【個人研究】

部分得点モデルにおける同時尺度調整法による垂直的等化の改訂報告

藤森　進*

Revised report of a simulation using vertical equating by concurrent calibration in Fujimori's partial test score model

Susumu FUJIMORI

There were serious errors in creation of simulation data in a previous study (Fujimori, 2009) that led to incorrect results. Consequently, these errors have been corrected in the current study. Using concurrent calibration in a polytomous item response model, this study examines the vertical equating of the averages of two populations through simulation. Simulated polytomous test data were generated using Fujimori's partial test score model with vertical equating of assumed conditions in which an individual test was conducted for two populations having significantly different ability (some anchor polytomous items were included in both tests). The population average and variance were estimated using the mean and variance of estimated ability scores. The following explains the simulation results. The difference between the averages of two populations under conditions in which concurrent calibration was used was readily reproducible. Results showed that it improved proportional to the number of anchor polytomous items among test items. Polytomous anchor items should account for more than a third of test items in order to achieve satisfactory results using vertical equating by concurrent calibration.

Key words: item response model, vertical equating, concurrent calibration, simulation, partial test score model
項目反応モデル　垂直的等価　同時尺度分析法　シミュレーション　部分得点モデル

1. 研究の目的

複数のテストの得点を、ある共通尺度上に位置付けることは、等化と呼ばれる。近年、テストの等化には、項目反応理論が利用されることが一般的である。同理論では、項目は正答を1、誤答を0とする２値のみをとることが仮定されており、この分析モデルとしては2母数ロジスティックモデルが代表的なものとして知られている。多値の得点を扱うモデルとしては、Samejima(1969)の段階反応モデルgradedresponse modelや評定尺度モデル(rating scale model;Andrich,1978) partial credit model (Masters,1982)などが代表的なものとして知られている。また藤森の部分得点モデル(partial test score model;藤森2002a)は簡明なモデルであり、項目母数の数も他の多値モデルと違って抑制されているなど実用性のあるもの

* ふじもり　すすむ　文教大学人間科学部心理学科

であり、英作文の採点などの実際のテスト結果への適用と運用も既に行われている実績のあるものである。部分得点モデルと他の代表的なモデルとの比較については、藤森(2002b)を参照のこと。多値の得点を扱うこれらのモデルは項目反応モデルの一群であるが、テスト得点の分析で本格的実務での運用を目指す場合には、これらのモデルを利用したときのテスト結果の等化が不可欠である。2値テスト得点の等化に関しては、例えば藤森(1999)などに見られるように、ある程度の結論が得られているように思われるが、多値のモデルを利用した場合の等化については十分わかっている状況ではない。昨年の研究(藤森,2009)では、部分得点モデルの垂直的等化が旨く行える基礎的条件(例えば等化情報を持つテストの項目数や人数など)をシミュレーションにより検討し、多集団の能力平均の差の再現性はないと報告したが、実はシミュレーションデータの作成プログラムに大きなミスがあり、全く間違った結論であった。このため、本研究では、昨年の研究に目を通してくださった読者にお詫びするとともに、昨年の資料とプログラムの誤りを訂正し、部分得点モデルの適用に関し正しい分析結果を提示せんとするものである。

2. 方法

方法は、藤森(2009)と同様であるが、部分得点モデルについて簡単に紹介するため、初めに代表的な項目反応モデルである2母数ロジスティックモデルについて説明し、その後部分得点モデルについて述べる。

2.1. 項目反応モデル

2.1.1. 2母数ロジスティックモデル

項目反応モデルに属するものは数多くあるが、式の2母数ロジスティックモデル(Birnbaum,1968)は、その代表的なものである。

$$P(\theta) = P\left(x_{ij}=1 \middle| \theta_i, a_j, b_j\right) = \frac{1}{1+\exp\left(-Da_j\left(\theta_i - b_j\right)\right)} \quad (1)$$

ここでは被験者、θはその能力を表す母数、D=1.7の定数、は項目番号、はその識別力、は困難度を表すモデルの母数である。または、被験者の項目に対する正誤を表し、正答のとき1、誤答のとき0となるダミー変数である。

2.2. 部分得点モデル

藤森(2002)の部分得点モデルを以下で簡単に紹介する。まず受験者のテスト項目jの得点が多値の得点 r_j によって表現されることを仮定する。部分的な得点を取り得る各テスト項目jは、各項目に固有であり潜在的な2値の正誤問題の合計得点(ただし、どの範囲の部分得点も扱えるようにするため平均化して０から１の範囲の部分得点になるようにする。採点の実務上は最低点を０、最高点を１となるようにする。即ち、採点された各得点を配点上の満点で割ればよい。)であることを仮定する。受験者は、テスト項目jの部分得点を、この潜在的な2値の項目に反応しながら得ることを仮定する。

またその潜在的な2値の正誤反応は式の2母数ロジスティックモデルが当てはまることを仮定し、しかもその母数は全て同一であることを仮定する。実際にはこの同一母数の仮定は、類似母数を持つ項目であれば近似的に成立するため、過度にこの仮定を重んじる必要はない。

式の $P_j(\theta)$ は、受験者が正答１又は誤答０のいずれか一方の値を取り得る潜在的問題に反応する確率である。多値項目jは2値の潜在項目を回繰り返し受験したときに、受験者が取りうる潜在的正誤反応パターンに基づいて生じると考える。項目jに関する尤度は潜在項目の母数が同一であるため

$$L(\theta) = P_1^{x_1}\left(1-P_1\right)^{1-x_1} \times \cdots \times P_s^{x_s}\left(1-P_s\right)^{1-x_s}$$

$$= \left\{ P^{\sum_{k=1}^{s} x_k} \left(1-P\right)^{\sum_{k=1}^{s}\left(1-x_k\right)} \right\} \quad (2)$$

となる。潜在的正誤得点の平均を考えるためには式のs乗根をとれば

$$\sqrt[s]{{}_1^{x_1}\left(1-P_1\right)^{1-x_1}\times\cdots\times P_s^{x_s}\left(1=P_s\right)^{1-x_s}}$$

$$=\left\{P^{\frac{\sum_{k=1}^{s}x_k}{s}}\left(1-P\right)^{1-\frac{\sum_{k=1}^{s}x_k}{s}}\right\}\quad(3)$$

となる。式の最尤推定値は一致するので、2値の和得点でなく、平均化した0から1の部分得点ｒ $_j$ による分析が可能となる。

$$Q_j\left(\theta\right)=1-P_j\left(\theta\right)\qquad(4)$$

$$\frac{\sum_{i=1}^{s}X_i}{s_j}=r_j\qquad(5)$$

として全て同一の母数であることを考慮して式の対数をとって、

$$\ell_{j,part}\left(\theta\right)=s_j\left(r_j\ln\left(P_j\left(\theta\right)\right)+\left(1-r_j\right)\ln\left(Q_j\left(\theta\right)\right)\right)\qquad(6)$$

となる。即ちテスト全体でθを推定するための尤度関数は

$$\ell_{part}\left(\theta\right)=\sum_{j=1}^{n}s_j\left(r_j\ln\left(P_j\left(\theta\right)\right)+\left(1-r_j\right)\ln\left(Q_j\left(\theta\right)\right)\right)\quad(7)$$

によって表される対数尤度 $\ell_{part}(\theta)$ を用いて受験者の能力θが推定される。ここではテストの項目数である。

これは一般の2母数ロジスティックモデルの対数尤度

$$\ell\left(\theta\right)=\log\left(\prod_{j=1}^{n}P_j^{x_j}Q_j^{1-x_j}\right)$$
$$=\sum_{j=1}^{n}\left\{x_j\log P+\left(1-x_j\right)\log Q\right\}\ (8)$$

の各項目jの尤度を式で置き換えたものになる。

ここで注意すべきは、観測可能なものは、受験者が問題ｊに対して獲得する０から１までの間の値を取りうる部分得点ｒ $_j$ であり、潜在的問題に対する受験者の潜在的な2値反応は観測できないという点である。また、同じ母数でなくても類似母数であるような潜在項目の場合も能力推定は近似的に一致することが示せるが、本研究では、同一母数の繰り返しと考えることにして、以下のシミュレーションによる検討を行っている。

2.2.1.　モデル母数と繰り返し数Sの推定

項目反応理論では、母数の推定を最尤法あるいはベイズ法によるのが一般的である。本研究では、モデル母数を推定するため

$$\ell_{part}\left(\theta\right)\varphi\left(\theta\right)$$
$$=\varphi\left(\theta\right)\sum_{j=1}^{n}s_j\left(r_j\ln\left(P_j\left(\theta\right)\right)+\left(1-r_j\right)\ln\left(Q_j\left(\theta\right)\right)\right)\ (9)$$

を最大とするような能力母数θと(潜在項目の正誤反応を2母数ロジスティックモデルとしたため)項目困難度と識別力を項目母数の推定値とする交互同時推定法を利用した。ここでは能力母数の事前分布であり、本研究では正規分布を仮定している。なお交互同時推定は問題があることが知られている。能力母数を尤度関数から積分により除外して得られる周辺尤度による項目母数の推定が適当とされるが、本研究では、能力母数と項目母数の最尤法による同時推定を利用した。

なお項目母数と能力母数の交互同時推定は自作のFORTRANプログラムで行った。

2.2.2. 等化方法

本研究では、初めに述べたように受験者集団の能力分布に差がある場合のテスト得点の等化、即ち垂直的等化が必要となる場面を問題としている。先に述べたように、本研究で採用したモデル母数の推定方法は、能力母数と項目母数の交互同時推定であり、この過程で得られた能力母数の推定値を利用して、テストの受験者集団の能力母数の平均、標準偏差について推定している。藤森(1998)などの2母数ロジスティックモデルでの同時尺度調整法での垂直的等化の分析結果を見る限り、良好な成績を示すとは期待し難いが、本研究ではこの方法によった。ベイズ推定の際は、このようにして推定された母集団分布を利用して、(9)式のφ(θ)を確定している。

2.3. シミュレーションデータ

等化成績の検証のためのシミュレーションデータは、以下のようにして作成した。今回は、昨年と同様、部分得点モデルの垂直的等化の基礎的条件を検討する研究目的であるので、テスト及び集団数は2ないし3とする。テスト数が4以上となる

ケースは機会を改めて検討したい。

2.3.1.　シミュレーションA1

データA1は、集団数が2で被験者数はいずれも3000人で、集団1の能力分布は、平均0、分散1の正規分布に従っている。また集団1は潜在的正誤項目として２値データ換算で40項目のテストを受験するとした。40項目の識別力母数は、平均0.8、標準偏差0.25、下限0.25、上限2.0の切断正規分布に従って作成された、困難度母数は平均0、標準偏差0.5の正規分布に従って作成され、これらの母数から式の正答確率を求め、0から1の一様乱数と比較して正答確率がより大きいときは、潜在的項目について正答1とし、小さいときは誤答0として2値の正誤データを作成した。2.2で述べたように部分得点モデルは同一母数の和得点であることを前提としているため、部分得点データは40項目の潜在的項目から同一母数5項目を選び(同一母数の潜在的項目を5つシミュレーションで作成して)、その正誤1-0データの平均を部分得点データとした。即ち、第1集団のテスト結果は、2値の潜在的40項目のテストであったため8項目の顕在的部分得点テストデータとなる。データA1の集団2は、集団1と同様、人数3000であり、垂直的等化を検討するため、能力母数の平均値が0.5と集団1よりやや高い能力分布としてある。第2集団に対するテスト項目は、項目困難度の平均が0.5と、第1集団に対するテストと比較して、やや難しい問題になっている点だけが異なる条件で作成されている。テストの難しさは受験者集団の能力分布に合わされるのが実際のテスト場面では当然だからである。第1テストと第2テストの共通項目は、8項目のうち4項目とした。即ち、テスト1,2を通して考えると12項目の部分得点問題があり、受験者集団1と2がどちらも受ける共通テスト項目が1から4の部分得点項目、5から8の部分得点項目は、受験者集団1だけが受験する項目となり、9から12の部分得点項目は、受験者集団2のみが受験する。データ全体での部分得点項目数は12問となる。以上のデータセットをA1-A5まで作成し、同時尺度調整法を適用した。

2.3.2.　シミュレーションA2

データA6からA13は、集団１と2のテストの共通項目数が5から7と増えるだけで2.3.1のシミュレーションデータと本質的に異なるところはない。共通項目数の増加は垂直的等化の成績を向上させることが期待されるが、これを確認するためのシミュレーションである。

2.3.3.　シミュレーションB

シミュレーションBは、シミュレーションA1と、共通項目数は同一の4項目であるが、データに含まれる項目数を増やし20としてテストデータの中で共通項目の占める割合を0.2と小さくしている点が2.3.1節と異なる。2.3.1節では共通項目の占める割合は0.333である。このシミュレーションは当初のシミュレーション計画にはなかったものであるが、シミュレーションA1、A2の結果を見て追加したものである。

2.3.4.　シミュレーションC

シミュレーションCは、2つのテストに含まれる共通項目数が2、データの項目数が14とそれぞれ少なくなっている点が2.3.1と異なるが他は同一である。

2.3.5.　シミュレーションD

シミュレーションDは、集団数が3となり、第１集団と第２集団の能力分布は2.3.1と変わらないが、第３集団は能力母数の平均が1.0となり、潜在的正誤２値のテスト項目の困難度平均が1.0となる点が2.3.1と異なる。更に共通項目の数が8と増加させてある。人数は集団1,2と同様に3000としてある。これは、これまでの経験から、集団数が増加すると同時尺度調整法による等化成績が、相当悪化することが予想されたためであり、たとえば、2.3.4のような共通項目数が2で良好な

表1　集団数2の同時尺度調整法の結果(共通項目数４)その1

データ	データA1	データA2	データA3	データA4	データA5	平均
共通項目数	4	4	4	4	4	4
全項目数	12	12	12	12	12	12
集団2の平均0.5	0.247	0.431	0.483	0.484	0.454	0.420
集団2のSD1.0	0.987	1.042	1.016	1.012	1.023	1.016
共通項目数比率	0.333	0.333	0.333	0.333	0.333	0.333

等化成績は期待できないと考えたからである。

3. 結果と考察

項目反応モデルでは、一般に尺度の平均と標準偏差の情報に関して自由度があるため本研究では、全ての分析において、能力母数と項目母数の交互同時推定の繰り返しごとに、集団1の能力分布の平均を0、標準偏差を1とするよう、すなわち標準正規分布となるよう標準化している。

3.1.1. シミュレーションA1の分析

シミュレーションA1データでは、全て集団1は標準正規分布としたため、第2集団以下の集団平均や標準偏差がシミュレーションの真値近くを再現しているか否かが垂直的等化の成否の目安となる。表 1は、データA1の集団数2の能力母数の平均値である。集団２の母集団平均値はいずれも0.5であり、能力母数と項目母数の推定に続いて母集団平均と標準偏差の推定を行って1回と数えている。推定法はいずれの場合も最尤推定を利用している。本研究の他の部分ではベイズ推定を行っている場合もあるが、特に注釈を付けない限り最尤法である。現時点で部分得点モデルの垂直的等化における交互同時推定の繰り返し数の目安は得られていないため、本研究では、とりあえず1000回以上としている。これで十分という理論的保証はないが、繰り返し推定における母数の推定値の変化の状態からは、実際には十分と思われる。

表 1は、集団数2のシミュレーションデータに於いて同時尺度調整法を適用して得られた第2集団の能力母数の平均値と標準偏差である。共通項目数が4の同時尺度調整では、第2集団の能力母数の平均値の５つのデータセットの平均は0.420と真値の0.5とは差が少しある。但し、いずれの場合も第２集団の能力母数の標準偏差は1.0に近く、５つのデータセットの平均は1.016となっていることが分かる。このシミュレーションに於ける1つの部分得点項目は、２値の正誤得点換算では5つのテスト項目となるので4項目換算では20項目分となるが、２値の同時尺度調整法の結果では藤森(1999)からは、最低8項目必要とされており、２値項目換算レベルでは、表 1の部分得点4項目は比較的多めであるが、第１集団と第２集団の能力分布の平均の差は、残念ながら、藤森(1999)ほどには、旨く再現できていないことが分かる。

3.1.2. シミュレーションA2の分析

2.3.1のシミュレーションA2データの分析(部分得点の共通項目数が5以上のシミュレーションデータの分析)表 2より、5項目の部分得点共通項目があるときは、第2集団の平均値は8つのデータセットの平均は0.372と表 1の結果と比較して第２集団の能力母数の平均値の真の値0.5よりやや悪い値となっている。表 2のデータA6、A13などの成績が悪くなっているのが目立つ。各表の最下段には各テスト項目に占める共通項目の比率を示してある。これをみると、この2つのデータセットでは、共通項目数比率が0.263，0.212などと低くなっており、これが等化成績の低さの原因の一つとなっているかもしれない。

3.1.3. シミュレーションBの分析

表 2からは、データA6、A12では、他のデータセットと比較して、その共通項目比率が1/3より小さく、これが等化に影響しているのかもしれな

表2　集団数2の同時尺度調整法の結果(共通項目数5以上)

データ	データA6	データA7	データA8	データA9	データA10	データA11	データA12	データA13	平均
共通項目数	5	5	5	5	5	5	7	7	5.5
全項目数	19	15	15	15	15	15	17	33	18
集団2の平均0.5	0.189	0.450	0.462	0.427	0.453	0.487	0.372	0.134	0.372
集団2のSD1.0	0.985	1.022	0.981	1.002	1.028	1.018	1.048	0.988	1.009
共通項目数比率	0.263	0.333	0.333	0.333	0.333	0.333	0.412	0.212	0.319

表3　集団数2同時尺度調整法の結果(共通項目数4)その2

データ	データB1	データB2	データB3	データB4	データB5	平均
共通項目数	4	4	4	4	4	4
全項目数	20	20	20	20	20	20
集団2の平均0.5	0.143	0.066	0.165	0.498	0.138	0.202
集団2のSD1.0	0.975	0.943	1.002	1.038	1.001	0.992
共通項目数比率	0.200	0.200	0.200	0.200	0.200	0.200

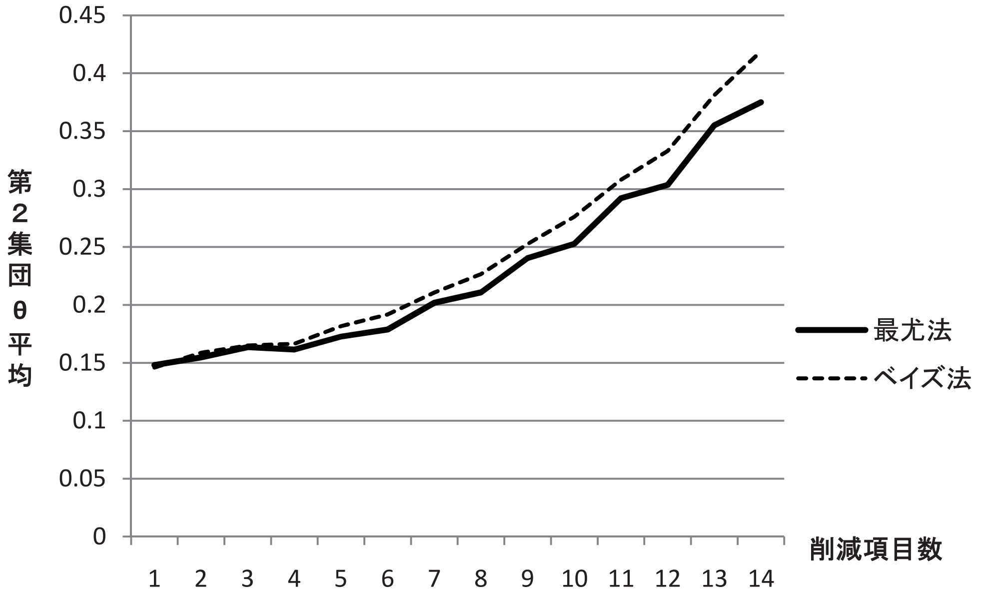

図1　　B1データのテスト項目数の削減に伴う第2集団の能力母数推定値の平均

表4　集団数2の同時尺度調整法の結果(共通項目数２)

データ	データC1	データC2	データC3	データC4	平均
共通項目数	2	2	2	2	2
全項目数	14	14	14	14	14
集団2の平均0.5	0.475	0.490	0.482	0.408	0.464
集団2のSD1.0	1.000	1.072	1.036	0.972	1.020
共通項目数比率	0.143	0.143	0.143	0.143	0.143

表5 集団数3の同時尺度調整法の結果(共通項目数10)その1

データ	データD1	データD2	データD3	データD4	データD5	平均
共通項目数	20	20	20	20	20	20
全項目数	40	40	40	40	40	40
集団2の平均0.5	0.290	0.394	0.382	0.430	0.414	0.382
集団2のSD1.0	1.009	0.611	0.606	0.653	0.628	0.701
集団3の平均1.0	0.574	0.821	0.746	0.855	0.876	0.774
集団3のSD1.0	1.208	0.901	0.886	0.979	0.961	0.987
共通項目数比率	0.500	0.500	0.500	0.500	0.500	0.500

表6 集団数3の同時尺度調整法の結果(共通項目数16+2)その1

データ	データEX1
共通項目数	16+2
全項目数	32
集団2の平均0.5	0.285
集団2のSD1.0	0.938
集団3の平均1.0	0.593
集団3のSD1.0	0.986
共通項目数比率	0.563

い。表 3は共通項目数4として共通項目以外の他の項目と合わせた全項目数が20とした時の同時尺度調整結果である。この意味は、共通項目比率が0.2と1/3より低くなっているデータセット群であるため、等化成績が悪いだろうということの確認のために行ったものである。予想通り、表 3の平均は0.202と表 1や表 2より悪くなっている。しかし、表 3のデータセットの中でも成績に大きなバラツキがあり、たとえば、データB4は、ほとんど真値0.5と近い集団平均0.498を得ているのに、データB2では、0.066と全く見当外れの値となっている。共通項目比率は等化の成績に係る一つの要因かもしれないが、共通項目数と共通項目比率だけで成績が決まるものでもないようである。共通項目比率の影響をより明確に確認するために表 3のデータB1において分析に投入する項目数を１項目ずつ削減して第２集団の能力母数の推定値の平均をグラフ化したのが図 1である。同図より、項目数の削減により、共通項目比率が高まり、第２集団の平均値が、真値0.5に接近する様子が明らかに見て取れる。同図には最尤法の結果だけでなく、能力母数のベイズ推定による結果も載せてある。明らかに最尤法よりベイズ推定の結果が良いが、ベイズ推定では、場合によって過剰に、つまり、たとえばこのシミュレーションで言えば真値0.5を超えてしまう場合もあり、その優劣の原因が現時点で不明なことから、本研究では最尤法の結果だけ報告している。

3.1.4. シミュレーションCの分析

共通項目比率は等化の成績に係る一つの要因かもしれないが、共通項目数と共通項目比率だけで成績が決まるものでもないことは、表 4の結果からも、明らかである。表 4は、共通項目数僅か2の結果である。４つのデータセットの第２集団の能力母数の平均は0.464と、真値0.5に近く、かなり良い成績を示している。表 4の共通項目数比率は、僅か0.143と表 1～表 3と比較してかなり低いにも関わらず、相対的に推定成績が良いという現象が存在することが分かった。シミュレーションの偶然から正誤の発生に表 4などで偏りが生じたのであろうか。この可能性もシミュレーションデータの中で成績の良いデータと悪いデータを比較してみると完全には否定できないが、よりシステマティックな要因が無いのかどうかは、本研究に許された時間内では、この問題の解決は困難であったため、今後の研究課題としたい。

3.1.5. シミュレーションDの分析：

最後に集団数を3とした(表 5)同時尺度調整法の結果である。集団数を1つ増やしたため、交互推定の繰り返し回数は3500回に増やしてある。共通項目数が16と多すぎる気がするが、これは隣り合う集団(たとえば集団1と2、集団2と3)のテストの共通項目数は半分の8個であり、データ全体では8+8の16個と数えたためである。表 5のデータでは隣り合う集団(たとえば集団1と2の間、集団2と3の間)にのみ共通項目を設置し、離れた集団間(たとえば集団1と3の間)には共通項目が存在しないような共通項目の設置パターンである。隣り合う集団間の等化情報に1か所でも問題があれば、多数の集団の等化は、全体としては失敗に終わるため、集団数が2の場合に比して、より多くの集団の等化は困難さが集団数に比例するというよりは、もっと等化の困難さの度合いの増加が急峻な印象がある。表 5の集団2の5つのデータセットの能力母数の平均値は0.383であるから、表 4までの集団数2の場合と比較して能力平均値の推定成績に遜色はないが、標準偏差の値が小さく推定されている問題がある。集団数2では、概ね能力母数の推定値の標準偏差は真の値の1に近いものとなっていたから、これは集団数が増えたことにより新たに生じた問題と考えて良いだろう。また表 5に於いて集団3の5つのデータセットの平均値は、0.774となっており、集団２の平均や標準偏差の推定が十分正確でない中で比較的良い値を維持している。比較的良いというのは、表 3の真値0.5に対して平均値0.202などと比較してみての話である、真の値1.0に対して0.774という値で実際の運用場面で利用可能と言っているわけではない。そして、面白いことに集団３の標準偏差は、真の値1.0に比較的近い0.983を得ており、集団2と異なり真の値1にとても近い結果となっている。集団２は下位の集団1と上位の集団3に挟まれた集団であることが、集団の標準偏差の結果に影響しているのかもしれない。２値のテストデータの分析でも、他の集団間に挟まれた集団の能力母数の推定結果に奇妙な結果が得られることは経験上何例かあるので、多値のデータに於いても類似現象があるのかもしれない。

表 6は、3.1.5の補足的分析であり、集団数３で隣接集団の8項目だけ共通項目としただけでなく、互いに離れた関係にある集団1と3に共通項目を2項目追加して同時尺度調整法を適用した結果である。第２集団と第３集団の能力母数の平均値は、表 5と比較して特に良くも悪くもないように思われる。わずか１例であるから、結論とすることはできないが、少なくともこの例では離れた集団を直接共通項目でつなぐ意味は無さそうといえる。

以上をまとめると、本研究では、2.2で述べた部分得点モデルに関して、同時尺度調整法による等化についてある程度の状況が判断できる知見を得ることができた。すなわち、同時尺度調整法による部分得点モデルの等化では、共通項目数の単純な増加のみによる等化成績の改善は観測されないことやテストの総項目数に占める共通項目の比率が等化に影響しているかもしれないことが示唆された。

文献

Andrich,D. A rating formulation for ordered response categories. *Psychometrika*,**43**,561-573.

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

藤森進　1999　算数・数学学力の到達度水準に関する発達的研究（研究課題番号08610130）　平成８年度～平成１０年度科学研究費補助金（基盤研究（C）(2)）研究成果報告書.

藤森進　2002a 項目反応理論におけるテストの部分得点の処理方法について　未発表論文

藤森進　2002b 部分得点モデルとその応用　第1回心理測定研究会.

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

Masters,G.N. 1992 A Rasch model for partial credit scoring. *Psychometrika*,47,2,147-174.

Mislevy,R.J. 1984 Estimating latent distributions. Psychometrika,**49**,359-381.
Samejima,F. 1969 Estimation of latent trait ability using a response pattern of graded scores. *Psychometric monograph*,No**17**.

［抄録］
昨年の研究には、シミュレーションデータの作成に重大な誤りがあり、その結果は正しくなかった。結論は本研究で正される。多値の項目反応モデルにおける同時尺度調整法を用いて、シミュレーションによって2つの母集団の能力母数の平均の垂直的等化を検討する。このため、藤森の部分得点モデルを利用して、多値のテストデータをシミュレーションによって作成し(２つのテストには共通する幾つかのアンカー項目が含まれている)、母集団平均と分散は、能力母数の推定値の平均と分散を利用してる。研究結果から、以下のことが示唆された。2つの母集団の能力の平均値の再現性は、比較的良く、テスト項目に占めるアンカー項目の比率が高くなると再現性は良くなることが示された。同時尺度調整法による垂直的等化でよい成績をあげるには、テスト項目数の三分の一以上の比率でアンカー項目を入れることが勧められる。