchor items with fixed common item parameters during the calibration process. *Paper presented at the Psychometric Society Meeting*, Knoxville, TN.

[29] Li, Y-H., Tam, H-P., & Tompkins, L.J. (2004). A comparison of using the fixed common-

precalibrated parameter method and the matched characteristic curve method for linking multiple-test items. *International Journal of Testing*, **4(3)**, 267-293.

[30] Lord, F.M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum.

[31] Lord, F.M., & Novick, M.R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

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

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

[34] Mislevy, R.J. (1984). Estimating latent distributions. *Psychometrika,* **49(3)**, 359-381.

[35] 光永悠彦・星野崇宏・繁桝算男・前川眞一 (2005). 因子スコアや潜在変数得点を用いた構造方程式モデルの母数推定の偏りの解決 行動計量学, **32(1)**, 21-33

[36] 村木英治 (2011). 項目反応理論　シリーズ＜行動計量の科学＞8 朝倉書店

[37] Muraki, E., Hombo, C.M., & Lee, Y-W. (2000). Equating and linking of performance assessments. *Applied psychological measurement*, **24(4)**, 325-337.

[38] Muraki, E. (2000). RESGEN4: Item Response Generator [Computer Program].

[39] Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement,16*, 159-176.

[40] Muraki, E., & Bock, R.D. (2003) PARSCALE [Computer Program]. Lincolnwood, IL: Scientific Software International.

[41] 室橋弘人 (2005). 多母集団 IRT モデル　豊田秀樹 (編)　項目反応理論 [理論編] –テストの数理–　朝倉書店

[42] Nakamura, Y., & Mitsunaga, H. (2011) Five-consecutive-year analysis of Japanese students' English proficiency using a large-scale in-house placement test. *The 16th World Congress of Applied Linguistics (AILA2011)*, Beijing, China.

[43] 日本テスト学会 (編) (2010). 見直そう、テストを支える基本の技術と教育　金子書房

[44] 日本テスト学会 (編) (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

[45] Ogasawara, H. (2001). Marginal maximum likelihood estimation of item response theory (IRT) equating coefficients for the common-examinee design. *Japanese Psychological Research,* **43(2)**, 72-82.

[46] 尾崎幸謙 (2005)．周辺最尤推定法　豊田秀樹 (編)　項目反応理論 [理論編] –テストの数理–　朝倉書店

[47] 尾崎幸謙 (2003)．項目反応理論の等化　豊田秀樹 (編)　共分散構造分析 [技術編] –構造方程式モデリング–　朝倉書店

[48] Pang,X., Madera, E., Radwan,N., & Zhang,S. (2010). *A comparixon of four test equating methods.* Report prepared for EQAO. Toronto, ON: Education Quality and Accountabil-

ity Office (EQAO).

[49] Patz, R.J., & Junker, B.W. (1999). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of educational and behavioral statistics*, **24(2)**, 146-178.

[50] Petersen, N.S., Kolen, M.J., & Hoover, H.D. (1989) Scaling, Norming and Equating. in Linn, R.L. (Ed) *Educational measurement, third ed.*, New York: Macmillan.
（前川眞一・訳 (1992)　尺度化、規準化、および等化 ロバート・L・リン（編）教育測定学 第 3 版.　みくに出版.）

[51] Reckase, M.D. (2009) *Multidimensional item response theory.* New York: Splinger.

[52] 斉田智里 （2003）. 高校入学時の英語能力値の年次推移– 項目反応理論を用いた県規模英語学力テストの共通尺度化–. 第 15 回英検研究助成報告, 12-24. 日本英語検定協会.

[53] 斉田智里・柳川浩三 （2011）. 共通項目デザインによる神奈川県高等学校「県下一斉英語学力テスト」の開発–項目応答理論を用いた等化によるテストの再評価と展望–. 日本テスト学会誌, **7**, 121-132.

[54] Samejima, F. (1969). Estimation of latent trait ability using a response pattern of graded scores. *Psychometrika, Monograph Supplement, 17.*

[55] 芝祐順 (編) (1991). 項目反応理論　基礎と応用　東京大学出版会

[56] 芝祐順・野口裕之 (1982). 語彙理解尺度の研究 I – 追跡データによる等化– 東京大学教育学部紀要, **22**, 31-42.

[57] Shigemasu, K., & Nakamura, T. (1996). A bayesian marginal inference in estimating item parameters using the Gibbs sampler. *Behaviormetrika,* **23**, 97-110.

[58] Skaggs,G., & Lissitz, R.W. (1986). IRT test equating: Relevant issues and a review of recent research. *Review of educational research* , **56(4)**, 495-529.

[59] Stocking, M.L., & Lord, F.M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, **7**, 201-210.

[60] Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis discretized variables. *Psychometrika*, **52**, 393-408.

[61] Thurstone, L.L. (1947). *Multiple Factor Analysis.* Chicago: University of Chicago Press.

[62] 豊田秀樹 (2012). 項目反応理論 [入門編] 第 2 版 朝倉書店

[63] 張　一平 （2009）. 2 パラメータと 3 パラメータの項目反応モデルにおける比較. 行動計量学, **36(1)**, 15-24.

[64] von Davier, A.A., Holland, P.W., & Thayer, D.T. (2004) *The kernel method of test equating* . New York: Springer-Verlag.

[65] 渡辺直登・野口裕之 （1999）. 組織心理測定論–項目反応理論のフロンティア–.　白桃書房

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

[67] 吉村宰 (2009) 大学入試センターによるテストデータベースによる項目分析　植野真臣・永岡慶三 (共編) e テスティング　培風館

[68] 吉村宰・荘島宏二郎・杉野直樹・野澤健・清水裕子・齋藤栄二・根岸雅史・岡部純子・フレイザー，サイモン. (2005). 大学入試センター試験既出問題を利用した共通受験者計画による英語学力の経年変化の調査　日本テスト学会誌,**1**, 51-58.

[69] Zimowski, M., Muraki,E., Mislevy, R., & Bock,R. (2003) BILOG-MG 3 [Computer Program]. Lincolnwood, IL: Scientific Software International.