電気通信大学情報理工学域
I 類 (情報系) 情報数理工学プログラム卒業論文

項目難易度制約付き等質適応型テスト

2023 年 1 月 30 日

情報数理工学プログラム

学籍番号 1910200

岸田 若葉

指導教員 植野 真臣

令和４年度 情報数理工学プログラム卒業論文概要

平成 31 年度 入学	学籍番号 1910200
指導教員 植野 真臣	氏名 岸田 若葉
題目 項目難易度制約付き等質適応型テスト	

概要

本論文では，能力推定値近傍と対応した難易度をもつ項目に限定して出題する項目難易度制約付き等質適応型テストを提案する．適応型テストとは，受検者の能力値を逐次的に推定し，その能力値に応じた項目を出題するテスト形式である．これにより，適応型テストは従来の試験よりも受検者の能力値を少ない項目数で高精度に測定できる．しかし，受検者の能力値は一般に正規分布に従うため，能力分布の平均値の識別に適した項目が過度に暴露されやすい．過度に暴露された項目は受検対策される恐れがあり，テストの信頼性の低下につながる．この問題を解決するため，暴露数を制御する手法が多数提案されているが，暴露数の減少と測定精度の向上にはトレードオフの関係がある．測定精度を保ちつつ暴露数を制御する手法として，宮澤ら (2023) は 2 段階等質適応型テストを提案した．1 段階目は異なる項目から構成されるが測定精度が等質な項目集合を 1 つ受検者に割り当て項目選択を行い，テスト後半ではアイテムバンク全体から項目選択する．宮澤ら (2023) は 1 段階目で能力推定値が真値近傍に収束するため，2 段階目ではその能力推定値近傍の難易度をもつ項目のみが選択され，暴露数の偏りが軽減できると仮定した．しかし，実際は能力真値と難易度が乖離していても識別力の高い項目であれば選択される傾向があるため，暴露数の偏りの改善は限定的である．この問題を解決するため，本研究では，1 段階目で受検者の能力値の近傍を推定し，2 段階目で受検者の能力値の近傍にある難易度をもつ項目に限定して出題する項目難易度制約付き等質適応型テストを提案する．これにより，識別力の高い項目に偏って暴露される問題を緩和できる．本論文では提案手法の有効性をシミュレーションデータと実データを用いて示した．その結果，提案手法は従来手法と測定精度を同等に保ちつつ，従来手法よりも暴露数の標準偏差および最大値を減少できた．特に，提案手法は 2 段階等質適応型テストにおいて頻繁に暴露された識別力の高い項目の暴露数を削減できた．

1　まえがき

e テスティング [1, 2] の一種として，適応型テスト (CAT: Computerized Adaptive Testing) というテスト形式がある．適応型テストとは受検者の能力値を逐次的に推定し，その能力値に応じた項目を出題するテスト形式である．これにより，適応型テストは従来の試験よりも受検者の能力値を少ない項目数で高精度に測定できる．しかし，受検者の能力値は一般に正規分布に従うため，能力分布の平均値の識別に適した項目が過度に暴露されやすい．過度に暴露された項目は受検対策される恐れがあり，テストの信頼性の低下につながる [3]．特定項目の暴露は実際に適応型テストが導入されている Synthetic Personality Inventory(SPI) [4] や Global Test of English Communication(GTEC) [5] においても問題となっている．

この問題を解決するために，暴露数を制御する様々な手法が提案されている [6–12]．最も有名な手法として，van der Linden らは暴露数の上限を制約としたシャドーテストを逐次的に構成し，そのテストから項目選択を行う手法を提案した [9]．ここで，シャドーテスト [9] とは，(1) 試験の制約 (例えば，テストの長さや暴露数の最大値) を全て満たし，(2) 既に受検者に出題した項目を全て含め，(3) 推定能力値に対して項目反応理論におけるテスト情報量が最大となる項目集合である．他にも，Sympson-Hetter 法では，事前にシミュレーション実験を用いて項目の出題確率とその出題確率を制御するパラメータを算出し，そのパラメータを用いて確率的に項目の出題を制御する [10–12]．さらに，van der Linden らは，事前にシミュレーション実験を必要としない手法を提案している [6,13,14]．この手法では，項目の出題可能な確率を意味する適格確率 (Eligibility probability) を定義し，受検者ごとに適格確率に従ってアイテムバンクから項目を除くことで暴露数を制御する．他には，分割したアイテムバンクを用いる手法が提案されている [15]．しかし，これらの手法は暴露数の制御に伴い，能力値の推定精度が低下するというトレードオフの問題がある．

このトレードオフを解決するために，宮澤らは 2 段階等質適応型テストを提案した [16, 17]．2 段階等質適応型テストでは，事前に等質テスト構成手法により生成した等質な項目集合 (以降，等質アイテムバンクと呼ぶ．) を用いる．等質テストとは，異なる項目から構成されるが測定精度が等質な項目集合である．等質テスト構成は多数研究されており，近年では大規模な数の等質テストを生成できる手法が提案されている [18–21]．情報処理技術者試験や医療系教養試験などの実際のテスト運営でも実用化が検討されている [22, 23]．等質アイテムバンクの生成には，当時最大数の等質テストを構成できた石

井らの手法 [19] を用いる．1 段階目では等質アイテムバンクを各受検者に 1 つ割り当て，項目選択を行う．これにより，受検者ごとに異なる項目が出題されるため，暴露数が減少する．しかし，出題数が増加するにしたがって，能力値の識別に適した項目が不足する．そこで，2 段階目ではアイテムバンク全体から項目選択を行う．これにより，能力推定値が収束しはじめたテスト後半に，能力値の識別に適した項目を十分に確保することができる．結果として，この手法は推定精度を保ちつつ暴露数を減少できた．宮澤ら [16, 17] は，1 段階目で推定値が能力真値の近傍に収束するため，第 2 段階目ではその能力推定値近傍の難易度をもつ項目のみが選択され，暴露数の偏りは軽減できると仮定している．しかし，実際は能力真値と難易度が乖離していても識別力の高い項目であれば選択される傾向にある．そのため，暴露数の偏りの改善は限定的である．

この問題を解決するため，本研究では，2 段階目以降は能力推定値近傍の難易度パラメータをもつ項目に限定して出題する項目難易度制約付き等質適応型テストを提案する．1 段階目では，2 段階等質適応型テストと同様に等質アイテムバンクから項目選択を行う．等質アイテムバンクの生成には現在最も多くの等質テストが構成可能な Fuchimoto らの手法 [21] を用いる．出題項目数の増加にともない，真の能力値近傍に到達する．能力推定値の更新幅があらかじめ設定した閾値よりも小さくなったとき，1 段階目を終了する．それ以降は，アイテムバンク全体のうち能力推定値近傍に対応した難易度パラメータ区間に限定して項目選択する．この難易度パラメータ区間は能力推定値の事後標準偏差に応じ決定する．以降，出題する度に更新した難易度パラメータ区間に限定した項目選択を繰り返す．この手法では，1 段階目で受検者の能力値の近傍を推定し，2 段階目で受検者の能力値の近傍にある難易度をもつ項目に限定して出題する．前述のように，従来手法では受検者の能力値と項目難易度が乖離していても識別力の高い項目は選択される傾向にあり，暴露数の偏りの原因と考えられる．提案手法ではこの問題を緩和できると期待できる．

本論文では，提案手法の有効性をシミュレーションデータと実データを用いて示した．その結果，提案手法は従来手法と測定精度を同等に保ちつつ，従来手法よりも暴露数の標準偏差および最大値を減少できた．特に，提案手法は 2 段階等質適応型テストにおいて頻繁に暴露されていた識別力パラメータが大きい項目の暴露数を大幅に削減できた．

2　項目反応理論による適応型テスト

2.1　項目反応理論

項目反応理論 (Item Response Theory: IRT) は，数理モデルを用いたテスト理論のひとつであり，近年，コンピュータ・テスティングの普及とともに多様な評価場面で活用されている [24–28]．IRT は受検者の項目への正答確率をモデル化したものである．これにより，異なる項目への受検者の反応を同一尺度上で評価できる．

項目反応理論のモデルの 1 つである 3 母数ロジスティックモデル (3-Parameter Logistic Model: 3PLM) では，多肢選択式の問題において能力値の低い受検者が偶然正答する可能性等を考慮して，能力値 $\theta \in (-\infty, \infty)$ の受検者が項目 $i \in \{1, ..., N\}$ に正答する確率を次の式で示す．

$$p(u_i = 1|\theta) = c_i + \frac{1 - c_i}{1 + \exp(-1.7a_i(\theta - b_i))} \tag{1}$$

ここで，u_i は受検者が項目 i に正答するとき 1，それ以外のときは 0 である変数である．また，$a_i \in [0, \infty]$，$b_i \in [0, \infty]$, $c_i \in [0, 1]$ はそれぞれ i 番目の項目の識別力パラメータ，難易度パラメータ，当て推量パラメータである．受検者が偶然正答する可能性等を考慮しない場合 ($c_i = 0$)，2 母数ロジスティックモデル (2-Parameter Logistic Model: 2PLM) と呼ぶ．

他には，多肢選択式の問題の各選択肢への反応データを扱う一般化部分採点モデル (Generalized Partial Credit Model:GPCM) がある．GPCM では正答確率を次の式で示す．

$$p(u_i = 1|\theta) = \frac{\exp(\sum_{m=1}^{k} \alpha_i(\theta - \beta_{im}))}{\sum_{l=1}^{K}[\exp(\sum_{m=1}^{l} \alpha_i(\theta - \beta_{im}))]} \tag{2}$$

ここで，β_{im} は項目 i のカテゴリ $m - 1$ からカテゴリ m に遷移する難易度を示す．

2.2　フィッシャー情報量

項目反応理論において，能力推定値の標準誤差はフィッシャー情報量の逆数の値に漸近的に一致する [24]．能力値 θ をもつ受検者に対して項目 i が与えるフィッシャー情報量を以下の式で定義する．

$$I_i(\theta) = \frac{[\frac{\partial}{\partial\theta} p(u_i = 1|\theta)]^2}{p(u_i = 1|\theta)(1 - p(u_i = 1|\theta))} \tag{3}$$

フィッシャー情報量 $I_i(\theta)$ の高い項目は能力値 θ 付近でその能力値をよく識別する．従来の適応型テストでは受検者の能力値を逐次的に推定し，フィッシャー情報量の高い項目を受検者に出題することで，効率の良い能力推定を実現する．本論文では情報量はフィッシャー情報量を指す．なお，テスト T を構成する項目集合の情報量の総和をテスト情報量とよび，次のように示す．

$$I_T(\theta) = \sum_{i \in T} I_i(\theta) \tag{4}$$

テスト情報量の逆数が受検者の能力推定値の漸近分散に収束する [29].

2.3 能力値 θ の推定

受検者の能力値の推定には EAP 推定 (Expected a posteriori) を用いる [30]．$m-1$ 項目までの反応データのベクトル $\mathbf{u}_{m-1}$ と能力値 θ の事前分布 $f(\theta)$ を用いて，能力値の事後分布 $f(\theta|\mathbf{u}_{m-1})$ は次のように表される．

$$f(\theta|\mathbf{u}_{m-1}) = \frac{L(\mathbf{u}_{m-1}|\theta)f(\theta)}{\int_{-\infty}^{\infty} L(\mathbf{u}_{m-1}|\theta)f(\theta)d\theta} \tag{5}$$

ここで，$L(\mathbf{u}_{m-1}|\theta)$ は能力値を所与としたときに反応 $\mathbf{u}_{m-1}$ を返す尤度である．EAP 推定は事後分布 $f(\theta|\mathbf{u}_{m-1})$ の θ に関する期待値を能力推定値とする．EAP 推定を用いて能力推定値 $\hat{\theta}$ は次式より求められる．

$$\hat{\theta} = \int_{\infty}^{\infty} \theta f(\theta|\mathbf{u}_{m-1})d\theta \tag{6}$$

2.4 適応型テスト

項目反応理論を用いた適応型テストでは，項目パラメータが既知のアイテムバンクを所与として，次のアルゴリズムにしたがって項目選択する．

1. 能力推定値を $\hat{\theta} = 0$ に初期化する．
2. 能力推定値 $\hat{\theta}$ を所与として情報量が最大となる項目 i をアイテムバンクから選択し受検者に出題する．
3. 回答履歴から受検者の能力推定値 $\hat{\theta}$ を推定する．
4. テスト終了条件まで手順（2)，(3）を繰り返す．

この手順に従って，適応型テストは受検者の能力を逐次的に推定し，情報量の高い項目を受検者に出題することで少ない項目数で高精度な能力推定を実現する．しかし，受検者の

能力値は一般に標準正規分布に従うため，受検者の能力分布の平均である $\theta = 0$ において情報量の高い項目が過度に暴露されやすい．これらの項目は受検対策される恐れがあり，テストの信頼性の低下につながる [3]．そのため，適応型テストでは各項目を一様に出題することが望ましい．

3　暴露数を制御する適応型テスト

本章では，従来の適応型テストの問題を解決するために提案された暴露数を制御する手法を紹介する．

3.1　整数計画問題 (Integer Programming Problem) に基づく適応型テスト (IP)

van der Linden らは，各項目の暴露数に最大暴露数 R を制約としたシャドーテストを逐次構成し，その中から項目選択する手法を提案した (以降，IP と呼ぶ．) [9]．具体的には，次のアルゴリズムにしたがって項目選択する．

1. 受検者の能力推定値を $\hat{\theta} = 0$ に初期化する．
2. 次の整数計画問題を用いてシャドーテストを構成する．

   maximize

$$\sum_{i=1}^{N} I_i(\hat{\theta}) x_i \tag{7}$$

   subject to

$$r_i x_i \leq R (i = 1, \cdots, N) \tag{8}$$

   (項目 i の暴露数 r_i，最大暴露数 R)

$$\sum_{i=1}^{N} x_i = n (\text{テスト項目数}) \tag{9}$$

$$x_i = \begin{cases} 1 & \text{項目 } i \text{ がシャドーテストに含まれる} \\ 0 & \text{それ以外} \end{cases}$$

3. シャドーテストから情報量が最大の項目を受検者に出題する．
4. 受検者の能力値 $\hat{\theta}$ を推定する．
5. テスト終了条件まで手順（2）から（4）を繰り返す．

3.2　整数計画問題の制約にターゲット情報量を用いた適応型テスト (TI)

整数計画問題を用いた他の手法として，Choi and Lim は制約にターゲット情報量（Target Information）を用いた手法を提案した（以降，TI と呼ぶ）[31]．具体的には，次

のアルゴリズムにしたがって項目選択する．

1. 受検者の能力推定値を $\hat{\theta} = 0$ に初期化する．
2. 次の整数計画問題を用いてシャドーテストを構成する．ただし，T はターゲット情報量である．
   minimize y
   subject to

$$\sum_{i=1}^{N} I_i(\hat{\theta}) x_i \leq T + y \tag{10}$$

$$\sum_{i-1}^{N} I_i(\hat{\theta}) x_i \geq T - y \tag{11}$$

$$y \geq 0 \tag{12}$$

$$\sum_{i=1}^{N} x_i = n(\text{テスト項目数}) \tag{13}$$

3. シャドーテストから情報量が最大の項目を受検者に出題する．
4. 受検者の能力値 $\hat{\theta}$ を推定する．
5. テスト終了条件まで手順（2）から（4）を繰り返す．

ただし，従来の適応型テストを実施したときのテスト情報量の平均をターゲット情報量 T として用いる．このように，予め設定したターゲット情報量とシャドーテストのテスト情報量の差が最小となるようにすることで，識別力パラメータが過度に大きい項目の出題を防止し，受検者の能力推定に適した識別力をもつ項目を出題できる．

3.3 確率的に項目選択を制御する適応型テスト (Prob)

別のアプローチとして，確率的に項目選択を制御する手法が提案されている [6, 10–14]．van der Linden と Veldkamp は適格確率 (Eligibility probabiliry) を用いた手法を提案している（以降，Prob と呼ぶ）[6, 13, 14]．具体的には，次のアルゴリズムにしたがって項目選択する．

1. 受検者の能力推定値を $\hat{\theta} = 0$ に初期化する．
2. 受検者 j に対する項目 i の適格確率 $P^{(j)}(E_i)$ を計算する．

$$P^{(j)}(E_i) = \min\{\frac{r^{max}}{P^{(j-1)}(A_i)} P^{(j-1)}(E_i), 1\} \tag{14}$$

ここで，r^{max} は暴露率の上限値，$P^{(j-1)}(A_i)$ は受検者 $j-1$ までの項目 i の暴露率である．

3. 適格確率 $P^{(j)}(E_i)$ にしたがって，適格でない項目のみアイテムバンクから除外する．
4. アイテムバンクから情報量が最大の項目を選択する．
5. 受検者の能力値 $\hat{\theta}$ を推定する．
6. テスト終了条件まで手順（3），(4）を繰り返す．

ただし，最初の受検者 $(j=1)$ または $P^{(j-1)}(A_i)=0$ の場合には $P^{(j)}(E_i)=1$ とする．テスト終了後には全ての項目をアイテムバンクに戻す．

3.4 Kingsbury and Zara(1989) の適応型テスト (KZ)

Kingsbury and Zara は暴露数の偏りを軽減させるために分割したアイテムバンクを使用する手法を提案した（以降，KZ と呼ぶ）[15]．具体的には，次のアルゴリズムにしたがって項目選択する．

1. アイテムバンクをランダムに分割し，項目集合を複数構成する．
2. 受検者の能力値を $\hat{\theta}=0$ に初期化する．
3. 暴露数が最小の項目集合を選択し，その項目集合から情報量が最大の項目を受検者に出題する．
4. 受検者の能力値 $\hat{\theta}$ を推定する．
5. テスト終了条件まで手順（3)，(4）を繰り返す．

KZ は，特定の項目群の過度な暴露を防ぐことができた．しかし，各項目集合は測定精度が等質でないため，受検者間でテストの長さや測定誤差に偏りが生じる．さらに，これらの手法は出題する項目を制限するため，測定精度が低下する．

3.5 2 段階等質適応型テスト

暴露数の減少と測定精度の向上の両立のため，Ueno and Miyazawa は 2 段階等質適応型テストを提案している [16, 17]（以降，TUAT と呼ぶ．）．TUAT では，事前に等質テスト構成手法により生成した等質アイテムバンクを用いる．等質アイテムバンクの生成には，当時世界最大数の等質テストを構成できる Ishii et al.(2017) の手法 [19] を用いる．1 段階目では，各受検者に等質アイテムバンクをランダムに 1 つ割り当て，項目選択を

行う．これにより，受検者ごとに異なる項目が出題されるため，暴露数が減少する．しかし，出題数が増加するにしたがって，能力推定値の識別に適した項目が不足する．そこで，2 段階目ではアイテムバンク全体から項目選択を行う．これにより，能力推定値が収束しはじめたテスト後半に，能力推定値の識別に適した項目を十分に確保できる．この結果，従来の適応型テストと同等の測定精度を保ちつつ，従来手法よりも暴露数を減少できた．宮澤ら [16,17] は，1 段階目で推定値が能力真値の近傍に収束するため，第 2 段階目ではその能力推定値近傍の難易度をもつ項目のみが選択され，暴露数の偏りは軽減できると仮定している．しかし，実際は能力値と難易度が乖離していても識別力の高い項目であれば選択される傾向にある．図 1 は，TUAT の 2 段階目において出題された項目の難易度と真値の RMSE である．能力真値を-3.0 から 3.0 まで 0.1 刻みで離散化し，各区間の RMSE の平均をプロットする．図 1 より，真値が能力分布の平均値 $\theta = 0$ から離れれば離れるほど難易度と真値の RMSE が大きくなる．一般にアイテムバンクは能力分布の平均値 $\theta = 0$ をよく識別できるよう設計される．そのため，能力分布の平均値から離れた能力値に対しても，難易度が $b_i = 0$ 付近で識別力の高い項目が出題された．このように識別力の高い項目に偏って暴露されるため，暴露数の偏りの改善は限定的である．

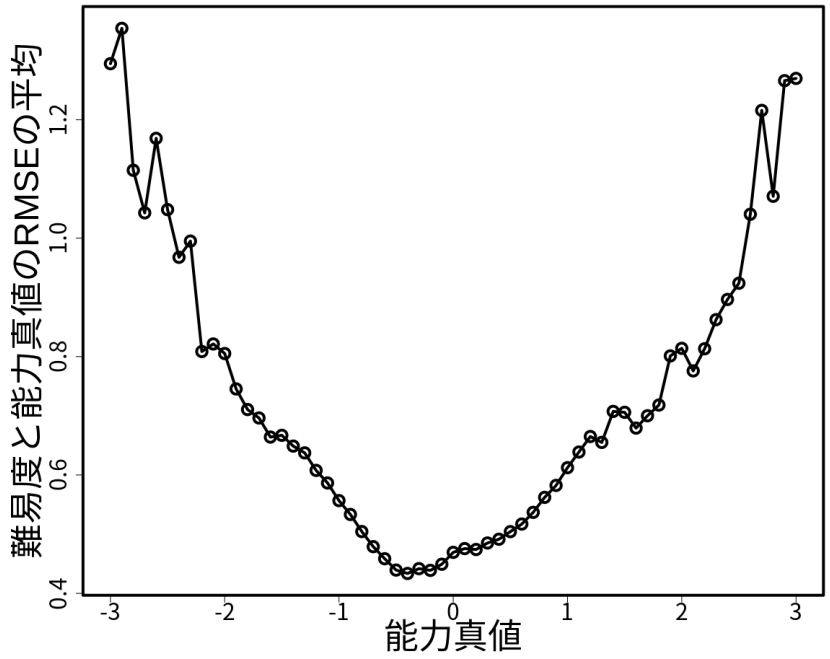


図 1: 2 段階目に出題された項目の難易度と真値の RMSE

4　提案手法

TUAT の問題を解決するため，本研究では，2 段階目以降は能力推定値近傍の難易度パラメータをもつ項目に限定して出題する項目難易度制約付き等質適応型テストを提案する．

本手法では，事前に等質テスト構成技術により生成した等質アイテムバンクを用いる．等質アイテムバンクの生成には，現在最も多くの等質テストを構成可能な Fuchimoto らの手法 [21] を用いる．この手法では，はじめに，時間計算量は小さいが空間計算量の大きい最大クリーク法 [18] により，メモリの限界まで多くの等質テストを構成する．次に，時間計算量は大きいが空間計算量が小さい整数計画法を用いた並列探索手法により，より多くの等質テストを生成する．本研究では，TUAT と同様に離散化した能力値 $\theta_k = \{\theta_1, \theta_2, ..., \theta_K\}$ について，テスト情報量の下限と上限を次のように制限してテストを構成する．

$$m(\theta_k)n \leq I_i(\theta_k)z_i \leq (m(\theta_k) + sd(\theta_k))n \tag{15}$$

ここで，z_i は項目 i がテストに含まれるとき 1，含まれないとき 0 をとる．また，$m(\theta_k)$ は能力値 θ_k についてのアイテムバンクに含まれる項目の情報量の平均，$sd(\theta_k)$ は θ_k ついての標準偏差，n は等質アイテムバンクの項目数である．

本手法では，まず，TUAT と同様に，等質アイテムバンクを受検者にランダムで 1 つ割り当て，項目選択を行う．出題項目数の増加にともない，能力推定値が真値近傍に到達する．能力推定値の更新幅が閾値よりも小さくなったとき，その能力推定値近傍に対応した難易度パラメータ区間を満たす項目から選択する．以降，項目を出題し能力値を推定するごとに更新した難易度パラメータ区間を満たす項目の選択を繰り返す．難易度パラメータ区間は，次式より表される能力推定値の事後標準偏差を用いて決定する．

$$SD(\hat{\theta}) = \sqrt{\int_{-\infty}^{\infty} (\theta - \hat{\theta})^2 f(\theta|\mathbf{u}_{m-1})} \tag{16}$$

事後標準偏差 $SD(\hat{\theta})$ は能力推定値 $\hat{\theta}$ の誤差の範囲を意味する．

本手法では，次の区間を能力推定値近傍とした．

$$\hat{\theta} - \alpha SD(\hat{\theta}) < \theta < \hat{\theta} + \alpha SD(\hat{\theta}) \tag{17}$$

そして，能力推定値近傍と対応する難易度パラメータ区間は次の通りである．

$$\hat{\theta} - \alpha SD(\hat{\theta}) < b < \hat{\theta} + \alpha SD(\hat{\theta}) \tag{18}$$

ただし，α は難易度パラメータ区間に対する事後標準偏差の影響度合いを決定する重み付けチューニングパラメータである．

具体的には，次のアルゴリズムにしたがって項目選択する．

1. 等質アイテムバンクのうちの 1 つを受検者にランダムに割り当てる．
2. 能力推定値 $\hat{\theta}$ を $\hat{\theta} = 0$ に初期化する．
3. 割り当てられた等質アイテムバンクから，能力推定値に対して情報量が最大の項目を出題する．
4. 受検者の能力推定値 $\hat{\theta}$ を推定する．
5. 能力推定値 $\hat{\theta}$ の更新幅が閾値 ϵ 未満になるまで手順（3），（4）を繰り返す．

能力推定値 $\hat{\theta}$ の更新幅が閾値 ϵ 未満になった後，出題アルゴリズムを次のように変更する．

1. 能力推定値 $\hat{\theta}$ と事後標準偏差 $SD(\hat{\theta})$ から難易度パラメータ区間を更新する．
2. 難易度パラメータ区間を満たす項目から，能力推定値に対して情報量が最大となる項目を出題する．
3. 受検者の能力推定値 $\hat{\theta}$ を推定する．
4. テスト終了条件まで手順（1），（2）を繰り返す．

本手法により，受検者の能力値の推定精度を保ちつつ，暴露数の標準偏差および最大値を減少させる．

5　評価実験

提案手法の有効性を示すため，従来手法と比較する．

5.1　パラメータチューニング

本節では提案手法の切替条件および等質アイテムバンクに含まれる項目数，難易度パラメータ区間を決定するパラメータ α を最適化する．提案手法は切替条件および等質アイテムバンクに含まれる項目数，能力推定値近傍と対応した難易度パラメータ区間を決定する α を適切に設定することで，測定精度を保ちつつ暴露数の偏りを軽減できる．

まず，切替条件を検討する．提案手法について，切替条件となる能力推定値の更新幅を 0.025 から 0.5 までステップ数を 0.025 として実験した結果を図 2 に示す．図 2 は横軸が切替条件となる能力推定値の更新幅，縦軸が暴露数の標準偏差および RMSE を示す．図 2 より，切替条件となる能力推定値の更新幅が大きくすると，暴露数の標準偏差が大きくなり，RMSE は小さくなる．このことから，暴露数の標準偏差と RMSE にはトレードオフの関係があることがわかる．これは TUAT と同様である．そのため，本実験では

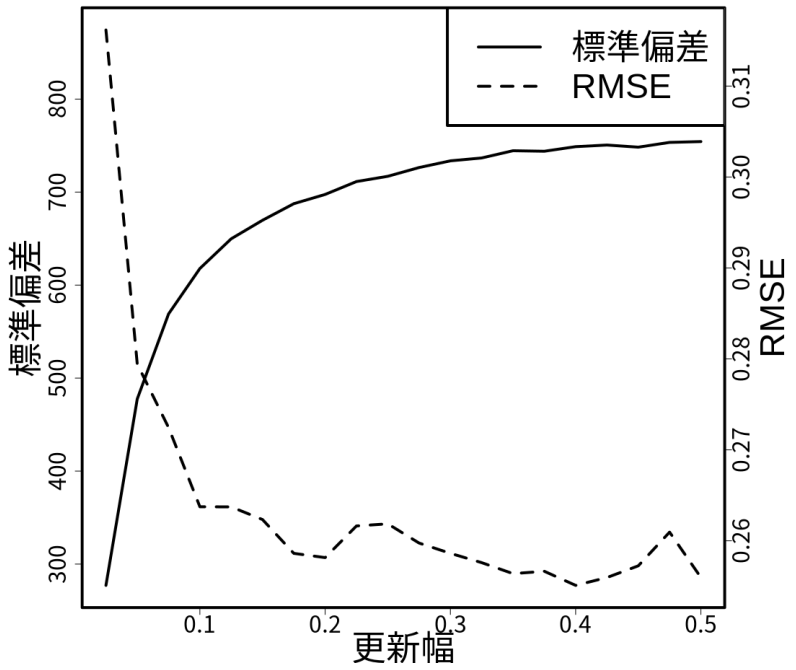


図 2: 切替条件である能力推定値の更新幅に対する暴露数の標準偏差と RMSE の変化

TUAT [16,17] と同様に，RMSE の減少幅が 0.01 以下となった更新幅を提案手法の切替条件とする．

次に，提案手法における難易度パラメータ区間を決定するパラメータ α について検討する．パラメータ α を 0.1 から 1.0 までステップ数を 0.1 として実験した結果を図 3 に示す．図 3 は横軸が α，縦軸が暴露数の標準偏差および RMSE を示す．図 3 より，パラメータ α を大きくすればするほど，暴露数の標準偏差が大きくなる．これは，難易度パラメータ区間が広くなることにより，難易度が乖離しているが識別力が高い項目が頻繁に暴露されるからである．一方，RMSE はパラメータ α を大きくすればするほど小さくなるが，大きな変化はない．そこで，本実験では RMSE の小数点第三位を四捨五入した値が同じであるパラメータ α の中で，暴露数の標準偏差が最小のものを採用する．

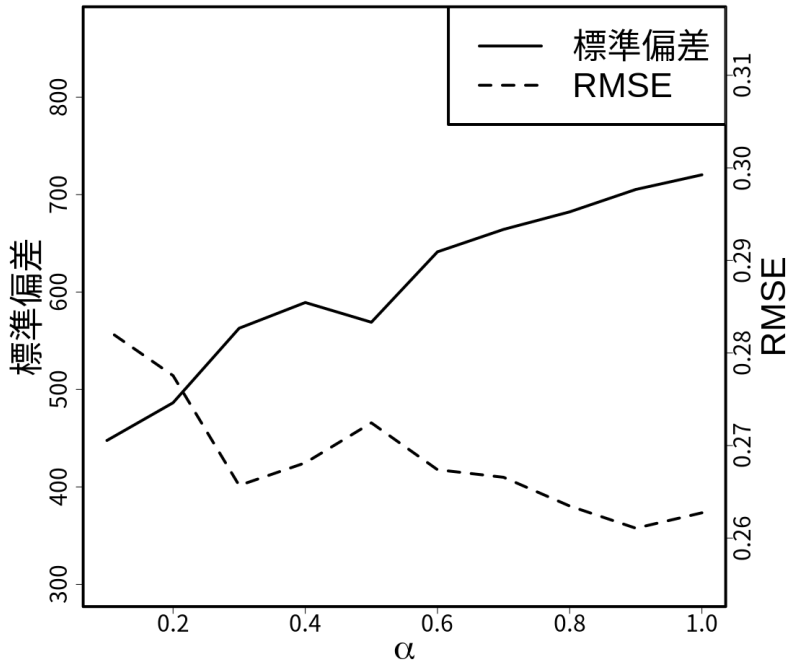


図 3: パラメータ α に対する暴露数の標準偏差と RMSE の変化

5.2 従来手法との比較実験

本実験では逐次的に能力推定値に対して情報量最大の項目を出題する適応型テスト (以降，CAT と呼ぶ．)，IP [9]，TI [31]，KZ [15]，TUAT [16,17] と比較する．実験の手順は以下の通りである．

1. 受検者の能力真値を $\theta \sim N(0,1)$ からサンプリングする．
2. 受検者の能力推定値を $\hat{\theta} = 0$ に初期化する．
3. 受検者の能力推定値を所与として，各手法のアルゴリズムにしたがい，出題項目を選択する．
4. 能力真値と出題項目のパラメータを所与として，出題項目への反応データを生成する．
5. 出題項目への反応データを所与として，受検者の能力推定値 $\hat{\theta}$ を求める．
6. テスト終了条件を満たすまで手順（4）から（6）を繰り返す．
7. 手順（2）から（7）を 10000 回繰り返す．生成された出題項目と反応データから，各項目の暴露数と能力真値と能力推定値の MSE を求めた．

実験には 2 つのシミュレーションアイテムバンクと，2 つの実データを用いた．シミュレーションアイテムバンクは，2PLM を用いて，次の分布から各項目のパラメータをサンプリングし，1000 項目からなるアイテムバンクを生成した．

- $\log a_i \sim N(-0.5, 0.2), b_i \sim N(0,1)$ (以降，このアイテムバンクを simu1 と呼ぶ．)
- $\log a_i \sim N(-0.75, 0.2), b_i \sim N(0,1)$ (以降，このアイテムバンクを simu2 と呼ぶ．)

実データは，リクルート（株）が開発した SPI [4] のアイテムバンクを用いた．なお，SPI のアイテムバンクは 2PLM を用いた 978 項目から構成される．

テスト終了条件は先行研究 [16,17] と同様にテストの長さとした．本実験ではテストの長さを 30 項目とした．また，KZ，TUAT および提案手法について，等質アイテムバンクに含まれる項目数はテストの長さと同一とした．

表 1 は暴露数の上限を設けない場合の実験結果である．なお，手法の横の括弧に TUAT は切替条件を，提案手法は切替条件と α の値を示した．

結果から，全てのデータセットにおいて TI の未出題項目数が最も小さいことが分かる．しかし，TI は暴露数の標準偏差が KZ，TUAT および提案手法よりも大きく，未出題項

表 1: 実験結果

アイテムバンク	手法	暴露数の標準偏差	暴露数の最大値	未出題項目数	MSE
simu1	CAT	1055.5	10000	832	**0.25**
	IP	1057.9	10000	833	**0.25**
	TI	997.6	10000	**0**	**0.25**
	KZ	918.0	6565	779	0.26
	TUAT(0.100)	864.7	6409	188	0.26
	Proposal(0.100，0.8)	**682.3**	**4520**	68	0.26
simu2	CAT	1167.6	10000	860	**0.32**
	IP	1162.9	10000	861	**0.32**
	TI	1078.9	10000	**0**	**0.32**
	KZ	1094.9	7816	829	**0.32**
	TUAT(0.100)	932.8	8104	251	0.33
	Proposal(0.100，0.7)	**702.8**	**5145**	128	0.33
SPI	CAT	1150.3	10000	836	**0.25**
	IP	1149.2	10000	836	**0.25**
	TI	1052.2	10000	**4**	0.26
	KZ	1032.0	7364	792	0.26
	TUAT(0.075)	937.6	7381	274	0.26
	Proposal(0.100，0.6)	**672.7**	**5031**	263	0.27

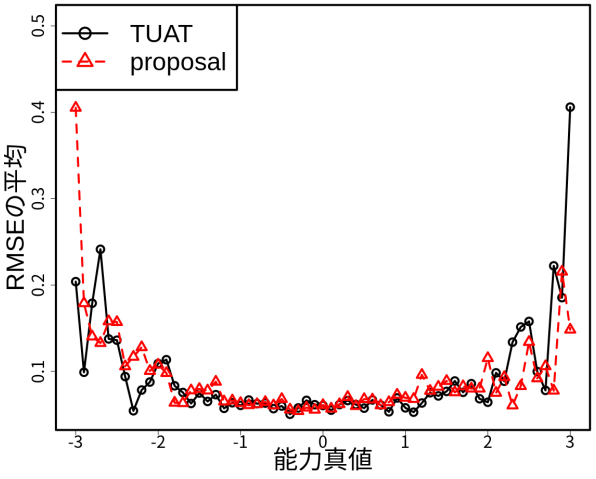
目数が他手法よりも大きい CAT，IP と近い値をとる．さらに，TI，CAT，および IP は暴露数の最大値と受検者数が一致するため，全ての受検者に出題された項目が存在する．このように，過度に暴露された項目は受検対策されやすく，テストの信頼性の低下につながる危険性がある．

一方，提案手法は全てのデータセットにおいて，暴露数の最大値および標準偏差を全ての手法の中で最小に抑えられた．特に，提案手法は全てのデータセットにおいて，暴露数の標準偏差および最大値，未出題項目数を TUAT よりも小さくできた．また，提案手法は SPI では測定精度が TUAT よりも僅かに劣るが，simu1，simu2 では測定精度を同等に保っている．

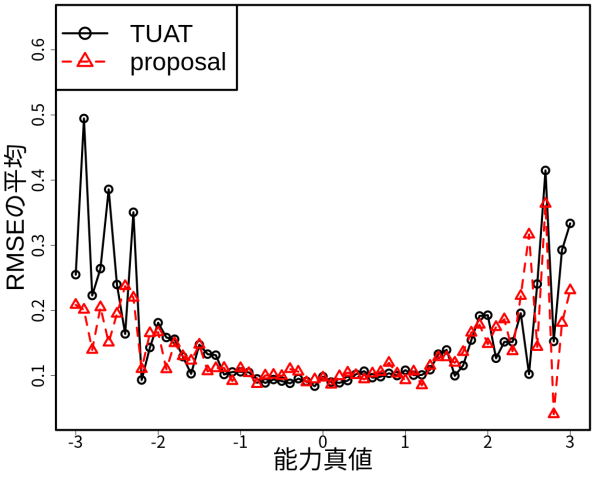
TUAT と提案手法の測定精度について分析する．図 4a，4b，4c はそれぞれ simu1，simu2，SPI の実験における能力真値と測定精度を示す RMSE の比較である．本研究では能力真値を-3.0 から 3.0 まで 0.1 刻みで離散化し，各区間の RMSE の平均をプロットする．これらの図から，TUAT と提案手法のいずれも，能力真値が受検者の能力分布の

平均値である $\theta = 0$ に近ければ近いほど RMSE が小さい．これは，アイテムバンク内に能力分布の平均値に対し高い識別力をもつ項目が多いからである．全ての実験において，提案手法の RMSE は，能力真値 $-2 < \theta < 2$ の範囲において TUAT と同等である．一方で，能力真値 θ が $|\theta| > 2$ の範囲において，提案手法の RMSE は TUAT よりもばらつきが大きくなる傾向がある．これは，この範囲で難易度パラメータ区間を満たす項目が少ないことに起因する．

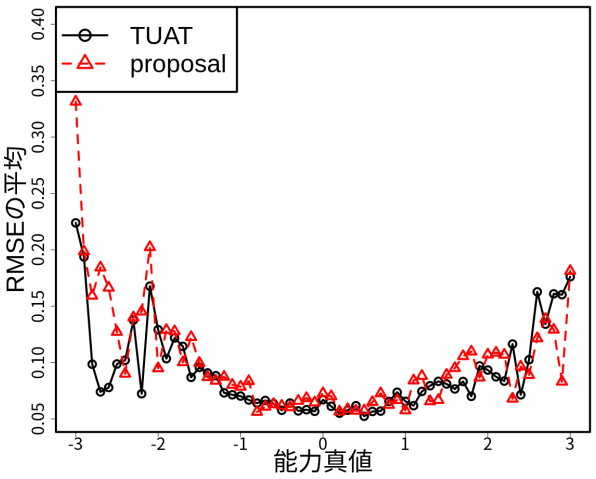
次に，TUAT と提案手法において暴露された項目を分析する．図 5a，5b，5c はそれぞれ simu1，simu2，SPI の実験における項目の識別力パラメータと暴露数の散布図である．



(a) simu1



(b) simu2

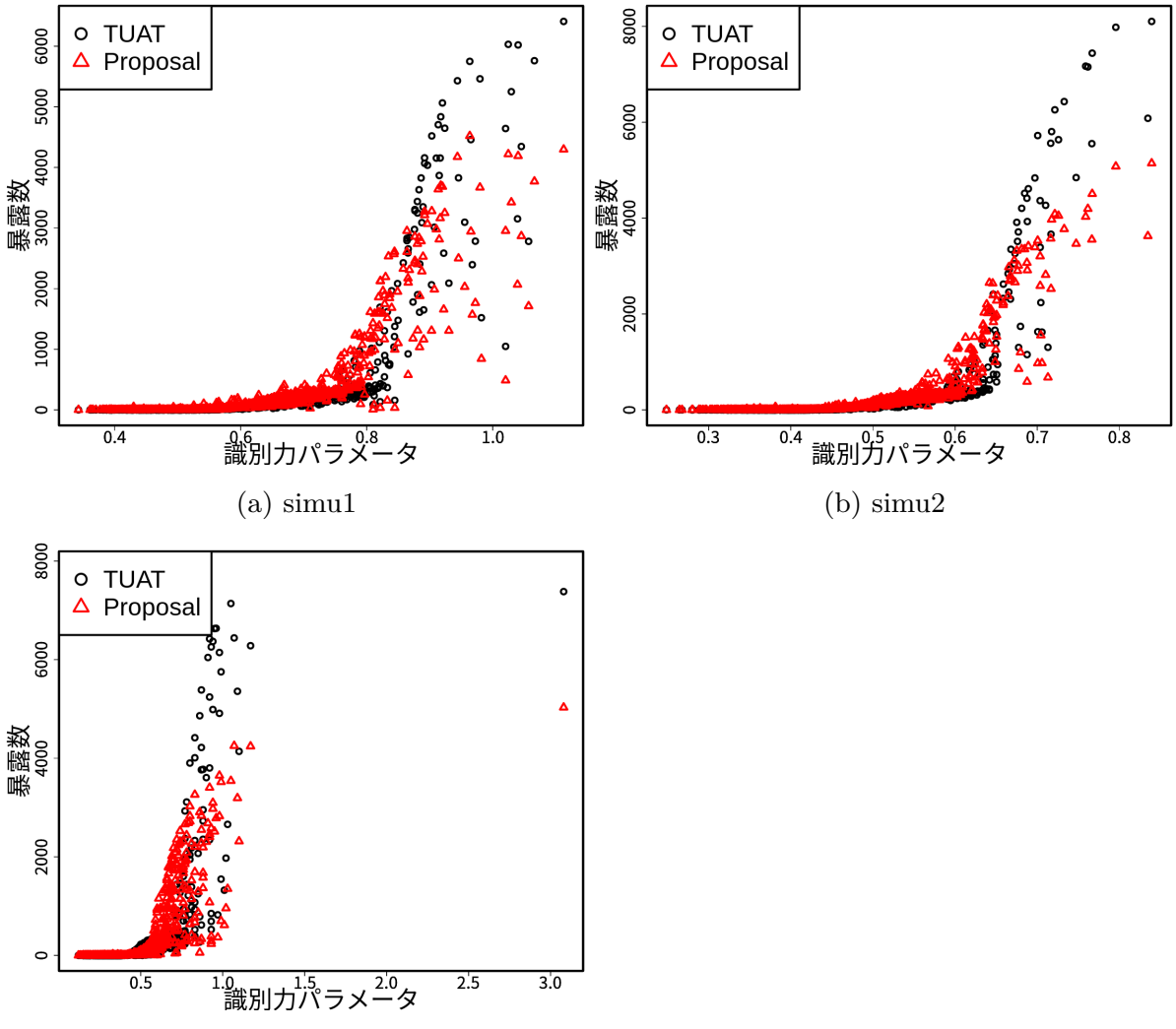


(c) SPI

図 4: 能力真値と測定精度の比較

これらの図から，TUAT は識別力パラメータが特に大きい項目を頻繁に暴露していることがわかる．一方で提案手法は，難易度パラメータ区間を満たす項目に限定して選択することで，TUAT で頻繁に暴露されていた識別力パラメータが大きい項目の暴露数を削減できた．

最後に，TUAT と提案手法において 2 段階目以降に出題された項目の難易度と真値の誤差を比較する．図 6a，6b，6c はそれぞれ 2PLM の項目のみから構成される simu1，simu2，SPI の実験における能力真値と，2 段階目以降に出題された項目の難易度と真値の RMSE の比較である．能力真値を-3.0 から 3.0 まで 0.1 刻みで離散化し，各区間の



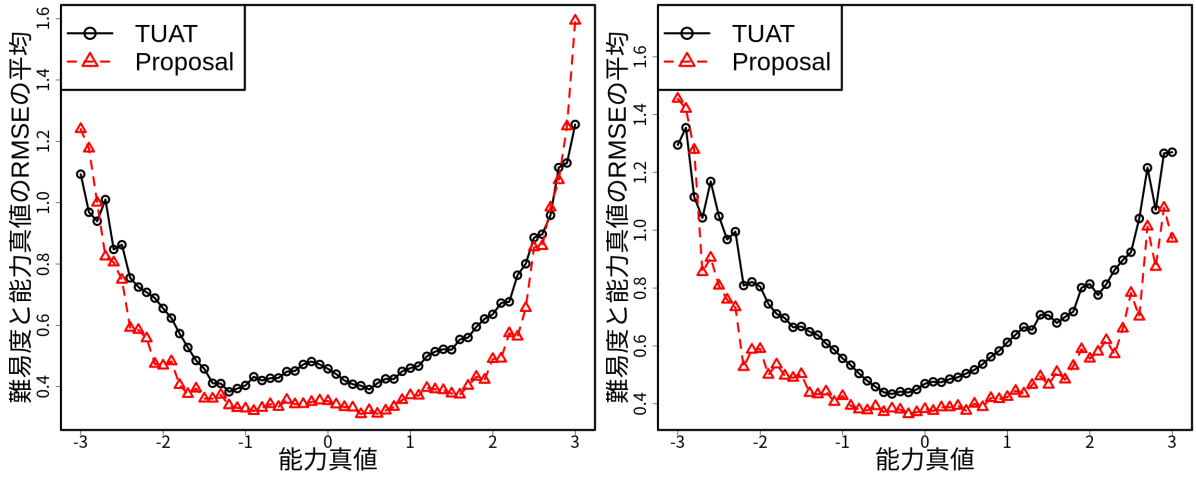
(a) simu1

(b) simu2

(c) SPI

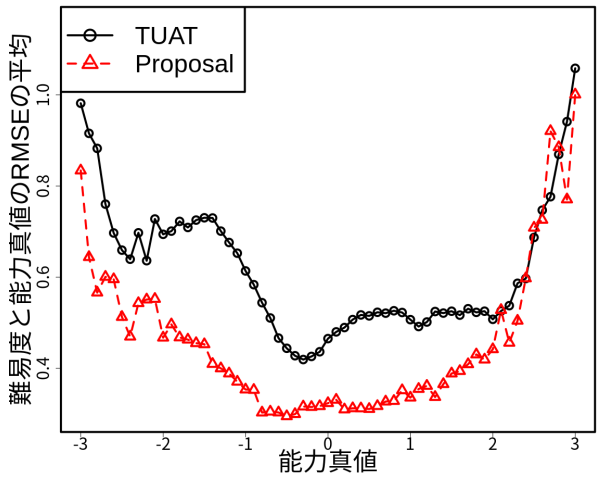
図 5: 暴露数と識別力パラメータの比較

RMSE の平均をプロットする．これらの図から，提案手法は $-2 < \theta < 2$ の範囲において，2 段階目以降に出題した項目の難易度と真値の誤差が TUAT よりも小さい．前述のように，能力真値と難易度が乖離していても識別力の高い項目であれば選択される傾向にあり，暴露数の偏りの原因と考えられている．提案手法は，1 段階目で受検者の能力真値の近傍を推定し，2 段階目で受検者の能力値の近傍にある難易度をもつ項目に限定して出題することで，この問題を緩和できた．一部のデータセットでは，$|\theta| > 2$ の範囲において，提案手法と項目の難易度の誤差が TUAT と同等またはそれよりも大きい．これは，能力真値が外れ値の場合事後標準偏差が大きいため，難易度パラメータ区間が広く，難易



(a) simu1

(b) simu2



(c) SPI

図 6: 2 段階目以降に出題された項目の難易度と真値の RMSE

度が乖離しているが識別力が高い項目が出題できたことが原因である．また，前述の通り外れ値の能力値に対して推定精度がばらつく傾向があることから，他の原因として，能力推定値が真値近傍にないため，難易度パラメータ区間が真値から離れたことが考えられる．

6　むすび

本研究では，2 段階等質適応型テストの問題を解決するため，能力推定値近傍と対応した難易度パラメータをもつ項目に限定して出題する項目難易度制約付き等質適応型テストを提案した．1 段階目では，2 段階等質適応型テストと同様に事前に生成した等質アイテムバンクを各受検者にランダムに 1 つ割り当て，項目選択を行う．等質アイテムバンクの生成には Fuchimoto et al. [21] を用いた．能力推定値が収束しはじめたとき，1 段階目を終了する．それ以降は，アイテムバンク全体のうち能力推定値近傍に対応した難易度パラメータ区間を満たす項目から選択する．この難易度パラメータ区間は能力推定値の事後標準偏差に応じ決定する．シミュレーションデータと実データを用いた実験により，提案手法は従来手法と比較して測定精度を同等に保ちつつ，暴露数の標準偏差および最大値を減少できた．特に，提案手法は 2 段階等質適応型テストにおいて頻繁に暴露されていた識別力パラメータが大きい項目の暴露数を削減できた．

本手法の問題として，未出題項目の存在があげられる．ハイ・ステークスなテストにおいてアイテムバンクを設計する際に，ひと項目あたりの開発コストが高い．出題されない項目が存在すると，その項目の開発コストが無駄になるため，全ての項目を活用することが望ましい．さらに，本手法は測定精度がアイテムバンクに含まれる項目のパラメータ分布の影響を受けやすく，特に受検者の能力分布の外れ値の測定精度にばらつきがあるという欠点がある．今後は，全ての項目を活用し，いかなるアイテムバンクにおいても高い測定精度を保ちつつ，項目ごとの暴露数がより一様に近い適応型テストの実現を目指す．

参考文献

[1] Maomi Ueno, Kazuma Fuchimoto, and Emiko Tsutsumi. e-testing from artificial intelligence approach. *Behaviormetrika*, Vol. 48, No. 2, pp. 409–424, 2021.

[2] Maomi Ueno. Ai based e-testing as a common yardstick for measuring human abilities. In *2021 18th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, pp. 1–5. IEEE, 2021.

[3] Walter D Way. Protecting the integrity of computerized testing item pools. *Educ. Meas. Issu. Pr.*, Vol. 17, No. 4, pp. 17–27, December 1998.

[4] Recruit. Synthetic Personality Inventory. `https://www.spi.recruit.co.jp/`.

[5] 株式会社ベネッセコーポレーション. Global Test of English Communication, 2014. `https://www.benesse.co.jp/gtec/`.

[6] Wim J van der Linden and Cees AW Glas. *Elements of adaptive testing*, Vol. 10. Springer, 2010.

[7] Len Swanson and Martha L Stocking. A model and heuristic for solving very large item selection problems. *Appl. Psychol. Meas.*, Vol. 17, No. 2, pp. 151–166, June 1993.

[8] Ying Cheng and Hua-Hua Chang. The maximum priority index method for severely constrained item selection in computerized adaptive testing. *Br. J. Math. Stat. Psychol.*, Vol. 62, No. Pt 2, pp. 369–383, May 2009.

[9] Wim J van der Linden and Lynda M Reese. A model for optimal constrained adaptive testing. *Appl. Psychol. Meas.*, Vol. 22, No. 3, pp. 259–270, September 1998.

[10] Rebecca D Hetter and J Bradford Sympson. *Item exposure control in CAT-ASVAB.* American Psychological Association, 1997.

[11] Martha L Stocking and Charles Lewis. Controlling item exposure conditional on ability in computerized adaptive testing. *J. Educ. Behav. Stat.*, Vol. 23, No. 1, pp. 57–75, March 1998.

[12] Martha L Stocking and Charles Lewis. Methods of controlling the exposure of items in CAT. In Wim J van der Linden and Gees A W Glas, editors, *Computerized Adaptive Testing: Theory and Practice*, pp. 163–182. Springer Netherlands, Dordrecht, 2000.

[13] Wim J van der Linden and Bernard P Veldkamp. Constraining item exposure in computerized adaptive testing with shadow tests. *J. Educ. Behav. Stat.*, Vol. 29, No. 3, pp. 273–291, September 2004.

[14] Wim J Linden and Seung W Choi. Improving item ‐ exposure control in adaptive testing. *J. Educ. Meas.*, No. jedm.12254, September 2019.

[15] G Gage Kingsbury and Anthony R Zara. Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education*, Vol. 2, No. 4, pp. 359–375, October 1989.

[16] 宮澤芳光, 植野真臣. 高精度能力推定を保証する 2 段階等質適応型テスト. 電子情報通信学会論文誌 D, Vol. J106, pp. 34–46, 2023.

[17] Maomi Ueno and Yoshimitsu Miyazawa. Two-Stage uniform adaptive testing to balance measurement accuracy and item exposure. In *Artificial Intelligence in Education*, pp. 626–632. Springer International Publishing, 2022.

[18] Takatoshi Ishii, Pokpong Songmuang, and Maomi Ueno. Maximum clique algorithm and its approximation for uniform test form assembly. *IEEE Transactions on Learning Technologies*, Vol. 7, No. 1, pp. 83–95, 2014.

[19] Takatoshi Ishii and Maomi Ueno. Algorithm for uniform test assembly using a maximum clique problem and integer programming. In *International Conference on Artificial Intelligence in Education*, pp. 102–112. Springer, 2017.

[20] 渕本壱真, 植野真臣. 等質テスト構成における整数計画法を用いた最大クリーク探索の並列化. 電子情報通信学会論文誌 D, Vol. J103, pp. 881–893, 2020.

[21] Kazuma Fuchimoto, Takatoshi Ishii, and Maomi Ueno. Hybrid maximum clique algorithm using parallel integer programming for uniform test assembly. *IEEE Trans. Learn. Technol.*, Vol. 15, No. 2, pp. 252–264, April 2022.

[22] 仁田善雄, 斎藤宣彦, 後藤英司, 高木康, 石田達樹, 江藤一洋. 医療系大学間共用試験における e テスティング. 日本テスト学会第 12 回大会発表論文抄録集, pp. 58–59, 2014.

[23] 谷澤明紀, 本多康弘. 情報処理技術者試験における e テスティング. 日本テスト学会第 12 回大会発表論文抄録集, Vol. 33, No. 2, pp. 54–57, 2014.

[24] F M Lord. *Applications of item response theory to practical testing problems*. Routledge, London, England, November 2012.

[25] Frederic M Lord and Melvin R Novick. *Statistical Theories of Mental Test Scores*. IAP, November 2008.

[26] F B Baker and S H Kim. *Item response theory: Parameter estimation techniques*. CRC Press, 2004.

[27] W J Van der Linden and others. *Handbook of item response theory, volume one: Models*. Chapman and Hall/CRC, 2016.

[28] W J Van der Linden and others. *Handbook of item response theory, volume two: Statistical Tools*. Chapman and Hall/CRC, 2016.

[29] 植野真臣, 永岡慶三. e テスティング. 培風館, 2009.

[30] R Darrell Bock and Robert J Mislevy. Adaptive EAP estimation of ability in a microcomputer environment. *Appl. Psychol. Meas.*, Vol. 6, No. 4, pp. 431–444, September 1982.

[31] S W Choi and S Lim. Adaptive test assembly with a mix of set-based and discrete items. *Behaviormetrika*, Vol. 49, pp. 231–254, 2022.

[32] Seung W Choi, Sangdon Lim, and Wim J van der Linden. TestDesign: an optimal test design approach to constructing fixed and adaptive tests in R. *Behaviormetrika*, Vol. 49, No. 2, pp. 191–229, July 2022.