【個人研究】

共通項目の部分得点モデル化によるテストの等化

藤　森　　進*

Equating by Modeling of Partial Scores of Test Anchor Items

Susumu FUJIMORI

The process of awarding points for a single anchor item across multiple tests on the basis of a common scale is referred to as “equating.” In general, the item response theory has been used for equating. This theory assumes that an item has a binary response whose score is 1 for a correct answer and 0 for an incorrect answer. This is a well-known example of a typical 2-parameter logistic model. In the present study, a new polytomous anchor item was created by calculating the sum of the scores for multiple anchor items. This was then equalized by applying Fujimori’s partial test score model and was compared with the score that was equalized according to the 2-parameter logistic model. No difference was observed between the results of the two models. As compared with the conventional methods, equalization by the modeling of partial scores for anchor items requires the deduction of fewer model parameters in the anchor. Despite this disparity, equivalent results were obtained. Therefore, it is concluded that modeling partial scores is a more efficient method.

Key words: item response theory, partial test score model, equating, anchor item
項目反応理論、部分得点モデル、等化、共通項目

1. 研究の目的

共通項目を持つ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、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。たとえば、藤森（1999）によれば、2値の場合で6から8項目程度の共通項目が等化には必要とされている。これに対して多値の解答データを許容する項目反応モデルの等化は十分でない。

この研究では、多値データを許容する部分得点モデル（藤森，2001）により等化が可能であることを示すとともに、2値データの多値化の持つ意義を検討する。

具体的には、1または0の値をとる複数の共通項目に関する得点の和を求めることにより、多値の新しい共通項目を1つだけ作成する。これに藤森の部分得点モデルを適用することによりテストの等化を行い、通常の2値の複数項目を利用する2母数ロジスティックモデ

*ふじもり　すすむ　文教大学人間科学部人間科学科

ルによる等化結果とシミュレーションによる比較を試みる。

2. 方法

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、本研究では式の2母数ロジスティックモデル（Birnbaum, 1968）を利用する。

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

ここでiは被験者、θはその能力を表す母数、D=1.7の定数、jは項目番号、a_jはその識別力、b_jは困難度を表すモデルの母数である。またx_{ij}は、被験者iの項目j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。モデルや母数の持つ意味などについては、たとえば藤森（2002a）などを参照されたい。

2.1.2. 部分得点モデル

藤森の部分得点モデルでは、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。またその得点は、潜在的な2値の正誤反応の和によって表されることを想定し、この潜在的な2値の項目に関して（1）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。この場合、潜在的2値項目の正誤得点が仮に分かっていたとすると、最尤法で推定される能力母数の推定値と部分得点モデルによる推定値は一致することを示すことができる。類似母数である場合もこの関係は近似的に成立する。

部分得点モデルでは、問題jの潜在的項目kに対して受験者が潜在的正答反応を取る確率を（1）式で想定しP_j（θ）とする。これをs_j回繰り返し受験したときに、受験者が潜在的に取りうる正誤反応の平均が、顕在的部分得点r_jとなる。

$$Q_j\left(\theta\right)=1-P_j\left(\theta\right) \quad (2)$$

とすると

$$\ell_{part}\left(\theta\right)=\sum_{j=1}^{n}s_j\left(r_j\ln\left(P_j\left(\theta\right)\right)+\left(1-r_j\right)\ln\left(Q_j\left(\theta\right)\right)\right) \quad (3)$$

によって表される対数尤度$\ell_{part}\left(\theta\right)$を用いて受験者の能力θが推定される。

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。母数を共通とする複数の潜在的項目の正誤パターンによる能力推定と、（3）式による能力推定は一致することが示せる。また、類似母数をとる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。

またテストが実施された集団の能力分布を仮定した上で、s_j回の正誤の和である2項分布と仮定された能力分布との積を能力の次元で積分して部分得点の理論的な分布関数を求め、求められた理論的な分布関数と実際のデータの部分得点の経験的な分布関数が最もよく一致する値として潜在的な問題の繰り返し回数であるs_jを推定することも可能であり、シミュレーションによる検討が行われた限りでは極めてその推定成績は良い（藤森，2002b）。

2.2. 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。等化するテスト及び集団数は2とする。テスト数が3以上となるケースは機会を改めて検討したい。被験者数はいずれも3000人、30項目とし、2つの集団の能力（2つのテストの困難度水準）が互いに異なる垂直的等化場面を想定する。

シミュレーションデータでは、被験者の能力分布は、下位群は標準正規分布に従うと仮定し、上位群は平均のみ異なる正規分布$N(0.5, 1^2)$とした。

テスト項目の2母数ロジスティックモデルの項目母数の分布型は以下のように定めた。識別力母数は、平均0.65、標準偏差0.25、下限0.3、上限2.0の切断正規分布、また困難度母数は、下位群に実施するテスト（下位テスト）に関しては平均0、標準偏差0.5の正規分布に従うと仮定し、また上位群に実施するテスト（上位テスト）の困難度に関しては平均0.5、標準偏差0.5の正規分布に従うと仮定した。能力母数と困難度の平均を一致させることにより正答率等に偏りが生じないようにしたわけである。ただし上位と下位のテストの共通項目とした6項目は、下位テストから選んでいる。下位テスト項目を共通項目とした理由は、学校教育などのテストの現実の実施場面では、下位のテスト項目に関する内容は、教育済みであり、これらを上位群に実施することは容易であるのに対し、その逆は容易でないからである。

以上のようにして全ての母数を定めた後、能力母数θの被験者のある項目に対する正誤は、2母数ロジスティックモデルから予想される正答確率を、範囲0～1の一様乱数と比較し、前者が下回る場合被験者の反応を正答1、上回る場合誤答0として作成した。2母数ロジスティックモデルに従うこの2値データパターンを、項目数30として10回繰り返し作成し（データ1～10）2値テストデータとした。

一方、データの共通項目の部分得点化は次のようにして行った。上位と下位のテストの共通項目とした6項目を部分得点モデルの潜在的項目と仮定し、これらの正誤の平均を求めることにより、多値のテスト項目を1つの共通項目として扱うこととした。なお共通項目以外の他の項目は、そのまま2値データとして処理している。すなわち、部分得点化と言ってもすべてのデータを部分得点化しているのではなく共通項目部分に限って行ってデータを作成した。

2.3. 母数の推定と等化

2.3.1. モデル母数の推定

項目反応理論では、モデル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。

2値のままのデータを扱う2母数ロジスティックモデルの項目母数の推定は尤度の周辺化を行い、Bock & Aitkin（1981）のEMアルゴリズムを利用している。また能力母数の推定は、最尤推定によっている。最尤法によると全問正答や誤答のとき、推定値を得るのが困難となるが、本研究では、後述するように、受験者の能力水準やテストの困難度水準が異なる2つのテストデータの等化（いわゆる垂直的等化）を取り扱う関係上、受験者の能力分布を能力母数の推定に利用するベイズ推定は、等化の成績に影響を及ぼす可能性があるため採用しなかった。

一方、部分得点モデルの項目母数の推定は、能力母数所与として項目母数を推定し、その後項目母数所与として能力母数を推定する過程を繰り返す同時推定法を利用している。本来は、2母数ロジスティックモデルと同様に尤度の周辺化を行いEMアルゴリズムを適用するのが比較の上からは適当であるが、プログラム作成が間に合わなかったため行うことができなかった。能力母数の推定は、2母数ロジスティックモデルと同様に最尤推定である。

母数の推定は、いずれも自作のFORTRANあるいはpascalプログラム（delphi6）によった。

2.3.2. 2値データの等化

本研究では、等化の前に上位、下位どちらも個別に項目母数の推定値を求めておくことにする。このような場合、上位と下位の2つのテストデータにおける共通項目の項目母数の推定値は、各データ別に推定を行っているわけであるから、見かけ上一致しない。本来1つの項目の特性を母数は表しているわけであるから見かけ上の不一致を共通尺度上にのせることによって解消する必要がある。項目反応理論では、これを等化と呼んでいる。項

目反応理論では、2つのテストデータにおける能力母数をθと**θ***とするとき両者には（4）の関係があり、式中のk, ℓを等化係数と呼んでいる。

$$\theta = k\theta^* + \ell \qquad (4)$$

さて共通項目が2値のテストデータの等化は、芝（1978）の主軸法によって行った。この方法は、2つのテストデータで個別に項目母数の推定値が得られた場合に適用できるものであり、詳細は省略するが、同法によるとkの推定値は（5）となる。

$$k = \frac{\sqrt{1+\omega^2}-1}{\omega} \qquad (5)$$

ここでωは（6）である。

$$\omega = \frac{2\sum_{j=1}^{m} a_j a_j^*}{\sum_{j=1}^{m}\left(a_j^2 - a_j^{*2}\right)} \qquad (6)$$

（5）で得られたkを利用してℓは（7）より得られる。

$$\ell = \bar{b} - k\bar{b}^* \qquad (7)$$

ここで、$\bar{b}$, $\bar{b}^*$は、それぞれのデータで得られた困難度母数の推定値の平均である。

2.3.3.　部分得点化したデータの等化

部分得点化した共通項目の等化でも、等化の前に上位、下位どちらも個別に項目母数の推定値を求めておくことは2値の場合と同様である。部分得点化した場合、上位と下位のテストとも、共通項目の数は部分得点化の結果として1項目となるので、前述の芝の方法などは適用できない。このため、ここでは、上位の共通項目の2つの項目母数、すなわち識別力と困難度の値を、下位のテストの項目母数の値と一致させることにより等化係数kとℓを推定し、この値を上位テストの他の項目に適用することによって等化をおこなっている。

3.　結果と考察

シミュレーションにより作成した2値のデータ及び部分得点化したデータに、それぞれのモデルを適用した。どちらのモデルも、上位及び下位のテストデータについて項目母数を推定した。表1の値は、いずれも上位テストの全ての項目の真値と推定値の平均二乗誤差（MSE）である。やや2値データの方が小さい誤差を得ているが、2値データはEMアルゴリズムであり、部分得点化したデータは、EMアルゴリズムによるプログラム開発が間に合わなかったため、項目母数と能力母数の同時推定となったことを考えれば不思議は無い。部分得点化したデータで項目母数の推定値の誤差が大きいことは、等化の成績にも悪影響を及ぼすので、本研究結果を評価する場合にはその点を考慮に入れなければならない。

項目母数を得た後に下位データの推定値を固定して上位データの項目母数を等化し、等化された上位データの項目母数を所与として、上位群の能力母数の推定値を求めた。このようにして推定された能力母数の平均と真の能力母数の平均の比較により、等化成績を判断することとした。能力母数の推定は、最尤推定としたため全問正答あるいは誤答の受験者の能力母数は推定できない。このため、これらの推定できない受験者を除いて求めた結果が表2である。

表2の真値との差の絶対値からは、部分得点化したデータの誤差の方が小さいのが10組のデータの中で6組となっており両モデルの結果に大きな差はないことがわかる。表2の平均欄からは、2値データの推定値の平均は、ほぼ真値の平均近くであるのに対して、部分得点化した場合はやや低い推定値となっていることがわかる。

表1に示したように、項目母数の推定値はやや部分得点化した場合の方が悪く、その分等化にも不利であったわけであるから、両等化結果はほぼ同等の成績と見てよいのではな

表1 上位テストの項目母数の推定成績（MSE）

		データ 1	2	3	4	5	6	7	8	9	10	平均
2値データ	識別力	0.0025	0.0026	0.0040	0.0018	0.0014	0.0021	0.0010	0.0029	0.0047	0.0019	0.0025
	困難度	0.0065	0.0077	0.0047	0.0079	0.0053	0.0024	0.0038	0.0036	0.0019	0.0048	0.0049
部分得点化データ	識別力	0.0144	0.0039	0.0131	0.0332	0.0016	0.0544	0.0009	0.0048	0.0107	0.0105	0.0148
	困難度	0.0087	0.0033	0.0092	0.0109	0.0196	0.0097	0.0061	0.0033	0.0012	0.0019	0.0074

表2 上位群のθの真値及び推定値の平均

	データ 1	2	3	4	5	6	7	8	9	10	平均
真値	0.4768	0.5131	0.5056	0.5092	0.5043	0.4944	0.5050	0.4763	0.5302	0.5113	0.5026
	(1.026)	(0.996)	(0.994)	(0.991)	(1.038)	(0.996)	(1.020)	(0.988)	(1.009)	(1.003)	
2値データ	0.3937	0.5411	0.4495	0.5946	0.5487	0.4995	0.5320	0.4331	0.5229	0.5445	0.5060
	(1.131)	(1.097)	(1.077)	(1.111)	(1.202)	(1.115)	(1.169)	(1.090)	(1.064)	(1.154)	
真値との差の絶対値	0.0831	0.0280	0.0561	0.0854	0.0444	0.0051	0.0270	0.0432	0.0073	0.0332	0.0413
部分得点化データ	0.3953	0.5218	0.4384	0.5008	0.5369	0.4542	0.4790	0.4174	0.5164	0.4863	0.4747
	(1.070)	(1.041)	(1.010)	(0.929)	(1.130)	(0.911)	(1.077)	(1.055)	(1.025)	(1.032)	
真値との差の絶対値	0.0815	0.0087	0.0672	0.0084	0.0326	0.0402	0.0260	0.0589	0.0138	0.0250	0.0362
有効データ数	2957	2979	2991	2978	2986	2988	2986	2995	2983	2992	

注：()は標準偏差

いだろうか。

以上により、共通項目の部分得点化による等化は、少なくとも2値のデータの等化と同程度に有効であることは示されたと言えよう。

さて本研究のシミュレーション条件では、2値のままであれば共通項目数が6であるのだから12個の項目母数が必要となるが、部分得点化した場合は2個であり、繰り返し数sを母数としても3個に過ぎない。共通項目の部分得点モデル化による等化では、従来の方法より共通項目の推定されるべきモデルの母数の数が少ないのに、同程度の結果が得られるのであるから、パーシモニー（節約）の原理より効率的であると結論できる。また実際のテスト等化場面で、2つのテストの共通項目として部分的得点を許容する1項目しか存在しない場合でも適用できる利点もあると指摘できよう。また2値データの複数項目で等化を行う場合は、等化方法に幾つもの候補があるため、どの等化法を利用するかにより結果に違いが生まれ、また共通項目の母数に2つのテスト版の推定値のどちらの値を用いるべきか明確な論拠はないが、部分得点モデル化による等化では、結果として1項目であり、2つの版で同一の推定値になるため、この種の問題は生じないことも利点である。以上より、共通項目の部分得点化の有用性はあると判断できる。

文献

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.

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　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　2001　項目反応理論における部分得点の処理について　日本教育心理学会第43回総会発表論文集，394.

藤森進　2002a　テスト得点を統計的枠組みで分析する―項目反応理論―　渡部洋編「心理統計の技法」第7章　福村出版.

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.