【原著】

AO 入試における評定者の寛大化傾向の測定・評価

――古典的項目反応理論によるアプローチ――

大澤公一（山口大学）

小論文や面接試験などのパフォーマンスアセスメントにおける評定者の行動特性をテスト項目の困難度から分離し，定量的に評価するための統計学的手法として項目反応理論を応用する。山口大学の AO 入試データを整形した擬似入試データに対して線形ロジスティックテストモデル（Linear Logistic Test Model, LLTM）を適用し，評定者の寛大化傾向の定量的な検出方法を例示する。

1 問題背景および目的

小論文や面接試験，ポートフォリオ評価など，一般にパフォーマンスアセスメントと呼ばれる試験では自由回答形式のテスト項目が用いられる。こうした試験では正答あるいは誤答という二値型（例：正答＝1，誤答＝0）の客観的採点ではなく，多段階（例：1 点～5 点）の主観的採点が行われることが一般的である。

能力評価における評定は，原則的に人間の評定者が行うことになる。ここで，評定者はそれぞれに異なる評定の傾向をもつものと考えられる。ある評定者は一貫して厳しい評定を行うかもしれないし，別の評定者は逆に寛容な評定を続けるかもしれない。このような，評定者個人について固有の行動特性が存在することを前提とする。

評定者効果については，従来より「評定者バイアス」という表現で産業組織心理学を中心に研究されてきた。評定者バイアスとは主に人事考課において評定者が無意識に犯しやすい評定の誤りや偏りを指しており，具体的には以下のようなバイアスの存在が知られている（友安, 1978）。

(1) 寛大化傾向（Leniency）

評定者が，受験者の真の能力水準よりも全体的に寛大な，より高い評価を下してしまうバイアスである。

(2) 中心化傾向（Central Tendency）

評定尺度上の中間値（例：5 段階の評定尺度上の 3 点）に評定が集まるバイアスである。受験者集団に真の評定値（能力値）の個人差が存在するにも関わらず中心化傾向が生じるのは，評定者が自身の評定能力に自信がないか，採点において平均点規制が行われており，優秀な者に高い得点を与えると他の者の評価をその分低くしなければならず，結果的に平均点周りに評定値が集まるといった理由が考えられる。

中心化傾向とは逆に中間値が少なく高評価と低評価に評定が集中する現象を 2 極化傾向という。2 極化傾向が発生する原因としては，評定者の要求水準が過大に高く，基準を達成できない者を全てふるい落としたり，あるいは自分の気に入った者だけを極端に高く評価し，そうでない者を退けてしまうといった理由が考えられる。

(3) ハロー効果（Halo Effect）

受験者の能力のある一面に対する評価を行わなければならないのに，受験生の全体的な印象が評定値に強く影響するバイアスである。例えば，面接試験などで最初の質問項目に対する回答から得られた印象がその後の質問項目の評定に強い影響を与えるようであれば，質問項目の内容や難易度などの特性に関係なく，評定値の間に強い相

関関係が発生し，受験者の能力を正常に測定できなくなる。

(4) 論理的誤差（Logical Error）

テスト項目の間に論理的な関係あるいは類似性がある場合に，双方の項目に対して類似した評定値を与えるバイアスである。論理的誤差は受験者の能力や特性を反映した反応（例：面接試験中の受験者の実際の回答内容）とは無関係であり，評定者の頭の中で論理的に考えて評定した結果生じるバイアスである。論理的誤差が生じた評定値の間には強い相関関係が発生する。論理的誤差はハロー効果と類似した評定バイアスであるが，ハロー効果がある同一の人物に対して発生するバイアスであるのに対し，論理的誤差は人物の区別なしに，特定の評定項目間に見られるバイアスである。

(5) 対比誤差（Contrast Error）

あるテスト項目について，評定者自身が持っている特性傾向の反対方向に受験者を評価するバイアスである。例えば，評定者が自分自身のことを論理的な人間であると思っていると，受験者のことを論理的でないと判断する傾向があるといった具合である。対比誤差は，相手の能力や特性を見ずに自分自身との比較によって評定を行ってしまうバイアスである。

テスト理論においては，パフォーマンスアセスメントにおける評定者の行動特性を記述・評価するための測定モデルとして，古典的テスト理論（Lord andNovick, 1968; 池田，1973）および一般化可能性理論（Brennan, 2001）の枠組の中で様々な研究が行われてきた。近年では，両理論に続く現代テスト理論とも称される項目反応理論（Item Response Theory, IRT; Lord, 1980; 芝，1991）の枠組での研究が主流となってきている。

項目反応理論の枠組みにおける評定者効果とは，一般的に**(1)**の寛大化傾向，すなわち各評定者の平均的な評定の甘さや厳しさを指す。評定者効果（寛大化傾向）の検出方法には大きく 2 つのアプローチ方法が存在する。1 つは線形ロジスティックテストモデル（Linear Logistic Test Model, LLTM; Fischer, 1983）であり [1)]，いま 1 つは階層評定者モデル（Hierarchical Rater Model, HRM; Patz, Junker, Johnson& Mariano, 2002）である。

本研究では，平成 23 年度山口大学 AO 入試データに対して線形ロジスティックテストモデル（LLTM）を適用する。受験者の能力尺度上における評定者の寛大化傾向を推定し，項目反応理論に基づく評定者効果の計量的な測定・評価方法を例示することが本研究の目的である。

2 項目反応モデル

2.1 2 母数ロジスティックモデル

IRT において，受験者iは学力や能力として定義される一次元の潜在特性θ_iの値の大小によって個人差を定義される。2 母数ロジスティックモデル（2 Parameter Logistic Model, 2PL; Lord, 1980）は，人為的な配点を設けずに正答・誤答の 2 値型で採点される（i.e. 正答＝1，誤答＝0 と採点される）テスト項目jに対する，受験者iの正答確率$P_j(\theta_i)$を(1)式のように定義する潜在変数モデルである。

$$(1)\quad P_j(\theta_i) = \frac{1}{1+\exp\{-D\alpha_j(\theta_i-\beta_j)\}}$$

2PL における正答確率$P_j(\theta_i)$は，受験者i（i=1, 2, … ,N ;N は総受験者数）の潜在特性値 θ_i，テスト項目j（j=1, 2, … ,n ;n は総項目数）の項目識別力 α_jおよび項目困難度 β_jという 3 つのパラメタによって定義される。Dは尺度定数であり，IRT のオリジナルである正規累積モデルに近似するために

は$D \cong 1.702$とする。項目困難度β_jは正答確率$P_j(\theta_i)$が 50%となるときの潜在特性値θとして定義されている。言い換えれば，受験者iの能力θ_iとテスト項目jの困難度β_jが同じ値であるとき，その項目の正答確率は 50%となる。あるいは，困難度β_jのテスト項目に 50%の正答確率を得るために必要な能力θのレベルがβ_jであるとも言える。

項目識別力α_jは潜在特性値θの僅かな差を正答確率$P_j(\theta_i)$の差異としてどの程度敏感に反映するのかを定義する項目パラメタである。識別力α_jが大きい項目は困難度β_j付近での潜在特性値θの高低による正答確率$P_j(\theta_i)$の変動が大きい。逆に，識別力の低い項目では潜在特性値の高低によって正答確率が変化しないため，個人の能力の識別が十分にできず，その項目をテストに含める意味がなくなる。

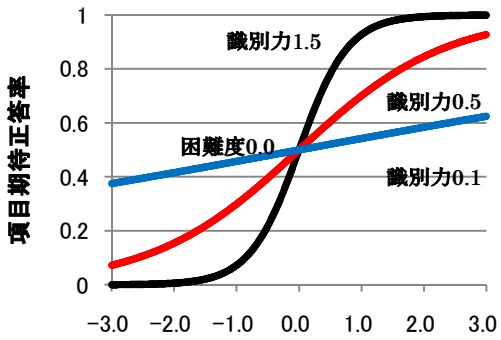

図1. 2PLの項目特性曲線

2PL の項目特性曲線（Item Characteristic Curve, ICC）の例を図 1 に示した。ここでは 3 つの仮想的なテスト項目を例とし，項目困難度β_jは 0.0 で共通とした。すなわち，受験者の潜在特性値θが 0.0 であれば，これらのテスト項目に対する期待正答率が 0.5（50%）となっている。3 項目の識別力α_jは 1.5，0.5，0.1 とした。識別力が 1.5 の項目は，潜在特性値（項目困難度）0.0 付近での曲線の勾配が最も大きく，微小な能力の高低による項目正答率の変動が大きいことをよく表している。識別力が 0.5 の項目は，識別力 1.5 の項目と比較して潜在特性値 0.0 付近での勾配がやや平坦となり，能力水準 0.0 を境に低能力水準においては期待正答率が高く，高能力水準においては正答率が低くなっている様子が観察できる。一方で，識別力が 0.1 の項目では潜在特性値 0.0 付近での曲線の勾配がかなり平坦であり，能力の高低によって項目の期待正答率がほとんど変化していない，すなわち個人の能力を識別していない良くないテスト項目であることが読み取れる。

2.2 一般化部分得点モデル

Muraki（1992）はk個のカテゴリ（$k = 1,2,\dots,v,\dots,K_j$；小問に含まれる部分得点に相当する段階）を通過することで最終的な正答（完全正答）に至る構造のテスト項目jについて，能力水準がθ_iである受験者iが，テスト項目jのカテゴリkに反応する確率$P_{jk}(\theta_i)$を一般化部分得点モデル（Generalized Partial Credit Model, GPCM）として(2)式のように定義した。

$$(2)\quad P_{jk}(\theta_i) = \frac{1+\exp\left\{\sum_{v=1}^{k} D\alpha_j(\theta_i-\beta_j+\gamma_v)\right\}}{1+\sum_{c=1}^{K_j}\exp\left\{\sum_{v=1}^{c} D\alpha_j(\theta_i-\beta_j+\gamma_v)\right\}}$$

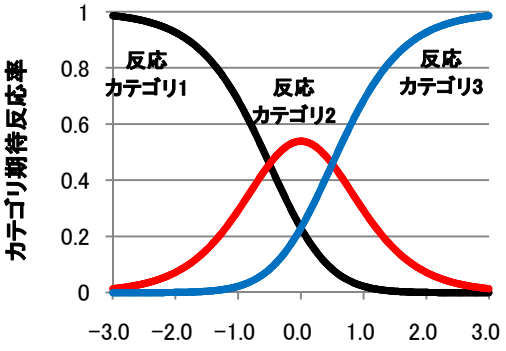

図2. GPCMの項目カテゴリ特性曲線

GPCM の項目カテゴリ特性曲線（Item Category Characteristic Curve, ICCC）の例を図 2 に示した。GPCM では能力レベル

θ_iと項目カテゴリパラメタ$\beta_{jk}=\beta_j-\gamma_k$が同値であれば，受験者$i$はテスト項目$j$の評定段階$k-1$と$k$とに等確率で反応する。図2の項目例では，$\beta_{12}=-0.5$であるので能力水準が－0.5の受験者は，この項目の反応カテゴリ1と反応カテゴリ2に等確率で反応することが期待される。θ_iがβ_{jk}を上回れば受験者iは段階$k-1$より段階kを選択する確率が大きくなり，逆にθ_iがβ_{jk}を下回れば，受験者iは段階kより段階$k-1$を選択する確率が大きくなる。図2の項目例では，$\beta_{23}=0.5$であるので，能力水準が0.5を超える受験者はカテゴリ2と比較してカテゴリ3に反応する確率が高くなることが期待される。

本研究では，項目識別力α_jは1.0に固定する。これを次節のLLTMへ拡張し，項目困難度から評定者効果（寛大化傾向）を分離したモデルを構築する。

2.3 線形ロジスティックテストモデル

評定者r（$r=1,2,\ldots,R$；Rは総評定者数）が，能力レベルθ_iである受験者iのテスト項目jに対する反応（回答，答案）にカテゴリkと評定・採点する確率を$P_{jkr}(\theta_i)$と定義する。$P_{jkr}(\theta_i)$のロジット部分の自然対数を取ったものを(3)式のように表現する。

$$(3)\quad \ln\frac{P_{jkr}(\theta_i)}{1-P_{jkr}(\theta_i)}=\theta_i-\beta_j-\gamma_k-\delta_r$$

この基本形に局所独立の仮定とモデル識別のための制約を課したものが基本的な線形ロジスティックテストモデル（LLTM; Fischer, 1983）である。モデルのパラメタ推定に関しては，標準的な項目反応モデルに適用される推定方法を拡張・一般化して用いることができる。本研究では周辺最尤推定法によってパラメタの推定を行った。なお，LLTMでは不完全なテスト評定デザイン，すなわち全ての評定者が全てのテスト項目や受験者を評定していないネストデザインを取り扱うことが可能である。

本研究では，多次元ランダム係数多項ロジットモデル（Multidimensional Random Coefficient Multinomial Logit Model, MRCML; Wu et al, 2007: 133-149）の枠組で一次元の古典的なLLTMを構成した。ここで，評定者に関連する相（facet）では，評定者rに固有の主効果（寛大化傾向δ_r）のみを推定パラメタとした。

3 方法

山口大学の平成23年度AO入試において，複数の評定者によって採点される書類審査および面接試験をテスト項目群とするLLTMによってIRT尺度化を行った。実際のAO入試においては志願者総数557人に対して評定者の数が50人を超える複雑な測定デザインであること，加えて数値計算上の理由から次節以下で述べる仮定ないしは制約を置いたモデルによってパラメタの推定を行った。

3.1 テスト項目の操作的定義

山口大学AO入試におけるテスト項目は，調査書，志望理由書（自己PR書），講義等理解力試験および面接試験である。しかし，本研究においては調査書や講義等理解力試験といった客観項目はモデルから除外し，テスト項目として「AO入試」という大項目β_jをただ1つ設定し（$j=1$），書類審査および面接試験は大項目にネスト[2)]している小項目ζ_l（$l=1,2$）であると仮定した。従って，モデルパラメタの推定においては大項目「AO入試」を構成する小項目群のプールから「書類審査」と「面接試験」の2項目（ζ_1, ζ_2）のみをサンプリングしたという状況を想定する。なお，面接試験については個人・集団面接の区別を行わず，最終的に得られた面接評定値を「面接試験」の得点として取り扱った。本研究で構成されたLLTMの具体的な相構成は**3.3**節において提示する。

表 1. 線形ロジスティックテストモデルの相構成と統計量

LLTM の相の名称	パラメタ	個数	推定値	SE	平均	分散	SD
受験者の潜在特性値	θ	557	−	−	0.000	1.190	1.091
AO 入試の困難度	β	1	−0.640	0.039	−	−	−
書類選考の相対困難度 面接試験の相対困難度	ζ	2	±0.450	0.039	0.000	−	−
評定段階の困難度（書類） 評定段階の困難度（面接）	γ	NA	NA	NA	NA	NA	NA
評定者効果（寛大化傾向）	δ	49	−	−	0.000	0.550	0.742

3.2 受験者

受験者（N=557）については，志望学部や学科・コースの区別を行わず，全体として 1 グループの山口大学 AO 入試受験者として取り扱った。この措置は，志望学部を区別することによってサンプルサイズが極端に小さい募集単位において MRCML の数値計算が収束しない，あるいは分析結果が受験者個人の特定に繋がる可能性を排除するためである。

3.3 LLTM の相構成（項目反応モデル）

推定される LLTM の各相は，AO 入試全体の困難度 β_j，書類審査および面接試験の相対困難度ζ_l，書類審査および面接試験の評定段階ごとのカテゴリ困難度γ_{lk}および評定者の寛大化傾向 δ_r（$r = 1, 2, \ldots, 49$）である。LLTM の対数ロジット表現は以下の(4)式の通りである。

$$(4)\quad \ln \frac{P_{jlkr}(\theta_i)}{1 - P_{jlkr}(\theta_i)} = \theta_i - \beta_j - \zeta_l - \gamma_{lk} - \delta_r$$

ところで，本研究は実際の入試データを用いて評定者効果を推定・検出する方法を例示することが目的であるため，書類審査や面接試験の相対困難度の推定結果ζ_lは提示するが，そのどちらが書類審査であり，どちらが面接試験であるかは非公開とする。また，個別の評定者の特定ができないように，評定者の並び順は分析に際して順不同に並び変えてある。

その他にも，MRCML の利点として受験者の性別や志望学部，高等学校の教育課程（例：普通科，商業科，工業科 etc）等の外的変数を用いて能力値の潜在回帰を行うことも可能であるが，本研究においては応用・拡張モデルは取り扱わず，古典的な LLTM 構成で尺度化を行った。

3.4 パラメタ推定

モデルパラメタの推定は，ConQuest ver. 2.0（Wu et al, 2007）を用いて行われた。MRCML は Rasch モデル族 [3]であるため，テスト項目の識別力が 1.0 であるという制約が置かれている。MRCML のモデル識別条件に関しては，ソフトウェアのデフォルトである項目困難度（本研究では「AO 入試」の困難度 β_j）を 0.0 に固定する方法ではなく，受験者の能力θの分布の平均を 0.0 とすることにより項目困難度の自由な推定を行った。数値計算上の詳細な仕様は紙幅の関係で省略するが，EM アルゴリズムによる数値計算は 72 回で収束基準に到達し，LLTM のパラメタ推定値が得られた。

4 結果

表 1 に，MRCML によるパラメタ推定結果を示す。本稿において報告可能なものは AO 入試の困難度 β_j [4]，書類審査および面接試験の相対困難度ζ_l，および評定者効果 δ_r（各評定者rについて，平均的な評定の甘さ・厳しさ）である [5]。

4.1AO 入試の全体困難度

テスト項目相として定義した「AO 入試」の困難度 β_j の推定値は－0.640（SE=0.039）であった。困難度の推定においては受験者の能力値θ の分布の平均に 0.0 という制約を置いたため（分散の推定値は 1.190 であった),困難度の自由な推定値が得られている。負の推定値は，受験者の平均的な能力水準 0.0 と比較して AO 入試の全体的な困難度が易しかったことを意味している。ただし，ここでの AO 入試の困難度は書類審査と面接試験の総合的な効果として操作的に定義されており，実際の山口大学 AO 入試の仕様や難易度を正確に反映したものではない。受験者の潜在特性値θ の最尤推定値の分布を以下の図 3 に参考として示す。

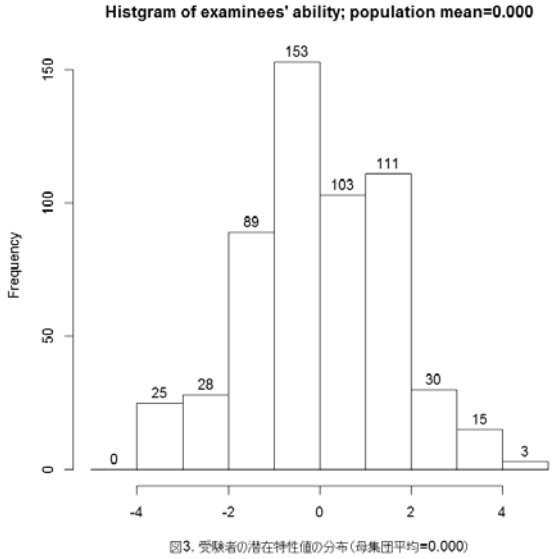

図3. 受験者の潜在特性値の分布（母集団平均=0.000）

4.2 書類審査・面接試験の相対困難度

「書類審査」および「面接試験」の相対困難度ζ_lは 0.450 および－0.450 であったが（SE=0.039），どちらがどの試験種別を示すかは非公開とする。また，MRCML による数値計算の制約上，両者の困難度の総和が 0.0 になるという制約を置く必要があるため，自由なパラメタ推定が行われたのはζ_1についてのみである。推定値を見ると，相対的に書類審査と面接試験との間には難易度差があることが分かる。なお，書類審査および面接試験の評定段階ごとのカテゴリ困難度 γ_{lk}についても，推定結果を提示することによって試験種別が特定されてしまうため，本稿では非公開とする。

表 2. 評定者効果 δ_r の要約統計量

最小値	−1.430
第一四分位	−0.487
中央値	−0.167
平均値	0.000
第三四分位	0.408
最大値	1.941
分散	0.550
標準偏差	0.742

4.3 評定者効果

分析に投入された 49 名の評定者について，受験者の能力尺度上で項目困難度と加算分割された評定者効果 δ_r の推定値が得られた。表 2 に評定者効果の要約統計量を示している。

評定者効果は 0.0 に近ければ平均的に甘くも辛くもない評定を行っていることを意味する。効果の絶対値が大きくなればなるほど，その評定者は平均的に極端な評定行動を示していることになる（正は厳しい，負は甘い評定傾向を示す)。

本研究において，評定者効果の最大値および最小値はそれぞれ 1.941（SE=0.304)，－1.430（SE=0.342）であった。受験者の能力分布（平均 0.0，標準偏差 1.091）と照らし合わせると，両者についてそれぞれが（1 標準偏差を超える）比較的大きな評定の偏りを持つと解釈することができる。なお，MRCML のモデル識別上の理由から，全評定者に渡って評定者効果の総和が 0.0 になるという数値制約を置く必要があり，自由なパラメタ推定が行われたのは分析に投入された評定者の総数 49－1=48 人分についてである [6)]。評定者効果 δ_rの推定値の分布を以下の図 4 に示す。

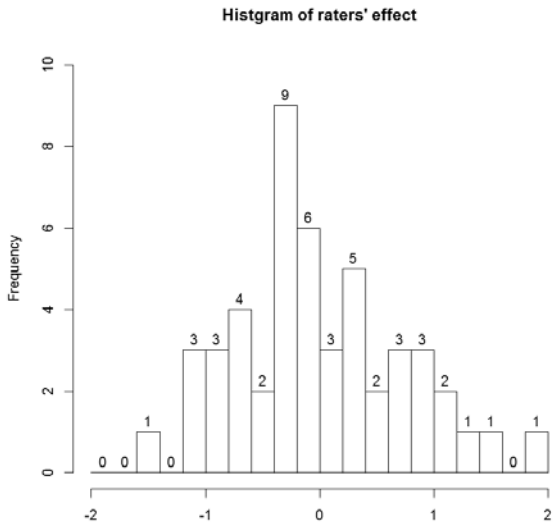

図4. 評定者効果(寛大化傾向)の分布

5 考察および今後の課題

本研究では，単純な形に整形した山口大学の平成 23 年度 AO 入試データに対して項目反応理論（IRT）を適用し，線形ロジスティックテストモデル（LLTM）の観点から評定者の特性（寛大化傾向）を潜在変数θの尺度上で計量的に測定・評価することを試みた。IRT 尺度化に際しては書類審査および面接試験の評定データのみを分析の対象とし，その他の入試資料は変数として一切含めなかった。本研究における LLTM モデリングは相当程度に単純化されたデザインにおける試行的なものではあるものの，評定者効果の推定値の分布を観察すると，評定者個々人の評価の甘さや辛さには大きな個人差が存在することが示唆された。このことは，書類審査や面接試験における評価基準の解釈が評定者によって異なる程度が大きいために，このような結果が得られたものと推測される。

入試における現実問題としては，如何に精緻な評定基準を作成しようとも，個々人の評定行動から個人差を完全に排除することは困難なものである。また，評定者トレーニングを如何に入念に行っても，その効果が見られない評定者も存在する。また，入試の最中に様々な理由によって評定者の評価基準の変動が発生する事態もあり得るだろう。そうした状況の中で入試を行う上で重要なことは，選抜における公平性を可能な限り担保するために，一人の受験者に対しては複数の（可能な限り多くの）評定者を割り当て，評価に大きな差異が生じた際にはベテランの第三者による再評定を行うなど，評価を適正に修正するためのプロセスを確認・確立しておくことである。

本研究では，書類審査および面接試験といった評定者が関わる試験科目のみをテスト項目として解析の対象としたが，今後は高校時代の成績や二次試験の結果を含めてモデル化したり，性別や高校の教育課程などの外的変数を利用した潜在回帰分析を行うなど，より包括的な入試データの解析を行うことが期待される。

また，テスト理論の観点からは，本研究で扱うような評定データにおいては IRT で一般的に仮定されている局所独立性が侵犯されているという状況を無視した尺度化を行っているため，分析結果においては困難度パラメタの推定値にバイアスが掛っていると同時に評定者数の増加に伴う情報量の過剰集積が発生している。本研究では IRT による評定者効果の測定の単純な適用例を紹介することが目的であったが，今後は評定データにおける局所従属性を考慮に入れた，より厳密で精緻な IRT モデリング（例：Wilson and Hoskins, 2001; Patz et al, 2002）を適用することが課題となる。

注

1）言語テストの分野では，多相モデル（Many-Faceted model, Facets model; Linacre, 1989）として参照されることが多い。
2）入れ子の構造を指す。
3）2 母数ロジスティックモデルにおいて，項目識別力の値がすべてのテスト項目間で共通となるような制約を課したモ

デルである。1 母数モデルとも呼ばれる。

4）「AO 入試」は，書類審査と面接試験の総合的な効果として操作的に定義されたテスト項目である。

5）総評定者数は 50 名を超えるが，共通評定者が存在しないなど，項目反応理論による測定デザインに従わない評定者は分析から除外されているため，合計 49 名が分析に投入された。

6）そのため，パラメタ値の制約を受ける評定者を変更すれば，全体のパラメタ推定結果にも影響が出ることに注意する必要がある。

参考文献

Brennan, R.L. (2001). Generalizability Theory, Springer.

Fischer, G.H. (1983). Logistic latent trait models with linear constraints. Psychometrika, **48**, 3-26.

池田央 (1973).『心理学研究法 8 テストⅡ』東京大学出版会.

Linacre, J. M. (1989). Multi-facet Rasch measurement, Chicago: MESA Press

Lord,F.M. &Novick,M. (1968). Statistical theories of mental test scores, Reading, MA: Addison-Wesley.

Lord,F.M. (1980). Applications of item response theory to practical testing problems, Hillsdale,NJ: Lawrence Erlbaum Associates.

Muraki, E. (1992). A Generalized Partial Credit Model: Application of an EM Algorithm, Applied Psychological Measurement,**16**(2), 159-176.

Patz R.J., Junker B.W., Johnson M.S.& Mariano, L.T. (2002). The Hierarchical Rater Model for Rated Test Items and its Application to Large-Scale EducationalAssessment Data, Journal of Educational and Behavioral Statistics, **27**(4), 341-384.

芝祐順（編）(1991).『項目反応理論－基礎と応用』東京大学出版会.

友安一夫 (1978).「評定者バイアス」東洋・大山正・詫摩武俊・藤永保編『心理学の基礎知識』有斐閣ブックス, 384-385.

Wilson, M., & Hoskins, M. (2001). The rater bundle model. Journal of Educational and Behavioral Statistics, **26**, 283-306.

Wu, M.L., Adams, R.J., Wilson, M.R.,& Haldane, S. (2007). ACER ConQuest Version 2.0General item response modelling software, ACER Press.