自動採点技術と項目反応理論に基づくテスト等化を通じた論述式回答評価の高精度化

荒巻洸太 [1]　宇都雅輝 [1]
[1] 電気通信大学大学院
{aramaki, uto}@ai.lab.uec.ac.jp

概要

論述式試験の問題として，評価結果が評価者の特性に依存してしまう点が挙げられる．この問題を解決するために，評価者の特性を考慮した項目反応理論が近年多数提案されている．他方で，現実の試験運用ではしばしば，複数の異なる受検者集団に同一の論述式問題を出題し，各集団の回答をそれぞれ異なる評価者集団が採点する場合がある．そのような評価結果に項目反応理論を適用して結果を比較するためには，等化と呼ばれる手続きが必要になる．等化を行うためには，集団間に共通する受検者や評価者が一般に必要となる．これに対し，本研究では，自動採点技術を用いて，共通受検者や共通評価者なしに等化を実現する手法を提案する．また，この方法を用いて，人間の評価者と自動採点を連携させた高精度な論述式回答評価を目指す．

1　はじめに

近年，様々な評価場面において，論理的思考力や表現力といった高次の能力を測定するニーズが高まっており，これらの能力を測定する手法の一つとして論述式試験が注目されている [1, 2]．一方，論述式試験の問題点として，評価結果が評価者の甘さ・厳しさなどの特性に依存してしまう点が指摘されてきた．この問題を解決する手法の一つとして，評価者の特性を考慮して受検者の能力を推定できる項目反応理論（Item response theory: IRT）と呼ばれる統計数理手法が広く研究されている [3, 4, 5, 6, 7, 8, 9]．例えば，最先端の IRT モデルの一つである一般化多相ラッシュモデル（Generalized Many Facet Rasch Model: GMFRM）[6, 8] では，多様な評価者特性を考慮した高精度な能力測定が実現できる．

一方で，現実の試験運用では，複数の異なる受検者集団に同一の論述式問題を出題し，各集団の回答をそれぞれ異なる評価者集団が採点する場合があり，そのような評価結果に GMFRM のような IRT モデルを適用し，推定された IRT パラメータ（受検者の能力値や評価者特性値）を集団間で比較したいニーズがしばしば生じる．これを実現するためには，等化と呼ばれる手続きが必要になる．一般に等化は，集団間で一部の受検者や評価者が共通するデザインで試験を実施し，それらの共通受検者や共通評価者の IRT パラメータ値を基準として，集団間でパラメータの尺度を合わせることで行われる [10, 11, 12]．しかし，実際には共通受検者や共通評価者を用意することが難しい場合もある．

そこで本研究では，共通受検者や共通評価者が存在しない状況下での等化を実現するために，近年高精度化が目覚ましい深層学習自動採点手法 [13, 14, 15] と GMFRM の統合技術である Uto & Okano[16] の手法を用いた等化手法を提案する．具体的には，尺度の基準となる集団のデータを用いて，GMFRM に基づく各受検者の能力パラメータを予測できる深層学習自動採点モデルを構築し，それを用いて等化対象集団の能力値を予測する．そして，その予測値を事前分布に反映させて，等化対象集団に対する得点データから GMFRM のパラメータをベイズ推定する．これにより，等化対象集団のパラメータ推定値を，基準集団の尺度に近づけることができる．さらに，この方法を用いて，人間評価者と自動採点を連携させた高精度な論述式回答評価を行う手法も提案する．

なお，これまでにも自動採点技術を用いて等化を試みる研究はいくつかなされているが [17, 18]，本研究のように評価者特性を考慮した IRT モデルを等化する手法は見当たらない．

2　データ

本研究では，二つの異なる受検者集団に，ある論述式問題を出題し，各集団の回答をそれぞれに異なる評価者集団が同一のスコアリングルーブリックで

採点する場合を考える．各集団から得られた得点のデータをそれぞれ $\boldsymbol{U}_1$，$\boldsymbol{U}_2$ とし，次式で定義する．

$$\boldsymbol{U}_1 = \left\{U_{jr} \in \{\mathcal{K} \cup -1\} | j \in \mathcal{J}, r \in \mathcal{R}\right\} \tag{1}$$

$$\boldsymbol{U}_2 = \left\{U_{j'r'} \in \{\mathcal{K} \cup -1\} | j' \in \mathcal{J}', r' \in \mathcal{R}'\right\} \tag{2}$$

ただし，U_{jr} は受検者 j の回答文を評価者 r が K 段階 $\mathcal{K} = \{1, \ldots, K\}$ で採点した得点を表し，$U_{jr} = -1$ は欠測データを表す．また，$\mathcal{J} = \{1, \ldots, J\}$ と $\mathcal{R} = \{1, \ldots, R\}$ は $\boldsymbol{U}_1$ における受検者と評価者の集合を，$\mathcal{J}' = \{1, \ldots, J'\}$ と $\mathcal{R}' = \{1, \ldots, R'\}$ は $\boldsymbol{U}_2$ における受検者と評価者の集合を表す．

本研究の目的は，このように受検者や評価者が異なる 2 集団以上の試験の得点データから GMFRM に基づいて推定される受検者の能力値や評価者特性値を比較できるように等化することにある．

3　項目反応理論

IRT は近年様々な試験で実用化が進められている数理モデルを用いたテスト理論の一つである．本研究では，評価者パラメータを加えた IRT モデルの中で最先端のモデルの一つである一般化多相ラッシュモデル（GMFRM）[6, 8] を用いる．

3.1　一般化多相ラッシュモデル (GMFRM)

GMFRM は IRT モデルの一つであり，テスト問題（本研究では論述式問題に対応する）i に対する受検者 j の回答文に，評価者 r が得点 k を与える確率 P_{ijrk} を次式で定義する．

$$P_{ijrk} = \frac{\exp \sum_{m=1}^{k} \left[D\alpha_i\alpha_r(\theta_j - \beta_i - \beta_r - d_{rm})\right]}{\sum_{l=1}^{K} \exp \sum_{m=1}^{l} \left[D\alpha_i\alpha_r(\theta_j - \beta_i - \beta_r - d_{rm})\right]} \tag{3}$$

ここで，θ_j は受検者 j の能力値，α_i はテスト問題 i の識別力，α_r は評価者 r の一貫性，β_i はテスト問題 i の困難度，β_r は評価者 r の厳しさ，d_{rk} は得点 k に対する評価者 r の厳しさを表すパラメータであり，D は定数 1.7 を表す．ただし，パラメータの識別性のために，$\prod_{i=1}^{I} \alpha_i = 1$, $\sum_{i=1}^{I} \beta_i = 0$, $d_{r1} = 0$, $\sum_{k=2}^{K} d_{rk} = 0$ を仮定する．GMFRM は複数の問題を出題して得られる三相データに適用できるが，2 章で述べたように，本研究では論述式問題が 1 つの場合を想定している．この場合，式 (3) は次式で表せる．

$$P_{jrk} = \frac{\exp \sum_{m=1}^{k} \left[D\alpha_r(\theta_j - \beta_r - d_{rm})\right]}{\sum_{l=1}^{K} \exp \sum_{m=1}^{l} \left[D\alpha_r(\theta_j - \beta_r - d_{rm})\right]} \tag{4}$$

本研究では，式 (4) を二つの試験結果データ $\boldsymbol{U}_1$，$\boldsymbol{U}_2$ に適用して得られるパラメータ推定値の等化を目指す．

3.2　等化

IRT モデルはパラメータの尺度に不定性があるため，異なる受検者集団・評価者集団のデータに GMFRM のような IRT モデルを適用した場合，推定されたパラメータ値はデータごとに異なる尺度で推定される．等化とは，このように異なる尺度で推定されたパラメータを共通の尺度へ変換する手続きのことである．一般的な等化手法は，共通受検者や共通評価者が存在する場合を仮定している．しかし，**1** 章でも説明したように，現実の採点場面では共通の受検者や評価者を用意するのが難しい場合がある．本研究ではこの問題を解決するために，受検者の回答文を共変量として用いることで，共通受検者や共通評価者なしで等化を行うことを目指す．

3.3　項目反応理論と共変量によるパラメータ推定手法の統合

IRT のパラメータ推定精度を向上するために，受検者由来の共変量を用いる手法が提案されてきた [19, 20, 21, 22, 23]．具体的には，各受検者の国籍や年齢といった情報を共変量として利用し，その情報を各受検者の能力値の事前分布に反映する．事前分布は，共変量を基にした正規回帰モデルとして設計される．そして，この事前分布を所与として，各受検者の能力値をベイズ推定する．この方法は，IRT のパラメータ推定精度の改善が主目的であるが，共変量が受検者の能力をある程度適切に識別できるとすれば，共変量から求めた各受検者の事前分布をもとに，異なる集団間の尺度の違いを調整することができ，等化を実現できる可能性がある．そこで，本研究のアイディアは，回答文から自動採点モデルで予測される能力値を共変量として各受検者の能力事前分布を得ることで，異なる集団のパラメータの等化を目指すというものである．

4　GMFRM を組込んだ自動採点手法

近年，GMFRM で得られる受検者の能力値 θ_j を回答文から予測できる深層学習自動採点手法が提案されている [16]．この手法では，各回答文を複数の評価者で採点した得点データ U を用いて，次の手順で深層学習自動採点モデルを訓練する．

1. 訓練データ $\boldsymbol{U}$ に GMFRM を適用し，各受検者の能力値 θ_j を推定する．
2. 受検者 j の回答文が入力，手順 1 で得られた能力値 θ_j が出力となるように，任意の自動採点モデルを訓練する．具体的には，以下の平均二乗誤差 (Mean Square Error:MSE) の最小化により

モデル学習を行う．

$$MSE(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = \frac{1}{J}\sum_{j=1}^{J}(\theta_j - \hat{\theta}_j)^2 \tag{5}$$

ここで，$\hat{\theta}_j$ は自動採点モデルによる予測値を表す．このモデルは様々な自動採点モデルで利用することができるが，本研究では先端的な深層学習モデルである BERT（Bidirectional Encoder Representations from Transformers）[24, 25] を自動採点モデルとして利用する．

この手法は，自動採点モデルの訓練データ中に混在する評価者バイアスの影響 [26, 27] を取り除くことで，安定的に自動採点モデルを訓練する目的で提案されたが，この手法を利用すると，新たな回答文を入力として GMFRM に基づく受検者の能力値を訓練データの尺度に沿って予測することが可能となる．本研究ではこの手法で予測される能力値を事前分布に用いて等化を目指す．

5 提案手法

提案手法では，基準となる集団のデータ（ここでは U_1 を基準集団とする）を用いて **4** 章で説明した能力値を予測する自動採点モデルを訓練する．その後，訓練した自動採点モデルを用いて等化対象集団（本研究では U_2 とする）における各受検者の能力値を予測し，各受検者の能力値の事前分布を設計する．そして，その事前分布と等化対象集団に対する得点データを用いてベイズ推定を行うことで，等化対象集団の GMFRM パラメータを求める．提案手法の具体的な手順は次の通りである．

1. 基準集団のデータ $\boldsymbol{U}_1$ に式 (4) の GMFRM を適用し，受検者集団 $\mathcal{J}$ の能力値 $\boldsymbol{\theta}_1$ を推定する．
2. 基準集団の各受検者の回答文と手順 1 で求めた能力推定値 $\boldsymbol{\theta}_1$ を用いて 4 章で紹介した自動採点モデルを訓練する．本研究では BERT を基礎モデルとして利用した．ここのポイントは，この自動採点モデルが基準集団の尺度に沿った能力値を回答文から予測できる点にある．
3. 訓練された自動採点モデルに等化対象集団の各受検者の回答文を入力として与え，受検者集団 $\mathcal{J}'$ に対する能力値 $\boldsymbol{\theta}_2^{pred}$ を予測する．
4. 受検者集団 $\mathcal{J}'$ の各受検者の能力値の事前分布を以下の正規分布に従うようにした上で，等化対象集団のデータ $\boldsymbol{U}_2$ を用いて GMFRM をベイズ推定することで，受検者集団 $\mathcal{J}'$ の能力値 $\boldsymbol{\theta}_2$ と評価者集団 $\mathcal{R}'$ のパラメータ $\boldsymbol{\alpha}_2$，$\boldsymbol{\beta}_2$，$\boldsymbol{d}_2$ を推定する．

$$\theta_{2j'} \sim N(\theta_{2j'}^{pred}, RMSE(\boldsymbol{\theta}_1, \hat{\boldsymbol{\theta}}_1)) \tag{6}$$

ここで，$RMSE(\boldsymbol{\theta}_1, \hat{\boldsymbol{\theta}}_1)$ は自動採点モデルの訓練時に得られた平均平方二乗誤差 (Root Mean Square Error：RMSE) を表す．

以上の手順により，等化対象集団のデータ $\boldsymbol{U}_2$ に基づくパラメータ推定時に，基準集団のデータ $\boldsymbol{U}_1$ の尺度に沿って等化対象集団の受検者に対して予測される能力値を活用できるため，共通受検者や共通評価者をもたない場合でも等化対象集団の IRT パラメータの尺度を基準集団のそれに近づけることができると期待できる．

ここで，提案手法によって等化対象集団に対して推定される能力値は，自動採点の予測と人間評価者の採点データを同時に加味して推定された標準化得点とみなせることに注意してほしい．つまり，この方法は，論述式採点でしばしば行われる，自動採点と人間評価者が連携した評価を，異なる集団間でも妥当な形で行えるようにした手法と解釈することもできる．加えて，通常であれば等化対象集団内については等化が可能なデザイン（全ての受検者に複数名の評価者が適切な条件で割り当てられる [12]）でデータを収集する必要があるが，提案手法では自動採点モデルで予測した能力事前分布によって尺度調整を目指すため，等化対象集団内部での等化ができないデザインであっても尺度調整を実現できる．以上を踏まえると，提案手法は，等化対象集団における評価者数が 1 人であっても，等化を実現しつつ，自動採点単体よりも標準化スコアの推定精度を向上できる可能性がある．

6 提案手法の有効性評価実験

6.1 実データ

本実験では，自動採点モデルのベンチマークデータとして広く利用されている Automated Student Assessment Prize (ASAP)[28] をもとにして Uto & Okano[16] が作成したデータを利用する．このデータは，ASAP に含まれる 1805 個のエッセイに対して，Amazon Mechanical Turk で募集した英語ネイティブ 38 名を一つのエッセイあたり 3〜5 名割り当てて 5 段階で採点させることで作成されている．

6.2 等化精度評価

本節では，実データを利用して提案手法の等化精度を評価する以下の実験を行った．

表 1　データ分割条件

		能力値 θ		評価者の厳しさ β	
		平均	標準偏差	平均	標準偏差
条件 1	基準群	高い	大きい	甘い	大きい
	等化対象群	低い	大きい	厳しい	大きい
条件 2	基準群	高い	大きい	甘い	大きい
	等化対象群	低い	小さい	厳しい	小さい
条件 3	基準群	高い	大きい	厳しい	大きい
	等化対象群	低い	大きい	甘い	大きい
条件 4	基準群	高い	大きい	厳しい	大きい
	等化対象群	低い	小さい	甘い	大きい

1. データセットを基準群と等化対象群の 2 つに分割した．この分割では，各群の受検者の能力値と評価者の厳しさが異なる分布になるように設定した．分割条件は表 1 に示し，詳細な分割手順は付録 A に記載した．等化対象群については，各受検者に対して評価者 1 人のみが採点を行った場合と，各受検者に対して評価者複数人が採点を行った場合のデータを作成し，それぞれのデータについて実験を行った．
2. 手順 1 で作成した基準群のデータを $\boldsymbol{U}_1$，等化対象群のデータを $\boldsymbol{U}_2$ とみなして，提案手法で等化対象群のパラメータを推定した．
3. 提案手法で予測したパラメータ値と，全データで推定されたパラメータ値（以降では「真値」と呼ぶ）との RMSE を求めた．また，比較のために，等化を適用せず各群のデータだけで推定したパラメータ値とパラメータ真値との RMSE も計算した．さらに，基準集団で訓練した BERT を用いて予測した能力値（これは等化対象集団に対する得点データがない場合に対応する）と能力真値との RMSE も計算した．
4. 手順 1 でのデータ分割を変えながら以上の手順を 10 回繰り返し，手順 3 で得られる RMSE の平均を求めた．

実験結果を表 2，3 に示す．表 2 より，提案手法により，全てのパラメータ値と真値との RMSE が，等化前と比べて，大幅に減少していることがわかる．また，表 3 より，各受検者に対して評価者 1 人のみが採点を行った場合についても，等化前に比べて RMSE が小さくなっていることが確認できる．複数評価者時と比べると，特に能力値（標準化スコア）については，複数名評価時と同程度の精度が達成できていることがわかる．このことから，提案手法を利用すると，等化対象群については受検者 1 人に対して評価者が 1 名採点を行うことで一定の精度で等化が可能であり，2 名以上の評価者がいる場合と近い精度で標準化スコアを求められることがわかる．これは，評価者を減らしても精度を維持できることを意味し，評価コストの低減につながる．また，BERT で予測した能力値を利用した場合と比べても，提案手法の性能が向上していることがわかる．このことから，4 章で紹介した自動採点手法を用いて直接に能力を予測することに加え，最低でも 1 名の人間評価者を入れて評価を行うことでより高精度な評価が可能になることが示せた．

表 2　各受検者に評価者複数人が採点をした際の各条件下における等化前後のパラメータ値と真値との RMSE

		条件 1	条件 2	条件 3	条件 4
θ	等化前	0.78	0.78	0.77	0.82
	BERT	0.61	0.54	0.59	0.51
	提案手法	**0.41**	**0.42**	**0.40**	**0.38**
α	等化前	0.44	0.59	0.34	0.50
	提案手法	**0.38**	**0.50**	**0.28**	**0.43**
β	等化前	0.77	0.74	0.75	0.75
	提案手法	**0.36**	**0.35**	**0.34**	**0.27**
d	等化前	0.19	0.38	0.17	0.33
	提案手法	**0.18**	**0.32**	**0.16**	**0.27**

表 3　各受検者に評価者 1 人が採点をした際の各条件下における等化前後のパラメータ値と真値との RMSE

		条件 1	条件 2	条件 3	条件 4
θ	等化前	0.88	0.83	0.88	0.87
	BERT	0.61	0.54	0.59	0.51
	提案手法	**0.51**	**0.48**	**0.50**	**0.45**
α	等化前	0.75	0.82	0.65	0.70
	提案手法	**0.60**	**0.69**	**0.51**	**0.58**
β	等化前	0.81	0.75	0.79	0.78
	提案手法	**0.40**	**0.36**	**0.38**	**0.31**
d	等化前	0.37	0.40	0.36	0.39
	提案手法	**0.35**	**0.36**	**0.33**	**0.36**

7　まとめ

本研究では，異なる受検者・評価者集団で構成される論述式試験の結果を比較可能にするために，回答文を共変量として，深層学習自動採点モデルを用いて GMFRM のパラメータを等化する手法を提案した．また，実データ実験により，等化による標準化スコアの推定と，人間と自動採点が協力した高精度な評価を実現できることが示せた．

謝辞

本研究は JSPS 科研費 JP23K17585, JP21H00898, JP19H05663 の助成を受けたものです.

参考文献

[1] Rebecca Schendel and Andrew Tolmie. Beyond translation: adapting a performance-task-based assessment of critical thinking ability for use in rwanda. **Assessment & Evaluation in Higher Education**, Vol. 42, No. 5, pp. 673–689, 2017.

[2] Yousef Abosalem. Assessment techniques and students' higher-order thinking skills. **International Journal of Secondary Education**, Vol. 4, pp. 1–11, 01 2016.

[3] Richard J. Patz, Brian W. Junker, Matthew S. Johnson, and Louis T. Mariano. The hierarchical rater model for rated test items and its application to large-scale educational assessment data. **Journal of Educational and Behavioral Statistics**, Vol. 27, No. 4, pp. 341–384, 2002.

[4] 宇佐美慧. 採点者側と受験者側のバイアス要因の影響を同時に評価する多値型項目反応モデル. 教育心理学研究, Vol. 58, No. 2, pp. 163–175, 2010.

[5] 宇都雅輝, 植野真臣. パフォーマンス評価のための項目反応モデルの比較と展望. 日本テスト学会誌, Vol. 12, No. 1, pp. 55–75, 2016.

[6] Masaki Uto and Maomi Ueno. Item response theory without restriction of equal interval scale for rater's score. In **Proceedings of International Conference on Artificial Inteligence in Education (AIED)**, pp. 363–368, 2018.

[7] Masaki Uto and Maomi Ueno. Empirical comparison of item response theory models with rater's parameters. **Heliyon, Elsevier**, Vol. 4, No. 5, pp. 1–32, 2018.

[8] Masaki Uto and Maomi Ueno. A generalized many-facet Rasch model and its Bayesian estimation using Hamiltonian Monte Carlo. **Behaviormetrika, Springer**, Vol. 47, No. 2, pp. 469–496, 2020.

[9] Masaki Uto. A multidimensional generalized many-facet Rasch model for rubric-based performance assessment. **Behaviormetrika, Springer**, Vol. 48, No. 2, pp. 425–457, 2021.

[10] Gary L. Marco. Item characteristic curve solutions to three intractable testing problems. **Journal of Educational Measurement**, Vol. 14, No. 2, pp. 139–160, 1977.

[11] Brenda H. Loyd and H. D. Hoover. Vertical equating using the rasch model. **Journal of Educational Measurement**, Vol. 17, No. 3, pp. 179–193, 1980.

[12] Masaki Uto. Accuracy of performance-test linking based on a many-facet Rasch model. **Behavior Research Methods, Springer**, Vol. 53, No. 4, pp. 1440–1454, 2021.

[13] Zixuan Ke and Vincent Ng. Automated essay scoring: A survey of the state of the art. In **Proceedings of International Joint Conference on Artificial Intelligence, IJCAI-19**, pp. 6300–6308, 2019.

[14] Uto Masaki. A review of deep-neural automated essay scoring models. **Behaviormetrika**, Vol. 48, No. 2, pp. 459–484, 07 2021.

[15] Paraskevas Lagakis and Stavros Demetriadis. Automated essay scoring: A review of the field. In **Proceedings of International Conference on Computer, Information and Telecommunication Systems (CITS)**, pp. 1–6, 2021.

[16] Masaki Uto and Masashi Okano. Learning automated essay scoring models using item-response-theory-based scores to decrease effects of rater biases. **IEEE Transactions on Learning Technologies**, Vol. 14, No. 6, pp. 763–776, 2021.

[17] Russell G Almond. Using automated essay scores as an anchor when equating constructed response writing tests. **International Journal of Testing, Taylor & Francis**, Vol. 14, No. 1, pp. 73–91, 2014.

[18] Suleyman Olgar. **The integration of automated essay scoring systems into the equating process for mixed-format tests**. Florida State University, 2015.

[19] Lale Khorramdel, Matthias von Davier, Eugenio Gonzalez, and Kentaro Yamamoto. **Plausible Values: Principles of Item Response Theory and Multiple Imputations**. Springer International Publishing, 2020.

[20] Robert J. Mislevy. Estimation of latent group effects. **Journal of the American Statistical Association**, Vol. 80, No. 392, pp. 993–997, 1985.

[21] Robert J Mislevy. Randomization-based inference about latent variables from complex samples. **Psychometrika**, Vol. 56, No. 2, pp. 177–196, 1991.

[22] OECD. **PISA 2018 Assessment and Analytical Framework**. OECD publishing, 2019.

[23] Janine Buchholz, Marta Cignetti, and Mario Piacentini. Developing measures of engagement in pisa. **OECD Education Working Papers**, No. 279, 2022.

[24] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In **Proceedings of naacL-HLT**, Vol. 1, p. 2, 2019.

[25] Fei Dong, Yue Zhang, and Jie Yang. Attention-based recurrent convolutional neural network for automatic essay scoring. In **Proceedings of Conference on Computational Natural Language Learning**, pp. 153–162, 2017.

[26] Evelin Amorim, Marcia Cançado, and Adriano Veloso. Automated essay scoring in the presence of biased ratings. In **Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics**, pp. 229–237, 2018.

[27] Jinchi Huang, Lie Qu, Rongfei Jia, and Binqiang Zhao. O2u-net: A simple noisy label detection approach for deep neural networks. In **Proceedings of International Conference on Computer Vision (ICCV)**, pp. 3325–3333, 2019.

[28] Ben Hamner, Jaison Morgan, lynnvandev, Mark Shermis, and Tom Vander Ark. The hewlett foundation: Automated essay scoring, 2012. https://kaggle.com/competitions/asap-aes.

A　データ分割手順

ここでは，6.2 節の実験手順 1 で概説した受検者の能力値と評価者の厳しさが異なる分布になるように設定された基準群と等化対象群を構築するための詳細な手順を説明する．

全データセットに対して GMFRM でパラメータを推定し，推定された受検者の能力値の平均と標準偏差を μ_θ^{all}, および σ_θ^{all} とし，同様に推定された評価者の厳しさパラメータの平均と標準偏差を μ_β^{all}, および σ_β^{all} とする．これらの値を使用して，基準群と等化対象群に対する受検者の能力値，および評価者の厳しさの平均と標準偏差の値を設定する．ここで，μ_θ^{ref}，および σ_θ^{ref} は，基準群の受検者の能力値の平均と標準偏差を示し，μ_β^{ref}，および σ_β^{ref} は，基準群の評価者の厳しさの平均と標準偏差を示す．同様に，μ_θ^{foc}，σ_θ^{foc}，μ_β^{foc}，および σ_β^{foc} は，等化対象群の受検者の能力値と評価者の厳しさの平均と標準偏差を示す．表 1 の 4 つの条件は，$\mu_\theta^{all}, \sigma_\theta^{all}$，$\mu_\beta^{all}, \sigma_\beta^{all}$ を基に決められており，表 4 にまとめられている．

以下の手順で，それぞれの条件に従って，基準群と等化対象群を構成した．

1. GMFRM で推定された全受検者の能力値と全評価者の厳しさを用意する．ここで，$\hat{\boldsymbol{\theta}}$ を全受検者の能力値の集合，$\hat{\boldsymbol{\beta}}$ を全評価者の厳しさの集合とする．
2. 正規分布 $N(\mu_\theta^{ref}, \sigma_\theta^{ref})$ からランダムに値をサンプリングし, サンプルされた値に最も近い $\hat{\theta}_j \in \hat{\boldsymbol{\theta}}$ に従う受検者を選択する．その受検者を基準群に追加し，$\hat{\boldsymbol{\theta}}$ から $\hat{\theta}_j$ を削除する．
3. 同様に，$N(\mu_\beta^{ref}, \sigma_\beta^{ref})$ からランダムに値をサンプリングし, サンプルされた値に最も近い $\hat{\beta}_r \in \hat{\boldsymbol{\beta}}$ に従う評価者を選択する．その評価者を基準群に追加し，$\hat{\boldsymbol{\beta}}$ から $\hat{\beta}_r$ を削除する．
4. $N(\mu_\theta^{ref}, \sigma_\theta^{ref})$ および $N(\mu_\beta^{ref}, \sigma_\beta^{ref})$ をサンプリング分布として使用し，手順 2 および 3 を実行し，等化対象群に受検者と評価者を追加する．
5. $\hat{\boldsymbol{\theta}}$，$\hat{\boldsymbol{\beta}}$ が空になるまで手順 2〜3 を繰り返す．
6. 基準群と等化対象群について，選択された受検者と評価者を基に，$\boldsymbol{U}_{ref}$，および $\boldsymbol{U}_{foc}$ のデータを作成する．
7. 各受検者が少なくとも二人の評価者から評価されることを担保するために，任意の受検者が一人の評価者からの評価しか受けていない場合，その受検者を各群から削除する．

表 4　データ分割条件の詳細

		能力値		評価者の厳しさ	
		平均	標準偏差	平均	標準偏差
条件 1	基準群	$\mu_\theta^{all} + 0.5\sigma_\theta^{all}$	1.0	μ_β^{all} - $0.5\sigma_\beta^{all}$	1.0
	等化対象群	μ_θ^{all} - $0.5\sigma_\theta^{all}$	1.0	$\mu_\beta^{all} + 0.5\sigma_\beta^{all}$	1.0
条件 2	基準群	$\mu_\theta^{all} + 0.5\sigma_\theta^{all}$	1.0	μ_β^{all} - $0.5\sigma_\beta^{all}$	1.0
	等化対象群	μ_θ^{all} - $0.5\sigma_\theta^{all}$	0.5	$\mu_\beta^{all} + 0.5\sigma_\beta^{all}$	0.5
条件 3	基準群	$\mu_\theta^{all} + 0.5\sigma_\theta^{all}$	1.0	$\mu_\beta^{all} + 0.5\sigma_\beta^{all}$	1.0
	等化対象群	μ_θ^{all} - $0.5\sigma_\theta^{all}$	1.0	μ_β^{all} - $0.5\sigma_\beta^{all}$	1.0
条件 4	基準群	$\mu_\theta^{all} + 0.5\sigma_\theta^{all}$	1.0	$\mu_\beta^{all} + 0.5\sigma_\beta^{all}$	1.0
	等化対象群	μ_θ^{all} - $0.5\sigma_\theta^{all}$	0.5	μ_β^{all} - $0.5\sigma_\beta^{all}$	1.0