

Rasch 項目分析モデル測定母数推定値分布幅、テスト次元性、 並びに、Rasch 測定適用妥当性

Rasch Parameter Estimates' Ranges Affected by Test Multidimensionality, and the Model Application Validity

平越裕之*、井澤廣行**

Hiroyuki Hirakoshi, Hiroyuki Izawa

This paper articulates a significant feature of the joint maximum likelihood estimation utilized in the Rasch item analysis model measurement; i.e., the decreasing distributional ranges of a set of parameter estimates affected by test multidimensionality. This feature is explicated in a theoretical analysis. The decreasing ranges of parameter estimates are also captured by the lowering validity of the model application. Furthermore, the Rasch residual fit statistics are shown to be insensitive or oversensitive at some points to the range variation of estimates.

Key words: joint maximum likelihood estimation, parameter estimates' ranges, multidimensionality, validity.

I. はじめに

確率論に基づく Rasch 項目分析モデル¹⁾ (Rasch, 1960) (以降 Rasch モデル) は、単純なモデルであるがゆえの解析の容易性やモデルの意味のわかりやすさ等の特徴がある²⁾。項目反応理論 (Item Response Theory³⁾: IRT) 体系では、Rasch モデルを一番単純なモデルとして位置づけることもあるが、Rasch モデルでの議論は、例えば複数の母数(Parameter)を持つロジスティックモデル³⁾や段階反応モデル³⁾といった、他のより複雑な発展型モデルにおいても基礎的な議論として成立するため、Rasch モデルにおける種々の議論は重要である。本稿では、Rasch モデルを用いた項目分析において仮定される、項目の一次元性が崩れた場合の推定値の振舞いについて考察する。さらに、一次元性が充足されている状況においても、結果的に Rasch モデル分析を用いて一次元性が議論できない場合があることを例証し、Rasch モデル測定 (以降 Rasch 測定) を適用する上での一つの指針を述べる。

まず、Rasch モデルの特徴に触れ、一次元性が満たされている場合に比べて、満たされない場合には、母数推定値がどのように振舞うかについて述べる。次いで、一次元性が満たされていても、母数推定値の分布幅が狭小である場合の Rasch モデル適合度や一次元性充足度指標値の振舞いについて例証し、Rasch モデルを適用する上で検討すべきことについて言及する。なお、ここでいう一次元性とは、分析対象の項目や能力の程度は同じ尺度で規定可能で、その尺度は唯一の変数で表現で

* 流通科学大学情報学部、〒651-2188 神戸市西区学園西町 3-1

** 流通科学大学サービス産業学部、〒651-2188 神戸市西区学園西町 3-1

きることを指している。

II. Rasch モデルにおける母数推定

1. Rasch モデルの項目特性曲線

Rasch モデルでは、項目困難度および受験者能力の二つの母数から導かれる正答確率を表すため、以下のような項目特性曲線を定義している。

$$P_j(\theta_i) = \frac{1}{1 + e^{-(\theta_i - \beta_j)}} \quad (1)$$

θ_i : 受験者*i*の能力値
 β_j : 項目*j*の困難度

この項目特性曲線は、ロジスティックモデル^{3) 4)}と呼ばれ、

$$P_j(\theta_i) = \frac{e^{\theta_i - \beta_j}}{e^{\theta_i - \beta_j} + e^{\theta_i - \beta_j} e^{-\theta_i + \beta_j}} = \frac{e^{\theta_i - \beta_j}}{e^{\theta_i - \beta_j} + 1} \quad (2)$$

とあらわされる。Log オッズ比⁵⁾であるロジットコレクト (Logit Correct: 正解確率を不正解確率で除して対数をとったもの) は、

$$\ln \left(\frac{P_j(\theta_i)}{1 - P_j(\theta_i)} \right) = \ln \left(\frac{e^{\theta_i - \beta_j}}{e^{\theta_i - \beta_j} + 1} \frac{e^{\theta_i - \beta_j} + 1}{1} \right) = \theta_i - \beta_j \quad (3)$$

となり、簡単に θ_i, β_j の和差だけで表わされると示すことができる。なお、 $\ln(\cdot)$ は自然対数を表す。この特性を利用して、項目困難度と受験者能力をロジットコレクトで表そうと考えたのが Rasch モデルである。つまり、このように定義すれば、項目困難度と受験者能力を一次元の間隔尺度として表すことができるからである。ロジットコレクトそのものは Rasch モデルで初めて開発されたわけではなく、それまでも様々な分野で利用されてきたものである(Berkson⁶⁾, 1944; Cramer⁷⁾, 2003; O'Neil⁸⁾, 2005; Peirce⁹⁾, 1878、参照)。Rasch モデルはロジットコレクトが受験者能力と項目困難度の関係を和差で表現でき、人間の感覚によく当てはまること、さらに数理的にその取り扱いが比較的容易であるため、これを利用したと考えることができる。

ロジットコレクトを用いると、能力および困難度の定義域を $-\infty \sim \infty$ とした上で和差表現できるので、ある意味で人間の感覚を正しく線形化しているように捉える事も出来る。しかし、ロジットコレクトは、受験者能力を線形尺度として測定する際の唯一の正しい数量化尺度ではない。数理的に利用しやすく、一方法として現実を線形化している尺度であると考えべきである。困難度や能力という値は、例えば温度や長さのような物理的なものではなく、あくまで仮想上のものであることからこのことはわかる。しかし、何らかの方法で困難度や能力を測り、仮想的な尺度を、可能な限り物理的に測定可能な尺度に成し得ようと試みられているのである。この命題に対し、ロジットコレクト以上

に妥当性があり数理的処理に適当な尺度が発見・考案されていないため、**Rasch** モデルでは、ロジットコレクトを基礎とした理論構築が行われている。**Rasch** モデルにおける項目困難度と受験者能力が線形尺度として認められ利用されているのはこのような理由からである。

2. Rasch モデルの一次元性と最尤推定

受験者能力と項目困難度が間隔尺度として配置されるということは、受験者能力と項目困難度が一次元であり、かつその間隔に意味があるということである。これにより、**Rasch** モデル適用上の大きな条件である二つのうちの一つの「項目および能力の一次元性」が導かれることになる。

「項目困難度及び受験者能力は一次元であり、困難度及び能力が定められた場合には、その大小と間隔に意味が見出せる。」

従って、「困難度の高い問題と低い問題では、どのような能力を持った受験者が受験しても困難度の高い問題のほうが常に正解確率は低く、逆にどのような困難度の問題であっても、能力の高い受験者と低い受験者では、高い受験者のほうが常に正解確率は高い」という条件が導かれる。なお、前述のもう一つの条件は各受験者の各項目に対する正誤反応は確率的に独立であるという局所独立性である。

実際にテストを行い、項目困難度と受験者能力を推定する場合には、**Rasch** モデルが条件とする一次元性が保たれている必要があるが、現実的にこの条件を満たしているかどうかを判定することが難しいことも事実である。もちろん、テスト前に項目・受験者を十分吟味することは必須である。現実的な方法として、まず、テストが **Rasch** モデルに適合しているとの前提で推定を行い、その結果の解析から困難度及び能力の一次元性の尺度となるような指標を取り出し、テストの一次元妥当性を吟味することも行われている。これには標準化残差を基にした **Outfit** 平均平方和と **Infit** 平均平方和を用いる方法などがある。これは後の章で **Rasch** モデル適合程度との関連も含めて述べることにする。

さて、実際のテストが **Rasch** モデルに適合していると一旦仮定して進めると、目標は各項目の困難度と各受験者の能力の二つの母数ベクトル（各項目困難度と各受験者能力をそれぞれベクトルと考えたもの）を推定することになる。母数の推定方法には、同時最尤推定、条件付き最尤推定、周辺最尤推定、ベイズ推定などさまざまな方法⁹⁾が考えられている。**Rasch** モデルに適用することを考えた場合には、項目数と受験者数がある程度大きいという仮定で、母数の事前分布を与えない条件付き最尤推定や同時最尤推定が用いられることも多い。以降は同時最尤推定の場合を仮定して議論を進めるが、統計的により良質な推定値が得られる条件付き最尤推定でも、傾向は同じになる。

同時最尤推定による母数推定の考え方は次のようである。まず、困難度及び能力の真値から構成される確率的な振舞いの結果として、正答・誤答反応が現れると考える。これを逆に、得られた正答・誤答反応は困難度と能力がどのような値である場合にどの程度の尤度（可能性）で生起するかを考える。つまり、項目困難度ベクトルと受験者能力ベクトルから正答・誤答反応が得られる尤度を計算する尤度関数を構成し、尤度の最も高い項目困難度ベクトルと受験者能力ベクトルを推定値として各困難度と各能力を推定するのである。数理詳細は本稿では触れないが、その方法を以下に簡潔にまと

める。

(1) 各項目困難度及び各受験者能力がある一定の値であれば、得られた正答・誤答反応データが生起する確率はいくらかを計算する。これが尤度で、各項目困難度及び各受験者能力から導かれるので尤度関数と呼ぶ。

(2) 尤度関数を最大にする各項目困難度及び各受験者能力を求める。

つまり、項目困難度と受験者能力は実際に得られた正答・誤答反応が一番起こりやすい値に推定されることになる。項目困難度と受験者能力から尤度を導出する際には、各反応が独立であるとの仮定から、各項目に対する各受験者の正誤反応が生起する同時確率を、確率の積を用いて求め、

$$L(\mathbf{U}|\boldsymbol{\theta}, \mathbf{B}) = \prod_{i=1}^M \prod_{j=1}^N [p_j(\theta_i)^{u_{ij}} (1-p_j(\theta_i))^{(1-u_{ij})}] \quad (4)$$

で表わされる尤度関数式 L を得る。但し、

\mathbf{U} : 正答・誤答反応行列

$\boldsymbol{\theta}$: 受験者能力値ベクトル

\mathbf{B} : 項目困難度ベクトル

u_{ij} : 受験者 i の項目 j に対する正答・誤答反応 (正答 = 1, 誤答 = 0)

M : 受験者人数

N : 項目数

である。しかし、積の形は扱いが難しいので、尤度関数 L の大小関係を保存したまま和の形で表すことができる対数変換を施す。結果的に、最大尤度を求める問題は、次のような対数尤度関数式 LL を最大化する $\boldsymbol{\theta}$ と \mathbf{B} を求める問題に帰着させることができる。

$$LL(\mathbf{U}|\boldsymbol{\theta}, \mathbf{B}) = \sum_{i=1}^M \sum_{j=1}^N [u_{ij} \ln p_j(\theta_i) + (1-u_{ij}) \ln(1-p_j(\theta_i))] \quad (5)$$

対数尤度 LL の意味は、各項目に対する各受験者の正答・誤答反応全域にわたり、正答であれば $u_{ij}=1, (1-u_{ij})=0$ となるので正解確率の対数値 $\ln p_j(\theta_i)$ を、誤答であれば $u_{ij}=0, (1-u_{ij})=1$ となるので不正解確率の対数値 $\ln(1-p_j(\theta_i))$ を、正誤に応じて選択的に総和計算したものとなる。

また、この対数尤度関数 LL 中の、受験者 i にのみ関係する部分対数尤度 LL_i は、

$$LL_i(\mathbf{u}_i|\theta_i, \mathbf{B}) = \sum_{j=1}^N [u_{ij} \ln p_j(\theta_i) + (1-u_{ij}) \ln(1-p_j(\theta_i))] \quad (6)$$

\mathbf{u}_i : 受験者 i の正答・誤答反応ベクトル

θ_i : 受験者 i の能力

と表すことができる。なお、元の対数尤度関数 LL との関係は、次のようになる。

$$LL(\mathbf{U}|\boldsymbol{\theta}, \mathbf{B}) = \sum_{i=1}^M LL_i(\mathbf{u}_i|\theta_i, \mathbf{B}) \quad (7)$$

式(6)においても、前述の通り受験者 i の各項目に対する正誤に応じて $\ln p_j(\theta_i)$ または $\ln(1-p_j(\theta_i))$

を選択的に総計していることになる。対数尤度 LL で総計される、一つの正誤反応に対する項のことを以下では対数尤度項と呼ぶことにする。 $p_j(\theta_i)$ は項目 j に対する受験者 i の正答確率であり、 $p_j(\theta_i)$ は θ_i の関数であるから、各項目に対する $\ln p_j(\theta_i)$ あるいは $\ln(1 - p_j(\theta_i))$ の振る舞いを、仮に $\beta_j = 0$ として θ_i だけで表わすと、図 1 のようになる。グラフの下部には θ_i から導かれる正解確率と不正解確率を併記している。どの項目の困難度も $\beta_j = 0$ と仮定したが、能力 θ と困難度 β は同次元の間隔尺度であるため、 β の値が a 増加すれば、グラフの θ_i 軸を左に a 平行移動するだけである。また $\beta = \theta$ のときに正解確率は 0.5 となる。

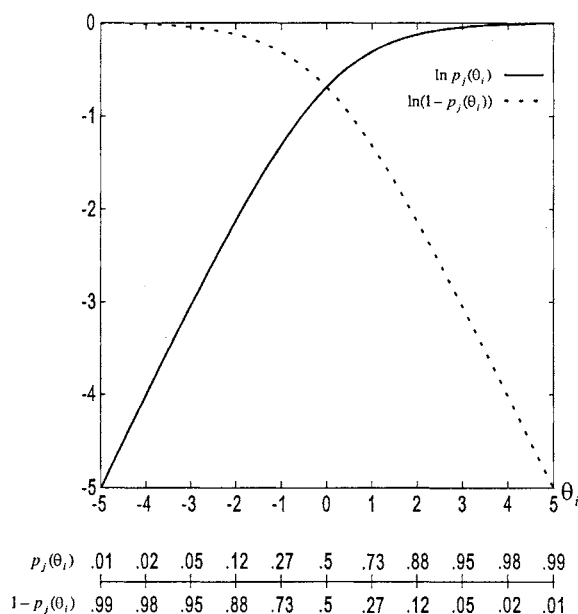


図 1 正誤による対数尤度個別項の値

図 1 より、正解確率が高い場合に正答している場合と正解確率が低い場合に誤答している場合には、対数尤度項はそれほど小さくならないが、逆に正解確率が低い場合に正解している、あるいは、正解確率が高いのに誤答している場合には、対数尤度項は大きく下がり、このような場合には最適化に関して大きなペナルティを課していることがわかる。表 1 には正誤による対数尤度項の変化を示す。表 1 より、 $\beta_j = 0$ で θ_i の値が 0 である場合、正解確率 $p_j(\theta_i)$ も不正解確率 $(1 - p_j(\theta_i))$ も 0.5 であるが、この場合には受験者が正答しようが誤答しようが、対数尤度項は -0.69 となる。しかし、 $\beta_j = 0$ で θ_i の値が 3 の場合を考えてみると、正答した場合には、対数尤度項は -0.05 であるが、誤答した場合には -3.05 と誤答に対して大きなペナルティがあることがわかる。逆に $\beta_j = 0$ で θ_i の値が -3 の場合を考えてみると、正答した場合には対数尤度項は -3.05 となり、誤答した場合には -0.05 と逆転して、正答に対して大きなペナルティがあることがわかる。例えば、 θ_i が 0 の時を基準に考えると、 θ_i が 3 に増えれば、仮に正答していれば対数尤度項は 0.64 増加するが、誤答していれば 2.64 も減少することとなる。逆に言うと、正答に対して、そこから導かれる正解確率を 0.4~0.5 辺りまで下げたとして

も対数尤度項の減少は緩やかだが、正解確率を 0.1～0.2 辺りまで下げると、対数尤度項は急激に減少してしまう傾向がある。これは、誤答に対しても同様である。

つまり、最尤推定時に対数尤度を下げる大きな要因となる対数尤度項があると、局所的にはこのような正答データ（誤答データ）に対して推定される正解確率（不正解確率）が、極力 0.4～0.5 を下回らないようにする作用が働くということがわかる。

表 1 正誤による対数尤度項の変化

θ	$p(\theta)$	$\ln(p(\theta))$	$1-p(\theta)$	$\ln(1-p(\theta))$
-5.0	0.007	-5.007	0.993	-0.007
-4.5	0.011	-4.511	0.989	-0.011
-4.0	0.018	-4.018	0.982	-0.018
-3.5	0.029	-3.530	0.971	-0.030
-3.0	0.047	-3.049	0.953	-0.049
-2.5	0.076	-2.579	0.924	-0.079
-2.0	0.119	-2.127	0.881	-0.127
-1.5	0.182	-1.701	0.818	-0.201
-1.0	0.269	-1.313	0.731	-0.313
-0.5	0.378	-0.974	0.622	-0.474
0.0	0.500	-0.693	0.500	-0.693
0.5	0.622	-0.474	0.378	-0.974
1.0	0.731	-0.313	0.269	-1.313
1.5	0.818	-0.201	0.182	-1.701
2.0	0.881	-0.127	0.119	-2.127
2.5	0.924	-0.079	0.076	-2.579
3.0	0.953	-0.049	0.047	-3.049
3.5	0.971	-0.030	0.029	-3.530
4.0	0.982	-0.018	0.018	-4.018
4.5	0.989	-0.011	0.011	-4.511
5.0	0.993	-0.007	0.007	-5.007

3. 一次元性が満たされない場合の推定値への影響

Rasch モデルの適用条件は項目困難度及び受験者能力の一次元性が保たれていることであるが、一次元性の確認は非常に困難な問題でもある。ここでは仮にある実際のテストで一次元性が保たれている場合に、他次元を測定する項目が追加された場合、推定値にどのような影響が出るのか考察する。

まず、一次元性を満たして、なおかつ項目数、受験者数とも十分に大きい理想的なテストが実施できたとすると、最尤推定の漸近性により推定値は真値に近いと考えることができる。ここで推定されるのは、対数尤度を最大にする項目困難度と受験者能力である。つまり、対数尤度の意味である「得られた反応（正答あるいは誤答）が起こる確率の対数の総和」が最大になる項目困難度と受験者能力が得られることになる。また、対数尤度式で総計される各項目 $\ln p_j(\theta_i)$ 又は $\ln(1-p_j(\theta_i))$ の値は、概ね、-0.5 以上程度、つまり得られた正答・誤答反応は概ね確率 0.6 以上で生起しているであろうと予想することに無理はない。なお、この予想は各個別反応によっては例外もあるが、傾向の問題として議論しているので、例外があることで議論の方向性に影響はない。

一旦推定値を得た後、このテストにまったく測定次元の異なる項目を追加したテストを考えると、追加した項目はすでに存在する各項目と正解関連性を持たない。従って、先ほど推定した値を用いて

計算される対数尤度関数の追加した項目に関する項には、比較的多くの絶対値の大きい負値が混入することになる。絶対値の大きさの程度は、正答・誤答反応が正解確率・不正解確率から外れれば外れるほど急激に増加して絶対値の大きな負の値になることは前節より明らかである。絶対値の大きい負値の存在は対数尤度値の減少を招き、推定値の変更を引き起こす作用を持つ。この時、追加したテストを含めて再推定すると、推定値の推移にどのような傾向が出るのかは次のように論説することができる。

測定次元の異なる項目をもまとめて一次元性を有していると考える **Rasch** モデルの前提から、一次元適合性の高い状況に、他次元を測る項目が追加された場合、正誤反応の正解確率を 0.5 近傍に近づけることが効果的となる。結果的に、困難度がさほど変わらない項目を集めて、能力のさほど変わらない受験者に与えた場合と同じような状態が尤もらしいとの結論を得ることになる。

この結論に具体的な説明を加えると、次のようになる。追加された項目の困難度は、そもそも推定されていない。しかし、その項目に正解した受験者数が項目困難度の最小十分統計量⁵⁾となっているので、他の項目と比較して追加項目の困難度推定値の範囲を定め、仮に範囲の中心を推定値と考えても以降の議論に差し支えはない。また、追加項目に正解した受験者数と同じ正解受験者数を持つ項目があれば、その項目の困難度推定値と等しくなる。ここでは、追加前の項目のうち困難度推定値が最大のものと同じ正解者数である項目を追加したとする。正解者数は全項目中最少である。追加した項目に対する各受験者の正誤反応は、すでに推定された受験者能力に関係なく、多くの受験者が誤答で、正答がいくつかある状態となる。従って、項目が追加される前の推定値を用いて追加された項目の対数尤度項を足し合わせると、対数尤度に「能力よりも困難度が相当高いのに正解している」ゆえに起こる絶対値の大きな負値混入の可能性が高くなる。対数尤度をさらに大きくするために考えられることは、項目困難度推定値の移動である。項目追加前には **Rasch** モデルによく適合した正誤反応を示していることから、各正誤反応の対数尤度項は-1 を大きく下回らない結果となっているであろうが、追加した項目についてはこれまでの正誤反応とは関連せずに正誤が出現する。特にここでは困難度が全項目中最大であるので、誤答はあり得ることだが、正答は確率的に困難である。このような状況でも受験者能力が低いのに正答しているといった、予期せぬ正答反応がいくつか生じることになり、対数尤度減少の可能性が高くなる。モデル上は追加した項目も一次元性が保たれている前提であるから、追加項目による対数尤度の減少が推定値の移動によって補えるならば、再推定による推定値の移動が起こることになる。この際、追加項目による対数尤度項の負値の絶対値を小さくする、つまり、予期せぬ正答反応に対して項目の正解確率を上げるために、追加した項目の困難度を減少させる推定値移動の可能性を順を追って考えてみる。

まず、追加した項目の予期せぬ正答反応は対数尤度の大きな減少を生じさせる。追加した項目に誤答している場合には、追加項目の困難度推定値が最大であることを考えると、十分に妥当性のある結果となっている。つまり追加項目への正答反応は対数尤度を大幅に減少させるが、誤答反応からの影

響は相当緩やかである。このため、対数尤度の大幅な増加を狙って追加項目の困難度を減少させ、正答反応に対する正解確率を増加させる方策を考えてみる。正答反応に対しては、対数尤度項の値が、 -1 を大きく下回る対数尤度項が出現する可能性が高いが、誤答反応からは絶対値の大きな負値はなく、正解確率がきわめて低い状態で誤答しているという状況になっている。追加項目の困難度を少し下げれば、正答反応に対する対数尤度項は鋭敏に反応して大きく増加する。誤答反応に対しては、対数尤度項は減少することとなるが、もともと緩やかに反応する領域にあるので、減少の程度は緩やかである。つまり、正答反応に対する対数尤度の増加が誤答反応に対する対数尤度の減少を上回っている限り、追加した項目の困難度は減少することになる。追加項目の困難度は全項目中最大であるので、追加した項目の困難度が下がれば、同じ程度の困難度を持つ項目は必然的に困難度を下げることになる。困難度減少の割合は困難度が近いほど顕著であり、困難度が遠くなればその割合はそれほど影響を受けないことになる。

また、項目困難度推定値が移動すれば各受験者能力推定値にも影響して、再推定時には同じ傾向で移動することとなる。現実には即した表現を用いると、次のようになる。

「一次元性適合データに測定次元の異なる項目を追加して再推定を行えば、項目追加前の各項目困難度推定値は再推定によって分布幅が狭まる。同時に、各受験者能力推定値も分布幅が狭まる。」

上記に関する個別の状況についての詳細は、以下のようにまとめることができる。

一次元適合性の高い項目が N 個、受験者が M 人のテストの結果から得られた各項目困難度推定値を $\hat{\beta}_j (j=1,2,\dots,N)$ 、各受験者能力推定値を $\hat{\theta}_i (i=1,2,\dots,M)$ とし、便宜上 $\hat{\beta}_1 \leq \hat{\beta}_2 \leq \dots \leq \hat{\beta}_N$ 、及び、 $\hat{\theta}_1 \leq \hat{\theta}_2 \leq \dots \leq \hat{\theta}_M$ となるように並べ替えておくことにする。ここに測定次元の異なる $N+1$ 個目の項目を追加し、同じ受験者にこの項目を追加受験させる、あるいは、仮想的に追加前と追加後の二種類のテストを行ったと仮定する。 $N+1$ 個目の項目の正解者数 C_{N+1} が、 1 から N 個目の項目の正解者数 C_1, C_2, \dots, C_N の配置される場所によって、仮に追加項目の困難度推定値 $\hat{\beta}_{N+1}$ を決定しておく。項目追加による再推定値の移動傾向は次のものとなる。

(1) $\hat{\beta}_{N+1} \leq \hat{\beta}_1$ の場合：

項目困難度推定値のうち、最大値 $\hat{\beta}_N$ はあまり移動せず、最小値 $\hat{\beta}_1$ が大きく増加する。各推定値 $\hat{\beta}_i$ は分布の範囲が狭くなった影響でそれぞれ増加方向に移動する。

(2) $\hat{\beta}_N \leq \hat{\beta}_{N+1}$ の場合：

項目困難度推定値のうち、最小値 $\hat{\beta}_1$ はあまり移動せず、最大値 $\hat{\beta}_N$ が大きく減少する。各推定値 $\hat{\beta}_i$ は分布の範囲が狭くなった影響でそれぞれ減少方向に移動する。

(3) $\hat{\beta}_1 < \hat{\beta}_{N+1} < \hat{\beta}_N$ の場合：

$\hat{\beta}_{N+1}$ と同等の項目困難度推定値を持つ項目を中心に、各項目困難度推定値が集まってきて分布

幅が減少する。

項目困難度の移動と共に、同時に、各受験者能力も同様に移動する。また、以上の議論は、Raschモデルの対称性により、項目困難度と受験者能力を入れ替えても成立することは明らかである。

なお、受験者全員が正答あるいは誤答した項目、および全項目に正答あるいは誤答した受験者については、同時最尤推定においては推定値が求められない（推定値は ∞ または $-\infty$ ）が、事前分布を利用する、例えばベイズ推定では推定値が求まる。しかし、このような項目・受験者が他の項目・受験者と次元性を有しているかどうかは、正誤反応を利用した如何なる方法を用いても確認することができない。従って、このような項目・受験者を除いた正誤反応を用いるか、あるいは、正誤反応を用いない別の方法で次元性を吟味するといった手段が必要となる。

本節の最後に、同時最尤推定や条件付き最尤推定以外の母数事前分布を利用する推定方法を用いる場合には、これまでの議論は成り立たない場合があることを付記しておく。

4. 同時最尤推定値と正答数の関係

ここでは、最尤推定時に、各項目期待正答者数と各項目正答者数が、また、各受験者期待正答数と各受験者正答数が、それぞれ等しくなるという関係についての確認を補足的に行う。

まず、式(5)を、項目困難度 β_j および受験者能力 θ_i でそれぞれ偏微分して、

$$\frac{\partial}{\partial \beta_j} LL(\mathbf{U}_j | \boldsymbol{\theta}, \beta_j) = \sum_{i \in \{i | U_{ij}=1\}} \frac{\frac{\partial}{\partial \beta_j} p_j(\theta_i)}{p_j(\theta_i)} + \sum_{i \in \{i | U_{ij}=0\}} \frac{\frac{\partial}{\partial \beta_j} p_j(\theta_i)}{1 - p_j(\theta_i)} \quad (8)$$

$$\frac{\partial}{\partial \theta_i} LL(\mathbf{U}_i | \theta_i, \mathbf{B}) = \sum_{j \in \{j | U_{ij}=1\}} \frac{\frac{\partial}{\partial \theta_i} p_j(\theta_i)}{p_j(\theta_i)} + \sum_{j \in \{j | U_{ij}=0\}} \frac{\frac{\partial}{\partial \theta_i} p_j(\theta_i)}{1 - p_j(\theta_i)} \quad (9)$$

を得る。ここで、

$$\frac{\partial}{\partial \beta_j} p_j(\theta_i) = -\frac{e^{-\theta_i + \beta_j}}{(1 + e^{-\theta_i + \beta_j})^2}, \quad \frac{\partial}{\partial \theta_i} p_j(\theta_i) = \frac{e^{-\theta_i + \beta_j}}{(1 + e^{-\theta_i + \beta_j})^2} \quad (10)$$

であるから、

$$\begin{aligned} \frac{\partial}{\partial \beta_j} LL(\mathbf{U}_j | \boldsymbol{\theta}, \beta_j) &= -\sum_{i \in \{i | U_{ij}=1\}} \frac{(1 + e^{-\theta_i + \beta_j}) e^{-\theta_i + \beta_j}}{(1 + e^{-\theta_i + \beta_j})^2} + \sum_{i \in \{i | U_{ij}=0\}} \frac{e^{-\theta_i + \beta_j}}{\left(1 - \frac{1}{1 + e^{-\theta_i + \beta_j}}\right) (1 + e^{-\theta_i + \beta_j})^2} \\ &= -\sum_{i \in \{i | U_{ij}=1\}} \frac{e^{-\theta_i + \beta_j}}{1 + e^{-\theta_i + \beta_j}} + \sum_{i \in \{i | U_{ij}=0\}} \left(\frac{e^{-\theta_i + \beta_j}}{1 + e^{-\theta_i + \beta_j}} - \frac{1 + e^{-\theta_i + \beta_j}}{1 + e^{-\theta_i + \beta_j}} \right) (1 + e^{-\theta_i + \beta_j}) \\ &= -\sum_{i \in \{i | U_{ij}=1\}} \left(\frac{1 + e^{-\theta_i + \beta_j}}{1 + e^{-\theta_i + \beta_j}} - \frac{1}{1 + e^{-\theta_i + \beta_j}} \right) + \sum_{i \in \{i | U_{ij}=0\}} \left(\frac{e^{-\theta_i + \beta_j}}{1 + e^{-\theta_i + \beta_j}} - \frac{1 + e^{-\theta_i + \beta_j}}{1 + e^{-\theta_i + \beta_j}} \right) (1 + e^{-\theta_i + \beta_j}) \\ &= -\sum_{i \in \{i | U_{ij}=1\}} \left(1 - \frac{1}{1 + e^{-\theta_i + \beta_j}} \right) + \sum_{i \in \{i | U_{ij}=0\}} \frac{1}{1 + e^{-\theta_i + \beta_j}} \end{aligned}$$

$$\begin{aligned}
&= -\sum_{i \in \{i|U_{ij}=1\}} (1 - p_j(\theta_i)) + \sum_{i \in \{i|U_{ij}=0\}} p_j(\theta_i) \\
&= -\sum_{i=1}^M u_{ij} + \sum_{i=1}^M p_j(\theta_i)
\end{aligned} \tag{11}$$

となる。同様に、

$$\begin{aligned}
\frac{\partial}{\partial \theta_i} LL(\mathbf{U}_i | \theta_i, \mathbf{B}) &= \sum_{j \in \{j|U_{ij}=1\}} \frac{(1 + e^{-\theta_i + \beta_j}) e^{-\theta_i + \beta_j}}{(1 + e^{-\theta_i + \beta_j})^2} - \sum_{j \in \{j|U_{ij}=0\}} \frac{e^{-\theta_i + \beta_j}}{\left(1 - \frac{1}{1 + e^{-\theta_i + \beta_j}}\right) (1 + e^{-\theta_i + \beta_j})^2} \\
&= \sum_{j \in \{j|U_{ij}=1\}} \frac{e^{-\theta_i + \beta_j}}{1 + e^{-\theta_i + \beta_j}} - \sum_{j \in \{j|U_{ij}=0\}} \left(\frac{e^{-\theta_i + \beta_j}}{1 + e^{-\theta_i + \beta_j} - \frac{1 + e^{-\theta_i + \beta_j}}{1 + e^{-\theta_i + \beta_j}}} \right) (1 + e^{-\theta_i + \beta_j}) \\
&= \sum_{j \in \{j|U_{ij}=1\}} \left(\frac{1 + e^{-\theta_i + \beta_j}}{1 + e^{-\theta_i + \beta_j}} - \frac{1}{1 + e^{-\theta_i + \beta_j}} \right) - \sum_{j \in \{j|U_{ij}=0\}} \left(\frac{e^{-\theta_i + \beta_j}}{1 + e^{-\theta_i + \beta_j} - \frac{1 + e^{-\theta_i + \beta_j}}{1 + e^{-\theta_i + \beta_j}}} \right) (1 + e^{-\theta_i + \beta_j}) \\
&= \sum_{j \in \{j|U_{ij}=1\}} \left(1 - \frac{1}{1 + e^{-\theta_i + \beta_j}} \right) - \sum_{j \in \{j|U_{ij}=0\}} \frac{1}{1 + e^{-\theta_i + \beta_j}} \\
&= \sum_{j \in \{j|U_{ij}=1\}} (1 - p_j(\theta_i)) - \sum_{j \in \{j|U_{ij}=0\}} p_j(\theta_i) \\
&= \sum_{j=1}^N u_{ij} - \sum_{j=1}^N p_j(\theta_i)
\end{aligned} \tag{12}$$

となる。式(11)、式(12)とも、 $= 0$ と置くと、

$$\begin{aligned}
\frac{\partial}{\partial \beta_j} LL(\mathbf{U}_j | \theta, \beta_j) &= -\sum_{i=1}^M u_{ij} + \sum_{i=1}^M p_j(\theta_i) \\
&= -(\text{正答者数}) + (\text{期待正答者数}) \\
&= 0
\end{aligned} \tag{13}$$

$$\begin{aligned}
\frac{\partial}{\partial \theta_i} LL(\mathbf{U}_i | \theta_i, \mathbf{B}) &= \sum_{j=1}^N u_{ij} - \sum_{j=1}^N p_j(\theta_i) \\
&= (\text{正答数}) - (\text{期待正答数}) \\
&= 0
\end{aligned} \tag{14}$$

となり、式(5)の対数尤度関数を最大化する項目困難度 $\beta = \hat{\beta}$ 、受験者能力 $\theta = \hat{\theta}$ は式(13) および式(14)を満たすことがわかる。つまり、得られた正答・誤答反応の各項目正答者数、各受験者正答項目数は、推定する項目困難度 $\hat{\beta}$ 、受験者能力 $\hat{\theta}$ を基にして得られる期待値とそれぞれ等しくなることがわかる。

なお、受験者全員が正答している項目困難度、あるいは、全ての項目に正答している受験者能力が最尤推定できないことは式(13)、式(14)より明らかである。

Ⅲ. Rasch 測定母数推定値分布幅、Rasch 測定適用妥当性、及び Outfit・Infit モデル適合度指標の相互連関性

一次元性充足度の高いデータに異次元性の強い一項目追加による Rasch 測定上での項目群・受験者群母数推定値分布幅の前章で論証された狭小性の一端を例示するものが表 2 である。なお、その仮想元データ作成法は静¹⁰⁾(2007)によって記述されている方法(pp. 188-191)であり、以下の通りである。Excel 上で行列それぞれに-7.0 から 7.0 に渡って 1.0 を間隔とする 15 項目、及び、-5.0 から 5.0 に渡って 0.5 を間隔とする 21 名を配置する。各項目設定値と各受験者設定値の組み合わせの上でモデル期待正答確率を出力する。その一対すべてに Excel 上で RAND 関数を適用してデータ 0・1 生起が与えられる。この 0・1 データの事前設定値は項目群と受験者群それぞれについて広い分布幅で以って構成されている故に、一次元性充足度の大変に高いデータとなっている。その元データに、難度最低第 1 番項目の 01 配置逆転項目、難度最高第 15 番項目の 01 配置逆転項目、難度中位第 8 番項目の 01 配置変更極端項目を第 16 番目にそれぞれ追加して、二次元性を特徴とする三つのデータを新たに作成する。各データ Rasch 測定上での項目群・受験者群母数推定値と各データ基本統計量を表 2 に与えている。異次元一項目追加による母数推定値分布幅狭小性、特に、受験者群についてのその傾向が観察される。

表 2 II 章論証異次元追加一項目による項目群・受験者群母数推定値分布幅の狭小性

番号	一次元性の高い 元データ		困難度-6.09の01配置 逆転項目追加データ		困難度6.48の01配置 逆転項目追加データ		困難度0.06の01配置 変更極端項目追加データ	
	15項目	21名	16項目	21名	16項目	21名	16項目	21名
1	-6.09	-6.21	-4.92	-4.67	-5.53	-4.80	-4.06	-3.57
2	-4.82	-5.07	-3.81	-3.89	-4.48	-4.80	-3.09	-2.85
3	-4.82	-4.15	-3.81	-3.17	-4.48	-3.98	-3.09	-2.23
4	-3.83	-3.29	-2.98	-2.46	-3.68	-3.22	-2.41	-1.66
5	-2.97	-2.40	-2.26	-1.73	-2.98	-2.46	-1.87	-1.11
6	-2.20	-1.48	-1.63	-0.96	-2.35	-1.67	-1.42	-0.59
7	-0.95	-0.54	-0.56	-0.16	-1.27	-0.87	-0.66	-0.06
8	0.06	-0.54	0.34	-0.16	-0.38	-0.87	-0.01	-0.06
9	0.97	0.38	1.16	0.64	0.45	-0.07	0.57	0.46
10	1.42	0.38	1.56	0.64	0.85	-0.07	0.86	0.46
11	2.85	0.38	2.83	0.64	2.11	-0.07	1.77	0.46
12	3.94	1.30	3.78	1.45	3.08	0.74	2.50	1.00
13	4.59	1.30	4.33	1.45	3.65	0.74	2.96	0.46
14	5.37	2.24	4.98	2.28	4.34	1.56	3.54	1.00
15	6.48	2.24	5.91	2.28	5.33	1.56	4.42	1.00
16		3.22	-4.92	3.13	5.33	2.40	-0.01	1.58
17		3.22		3.13		2.40		1.58
18		4.19		4.00		3.22		2.20
19		4.19		4.00		3.22		2.20
20		5.20		4.91		4.04		2.88
21		6.44		4.91		4.90		3.66
最大値	6.48	6.44	5.91	4.91	5.33	4.90	4.42	3.66
平均値	0.00	0.52	0.00	0.77	0.00	0.09	0.00	0.32
最小値	-6.09	-6.21	-4.92	-4.67	-5.53	-4.80	-4.06	-3.57
範囲	12.57	12.65	10.83	9.58	10.86	9.70	8.48	7.23
標準偏差	4.06	3.42	3.64	2.83	3.69	2.84	2.56	1.85

前章論証に基づいて表 2 に例示した Rasch モデルにおける同時最尤推定法の特性から帰結される母数推定値分布幅狭小性による示唆はデータ多次元性ないしはデータ Rasch 測定妥当性欠如である。こ

の Rasch モデル母数推定値の分布幅狭小性については、Rating Scale モデル測定上での閾値母数推定値分布幅との関連性で以って Andrich¹¹⁾ (1988)により次の様な言及がなされている。『大きい分布幅を持つ項目困難度推定値群が所与とされた全体 0・1 データ Rasch 測定において、推定値中間範囲に集中した隣接困難度推定値間の差が小さい項目群のみから成る部分テストは、推定値全体範囲を網羅する隣接困難度推定値間の差が大きい項目群のみから成る部分テストに較べて、受験者群能力弁別性の程度が大きい』(Andrich¹¹⁾, 1988, pp. 370-371)。著者はこの指摘を確認済みであり、Rasch 測定上での項目困難度母数推定値分布幅の狭小性が、テスト項目群全体としての受験者群能力弁別性増大及び受験者能力推定値分布幅の狭小化への方向性につながると示唆されている。これは、0・1 データ素点得点上での、例えば、易しい問題 100 項目から成るテスト結果として、平均 80 点を中心として 70 点から 90 点に渡って狭小に分布する受験者群得点の観察として容易に理解される。一方、Rasch 測定における母数推定値分布幅の狭小性と Rasch 測定妥当性欠如との連関性についての明確な言及は著者取得の Rasch 測定関連文献上で眼にしていない。その連関性は識者により当然とされているかも知れないが、その認識はテスト問題作成者、並びに、Rasch 測定適用者にとっては本質的認識最重要事項の一つである。又、データ次元性を意味するテスト構成概念妥当性の検証は Rasch 測定上では一般的に Outfit・Infit モデル適合度指標によってなされる (Linacre¹²⁾, 1989, p. 4; Wright and Masters¹³⁾, 1982, pp. 93-101)。この指標の特性と限界性についての代表的な論考は Karabatsos¹⁴⁾ (2000)と Smith, Schumacker, and Bush¹⁵⁾ (1998)であるが、そのモデル適合度指標と、Rasch 測定における母数推定値分布幅の狭小性及び Rasch 測定妥当性の関係についての明示的な言及はその二論考に見出されない。以上を踏まえて、Rasch モデルの母数推定値分布幅、Rasch 測定妥当性、及び、Outfit・Infit モデル適合度指標の相互連関性を本章で考察する。

以降の議論展開への準備段階として、Rasch モデル適合度指標の各項目に関するその算出式(Wright and Masters¹⁴⁾, 1982, pp. 98-101)再掲(井澤¹⁶⁾, 2007, p. 2、参照)の上でその数理脆弱性に触れておく。なお、受験者数 M に替えて項目総数 N を挿入して、総和の方法を $\sum_{i=1}^M$ から $\sum_{j=1}^N$ に変換すれば各受験者についてのモデル適合度指標の算出式となる。また、本稿全体に合わせて添え字などの表現は変更している。

$$\text{Rasch 測定期待値 (期待正答確率): } P_{ij} = \frac{\exp(\theta_i - \beta_j)}{\exp(\theta_i - \beta_j) + 1}$$

$$\text{残差: } y_{ij} = u_{ij} - P_{ij}$$

$$\text{標準化残差: } z_{ij} = \frac{u_{ij} - P_{ij}}{\sqrt{P_{ij}(1 - P_{ij})}}$$

$$\text{Outfit } \chi^2 \text{ 平均平方統計量: } O_j = \frac{\sum_{i=1}^M z_{ij}^2}{M}$$

$$\text{Outfit } t \text{ 値:} \quad \text{Outfit } t_i = (O_j^{\frac{1}{3}} - 1) \left(\frac{3}{\text{Outfit } \chi^2 \text{ 平均平方統計量標準偏差}} \right) + \left(\frac{\text{Outfit } \chi^2 \text{ 平均平方統計量標準偏差}}{3} \right)$$

$$\text{Infit } \chi^2 \text{ 平均平方統計量:} \quad I_j = \frac{\sum_{i=1}^M y_{ij}^2}{\sum_{i=1}^M (P_{ij}(1 - P_{ij}))}$$

$$\text{Infit } t \text{ 値:} \quad \text{Infit } t_i = (I_j^{\frac{1}{3}} - 1) \left(\frac{3}{\text{Infit } \chi^2 \text{ 平均平方統計量標準偏差}} \right) + \left(\frac{\text{Infit } \chi^2 \text{ 平均平方統計量標準偏差}}{3} \right)$$

上記 Outfit・Infit モデル適合度指標算出数理の本質的特徴は、Rasch 測定における標準化残差の分布が正規分布であるという仮定である (Linacre¹⁷⁾, 1998, p. 271; Mead and Wright, 1980, cited in Smith and Hedges¹⁸⁾, 1982, p. 46; Smith¹⁹⁾, 1988; Smith and Hedges¹⁸⁾, 1982, p. 50、参照)。『データが Rasch モデルに適合しているならば、Rasch 測定における標準化残差の約 5% 並びに約 1% が、それぞれ絶対値 2 並びに 3 の範囲外の値となる』と Linacre²⁰⁾ (1989-2001, p. 52) により指摘されている。但し、観測値から能力母数と困難度母数のいずれもが Rasch 測定により最尤推定された上で、そのモデル期待値が算出されるから、観測値と期待値は独立しておらず (Van den Wollenberg, 1979, cited in Smith and Hedges¹⁸⁾, 1982, p. 46)、又、観測値が二値変数であり期待値が連続変数である (George, 1979, cited in Smith and Hedges¹⁸⁾, 1982, p. 46) 故に、その標準化残差分布の正規性仮定に対する批判も多くなされている (Divgi²¹⁾, 1986, p. 293; George, 1979, Gustafsson, 1980, Hambleton et al, 1978, Reckase, 1981, Van den Wollenberg, 1979, cited in Smith and Hedges¹⁸⁾, 1982, pp. 45-46)。さらに、標準化残差の線形性欠如に係わる Outfit・Infit モデル適合度指標の限界性が Karabatsos¹⁴⁾ (2000) により例証・詳述されている (井澤¹⁶⁾、2007、p. 2、参照)。Rasch モデル適合度指標が標準化残差に数理基盤を置く限りにおいて、その絶対的な信頼性は保証され得ないとの結論が Karabatsos¹⁴⁾ (2000) により下された (p. 170) 上で、「最良適合法」を含む複雑な数理を伴ういくつかのモデル適合検定法への言及がなされている (pp. 171-173)。

上記の Outfit・Infit 統計量によるモデル適合度検知が不完全であることには留意するべきである。然しながら、Rasch 測定分析プログラムの出力においては、通常 Outfit・Infit 統計量と分離信頼性指標以外は出力されないために、Rasch 測定適用者が各項目と各受験者についてのモデル適合の上で参照する指標は主に Outfit・Infit 統計量であるというのが現状である。一般的に、 χ^2 平均平方統計量が 0.7 (Wright and Linacre²²⁾, 1994, p. 370) あるいは 0.75 (McNamara²³⁾, 1996, p. 173) から 1.3 (McNamara²³⁾, 1996, p. 173; Wright and Linacre²²⁾, 1994, p. 370) の範囲内にある値を持つ受験者

と項目は、Rasch モデルへの適合度において極端な過不足性がないとみなされる。但し、Smith and Suh²⁴⁾ (2003)により、受験者標本数の大きい、2000名における受験者群能力高低に基づいた二分割標本 t 値上で、その絶対値が 2 未満に対応する Outfit χ^2 平均平方統計量上での妥当な範囲は、80 個から成る項目群に関しては 0.9 以上 1.1 以内である (p. 161) と一事例上で実証的に示されている。受験者標本が大きくなるにつれて項目群に付される上記 χ^2 平均平方統計量標準化 t 値は過大評価される傾向があり (Wilson²⁵⁾, 2005, p. 129)、安定性の高い χ^2 平均平方統計量への参照が好ましいと McNamara²³⁾ (1996, p. 181)により指摘されている。一方、Rasch モデル不適合項目検出精度の観点からは、モデル適合度指標 t 値への参照が優るとの研究発表を Smith and Suh²⁴⁾ (2003)が与えている。更に、受験者数変動に対して項目群に付される t 値が、そのはずれ値個数の割合の点で χ^2 平均平方統計量によるそのはずれ値個数の割合に優って、より安定した分布属性を持っているとの人工模擬データ分析報告 (Smith, Schumacker, and Bush¹⁵⁾, 1998)もある。一般的に、 χ^2 平均平方統計量 0.7 未満ないし t 値 -2 以下で示唆されるモデル過剰適合項目は、他のある一つあるいは複数の項目との内容同一性 (Shen²⁶⁾, 1996, p. 485)による統計的情報力寄与欠如を間接的に示しており (Linacre²⁷⁾, 1997, p. 546)、『[構成された]尺度の意味を乱すものではない』(Linacre²⁸⁾, 2000, p. 755; Linacre and Wright²⁹⁾, 1994, p. 361)とされている。従って、モデル過剰適合項目は、一般的にデータからの除去対象とはならず (Henning³⁰⁾, 1987, p. 123)、むしろ、モデル適合項目 (Karabatsos³¹⁾, 2004, pp. 633-634; Wright³²⁾, 1996, p. 15)ないしはモデル最良適合項目 (Wright and Stone³³⁾, 2004, pp. 66-67)と識者により思慮されている節が窺われる。以上が Rasch 測定プログラム適用により各項目と各受験者に対して自動的に出力されるモデル適合度指標 Outfit・Infit 統計量についての識者認識上での著者による現況理解である。

標準化残差 $z_{ij} = (u_{ij} - P_{ij}) / \sqrt{P_{ij}(1 - P_{ij})}$ は Rasch 項目分析モデル上での期待正答確率 $P_{ij} = \exp(\theta_i - \beta_j) / (\exp(\theta_i - \beta_j) + 1)$ の関数である。その標準化残差分布の正規性仮定に基づいて、標準化残差平方和 $\sum z_{ij}^2$ の平均が各項目と各受験者に付される Outfit χ^2 平均平方統計量であり、それに Wilson-Hilferty 変換 (静¹⁰⁾, 2007, p. 318, 参照)が施された標準化統計量が Outfit t 値である。 $P_{ij}(1 - P_{ij})$ が各項目に対する各受験者の応答上でのモデル期待分散である。標準化残差平方 z_{ij}^2 にそのモデル期待分散を乗じた上での、各項目と各受験者に関するその総和 $\sum (z_{ij}^2 P_{ij}(1 - P_{ij}))$ を各項目と各受験者について個別にモデル期待分散総和 $\sum (P_{ij}(1 - P_{ij}))$ で除した値が Infit χ^2 平均平方統計量であり、その標準化統計量が Infit t 値である。従って、Outfit・Infit χ^2 平均平方統計量は標準化残差平方 z_{ij}^2 の関数であり、そのモデル期待値は 1 とされて、その標準化統計量 t 値のモデル期待値は 0 とされる。さらに、標準化残差平方 z_{ij}^2 はモデル期待分散 $P_{ij}(1 - P_{ij})$ の関数であるから、Outfit と Infit の χ^2 平均平方統計量はいずれもモデル期待分散 $P_{ij}(1 - P_{ij})$ の関数でもある。モデル期待分散 $P_{ij}(1 - P_{ij})$ は受験者能力母数 θ_i に関するモデル期待正答確率 P_{ij} の変化率であり、それは Rasch モデル上での項目 j の

弁別力と定義されている (Andrich¹¹⁾, 1988, pp. 371-372、参照)。項目 j の弁別力は、『単一の数値ではなく、 θ_n の関数である』(静¹⁰⁾, 2007, p. 280) ことに留意される。Rasch モデル上での項目群と受験者群は対称であるから、項目 j に対応する受験者 i の弁別力についても同様である。これにより、Outfit と Infit の χ^2 平均平方統計量はいずれも項目 j と受験者 i それぞれについての弁別力に大きく依拠していると理解される。『Rasch 測定上での項目群 Infit 統計量によって供与される診断情報は、0・1 データ上での点双列相関係数と項目反応理論上での項目弁別力母数推定値それぞれによるものとはほぼ同じである』との Wright³¹⁾ (1996, p. 6) の指摘が上記に關係する記述である。但し、以降に示される様に、Outfit・Infit χ^2 平均平方統計量と点双列相関係数との相関は常に高いとは限らず、Rasch モデル上での弁別力と 0・1 データ上での弁別力が同等物であるとは見なされ難い。

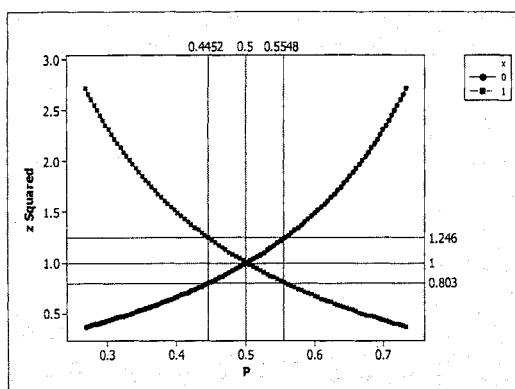


図2 標準化残差平方 対 モデル期待正答確率

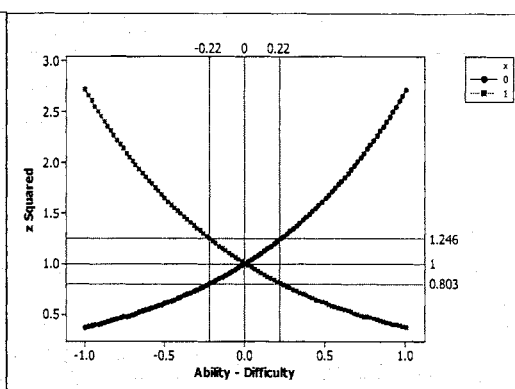


図3 標準化残差平方 対 能力と困難度の差

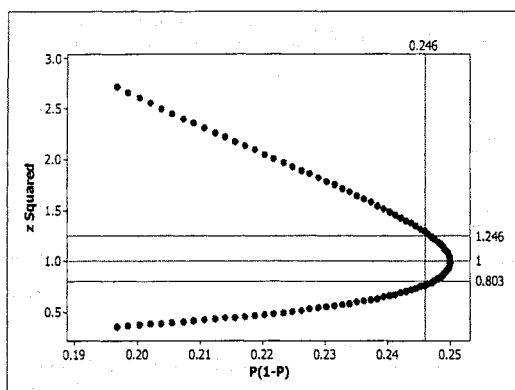


図4 標準化残差平方 対 モデル期待分散

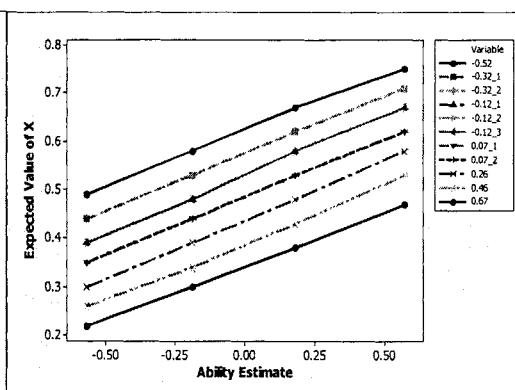


図5 下記人工データ項目群特性曲線

上述の如く、標準化残差平方和 $\sum z_{ij}^2$ の平均が各項目と各受験者に付される Outfit χ^2 平均平方統計量である。その値 1.3 以上を外れ値として、四捨五入上での上限臨界値 1.246 を基準とする標準化残差平方に対するモデル期待正答確率、受験者能力と項目困難度の差、及び、モデル期待分散それぞれ

の分布を示すものが図 2、3、4 である。その下限臨界値が自動的に 0.803 になることが示されて、その上限・下限臨界値内に入るモデル期待正答確率範囲、受験者能力と項目困難度の差の範囲、及び、モデル期待分散値範囲がそれぞれの図に与えられている。図 2 への目視により、Rasch 測定上で、各項目に対する各受験者の期待正答確率がすべて 0.4452 から 0.5548 までの範囲内にあれば、各項目と各受験者に付される Outfit χ^2 平均平方統計量はすべて 0.8 以上 1.2 以下の範囲内にある値になることが分かる。

表 3 受験者能力推定値・項目困難度推定値分布幅狭小人工データ Facets 出力と素点二種類統計量

Facets出力 受験者能力 推定値	Facets 出力 項目 困難度 推定値											素点 得点	Outfit MnSq	Outfit ZStd	Infit MnSq	Infit ZStd	点双列 相関 係数
	-0.52	-0.32	-0.32	-0.12	-0.12	-0.12	0.07	0.07	0.26	0.46	0.67						
-0.57	0	1	0	1	1	0	0	0	0	0	1	4	1.1	0	1.1	0	-0.31
-0.57	1	0	0	1	0	0	1	0	1	0	0	4	1.0	0	1.0	0	-0.11
-0.19	0	1	1	0	0	0	1	1	0	0	1	5	1.1	0	1.1	0	-0.35
-0.19	1	0	0	0	0	1	1	1	0	1	0	5	1.0	0	1.0	0	-0.26
-0.19	1	1	1	0	0	0	1	0	1	0	0	5	0.9	0	0.9	0	0.17
-0.19	1	1	1	0	0	1	1	0	0	0	0	5	0.9	0	0.9	0	0.44
-0.19	1	0	0	1	1	0	1	0	1	0	0	5	1.0	0	1.0	0	-0.06
-0.19	0	0	1	1	0	0	0	1	1	1	0	5	1.1	0	1.1	1	-0.44
-0.19	0	1	1	1	1	0	0	0	0	0	1	5	1.0	0	1.0	0	-0.16
-0.19	1	0	0	0	1	0	1	1	0	0	1	5	1.1	0	1.1	0	-0.35
-0.19	1	1	0	0	0	1	0	1	0	1	0	5	1.0	0	1.0	0	-0.06
-0.19	1	0	0	1	1	0	0	1	0	1	0	5	1.0	0	1.0	0	-0.16
0.18	1	0	1	1	0	1	0	1	1	0	0	6	0.9	0	0.9	0	0.13
0.18	1	1	1	0	1	1	0	0	0	1	0	6	0.9	0	0.9	-1	0.25
0.18	0	0	1	1	1	1	1	0	1	0	0	6	1.0	0	1.0	0	-0.10
0.18	0	1	1	1	1	1	0	0	0	0	1	6	1.0	0	1.0	0	-0.10
0.18	1	1	1	1	0	0	0	0	0	1	1	6	1.0	0	1.0	0	-0.20
0.18	1	0	1	0	1	1	0	1	1	0	0	6	0.9	0	0.9	0	0.13
0.18	0	1	0	1	1	1	0	0	0	1	1	6	1.1	1	1.1	1	-0.47
0.57	0	1	0	0	1	1	1	1	1	1	0	7	1.1	0	1.1	0	-0.41
0.57	1	1	1	0	0	1	1	1	1	0	0	7	0.9	0	0.9	0	0.23
素点得点	13	12	12	11	11	11	10	10	9	8	7						
Outfit MnSq	1.1	1.0	0.9	1.1	1.0	0.9	1.1	1.0	1.0	1.0	1.1						
Outfit ZStd	0	0	0	1	0	0	0	0	0	0	0						
Infit MnSq	1.1	1.0	0.9	1.1	1.0	0.9	1.1	1.0	1.0	1.0	1.1						
Infit ZStd	0	0	0	1	0	0	0	0	0	0	0						
点双列 相関係数	-0.57	-0.38	-0.28	-0.65	-0.43	0.01	-0.55	-0.40	-0.35	-0.39	-0.59						

表 3 は、一例として、Rasch 測定分析プログラム Facets(Linacre²⁰⁾, 1989-2001, Winsteps.com) の出力上で、項目群困難度推定値と受験者群能力推定値の分布幅が狭小となる意図で以って、0・1 無作為抽出を基にして作成した、11 項目と 21 名から成る人工データの Facets 出力に、素点得点と点双列相関係数を併せて示したものである。0・1 生起目視、及び、各項目と各受験者に付された点双列相関係数により、データの Rasch 測定妥当性欠如が一目瞭然である。然しながら、Outfit・Infit 統計量は大変に高い Rasch モデル適合度を示唆している。項目群困難度推定値と受験者群能力推定値の分布幅が狭小となれば、各項目と各受験者に付される Outfit・Infit 統計量のそれぞれがその期待値に近似する傾向が顕著に観察される。ちなみに、11×21=231 個の Rasch 測定標準化残差平方値、モデル期待正答確率、及び、モデル期待分散それぞれにおける最小値と最大値の組み合わせは、0.2916 と 3.4596、0.22 と 0.75、及び、0.17 と 0.25 である。又、同じく 231 個の標準化残差における最小値と最大値はそれぞれ -1.72 と 1.86 であり、その分布の正規性仮定の下では、231 個すべてが二標準偏差内の範囲に位置している。更に、図 5 が、『Rasch モデル母数はデータが常にモデルへの[完全な]適合

を果しているとの仮定に基づいて推定されている』との Linacre³⁴⁾ (1999, p. 676)の言葉を明示する当該人工データ項目群の特性曲線である。

表4が前章での考察に關係する Rasch モデル母数推定値分布幅とデータ Rasch 測定妥当性欠如の連関性一端を顕示するものである。各データ作成法は前掲表2についてと同じく、静¹⁰⁾ (2007, pp. 188-191)によって記述されたものである。「事前設定値」の「最小値」と「最大値」が31項目と301名について事前に定められた Rasch 測定事前設定値幅の両端値であり、各データ上での項目群と受験者群それぞれにおける設定値は等間隔とされている。なお、Excel 上での RAND 関数適用によるその0・1生起はすべて無作為であるが、得点0点と満点の受験者生起の場合には再試行によりその生起を防ぐものとしている。その作成法に基づいて、狭小から順次に拡大された事前設定値幅六種類各二つ、計12のデータそれぞれへの Facets 適用出力上での、各データにおける項目群と受験者群それぞれにおける母数推定値の「最小値」と「最大値」に併せてその「範囲」が与えられている。

Rasch 測定理論の真髄は「不変性」(invariance)である (Andrich¹¹⁾, 1988, pp. 21-22, and 2005³⁵⁾, p. 53)。Rasch 理論の数理基盤により母数分離定理が導かれて、モデル適合度の高いデータにより算出される Rasch 測定母数推定値は、局所独立性充足の下で誤差変動内での間隔尺度値不変性を属性とする (Smith and Suh²⁴⁾, 2003, p. 153、参照)。従って、モデル適合度の高いデータにおける項目群と受験者群それぞれに付される母数推定値が、それぞれの母集団において高い程度に保持されていると想定される。その母数推定値不変性の成立程度の確認は、一般包括的には、項目群に関しては受験者群能力高位群・低位群 (Smith and Suh²⁴⁾, 2003)、並びに、受験者群に関しては項目群困難度上位群・下位群 (Gustafsson³⁶⁾, 1980)それぞれにおける折半データ上での Rasch 測定母数推定値の同等性検証によってなされる。然しながら、井澤 (2005a³⁷⁾, p. 9-10、2006³⁸⁾, p. 7)により、上記のデータ折半法以上に、元データにおける受験者群と項目群それぞれについての Rasch 測定標準化残差の主成分分析における負荷量上での正負記号分別に基づく受験者群折半と項目群折半 (Linacre 1998¹⁷⁾; Wright 1996b³⁹⁾; Wright and Stone³³⁾, 2004、参照)が、母数推定値同等性下限信頼性が高いと示されている。これにより、表4での項目群と受験者群、それぞれについての母数推定値不変性成立程度を観察する簡易法としての各データ折半法を、受験者群と項目群それぞれについての Rasch 測定標準化残差主成分分析の負荷量記号正負分別に基づくものとしている。なお、項目群と受験者群それぞれについての推定値不変性成立程度が各データ全体としての Rasch 測定妥当性の高さを示唆する。

表4については、母数推定値分布幅の拡大につれて項目群と受験者群のいずれについても、注釈を必要としない程、推定値不変性成立程度が見事に正の相関係数値の上で高くなっている。それは、AとBの一对上で六種類のデータ群がその順序で Rasch 測定妥当性が高いとの示唆である。表4は、当然ながら、事象全体の生起についての平均を示すものではないが、前章での考察に關係する Rasch 測定母数推定値分布幅とデータ Rasch 測定妥当性の連関性一端を明示していることに疑念はない。

表4 Rasch 測定母数推定値分布幅とデータ Rasch 測定妥当性の連関性一端

データ 1A (31項目 x 301名)	項目困難度												
	事前 設定値	最小値 最大値	-0.050 0.052	データ 推定値	最小値 最大値	-0.32 0.22	データ 推定値範囲	0.54	推定値 不変性	有効数	31	ピアソン kendall-tau b	-0.841 -0.613
	受験者能力												
	事前 設定値	最小値 最大値	-0.050 0.052	データ 推定値	最小値 最大値	-1.08 0.90	データ 推定値範囲	1.96	推定値 不変性	有効数	301	ピアソン kendall-tau b	-0.209 -0.171
データ 1B (31項目 x 301名)	項目困難度												
	事前 設定値	最小値 最大値	-0.050 0.052	データ 推定値	最小値 最大値	-0.26 0.29	データ 推定値範囲	0.55	推定値 不変性	有効数	31	ピアソン kendall-tau b	-0.833 -0.567
	受験者能力												
	事前 設定値	最小値 最大値	-0.050 0.052	データ 推定値	最小値 最大値	-1.24 1.24	データ 推定値範囲	2.48	推定値 不変性	有効数	301	ピアソン kendall-tau b	-0.114 -0.118
データ 2A (31項目 x 301名)	項目困難度												
	事前 設定値	最小値 最大値	-1.50 1.50	データ 推定値	最小値 最大値	-1.65 1.51	データ 推定値範囲	3.16	推定値 不変性	有効数	31	ピアソン kendall-tau b	0.621 0.446
	受験者能力												
	事前 設定値	最小値 最大値	-1.50 1.50	データ 推定値	最小値 最大値	-2.57 3.03	データ 推定値範囲	5.60	推定値 不変性	有効数	296	ピアソン kendall-tau b	0.573 0.444
データ 2B (31項目 x 301名)	項目困難度												
	事前 設定値	最小値 最大値	-1.50 1.50	データ 推定値	最小値 最大値	-1.70 1.50	データ 推定値範囲	3.20	推定値 不変性	有効数	31	ピアソン kendall-tau b	0.640 0.504
	受験者能力												
	事前 設定値	最小値 最大値	-1.50 1.50	データ 推定値	最小値 最大値	-2.58 2.22	データ 推定値範囲	4.80	推定値 不変性	有効数	296	ピアソン kendall-tau b	0.612 0.491
データ 3A (31項目 x 301名)	項目困難度												
	事前 設定値	最小値 最大値	-2.50 2.60	データ 推定値	最小値 最大値	-2.50 2.56	データ 推定値範囲	5.06	推定値 不変性	有効数	31	ピアソン kendall-tau b	0.818 0.627
	受験者能力												
	事前 設定値	最小値 最大値	-2.50 2.60	データ 推定値	最小値 最大値	-4.35 3.58	データ 推定値範囲	7.93	推定値 不変性	有効数	292	ピアソン kendall-tau b	0.767 0.625
データ 3B (31項目 x 301名)	項目困難度												
	事前 設定値	最小値 最大値	-2.50 2.60	データ 推定値	最小値 最大値	-2.62 2.64	データ 推定値範囲	5.26	推定値 不変性	有効数	31	ピアソン kendall-tau b	0.834 0.679
	受験者能力												
	事前 設定値	最小値 最大値	-2.50 2.60	データ 推定値	最小値 最大値	-3.06 4.34	データ 推定値範囲	7.40	推定値 不変性	有効数	294	ピアソン kendall-tau b	0.762 0.625
データ 4A (31項目 x 301名)	項目困難度												
	事前 設定値	最小値 最大値	-3.40 3.50	データ 推定値	最小値 最大値	-3.22 3.58	データ 推定値範囲	6.80	推定値 不変性	有効数	31	ピアソン kendall-tau b	0.879 0.714
	受験者能力												
	事前 設定値	最小値 最大値	-3.40 3.50	データ 推定値	最小値 最大値	-4.95 4.24	データ 推定値範囲	9.19	推定値 不変性	有効数	288	ピアソン kendall-tau b	0.817 0.670
データ 4B (31項目 x 301名)	項目困難度												
	事前 設定値	最小値 最大値	-3.40 3.50	データ 推定値	最小値 最大値	-3.76 3.34	データ 推定値範囲	7.10	推定値 不変性	有効数	31	ピアソン kendall-tau b	0.882 0.721
	受験者能力												
	事前 設定値	最小値 最大値	-3.40 3.50	データ 推定値	最小値 最大値	-4.97 4.96	データ 推定値範囲	9.93	推定値 不変性	有効数	280	ピアソン kendall-tau b	0.813 0.666
データ 5A (31項目 x 301名)	項目困難度												
	事前 設定値	最小値 最大値	-5.00 4.90	データ 推定値	最小値 最大値	-5.43 4.75	データ 推定値範囲	10.18	推定値 不変性	有効数	31	ピアソン kendall-tau b	0.897 0.768
	受験者能力												
	事前 設定値	最小値 最大値	-5.00 4.90	データ 推定値	最小値 最大値	-6.27 6.11	データ 推定値範囲	12.38	推定値 不変性	有効数	274	ピアソン kendall-tau b	0.864 0.727
データ 5B (31項目 x 301名)	項目困難度												
	事前 設定値	最小値 最大値	-5.00 4.90	データ 推定値	最小値 最大値	-5.12 6.22	データ 推定値範囲	11.34	推定値 不変性	有効数	31	ピアソン kendall-tau b	0.989 0.943
	受験者能力												
	事前 設定値	最小値 最大値	-5.00 4.90	データ 推定値	最小値 最大値	-6.33 6.68	データ 推定値範囲	13.01	推定値 不変性	有効数	282	ピアソン kendall-tau b	0.909 0.781
データ 6A (31項目 x 301名)	項目困難度												
	事前 設定値	最小値 最大値	-10.5 10.5	データ 推定値	最小値 最大値	-10.28 11.12	データ 推定値範囲	21.40	推定値 不変性	有効数	30	ピアソン kendall-tau b	0.944 0.884
	受験者能力												
	事前 設定値	最小値 最大値	-10.50 10.50	データ 推定値	最小値 最大値	-10.60 11.03	データ 推定値範囲	21.63	推定値 不変性	有効数	274	ピアソン kendall-tau b	0.938 0.835
データ 6B (31項目 x 301名)	項目困難度												
	事前 設定値	最小値 最大値	-10.5 10.5	データ 推定値	最小値 最大値	-11.17 11.42	データ 推定値範囲	22.59	推定値 不変性	有効数	31	ピアソン kendall-tau b	0.951 0.856
	受験者能力												
	事前 設定値	最小値 最大値	-10.50 10.50	データ 推定値	最小値 最大値	-11.58 11.30	データ 推定値範囲	22.88	推定値 不変性	有効数	252	ピアソン kendall-tau b	0.920 0.811

補足情報として表5を与える。表4でのデータ5Aと5Bとの比較意図で以って、項目群と受験者群について一方のみの「事前設定値」分布幅を狭小にしたデータそれぞれのRasch測定妥当性程度を与えるものである。結果として、データ7の項目群と受験者群についてのRasch測定母数推定値分布幅がそれぞれデータ2A・2Bとデータ3A・3Bに相応するものとなっており、それぞれに準ずるRasch測定妥当性が示されている。データ8については、項目群推定値分布幅はデータ5A・5Bに相応してその項目群不変性成立程度は大変に高いけれども、受験者群推定値分布幅の狭小性がRasch測定妥当性低下と連関している。要するに、表5は表4での受験者群と項目群それぞれ個別単位での推定値分布幅とRasch測定妥当性程度の関係を示唆するものとなっている。

表5 項目群と受験者群について一方の母数推定値分布幅狭小性とデータRasch測定妥当性低下との関係一端

データ 7 (31項目 x 301名)	項目困難度												
	事前 設定値	最小値 最大値	-1.50 1.50	データ 推定値	最小値 最大値	-1.43 1.64	データ 推定値範囲	3.07	推定値 不変性	有効数	31	ピアソン ケンドールタウb	0.507 0.434
	受験者能力												
	事前 設定値	最小値 最大値	-5.00 4.90	データ 推定値	最小値 最大値	-3.76 3.79	データ 推定値範囲	7.55	推定値 不変性	有効数	160	ピアソン ケンドールタウb	0.844 0.696
データ 8 (31項目 x 301名)	項目困難度												
	事前 設定値	最小値 最大値	-5.00 4.90	データ 推定値	最小値 最大値	-6.15 5.65	データ 推定値範囲	11.80	推定値 不変性	有効数	30	ピアソン ケンドールタウb	0.989 0.960
	受験者能力												
	事前 設定値	最小値 最大値	-1.50 1.50	データ 推定値	最小値 最大値	-3.47 2.70	データ 推定値範囲	6.17	推定値 不変性	有効数	301	ピアソン ケンドールタウb	0.606 0.518

表6と表7は、表4における12のデータの項目群と受験者群それぞれに関するモデル適合度指標値出力結果の一覧表である。その意図は、右方向にデータRasch測定妥当性が高くなっている傾向性とモデル適合度指標値変動性の関連一端を窺うことにある。上段四行はSPSS (Version 15.0, SPSS Inc., 2006)出力上での素点に関するCronbachのアルファ係数と点双列相関係数の提示である。その四行以外はすべてFacets出力によるものであるが、下段二行にはOutfitとInfitの χ^2 平均平方統計量それぞれと点双列相関係数のピアソン相関係数を与えている。なお、表6における*印は、データ1Aと1BにおけるOutfit・Infit χ^2 平均平方統計量がすべて同値1.0であるから、相関係数出力不能との意味である。又、表6には下から三行目に「各受験者と各項目の組み合わせに対して標準化残差絶対値2以上の値が付された個数とその割合」を付加している。

各データ31項目に関する表6において、傾向性特徴として顕著に観察されることは次の二点である。第一に、データ3Aから6Bにかけて、AとBの一对単位ではば順次にRasch測定適用妥当性の程度が高くなっているにもかかわらず、Outfit χ^2 平均平方統計量上でのモデル不適合を示唆する1.3以上の値が付された項目数の6Bを除いてのかなり高い割合である。第二点としては、Wright (1996a³²⁾, p. 6)の指摘に反して、Outfit・Infit χ^2 平均平方統計量と素点上での点双列相関係数の指し示すものには大きな相違がある。これは各データ301名に関する表7においても同様である。

表7においては、データ4Aから6BにかけてRasch測定適用妥当性がかなり高いにもかかわらず、OutfitとInfitの χ^2 平均平方統計量いずれにおいても各データがモデル不適合と顯示されている。

表 6 12 データ項目群モデル適合度指標値

モデル適合度 指標値	1A	1B	2A	2B	3A	3B	4A	4B	5A	5B	6A	6B
アルファ係数	-0.139	0.089	0.835	0.838	0.914	0.912	0.935	0.935	0.952	0.952	0.962	0.962
点双列相関係数 最大値	0.112	0.135	0.430	0.495	0.624	0.567	0.691	0.688	0.757	0.785	0.832	0.833
点双列相関係数 平均値	-0.022	0.017	0.341	0.346	0.480	0.476	0.538	0.539	0.600	0.600	0.636	0.635
点双列相関係数 最小値	-0.136	-0.121	0.274	0.253	0.342	0.307	0.320	0.362	0.271	0.352	0.192	0.252
Outfit MnSq 最大値	1.0	1.0	1.2	1.3	1.6	1.8	2.8	1.8	9.0	9.0	1.8	1.9
Outfit MnSq 平均値	1.00	1.00	1.00	0.99	0.97	0.99	0.98	0.91	1.14	1.13	0.48	0.46
Outfit MnSq 最小値	1.0	1.0	0.9	0.8	0.6	0.6	0.5	0.4	0.3	0.2	0.1	0.1
Outfit MnSq ≥ 1.3 の 個数と割合	0 (0%)	0 (0%)	0 (0%)	1 (3.2%)	3 (9.7%)	3 (9.7%)	4 (12.9%)	3 (9.7%)	5 (16.1%)	4 (12.9%)	2 (6.5%)	1 (3.2%)
Outfit MnSq ≤ 0.7 の 個数と割合	0 (0%)	0 (0%)	0 (0%)	0 (0%)	4 (12.9%)	4 (12.9%)	9 (29.0%)	10 (32.3%)	17 (54.8%)	16 (51.6%)	27 (87.1%)	26 (83.9%)
Outfit ZStd 最大値	1	1	2	2	2	1	2	1	2	2	0	0
Outfit ZStd 平均値	0.03	0.00	0.13	-0.03	-0.13	0.00	-0.03	-0.13	0.06	0.06	0.00	0.00
Outfit ZStd 最小値	-2	-1	-1	-2	-2	-1	-1	-1	-1	0	0	0
Outfit ZStd ≥ 2 の 個数と割合	0 (0%)	0 (0%)	1 (3.2%)	1 (3.2%)	1 (3.2%)	0 (0%)	1 (3.2%)	0 (0%)	1 (3.2%)	1 (3.2%)	0 (0%)	0 (0%)
Outfit MnSq ≤ -2 の 個数と割合	1 (3.2%)	0 (0%)	0 (0%)	1 (3.2%)	2 (6.5%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Infit MnSq 最大値	1.0	1.0	1.1	1.1	1.2	1.2	1.1	1.2	1.3	1.2	1.2	1.3
Infit MnSq 平均値	1.00	1.00	0.99	1.01	0.99	1.00	0.99	0.99	0.98	0.96	0.92	0.89
Infit MnSq 最小値	1.0	1.0	0.9	0.9	0.8	0.9	0.8	0.8	0.8	0.7	0.7	0.6
Infit MnSq ≥ 1.3 の 個数と割合	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (3.2%)	0 (0%)	0 (0%)	2 (6.5%)
Infit MnSq ≤ 0.7 の 個数と割合	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (3.2%)	4 (12.9%)	4 (12.9%)
Infit ZStd 最大値	1	1	1	1	2	1	1	2	2	2	1	1
Infit ZStd 平均値	0.03	0.00	0.00	0.06	-0.10	-0.03	-0.10	-0.03	-0.23	-0.23	-0.32	-0.55
Infit ZStd 最小値	-2	-2	-1	-2	-2	-1	-2	-2	-2	-3	-2	-3
Infit ZStd ≥ 2 の 個数と割合	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (3.2%)	0 (0%)	0 (0%)	1 (3.2%)	1 (3.2%)	1 (3.2%)	0 (0%)	0 (0%)
Infit MnSq ≤ -2 の 個数と割合	1 (3.2%)	1 (3.2%)	0 (0%)	1 (3.2%)	2 (6.5%)	0 (0%)	1 (3.2%)	1 (3.2%)	1 (3.2%)	2 (6.5%)	2 (6.5%)	3 (9.7%)
分離信頼性係数	0.02	0.21	0.98	0.98	0.99	0.99	0.99	0.99	1.00	1.00	1.00	1.00
Outfit MnSqと 点双列相関係数との ピアソン相関係数	*	*	-0.656	-0.837	-0.552	-0.438	-0.140	-0.018	-0.156	-0.204	0.135	0.400
Infit MnSqと 点双列相関係数との ピアソン相関係数	*	*	-0.612	-0.730	-0.660	-0.378	-0.266	-0.519	-0.234	-0.290	-0.370	-0.241

表7 12 データ受験者群モデル適合度指標値

モデル適合度 指標値	1A	1B	2A	2B	3A	3B	4A	4B	5A	5B	6A	6B
アルファ係数	0.027	0.223	0.981	0.982	0.991	0.991	0.994	0.994	0.995	0.995	0.996	0.997
点双列相関係数 最大値	0.474	0.519	0.816	0.739	0.866	0.815	0.869	0.854	0.867	0.866	0.867	0.868
点双列相関係数 平均値	0.003	0.016	0.377	0.382	0.513	0.510	0.578	0.575	0.622	0.634	0.674	0.680
点双列相関係数 最小値	-0.481	-0.505	-0.329	-0.060	-0.010	0.079	0.051	0.215	0.106	0.120	0.267	0.232
Outfit MnSq 最大値	1.1	1.1	3.5	2.4	4.8	7.0	9.0	8.9	9.0	9.0	9.0	9.0
Outfit MnSq 平均値	1.00	1.00	1.00	0.99	0.97	0.99	0.94	0.90	0.83	0.84	0.47	0.45
Outfit MnSq 最小値	0.9	0.9	0.5	0.4	0.3	0.3	0.2	0.2	0.1	0.1	0.1	0.0
Outfit MnSq ≥ 1.3 の 個数と割合	0 (0%)	0 (0%)	36 (12.0%)	42 (14.0%)	62 (20.6%)	59 (19.6%)	56 (18.6%)	57 (18.9%)	44 (14.6%)	44 (14.6%)	20 (6.6%)	18 (6.0%)
Outfit MnSq ≤ 0.7 の 個数と割合	0 (0%)	0 (0%)	53 (17.6%)	45 (15.0%)	131 (43.5%)	131 (43.5%)	175 (58.1%)	171 (56.8%)	214 (71.1%)	212 (70.4%)	263 (87.4%)	265 (88.0%)
Outfit ZStd 最大値	1	2	2	2	3	2	2	2	2	1	0	0
Outfit ZStd 平均値	-0.02	-0.05	-0.01	-0.02	-0.01	-0.04	-0.01	0.00	0.00	0.02	0.00	0.00
Outfit ZStd 最小値	-1	-2	-2	-2	-2	-2	-1	-1	-1	0	0	0
Outfit ZStd ≥ 2 の 個数と割合	0 (0%)	1 (0.3%)	6 (2.0%)	1 (0.3%)	8 (2.7%)	5 (1.7%)	4 (1.3%)	2 (0.7%)	1 (0.3%)	0 (0%)	0 (0%)	0 (0%)
Outfit MnSq ≤ -2 の 個数と割合	0 (0%)	4 (1.3%)	4 (1.3%)	2 (0.7%)	4 (1.3%)	1 (0.3%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Infit MnSq 最大値	1.1	1.1	1.4	1.4	1.9	1.8	1.7	1.8	1.9	2.1	2.7	3.4
Infit MnSq 平均値	1.00	1.00	1.00	1.00	0.99	1.00	0.99	0.99	0.97	0.97	0.91	0.90
Infit MnSq 最小値	0.9	0.9	0.6	0.7	0.4	0.5	0.4	0.5	0.4	0.4	0.3	0.3
Infit MnSq ≥ 1.3 の 個数と割合	0 (0%)	0 (0%)	15 (5.0%)	12 (4.0%)	34 (11.3%)	32 (10.6%)	56 (18.6%)	43 (14.3%)	57 (18.9%)	62 (20.6%)	70 (23.3%)	72 (23.9%)
Infit MnSq ≤ 0.7 の 個数と割合	0 (0%)	0 (0%)	13 (4.3%)	9 (3.0%)	40 (13.3%)	38 (12.6%)	63 (20.9%)	48 (15.9%)	77 (25.6%)	88 (29.2%)	126 (41.9%)	143 (47.5%)
Infit ZStd 最大値	1	2	2	2	3	2	2	2	2	2	2	2
Infit ZStd 平均値	-0.02	-0.05	-0.04	0.00	-0.08	-0.07	-0.09	-0.07	-0.09	-0.15	-0.21	-0.27
Infit ZStd 最小値	-1	-2	-3	-2	-3	-2	-2	-2	-2	-2	-1	-1
Infit ZStd ≥ 2 の 個数と割合	0 (0%)	1 (0.3%)	3 (1.0%)	2 (0.7%)	6 (2.0%)	4 (1.3%)	2 (0.7%)	2 (0.7%)	2 (0.7%)	3 (1.0%)	2 (0.7%)	4 (1.3%)
Infit MnSq ≤ -2 の 個数と割合	0 (0%)	4 (1.3%)	6 (2.0%)	2 (0.7%)	9 (3.0%)	10 (3.3%)	8 (2.7%)	6 (2.0%)	8 (2.7%)	17 (5.6%)	0 (0%)	0 (0%)
分離信頼性係数	0.00	0.12	0.83	0.83	0.91	0.90	0.93	0.93	0.96	0.96	0.98	0.98
各受験者と各項目の 組み合わせに対して 標準化残差絶対値 2以上の値が 付された個数と その割合	0 0.0%	2 0.0%	389 4.2%	383 4.1%	438 4.7%	419 4.5%	385 4.1%	370 4.0%	276 3.0%	280 3.0%	150 1.6%	143 1.5%
Outfit MnSqと 点双列相関係数との ピアソン相関係数	-0.356	-0.516	-0.866	-0.854	-0.644	-0.576	-0.420	-0.296	-0.184	-0.120	-0.092	-0.036
Infit MnSqと 点双列相関係数との ピアソン相関係数	-0.331	-0.449	-0.936	-0.930	-0.797	-0.791	-0.646	-0.569	-0.383	-0.390	-0.204	-0.277

一方、その χ^2 平均平方統計量が標準化された Z_{std} 統計量基準値 2 未満の点からは 4A から 6B に渡る 6 データはすべて大変に高いモデル適合と表示されている。又、各データにおける「各受験者と各項目の組み合わせに対して標準化残差絶対値 2 以上の値が付された個数とその割合」を参照すれば、すべてが、前記参照 Linacre²⁰⁾ (1989-2001, p. 52)によればモデル適合を示唆する 5%未満の値である。

以上により、当該データ群に関しても井澤(2005a³⁷⁾、2005b⁴⁰⁾、2006³⁸⁾)においてと同様に Outfit・Infit 統計量による項目群・受験者群モデル不適合率と両群母数推定値不変性成立程度の間には大きな乖離が窺われる。Karabatsos¹⁴⁾ (2000)による Outfit・Infit モデル適合度統計量限界性、並びに、Smith, Schumacker, and Bush¹⁵⁾ (1998)及び Smith and Suh²⁴⁾ (2003)により標本数に依存するその統計量変動性が検証されており、テスト構成概念妥当性判断をその統計量のみで以て下すことは短絡である。受験者群と項目群それぞれについての Rasch 測定標準化残差主成分分析における負荷量記号の正負分別に基づく受験者群折半と項目群折半(Linacre 1998¹⁷⁾; Wright 1996b³⁹⁾; Wright and Stone³³⁾, 2004, 参照)による項目群と受験者群それぞれにおける母数推定値不変性成立程度の確認が、テスト構成概念妥当性判断の上では Outfit・Infit モデル適合度統計量に優る。項目群と受験者群それぞれに関するその折半法の上での下位データ間に算出される二分割標本 t 値(Smith⁴¹⁾, 1996; Smith and Suh²⁴⁾, 2003, p. 155; Wright and Stone⁴²⁾, 1979, pp. 94-95)ないしは χ^2 検定(McNamara⁴³⁾, 1991, pp. 153-154)による母数推定値同等性の統計的検証が、母数推定値不変性確認補完性の上で更に望まれる。なお、Rasch 測定標準化残差の線形性欠如の点から、その残差に優る線形性保持代替残差への主成分分析適用が Karabatsos¹⁴⁾ (2000)によって提案されている(pp. 160-161)。然しながら、データ折半法の上では、いずれの残差の主成分分析負荷量正負記号分別に基づく受験者群折半もほぼ同等な項目群一次元性充足度が一例上で示されている(井澤⁴⁴⁾, 2007, p. 12)ことを付記しておく。

IV. おわりに

本稿の Rasch 測定適用者への示唆は大変に重要である。同時最尤推定法上での母数推定値が項目群困難度と受験者群能力のいずれも、例えば-4.0 以下から 4.0 以上に渡る、広い幅で分布していなければ、Rasch 測定母数推定は妥当性に欠ける。特に、単一試験において受験者群の能力高低分別あるいは個人能力差峻別を重要視する場合にはなおさらである。Rasch 測定母数推定値幅が狭小である場合には、(1) データ一次元性充足度の低い可能性が高い、あるいは、(2) 一次元性には問題が無いが、項目群困難度と受験者群能力が事実として密集している可能性がある。一次元性充足度の低いデータが Rasch 測定に適合しないことは当然である。一方、項目群困難度と受験者群能力が密集している場合には、Rasch モデル適用の大前提である一次元性規定には本来的に適っているとしても、推定値を得る時点でその信頼性に問題が発生する。それは母数推定を行うこと自体の妥当性の低さを意味する。この場合には、データ次元性についての議論すら成立しないことも十分考えられる。然しながら、いずれの場合においても、Rasch 測定上でのモデル適合度を示す Outfit・Infit 指標に反映されず、低

いとは言えないモデル適合と判断される可能性も多分に存在する。つまり、Rasch 測定上での母数推定値分布幅に関心を寄せ、また、各推定値の標準誤差をも注視して、母数推定自体の妥当性についての判断が最優先されなければならないということである。上記に関連して、Rasch 測定分析プログラム上で一般的に出力される項目群・受験者群分離信頼性指標への参照も必須である。

上で述べたことは次のようにも説明される。教育機関が実施する大規模な実力検定試験や入学試験での本質的に能力選別厳格性が意図される試験以外の多くの場では、Rasch 測定は必要とされない可能性がある。一教師の授業期末等におこなわれる達成目標基準準拠テストの実施後、たとえそのテストが多く項目数で以って多くの受験者に与えられたとしても、その Rasch 測定結果は母数推定値の間隔尺度性を多分に保証しない。その理由は、一回限りの達成目標基準準拠テストにおいては、授業内容と関係するその実施目的の上で本質的に受験者全員が高得点となるように意図されているからである。従って、そのテストの能力選別を目的としない性質から素点得点分布に正規分布が期待されず、その Rasch 測定においても狭小な分布幅を持つ母数推定値が観察されることになるであろう。それは、その単一テスト結果の上で Rasch 測定適用妥当性あるいは能力推定値不変性が保証されていない可能性が高く、その推定値自体に素点上での順序尺度とは異なる意味は見出されないということである。如何なるテストにおいても能力測定自体の妥当性とデータ次元性を確認する上では Rasch 測定に有意義性が認められる。然しながら、一回限りの達成目標基準準拠テストにおいては Rasch 測定上での母数推定値分布幅狭小性がほとんどの場合に自明であるため、Rasch 測定適用による母数推定自体にその妥当性が欠けている。勿論、これは複数回の同質的なテスト試行において受験者能力伸長測定上で妥当性と信頼性の高い項目群を順次に収集する意図の下での Rasch 測定を否定するものではない。

最後に、本稿第Ⅱ章では同時最尤推定法上での母数推定値分布特徴の発生経緯分析を行ったが、その特徴を利用したデータ次元性確認の方法を検討することが今後の課題である。又、Rasch 測定母数推定値分布幅の狭小性がデータ多次元性によって生起するという本稿考察に加えて、一次元性充足度の高い本来的には Rasch モデル適合となるべきデータにおいて、その推定値分布幅狭小性と Rasch 測定適用妥当性低下の共起現象が前掲表 4 に窺われる。その事象への精察必要性も意識される。

参考文献

- 1) G. Rasch. 1960. *Probabilistic models for some intelligence and attainment tests*. The Danish Institute for Educational Research. (Reprinted in 1980 by the University of Chicago Press with a Foreword and Afterword by Wright, B. D.)
- 2) 豊田秀樹 2002. 『項目反応理論[入門編]』 東京： 朝倉書店.
- 3) W. J. V. D. Linden & R. K. Hambleton (Eds.). 1997. *Handbook of modern item response theory*. New York: Springer.
- 4) S. M. Ross. 2004. *Introduction to probability and statistics for engineers and scientists, 3rd Ed.* London: Elsevier Academic Press.

- 5) 豊田秀樹 (編著) 2005. 『項目反応理論[理論編]』 東京: 朝倉書店.
- 6) J. Berkson. 1944. Application of the logistic function to bio-assay. *Journal of the American Statistical Association*, 39, pp. 357- 365.
- 7) J. S. Crammer. August 2003. The origins and development of the logit model.
<http://www.cambridge.org/resources/0521815886/1208 default.pdf>.
- 8) T. O'Neil. February 2005. Definition of a logit. *NCLEX Psychometric Technical Brief*, 2
(http://www.ncsbn.org/pdfs/02_18_05_brief.pdf).
- 9) C. B. Peirce. 1878. Illustrations of the logic of science. *Popular Science Monthly*, pp. 705- 718.
- 10) 静 哲人 2007. 『基礎から深く理解するラッシュモデリング』 大阪: 関西大学出版部
- 11) D. Andrich. 1988. A general form of Rasch's extended logistic model for partial credit scoring. *Applied Measurement in Education*, 1, 4, pp. 363- 378.
- 12) J. M. Linacre. 1989. *Many-facet Rasch measurement*. Chicago: MESA Press.
- 13) B. D. Wright & G. N. Masters. 1982. *Rating scale analysis*. Chicago: MESA Press.
- 14) G. Karabatsos. 2000. A critique of Rasch residual fit statistics. *Journal of Applied Measurement*, 1, 2, pp. 152- 176.
- 15) R. M. Smith, R. E. Schumacker, & M. J. Bush. 1998. Using item mean squares to evaluate fit to the Rasch model. *Journal of Outcome Measurement*, 2, 1, pp. 66- 78.
- 16) 井澤廣行 2007. 「Karabatsos の Rasch モデル遊離度線形残差と Rasch 測定標準化残差の異同程度」『流通科学大学論集 — 人間・社会・自然編』第 19 巻、第 3 号、pp. 1- 14.
- 17) J. M. Linacre. 1998. Detecting multidimensionality: Which residual data-type works best? *Journal of Outcome Measurement*, 2, 3, pp. 266- 283.
- 18) R. M. Smith & L. V. Hedges. 1982. A comparison of likelihood ratio χ^2 and Pearsonian χ^2 tests of fit in the Rasch model. *Educational Research and Perspectives*, 9, pp. 44- 54.
- 19) R. M. Smith. 1988. The distributional properties of Rasch standardized residuals. *Educational and Psychological Measurement*, 48, pp. 657- 667.
- 20) J. M. Linacre. 1989-2001. *A user's guide to FACETS: Rasch measurement computer program*. Chicago: Winsteps.com.
- 21) D. R. Divgi. 1986. Does the Rasch model really work for multiple choice items? Not if you look closely. *Journal of Educational Measurement*, 23, 4, pp. 283- 298.
- 22) B. D. Wright & J. M. Linacre. 1994. Reasonable mean-square fit values. *Rasch Measurement Transactions*, 8, 3 in Linacre, J. M. (Ed.), 1996, *Rasch Measurement Transactions, Part 2* (p. 370). Chicago: MESA Press.
- 23) T. McNamara. 1996. *Measuring second language performance*. London: Longman.
- 24) R. M. Smith & K. K. Suh. 2003. Rasch fit statistics as a test of the invariance of item parameter estimates. *Journal of Applied Measurement*, 4, 2, pp. 153- 163.
- 25) M. Wilson. 2005. *Constructing measures: An item response modeling approach*. Mahwah, N.J.:

- Lawrence Erlbaum Associates, Inc.
- 26) L. Shen. 1996. Quantifying item dependency. *Rasch Measurement Transactions*, 10, 1, p. 485.
 - 27) J. M. Linacre. 1997. Investigating judge local independence. *Rasch Measurement Transactions*, 11, 1, pp. 546- 547.
 - 28) J. M. Linacre. 2000. Redundant items, overfit and measure bias. *Rasch Measurement Transactions*, 14, 3, p. 755.
 - 29) J. M. Linacre & B. D. Wright. 1994. Chi-square fit statistics. *Rasch Measurement Transactions*, 8, 3 in Linacre, J. M. (Ed.), 1996, *Rasch Measurement Transactions, Part 2* (pp. 360- 361). Chicago: MESA Press.
 - 30) G. Henning. 1987. *A guide to language testing*. Boston: Newbury House Publishers.
 - 31) G. Karabatsos, G. 2004. The Rasch model, additive conjoint measurement, and new models of probabilistic measurement theory. In Smith, Jr., E. V. & Smith, R. M. (Eds.), *Introduction to Rasch Measurement*, (pp. 630- 664). Maple Grove, MN: JAM Press.
 - 32) B. D. Wright. 1996a. Comparing Rasch measurement and factor analysis. *Structural Equation Modeling*, 3, 1, pp. 3- 24.
 - 33) B. D. Wright & M. H. Stone. 2004. *Making measures*. Chicago: The Phaneron Press, Inc.
 - 34) J. M. Linacre. 1999. Category disordering vs. step disordering. *Rasch Measurement Transactions*, 13, 1, pp. 675- 678.
 - 35) D. Andrich. 2005. The Rasch model explained. In Alagumalai, S, Curtis, D. D., & Hungi, N (Eds.), *Applied Rasch measurement: A book of exemplars* (pp. 27- 59). The Netherlands: Springer.
 - 36) J. E. Gustafsson. 1980. Testing and obtaining fit of data to the Rasch model. *British Journal of Mathematical and Statistical Psychology*, 33, pp. 205- 233.
 - 37) 井澤廣行 2005a. 「Rasch 測定標準化残差主成分分析に基づく母数推定値不変性成立程度確認の重要性と有効性」『流通科学大学論集 — 人間・社会・自然編』第 18 巻、第 1 号、pp. 1- 14.
 - 38) 井澤廣行 2006. 「Rasch モデル母数推定値同源性検証のためのカイ二乗検定有用性」『流通科学大学論集 — 人間・社会・自然編』第 18 巻、第 3 号、pp. 1- 14.
 - 39) B. D. Wright. 1996b. Local dependency, correlations and principal components. *Rasch Measurement Transactions*, 10, 3, pp. 509- 511.
 - 40) 井澤廣行 2005b. 「Rasch モデル母数一元性規定に離反する項目・受験者弁別力異同への Infit・Outfit 統計量検知鋭敏性」『流通科学大学論集 — 人間・社会・自然編』第 18 巻、第 2 号、pp. 1- 14.
 - 41) R. M. Smith. 1996. A comparison of the Rasch separate calibration and between-fit methods of detecting item bias. *Educational and Psychological Measurement*, 56, 1, pp. 403- 418.
 - 42) B. D. Wright & M. H. Stone. 1979. *Best test design*. Chicago: MESA Press.
 - 43) T. McNamara. 1991. Test dimensionality: IRT analysis of an ESP listening test. *Language Testing*, 8, 2, pp. 139- 159.
 - 44) 井澤廣行 2007. 「Karabatsos の Rasch モデル遊離度線形残差と Rasch 測定標準化残差の異同程度」『流通科学大学論集 — 人間・社会・自然編』第 19 巻、第 3 号、pp. 1- 14.