「被験者の推定尺度値を利用した潜在特性尺度 等化法」のシミュレーションによる検討

野 口 裕 之*

SIMULATION STUDY OF A NEW EQUATING METHOD
IN THE LATENT TRAIT MODEL

Hiroyuki NOGUCHI

An equating method of two scales based upon two separate tests using common subjects was proposed in the previous paper (Noguchi, 1983). In that method, each subject's scale values are estimated on the two separate scales, and then the principal component for the whole group of examinees is extracted, which corresponds to the equating equation.

In this paper, using computer simulation, the validity of Noguchi's method is examined in various situations, changing the difficulty levels of the two tests, the distribution of subjects and the number of subjects.

As a consequence, Noguchi's equating method proved to be valid in situations where the two tests are not so different in difficulty levels and the subjects' ability distributes for a wide range.

1. 尺度の等化

1.1 尺度を等化する必要性

教育心理測定では異なる尺度で測定された結果を相互に比較する必要がある場合が多い。例えば、大学入学者選抜共通第1次学力試験では、理科については「物理 I」、「化学 I」、「生物 I」及び「地学 I」のうちから2科目を試験室で選択解答、又は「基礎理科」1科目を解答する。(昭和59年度大学入学者選抜共通第1次学力試験受験案内) ことになっているが、この場合には合否の判定にあたって異なる科目を選択した受験者について、同じ理科の得点として相互に比較することになる。また、数は少ないが追試験の受験者について、その得点を本試験の受験者の得点と相互に比較する必要も出てくる。

また、TOEFL や TOEIC など、わが国で広く知られている英語力に関するテストは年に複数回実施されるが、いずれの回に受験しても結果は相互に比較可能なように同じ尺度上に表わされる必要がある。

このように、異なるテストから得られた結果を相互に比較可能なように尺度(得点)を調整することを尺度の「等化」と呼ぶ。そして、今述べた例はいずれも等化する各テストの受験者間で能力水準にそれ程大きな違いがあると考えられない場合で、特に「水平的等化」と呼ぶ。

さらに、実用的なテスト場面だけではなく、心理学で発達の研究を行なおうとする場合には、発達の状況を測定する尺度は広範囲な年齢のいずれに於ても同一の原点と単位を持つように構成されていなければならない。そうでない場合には、発達曲線の型についての議論は全く無意味なものになってしまう。

ところで、実際の研究では例えば各年齢水準毎に当該年齢の被験者を測定対象としたテストを構成し実施することになるが、この場合には困難度の異なる複数のテストを実施した結果を相互に比較可能な同一の尺度上に表わす工夫、すなわち尺度の等化が必要となる。この場合、等化する各テストの受験者群間で能力水準に発達の差があるため、特に「垂直的等化」と呼ぶ。発達心理学では最近、縦断的研究の重要性が見直されつつある(例えば Baltes et al, 1980) が、定量的な研究を行なおうとする時には、広い範囲の年齢で等化さ

* 東京学芸大学教育学部教育心理学教室 (Department of Educational Psychology, Faculty of Education, Tokyo Gakuhei University)

れている心理測定尺度が存在することが不可欠である。

1.2 尺度を等化する方法

具体的に尺度を等化する方法として従来から検討されて来たものに

(i) 等パーセンタイル法 (equipercentile equating)

(ii) 線型変換による等化法 (linear equating) がある。

等パーセンタイル法というのは、基本的には等化すべきテスト X とテスト Y とが等しい信頼性を持ち、平行テストになっている場合に、いかなる被験者集団においても得点 x と得点 y のパーセンタイル順位が等しいならば、テスト X における得点 x とテスト Y における得点 y とは同等である と考えるもので、パーセンタイル順位を媒介としてテスト X とテスト Y を等化する。

また、線型変換による等化法は、同じく基本的には等化すべきテスト X とテスト Y とが等しい信頼性を持ち、平行テストになっている場合に、いかなる被験者集団においても得点 x を標準化した得点 Z_x と得点 y を標準化した得点 Z_y とが等しいならば、テスト X における得点 x とテスト Y における得点 y とは同等である と考えるもので、標準得点を媒介としてテスト X とテスト Y を等化する。

実際のテストを等化するために状況に応じて等化法にも様々な工夫がされている(例えば、Angoff: 1971, 塗師: 1981などを参照)。

これらの古典的な方法に代わって、最近特に注目されている方法に潜在特性モデルを用いた方法がある。潜在特性モデルには、項目の特性を表わすパラメータが特定の集団の特性とは無関係に定義できる (Lord & Novick: 1968) ことや、測定結果が正答数得点ではなくて潜在特性尺度値 θ で表わされるため、項目パラメータが同一の尺度上で表現されていたならば、異なる項目から構成されたテスト間で相互に比較が可能であることなどの利点がある。

潜在特性モデルを用いた等化法の一つに "真の得点を用いた等化法 (True-score equating)" がある。これは、潜在特性尺度値が θ である被験者のテスト X 及びテスト Y の真の得点をそれぞれ ξ , η とすると、 ξ 及び η は

$$\begin{cases} \xi \equiv \sum_{i=1}^n P_i(\theta) \\ \eta \equiv \sum_{j=1}^m P_j(\theta) \end{cases} \quad (1)$$

と表わされることを利用する。ここで、 $P_i(\theta)$, $P_j(\theta)$ は各項目の項目特性関数を表わす。(1)式は θ を媒介としてテスト X の真の得点 ξ とテスト Y の真の得点 η との関係を示しており、 θ の値を変化させることによって、 ξ と η との関係を求めることができる。

ところで、この方法ではテスト X 及びテスト Y に含まれる全ての項目のパラメータ値が同一の尺度上で表現されていることが必要であるが、実際にテストを等化する場合には二つのテストの項目パラメータ値がそれぞれ別の原点と単位を持つ尺度上に表現されることがある。これは、各項目のパラメータ値を推定する際に便宜上、推定のために用いられた被験者集団で潜在特性尺度値の平均が0、標準偏差が1となるように尺度の原点と単位を定め、その尺度上の値としてパラメータ値を推定することが多いからである。従って、異なる被験者集団を用いて推定された項目パラメータの値は共通尺度上の値に等化する必要がある。

いま、被験者集団 A が平均0、標準偏差1の分布をするように作られた尺度 θ と被験者集団 B が平均0、標準偏差1の分布をするように作られた尺度 θ^* とを等化するという状況について考える。もし、被験者集団 A が尺度 θ^* 上で平均 l 、標準偏差 k の分布をすることが既知ならば、

$$\theta^* = k\theta + l \quad (2)$$

で尺度 θ は尺度 θ^* に等化できる。この時、尺度 θ で識別力パラメータが a 、困難度パラメータが b という値を持つ項目については、

$$a^* = \frac{1}{k} a \quad (3)$$

$$b^* = kb + l \quad (4)$$

によって尺度 θ^* 上の項目パラメータ値が得られる。実際には平均値 l 及び標準偏差の値 k は未知であるから(2)式中の等化係数 k , l の値は推定する必要がある。その際に利用する情報に応じて、等化法には、

- (i) 同一の項目群について2つの尺度上でのパラメータ推定値が得られている時、これらを利用して等化係数の値 k , l を推定する方法
- (ii) 同一の被験者集団について2つの尺度上での推定尺度値が得られている時、これらを利用して等化係数の値 k , l を推定する方法

の2種類がある。

2. 共通被験者を利用した尺度の等化

野口 (1983) は前節の (ii) の方法に相当する、共通被験者の推定尺度値を利用して二つの潜在特性尺度を等化する方法を提案している。

いま、テスト A 及びテスト B という二つのテストがあり、テスト A の結果は尺度 θ 上の値で、テスト B の結果は尺度 θ^* 上の値で表現されるものとする。この時、テスト A の結果を尺度 θ^* 上の値で表現しようとするならば尺度 θ を尺度 θ^* に等化することが必要になる。

野口 (1983) の方法では、

① 先ず、テスト A (尺度 θ) 及びテスト B (尺度 θ^*) の適用範囲 $[\theta_L, \theta_U]$ 及び $[\theta_L^*, \theta_U^*]$ を各テストのテスト情報量に基づいて決定する。

② 次に、 N' 名の被験者にテスト A 及びテスト B の両方を実施して、被験者 i ($i=1, \dots, N'$) の推定尺度値 $\hat{\theta}_i$ 及び $\hat{\theta}_i^*$ を求め、これら双方の値が各テストの適用範囲に入った被験者 (有効被験者) N 名を残す。

③ 残った被験者の推定尺度値 $\hat{\theta}_i, \hat{\theta}_i^* (i=1, \dots, N)$ から各テスト毎の平均 m_A, m_B 及び分散 S_A^2, S_B^2 を、さらに二つのテストの共分散 S_{AB} を求める。

$$\textcircled{4} \quad K \equiv \frac{S_{AB}}{S_A^2 - S_B^2} \quad (5)$$

として、

$$\hat{k} = \frac{-1 + \text{sign}(K)\sqrt{1+K^2}}{K}, \quad (6)$$

$$\hat{l} = m_B - \hat{k}m_A \quad (7)$$

から等化係数 k 及び l の推定値を求める。

という手順をとっている。

この方法は、二つのテストを N 名の同一被験者に実施することによって、

$$\theta_i^* = k\theta_i + l \quad (i=1, \dots, N) \quad (8)$$

が厳密に成立する状況を設定して、各被験者から実際に得られた推定尺度値 $\hat{\theta}_i, \hat{\theta}_i^* (i=1, \dots, N)$ を用いて等化係数 k, l の値を推定しようとするものである。

この時、推定尺度値 $\hat{\theta}_i, \hat{\theta}_i^* (i=1, \dots, N)$ は真値 θ_i, θ_i^* と誤差 ε_i, δ_i との和で、すなわち

$$\begin{cases} \hat{\theta}_i = \theta_i + \varepsilon_i \\ \hat{\theta}_i^* = \theta_i^* + \delta_i \end{cases} \quad (i=1, \dots, N) \quad (9)$$

で表わされるものとし、誤差 ε_i, δ_i についてはそれぞれの期待値が0、両者の共分散の値が0であることを全ての被験者 i について仮定している。ここで、尺度

値の推定に最尤推定法を用いた場合には、誤差 ε_i, δ_i の分散は漸近的にはそれぞれのテスト情報量の逆数に等しくなる (例えば、Lord: 1980)。すなわち、

$$\begin{cases} \text{var}(\varepsilon_i) = 1/I(\theta_i) \\ \text{var}(\delta_i) = 1/I^*(\theta_i^*) \end{cases} \quad (i=1, \dots, N) \quad (10)$$

ところで、実際にテストを用いて測定する際には当該被験者に対してテスト情報量がある程度以上の値をとるテストが選ばれる。逆に、テストの方から考えると当該テストで精度の高い測定が可能な被験者はある範囲内の尺度値を持つ者に限定されることになる。この範囲のことを野口 (1983) ではテストの“適用範囲”と呼んでいるが、この適用範囲内の尺度値については実際のテストではテスト情報量にそれ程大きな違いが見られないことが多いため、誤差の大きさは適用範囲内では一定である。すなわち

$$\begin{cases} \text{var}(\varepsilon_i) = \sigma_\varepsilon^2 \\ \text{var}(\delta_i) = \sigma_\delta^2 \end{cases} \quad (i=1, \dots, N) \quad (11)$$

と仮定し、さらに、項目数がほぼ等しい二つのテストの間でもテスト情報量にそれ程大きな違いが見られないことが多いため、二つのテストの間で、それぞれの適用範囲内では誤差分散の大きさは等しい、すなわち

$$\sigma_\varepsilon^2 = \sigma_\delta^2 = \sigma^2 \quad (12)$$

と仮定している。これらの仮定は厳密な意味で成り立つものではなく、実用的な近似である。

これらの二つの仮定を置かない場合には、統計学で言う変数内誤差モデルに相当し、係数 k, l の推定に関して様々な議論が含まれる (例えば、Kendall: 1980) が、野口 (1983) の等化法ではこれらの仮定を置くことによって、問題を主成分を表わす直線式を推定することに単純化しているわけである。

野口 (1983) では、項目パラメータの値が既知である。芝 (1978) の“語彙理解力尺度”のJ1版の52項目を折半して二つのテストを構成し、各テストから得られる被験者の推定尺度値を用いて等化係数 k, l の値を推定し、この等化法の妥当性を確かめている。この場合、二つのテストは同一の尺度上で構成されているのであるから、いずれのテストから得られる推定尺度値も同一尺度上に表現されるわけで、結果的に得られる等化係数の推定値は $\hat{k}$ がほぼ1に、 $\hat{l}$ がほぼ0になることが期待される。そして、実際に、二つのテストがほぼ平行になるように折半した場合には、約300名の被験者で、また、二つのテストが比較的易しい項目から成るテストと比較的難しい項目から成るテストで両者の困難度の分布が一部重なるように折半した場合

には、約 350 名の被験者で一応満足な結果が得られたとしている。

3. 野口 (1983) の方法の検討

3.1 目的

2 章でも述べたように、野口 (1983) では「語彙理解力尺度」を用いて等化係数の推定値 $\hat{k}$, $\hat{l}$ について検討しているが、その際、被験者には芝 (1978) の「J 1 版に対する標準データ」を用いている。このため、得られた結果は特定の被験者集団に基づいたものに限定され、推定値 $\hat{k}$, $\hat{l}$ がどのような分布をするかについてはさらに詳しく検討することが必要である。しかしながら、 $\hat{k}$, $\hat{l}$ の分布について直接議論することは複雑であるし、野口 (1983) の方法はテストの適用範囲内で誤差分散について (11), (12) 式が成立するものと近似的にみなしているため、厳密な標本分布を用いて検討することはかならずしも適当とは言えない。このため、本研究では複数の異なる条件を設定して、コンピュータ・シミュレーションを用いて、野口 (1983) の方法による等化係数推定値の偏り (bias) 及び安定性について検討する。

3.2 方法

本研究では、野口 (1983) で用いられたテストと同じものを用いて検討を進める。すなわち、芝 (1978) の「語彙理解力尺度」の J 1 版の中から 52 項目を抽出し、それを折半して 26 項目から成るテストを二つ用意するが、その際、二つのテスト相互間の関係が

- (i) ほぼ平行テストになる場合 (テスト A, テスト B),
- (ii) 比較的易しい項目から成るテストと比較的難しい項目から成るテストで、両者の困難度が一部重なっている場合 (テスト C, テスト D),

- (iii) 易しい項目から成るテストと難しい項目から成るテストで、両者の困難度が全く重ならない場合 (テスト E, テスト F),

という 3 組用意する。そして、テスト A をテスト B に等化する場合をケース I, テスト C をテスト D に等化する場合をケース II, テスト E をテスト F に等化する場合をケース III と呼ぶ。

J 1 版の項目は、標準的な中学校 1 年生の語彙理解力分布の平均が 0, 標準偏差が 1 となるような尺度上で、項目パラメータ値が推定されている。従って、いずれの 2 組のテストを等化した場合でも、等化係数の推定値は $\hat{k}$ がほぼ 1.0, $\hat{l}$ がほぼ 0.0 になることが期待される。

各テストの項目パラメータ推定値及び適用範囲は表 1 に示した通りである。また、各テストの情報量は図 1 に示した通りである。なお、各テストの適用範囲には、野口 (1983) と同じく、そのテストが持つ最大情報量を $I_{\max}$ として、 $\sqrt{I(\theta)}$ が $\sqrt{I_{\max}}$ の 70% 以上の値を示す語彙理解尺度値の区間をあてているが、等化係数の計算に際しては若干の修正を加えた。というのは、当該被験者について得られるのは推定尺度値 $\hat{\theta}$ 及び $\hat{\theta}^*$ であって、真の尺度値 θ 及び θ^* はわからず、従って当該被験者の尺度値が適用範囲内にあるか否かの判断ができないからである。野口 (1983) では、真の尺度値は適用範囲内にあるにもかかわらず適用範囲外の推定尺度値が得られた被験者は計算から除外してしまうことになる。逆に真の尺度値は適用範囲外にあるにもかかわらず推定尺度値が適用範囲内に入る場合も当然あるわけであるが、ここではデータを利用しないことによる損失を重視して、実際の計算には、テストの適用範囲を $[\theta_L, \theta_U]$ とした時、上限は θ_U に $1/\sqrt{I(\theta_U)}$ の t_U 倍を加えた値に、下限は θ_L から $1/\sqrt{I(\theta_L)}$ の t_L 倍を減じた値にした範囲内の推定尺度値を持つ被験者を

表 1. 各テストの項目パラメータ値 及び 適用範囲

テスト	識 別 力			困 難 度			適用範囲	計算に利用する 推定尺度値の範囲
	最 大	最 小	平 均	最 大	最 小	平 均		
A	.90	.27	.57	1.54	-2.05	-.38	-2.6 ~ 1.5	-3.0 ~ 2.0
B	.91	.37	.55	1.70	-1.87	-.30	-2.4 ~ 1.6	-3.0 ~ 2.0
C	.91	.38	.59	.09	-2.05	-.93	-2.8 ~ 1.0	-3.5 ~ 1.5
D	.81	.27	.53	1.70	-.93	.25	-2.0 ~ 2.2	-2.5 ~ 2.5
E	.91	.38	.58	-.45	-2.05	-1.08	-2.9 ~ .7	-3.5 ~ 1.0
F	.82	.27	.55	1.70	-.40	.40	-1.7 ~ 2.1	-2.0 ~ 2.5

図 1-a. テスト A とテスト B の情報曲線 (ただし、縦軸には $\sqrt{I(\theta)}$ を表示した。
b, c も同じ)

図 1-b. テスト C とテスト D の情報曲線

図 1-c. テスト E とテスト F の情報曲線

用いることにする。 t_u, t_l は 1 に近い定数で、結果的に得られる適用範囲が比較的区切りの良い数になるように定めた。この計算に利用する推定尺度値の範囲は表 1 に示した通りである。なお、他の一般の測定尺度を等化するのに本等化法を用いる場合には、 $\sqrt{I(\theta)}$ が $\sqrt{I_{\max}}$ の 70% 以上の値を示す区間を用いるというのではなく、当該測定場面が必要となる精度から適用範囲を定めなければならないことは言うまでもない。

また、被験者集団としては、各テストの項目パラメータ値を推定するのに用いられた被験者集団と同じ性質を持つものとして、尺度値が平均 0.0、標準偏差 1.0 の正規分布する集団（集団 I）と、尺度値が平均 0.8、標準偏差 1.2 の正規分布する集団（集団 II）の 2 つを設定した。集団 II は語彙理解尺度で 1 学年上の集団の分布に相当するが、本等化法で用いられる被験者集団に一種のかたよりがある場合に生ずる結果への影響を確認するために設定した。

さらに、有効被験者の数がどのような影響を持つかを確認するために、有効被験者数が 50, 100, 200, 400, 800 名の 5 通りの場合を設定した。

結局、全部でテストの難易度の差 (3) × 被験者の分布状況 (2) × 有効被験者数 (5) の 30 通りの条件で検討することになる。

シミュレーションの具体的な手続は次の通りである。

- (i) 集団 I または集団 II の分布に相当する正規乱数を発生させる。これが当該被験者 i の尺度値 θ_i になる。
 - (ii) 一様乱数 $[0, 1)$ を各テストの項目数分だけ発生させる。そして各項目 j について、乱数 r_j と項目特性関数から計算される被験者 i の正答確率 $P_j(\theta_i)$ とを比較する。この時、 r_j が $P_j(\theta_i)$ よりも大ならば被験者 i は項目 j に誤答したとみなし、 r_j が $P_j(\theta_i)$ よりも小ならば被験者 i は項目 j に正答したとみなす。
 - (iii) 二つのテストについて、それぞれ正答及び誤答の反応パターンが得られたならば、各テスト毎に最尤推定法を用いて尺度値の推定値 $\hat{\theta}_i$ 及び $\hat{\theta}_i^*$ を求める。これら二つの推定値の少なくとも一つが当該テストで計算に利用する推定尺度値の範囲からはずれているならば、この被験者は無効被験者として以後の計算には用いない。そうでなければ、有効被験者として残す。
- ここまでの手続を有効被験者数が所定の数 N になるまで繰り返す。
- (iv) この結果得られた $\hat{\theta}_i, \hat{\theta}_i^* (i=1, \dots, N)$ を用いて等化係数 k, l の値を計算する。
 - (v) 以上の手続を各条件毎に 100 回ずつ繰り返し実行する。

表 2. ケース I の結果 (集団 I)

有効被験者数	等化係数 k		等化係数 l		無効被験者数	
	平均	標準偏差	平均	標準偏差	平均	標準偏差
50	1.027	.121	.006	.087	3.1	1.7
100	1.014	.093	.012	.069	6.6	3.0
200	1.006	.056	-.008	.038	11.8	4.0
400	0.996	.042	-.006	.030	23.8	4.5
800	1.000	.026	.003	.020	48.0	6.7

表 3. ケース II の結果 (集団 I)

有効被験者数	等化係数 k		等化係数 l		無効被験者数	
	平均	標準偏差	平均	標準偏差	平均	標準偏差
50	1.072	.129	.026	.094	5.1	2.3
100	1.080	.103	.051	.068	10.4	3.2
200	1.074	.070	.050	.042	20.7	4.6
400	1.073	.047	.044	.033	40.3	6.6
800	1.075	.034	.046	.023	83.4	9.7

表 4. ケースⅢの結果 (集団Ⅰ)

有効被験者数	等化係数 k		等化係数 l		無効被験者数	
	平 均	標準偏差	平 均	標準偏差	平 均	標準偏差
50	1.170	.168	.117	.093	12.7	3.9
100	1.147	.122	.112	.056	24.4	5.8
200	1.127	.093	.121	.051	50.5	8.1
400	1.123	.060	.109	.031	99.0	10.8
800	1.123	.035	.109	.023	199.0	15.7

表 5. ケースⅠの結果 (集団Ⅱ)

有効被験者数	等化係数 k		等化係数 l		無効被験者数	
	平 均	標準偏差	平 均	標準偏差	平 均	標準偏差
50	1.009	.104	.002	.091	13.5	4.3
100	.986	.070	.010	.066	26.1	5.6
200	.997	.052	-.003	.046	53.3	8.3
400	1.003	.033	-.008	.030	106.7	10.5
800	.999	.025	-.003	.023	216.5	16.2

表 6. ケースⅡの結果 (集団Ⅱ)

有効被験者数	等化係数 k		等化係数 l		無効被験者数	
	平 均	標準偏差	平 均	標準偏差	平 均	標準偏差
50	1.194	.161	.032	.109	18.0	5.0
100	1.208	.106	.046	.068	37.2	7.3
200	1.185	.074	.055	.051	73.5	10.1
400	1.184	.052	.056	.037	148.9	12.8
800	1.185	.037	.056	.026	300.1	19.8

表 7. ケースⅢの結果 (集団Ⅱ)

有効被験者数	等化係数 k		等化係数 l		無効被験者数	
	平 均	標準偏差	平 均	標準偏差	平 均	標準偏差
50	1.396	.250	.159	.106	37.0	8.6
100	1.337	.146	.146	.067	75.2	11.2
200	1.318	.098	.153	.047	145.6	14.3
400	1.324	.074	.155	.032	296.0	21.3
800	1.311	.047	.155	.023	585.7	34.1

3.3 結 果

以上の方法に従って、集団Ⅰを用いて計算した結果は表2, 3, 4に、集団Ⅱを用いて計算した結果は表5, 6, 7に示した通りである。表2と表5は二つのテストがほぼ平行であるケースⅠについて、表3と表6は二つのテストの難易度差が小さいケースⅡについて、表

4と表7は二つのテストの難易度の差が大きいケースⅢについての結果をそれぞれ示している。各表とも、等化係数 k 及び l の推定値の平均及び標準偏差、無効被験者数の平均及び標準偏差が示されている。

3.4 考 察

等化係数 k, l の推定値の平均については、 $\hat{k}, \hat{l}$ とも

表 8. 等化による尺度値のズレ

等化の方向		尺 度 値			
		0.0	1.0	2.0	3.0
集 団 I	AをBに	0.0	1.0	2.0	3.0
	CをDに	0.05	1.12	2.19	3.26
	EをFに	0.11	1.23	2.35	3.47
集 団 II	AをBに	0.0	1.0	2.0	3.0
	CをDに	0.06	1.25	2.44	3.63
	EをFに	0.16	1.48	2.80	4.12

にケース×分布の六つの状況のそれぞれの中では有効被験者数の変化にもかかわらず、ほとんど安定している。

集団 I で、ケース I の場合はおよそ $\hat{k}$ が 1.00, $\hat{l}$ が 0.00, ケース II の場合はおよそ $\hat{k}$ が 1.07, $\hat{l}$ が 0.05, ケース III の場合はおよそ $\hat{k}$ が 1.12, $\hat{l}$ が 0.11 である。また、集団 II でケース I の場合はおよそ $\hat{k}$ が 1.00, $\hat{l}$ が 0.00, ケース II の場合はおよそ $\hat{k}$ が 1.19, $\hat{l}$ が 0.05, ケース III の場合はおよそ $\hat{k}$ が 1.32, $\hat{l}$ が 0.16 である。

ほぼ平行なテストの場合には、集団の分布にかかわらず等化係数の真値に近い推定値が得られるが、等化するテストの平行度が下がるにつれて、そして集団の分布がかたよるにつれて推定値が悪くなっている。すなわち、 $\hat{k}$ の値がバイアスを持ってしまう。

これらの値を用いて、等化のための式を状況毎に表わすと、順に次のようになる。

$$\text{集団 I} \cdot \text{ケース I} : \theta^* = 1.00\theta + 0.00 \quad (13)$$

$$\text{集団 I} \cdot \text{ケース II} : \theta^* = 1.07\theta + 0.05 \quad (14)$$

$$\text{集団 I} \cdot \text{ケース III} : \theta^* = 1.12\theta + 0.11 \quad (15)$$

$$\text{集団 II} \cdot \text{ケース I} : \theta^* = 1.00\theta + 0.00 \quad (16)$$

$$\text{集団 II} \cdot \text{ケース II} : \theta^* = 1.19\theta + 0.06 \quad (17)$$

$$\text{集団 II} \cdot \text{ケース III} : \theta^* = 1.32\theta + 0.16 \quad (18)$$

これらの式を用いて、 θ の値が 0.0, 1.0, 2.0, 3.0 の時に等化した結果得られる θ^* の値を表 8 に示した。 θ の値と θ^* の値との差が 0.3 までを許容することになると、ケース I の場合にはいずれの集団でもこの条件を満たすが、ケース II の場合には集団 I のみでしかこの条件は満たされず、ケース III の場合にはいずれの集団でもこの条件は満たされない。基準にとった値 "0.3" は、信頼性係数が 0.91 であるテストを真の得点が平均 0.0, 標準偏差 1.0 の分布をする集団に実施した時の標準誤差の大きさに相当するが、Z 得点に

換算すると 3 点であり、実用的にはこの程度の大きさの差はほとんど問題にはならない。

等化係数 k, l の推定値の安定性については、いずれの集団・ケースについても有効被験者数が多くなる程、標準偏差の値が小さくなることから、安定性が良くなっていることがわかる。細かく見ると、集団 II でケース III の場合に安定性が最も悪く、集団 I でケース III の場合がこれに次いで悪い。その他の四つの場合はほとんど同じ傾向を示している。語彙理解尺度では項目パラメータ値を小数第 1 位までで示している（芝・野口：1982）ことから、ここでは標準偏差の値がおよそ 0.05 以下になった時に安定した結果が得られたものと考え、いずれの集団に於ても、二つのテストが平行であるケース I では、 $\hat{k}, \hat{l}$ ともに 200 名の有効被験者が必要なことになる。同じく、二つのテストの難易度が小さいケース II では、 $\hat{k}$ については 400 名、 $\hat{l}$ については 200 名が、二つのテストの難易度差が大きいケース III では、 $\hat{k}$ については 800 名、 $\hat{l}$ については 200 名の有効被験者が必要である。

従って、等化係数の推定値の安定性に関しては、 $\hat{l}$ についてはいずれの場合にも 200 名程度の有効被験者が必要であるが、 $\hat{k}$ については等化するテスト間の平行度が低くなるにつれて安定性が悪くなり、同じ安定性を得るためにはより多くの有効被験者が必要となる。

無効被験者数については、集団 II の場合の方が集団 I の場合よりも数が多くなっているが、これは集団 II の分布の位置が各テストの適用範囲に比較して全体として尺度値の高い方に寄っているために生じたものである。集団 I では、ケース I で有効被験者数の約 6% の人数が、ケース II で同じく約 10% が、ケース III では約 25% が無効被験者として計算から除外されている。また、集団 II では順に、約 25%、約 36%、約 75% が除外されている。

実際に等化を行なう状況を考えると、被験者数をそれほど多く確保することは難しく、集められたデータはできるだけ効率よく利用することが望まれる。このため、実用的には無効被験者の数は有効被験者数の 20~25% 以内にとどまることが望ましい。本研究では、集団 I でケース I 及びケース II の場合が、集団 II でケース I の場合がこれにあたる。

以上で、等化係数 k, l の推定値のかたより (平均), 安定性 (標準偏差), そして無効被験者数の 3 つの点から考察を進めたが、全体として見ると、ケース I では集団の分布にかかわらず少数の有効被験者で良い推定結果が得られるが、ケース II では集団の分布にかたよりがある場合にあまり良い推定結果は得られず、さらに、ケース III ではいずれの集団についてもあまり良い結果が得られていない。特に等化係数 k の推定に関しては集団の分布のかたよりの影響が強く、等化係数 l の推定に関しては二つのテストの平行度の影響が強いことがわかる。従って、ケース III のように等化する二つのテストの難易度が極端に離れている場合に本方法を用いて等化することは容易ではない。しかし、現実にはこのようなケースはまれであって、ケース II のように項目の困難度の分布がある程度重なっているテストどうしを等化する場合が多い。この場合には集団 II のように一方のテストの適用範囲内に主として分布する被験者集団を用いて等化したのでは具合が悪い。例えば、一方のテストが小学校 3 年生を対象としたもので、もう一方のテストが小学校 4 年生を対象としたものである場合には、いずれか一方の学年の児童に対してのみ両方のテストを実施し、その結果から等化係数を推定するというのでは具合が悪いというのである。この場合は例えば、両方の学年の児童に対して両方のテストを実施して、その結果から等化係数を推定するなど分布になるだけ偏りのないように配慮しておく必要がある。

3.5 結 論

本等化法は等化係数を推定する際に複雑な数値計算を必要としない簡便な方法である。特に、等化する二つのテストの難易度が極端に異ならない場合で、等化係数を推定する際に用いる被験者集団の分布に偏りが

ないように配慮されている場合に有効であることが示された。

謝 辞

本研究を実施するにあたって、東京大学教育学部の芝 祐順教授から有益な助言をいただきました。また、テネシー大学の鯨島史子教授からは英文要約の作成に際してご指導いただきました。ここに深く感謝の意を表します。

附 記

本研究の計算は東京大学大型計算機センターの M 280H システムを使用した。乱数の発生については同センターのライブラリー・プログラム G 5/TC/UNIF 及び G 5/TC/NORM (手塚 集・伏見正則両氏作成) を使用した。

参 考 文 献

- Angoff, W.H. (1971) Scales, Norms and Equivalent Scores. In Thorndike, R.L. (Ed.) *Educational Measurement*, 2nd ed. Washington, D.C., American Council on Education.
- Baltes, P.B., Reese, H.W. & Lipsitt, L.P. (1980) Life-span Developmental Psychology. *Annual Review of Psychology*, **31**, 65-110.
- Kendall, M.G. (1980) *Multivariate Analysis*, 2nd ed., Charles Griffin., 奥野忠一・大橋靖雄 (訳) 多変量解析. 培風館.
- Lord, F.M. (1980) *Applications of Item Response Theory to Practical Testing Problems*. Lawrence Erlbaum Associates.
- Lord, F.M. & Novick, M.R. (1968) *Statistical Theories of Mental Test Scores.*, Addison-Wesley.
- 野口裕之 (1983) 被験者の推定尺度値を利用した潜在特性尺度等化法. 教育心理学研究, **31**, 233-238.
- 塗師 斌 (1981) テスト得点の等化 (equating) の方法. 教育心理学年報, 第 20 集, 133-140, 日本教育心理学会.
- 芝 祐順 (1978) 語彙理解尺度作成の試み. 東京大学教育学部紀要, 第 17 巻, 47-58.
- 芝 祐順・野口裕之 (1983) 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, 第 22 巻, 31-42.

—1984 年 9 月受付, 1985 年 1 月再受付—