【原著】 大学入試研究ジャーナル第 36 号，114–121，2026

密度比に基づいた Haebara 法の拡張等化手法

今泉 智子（東京大学），作村 建紀（法政大学），岡田 謙介（東京大学）

項目反応モデルを用いた等化方法の一つに，Haebara 法がある。Haebara 法は個別推定において等化誤差が小さく安定した推定ができることが示されている反面，理論的には確率単体の上での不一致度の指標として，本来確率を対象とするわけではない L2 ノルムを損失関数に用いるという不自然さがある。そこで本研究では，Haebara 法を拡張した密度比に基づく損失関数を用いた等化係数の推定手法を提案し，その性能を IRF 一致度の観点から数値実験によって，等化時よりも多数の能力点のもとで評価した。その結果，概ね提案手法が目的とする損失の意味で，IRF を最適化する等化が実施できていることが確認された。さらに，提案手法を用いてブラジルの大学入学共通テスト ENEM での等化を実施し異なるテストフォームにおける能力値を同一尺度上で表現することを試みた。

キーワード：等化，項目反応理論，個別推定，Haebara 法，ENEM

1 序論

異なるテストの尺度・得点を比較可能にする操作のことを尺度の等化（equating）という。等化により評価基準を一定に保つことは，異なる実施回にテストを受けた受験者を同一尺度上で採点し，結果の一貫性や公平性を確保する上で極めて重要である。

テストの標準的な統計モデルは項目反応理論（item response theory，IRT）であり，等化にあたっても IRT による方法が広く用いられている。なかでも，2 つの平行テストを双方受験した受験者の能力パラメータ θ，あるいは，1 つのテストを受験した 2 集団から推定される項目パラメータを等化するための変換式を考え，等化係数 A, B を推定する方法が最も標準的に利用されている。

等化は，大学入学共通テストのような大規模試験の実施およびその発展において重要な役割を果たす。たとえば，近年導入が進むComputer Adaptive Testing（CAT）型のcomputer-based tests（CBT）を新たに導入する際には，過去に出題した問題項目の内容や項目パラメータをデータベース化したアイテムバンクの構築が必要となるが，このアイテムバンク内の項目はすべて等化済みであることが前提となる。等化が適切に行われることで，異なる問題セットを受験したとしても，その得点が同等の尺度上で比較可能となり，受験者間の公平性が保たれる。また，従来の紙ベースのテストからCBT へ移行する場合には，各テストフォームの得点の比較可能性を慎重に検討する必要があり（Kolen and Brennan, 2014），この段階でも等化が求められる。

また，IRT ベースでの大規模テストの運用は，公平かつ精度の高い得点等化を実現できるという利点があり，導入の重要性が高まっている。日本の大学入学共通テストは現在のところIRT を用いて運用されていないが，例えばブラジルにおける大学入学のための共通テスト ENEM は毎年約 400 万人近くが受験し，日本の共通テストと類似した社会的影響力を有するテストであり，2009 年以降はIRT ベースの等化を導入して経年比較が可能な形で運用されている。

優れた等化方法の一つとされる Haebara 法（Haebara, 1980）は，Stocking-Lord 法（Stocking and Lord, 1983）と並ぶ特性曲線等化法のひとつで，Mean/Sigma 法（Marco, 1977）やMean/Mean 法（Loyd and Hoover, 1980）のようなモーメント法に比べ等化誤差が小さく，安定した推定ができることが知られている（Ogasawara, 2001）。一方で，理論的には確率単体の上での不一致度の指標として，本来確率を対象とするわけではないL2 ノルム（ユークリッド距離）を損失関数として用いるという不自然さがある。

そこで本研究では，Haebara 法の拡張として考えられる，密度比に基づく等化係数の推定を複数提案する。提案手法を既存手法と比較し，等化精度について数値シミュレーションによる検討を行う。その上で，この手法をブラジルにおける高等教育機関の入学者選抜のための大規模な共通試験であるENEM のデータを利用したIRT 分析に適用し，等化の結果を示す。

2 等化の方法

2.1 テストの統計モデル（IRT モデル）

等化の対象となる 2 つのテストは，何らかの共通要素を含むようにデザインされる（光永, 2024）。本研究で想定する共通項目デザインでは，等化する 2 つのテストに共通して出題される共通項目に対する反応

データを利用して等化を行う。

異なるテストX, Yがあり，それぞれのテストを2集団が解答し，尺度が構成されている場面を考える。X, Yはそれぞれ固有の尺度上で項目パラメータが算出されているが，ここでテストX, Yを別の集団Zに提示して両方のテストに解答させた場合，それぞれのテストにおける能力値の母平均を$\mu_{\theta X}, \mu_{\theta Y}$母分散を$\sigma^2_{\theta X}, \sigma^2_{\theta Y}$とする。このとき，$Z$に含まれる任意の受験者$i$の両テストにおける能力パラメータ$\theta_{iX}, \theta_{iY}$を標準化すると

$$\frac{\theta_{iX} - \mu_{\theta X}}{\sigma_{\theta X}} = \frac{\theta_{iY} - \mu_{\theta Y}}{\sigma_{\theta Y}} \tag{1}$$

の関係が成り立つ。式（1）をθ_{iY}について解いて$A = \frac{\sigma_{\theta Y}}{\sigma_{\theta X}}$, $B = \mu_{\theta Y} - A\mu_{\theta X}$とおくと，変換式

$$\theta_{iY} = A\theta_{iX} + B \tag{2}$$

が得られ，A倍してBを足すという操作で，テストYを基準にして，テストXの受験者の能力値尺度をテストYの受験者の能力値尺度上に表現できる。

IRTでは，ある問題（項目）に正答する確率をロジスティック曲線を用いてモデリングする。この曲線を項目反応関数（item response function, IRF）という。本研究で用いる具体的なIRTモデルとしては，最も標準的な設定の一つである，識別力と困難度の2つの項目パラメータを持つ2パラメータロジスティックモデル（2PLM）を用いる。この確率モデルは次のように書くことができる：

$$y_{ij} \sim Bernoulli(P_j(\theta_i, a_j, b_j)), \tag{3}$$

$$P_j(\theta_i) = \frac{1}{1 + \exp(-Da_j(\theta_i - b_j))}. \tag{4}$$

ここでy_{ij}は受験者iの項目jへの正誤を表す2値観測変数であり，Dは尺度因子と呼ばれ，正規累積分布を用いた項目特性曲線に近似させる場合D=1.702を用いる。また識別力$a_j \in \mathbb{R}^+$は正誤の分かれやすさを，困難度$b_j \in \mathbb{R}$は項目の難しさを表す項目パラメータである。

2.2 既存の等化方法

IRTに基づいた等化の方法には以下に述べる複数の方法が提案されている。

Mean/Sigma法（MS法）では，式（2）の関係性が，θと同様に項目困難度bにも当てはまることを利用し，次式をA, Bの推定量とする：

$$\hat{A} = \frac{s_{bY}}{s_{bX}}, \tag{5}$$

$$\hat{B} = \bar{b}_Y - \hat{A}\bar{b}_X. \tag{6}$$

ただし，$\bar{b}_X, \bar{b}_Y$は集団X，集団Yから得られたbの推定値の平均，s_{bX}, s_{bY}はそれぞれの標準偏差を表す。MS法に対して外れ値の影響が小さく，識別力パラメータも加味した方法がMean/Mean法（MM法）であり，次式をA, Bの推定量とする：

$$\hat{A} = \frac{\bar{a}_X}{\bar{a}_Y}, \tag{7}$$

$$\hat{B} = \bar{b}_Y - \hat{A}\bar{b}_X. \tag{8}$$

ただし，$\bar{a}_X, \bar{a}_Y, \bar{b}_X, \bar{b}_Y$は集団$X$，集団$Y$から得られた$a$, bの推定値の平均を表す。これらの方法は項目パラメータの分布の「モーメント」（平均と標準偏差）に依存して等化係数を算出するため，「モーメント法」と呼ばれている。

IRFの形を等化前後で一致させるように等化係数を求める方法がHaebara法であり，等化係数は以下で推定される：

$$\begin{aligned}(\hat{A}, \hat{B}) = &\\ \arg\min_{A,B} &\sum_{m=1}^{M}\sum_{j=1}^{J}[P_j(\theta_{mY}) - P^*_j(\theta_{mY})]^2 g(\theta_{mY}) \\ + &\sum_{m=1}^{M}\sum_{j=1}^{J}[P_j(\theta_{mX}) - P^{**}_j(\theta_{mX})]^2 g(\theta_{mX}),\end{aligned} \tag{9}$$

$$P_j(\theta_m) = \frac{1}{1 + \exp(-Da_j(\theta_m - b_j))},$$

$$P^*_j(\theta_m) = \frac{1}{1 + \exp\left(-\frac{Da_j}{A}(\theta_m - Ab_j - B)\right)},$$

$$P^{**}_j(\theta_m) = \frac{1}{1 + \exp\left(-Da_jA\left(\theta_m - \frac{b_j - B}{A}\right)\right)}.$$

ここで，式（9）第2項の$P^{**}_j(\theta_{iX})$は尺度Xを基準にして尺度Yを等化した結果得られたIRFを表す。また$g(\theta_X), g(\theta_Y)$はそれぞれ尺度X, Yにおける能力パラメータθの確率密度関数であり，IRFと同様にM個の求積点（m=1, 2,..., M）に等分されているものとする。また，$g(\theta_m)$は求積点mに対する重みで，$g(\theta_m) = p(\theta_m)\Delta\theta_m, \sum_{m=1}^{M} g(\theta_m) \approx 1$で求められる。$p(\theta_m)$は$g(\theta)$における求積点$\theta_m$での確率密度，$\Delta\theta_m$は$\theta$を$M$個の求積点に等分したときの求積点間隔を表す。

Stocking-Lord法では，テスト特性曲線（test characteristic curve, TCC）の差を最小にするように等化係数を求める。TCCはあるテストフォームに含まれる項目すべてについてIRFを合計したもので定義され，等化係数は以下で推定される：

$$(\hat{A}, \hat{B}) = \arg\min_{A,B} \sum_{m=1}^{M}\left[\sum_{j=1}^{J} P_j(\theta_{mY}) - \sum_{j=1}^{J} P^*_j(\theta_{mY})\right]^2 g(\theta_{mY})$$

$$+\sum_{m=1}^{M}\left[\sum_{j=1}^{J}P_j(\theta_{mX})-\sum_{j=1}^{J}P_j^{**}(\theta_{mX})\right]^2 g(\theta_{mX}), \tag{10}$$

ここまで述べた既存手法は個別のテストフォームから求められた項目パラメータの推定値のみを用いて等化を行う個別推定という枠組みに含まれる一方で，複数のテストフォームに含まれる正誤データすべてを用いて，集団間で共通の能力尺度を仮定し，この上で項目パラメタを推定する方法は共時推定法（concurrent calibration）と呼ばれる。多母集団を仮定したIRT モデルにより集団間で共通の項目パラメタ及び母集団における能力値分布を各集団において推定する。

既存手法のうち，Haebara 法は正答確率の二乗損失を能力分布で重み付けした式を最小化するパラメータ A, B を推定する重み付け最小二乗法といえ，確率単体（非ユークリッド空間）上の損失にL2 ノルム用いていることが分かる。しかし，L2 ノルムは確率の単なる数値の差を評価しているにすぎず，確率分布が持つ情報的な違いには鈍感である。特に 0 や 1 といった端点に近い確率では，僅かなずれでも情報的には大きな誤差が生じるが，L2 ノルムではそれが反映されない。

3 提案方法

3.1 提案手法における損失関数の定式化

Haebara 法は損失に確率の二乗損失を用いるが，本研究では，この損失関数を密度比をベースにした他の損失に置き換えることで，Haebara 法を拡張する枠組みを提案する。密度比は，2 つの確率分布の相対的な偏りを直接捉える尺度であり，特に端点付近でのわずかな差異も大きく評価できる特性を持つ。これにより，従来法では見落とされがちだった分布間の情報的なずれを適切に損失関数に反映させることが可能となる。

より具体的には，本研究では，(1) 密度比を直接用いた場合（ratio, RA 法），(2) 密度比の対数に基づく，自己情報量（information content）を用いた場合（IC 法），(3) 密度比に測定精度を考慮した重みをつけた場合に対応する，正答確率をパラメータに持つベルヌーイ分布間のPearson の χ^2 ダイバージェンスを用いた場合（PE 法）を検討する。ベルヌーイ分布間のPearson の χ^2 ダイバージェンスは式（14）のように正答確率の誤差の二乗を基準集団に向け等化済みの正答確率の分散の逆数で重みづけした形になっている。これにより異分散性に対処し，情報量が適切に考慮されることで推定精度の向上が期待される。PE は非対称のため，その双対の，Neymann の χ^2 ダイバージェンス（PE-Reverse, PER 法）も検討する。特に (3) では, 各能力水準における正答確率の差異を, ベルヌーイ分布間の χ^2 ダイバージェンスを用いて確率論的に評価し，その距離が最小となるように等化係数を推定する。このアプローチは，能力水準ごとの推定精度を反映した重み付けに基づき，基準フォームと等化フォームのIRF の一致度をより精密に評価するものである。IRF の整合性が高まることで，等化後に推定される能力値の一貫性が向上し，結果としてフォーム間で得られる能力分布の差異が抑制される点で，Haebara 法など従来のIRF 誤差最小化手法とは評価の基準と目的が異なる。

3.2 定式化

提案法における等化係数の推定量は，

$$RA : (\hat{A},\hat{B}) = \arg\min_{A,B}\sum_{m=1}^{M}\sum_{j=1}^{J}\left[\frac{P_j(\theta_{mY})-P_j^*(\theta_{mY})}{P_j(\theta_{mY})}\right]^2 g(\theta_{mY}) + \sum_{m=1}^{M}\sum_{j=1}^{J}\left[\frac{P_j(\theta_{mX})-P_j^{**}(\theta_{mX})}{P_j(\theta_{mX})}\right]^2 g(\theta_{mX}), \tag{11}$$

$$IC : (\hat{A},\hat{B}) = \arg\min_{A,B}\sum_{m=1}^{M}\sum_{j=1}^{J}[\log P_j(\theta_{mY})-\log P_j^*(\theta_{mY})]^2 g(\theta_{mY}) + \sum_{m=1}^{M}\sum_{j=1}^{J}[\log P_j(\theta_{mX})-\log P_j^{**}(\theta_{mX})]^2 g(\theta_{mX}), \tag{12}$$

$$PE, PER : (\hat{A},\hat{B}) = \arg\min_{A,B}\sum_{m=1}^{M}\sum_{j=1}^{J} D_{PE(PER)}(p,p^*) g(\theta_{mY}) + \sum_{m=1}^{M}\sum_{j=1}^{J} D_{PE(PER)}(p,p^{**}) g(\theta_{mX}), \tag{13}$$

$$D_{PE}(p,p^*) = (P_j^*(\theta_m)-P_j(\theta_m))^2\frac{1}{P_j^*(\theta_m)(1-P_j^*(\theta_m))}, \tag{14}$$

$$D_{PER}(p,p^*) = D_{PE}(p^*,p) \tag{15}$$

と定められる。ここで $p(y)\sim Bernoulli(P_j(\theta_m))$, $p^*(y)\sim Bernoulli(P_j^*(\theta_m))$ である。上記の目的関数は A, B に関して連続かつ微分可能であるため、数値的な最適化に適している。

4 シミュレーション

4.1 試験場面の設定

提案手法における等化精度を確認し，既存手法と比

較することを目的にシミュレーションを行った。シミュレーションの設定にあたっては，光永・前川（2012）を参考にした。まず，基準集団（等化先）を定義するための「テスト1」及びそれに続く「テスト2」の2種類のテストフォームを構成し，これらの項目に対し，毎試験後に2PLM（尺度因子はD=1.0とおく）を当てはめてIRTに基づく分析を行うものとした。テスト2には，テスト1の項目の一部である共通項目と実施時点では項目特性が未知である「独自項目」を含むものとし，これらを同時に受験者に提示し反応を得たと想定した。そして，テスト1，2に共通して含まれる共通項目に対してそれぞれ推定された項目パラメータを用いてテスト2をテスト1に等化する場面を考えた。

4.2 テストフォームと受験者集団

テスト1，2ともに総項目数50を想定し，それぞれで独立した受験者のθの分布を仮定した。テスト1の真の困難度を平均0，標準偏差1の正規分布に従う乱数から発生させ，真の識別力を，$\log(0.5)$を平均，0.2を標準偏差とする正規分布に従う対数正規乱数（平均$\fallingdotseq 0.5101$，分散$\fallingdotseq 0.0106$）から生成した。これらの項目に対し，真の能力値$\theta_1 \sim N(0, 1^2)$をとる受験者I_1人からの反応を得たものとした。

テスト2においては，テスト1実施後の項目パラメータ既知の項目をJ_{anc}項目，共通項目として推定された項目パラメータのうち困難度を降順に並べ替え，共通項目数分のブロックに分けたのち，ブロックごとに識別力が最大のものを1項目ずつ抽出し，それにテスト2の独自項目をJ_{new}追加した50項目を用いた。独自項目の真の困難度と識別力はテスト1と同様の分布から生成した。これらの項目に対し，真の能力値$\theta_2 \sim N(1.2, 0.5^2)$をとる受験者$I_2$人からの反応を得たものとした。

4.3 シミュレーションデザイン

共通項目数J_{anc}は（10, 25）の2条件とし，テスト1，2それぞれの受験者数に関して，大集団を仮定しI_1, I_2=3000とした。

それぞれの条件でm=500回ずつのテスト実施場面を実行した。まず，2PLMにおいて正答確率を求め，一様乱数と比較することで項目反応行列を生成した。本実験では正答確率を導出する際，数値安定性のため最小値1e-6、最大値1 - (1e-6)に制限（クリッピング）している。そして等化処理として，MS法，MM法，Haebara（HA）法, Stocking-Lord（SL）法 , さらに，実務上用いられることの多い共時等化（concurrent calibration, CC）法も比較対象に含める。CC法では基準群の（μ, σ^2）を（0, 1）とし, 共通項目に同一パラメータ制約を課した上で2PLモデルの同時推定を行って, 推定された他群の潜在分布の(μ, σ^2)から式(1)，(2）より等化係数（A, B）を求める。提案手法として，RA法，IC法，PE法，PER法を用いて（A, B）を推定した。全ての目的関数は離散化した能力分布で重みづけされているが，その際$[\min(\theta_1, \theta_2), \max(\theta_1, \theta_2)]$の範囲で，$M$=50の求積点を用意した。以上の実験をR version4.4.1で行い，項目パラメタは，mirt（）関数を用いて$\theta \sim N(0, 1)$と仮定した周辺最尤推定法により推定した。最適化にはoptim（）関数を利用し，最適化手法には準ニュートン（BFGS）法を指定した。収束基準は，reltol（相対収束許容値）= 1e-8とした。

4.4 評価

等化の目的は，異なるフォーム上で推定された能力値θを，共通の尺度へ線形変換することによって比較可能な状態に揃えることである。理想的には，等化後の尺度上では，同一受験者がどのフォームを受験したとしても，同じθ値が得られることが期待される。そのためには，等化係数A, Bが，基準フォームと等化フォームの項目特性曲線（IRF）の形状を適切に一致させている必要がある。

本研究における評価の目的は，等化手法によって推定された等化係数A, Bが，項目特性の一致という観点からどれだけ妥当であるかを，推定に用いたデータとは独立した基準のもとで検証することである。推定時に得られたIRFの一致だけを見るのではなく，より客観的で厳密な評価を行うために，推定時とは別に新たな能力値点を用いた検証が必要となる。そこで本研究では，各手法で推定されたA,Bを用いて，共通項目のIRFを高密度に配置した多数の能力値θ(500点)において再計算した。これは，IRFの形状のズレをより精緻に検出するためである。こうして得られた基準フォームと等化フォームのIRFを比較し，一致度の指標としてχ^2ダイバージェンス（Pearson, Neyman）を用いて評価する。

また，補足的な評価として得られた等化係数の推定値に関して，真値とのずれを表すRMSEとバイアス

$$RMSE = \sqrt{\frac{1}{m}\sum_{k=1}^{m}(A - \hat{A}_k)^2}, \sqrt{\frac{1}{m}\sum_{k=1}^{m}(B - \hat{B}_k)^2},$$

$$BIAS = E_A[\hat{A}] - A, E_B[\hat{B}] - B.$$

を算出し，係数自体の推定に関しても検討する。

4.5 結果

IRF 類似度評価の結果を図 1，図 2，等化係数推定値におけるRMSE とバイアスの結果を図 3 に示す。

IRF 類似度評価について，共通項目数 10 のとき，Pearson ダイバージェンスを類似度評価とした場合，評価と同様に，損失関数にPearson ダイバージェンスを用いたPE 法で最小値をとる。評価にNeyman ダイバージェンスを用いたときも損失関数にNeyman ダイバージェンスを用いたPER 法が上から 2 番目の一致度となっている。他の提案手法をみても，どちらのダイバージェンスを評価としても提案手法がIRF 一致度の上位に位置していることが確かめられる。

RMSE の観点から，共通項目数が少ない段階から提案手法のうちPE 法が既存手法には劣るが，ほぼ同程度の精度を示した。バイアスについても同様の傾向がみられ，特に共通項目数が少ない条件下においてもIC 法とPE 法は比較的安定した推定が得られた。 密度比（尤度比）を損失とした場合は，Haebara 法で用いられる正答確率の二乗損失（MSE），つまり絶対誤差を損失としているのに対してMSE を相対誤差で評価したものである。そのため，正答確率の小さい領域に対して過剰に誤差を重視し，オーバーフィッティングが生じる。自己情報量を損失とした場合は，正答確率の対数変換後の損失を評価する。この点は，予測値と真値の対数差に基づいて誤差を測定する平均二乗対数損失（MSLE）と類似している。MSLE は，本来外れ値や非対称分布をもつデータに対して有効であり，大きな誤差を圧縮し，小さな誤差を相対的に強調するという特性をもつ（Li, K. Liu, and S. Liu 2025）。しかし，本研究での自己情報量損失では，正答確率が極端に小さい領域において対数変換による過大なペナルティ（誤差の増幅）が生じやすい。また，最適化中にわずかに外れた推定をしても損失が極端に増え，最適化が不安定になりやすい。結果としてこれらは，RMSE やバイアスが他の損失関数よりも大きくなったと考えられる。

5 ENEM データの分析

前節ではシミュレーションデータにより，本研究で提案した方法の検討を行ったが，ここでは，実際に実施されたテストデータを用いた検討を行う。具体的には，ブラジル教育省傘下の国家教育調査研究所（INEP）が実施している中等教育修了（予定）者を対象とし，高等教育機関の入学者選抜の主流な方法である全国中等教育検定試験（Exame Nacional do Ensino Medio: ENEM）のテストデータを利用する。ENEM は，実際に実施されたテストの反応データをウェブサイトで公開しているが，本分析ではその中で 2023 年度に実施されたテストデータを対象とした。

5.1 テストデータの詳細

ENEM の試験科目と問題数については，人文科学 45 問，自然科学 45 問，言語 45 問，数学 45 問の合計 180 問の選択問題と，小論文（1000 点）で構成されている。選択問題の採点方法としてIRT が採用されており，各問題の点数を，受験者の正解と不正解の数で，問題の難易度が「簡単」,「普通」,「困難」に定義され,「簡単」の方が「困難」な問題よりも配点が小さくなる仕組みとなっている。評価は 500 点が基準点となるように設定されている。

またENEM では，病気・負傷や障害等のために，受験に際して配慮を希望する志願者に対し，個々の症状や状態等に応じた受験上の配慮を行っている。2023 年度実施分は，38,101 件の専門対応の申請があり，試験時間の延長（18173 件）や読み上げ（10721 件）･文字起こし（7507 件）支援，回答用紙の拡大（2 フォーム，拡大：5194 件，超拡大：996 件），点字対応（181 件）等のアクセシビリティ資源利用の申請 70,411 件を認めた。よってENEM で受験者に提示されるテストフォームには，一般的な受験者向けの通常フォームとは別にアクセシビリティ配慮フォームが存在する。

5.2 目的

本分析では，数学における通常フォーム（Form 1213）と 2 つのアクセシビリティ配慮（回答用紙の拡大・超拡大）フォーム（Form 1215, Form 1216）で各フォームで受験者母集団は異なると考えられるため別々の母集団とみて，まず各フォームごとに個別にパラメータ推定した後に，比較可能性のために等化を行う。この際，通常フォームを基準とし，各フォームに含まれる項目すべてを共通項目とする。正誤を二値化したのち最終的に利用したテストデータは受験者数 3500 名（Form 1213），3319 名（Form 1215），646 名（Form 1216），項目数は 1 項目（Q3）が無効（X）となっていたため，除外し 44 項目であった。具体的な分析手続きを次に示す。

5.3 IRT 分析と等化係数の推定

まずフォームごとにパラメータ推定を行った。本研究では確率的プログラミング言語 Stan を用いて，マルコフ連鎖モンテカルロ法（Markov chain Monte

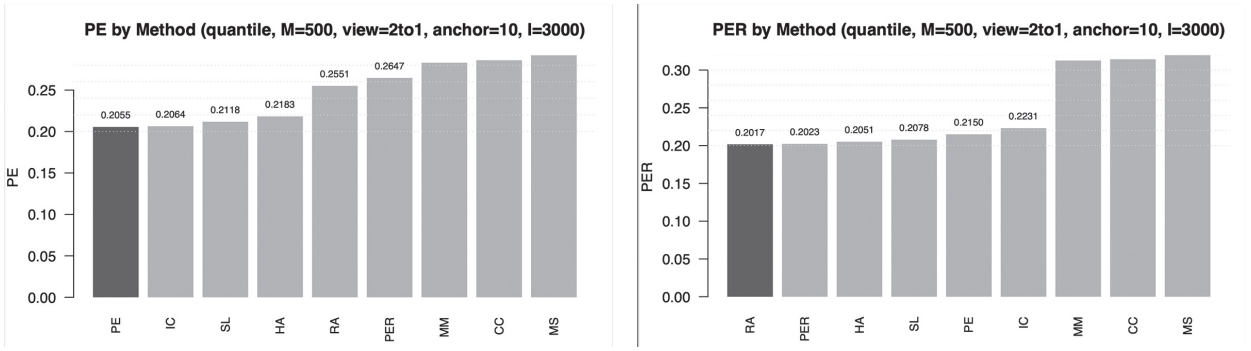


図 1　IRF 比較尺度（左が Pearson ダイバージェンス，右が Neyman ダイバージェンス。共通項目数 10）

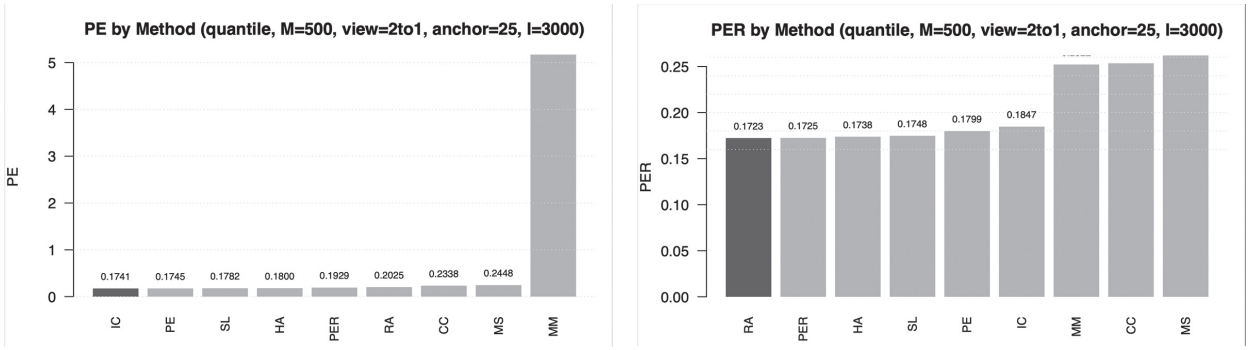
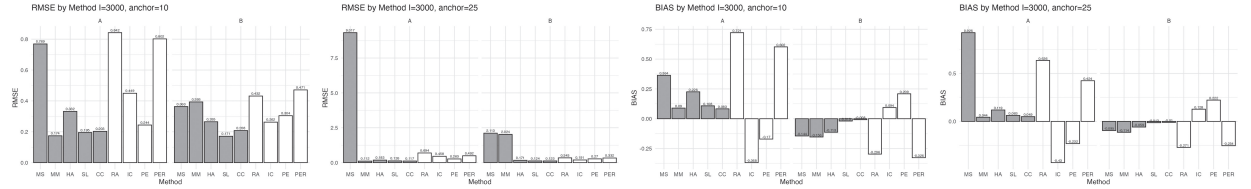


図 2　IRF 比較尺度（左が Pearson ダイバージェンス，右が Neyman ダイバージェンス。共通項目数 25）



(a) RMSE（共通項目数 10）(b) RMSE（共通項目数 25）(c) バイアス（共通項目数 10）(d) バイアス（共通項目数 25）
図 3　推定された等化係数（A, B）の性能比較（共通項目数別）。左 2 図が RMSE，右 2 図がバイアス

Carlo, MCMC）による期待事後確率推定法（Expected a posteriori, EAP）により推定値を求めた。項目パラメータの推定結果を図 4 に示す。次にForm 1213 を基準にしてForm 1215 と 1216 を等化するための等化係数の推定を行った。このとき，推定方法には前節のシミュレーション実験において提案手法内で最も等化精度が高いと判断したPE 法を用いた方法を採用した。

5.4 結果

PE 法により等化係数は，Form 1215 の場合，A=0.938, B=-0.240，Form 1216 の場合，A=1.026, B=-0.331 と推定された。図 5a は，等化前の各フォームにおける能力値（θ）の推定分布を示している。いずれのフォームにおいても, 分布は右に裾を引く非対称な形状を示しており，Form 1213 とForm 1215 では類似した分布形状である一方、Form 1216 はやや能力値が低い傾向を示していた。

一方，図 5b は，等化後の能力分布を重ねて比較したものである。Form 1213 とForm 1215 の分布（上段）は，ほぼ重なっており，等化処理によって両フォーム間で測定尺度の一貫性が保たれていることが確認できる。これに対して, Form 1213 とForm 1216 の分布（下段）では，Form 1216 の分布がForm 1213 より全体的に左側に位置している。これは，等化後においてもForm 1216 の受験者の能力値が相対的に低く推定されていることを示しており，視覚支援を要する受験者が本テストにおいて相対的に低いパフォーマンスを示していた可能性がある。背景には，受験者特性に起因する学習上の困難，あるいは支援方法の限界が考えられる。

6 結論

本論文では，Haebara 法を拡張した新たな等化手法として密度比をベースにした 3 つの損失における等化係数の推定を提案した。シミュレーションの結果より，提案手法はIRF 一致度の観点から，概ね提案手法が目的とする損失の意味で，IRF を最適化する等化が実施できていることが確認された。推定等化係数のRMSE とバイアスにおいても，Pearson のχ^2ダイバージェンスを損失としたもの（PE 法）は既存手法と同程度の精度を得られることが示された。また，ENEM データの分析では等化処理で能力値を同一尺度上に表現できたことにより，受験者母集団の特性の違いが確かめられただけでなく，「支援の効果・限界が得点や推定値に及ぼす影響」を検討する手がかりになりうることが示唆された。これは，支援の設計や公平な評価の観点からも等化が果たす役割の重要性を再認識させる結果である。

本研究では，2 値型 IRT モデルである 2PLM を仮定した等化であったが，実際の大規模テストでは記述式や段階的得点を含む多値型データも一般的であることから，多値型 IRT モデルにおける等化手法の適用と拡張は，実用性を高めるうえで今後重要な課題となる。また，よりロバストな損失関数の設計，等化係数の推定値の評価方法についても今後さらなる検討を要すると考えられる。

注

本論の一部は日本テスト学会第 23 回大会（2025）において発表された。

参考文献

Haebara, T.（1980）. “Equating logistic ability scales by a weighted least squares method,” *Japanese Psychological Research*, **22**（3）, 144–149.

Kolen, M. J. and Brennan, R. L.（2014）. “Linking,” In *Test equating, scaling, and Linking: methods and Practices*, 487–536. New York, NY: Springer New York.

Li, C., Liu, K., and Liu, S.（2025）. “A Survey of Loss Functions in Deep Learning,” *Mathematics*, **13**（15）, 2417.

Loyd, B. H. and Hoover, H. D.（1980）. “Vertical equating using the Rasch model,” *Journal of Educational Measurement*, **17**, 179–193.

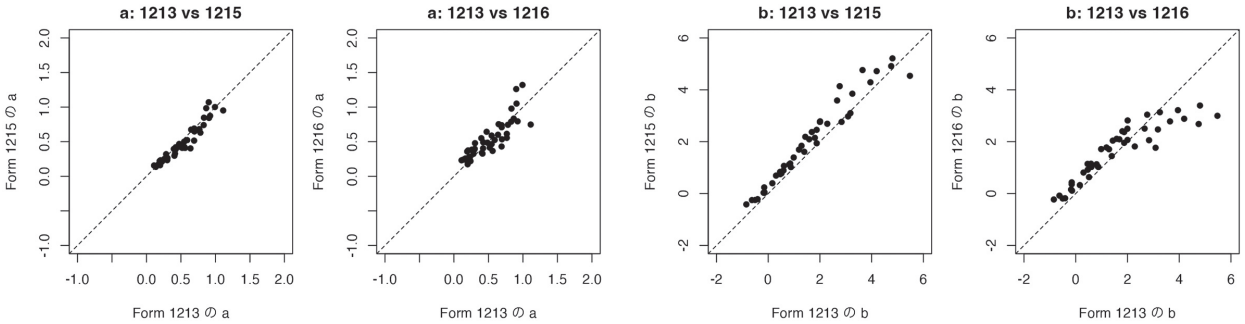
Marco,G.L.（1977）. “Item characteristic curve solutions to three intractable testing problems 1,” *ETS Research Bulletin Series*, 1977（1）, i–41.

光永悠彦・前川眞一（2012）.「項目反応理論に基づくテストにおける項目バンク構築時の等化方法の比較」『日本テスト学会誌』**8**（1）, 31–48.

光永悠彦（2024）.「心理尺度の統計的共通化：等化とリンキングの方法と実践」『統計数理』**72**（1）, 61–78.

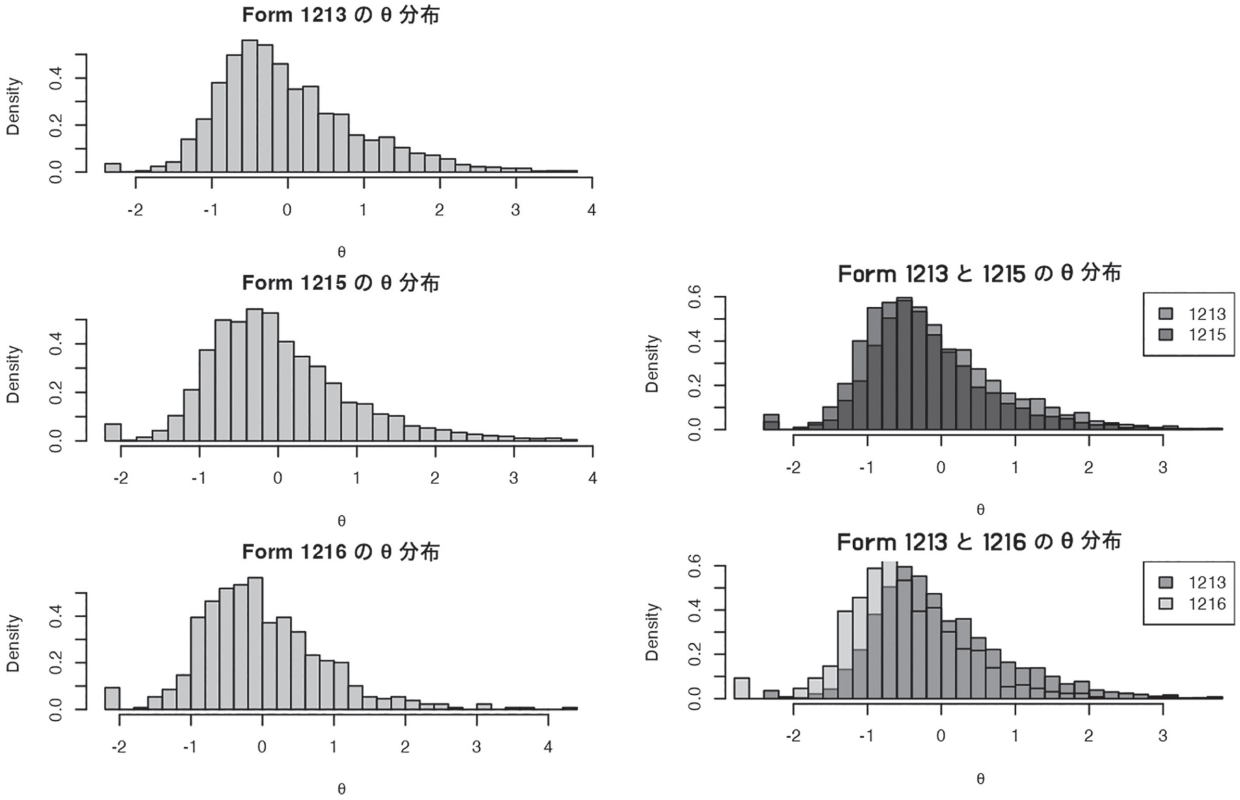
Ogasawara, H.（2001）. “Standard errors of item response theory equating/linking by response function methods,” *Applied Psychological Measurement*, **25**（1）, 53–67.

Stocking, M. L. and Lord, F. M.（1983）. “Developing a common metric in item response theory,” *Applied psychological measurement*, **7**（2）, 201–210.



(a) 等化前の各フォームにおける a 散布図　　(b) 等化前の各フォームにおける b 散布図

図 4　各フォームの等化前の項目パラメータ散布図



(a) 等化前の各フォームにおける θ の分布　　(b) 等化後の各フォームにおける θ の分布

図 5　各フォームの等化前後 θ 分布の比較