【個人研究】

同時尺度調整法による垂直的等化の検討

藤　森　　進 *

Simulation Study for Examining the Vertical Equating by Concurrent Calibration

Susumu Fujimori

Using the concurrent calibration in the item response theory, this research examines the vertical equating of the average of two populations through simulation. Simulated data were generated using 2-parameter logistic model with the vertical equating on the assumption of conditions in which an individual test was conducted for two populations having a significantly different ability (anchor items included in both tests). The population average and variance were estimated by the Mislevy 's (1984) method and the method of averaging the maximum likelihood estimates of subjects 'abilities, and compared the results of both methods. The following explains the simulation results. The Mislevy 's method showed the better results than the other method. For the reproduction of difference between the averages of two populations under conditions in which concurrent calibration was used, it was also found that the results would not become better proportionally to the number of test items increased. This finding disproves the general concept that increasing the number of test items will bring better results. Another finding is that estimated results improve proportionally to the number of anchor items between the two tests increased.

Key words: item response theory, vertical equating, concurrent calibration, maximum likelihood, EM-algorithm, simulation

1 序

現在使用されている多くの心理学テストでは、標準化がなされているため、特定のテスト結果に限ればテスト結果の比較に困難さは生じない。しかし、そのテスト結果を他のテストの結果と比較しようとしたりする場合には、たちまち困難に直面する。同一の心理学的特性を測定するテストが複数存在するときに、互いのテスト得点を適切に対応づけることを等化（equating）と呼んでいる。この研究は、テスト得点を分析する理論として古典的テスト理論にとって代わり一般的に利用されるに至った項目反応理論における等化について、藤森（1997）に続いてシミュレーションにより検討を加えようとするものである。

1.1 項目反応理論

項目反応理論（item response theory）では、テスト項目に対して被験者が正答したり誤答したりする確率は、テスト項目の困難さなどの諸特徴と被験者の能力の関数によって表現される。次式は、項目反応理論において良く利用されている２母数ロジスティックモデルである。

＊ふじもり　すすむ　文教大学人間科学部人間科学科

$$P(x_{ij}=1|\theta_i)=\frac{1}{1+\exp(-Da_j(\theta_i-b_j))} \quad (1)$$

ここでθは被験者iの能力を表す母数である。またDは1.7の定数であり、jは項目の番号を表す。また、xは正答のとき１、誤答のとき０となるダミー変数である。aは項目の識別力と呼ばれる項目母数であり、b は項目の困難度を表す項目母数である。a, b２つの項目母数を持つため、２母数ロジスティックモデルと呼ばれる。この他にも困難度のみを項目母数とするラッシュモデルや、識別力と困難度の他にあて推量に関する項目母数を導入した３母数ロジスティックモデルも一般に良く利用されている。しかしラッシュモデルは、本研究で対象とする垂直的等化（後述）に関して問題があるという報告（たとえばHolmes,1982）がある等の理由により本研究では検討の対象としない。一方、３母数ロジスティックモデルは垂直的等化で良い成績を示すという報告（Skaggs & Lissitz,1986）はあるものの、同モデルはあて推量の入り込む余地のあるテスト、例えば多肢選択形式のテストにしか適用できず、その利用は限定的なものにならざるを得ない。以上のような理由により、ここでは２母数ロジスティックモデルのみを分析の対象にする。

1.2 本研の目的

等化の必要な状況には様々なケースがあるが、ここでは芝（1978）の語彙理解力の研究や藤森（1991）、藤森・中野（1994）などの算数・数学学力の研究で問題となるようなタイプの等化について検討する。これらの研究では、集団が異なることによる能力水準の違いと各集団に実施されるテストの違いが同時に存在している。この種の状況における等化を垂直的等化（vertical equating）と呼んでいる。垂直的等化を実施するための研究デザインとして、ここでは、芝の研究などで利用されている共通項目デザインを取り上げる。すなわち各集団に実施するテストに共通項目を置いて実施するデザインを研究の対象にする。ただし教育内容の履修順序の存在も考慮して下位集団のテスト項目を共通項目として利用するケースをここでは想定する。異なる２つのテスト結果を共通尺度にのせる等化方法としては、 藤森（1997）と同様にテストの項目及び被験者の母数を一度に推定して共通尺度上に表現し、等化の作業をこの過程に織り込んでしまう同時尺度調整法（concurrent calibration）を取り上げる。他の方法については、例えばPetersen, Kolen, & Hoover（1989）などを参照されたい。

本研究では、藤森（1997）で十分には取り上げられなかった要因について更に様々なケースを想定して分析を加えた。すなわちシミュレーションＡでは、①被験者能力の母集団分布のMislevy（1984）の方法による推定と項目母数のＥＭアルゴリズムによる推定を交互に行う方法と、②能力母数の母集団分布を能力母数の最尤推定値の平均・分散によって推定することと項目母数のＥＭアルゴリズムによる推定を交互に行なう方法との比較検討を行なう。シミュレーションＢでは、①の方法におけるテスト項目数の影響を検討し、シミュレーションＣでは、①の方法に関してテスト項目の識別力と困難度の分布や、能力母数の平均の集団間の差の大きさや能力母数の母集団分布の標準偏差が垂直的等化に与える影響などを検討する。

2 母数の推定方法

2.1 項目母数の推定及び能力母数の母集団分布に関する推定

Birnbaum（1968）による項目母数と能力母数の同時最尤推定法は項目反応理論で古くから利用されてきた。しかし同時最尤推定法の項目母数の推定の一致性に関する問題点が指摘され、これを克服するため　Bock & Aitkin（1981）により採用されたＥＭアルゴリズムを利用した周辺最尤推定による項目母数の推定法は、実用性もあったため普及してきており、本研究でもこの方法を利用している。

能力母数の母集団分布の推定は、ＥＭアルゴ

リズムにより項目母数の推定値が得られた後に、つまり項目母数を所与として行っている。能力母数の母集団分布の推定方法にも様々なものがあるが、本研究では、主たる方法として Mislevy（1984）の方法を利用している。Mislevy の方法では、母集団分布として正規分布を仮定し、平均μと分散σ^2の推定をする際にθを介さず正誤データから直接μとσ^2を推定する。いまθが平均μ、分散σ^2の正規分布 h に従っていると仮定する。N 人のテストデータが得られたとき、能力母数θを積分によって排除したμ、σ^2の周辺化された尤度は、

$$L(\mu,\sigma|x_1,x_2,...,x_N)=\prod_{i=1}^{N}\left\{\int_{-\infty}^{\infty} f(x_i|\theta)\,h(\theta|\mu,\sigma)\,d\theta\right\} \quad (2)$$

で表される。ここで f は

$$f(x_i|\theta)=\prod_{j=1}^{n} P_j^{x_{ij}}(1-P_j)^{1-x_{ij}} \quad (3)$$

であり、x_i は被験者 i の正誤を表す n 個の成分x_{ij}からなる列ベクトル、nはテストの項目数である。また式中のP は、（1）の正答確率である。L を最大にするμとσ^2を数値計算により求め、母集団平均と分散の推定値とするのがMislevyの方法の骨子である。なお同方法については前川（1991）を参照されたい。

またMislevyの方法との比較のため項目母数所与として、各被験者の能力母数の最尤推定値を求め、その平均と分散を計算し、これを母集団分布の平均値と分散の推定値とする方法も試みている。これは藤森（1997）で行なった方法である。

2.2 プログラム

母数の実際の推定は、Bock & Aitkinによる前記のＥＭアルゴリズム及びMislevyの方法（あるいは能力母数の最尤推定値の平均、分散で母集団平均を推定する方法）を採用した自作のFORTRANプログラムによって行っている。 同プログラムでは、各集団の能力母数の分布を推定しながら項目母数との交互推定を行っている。ＥＭアルゴリズムによる推定は比較的収束が遅いことが知られているためＥＭステップと母集団分布の推定との繰り返しは最低100回以上としている。なお、この他に尤度関数による収束基準も定めているが、くり返し毎の変化量が小さいため、実質的には最大繰り返し数が計算の打ち切り基準として機能している。

3 シミュレーション

3.1 シミュレーションデータの作成

この研究では複数のシミュレーションを行なうのでここでは全てのシミュレーションにかかわる事、あるいは代表的なケースに関する事を述べ、各シュミレーションに固有のことについては該当する節で述べる事にする。被験者の項目に対する正誤は、（1）の2母数ロジスティックモデルに従っていると仮定する。等化には、テストの項目数、被験者数、集団間の能力差の大きさ、テスト間の共通項目の数など様々な要因が関与している。これらの要因についてこの研究で全てとり上げることは困難であり、かなり限定を加える必要がある。まず初めに、集団数は２とする。集団数が３以上のデータを等化する場合は、能力尺度上で中位の集団が上位・下位の両方の集団のデータと共通部分を持ち得るという点で集団数が 2 の場合と状況が異なるものの、集団の数が 2という状況は等化の最も基本的なものであることを考慮し、この設定としている。シミュレーションにおける標準的な条件としては、下位の集団の能力母数は、平均－0.25、標準偏差1の正規分布に従い、上位の集団の能力母数は平均0.25、標準偏差１の正規分布に従うとする。集団間の差の大きさの決め方は恣意的と感じる向きもあるかもしれないが、芝（1978）の研究や、藤森（1991）などにおける算数・数学学力の分析結果から考えて0.5標準偏差という値は大きくもなく小さくもない水準であると考えている。擬似的な被験者の人数は1集団あたり300、500あるいは1500人とする。各集団毎に難易

度水準の異なる別々のテストを実施することを想定するが、その項目数は、標準で20〜60であり、最大120項目までとする。また2つのテストの共通項目数は2,4,6,8,10,12の6通りとする。下位集団のテスト項目の困難度母数は、平均−0.25、標準偏差0.5の正規分布に従い、上位集団のそれは平均0.25、標準偏差0.5の正規分布にそれぞれ従うと仮定する。識別力母数は、どちらの集団も平均0.85、標準偏差0.25、下限0.3、上限2.0の切断正規分布に従うとする。共通項目は、先に述べたように下位集団の分布に従って項目母数が決められている。

以上のようにして２母数ロジスティックモデルの全ての母数を定めた後に、被験者 i と項目 j を固定して、０から１の一様乱数 R を作り、これと（1）の値を比較して R が大きければ誤答、小さければ正答とし、以下同様の手続きによって全ての正誤パターンを決定する。このようにして被験者及び項目の各組み合わせについてデータを作成し１つのデータセットとする。被験者及び項目母数を生み出す分布を確定する母数（ベイズ流に言えば超母数）の組み合わせ各々について、異なる乱数系列の下で10回繰り返してデータセットを作成した。この繰り返しでは、被験者や項目の母数まで含めて再度発生させている点に注意されたい。被験者と項目の各組み合わせごとに10回というシミュレーションの繰り返し回数は少なく感じられるかもしれないが、計算時間という制約を考慮してこの回数に決定した。参考までに1回の計算時間は、200メガヘルツのペンティアムプロセッサを利用しても、短いもので1時間、長いものでは8時間程度かかっている（データの大きさによって異なる）。

3.2 シミュレーションA

3.2.1 シミュレーションAの目的と方法

シミュレーションＡは、各母集団の平均と分散を推定する方法について検討する。すなわち各被験者の能力母数の推定値を得てその平均と分散を計算し、これを母集団平均と分散の推定値とする方法と、尤度から直接的に被験者母集団の能力分布の平均と分散を推定するMislevy（1984）の方法の比較を行なう。

3.2.2 シミュレーションAの結果と考察

表 1〜表 3は、標準的条件におけるシミュレーションの結果である。すなわち Mislevy の方法による集団あたり300、500、及び1500人の被験者、項目数20、40、及び60項目のテストを分析した結果である。項目母数及び能力母数も3.1節で述べた標準的条件にしたがって作成されたものである。

さて本研究の直接的な関心対象は集団ごとの能力母数の平均値の差である。シミュレーションの結果は膨大なものであることもあり、以下のようなものにまとめた（表１以下を参照）。各データセットにおける結果とも下位集団の能力母数の平均を０、分散を１と標準化する。2集団間の母集団平均の推定値の差を δ_{12} とし、各集団の被験者の真値の平均を求めその差を Δ_{12} とするとき、表中の平均値と記されている欄は $\gamma = \delta_{12}/\Delta_{12}$ の10回の平均である（以下同様）。この γ が１に近ければ近いほど等化の精度が高いことになる。SD、レンジ及びMSEも10回のくり返し毎の γ 値に基づいて計算されたものである。ただし、SDは不偏分散の正の平方根であり、MSEとは以下のような指標である。

$$MSE = \frac{1}{10}\Sigma\left(1-\frac{\delta_{12}}{\Delta_{12}}\right)^2 \qquad (4)$$

なおΣは、10回のシミュレーションに関する和を表す。

表1より、先ず共通項目数の影響を見てみよう。人数300人の場合、テスト項目数が20、40、そして60項目いずれの場合も、共通項目数が2から12に増加するに従って推定成績、特にMSEの結果に改善が見られる。表 ２及び表３より人数が500及び1500の場合も同様の傾向が明確である。続いてテスト項目数の影響を見てみよう。人数300人の場合、テスト項目数が20、40、そして60項目に増加して

表1　Mislevyの方法（被験者数300人）

テスト		共通項目数					
項目数		2	4	6	8	10	12
20項目	平均	0.925	0.910	0.967	1.021	0.980	0.982
	ＳＤ	0.181	0.134	0.088	0.162	0.049	0.088
	レンジ	0.432	0.394	0.328	0.545	0.137	0.229
	ＭＳＥ	0.035	0.024	0.008	0.024	0.003	0.007
40項目	平均	0.955	0.966	0.974	0.964	1.068	1.042
	ＳＤ	0.305	0.143	0.211	0.084	0.140	0.100
	レンジ	1.046	0.475	0.619	0.238	0.497	0.318
	ＭＳＥ	0.086	0.020	0.041	0.008	0.022	0.011
60項目	平均	0.922	0.858	0.906	0.880	0.988	0.967
	ＳＤ	0.247	0.351	0.183	0.174	0.124	0.134
	レンジ	0.811	1.104	0.606	0.591	0.398	0.445
	ＭＳＥ	0.061	0.131	0.039	0.042	0.014	0.017

表2　Mislevyの方法（被験者数500人）

テスト		共通項目数					
項目数		2	4	6	8	10	12
20項目	平均	1.085	0.994	0.941	1.032	1.017	1.005
	ＳＤ	0.185	0.146	0.097	0.075	0.059	0.078
	レンジ	0.612	0.506	0.263	0.195	0.181	0.254
	ＭＥＳ	0.038	0.019	0.012	0.006	0.004	0.006
40項目	平均	0.912	1.002	1.015	1.003	0.983	0.974
	ＳＤ	0.190	0.153	0.112	0.086	0.080	0.076
	レンジ	0.603	0.486	0.411	0.312	0.243	0.227
	ＭＳＥ	0.040	0.021	0.012	0.007	0.006	0.006
60項目	平均	0.849	0.997	1.015	0.966	0.949	0.995
	ＳＤ	0.276	0.139	0.090	0.097	0.092	0.049
	レンジ	0.775	0.378	0.345	0.289	0.304	0.157
	ＭＳＥ	0.091	0.017	0.008	0.010	0.010	0.002

表3 Mislevyの方法（被験者数1500人）

テスト項目数		共通項目数 2	4	6	8	10	12
20項目	平均	1.036	1.016	0.997	1.009	1.015	1.025
	ＳＤ	0.156	0.045	0.063	0.019	0.053	0.027
	レンジ	0.448	0.155	0.213	0.060	0.172	0.073
	ＭＳＥ	0.023	0.002	0.004	0.000	0.003	0.001
40項目	平均	0.915	1.044	0.976	0.983	1.022	1.013
	ＳＤ	0.146	0.065	0.055	0.032	0.049	0.050
	レンジ	0.407	0.225	0.149	0.104	0.124	0.160
	ＭＳＥ	0.027	0.006	0.003	0.001	0.003	0.003
60項目	平均	0.951	0.961	0.954	0.978	0.990	1.008
	ＳＤ	0.235	0.120	0.063	0.045	0.042	0.053
	レンジ	0.777	0.360	0.219	0.140	0.150	0.160
	ＭＳＥ	0.052	0.014	0.006	0.002	0.002	0.003

も必ずしも良くなるという事はない。人数が500、1500の場合も同様である。次に人数の影響を見てみよう。人数が300から500に増加してもそれほど成績の差は明瞭ではないが、人数が1500では推定成績の改善がある程度見られる。

表４～表６は、Mislevyの方法の代わりに、個々の被験者の能力母数を最尤推定し、その平均及び分散を求めて母集団平均及び分散の推定値としたものである。共通項目数、テスト項目数及び人数の影響は表 1～表 3とほぼ同様の結果と言える。表１～表３と表４～表６のMSEを比較するに、人数300の5つのケース及び500の2つのケースを除いた大多数のケースではMislevyの方が良い結果を与えている事が分かる。この結果より個々の被験者の能力母数の推定値に基づいて母集団平均及び分散の推定を行うよりもMislevyの方法の方が良い事は明らかであろう。

3.3 シミュレーションＢ

3.3.1 シミュレーションＢの目的と方法

一般に測定精度の観点からは被験者に実施するテスト項目数は多ければ多いほど良いと考えられている。しかし前節に示したシミュレーションの結果からは必ずしも項目増の効果は明瞭でなかった。3.2.2節の結果よりMislevyの方法を利用する事の優位性は明らかと判断し、本節では同方法に限ってテスト項目数を更に80、100、そして120と増加させ、その等化に対する影響を検討する事にする。

3.3.2 シミュレーションＢの結果と考察

表７～表９は、Mislevyの方法を利用したときのテスト項目数が80、100、120の場合のシミュレーション結果である。テスト項目数以外の他の条件は表１～表３と同一である。

表７～表９より、共通項目数の増加及び被験者人数の増加につれてMSEに改善傾向が見られる点は前節の結果と同様である。さてテスト項目数の影響を考えるために表１と表７

表4　θの最尤推定値の平均（被験者数300人）

テスト		共通項目数					
項目数		2	4	6	8	10	12
20項目	平均	0.652	0.780	0.796	0.804	0.840	0.914
	ＳＤ	0.167	0.178	0.103	0.129	0.108	0.115
	レンジ	0.533	0.604	0.356	0.417	0.305	0.308
	ＭＳＥ	0.146	0.077	0.051	0.053	0.036	0.019
40項目	平均	0.816	0.871	0.888	0.931	0.926	0.906
	ＳＤ	0.152	0.159	0.072	0.119	0.116	0.100
	レンジ	0.440	0.458	0.245	0.330	0.388	0.373
	ＭＳＥ	0.055	0.039	0.017	0.018	0.018	0.018
60項目	平均	0.855	0.979	0.808	0.870	0.862	0.914
	ＳＤ	0.319	0.197	0.167	0.085	0.135	0.138
	レンジ	1.117	0.612	0.491	0.284	0.393	0.424
	ＭＳＥ	0.113	0.036	0.062	0.023	0.036	0.025

表5　θの最尤推定値の平均（被験者数500人）

テスト		共通項目数					
項目数		2	4	6	8	10	12
20項目	平均	0.655	0.824	0.851	0.834	0.869	0.876
	ＳＤ	0.142	0.130	0.129	0.117	0.061	0.065
	レンジ	0.490	0.375	0.436	0.430	0.214	0.198
	ＭＳＥ	0.137	0.046	0.037	0.040	0.021	0.019
40項目	平均	0.706	0.772	0.914	0.940	0.895	0.950
	ＳＤ	0.128	0.154	0.135	0.065	0.080	0.069
	レンジ	0.401	0.532	0.380	0.201	0.265	0.238
	ＭＳＥ	0.101	0.074	0.024	0.007	0.017	0.007
60項目	平均	0.772	0.802	0.906	0.936	0.962	0.928
	ＳＤ	0.327	0.170	0.121	0.062	0.085	0.089
	レンジ	1.041	0.472	0.343	0.197	0.260	0.249
	ＭＳＥ	0.149	0.065	0.022	0.008	0.008	0.012

表6　θの最尤推定値の平均（被験者数1500人）

テスト項目数		共通項目数 2	4	6	8	10	12
20項目	平均	0.645	0.787	0.821	0.827	0.843	0.887
	ＳＤ	0.095	0.078	0.045	0.057	0.046	0.069
	レンジ	0.327	0.256	0.152	0.204	0.145	0.250
	ＭＳＥ	0.134	0.051	0.034	0.033	0.026	0.017
40項目	平均	0.795	0.865	0.875	0.890	0.903	0.883
	ＳＤ	0.119	0.047	0.061	0.074	0.052	0.051
	レンジ	0.417	0.130	0.188	0.258	0.148	0.142
	ＭＳＥ	0.055	0.020	0.019	0.017	0.012	0.016
60項目	平均	0.765	0.831	0.934	0.908	0.883	0.918
	ＳＤ	0.197	0.086	0.049	0.053	0.052	0.028
	レンジ	0.658	0.292	0.170	0.161	0.173	0.087
	ＭＳＥ	0.090	0.035	0.007	0.011	0.016	0.007

の結果を比較すると、テスト項目数が増加するにつれて推定成績の改善が見られるどころか逆に悪化する傾向が認められる。特に共通項目数が2〜4の場合、かなり悪化傾向が明瞭であり、MSEが0.1より大きい値となっている。また表３及び表９に示された被験者人数1500で、共通項目数が12項目の結果の中でMSEが0.01を越えてしまっているのは、テスト項目数20〜120項目の6通りの内で120項目の場合だけである。同時尺度調節法による垂直的等化では、テスト項目数の増加は必ずしも母集団平均間の差の推定に良い影響をもたらさず、かえって悪影響が生じる場合がある事が示されたと言えよう。

3.4 シミュレーションC

3.4.1 シミュレーションCの目的と方法

本節でもMislevyの方法に限って以下のような検討を行なった。検討する要因として初めに、能力分布の母集団平均の差を取り上げる。3.1節ではその大きさを0.5としたが、これを0.75へ変更した表10の結果ではどのように違いが生じるかを検討する。もちろんその差の大きさだけでなく、差の変更の仕方にも色々な方法があるが、本研究では、下位集団の能力分布の母集団平均を－.375、上位のそれを.375とした。項目困難度は、そののままであるから下位集団にとっては項目がやや難しくなり、上位集団にとってはやや易しくなったことになる。このような変更は共通項目における正答率の群間差に影響を与えるであろうから、3.1節の結果より両集団の弁別自体は良くなるであろう。しかし能力分布の平均の差の再現率はどうなるかは不明であるし、集団間の距離が遠くなれば等化が困難となることも予想される。続いて第2の要因には、能力分布の母集団の標準偏差を取り上げる。3.1節では1.0としたが、これを1.25へ変更した表11の結果が表３とどのように違いが生じるかを検討する。この変更は、上位群の能力母数の標準偏差が大きくなる事を意味するが、数学などにおいて小学生から中学生に

表7　テスト項目数80～120（被験者数300人）

テスト項目数		共通項目数 2	4	6	8	10	12
80項目	平均	0.801	0.906	0.940	1.031	0.930	1.036
	ＳＤ	0.294	0.431	0.255	0.184	0.123	0.134
	レンジ	0.972	1.130	0.950	0.563	0.426	0.426
	ＭＳＥ	0.117	0.176	0.062	0.031	0.019	0.017
100項目	平均	0.823	1.059	0.978	0.946	0.926	0.964
	ＳＤ	0.646	0.560	0.329	0.220	0.288	0.162
	レンジ	2.046	1.604	1.035	0.608	0.940	0.481
	ＭＳＥ	0.407	0.286	0.098	0.047	0.080	0.025
120項目	平均	0.522	0.752	0.914	0.793	1.085	1.010
	ＳＤ	0.426	0.508	0.304	0.274	0.311	0.154
	レンジ	1.584	1.548	0.903	0.906	0.951	0.500
	ＭＳＥ	0.392	0.294	0.091	0.111	0.094	0.021

表8　項目数80～120（被験者数500人）

テスト項目数		共通項目数 2	4	6	8	10	12
80項目	平均	0.718	1.025	0.956	0.982	0.904	0.980
	ＳＤ	0.321	0.277	0.080	0.119	0.069	0.117
	レンジ	0.926	0.720	0.252	0.347	0.188	0.333
	ＭＳＥ	0.172	0.070	0.008	0.013	0.014	0.013
100項目	平均	0.952	0.975	0.927	1.013	0.951	0.937
	ＳＤ	0.592	0.272	0.183	0.180	0.087	0.092
	レンジ	1.720	0.746	0.526	0.479	0.300	0.266
	ＭＳＥ	0.318	0.067	0.036	0.029	0.009	0.012
120項目	平均	0.544	0.777	0.938	0.857	0.928	0.945
	ＳＤ	0.895	0.379	0.232	0.347	0.111	0.178
	レンジ	3.207	1.311	0.766	1.182	0.321	0.564
	ＭＳＥ	0.930	0.179	0.052	0.129	0.016	0.032

表9　テスト項目数80～120（被験者数1500人）

テスト項目数		共通項目数 2	4	6	8	10	12
80項目	平均	0.770	0.925	0.974	0.965	0.987	0.992
	ＳＤ	0.255	0.135	0.091	0.082	0.062	0.079
	レンジ	0.850	0.501	0.328	0.253	0.205	0.254
	ＭＳＥ	0.111	0.022	0.008	0.007	0.004	0.006
100項目	平均	0.700	0.839	0.902	0.947	0.955	0.953
	ＳＤ	0.185	0.189	0.121	0.095	0.067	0.061
	レンジ	0.595	0.641	0.392	0.286	0.217	0.159
	ＭＳＥ	0.121	0.058	0.023	0.011	0.006	0.006
120項目	平均	0.651	0.869	0.974	0.920	0.951	0.931
	ＳＤ	0.334	0.237	0.104	0.087	0.116	0.092
	レンジ	1.019	0.853	0.366	0.276	0.422	0.260
	ＭＳＥ	0.222	0.068	0.010	0.013	0.015	0.012

なるにつれて「落ちこぼれ」などの表現に伺えるような集団内の学力の差の拡大が現れる事は良く知られている現象である（例えば中垣（1985）や黒須（1987）など）。この意味で本シミュレーションは現実場面との関わりも大きい課題を検討するものである。更に第3の要因には、項目識別力を作り出す分布の平均を取り上げ、その平均を3.1節の0.85から0.6に低下させた場合にどのような影響が生じるかを検討する（表12）。他の要因を一定にして識別力の平均を低下させれば、共通項目もその例外ではないのだから、能力分布の平均の差の再現に悪影響があるものと予想されるが、その程度について検討する。最後の要因としては、項目困難度を作り出す分布の標準偏差を取り上げ、その値を3.1節の0.5から0.25にした場合の影響を検討する（表13）。他の要因を3.1節と変更せずに項目困難度の分布の標準偏差を小さくする事は、ピーク型テスト（テスト情報量がθ軸上のある点で高くなり離れるに従って速やかに低下する）と非ピーク型テストのどちらが能力分布の平均の差の再現に有効かを判断する一つの手がかりになる。少なくとも共通項目については能力差のある２つの集団に実施するわけであるから両集団をある程度カバーできる項目反応曲線を持ったものでなければ等化はうまく行かないであろう。この意味では非ピーク型の場合、たまたま２つの集団をカバーできるような項目（本シミュレーションでは下位集団のテストの難しい項目）が共通項目として選択されれば等化の成績は相対的に良くなるであろうし、逆に下位集団の易しい問題が共通項目として選択されれば等化の成績はかなり悪化するであろう。ピーク型では等化の成績の絶対的水準がどうなるかは不明としても、推定成績のバラツキは非ピーク型より小さくなる傾向を持つと予想される。もちろんピーク型か否かによってだけでなく、等化の成績には、他の要因、例えばテスト項目の困難度の分布する範囲と能力母数の分布範囲の関係なども結果に影響するであろうが、本シ

ミュレーションで設定したような被験者の能力分布とテストとの組み合わせでは、困難度の分布の標準偏差0.25は、能力分布の標準偏差1に比較してかなり小さく、能力分布の平均に合わせたピーク型テストと見なして良いだろう。

3.4.2 シミュレーションCの結果と考察

本節で検討された要因についてのシミュレーションは、いずれも限定的なものであり十分とは言えないが参考にはなるであろう。

まず第1に取り上げた能力分布の母集団平均の差を拡大した影響であるが、表 10の結果を見ると全体的にMSEが表３と比較して小さく、特に共通項目数が少ない場合の成績が良くなっている。２つの群の能力平均値の差が大きくなったため、分離し易くなった事が効いていると思われる。なお本節での結果の表示は紙面の都合上被験者数1500の場合に限る事にする。第2に取り上げた上位集団の能力分布の標準偏差を大きくした影響であるが、表11と表３を比較してみると、MSEの成績はまちまちで特に目立った傾向は認められない。すなわち上位集団の標準偏差が大きくなっても影響は特に認められなかった。第3に取り上げたテスト項目の識別力を低下させた場合の影響についてであるが、表12に示したように表 3と比較して、共通項目数が少ない場合の相対的な成績の優劣はあるものの、共通項目数が増加した場合には表3と同様に概ね良い成績を示し、識別力のこの程度の低下が等化に特に悪い影響を与えているとは言えない。最後に取り上げたものがピーク型のテスト条件にした場合の影響を探るものであるが、表13と表3の比較から、共通項目数が2の場合のMSEの改善が目立つ。すなわち、ピーク型にした方が、成績は良くなっているようである。ただしこのことが成り立つためには当然のことながら、テストのピークと測定対象となる集団の平均とが近いものであることが必要だろう。推定成績のバラツキは共通項目数が2のときは、予想通りSDあるいはレンジなどが小さくなっているが、共通項目数が増えると絶対的な誤差が小さい事もあり、SDなどが小さくなっているとは言えない。

4 藤森（1997）の結果との比較及び全体的考察

藤森（1997）のシミュレーションでは集団間の学力差が全体的に過小推定になっていたが、本研究の結果からはMislevyの方法で母集団分布の推定をする場合は共通項目数が少ない場合も含めそのようなことがないことがわかった。ただし、傾向として共通項目数が多くなるにつれて等化の成績が改善する事自体は同様である。個々の被験者の最尤推定値に基づいて母集団分布を推定する方法に関して、藤森（1997）の結果では、γの平均は、共通項目数が２のときは0.3～0.4程度であり、共通項目数が6～8で0.7程度となり、共通項目数が12程度で0.8近くと、いずれも表４～表６に示したように、やはり1より小さくなる傾向はあるものの今回の結果の方が良い成績となっている。これは、以前の研究ではＥＭアルゴリズムを用いた項目母数の推定の際に、能力分布を１つの標準正規分布と仮定していたのに対し、今回は2つの母集団の分布を考慮した推定であった事が両者の差が生まれた主たる原因であろう。

藤森（1997）では必ずしも明確ではないとしていたテスト項目数の増加の効果は、80項目を越えてくると、少なくとも垂直的等化に関しては一般的な期待に反して害になる場合があることが本研究では示された。テストの項目数が大きいほど等化の成績が良いことをVale（1986）は報告しているが、本研究のように能力的に等質でない集団を同時尺度調節法で垂直的に等化するときには、複数のテスト結果をリンクする情報と個々のテストの情報の量的バランスが必要な事を本研究の結果は示唆しているものと思われる。また、この問題とＥＭアルゴリズムの収束の緩慢性を考え合わせると、計算の繰り返し回数を、例えば10～100倍にしたならば、テスト項目数120などのように今回の研究で良い成績を出せな

表10　平均差0.75（被験者数1500人）

テスト		共通項目数					
項目数		2	4	6	8	10	12
20項目	平　均	1.044	1.023	0.999	0.983	0.990	1.000
	ＳＤ	0.069	0.064	0.038	0.048	0.031	0.025
	レンジ	0.222	0.183	0.115	0.184	0.100	0.084
	ＭＳＥ	0.006	0.004	0.001	0.002	0.001	0.001
40項目	平均	0.986	1.000	0.985	0.996	0.997	1.020
	ＳＤ	0.095	0.058	0.058	0.028	0.030	0.040
	レンジ	0.277	0.188	0.194	0.081	0.099	0.105
	ＭＳＥ	0.008	0.003	0.003	0.001	0.001	0.002
60項目	平均	0.925	0.981	0.978	0.969	1.001	0.976
	ＳＤ	0.134	0.085	0.068	0.048	0.042	0.055
	レンジ	0.341	0.291	0.260	0.136	0.138	0.175
	ＭＳＥ	0.022	0.007	0.005	0.003	0.002	0.003

表11　上位集団θの標準偏差1.25（被験者数1500人）

テスト	共通項目数						
項目数		2	4	6	8	10	12
20項目	平均	0.965	0.987	0.984	1.007	0.986	0.980
	ＳＤ	0.118	0.054	0.046	0.056	0.044	0.035
	レンジ	0.333	0.159	0.126	0.169	0.136	0.115
	ＭＳＥ	0.014	0.003	0.002	0.003	0.002	0.002
40項目	平均	0.940	0.951	0.961	0.967	0.994	0.992
	ＳＤ	0.105	0.056	0.044	0.060	0.045	0.071
	レンジ	0.327	0.199	0.175	0.172	0.137	0.203
	ＭＳＥ	0.014	0.005	0.003	0.004	0.002	0.005
60項目	平均	0.803	0.956	0.965	0.972	0.951	0.979
	ＳＤ	0.191	0.074	0.115	0.102	0.064	0.032
	レンジ	0.547	0.227	0.412	0.270	0.185	0.122
	ＭＳＥ	0.072	0.007	0.013	0.010	0.006	0.001

表 12　識別力0.6（被験者数1500人）

テスト項目数		共通項目数 2	4	6	8	10	12
20項目	平均	1.007	1.068	1.013	1.011	0.990	1.001
	ＳＤ	0.124	0.089	0.106	0.049	0.041	0.051
	レンジ	0.345	0.296	0.294	0.161	0.141	0.156
	ＭＳＥ	0.014	0.012	0.010	0.002	0.002	0.002
40項目	平均	0.915	0.967	1.017	1.024	1.013	1.021
	ＳＤ	0.156	0.074	0.060	0.051	0.037	0.036
	レンジ	0.580	0.247	0.194	0.169	0.113	0.102
	ＭＳＥ	0.029	0.006	0.004	0.003	0.001	0.002
60項目	平均	0.929	0.995	0.994	0.958	0.981	1.000
	ＳＤ	0.075	0.064	0.068	0.064	0.066	0.078
	レンジ	0.267	0.191	0.205	0.175	0.211	0.260
	ＭＳＥ	0.010	0.004	0.004	0.006	0.004	0.006

表13　ピーク型テスト（被験者数1500人）

テスト項目数		共通項目数 2	4	6	8	10	12
20項目	平均	0.996	1.006	0.983	0.984	1.004	0.977
	ＳＤ	0.096	0.073	0.048	0.041	0.060	0.041
	レンジ	0.318	0.194	0.143	0.135	0.225	0.129
	ＭＳＥ	0.008	0.005	0.002	0.002	0.003	0.002
40項目	平均	0.954	0.948	0.984	0.996	1.005	1.013
	ＳＤ	0.100	0.061	0.050	0.050	0.040	0.067
	レンジ	0.294	0.185	0.176	0.160	0.121	0.175
	ＭＳＥ	0.011	0.006	0.003	0.002	0.001	0.004
60項目	平均	0.972	0.979	0.948	0.974	0.962	1.016
	ＳＤ	0.104	0.085	0.094	0.076	0.065	0.041
	レンジ	0.358	0.284	0.322	0.259	0.208	0.142
	ＭＳＥ	0.011	0.007	0.011	0.006	0.005	0.002

かったケースの成績も改善される可能性もある。しかし、それでは余りに時間がかかりすぎ実用上差し支えるかもしれないし、収束が遅いという問題点を持つことには変わりがない。またValeは、同時に共通項目数が２でも良いのではないかとしている。２という項目数は、等化する尺度の原点と分散を決定するのに最低限の数となるが、それでは十分でなく６～８個の共通項目が、（本研究が扱ったような条件下で行なわれる）等化では必要な事が本研究では示唆された。なお項目母数の推定の問題、及び平均と並んで母集団分布のもう一つの母数である分散の推定については、紙数の都合により全く触れる事が出来なかったが機会を改めて報告することにしたい。

文献

Birnbaum,A. 1968 Some latent trait models and their uses in inferring an examinee 's ability. In F.M.Lord and M.R.Novick, *Statistical theories of mental test scores* (pp.397-479). Reading, MA: Addison-Wesley.

Bock, R. D. and Aitkin, M. 1981 Marginal maximum likelihood estimation of item parameters: An application of an EM algorithm. *Psychometrika*, 46, 443 - 459.

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

藤森　進 1997 同時尺度調節法による垂直的等化のシミュレーションによる検討　岡山大学教育学部研究集録,106, 173 - 177.

藤森　進・中野和代　1994　中学生の数学学力尺度の作成　岡山大学教育学部研究集録, 96, 115 - 121.

Holmes,S.E. 1982 Unidimensionality and vertical equating with the Rasch model. *Journal of Educational Measurement,*19,2,139-147.

黒須　俊夫 1987 小学生の算数学力の進展と遅滞（Ⅱ）―計算（その２）― 宮城教育大学紀要,22,61-83.

前川　眞一 1991 パラメタの推定　芝祐順（編）「項目反応理論」第4章 東京大学出版会.

Mislevy,R.J. 1984 Estimating latent distributions. *Psychometrika*, 49, 359 - 381.

中垣　啓 1985 計算問題より見た児童・生徒の学力と発達 国立国語教育研究所研究集録,11,51-64.

Petersen,N.S., Kolen,M.J., & Hoover,H.D. 1989 Scaling, Norming, and Equating. In Linn, R. L. (ed.). *Educational Measurement* (3rd ed.), Macmillan. （池田ほか（編訳）1992 「教育測定学」第3版　みくに出版）

芝　祐順 1978 語彙理解力尺度作成の試み 東京大学教育学部紀要,17,47-58.

Skaggs, R. K. and Lissitz,R.W. 1986 An exploration of the robustness of four test equating models. *Applied Psychological Measurement,* 10, 3, 303 - 317.

Vale, C. D. 1986 Linking item parameters onto a common scale. *Applied Psychological Measurement,* 10, 4, 333 - 344.

本研究は文部省科学研究費補助金（基盤研究Ｃ）課題番号08610130研究代表者藤森進による研究の一部である。