統計数理(2024)
第 72 巻 第 1 号 61–78

特集「心理統計学の新展開」
[原著論文]

心理尺度の統計的共通化：等化とリンキングの方法と実践

光永 悠彦†

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

要　　旨

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

キーワード：テスト理論，項目反応理論，大規模テスト，教育測定．

1. はじめに

1.1 テスト得点の尺度化と項目反応理論(IRT)

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

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

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

である(詳細は後述する).

1.2 共通尺度化の必要性とその方法

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

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

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

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

1.3 本論の目的

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

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

すく説明する方法について述べる．

2. 等パーセンタイル法による尺度調整

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

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

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

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

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

で求められる．

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

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

て，使用する等化・尺度調整の手法も適切に選択する必要がある．

3. IRT による尺度化と等化

3.1 IRT に基づく尺度化

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

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

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

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

項目パラメタの推定においては，尤度関数として

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

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

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

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

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

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

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

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

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

3.2 多母集団を仮定した場合の尺度化

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

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

3.2.1 IRT の尤度関数を多母集団に拡張する方法

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

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

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

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

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

3.2.2 潜在回帰モデルによる方法

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

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

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

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

4. 異なるテスト間における尺度の等化方法

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

4.1 等化に必要な共通要素

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

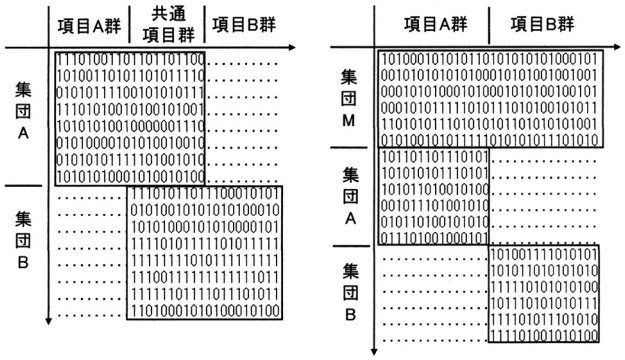

図 1. 共通項目デザインと共通受検者デザイン．

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

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

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

4.2 IRT による等化 (1) 個別推定

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

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

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

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

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

となり，θ_{iX} から θ_{iY} への変換式は

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

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

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

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

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

4.2.1 Mean/Sigma 法及び Mean/Mean 法

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

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

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

を用いて，

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

4.2.2 Haebara 法及び Stocking-Lord 法

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

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

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

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

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

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

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

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

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

4.3 IRT による等化(2)同時推定

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

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

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

4.4 IRT による等化(3)固定項目パラメタ法

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

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

4.5 θ の尺度得点への変換

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

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

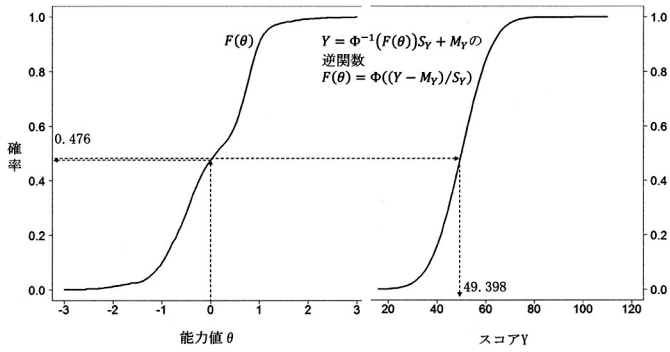

図 2. 正規分布していない θ を正規分布に従う尺度得点 Y に変換する方法（光永, 2017, p.137）.

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

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

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

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

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

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

5. 大規模テストにおける等化

5.1 大規模テストで求められる要件と項目バンク

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

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

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

5.2 等化手続きのブロックダイヤグラムによる図示

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

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

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

5.3 大規模テストにおけるテストデザイン例

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

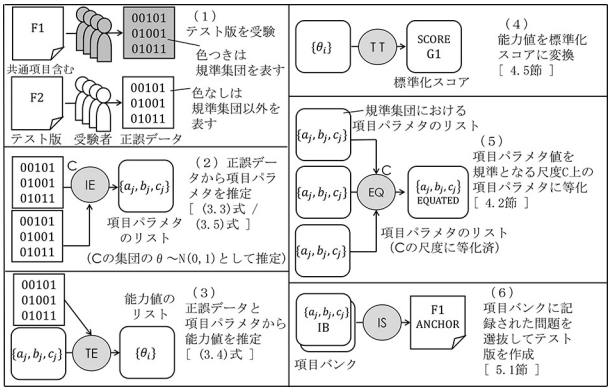

図 3. 等化にあたってデータ処理の流れを図示するためのアイコン（光永, 2022 を一部改変）.

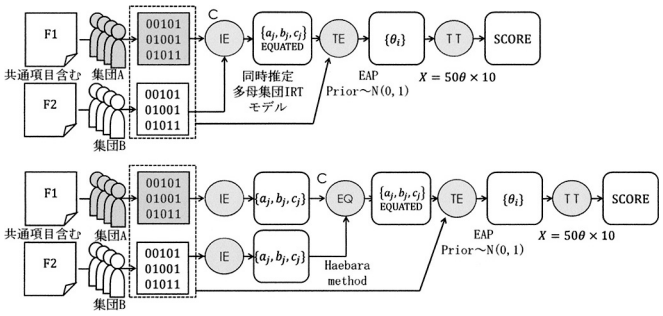

図 4. 共通項目デザインによる等化のブロックダイヤグラム．上は同時推定，下は個別推定の場合を表す．

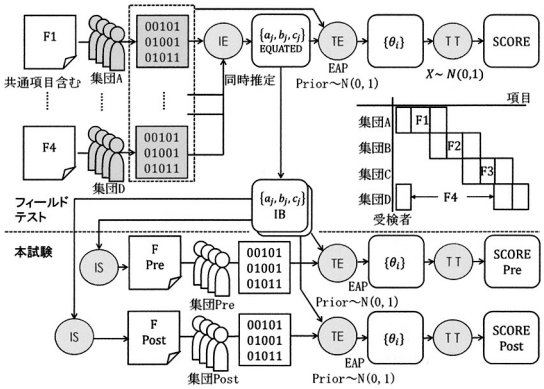

図 5. 事前テストと事後テストの間で比較可能な得点を得るテストのためのブロックダイヤグラム．フィールドテストの重複テスト分冊法による項目割り付け図とともに記した．

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

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

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

6. 大規模テストにおける等化の例

6.1 テストデザイン及び調査の実施

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

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

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

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

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

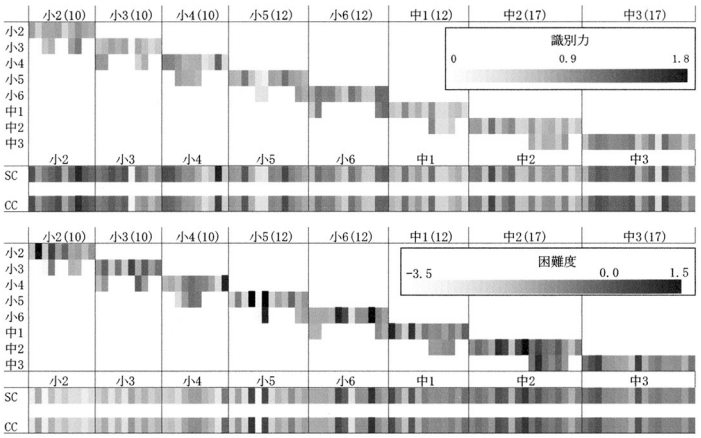

図 6. 上から，個別のテスト版の出題状況及び項目ごとに推定された等化前の識別力パラメタ，個別推定により等化した識別力パラメタ(SC)，同時推定により等化した識別力パラメタ(CC)，個別のテスト版から推定された困難度パラメタ，個別推定により等化した困難度パラメタ(SC)，同時推定により等化した困難度パラメタ(CC)．それぞれのセルの色は，識別力パラメタの値の大小を表している．( )内の数字は，当該学年向けに出題された項目数を示す．

6.2 項目パラメタの垂直尺度化結果

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

6.3 垂直尺度化における実践上の注意点

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

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

7. おわりに

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

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

謝　辞

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

参 考 文 献

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Haruhiko Mitsunaga

Graduate School of Education and Human Development, Nagoya University

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

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