速　報

ベイズ IRT と水平等化を用いた小規模テストの年度間比較

大石　信弘*

Year-to-year Comparison of Small-scale Tests by Means of Bayesian IRT and Horizontal Equating

Nobuhiro Oishi*

Item Response Theory (IRT) has come to be used in large-scale tests such as certification exams to clearly show the pass/fail of examinees. However, for a small number of examinees, the item parameter may become negative. Then I tried to apply Bayesian statistical modeling to IRT, item parameters and ability parameters can now be obtained. And a common item design was applied to make equating the abilities of examinee of this year to last year's scale. The IRT analysis and equating small-scale tests provide a strong grounds of an argument for credit units and quality assurance.

キーワード：学業成績、小規模テスト、項目反応理論、ベイズ統計モデリング、水平等化

Keywords：Scholastic performance, Small-scale tests, Item Response Theory, Bayesian Statistical Modeling, Horizontal equating

１．はじめに

高等教育機関の使命の一つとして、学生の能力を高めて社会に送り出すことが挙げられる。対象となる学生は、入試倍率も年によって異なり、志望動機もその年その年の社会情勢を反映して入学してくるわけで、毎年カラーが違っている。そのような学生を相手に、年度が変わっても、学生の能力がある程度のレベルまで到達できていることを示すことは、非常に重要である。

しかし、現状のテスト評価では、年度間の比較は平均点や成績ヒストグラムで行うしかなく、ましてや、留年かどうかの判定基準となる点数が、毎年同じレベルであると断言することは至難の業である。

そこで、項目反応理論（IRT : Item Response Theory[(1)]）を適用することで、科目の難易度と学生の能力を別々に分けて評価し、さらに、IRT の結果を等化することによって、年度間の比較が可能になるのではないかと考えた。

ただし、ここには問題点が 2 つある。

一つは、IRT の受験者数の問題だ。通常の IRT の場合、パラメータの推定値の精度を保証するためには、受験者数が最低でも 200～400 人程度必要である[(1)]。高専の場合、テストの受験者数は、少なく見積もって、一般必修科目で 40 名から 120 名程度、専門必修科目で 40 名程度、専門選択科目で 10 名程度である。そのため、通常の IRT の解法を適用すると、パラメータが不定になったり、非負であるはずのパラメータが負の値になったりするため、不都合が生じる。このことを解決するために、ベイズ統計モデリングを IRT に適用し、少人数（10 名程度でも可）の受験者でも IRT のパラメータが安定して求められると考える。

もう一つの問題点は、等化に必要な共通項目数である。ここで、IRT における「項目」とは、試験問題のことを指していることを注記しておく。共通項目デザインで等化を行う場合、必要な共通項目数は 8 項目程度といわれている[(2)]。認証評価の際、3 割程度であれば同じ問題を出題してもよいと指導されているため、8 問が同じ問題となるような試験問題の累計数は 24 問程度となる。これは、後期中間試験まで、または後期期末試験までの試験問題累計数に相当するため、前期を終了した時点での IRT 分析はできないことになる。そのため、本論文では、試験問題の本質は変わらない、数値や向きなどだけを変えた、いわゆる「類題」も共通項目とすることで、安定した等化係数を得られると考える。

筆者はこれまで、統計解析環境 R[(3)]を用いて、因果分析[(4),(5)]や教育ツールの開発[(6)]を行ってきた。それに倣い、本論文の分析には、統計統合環境 R ver. 4.0.2 を用いた。特に、ベイズ統計モデリングに用いるマルコフ連鎖モンテカルロ法（MCMC : Markov Chain Monte Carlo）には rstan パッケージ(2.21.5)を、等化には plink パッケージ(1.5-1)を用いた。PC は Core i5-7300U CPU(4core)、8GB RAM の Windows 10 Pro(64bits)21H2 である。

*　電子情報システム工学系
〒861-1102　熊本県合志市須屋 2659-2
Faculty of Electronics and Information Systems Engineering,
2659-2 Suya, Koshi-shi, Kumamoto, Japan 861-1102
E-mail address: oishi@kumamoto-nct.ac.jp (N. Oishi).

２．分析手法について

本論文では、おもに項目反応理論（IRT）、水平等化（horizontal equating）およびMCMC法（Markov Chain Mote Carlo Method）によるベイズ統計モデリング（Baysian Statistical Modeling）の3つの手法を用いている。以後、これらの概略を簡単に説明する。詳細については、参考文献を参照していただきたい。

2.1 項目反応理論（IRT）

従来のテストでは、受験者集団（標本）の能力が違えば正答率が変わるだろうし、出題された問題（項目）の難易度によって、同じ標本でも得点が違ってくることが予想される。つまり、テストの得点や標準偏差による評価では、標本依存性や項目依存性を克服できない。そこで、能力θの受験者が、項目jに正答する確率$P_j(\theta)$がθの漸増関数で与えられるとすると、項目に依存する問題と受験者に依存する問題とに分けて考えることができる。こう考えるのが、項目反応理論である。$P_j(\theta)$としていくつかの式が提案されているが、本論文では、次式に示す、２つの母数（パラメータ）を用いたロジスティック関数で表すモデル（2PLM : 2 Parameters Logistic Model）を用いることにする。

$$P_j(\theta) = \frac{1}{1+\exp\{-Da_j(\theta - b_j)\}} \qquad (1)$$

(1)式が描く曲線は、項目特性曲線（ICC : Item Characteristic Curve）と呼ばれており、S字曲線（シグモイド）である。微分した曲線は、項目情報曲線（IIC : Item Information Curve）と呼ばれており、IICが作るピークの位置付近の情報量が多いことを示している。(1)式の中のa_jが大きいと、受験者の能力をよく識別できるので、（項目）識別力パラメータと呼ばれている。また、式中のb_jは、シグモイド曲線をシフトする効果があり、この値が大きいと、同じ能力の受験者でも正答確率が低くなるため、（項目）困難度パラメータと呼ばれている。a_jとb_jはともに、項目j特有の値である点に留意してほしい。式中のDは尺度因子と呼ばれており、2PLMのIICを正規分布に近くするための定数である。一般的に$D = 1.7$という値が、用いられている。

通常のIRTであれば、与えられた受験者ごと、項目ごとの正誤データから、対数尤度関数を能力パラメータと項目パラメータを交互に最適化していくEMアルゴリズムを用いて最尤推定値を求めている。このEMアルゴリズムでは、2PLMの場合、最小標本数は200～400だと報告されている。

2.2 水平等化

２つの試験T-setとF-setを実施した場合、それぞれのテストの尺度で受験者の能力を測ることになる。別々の尺度なので原点や目盛が異なっているため、各試験の受験者を同列に比較することができない。そこで、２つの尺度について、どちらか一方の原点と目盛を他方のそれと合わせることになる。F-setの能力パラメータθをT-setの尺度に１次変換してθ^*にする場合を考える。

$$\theta^* = A\theta + B \qquad (2)$$

等化係数AとBを決定することが等化である。これらの係数から、等化後の項目パラメータは次式のようになる。

$$a^*_{jF} = a_{jF}/A \ = a_{jT} \qquad (3)$$

$$b^*_{jF} = Ab_{jF} + B = b_{jT} \qquad (4)$$

この等化係数の推定値は、識別力パラメータの平均値を用いて、次式で得られる。

$$\hat{A} = \frac{\overline{a_F}}{\overline{a_T}} \qquad (5)$$

$$\hat{B} = \overline{b_T} \ - \hat{A}\overline{b_F} \qquad (6)$$

このように、項目パラメータの平均値を用いる方法をMean & Mean法と呼んでいる。項目パラメータの平均値を用いない方法として、共通項目のテスト特性曲線（TCC）の差異を最小にするように、等化係数を求める方法がある。この方法をStocking & Lord法と呼んでいる。

２つの尺度を等化しようとする場合、２つのテストセットには、受験者の一部か項目の一部を共通に含んでいる必要がある。それぞれ、共通受験者デザインおよび共通項目デザインと呼んでいる。

2.3 MCMC法によるベイズ統計モデリング[7]

解答パターンを$\boldsymbol{u}$、パラメータをλとしたとき、解答パターンが与えられた時のパラメータの分布$f(\lambda|\boldsymbol{u})$は、ベイズの定理より次式で与えられる。

$$f(\lambda|\boldsymbol{u}) \propto \Upsilon(\boldsymbol{u}|\lambda)p(\lambda) \qquad (7)$$

ここで、左辺$f(\lambda|\boldsymbol{u})$は、事後分布と呼ばれる。右辺の$\Upsilon(\boldsymbol{u}|\lambda)$は、パラメータを指定したもとでの解答パターンの分布を表し、尤度関数と呼ばれている。$p(\lambda)$はパラメータの分布を表しており、事前分布と呼ばれている。本論文においては、識別力パラメータおよび困難度パラメータの事前分布として、正規分布を用いた。ただし、項目のバリエーションを広くとらえるために、標準偏差を2としている。また、尤度関数としては、(1)式を0.5以上で1、それ未満で0となるように二値化するために、ベルヌイロジット分布を用いた。

$$a \sim Norm(0,2) \qquad (8)$$

$$b \sim Norm(0,2) \qquad (9)$$

$$\Upsilon \sim BernoulliLogit\big(1.7a(\theta - b)\big) \qquad (10)$$

2.4 分析に用いるデータについて

共通項目デザインによる等化を行うことから、筆者が継続して担当している、今年度（R4）と昨年度（R3）の「TE1 基礎電気学 I 」の解答パターンを用いる。出題した問題は、IRT 分析を念頭に、6 択の多肢選択問題とする。解答はマークシートに解答してもらい、スキャナで画像化した後、PC で採点する。IRT 分析する際、さらに正誤（1/0）のデータに変換して分析を行う。

受験者数は R3 で 43 名、R4 で 42 名であったが、R3 に病欠で欠席した学生がいたため、その学生の全解答を分析用のデータから削除した。つまり、分析に用いた受験者数は、どちらの年度も 42 名である。

本論文執筆時点での今年度の実施分は、前期期末試験までである。R4 には、14 問（今後 F-set と呼ぶ）が出題されている。そのうち、昨年度と共通して出題した問題は、3 問である。昨年度、これら 3 問が出題されたのは後期中間試験までであったことから、R3 の IRT 分析のためのデータとしては、後期中間試験までの 24 問（今後 T-set と呼ぶ）とする。また、数値や向きなどだけが異なり、本質的に変わらない問題を、ここでは「類題」と呼ぶことにする。類題の数は、6 問である。等化の安定性を見極めるために、共通項目数として、同じ問題の 3 問とした場合と、類題を含めた 9 問とした場合の等化係数を比較する。なお、出題した問題は公開しておらず、受験の後先で不公平にならないよう配慮した。

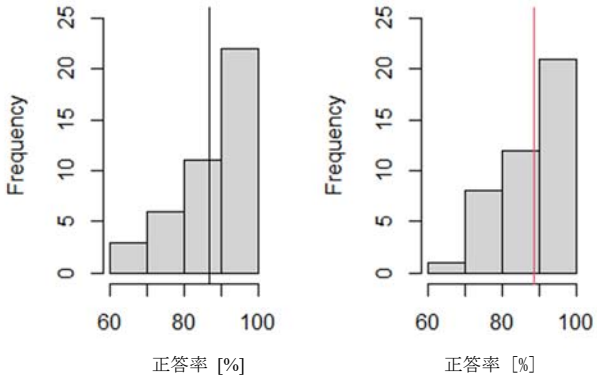
３．分析結果および考察

3.1 古典的テスト理論による分析

古典的テスト理論（CTT : Classical Test Theory）では、各受験者の正答率で能力を測り、各項目の正答率で項目困難度を表している。年度間の受験者の能力の比較を行うために、T-set ならびに F-set の正答率のヒストグラムを図 1 に示す。ヒストグラムの区切りは、学業の評価（S、A、B、C）に対応させている。T-set の平均正答率は 86.9 %、標準偏差は 9.13 であった。F-set では、88.6 %、8.93 であった。平均正答率は、図中に垂線で示す。F-sest の方が 1.7 %高いが、わずかである。最頻値は、両グループともに最高得点側の区分に表れている。CTT では、F-set のヒストグラムも T-set のヒストグラムもほとんど差異はないように思われる。

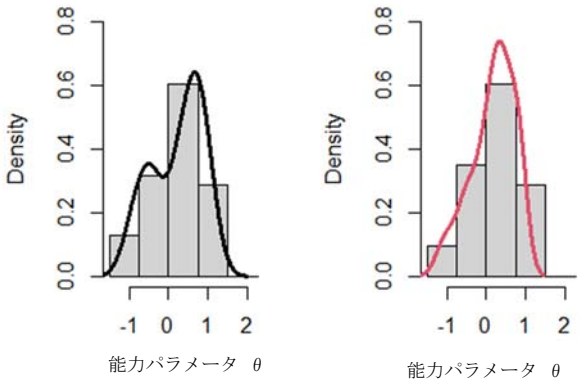
3.2 ベイズ IRT

2.3 節で述べた、ベイズモデリングを適用した IRT で、チェーン数 4、反復数 10^5、バーンイン期間 5000 として、Stan を用いて推定値を求めた。トレースプロットや$\hat{R}$、お



(a) T-set (b) F-set

図 1 正答率のヒストグラム



(a) T-set (b) F-set

図 2 ベイズ IRT による能力パラメータのヒストグラム

よび自己相関プロットで収束を判定したところ、問題なく収束していた。実行時間は T-set と F-set を合わせて、24 分程度であった。受験者数が少ない場合、EM アルゴリズムだと識別力パラメータが負になり、項目パラメータが求まらないが、今回提案のベイズ IRT では、少人数の受験者数でも問題なく項目パラメータが求まっている。ベイズ IRT による推定値のうち、年度間比較の対象となる能力パラメータのヒストグラムを図 2 に示す。分割数を図 1 と合わせたため、ヒストグラムの見え方が区間の区切りに影響を受けている。それを回避するために、図 2 には密度関数も合わせて示している。図 1 と比べて、最頻値が T-set、F-set ともに、中央付近に表れているのが特徴である。ヒストグラムでは、T-set も F-set も差異がないように見えるが、密度関数では T-set には$\theta = -0.5$付近にサブピークがあることが分かる。ピーク位置を比較すると、T-set のほうが 0.2 程大きいことが分かる。

3.3 共通項目デザインによる水平等化

共通項目デザインで採用した共通項目の正答率を表 1 に示す。全く同じ問題は、項目名の 1 文字目を”C”とし、

表 1　共通項目の正答率

項目名		CQ1	CQ2	CQ3	SQ1	SQ2	SQ3	SQ4	SQ5	SQ6
正答率	T-set	30.2	95.2	95.4	51.2	93.0	76.7	88.4	90.5	97.7
[%]	F-set	53.5	92.9	97.6	53.5	95.4	93.0	100	95.2	100.

表 2 共通項目数による等化係数の違い

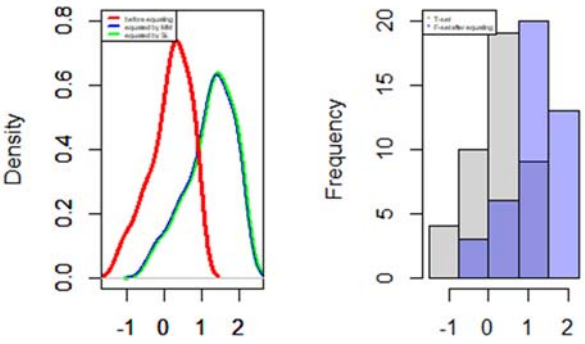
等化手法	C：3 項目		C+S：9 項目	
	A	B	A	B
Mean & Mean(MM)	0.9525	0.7158	1.1695	1.0128
Mean & Sigma(MS)	2.0572	2.1653	1.2381	1.1364
Haebara(HB)	1.1225	0.5224	0.8464	0.4333
Stocking & Lord(SL)	1.0695	0.5915	1.1602	1.0468

向きや数値などを変えた類題は、1 文字目を”S”にしている。F-set が T-set と比べて、正答率が 10％ 以上高くなっている項目は、CQ1、SQ3 および SQ4 である。多くは F-set の方が正答率が高いが、CQ2 は T-set の方が正答率が高くなっている。

共通項目として、CQ1-CQ3 の 3 項目のみとした場合と、SQ1-SQ6 の 6 項目を加えて合計 9 項目とした場合の等化係数を表 2 に示す。理想的には、手法が違っても等化係数は同じ値になるはずである。逆に、異なる手法間で同じ等化係数が見積もられれば、真の値に近づいていることを示していると考えられる。ここでは、(5)、(6)式から求められる Mean & Mean 法（MM 法）と、TCC の差異を最小にするように推定する Stocking & Lord 法（SL 法）という、まったく異なった観点の 2 つの手法に着目したい。この 2 つの手法の差の絶対値は、3 項目の場合、$\Delta A_3 = 0.117, \Delta B_3 = 0.1243$であるのに対し、9 項目に増やした場合は、$\Delta A_9 = 0.009, \Delta B_9 = 0.034$と改善されている。このことより、類題を共通項目として採用する意義は大きいと考えられる。手法としては、等化係数の定義に近い MM 法を採用する。つまり、今回の等化係数として、共通項目を 9 項目とした MM 法の係数$A_{MM9} = 1.1695, B_{MM9} = 1.0128$を採用する。

3.4 等化による年度間比較

採用した等化係数を用いて、F-set の能力パラメータを水平等化した結果を図 3 に示す。(a)図には、MM 法（青線）と SL 法（緑線）での等化後の密度関数がほぼ重なっており、同じ結果になっていることを示している。(b)図に、等化後のヒストグラムを紫で示している。灰色のヒストグラムは T-set のものである。等化前のヒストグラム(図 2）では、T-set も F-set も似たような分布になっていたものが、等化により F-set の分布が右側（高能力側）にシフトしたことが分かる。このことより、今年度（F-set）



(a) 等化後の密度関数　　(b)等化後のヒストグラム

図 3　等化による年度間比較

の受験者は、昨年度（T-set）の受験者より、実は能力が高いということが理解できる。特に、一番左の区分には、今年度の受験者が誰もいないことを示せていることは、単位の認定の際に強い根拠を与えるものと考えられる。

4．まとめ

今回、IRT にベイズモデリングを適用し、項目パラメータと能力パラメータを、40 名という少人数の受験者でも、安定して求めることができた。また、共通項目デザインのもとで、昨年度と今年度の受験者の能力パラメータを比較することができた。共通項目数を確保するためには、類題の組み込みが効果をもたらすことが分かった。

小規模テストの IRT および等化ができるようになったことで、単位の認定や質保証に対し、大きな根拠となることが期待できる。

（令和 4 年 9 月 16 日受付）
（令和 4 年 11 月 4 日受理）

参考文献

(1) 加藤健太郎，山田剛史，川端一光：「R による項目反応理論」, pp.70-281 (2014).
(2) 藤森進：「同時尺度調整法による垂直的等化の検討」，人間科学研究, 20, pp.34-47 (1998).
(3) “The R Project for Statistical Computing”, https://www.r-project.org/ , (Retrieved Sep. 16, 2022).
(4) N. Oishi, N. Yamamoto, A. Ishida and J. Murakami, “A Causal Analysis by Structural Equation Modeling of Sleep Monitoring Sensor Data”, IJEEE, Vol. 8, No. 3, pp. 58-62 (2020).
(5) 大石信弘：「TE 科 2 年生における欠点科目数の因果分析－構造方程式モデリングによるアプローチ－」, 熊本高等専門学校 研究紀要, 第 13 号, pp.67-70 (2021).
(6) 石田明男，扇崎和希，山本直樹，大石信弘，村上純：「第 87 号 60 頁の 4×4×4 立方体パズルについて(後編)」，初等数学，第 90 号，pp.18-22 (2021).
(7) 豊田秀樹：「基礎からのベイズ統計学」, pp.126-158 (2015).