# 評価者特性パラメータを付与した項目反応モデルに基づくパフォーマンス・テストの等化精度



| | |
|---|---|
| 著者 | 宇都 雅輝 |
| 雑誌名 | 電子情報通信学会論文誌．D，情報・システム |
| 巻 | J101-D |
| 号 | 6 |
| ページ | 895-908 |
| 発行年 | 2018-06-01 |
| URL | http://id.nii.ac.jp/1438/00008847/ |

doi: 10.14923/transinfj.2017LEP0027

# 評価者特性パラメータを付与した項目反応モデルに基づくパフォーマンス・テストの等化精度

宇都　雅輝[†a)]

# Accuracy of Performance Test Equating Based on Item Response Theory Models with Rater Characteristic Parameters

Masaki UTO[†a)]

**あらまし**　近年，受験者の実践的かつ高次の能力を測定する手法の一つとしてパフォーマンス評価が注目されている．一方で，パフォーマンス評価の問題として，能力測定の精度が評価者とパフォーマンス課題の特性に強く依存する点が指摘されてきた．この問題を解決する手法として，近年，評価者と課題の特性を表すパラメータを付与した項目反応モデルが多数提案され，その有効性が示されている．他方，現実の評価場面では，複数回の異なるパフォーマンステストの結果を比較するニーズがしばしば生じる．このような場合に項目反応モデルを適用するためには，個々のテスト結果から推定されるモデルパラメータを同一尺度上に位置付ける「等化」が必要となる．一般に，パフォーマンステストの等化を行うためには，テスト間で課題と評価者の一部が共通するように個々のテストを設計する必要がある．このとき，等化の精度は，共通課題や共通評価者の数，各テストにおける受験者の能力特性分布，受験者数・評価者数・課題数などの様々な条件に依存すると考えられる．しかし，これまで，これらの要因が等化精度に与える影響は明らかにされておらず，テストをどのように設計すれば高精度な等化が可能となるかは示されてこなかった．そこで本研究では，項目反応モデルをパフォーマンス評価に適用して等化を行う場合に，その精度に影響を与える要因を実験により明らかにし，その結果に基づき，高い等化精度を達成するために必要なテストのデザインについて基準を示す．

**キーワード**　パフォーマンス評価，項目反応理論，等化，テストデザイン，評価者特性，教育評価

## 1. まえがき

近年，論理的思考力や問題解決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，これを実現する手法の一つとしてパフォーマンス評価が注目されている [1]～[6]．パフォーマンス評価は，大学入試における論述式テストや外国語のリスニング試験，学習場面におけるレポート課題やプログラミング課題など，様々な形式で広く活用されてきた．また，将来の導入が検討されている「大学入学希望者学力評価テスト」においても記述式テストの採用が検討されており，パフォーマンス評価の重要性は今後ますます増加すると考えられる．

一方で，パフォーマンス評価の問題として，能力測定の精度が評価者やパフォーマンス課題の特性に強く依存する点が指摘されてきた [5]～[14]．この問題を解決する手法の一つとして，近年，評価者と課題の特性パラメータを付与した項目反応モデルが多数提案されている（e.g., [6], [7], [11]～[13]）．これらの項目反応モデルでは評価者と課題の特性を考慮して受験者の能力を推定できるため，素点の合計や平均といった単純な得点化法に比べて高精度な能力測定が可能となる [6], [7], [11], [13]．そのため，現実のパフォーマンス評価場面において，これらのモデルの実用化が強く期待されている [14]．

他方，現実の評価場面では，異なる受験者に実施された異なるパフォーマンステストの結果を比較するニーズがしばしば生じる [15], [16]．このような場合に項目反応モデルを適用するためには，個々のテスト結

† 電気通信大学，調布市
The University of Electro-Communications, Chofu-shi, 182–8585 Japan
a) E-mail: uto@ai.lab.uec.ac.jp
DOI:10.14923/transinfj.2017LEP0027

果から推定されるモデルパラメータを同一尺度上に位置付ける「等化」が必要となる．

一般に，パフォーマンステストを等化するためには，受験者・課題・評価者の一部が共通するように各テストを設計する必要がある [12], [16]～[18]．具体的には，受験者・課題・評価者の三つの相のうち，二つ以上の相について共通部分ができるようにテストを設計する必要がある [16], [18]．ただし，共通受験者を想定したテスト設計では，共通受験者の回答負担の増加や学習効果の影響が問題となるため [19]，一般には，共通評価者と共通課題を用いて等化を行うことが望ましいとされる [16], [18]．

このとき，等化が高精度に行えるかは，共通評価者と共通課題の数に強く依存する [18]．一般に共通評価者と共通課題を増やすほど等化の精度は向上する．しかし，共通評価者の増加は評価者の採点負担の増加を引き起こし，共通課題数の増加は課題内容の暴露によるテストの信頼性低下の要因となり得る（e.g., [20], [21]）．したがって，実践的には，高い等化精度を維持しつつ，できる限り共通課題と共通評価者が少なくなるようにテストを設計することが重要となる．

しかし，これまで，どの程度の共通課題と共通評価者を採用すれば高精度な等化が可能となるかは明らかにされてこなかった．Linacre [18] は，十分な等化精度を得るためには最低でも 5 名の共通評価者と五つの共通課題が必要であると述べているが，この基準を正当化する根拠は示されていない．また，客観式テストに項目反応理論を適用した場合の等化精度に関する先行研究（e.g., [19], [22]～[27]）では，どの程度の共通部分が必要であるかは，受験者の能力分布や受験者数，パラメータの推定精度などに依存することを指摘している．このことから，パフォーマンステストの等化にどの程度の共通課題と共通評価者が必要であるかは，次の要因に依存すると予想できる．

（1） 各テストの受験者・評価者・課題の特性値分布
（2） 受験者数・評価者数・課題数
（3） データの欠測率
（4） 利用する項目反応モデル

ここで，データ欠測率と利用する項目反応モデルはパラメータの推定精度に強く影響するため [6], [7]，上述の先行研究の知見から等化精度に影響を与えると推測できる．

以上の理由から，本研究では，項目反応理論に基づくパフォーマンステストの等化を共通評価者と共通課題を用いて行う場合を対象に，上記の 4 要因が等化精度に与える影響を分析し，その結果に基づいて，高精度な等化に必要な共通評価者と共通課題の数を明らかにする．具体的には，上記の 4 要因と共通評価者数・共通課題数を変えながら等化の精度を評価するシミュレーション実験を行う．その分析結果をもとに，各要因の設定条件ごとに，高精度な等化に必要となる共通評価者数と共通課題数を分析し，その基準を示す．

## 2. パフォーマンス評価データ

本研究では，パフォーマンス評価で得られるデータ $\boldsymbol{U}$ を，パフォーマンス課題 $i \in \mathcal{I} = \{1, \ldots, I\}$ に対する受験者 $j \in \mathcal{J} = \{1, \ldots, J\}$ のパフォーマンスに評価者 $r \in \mathcal{R} = \{1, \ldots, R\}$ が与える評価カテゴリー $k \in \mathcal{K} = \{1, \ldots, K\}$ の集合として次のように定義する．

$$\boldsymbol{U} = \{x_{ijr} \in \mathcal{K} \cup \{-1\} \mid i \in \mathcal{I}, j \in \mathcal{J}, r \in \mathcal{R}\}.$$

ここで，$x_{ijr} \in \mathcal{K}$ は，課題 $i$ に対する受験者 $j$ のパフォーマンスに評価者 $r$ が与える評価カテゴリーを表し，$x_{ijr} = -1$ は欠測データを表す．

また，現実のパフォーマンス評価では，評価者の採点負担を軽減するために，個々の評価対象物に数名の評価者を割り当てて採点を行わせることが多い [12], [16], [17]．特に，個々の対象物に 2 名の評価者を割り当てるデザインは，評価者ペアデザイン（Rater pair design）と呼ばれる [17]．ここで，評価者ペアデザインの例を表 1 に示す．表 1 では，チェックマークが表示された箇所は採点データが存在することを表し，空白箇所は欠測データを表す．表 1 の例では，例えば，課題 1 における受験者 1 のパフォーマンスを評価者 1 と 2 が採点し，受験者 2 のパフォーマンスを評価者 3 と 4 が採点することを表している．評価者ペアデザインの採用により，全てのパフォーマンスを全ての評価者が採点する場合と比べて，評価者の採点負担が大幅

表 1 評価者ペアデザインの例
Table 1 An example of rater pair design.

| | 課題 1 | | | | 課題 2 | | | | 課題 3 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 評価者 | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 |
| 受験者 1 | ✓ | ✓ | | | ✓ | | | ✓ | ✓ | | ✓ | |
| 受験者 2 | | | ✓ | ✓ | | ✓ | ✓ | | | ✓ | | ✓ |
| 受験者 3 | ✓ | | ✓ | | ✓ | ✓ | | | ✓ | | | ✓ |
| 受験者 4 | | ✓ | | ✓ | | | ✓ | ✓ | | ✓ | ✓ | |

に軽減される．

本研究では，このように得られるパフォーマンス評価データ $\boldsymbol{U}$ に項目反応理論を適用することを考える．

## 3. 項目反応理論

項目反応理論（Item Response Theory: IRT）とは，コンピュータ・テスティングの普及とともに，近年様々な分野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 [28]．IRT では，受験者のテスト項目への反応を，受験者の能力と項目の特性値（困難度や識別力など）で定義される確率モデルで表現する．IRT では，項目の特性と受験者の能力を分離して扱うことができるため，異なる項目群に回答した受験者の能力を，出題項目の特性や内容に依存せず同一尺度上で測定できる．また，テストや項目の測定精度を，情報量関数や標準誤差関数を用いて受験者の能力の関数として評価できる利点も有する．

このような利点から，IRT は，適応型テストや等質テスト自動構成などの現在のテスト理論の基礎として，情報処理技術者試験の一つである IT パスポート試験 [29] や医療系大学間共用試験実施評価機構による臨床実習開始前の共用試験 [30] をはじめとする様々な評価場面で活用されてきた．

IRT は，これまで，正誤判定問題や多肢選択式問題のように正誤が一意に定まる客観式テストに利用されることが一般的であった．一方で，近年では，多値型項目反応モデルを記述式テストや実技試験などのパフォーマンス評価に応用する研究も進められている（e.g., [15], [31]）．本研究で扱うようなリッカート型データに適用できる多値型項目反応モデルとしては，段階反応モデル（Graded Response Model: GRM）[32] や一般化部分採点モデル（Generalized Partial Credit Model: GPCM）[33] が知られている．しかし，これらのモデルは，課題に対する受験者の反応で構成される「受験者」×「課題」の二相データへの適用を想定しており，「受験者」×「課題」×「評価者」の三相データとなるパフォーマンス評価データに直接適用することはできない [7]．この問題を解決するアプローチとして，近年，評価者特性を表すパラメータを付与した項目反応モデルが多数提案されている（e.g., [6], [7], [11]〜[13]）．

## 4. 評価者特性パラメータを付与した項目反応モデル

本章では，本研究で用いる評価者特性パラメータを付与した項目反応モデルとして，多相ラッシュモデル（Many-facet Rasch model: MFRM）[34] と宇都・植野のモデル [11] を紹介する．

### 4.1 多相ラッシュモデル

MFRM は，評価者特性パラメータを付与した最も単純な項目反応モデルであり，パフォーマンス評価データの分析手法として古くから活用されてきた（e.g., [9], [10], [12], [17]）．MFRM には幾つかのバリエーションが存在するが [12], [35]，最も代表的なモデル化では，$x_{ijr} = k \in \mathcal{K}$ が得られる確率 $P_{ijrk}$ を次式で定義する．

$$P_{ijrk} = \frac{\exp \sum_{m=1}^{k} [\theta_j - \beta_i - \beta_r - d_m]}{\sum_{l=1}^{K} \exp \sum_{m=1}^{l} [\theta_j - \beta_i - \beta_r - d_m]}, \tag{1}$$

ここで，$\theta_j$ は受験者 $j$ の能力，$\beta_i$ は課題 $i$ の困難度，$\beta_r$ は評価者 $r$ の厳しさ，$d_k$ は評価カテゴリー $k-1$ から $k$ に遷移する困難度を表す．パラメータの識別性のために $\beta_{r=1} = 0$，$d_1 = 0$．$\sum_{k=2}^{K} d_k = 0$ を仮定する．

MFRM の特徴は，最小限の課題特性パラメータと評価者特性パラメータのみを採用している点である．一般に，項目反応モデルのパラメータ推定精度は，パラメータ数が少ないほど高くなるため [6], [7]，MFRM は類似モデルの中で最も高精度なパラメータ推定が期待できる．一方，評価者や課題の多様な特性の影響が想定される場合，MFRM ではそれらの特性を十分に表現できないため能力測定精度が低下する [6], [11]．そのため，近年では，課題と評価者のより多様な特性を考慮できるモデルが多数提案されている [6], [7], [11]．次節では，それらのモデルの中で最も多様な評価者特性を表現できる宇都・植野のモデル [11] を紹介する．

### 4.2 多様な評価者特性と課題特性を考慮した項目反応モデル

宇都・植野のモデル [11] は，評価者特性パラメータを付与した GPCM として次式で定義される．

$$P_{ijrk} = \frac{\exp \sum_{m=1}^{k} [\alpha_r \alpha_i (\theta_j - \beta_i - \beta_r - d_{rm})]}{\sum_{l=1}^{K} \exp \sum_{m=1}^{l} [\alpha_r \alpha_i (\theta_j - \beta_i - \beta_r - d_{rm})]} \tag{2}$$

ここで，$\alpha_i$ は課題 $i$ の識別力，$\alpha_r$ は評価者 $r$ の評価の一貫性，$d_{rk}$ は評価カテゴリー $k$ に対する評価者 $r$ の厳しさを表す．ただし，パラメータの識別性のため

に，$\alpha_{r=1}=1$，$\beta_{r=1}=0$，$d_{r1}=0$，$\sum_{k=2}^{K} d_{rk}=0$ を仮定する．

MFRM では評価者の厳しさと課題困難度しか考慮できなかったのに対し，このモデルでは，評価者特性として評価の一貫性と尺度範囲の制限（特定の評価カテゴリーを過剰に使用する傾向）の特性を表現できる．更に，課題特性として課題識別力の差異も捉えることができる．そのため，このモデルでは，多様な評価者特性・課題特性の生起が想定される場合に，MFRM より高精度な能力測定を実現できる [6], [11]．

### 4.3 評価者特性パラメータを付与した項目反応モデルの利点

本章で紹介した項目反応モデルをパフォーマンス評価データに適用することで，評価者と課題の特性を考慮した受験者の能力測定が可能となるため，素点の平均や合計などの単純な得点化法より高精度な能力測定値が得られる [6], [7], [11], [13]．また，通常の IRT と同様に，これらのモデルでは，受験者の能力と評価者特性，課題特性が分離して表現されるため，評価者や課題が変化しても同一尺度上で受験者の能力を測定できる．加えて，評価者や課題がもつ能力測定精度を情報量関数などで評価することも可能である．このような利点から，これらのモデルは，現実のパフォーマンス評価場面での実用化が強く期待されている [14]．また，最適な評価者の割り当て [36] や異質評価者の分析 [11] などの様々な応用も進められている．

一方で，現実の評価場面では，異なる受験者に実施された異なるパフォーマンステストの結果を比較するニーズがしばしば生じる [15]．このような場合に項目反応モデルを適用するためには，個々のテスト結果から推定されるモデルパラメータを同一尺度上に位置付ける「等化」が必要となる．

## 5. パフォーマンステストの等化

パフォーマンス評価データに項目反応理論を適用して等化を行う方法の一つとして，受験者集団の能力分布，課題集合の特性値分布，評価者集団の特性値分布が等しくなるように各テストを設計・実施するアプローチが考えられる [18]．全ての特性値分布がテスト間で等しい場合，テストごとの評価データから推定されたモデルパラメータは常に等化されているとみなせる．しかし，現実のテスト場面では，必ずしもこの仮定を満たすようなテスト実施が可能とは限らない．

特性値分布がテストごとに異なる場合に等化を行う方法としては，**1.** で議論したように，テスト間で評価者と課題の一部が共通するようにテストを設計する方法が一般的である [12], [16]～[18]．

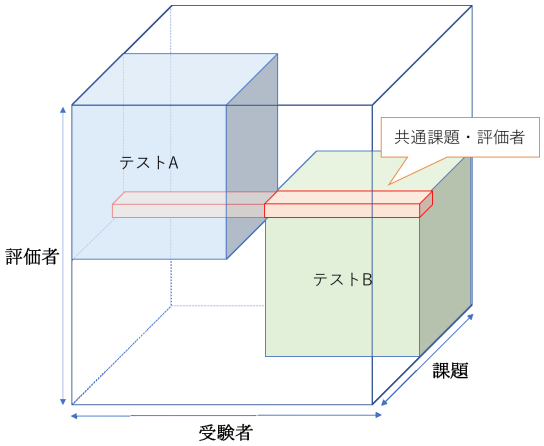


図 1 共通評価者・共通課題を用いた等化デザインの概念図
Fig. 1 Equating design using common raters and common tasks.

ここで，図 1 に，二つのパフォーマンステスト（テスト A，テスト B と呼ぶ）を共通評価者と共通課題を用いて等化する場合のデータ構造の概念図を示す．**2.** で定義したように，パフォーマンス評価データは「受験者」×「課題」×「評価者」の三相データとなるため三次元配列として図示できる．図では，色付けされた領域にデータが存在し，それ以外の領域は欠測データであることを表す．図のように，課題と評価者の一部がテスト間で共通するようにデータを収集し，得られたデータからモデルパラメータを推定することで，共通課題と共通評価者のパラメータ推定値を基準として，その他のパラメータ推定値を同一尺度上に位置づけることが可能となる．

このとき，等化が高精度に行えるかは，共通評価者と共通課題の数に強く依存する [18]．一般に共通評価者と共通課題を増やすほど等化の精度は向上するが，**1.** で述べたように，これらの数は，等化の精度を保ちつつできる限り少なくすることが実践的には望ましい．しかし，これまで，高精度な等化にどの程度の共通課題と共通評価者が必要であるかは明らかにされてこなかった．**1.** で議論したように，等化に必要な共通課題数や共通評価者数は，1）各テストの受験者・評価者・課題の特性値分布，2）受験者数・評価者数・課題数．3）データの欠測率，4）利用する項目反応モデル，の

4 要因に強く依存すると考えられる．そこで，本研究では，この 4 要因が等化の精度に与える影響の分析を通して，高精度な等化に必要な共通評価者と共通課題の数を明らかにする．

## 6. 等化精度の評価手法

本研究では，$I$ 個の課題で構成された二つのパフォーマンステスト（以降，それぞれをテスト A，テスト B と呼ぶ）を，それぞれ $J$ 人の受験者が受験し，$R$ 人の評価者集団が採点する場合に，共通評価者と共通課題を用いて等化を行うことを考える．本研究では，この設定のもとで，諸条件を変えながら等化精度を分析する．

等化精度の評価法には複数の方法が知られているが（例えば, [17], [23], [37], [38]），一般には，等化誤差を定義し，その誤差が小さいほど精度が高いと評価するアプローチが用いられる．ここでは，等化誤差の算出手法として，Ihan [17] のアプローチを採用する．この手法では，両テストにおいて受験者・課題・評価者の特性値分布が等しい場合，共通部分の有無にかかわらず全てのモデルパラメータが等化済みとみなせる性質を利用し，このときのパラメータ推定精度と等化を行ったときのパラメータ推定精度を比較することで等化誤差を評価する．これらのパラメータ推定精度の差異が小さいほど，等化誤差が小さい，すなわち，等化精度が高いと解釈する．

このアプローチに基づき，本研究では，各パフォーマンステストに任意の異なるパラメータ分布を与えたときの等化誤差を以下の手順で求める．

（1） テスト A とテスト B について，モデルパラメータの真値をランダムに生成する．具体的には，$I$ 個の課題，$J$ 人の受験者，$R$ 人の評価者のパラメータを，所与とした 2 種類のパラメータ分布に従って 2 テスト分発生させる．

（2） テスト間に，$N_{CR}$ 人の共通評価者と $N_{CI}$ 個の共通課題を設置する．具体的には，テスト A から選択した $N_{CR}$ 人の評価者と $N_{CI}$ 個の課題のパラメータ値を，テスト B からランダムに選択した $N_{CR}$ 人の評価者と $N_{CI}$ 個の課題のパラメータ値に置換する．

（3） モデルパラメータを所与として，各テストの評価データをランダムサンプリングする．

（4） 生成したデータを用いてモデルパラメータを推定し，その推定値と手順（1）で生成したパラメータ真値との平均平方 2 乗誤差（RMSE: Root Mean Square Error）を求める．ここで，パラメータ推定はマルコフ連鎖モンテカルロ法（Markov chain Monte Carlo: MCMC）[7], [13] を用いた期待事後確率推定法（EAP: Expected a posteriori）で行い，二つのテストのパラメータは同時に推定する．事前分布には表 2 の分布を用いる．ここで $N(\mu, \sigma)$ は平均 $\mu$，標準偏差 $\sigma$ の正規分布を表す．

表 2 パラメータ分布
Table 2 Parameter distributions.

| | |
|---|---|
| $\log \alpha_i \sim N(0.1, 0.4)$, | $\log \alpha_r \sim N(0.0, 0.5)$ |
| $\beta_i \sim N(0.0, 1.0)$, | $\beta_r \sim N(0.0, 1.0)$ |
| $d_{rk} \sim N(0.0, 1.0)$, | $d_k \sim N(0.0, 1.0)$ |
| $\theta_j \sim N(0.0, 1.0)$ | |

（5） 手順（2）から（4）と同様の手順を，テスト A とテスト B のパラメータ真値を同一分布（表 2 の分布）から発生させた場合についても行い，このときの RMSE を求める．

（6） 手順（4）と手順（5）で得られた RMSE の平均絶対誤差を求める．これまでの議論から，この値は等化誤差の指標の一つとみなせる．ただし，本研究では，データやパラメータの乱数生成による影響を考慮し，以上の手順を 10 回繰り返して求めた RMSE の平均絶対誤差を等化誤差として用いる．

本研究では，上記の手順で求められる等化誤差を用いて，共通評価者数 $N_{CR} = \{0, 1, 2, 3, 4, 5\}$，共通課題数 $N_{CI} = \{0, 1, 2, 3, 4, 5\}$ と変化させながら等化精度を評価することで，任意の条件下において高精度な等化を実現するために必要な共通評価者と共通課題の数を分析する．

ここで，上記の等化誤差の解釈を説明するために，$J = 100$，$I = 10$，$R = 10$ で構成される二つのテストに MFRM を適用して等化する場合を考える．パラメータ分布は以下のように与えられたと仮定する．

> テスト A では能力パラメータ $\theta_j$ が $N(1.0, 0.5)$ に従い，テスト B では $N(-1.0, 0.5)$ に従う．その他のパラメータは標準正規分布 $N(0.0, 1.0)$ に従う．

図 2 に，等化誤差の計算手順（4）を 10 回繰り返して得られた RMSE を，$N_{CR}$ と $N_{CI}$ の条件ごとに示した．また，図 3 に，手順（5）を 10 回繰り返して得られた RMSE を示した．すなわち，図 2 が，テストごとに分布が異なる場合に対応し，図 3 が両テストに同一の分布を仮定した場合に対応している．各図

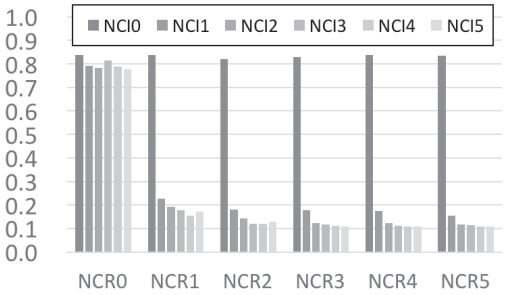


図 2　テストごとに分布が異なる場合のパラメータ推定誤差
Fig. 2　Parameter estimation errors when different ability distributions were assumed for two tests.

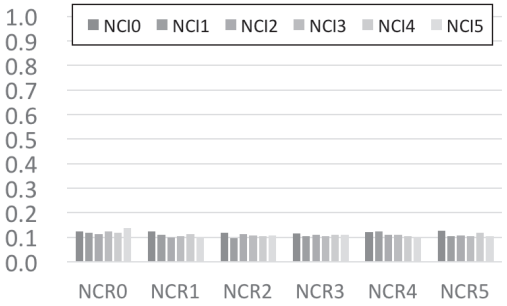


図 3　テスト間で分布が等しい場合のパラメータ推定誤差
Fig. 3　Parameter estimation errors when same ability distributions were assumed for two tests.

の縦軸は RMSE を表す．

図 3 から，テスト間でパラメータ分布が等しい場合には，共通評価者や共通課題の有無にかかわらず常に等化されているとみなせるため，$N_{CR}$ と $N_{CI}$ の値に依らず安定したパラメータ推定精度を示していることが分かる．一方で，テストごとに分布が異なる場合に対応する図 2 では，共通評価者と共通課題の両方が存在しないと等化がなされないため，どちらか一方でも共通部分が存在しないとき RMSE が極端に大きい値を示している．また，図 2 では，共通部分が増加するほど RMSE が減少していき，最終的にテスト間で分布が一致している場合と同程度の精度に収束していることも確認できる．本研究で用いる等化誤差は，これらの RMSE の差異で定義される．本研究では，等化誤差が十分に小さいとき（$\delta$ 未満のとき），高精度に等化がなされたと解釈する．

ここで，等化精度が高いと解釈する基準 $\delta$ には 0.04 を採用する．表 2 に示したように，項目反応理論における能力パラメータの分布には標準正規分布を仮定することが一般的である．標準正規分布を仮定すると，推定値の 95% は −2 から 2 の区間に収まる．能力パラメータの推定誤差 0.04 とは，能力値 −2 以下を 0 点，能力値 2 以上を 100 点とみなした線形変換において得点の差異が 1 点未満となる場合に対応しており，基準値の一つとして妥当であると解釈できる．

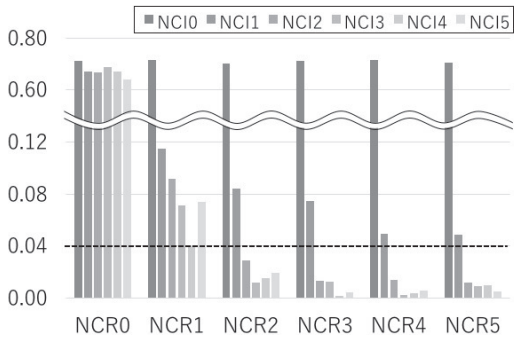


図 4　図 2 と図 3 から求めた等化誤差
Fig. 4　Equating error calculated from Fig. 2 and 3.

ここで，図 4 に，図 2 と図 3 から求めた等化誤差を示した．図の縦軸が誤差値を表し，破線は基準値 $\delta = 0.04$ を表している．図 4 より，共通部分の増加に伴い等化誤差が減少していることが分かる．また，この条件においては，共通評価者数 $N_{CR} \geq 2$ かつ共通課題数 $N_{CI} \geq 2$ において等化誤差が基準を下回っており，十分な精度で等化がなされたと解釈できる．

以降の実験では，上記の方法で等化精度の評価を行い，図 4 と同様のグラフでその結果について議論する．

## 7. 評価実験

**5.** で述べたように，本研究では，等化精度に影響を与えると想定される次の 4 要因と共通評価者・共通課題数を変化させながら **6.** の方法で等化精度を評価する．(1) 個々のテストにおける受験者・評価者・課題の特性値分布，(2) 受験者数・評価者数・課題数，(3) データの欠測率，(4) 利用する項目反応モデル．

このような評価実験は，実際にテストを設計・実施して収集した実データを用いて行うことが望ましいといえる [17]．しかし，このためには，様々な条件で繰り返しテストを実施する必要があり，膨大なコストと時間を要する．そのため，等化に関する研究では，テストの規模や実施方法，受験者の特性値分布などを現実に近い条件に設定して，シミュレーションにより評価を行うことが一般的である [22]〜[24], [26], [27], [37]．

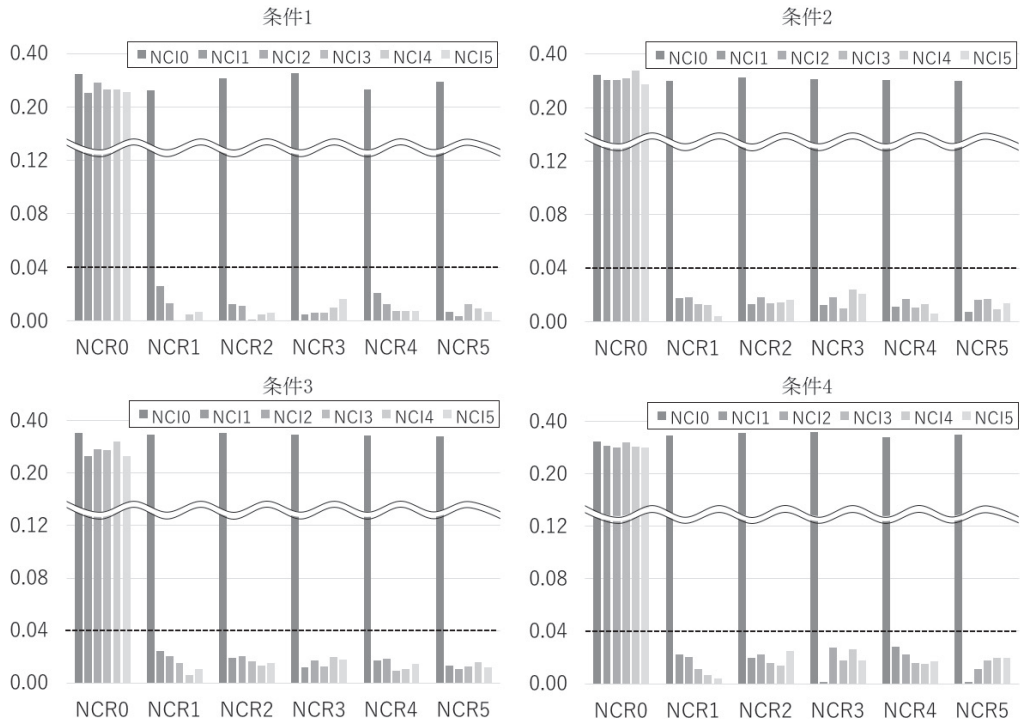


図 5　異なるパラメータ分布を用いた場合の等化誤差
Fig. 5　Equating error for different parameter distributions.

IRT を用いて客観式テストの等化精度を評価した先行研究では，このようなシミュレーションによる評価結果 [22] が実データを用いた評価結果 [19] とおおむね同様の結論を導いており，シミュレーション評価の妥当性が伺える．以上の理由から，本研究でも，シミュレーションにより等化精度の評価を行う．

### 7.1　分布の差異が等化精度に与える影響の分析

ここでは，テストごとの受験者の能力分布，評価者特性分布，課題特性分布を変化させながら等化精度を評価する．具体的には，各テストにおけるパラメータ分布を表 3 の四つの条件に変化させながら等化誤差を求めることで評価を行った．ここで，条件 1 はテスト間で受験者の能力値分布のみに差がある場合を表し，条件 2 と条件 3 は，受験者の能力値分布に加え，評価者または課題の特性分布がテスト間で異なる場合に対応している，また，条件 4 は，受験者の能力値分布，評価者特性分布，課題特性分布の全てがテスト間で異なる場合を表現している．等化の精度に関する先行研究では，分布の平均値がテスト間で 0.5 ～ 1 程度異なることを想定することが一般的であるため [19], [22], [24], [27]，本研究でも分布の平均値の差異がテスト間で 1 になるように各分布を設定した．本実験では，$J = 100$，$I = 10$，$R = 10$ に固定し，項目反応モデルには，最も広く利用される MFRM を用いた．

表 3　各テストのパラメータ分布
Table 3　Parameter distributions given to each test.

| | テスト A | テスト B |
|---|---|---|
| 条件 1 | $\theta_j \sim N(0.5, 0.5)$ | $\theta_j \sim N(-0.5, 0.5)$ |
| 条件 2 | $\theta_j \sim N(0.5, 0.5)$ | $\theta_j \sim N(-0.5, 0.5)$ |
| | $\beta_r \sim N(0.5, 0.5)$ | $\beta_r \sim N(-0.5, 0.5)$ |
| 条件 3 | $\theta_j \sim N(0.5, 0.5)$ | $\theta_j \sim N(-0.5, 0.5)$ |
| | $\beta_i \sim N(0.5, 0.5)$ | $\beta_i \sim N(-0.5, 0.5)$ |
| 条件 4 | $\theta_j \sim N(0.5, 0.5)$ | $\theta_j \sim N(-0.5, 0.5)$ |
| | $\beta_r \sim N(0.5, 0.5)$ | $\beta_r \sim N(-0.5, 0.5)$ |
| | $\beta_i \sim N(0.5, 0.5)$ | $\beta_i \sim N(-0.5, 0.5)$ |

*記載のないパラメータの分布には表 2 の分布を用いる

本実験の結果を図 5 に示す．図から，全ての条件において，共通評価者と共通課題のいずれか一方でも存在しない場合には等化誤差が著しく大きいことが分かる．前章でも議論したように，テスト間で分布が異なる場合には，共通評価者と共通課題が最低一つ以上存在しないと等化がなされないことが確認できる．

共通評価者と共通課題を一つ以上仮定した場合で比

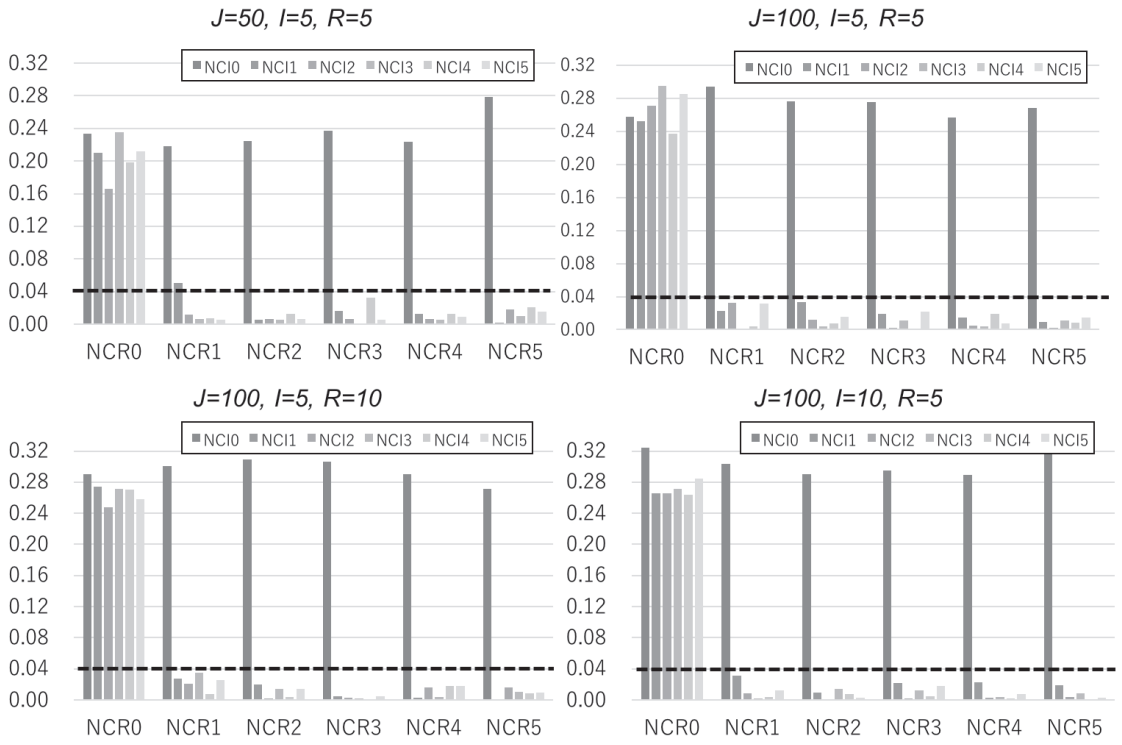


図 6　異なる受験者数・課題数・評価者数における場合の等化誤差
Fig. 6　Equating error for different number of examinees, tasks and raters.

較すると，共通評価者と共通課題をそれぞれ一つずつ導入した時点で等化誤差が劇的に低下したことが分かる．結果として，$N_{CR} \geq 1$ かつ $N_{CI} \geq 1$ の全ての場合において，等化誤差が基準値 $\delta = 0.04$ を下回っている．このことから，テスト間のパラメータ分布差が一般的な範囲であれば，共通評価者と共通課題は一つずつで十分に等化可能であることが示された．

他方，テスト間のパラメータ分布の差異が極端に大きい場合には，等化誤差は増加することが知られている [19], [27]．**6.** で等化誤差の解説に用いた図 4 は，パラメータ分布の平均値がテスト間で 2 離れており，分布差が極端に大きい例に対応している．図 5 の場合とは異なり，図 4 では，共通評価者と共通課題が一つずつのみでは必ずしも等化誤差が基準値 $\delta = 0.04$ を下回っておらず，等化誤差が増加したことが確認できる．この例の場合，高精度な等化精度を達成するためには，共通評価者と共通課題が最低でも二つずつ必要であることが読み取れる．

### 7.2　受験者数・課題数・評価者数が等化精度に与える影響の分析

本節では，テストごとの受験者数，課題数，評価者数が等化精度に与える影響を分析する．具体的には，次の四つに条件を変えながら等化誤差を比較した．

（1）　$J = 50$，$I = 5$，$R = 5$
（2）　$J = 100$，$I = 5$，$R = 5$
（3）　$J = 100$，$I = 5$，$R = 10$
（4）　$J = 100$，$I = 10$，$R = 5$

なお，テストごとのパラメータ分布は表 3 の条件 1 とし，モデルは前節の実験と同様に MFRM を用いた．

各条件における等化誤差を図 6 に示す．図 6 から，$J = 50$，$I = 5$，$R = 5$ 以外の条件では，$N_{CR} \geq 1$ かつ $N_{CI} \geq 1$ の全ての場合で等化誤差が基準値 $\delta = 0.04$ を下回ったことが分かる．$J = 50$，$I = 5$，$R = 5$ では，他条件と異なり，$N_{CR} = 1$，$N_{CI} = 1$ のときに等化誤差が基準値を超えており，微小ではあるが他の条件より等化精度が低いと解釈できる．この条件の特徴は，受験者数が他条件より少ない点にある．IRT を用いた客観式テストの等化に関する先行研究 [19], [22], [24], [37]

でも，受験者数が少ないときに等化精度が低下する傾向があることを報告しており，本研究の結果と一致する．このことから，パフォーマンステストにおいて高い等化精度を得るためには，一定数以上の受験者数が必要となることが示唆された．

他方，本実験においては，課題数と評価者数の増減は等化精度に大きな影響は与えていないことが確認できる．ただし，藤森 [22] は，客観式テストの等化において，テスト項目数が過度に増加すると等化精度が低下する傾向があることを報告している．したがって，パフォーマンス評価においても，課題数や評価者数を過度に増加させた場合には等化精度が低下すると予測される．しかし，パフォーマンス評価では，課題当りの回答負担が大きいことから，一度のテストに出題できる課題数は少数であることが一般的である．また，各テストを採点する評価者数も，テストの実行可能性の観点から極端に大きくすることは難しい場合が多い [39]．すなわち，現実のパフォーマンステストでは，課題数・評価者数が大規模になることは少なく，そのような条件下では課題数・評価者数の変化が等化精度に与える影響は小さいと解釈できる．

### 7.3 データ欠測が等化精度に与える影響の分析

これまでの実験では，各テスト内において，全ての課題に対する全ての受験者のパフォーマンスを全ての評価者が採点することを想定していた．一方で，**2.** で述べたように，実際の評価場面では，評価者の負担を減少させるために個々のパフォーマンスに数名の評価者のみを割り当てて採点を行うことが多い．このような採点デザインでは，データに多数の欠測が生じる．一般に，データ数の減少はパラメータ推定の精度低下を引き起こし，パラメータ推定精度の低下は等化精度の低下要因の一つとなることが知られている [19]．

そこで，本節では，このような採点デザインにより生じるデータ欠測が，パフォーマンステストの等化精度にどのように影響するかを評価する．ここでは，個々のパフォーマンスに $N_R$ 名（ただし，$R \geq N_R \geq 2$）の評価者を割り当てることを考える．この割り当てを行うために，本研究では，**2.** で紹介した評価者ペアデザインを，$N_R$ 人を割り当てる評価者集合デザインに拡張する．Ihan [17] は，評価者ペアデザインを等化可能な条件のもとで生成するための評価者割り当てアルゴリズムを提案している．このアルゴリズムでは，全ての評価者ペアを列挙し，各評価者ペアに採点対象物を一つずつ順に割り当てる．同様のアプローチに基づき，3 名以上の評価者の割り当てに対応させたアルゴリズムを Algorithm1 に示す．

**Algorithm 1**：評価者集合デザイン

Input: $\mathcal{I}$, $\mathcal{J}$, $\mathcal{R}$, $N_R$

評価者割当 $\boldsymbol{Z} = \{z_{ijr} \in \{0,1\} \mid i \in \mathcal{I}, j \in \mathcal{J}, r \in \mathcal{R}\}$ を初期化する．ここで，$z_{ijr}$ は課題 $i$ において評価者 $r$ が受験者 $j$ に割当られたとき 1，それ以外のとき 0 をとる変数を表す．

評価者集合 $\mathcal{R}$ を所与として，$N_R$ 人で構成される評価者集合の全ての組合せ $\boldsymbol{C} = \{\boldsymbol{C}_1, \cdots, \boldsymbol{C}_H\}$ を生成する．ただし，$H = {}_R C_{N_R}$．

$h = 0$.
**for** $i \in \mathcal{I}, j \in \mathcal{J}$ **do**
  **for** $r \in \boldsymbol{C}_h$ **do**
    Set $z_{ijr} = 1$.
  **end for**
  $h = h + 1$.
  **if** $h > H$ **then**
    $h = 0$.
    $\boldsymbol{C}$ の順序をシャッフル．
  **end if**
**end forreturn** $\boldsymbol{Z}$

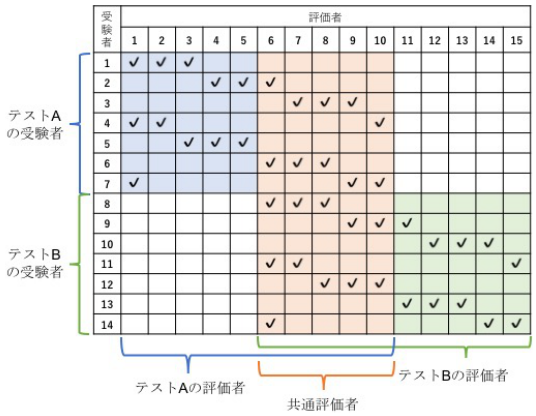


図 7　共通課題に対する評価データの概念図
Fig. 7　Example of rating data for a common task.

また，図 7 に，本アルゴリズムによる評価者割り当ての例を示す．図は，$R = 10$，$N_R = 3$，$N_{CR} = 5$ としたときの，共通課題に対する評価者割り当ての例である．図では，チェックマークが表示された箇所が評価者が割り当てられていることを表す．図から，各受験者に 3 名の異なる評価者が割り当てられており，各テストの 10 名の評価者のうち 5 名がテスト間で共通していることが分かる．

本実験では，本アルゴリズムで得られた評価者割り当てを所与として，等化誤差の算出手順（3）で生成したデータについて，評価者が割り当てられていない

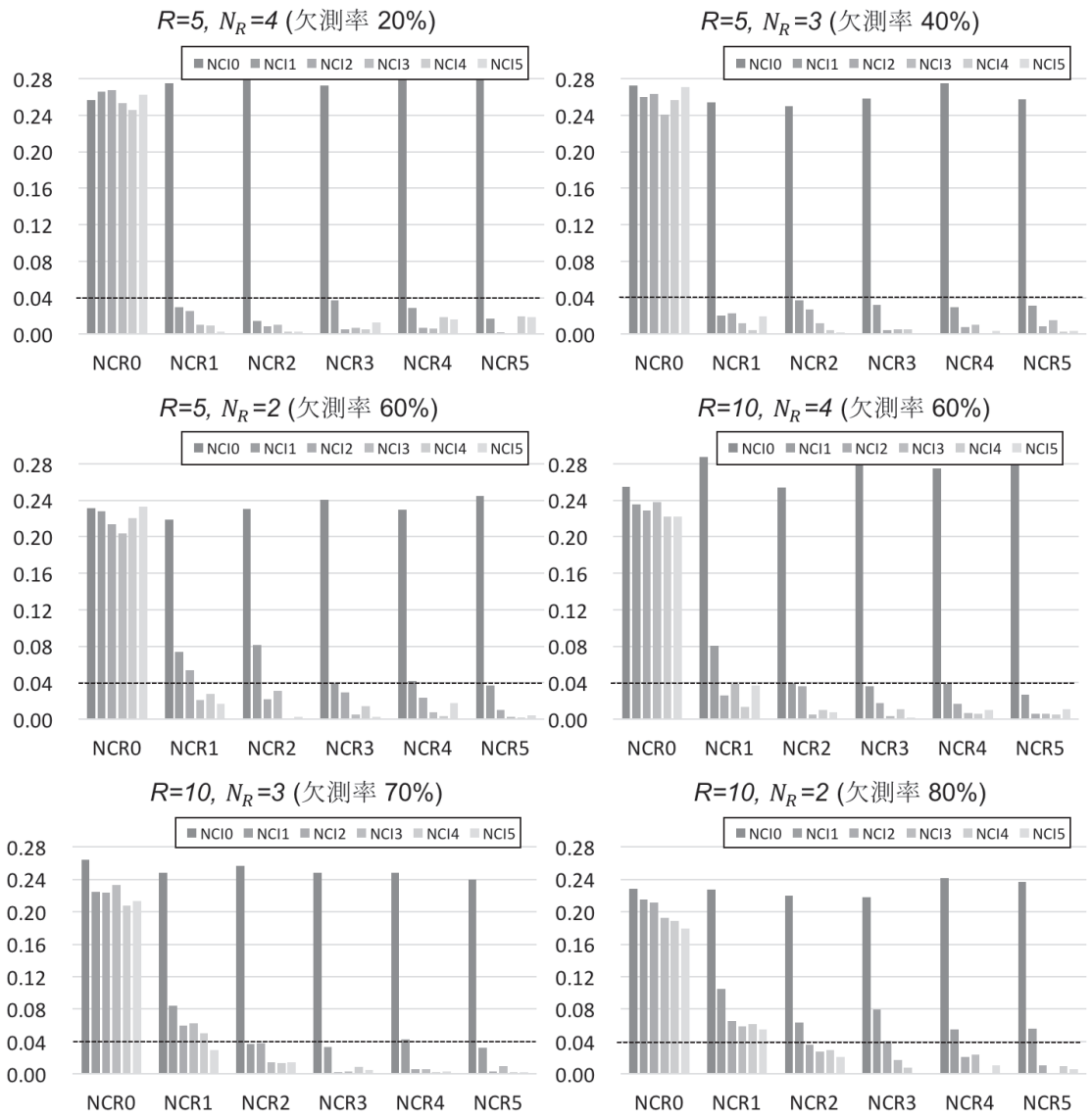


図 8　データ欠測率を変化させたときの等化誤差（$J = 100, I = 10$）
Fig. 8　Equating error for missing data.

箇所を欠測値に変更したデータを用いて等化誤差を求める．このとき，欠測率は $1-(N_R/R)$ として計算できる．ここでは，欠測率が変動するように，$R$ と $N_R$ を次の六つの値に変更させて等化誤差を評価した．

（1）　$R = 5$，$N_R = 4$（欠測率 20%）
（2）　$R = 5$，$N_R = 3$（欠測率 40%）
（3）　$R = 5$，$N_R = 2$（欠測率 60%）
（4）　$R = 10$，$N_R = 4$（欠測率 60%）
（5）　$R = 10$，$N_R = 3$（欠測率 70%）
（6）　$R = 10$，$N_R = 2$（欠測率 80%）

なお，本実験では，受験者数 $J = 100$，課題数 $I = 10$ とし，テストごとのパラメータ分布には表 3 の条件 1 を仮定し，項目反応モデルはこれまでの実験と同様に MFRM を用いた．

本実験の結果を図 8 に示す．図 8 及び欠測がない場合の結果（図 6 の右下）から，欠測率が一定以上増加すると等化誤差が増加する傾向が確認できる．具体的には，欠測率 40%までは，共通評価者・共通課題とも

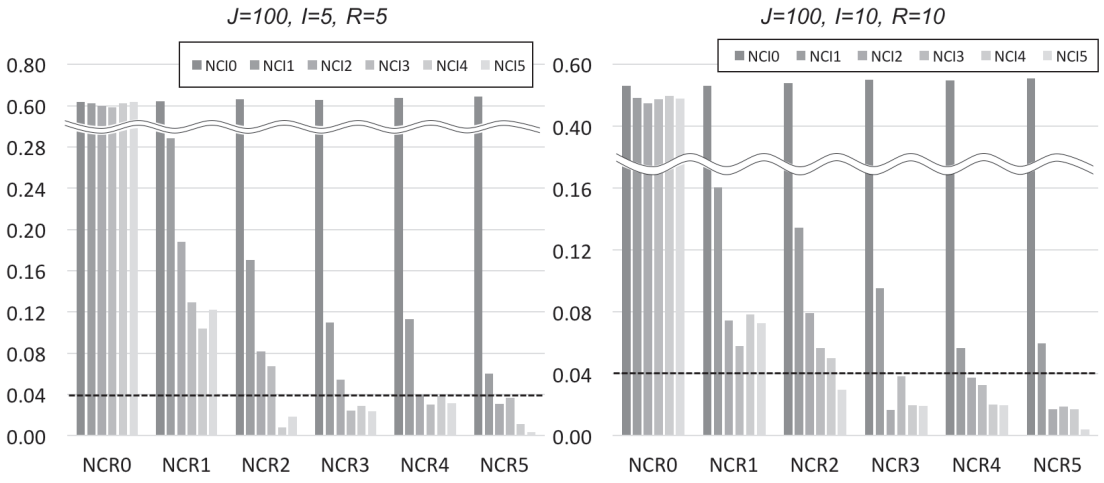


図 9　宇都・植野モデルの等化誤差
Fig. 9　Equating error for Uto and Ueno model.

に一つずつのみで等化誤差が基準値を下回っており，誤差の明らかな増加は認められない．しかし，欠測率60%以上では，基準値を下回るために共通評価者・共通課題のいずれか，または両方が二～三つ程度必要になっており，等化誤差が増加していることが確認できる．

等化誤差の増加の要因としては，欠測率の増加に伴うパラメータ推定精度の低下が考えられる．実際，$J = 100$，$I = 10$，$R = 10$ において，完全データからパラメータを推定した場合の RMSE が 0.103 であるのに対し，欠測率 80%では 0.240 であった．このことから，欠測率の大幅な増加がパラメータ推定精度の低下を引き起こし，それが等化精度低下の要因となることが分かる．

### 7.4　項目反応モデルの選択が等化精度に与える影響の分析

上記の実験では，評価者特性パラメータを付与した項目反応モデルの中で最も単純な MFRM を用いていた．一方で，**4.** で述べたように，近年では，課題と評価者の特性をより柔軟に表現できるモデルも提案されている．これらのモデルは MFRM より多くのパラメータを有するため，パラメータ数に対するデータ数が少なくなり，MFRM に比べてパラメータ推定精度が低下する傾向がある [6]．本節では，モデルの複雑化に伴うパラメータ推定精度の低下が等化精度に影響を与えるかを分析する．具体的には，**4.2** で紹介した宇都・植野のモデルを用いて等化精度を評価する．

ここでは，各テストのパラメータ分布として表 3 の条件 1 を仮定し，次の 2 条件について宇都・植野のモデルを用いて等化誤差を算出した．

（1）　$J = 100$，$I = 5$，$R = 5$

（2）　$J = 100$，$I = 10$，$R = 10$

実験結果を図 9 に示す．図 9 を，MFRM の結果（図 6 の右上が上記（1）に対応し，図 5 の条件 1 が上記（2）に対応）と比較すると，宇都･植野のモデルでは全体として等化誤差が増加していることが分かる．具体的には，MFRM ではどちらの条件下でも，$N_{CR} \geq 1$ かつ $N_{CI} \geq 1$ で等化誤差が基準値 $\delta = 0.04$ を下回っていたが，宇都・植野のモデルでは，$\delta = 0.04$ を下回るために $N_{CR} \geq 3$ かつ $N_{CI} \geq 2$ 程度の共通部分が必要となることが読み取れる．

等化誤差の増加要因としては，欠測率が増加した場合と同様に，パラメータ推定精度の低下が挙げられる．ここで，$J = 100$，$I = 5$，$R = 5$ における宇都・植野のモデルと MFRM の RMSE を確認したところ，それぞれ 0.149 と 0.103 であり，宇都・植野モデルの方が RMSE が大きかった．この結果から，複雑なモデルを採用した場合，データ数に対するパラメータ数の増加によりパラメータ推定精度が低下し，それが等化精度の低下要因となり得ることが示唆された．

## 8.　総 合 考 察

以上の実験により，等化に必要な共通評価者と共通課題の数について，以下の点が明らかとなった．

（1） MFRM のような単純なモデルを用いる場合，テスト間の特性値分布の差異が，等化に関する一般的な研究で想定される範囲であれば，共通課題・共通評価者ともに一つずつで十分な等化精度が得られる．分布の差異が極端に大きい場合でも，共通課題と共通評価者は二つ程度で等化可能である．

（2） 複雑な項目反応モデルを採用する場合には，より多くの共通課題と共通評価者が必要となる．具体的には，最も多様な評価者特性パラメータを付与した宇都・植野のモデルの場合，共通課題と共通評価者は 2 ～ 3 つ程度必要である．

（3） データの欠測が極端に増加する場合には，共通課題・共通評価者数を増加させる必要がある．具体的には，MFRM を用いる場合に欠測率が 60%を超えるときには共通評価者と共通課題が 2 ～ 3 つ程度必要となる．

（4） 一般的なパフォーマンステストでは，評価者数や課題数が等化精度に与える影響は小さいが，受験者数の増加は精度の改善に寄与する．したがって，欠測率が増加する場合や複雑なモデルを採用する場合には，受験者数が多くなるようにテストを運用することで，等化精度を改善することができる．

## 9. むすび

本研究では，項目反応理論に基づくパフォーマンステストの等化を共通評価者と共通課題を用いて行う場合に，等化精度に影響を与える要因について分析を行い，高精度な等化に必要な共通評価者と共通課題の数について一つの基準を示した．具体的には，パフォーマンステストの等化精度に影響を与えると想定される要因として，（1）各テストにおける受験者・評価者・課題の特性値分布，（2）受験者数・評価者数・課題数，（3）データの欠測率，（4）利用する項目反応モデル，を想定し，これらの 4 要因と共通評価者数・共通課題数を変えながら等化の精度を評価するシミュレーション実験を行った．実験結果に基づき，各要因の条件設定ごとに，高精度な等化に必要となる共通評価者数と共通課題数について述べた．

本研究の結果，MFRM のような単純な項目反応モデルを用い，欠測率が過度に多くなければ，共通評価者数・共通課題数はともに一つずつで高い等化精度が得られることが明らかとなった．また，より多様な状況を想定した場合でも，共通評価者数と共通課題数はともに二～三つずつ程度でおおむね十分な等化精度が得られることが示された．**1.** や **5.** で述べたように，実践的には，一定の等化精度を維持しつつ，共通評価者数と共通課題数をできる限り少なくするようにテストを設計することが望まれる．本研究の結果は，先行研究 [18] で恣意的に設定された共通部分数と比べて，大幅に少ない数でも十分に等化が可能であることをデータで示しており，実践的に有益な知見を与えたと考える．

本研究では，共通評価者と共通課題を用いた等化に着目したが，共通受験者を用いる等化のデザインについても設計は可能である．また，本研究では，二つのテストを等化する場合を考えたが，長期的あるいは大規模なテストにおいては，より多くのテストを等化する場合も想定できる．今後は，このようなデザインでの等化についても研究を行いたい．

また，本研究では，等化精度に関する研究で一般的なシミュレーションによる評価方法を採用したが，今後は実データを用いた等化精度の評価も行いたい．

**謝辞** 本研究は JSPS 科研費 17H04726 の助成を受けたものです．

## 文　献

[1] R. Schendel and A. Tolmie, “Assessment techniques and students’ higher-order thinking skills,” Assessment & Evaluation in Higher Education, vol.42, no.5, pp.673–689, 2017.

[2] Y. Abosalem, “Beyond translation: Adapting a performance-task-based assessment of critical thinking ability for use in rwanda,” Int. J. Secondary Education, vol.4, no.1, pp.1–11, 2016.

[3] Y. Rosen and M. Tager, “Making student thinking visible through a concept map in computer-based assessment of critical thinking,” J. Educational Computing Research, vol.50, no.2, pp.249–270, 2014.

[4] O.L. Liu, L. Frankel, and K.C. Roohr, “Assessing critical thinking in higher education: Current state and directions for next-generation assessment,” ETS Research Report Series, vol.2014, no.1, pp.1–23, 2014.

[5] H.J. Bernardin, S. Thomason, M.R. Buckley, and J.S. Kane, “Rater rating-level bias and accuracy in performance appraisals: The impact of rater personality, performance management competence, and rater accountability,” Human Resource Management, vol.55, no.2, pp.321–340, 2016.

[6] 宇都雅輝，植野真臣，“パフォーマンス評価のため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.

[7] M. Uto and M. Ueno, “Item response theory for peer assessment,” IEEE Trans. Learning Technologies, vol.9, no.2, pp.157–170, 2016.

[8] N.L.A. Kassim, “Judging behaviour and rater errors: An application of the many-facet Rasch model,” GEMA Online Journal of Language Studies, vol.11, no.3, pp.179–197, 2011.
[9] C.M. Myford and E.W. Wolfe, “Detecting and measuring rater effects using many-facet Rasch measurement: Part I,” J. Appl. Meas., vol.4, pp.386–422, 2003.
[10] T. Eckes, “Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis,” Language Assessment Quarterly, vol.2, no.3, pp.197–221, 2005.
[11] 宇都雅輝，植野真臣，“ピアアセスメントにおける異質評価者に頑健な項目反応理論，” 信学論（D），vol.J101-D, no.1, pp.211–224, Jan. 2018.
[12] T. Eckes, Introduction to Many-Facet Rasch Measurement: Analyzing and Evaluating Rater-Mediated Assessments, Peter Lang Pub., 2015.
[13] 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：MCMC アルゴリズムに基づく推定，” 教育心理学研究，vol.58, no.2, pp.163–175, 2010.
[14] 宇佐美慧，“論述式テストの運用における測定論的問題とその対処，” 日本テスト学会誌，vol.9, no.1, pp.145–164, 2013.
[15] E. Muraki, C.M. Hombo, and Y.W. Lee, “Equating and linking of performance assessments,” Applied Psychological Measurement, vol.24, pp.325–337, 2000.
[16] G. Engelhard, “Constructing rater and task banks for performance assessments,” J. Outcome Measurement, vol.1, no.1, pp.19–33, 1997.
[17] M. Ilhan, “A comparison of the results of many-facet Rasch analyses based on crossed and judge pair designs,” Educational Sciences: Theory and Practice, pp.579–601, 2016.
[18] J.M. Linacre, “A user’s guide to FACETS Rasch-model computer programs,” 2014.
[19] 泉　毅，山野井真児，山田剛史，金森保智，対馬英樹，“共通項目数が等化の精度に及ぼす影響：大規模学力テストデータを用いた探索的研究，” 教育実践学論集，vol.13, pp.49–57, 2012.
[20] W.D. Way, “Protecting the integrity of computerized testing item pools,” Educational Measurement: Issues and Practice, vol.17, no.4, pp.17–27, 1998.
[21] T. Ishii and M. Ueno, “Clique algorithm to minimize item exposure for uniform test forms assembly,” Proc. 17th International Conference on Artificial Intelligence in Education, pp.638–641, 2015.
[22] 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，vol.20, pp.34–47, 1998.
[23] 光永悠彦，前川眞一，“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，vol.8, no.1, pp.31–48, 2012.
[24] S. Kilmen and N. Demirtasli, “Comparison of test equating methods based on item response theory according to the sample size and ability distribution,” Social and Behavioral Sciences, vol.46, no.Supplement C, pp.130–134, 2012.
[25] M.J. Kolen and R.L. Brennan, Test Equating, Scaling, and Linking, Springer Verlag, 2014.
[26] 藤森　進，“共通項目の部分得点モデル化によるテストの等化，” 人間科学研究，vol.27, pp.77–81, 2005.
[27] S. Uysal and I. Kilmen, “Comparison of item response theory test equating methods for mixed format tests,” International Online Journal of Educational Sciences, vol.8, no.2, pp.1–11, 2016.
[28] F.M. Lord, Applications of item response theory to practical testing problems, Erlbaum Associates, 1980.
[29] 独立行政法人情報処理推進機構，“IT パスポート試験，” https://www3.jitec.ipa.go.jp/JitesCbt/.
[30] 公益社団法人医療系大学間共用試験実施評価機構，“臨床実習開始前の「共用試験」第 14 版（平成 28 年度），” http://www.cato.umin.jp/e-book/14/index.html.
[31] S.P. Reise and D.A. Revicki, Handbook of Item Response Theory Modeling: Applications to Typical Performance Assessment (Multivariate Applications Series), Routledge, 2014.
[32] F. Samejima, “Estimation of latent ability using a response pattern of graded scores,” Psychometrika Monography, vol.17, pp.1–100, 1969.
[33] E. Muraki, “A generalized partial credit model,” in Handbook of Modern Item Response Theory, ed. W.J. van derLinden and R.K. Hambleton, pp.153–164, Springer, 1997.
[34] J.M. Linacre, Many-faceted Rasch Measurement, MESA Press, 1989.
[35] 野口裕之，大隅敦子，テスティングの基礎理論，研究社，2014.
[36] M. Uto, N. Duc Thien, and M. Ueno, “Group optimization to maximize peer assessment accuracy using item response theory,” Proc. International Conference on Artificial Intelligence in Education, pp.393–405, 2017.
[37] S. Arai and S. Mayekawa, “A comparison of equating methods and linking designs for developing an item pool under item response theory,” Behaviormetrika, vol.38, pp.1–16, 2011.
[38] W.C. Lee and J.C. Ban, “A comparison of IRT linking procedures,” Applied Measurement in Education, vol.23, no.1, pp.23–48, 2009.
[39] 松下佳代，小野和宏，高橋雄介，“レポート評価におけるルーブリックの開発とその信頼性の検討，” 大学教育学会誌，vol.35, no.1, pp.107–115, 2013.

（平成 29 年 9 月 29 日受付，30 年 1 月 10 日再受付，
3 月 6 日早期公開）

**宇都　雅輝**　（正員）

2013 年電気通信大学大学院情報システム学研究科博士後期課程修了．博士（工学）．長岡技術科学大学を経て，2015 年より電気通信大学助教に着任，現在に至る．e テスティング，e ラーニング，人工知能，ベイズ統計，自然言語処理などの研究に従事．