論　文

高精度能力推定を保証する2段階等質適応型テスト

宮澤　芳光[†a)]　　植野　真臣[††b)]

Two Stages Uniform Adaptive Testing to Guarantee High Measurement Accuracy

Yoshimitsu MIYAZAWA[†a)] and Maomi UENO[††b)]

あらまし　適応型テストとは，受検者の能力値を逐次的に推定し，その能力値に応じて情報量が最も高い項目を出題するコンピュータ・テスティングの出題形式である．適応型テストは，測定誤差を増加させずに，テストの長さや受験時間を短縮できる．しかし，同一の受検者が複数回受験した場合には，同一の項目群が出題される傾向がある．また，特定の項目群が頻繁に暴露される傾向があるため，項目内容の暴露につながり，テストの信頼性の低下要因となりうる．これらの問題を解決するため，項目暴露を制御する適応型テストが提案されてきた．しかし，暴露数の減少と測定誤差の増加にはトレードオフの関係が存在する．本研究では，このトレードオフを制御するため，2段階等質適応型テストを提案する．2段階等質適応型テストは，事前にアイテムバンクを等質に分割して複数の項目集合を構成し，テストの前半に項目集合から情報量が高い項目を選択し，受検者の能力推定値が収束してきたテストの後半にアイテムバンクからその能力推定値付近で情報量が高い項目を選択する．本論では，シミュレーション実験と実データを用いた実験により2段階等質適応型テストの有効性を示す．

キーワード　適応型テスト，eテスティング，等質テスト構成，項目反応理論

1. まえがき

近年，eテスティングが実用化されつつある [1]〜[4]．eテスティングは，Web上でテストを受験するCBT（Computer based testing）であり，受検者が何度でも等質な情報量で異なる項目から構成されたテストを受検できる．このため，テストの結果が受検者に大きな影響を与えるハイ・ステークスなテストで導入が検討されている．eテスティングやCBTの技術として適応型テスト（CAT: Computerized Adaptive Testing）と呼ばれるテスト出題方式が知られている [5]．適応型テストは，受検者の能力値を逐次的に推定し，その能力値に応じて情報量が最大の項目を出題する．これにより，測定誤差を増加させずに，テストの長さや受験時間を短縮できる．

しかし，同一の受検者が複数回受験した場合には，同一の項目群が出題される傾向があり，実際に適応型テストを導入しているSynthetic Personality Inventory（SPI）[6] やGlobal Test of English Communication（GTEC）[7] の重要な問題になっている．更に，特定の項目群が多くの受検者に対して提示されてしまうため，一部の項目が過剰に暴露され，項目内容の漏洩につながり，テストの信頼性の低下要因となりうる [8], [9]．適応型テストの運用には，項目の難易度や識別力を事前に推定し，テスト項目集合（アイテムバンクと呼ばれる）を構築する必要がある．

ハイ・ステークスなテストでは，項目の作成に膨大な経済的・時間的コストを要するため，項目の一様な活用が望ましい．これらの問題を解決するため，暴露制御を用いた適応型テストが提案されている [5], [10]〜[15]．van der Lindenらの手法は，代表的な出題方略であり，暴露上限の制約を用いて項目集合（シャドーテストと呼ばれる）を逐次的に構成し，その項目集合から項目選択する [12]．シャドーテストとは，(1) テストの長さや暴露数の最大値といった試験の制約を全て満たし，(2) 受検者に出題した項目を全て含め，(3) 能力推定値に対して情報量が最大になるように構成された項目集合である．van der Lindenらの手法は，項目選択のたびにシャドーテストを構成して，その中か

† 大学入試センター，東京都
The National Center for University Entrance Examinations, 2–19–23 Komaba, Meguro-ku, Tokyo, 153–8501 Japan
†† 電気通信大学，調布市
The University of Electro-Communications, 1–5–1 Chofugaoka, Chofu-shi, 182–8585 Japan
a) E-mail: miyazawa@rd.dnc.ac.jp
b) E-mail: ueno@ai.is.uec.ac.jp
DOI:10.14923/transinfj.2021JDP7052

ら未出題で情報量が最大の項目を選択する．これにより，暴露数の最大値を制御できる．一方，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 [5], [13]〜[17]．Sympson-Hetter 手法では，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手法が提案されている [13]〜[15]．更に，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開発している [5], [16], [17]．van der Linden らの手法では，受検者に出題可能な確率として適格確率（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外することで暴露を制御する．また，アイテムバンクの分割を用いた手法が提案されている [18]．この手法では，アイテムバンクをランダムに複数の項目集合に分割し，最も暴露数が少ない項目集合から項目選択する．これらの手法では，特定の項目群の過度な暴露を防ぐことはできる．しかし，項目集合の等質性は保証されず，受検者間でテストの長さや測定誤差に偏りが生じる傾向がある．そこで，筆者らは，情報量を制約条件として暴露数が最小の項目集合を構成し，その中から逐次的に項目選択する手法を提案している [19]．同様に，Choi らは，近年，van der Linden らの手法 [12] を拡張して情報量を制約条件として定式化した手法を提案している [20], [21]．しかし，これらの手法では，暴露数の一様性は担保されるが，情報量を制約しているために測定誤差が増加してしまう問題点がある．すなわち，暴露数の減少と測定誤差の増加には，トレードオフの関係がある．

この問題を解決するために，筆者らは，このトレードオフを制御する等質適応型テストを提案してきた [22], [23]．等質適応型テストは，最新の等質テスト構成技術を用いて情報量が等質な項目集合にアイテムバンクを分割し，受検者ごとに異なる等質テストを割り当ててアイテムバンクとみなして適応型テストを実施する．受検者ごとに異なる項目集合をアイテムバンクとして用いるため，能力値が同等な受検者であっても異なる項目群を出題することができ，受検者間の測定誤差を等質にしつつ，暴露数を減少できた．本手法において，等質テストとは，複数のテストが異なる項目で構成されているにもかかわらず，等質な情報量をもつ項目集合である．等質テスト構成技術は，既に e テスティング技術の基幹技術として多数研究されている [24]〜[31]．当初は，その計算量の大きさから少数の等質テストしか生成できなかった．しかし，近年の研究では，大規模な数の等質テストを生成できる技術が開発され（例えば, [25], [28]〜[31]），情報処理技術者試験や医療系共用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 [32], [33]．等質適応型テストでは，当時，最大数の等質テストを構成できる Ishii et.al [28] の手法を用いてアイテムバンクを多数の項目集合に分割し，受検者ごとに異なる項目集合を割り当て，その項目集合から情報量が最大の項目を選択している．これにより，暴露数と測定誤差のトレードオフを制御できた．更に，先行研究 [23] では，等質適応型テストが過学習を避け，テストの長さを短縮できたことを報告している．しかし，一般的にテストの長さが短縮すると能力推定値の測定誤差が大きくなることが多い．先行研究 [22], [23] では，従来の適応型テストの収束基準に従い，受検者の能力推定値が収束することでテストを終了させ，フィッシャー情報量の平方根の逆数により求められる漸近誤差は評価しているが，受検者の真の能力値と能力推定値の差異による測定誤差に関して評価は行っていない．

本研究では，最初に等質適応型テストの測定誤差をシミュレーションにより評価し，能力推定値の測定誤差が大きいことを示す．実験の結果，等質適応型テストでは，アイテムバンクを分割することで項目候補が少なくなり，枯渇して見かけ上能力推定値が収束しており，能力真値と能力推定値が大きく乖離していることを発見した．

本研究では，能力真値と能力推定値が大きく乖離する問題を解決するため，2 段階等質適応型テストを提案する．本手法は，事前にアイテムバンクを分割して情報量が等質な項目集合を複数構成する．次に，テストの前半に項目集合から項目選択し，受検者の能力推定値が収束し始めるテストの後半にアイテムバンク全体から項目選択する．項目集合の項目難易度分布は疎ではあるが，第 1 段階では，過学習を避け，より高速に推定値が真の能力値近傍まで到達することが期待できる．第 2 段階ではアイテムバンクの項目集合全体を用いるが，能力推定値は収束し始めており，暴露分布の偏りは大きくはならないと考えられる．アイテムバンクの分割には，現時点で世界最大数の等質テスト構成を可能にする石井他 (2017) [31] を用いる．石井他 (2017) では，整数計画問題を用いて逐次的に最大クリークを探索し，効率的に等質テストを構成できる．

本研究では，石井他 (2017) の等質テスト構成手法を用いてアイテムバンクを分割し，情報量が等質な項目集合を多数構成する．本論では，シミュレーション実験と実データを用いた実験により 2 段階等質適応型テストの有効性を示す．

2. 項目反応理論による適応型テスト

本章では，本研究の基礎理論として用いる項目反応理論と適応型テストについて述べる．

2.1 項目反応理論

項目反応理論は，数理モデルを用いたテスト理論の一つであり，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 [34]〜[38]．項目反応理論の特徴としては，以下のような点が挙げられる [39]〜[41]．

（1） 測定誤差が大きい異質項目の影響を小さくして受検者の能力値を推定できる．

（2） 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．

（3） 欠測データから容易にパラメータを推定できる．

項目反応理論は，正誤判定問題や多肢選択式問題など，データが正誤の 2 値となる反応データに適用されることが一般的である．このような 2 値データに適用できる項目反応モデルとしては，2 パラメータロジスティックモデル（2PLM: 2-Parameter Logistic Model）が古くから広く利用されてきた．

2PLM では，能力値 $\theta \in (-\infty, \infty)$ の受検者が項目 $i \in \{1, \ldots, N\}$ に正答する確率を以下の式で表す．

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

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 0 の変数である．また，$a_i \in [0, \infty)$ は，項目 i の識別力パラメータと呼ばれ，どのくらいの精度で受検者の能力値を識別できるかを示す．$b_i \in (\infty, \infty)$ は，項目 i の難易度パラメータと呼ばれ，その項目の難しさを表す．適応型テストでは，2PLM を用いることにより，項目特性を考慮して受検者の能力値 θ を推定できる．

2.2 フィッシャー情報量

項目反応理論においては，能力推定の標準誤差がフィッシャー情報量の平方根の逆数の値に漸近的に一致することが知られている [34]．そのため，この漸近誤差を測定誤差の近似として用いることがある．

2PLM では，能力値 θ の受検者に対して項目 i のフィッシャー情報量を以下の式で表す [42]．

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

ここで，

$$p'(u_i = 1|\theta) = \frac{\partial}{\partial\theta} p(u_i = 1|\theta) \tag{3}$$

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ことを意味する．したがって，適応型テストでは，能力値を所与としてフィッシャー情報量の高い項目を各受検者に出題することで，効率のよい能力測定が実現できると期待される．本論において，情報量はフィッシャー情報量を指す．なお，テスト T に含まれる項目集合の情報量の総和であるテスト情報量 $I_T(\theta) = \sum_{i \in T} I_i(\theta)$ の平方根の逆数が能力推定値の漸近誤差に一致し，テストの測定誤差の近似として用いられることが多い．

2.3 能力値 θ の推定

受検者の能力値 θ の推定にはベイズ推定法を用いる [43]．ベイズ推定は，一致性及び漸近有効性をもつと同時に少数データからの推定にも適していることが知られている [44]．能力値 θ の推定は，$k-1$ 項目までの反応データのベクトルを $\mathbf{u}_{k-1}$，能力値 θ の事前分布を $g(\theta)$ として EAP (Expected a posteriori) 推定法を用いる．EAP 推定法は，以下の事後分布の θ に関する期待値を推定値とする方法である．

$$g(\theta|\mathbf{u}_{k-1}) = \frac{L(\theta|\mathbf{u}_{k-1})g(\theta)}{\int L(\theta|\mathbf{u}_{k-1})g(\theta)d\theta} \tag{4}$$

$L(\theta|\mathbf{u}_{k-1})$ は，$k-1$ 項目の反応データを用いた能力値のゆう度である．

2.4 適応型テストのアルゴリズム

適応型テストでは，項目特性が既知のアイテムバンクを所与として，以下の手順で受検者の能力値 θ を推定する．

（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て受検者に出題する．

（3） 項目 i に対する正誤データとそれまでの解答履歴から受検者の能力推定値 $\hat{\theta}$ を更新する．

（4） 受検者の能力推定値 $\hat{\theta}$ の更新幅がしきい値 ϵ

以下になるまで上記の手順 (2) と (3) を繰り返す．

適応型テストは，上記の情報量最大化に基づく項目出題と受検者の能力推定を交互に繰り返すことで高精度に能力値を推定できる [5], [45]．しかし，このアルゴリズムでは，能力推定値 $\hat{\theta}$ に対して出題される項目が一意に定まるため，同じ能力値をもつ受検者に対して同じ項目群を出題する傾向がある．したがって，同一の受検者が複数回受検した場合に同一の項目群が出題される傾向があり，実際に適応型テストを導入したときの深刻な問題になりえる．更に，受検者の能力値が標準正規分布に従うため，$\theta = 0$ 付近で高い情報量をもつ項目群が過度に受検者に暴露される傾向がある．試験問題の作成時には困難度パラメータが未知であるため，特定の困難度パラメータの項目を作成することは難しい．これにより，項目内容の暴露につながり，テストの信頼性の低下要因となりうる [8]．

3. 暴露制御を用いた適応型テスト

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

従来の適応型テストにおける項目暴露数バイアスの問題を解決するため，van der Linden らは，各項目の暴露数に最大暴露数 R という制約を用いて項目集合を逐次構成し，その中から項目選択する手法を提案している（以下，IP と呼ぶ）[12]．IP は，以下のアルゴリズムに従って項目選択する．

（1） 受検者の能力推定値を $\hat{\theta} = 0$ に初期化する．

（2） 整数計画問題を用いてシャドーテストを構成する．

$$\text{maximize} \sum_{i=1}^{N} I_i(\hat{\theta}) x_i \tag{5}$$

subject to

$r_i x_i \leq R; (i = 1, \cdots, N)$,

(項目 i の暴露数 r_i，最大暴露数 R)

$\sum_{i=1}^{N} x_i = n$; (テストの長さ),

$$x_i = \begin{cases} 1\text{:項目 } i \text{ がシャドーテストに含まれるとき,} \\ 0\text{:上記以外} \end{cases} \tag{6}$$

（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） 受検者の能力推定値 θ の更新幅がしきい値 ϵ 以下になるまで手順 (2)〜(4) を繰り返す．

3.2 van der Linden and Veldkamp の適応型テスト（LV）

別のアプローチとしては，確率的な項目選択制御を用いた手法が提案されている [5], [13]〜[17]．van der Linden と Veldkamp は，適格確率（Eligibility probability）を用いた手法を提案している（以下，LV と呼ぶ）[5], [16], [17]．適格である項目は，アイテムバンクに残し，適格でない項目はアイテムバンクから除外される．具体的には，受検者 j に対して項目 i の適格確率を $P^{(j)}(E_i)$，暴露率の上限値を $r^{\max}$，受検者 j までの項目 i の暴露率を $P^{(j)}(A_i)$ としたとき，LV は以下のアルゴリズムに従って項目選択する．

（1） 受検者の能力推定値を $\hat{\theta} = 0$ に初期化する．

（2） 適格確率 $P^{(j)}(E_i)$ に従ってアイテムバンクに項目を残す．適格でなければアイテムバンクから除外する．

$$P^{(j)}(E_i) = \min\left\{ \frac{r^{\max}}{P^{(j-1)}(A_i)} P^{(j-1)}(E_i), 1 \right\} \tag{7}$$

（3） アイテムバンクから情報量が最大の項目を選択する．

（4） 受検者の能力値 $\hat{\theta}$ を推定する．

（5） 受検者の能力推定値 θ の更新幅がしきい値 ϵ 以下になるまで手順 (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．

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

Kingsbury と Zara は，暴露数の偏りを軽減させるためにアイテムバンクの分割を用いた手法を提案している（以下，KZ と呼ぶ）[18]．KZ は，以下のアルゴリズムに従って項目選択する．

（1） アイテムバンクをランダムに分割して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） 受検者の能力推定値 θ の更新幅がしきい値 ϵ 以下になるまで上記の手順 (3) と (4) を繰り返す．

これらの手法では，特定の項目群の過度な暴露を防

ぐことができた．しかし，項目集合の測定誤差の等質性は保証されず，受検者間でテストの長さや測定誤差に偏りが生じる傾向がある．

3.4 Miyazawa and Ueno の適応型テスト（MU）

この問題を解決するために，筆者らは情報量の制約を用いた手法を提案している（以下，MU と呼ぶ）[19]．MU は，以下のアルゴリズムに従って項目選択する．

（1） 受検者の能力推定値を $\hat{\theta} = 0$ に初期化する．

（2） 整数計画問題を用いて項目集合を構成する．

$$\text{Minimize} \sum_{i=1}^{I} e_i x_i, (\text{項目 } i \text{ の暴露数 } e_i) \tag{8}$$

subject to

$$LB(\theta_l) \leq I(\theta_l) \leq UB(\theta_l)$$

$$(l = 1, \cdots, L), (\text{テスト情報量の下限値と上限値})$$

$$\sum_{i=1}^{N} x_i = n; (\text{テストの長さ}),$$

$$x_i = \begin{cases} 1: \text{項目 } i \text{ が項目集合に含まれるとき}, \\ 0: \text{上記以外} \end{cases} \tag{9}$$

（3） 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） 受検者の能力推定値 θ の更新幅がしきい値 ϵ 以下になるまで手順 (2)〜(4) を繰り返す．

MU は，テスト情報量の下限値と上限値を制約とすることで受検者間で漸近誤差の等質性を実現している．しかし，これらの手法では，暴露数の一様性は担保されるが，情報量を制約しているために測定誤差が増加してしまう問題点がある．すなわち，暴露数の減少と測定誤差の増加にはトレードオフの関係がある．

3.5 等質適応型テスト（UAT）

暴露数の減少と測定誤差の増加のトレードオフを制御するために，筆者らは等質適応型テスト（Uniform Adaptive Testing; 以降 UAT と呼ぶ）を提案してきた [22], [23]．UAT は，等質テスト構成技術を用いてアイムバンクを分割して項目集合を構成し，その項目集合から項目選択する．当初の UAT では，当時，最大数の等質テストを構成できる Ishii et.al [28] の手法を用いてアイテムバンクを多数の項目集合に分割していた．Ishii et al. (2014) [28] の手法では，等質テスト構成問題を最大クリーク問題として扱う．具体的には，次のグラフを考え，そこから最大のクリーク（その集合に含まれる任意の頂点が全て結合されている）構造を探索することで等質テストを構成する．

頂点：与えられたアイテムバンクから重複条件以外の全てのテスト構成条件を満たすテストを頂点とする．

エッジ：二つの頂点（テスト）が重複条件を満たしている場合（重複条件により指示される最大重複項目数より少ない重複項目しかもっていないなら），その二つの頂点間にエッジを張る．

このように作成されたグラフのクリークは，所望のテスト構成条件を満たした等質テストの集合と解釈できる．そのため，グラフの最大クリークを探索することで，最大数の等質テストを構成できる．

UAT は，以下のアルゴリズムに従って項目選択する．

（0） 事前に等質テスト構成技術を用いてアイムバンクを分割し，情報量が等質な項目集合を複数構成する．

（1） 受検者への項目集合の割り当てをランダムに行う．

（2） 受検者の能力推定値を $\hat{\theta} = 0$ に初期化する．

（3） 受検者に割り当てられた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） 受検者の能力推定値 θ の更新幅がしきい値 ϵ 以下になるまで手順 (3)〜(4) を繰り返す．

UAT では，受検者ごとに異なる項目集合を割り当てることで暴露数を低減させている．しかし，先行研究 [22], [23] では，フィッシャー情報量を用いた漸近誤差が評価されているが，受検者の真の能力値と能力推定値の差異による測定誤差については分析されていない．

4. 等質適応型テストの測定誤差評価

ここでは，受検者の真の能力値と能力推定値の差異による測定誤差の評価実験を行う．具体的には，通常の適応型テスト (CAT)，整数計画問題（Integer Programming Problem）を用いた制約付き適応型テスト (IP) [12]，van der Linden and Veldkamp の適格確率を用いる手法（LV）[5], [16], [17]，Kingsbury and Zara のアイテムバンクの分割を用いた手法 (KZ) [18]，等質適応型テスト（UAT）[22], [23] の測定誤差に関する比較を行う．評価実験の手順は，以下のとおりである．

（1） 1000 項目で構成されるアイテムバンクを生成する．このとき，各項目のパラメータは，$\log a_i \sim N(-0.5, 0.2)$，$b_i \sim N(0, 1)$ からサンプリングした．

表 1　既存手法の検証実験
Table 1　Experimental results of the previous methods.

手法	テストの長さの平均		暴露数の平均		RMSE
CAT	19.85	(2.40)	174.14	(201.73)	0.30
IP	23.81	(3.63)	67.65	(40.37)	0.32
LV	21.98	(3.25)	86.55	(43.49)	0.31
KZ	19.74	(3.32)	109.69	(113.14)	0.33
UAT	16.18	(2.87)	17.91	(12.61)	0.43

（2） 受検者の真の能力値を $\theta \sim N(0,1)$ からサンプリングする．

（3） 受検者の能力推定値を $\hat{\theta} = 0$ に初期化する．

（4） 各手法に基づき項目を選択し，能力真値と項目パラメータを所与として項目への反応データを発生させる．

（5） EAP 法 [44] を用いて解答履歴データから能力推定値 $\hat{\theta}$ を推定する．

（6） 能力推定値の更新幅が 0.05 以下になるまで手順 (4) と (5) を繰り返す．

（7） 手順 2 から 6 を 1000 回繰り返し，得られた出題パターンと解答履歴を用いて，a）テストの長さ，b）各項目の暴露数，c）能力真値と能力推定値の RMSE（Root Mean Square Error），に関する統計量を求めた．

実験結果を表 1 に示す．表 1 の括弧内には標準偏差の値を表した．UAT は，先行研究 [22], [23] で報告されたとおり，テストの長さが短く，暴露数が小さいことがわかる．しかし，先行研究 [22], [23] では，漸近誤差を評価しているが，受検者の真の能力値と能力推定値の差異による測定誤差の評価を行っていない．本シミュレーションの結果から能力真値と能力推定値の RMSE については，その値が大きいことがわかる．UAT では，アイテムバンクを分割することで項目候補が少なくなり，枯渇して見かけ上能力推定値が収束していた．このため，能力真値と能力推定値が大きく乖離している．本研究では，UAT の問題を解決し，測定誤差の小さい等質適応型テストを開発する．

5. 2 段階等質適応型テスト

本研究では，従来の等質適応型テストの問題を解決するため，2 段階等質適応型テストを提案する．2 段階等質適応型テストでは，事前にアイテムバンクを分割して情報量が等質な項目集合を複数構成し，テストの前半に項目集合から項目選択し，受検者の能力推定値が収束し始めたテストの後半にアイテムバンク全体から項目選択する．項目集合の項目難易度分布は疎ではあるが一様に分布しており，第 1 段階では，過学習を避け，より高速に推定値が真の能力値近傍まで到達することが期待できる．第 2 段階では，アイテムバンク全体から項目選択するが，能力推定値が収束し始めており，暴露分布の偏りはそれほど大きくはならないと考えられる．

アイテムバンクの分割には，現時点で世界最大数の等質テスト構成を可能にする石井他 (2017) [31] を用いる．石井他 (2017) では，整数計画問題を用いて最大クリークを逐次探索し，効率的に等質テストを構成できる手法を提案している．石井他 (2017) の手法では，構成中の等質テスト群を $\boldsymbol{C}$，構成済みの等質テスト数を $|\boldsymbol{C}|$ とし，以下の整数計画問題を用いて等質テストを構成する．

$$
\begin{aligned}
&\text{maximize} \sum_{i=1}^{N} \lambda_i x_i \qquad (10)\\
&\text{subject to}\\
&\sum_{i=1}^{N} y_{k,i} x_i \le O(\text{項目の重複上限数}); (k = 1, \cdots, |C|)\\
&\sum_{i=1}^{N} x_i = n; (\text{テストの長さ}),\\
&\sum_{i=1}^{N} I_i(\theta_l) x_i = I(\theta_l)\\
&LB(\theta_l) \le I(\theta_l) \le UB(\theta_l)\\
&(l = 1, \cdots, L)\\
&x_i = \begin{cases} 1\text{:項目 } i \text{ が等質テストに含まれるとき,} \\ 0\text{:上記以外} \end{cases}\\
&y_{k,i} = \begin{cases} 1\text{:}i \text{ 番目の項目が } \boldsymbol{C} \text{ 中の } k \text{ 番目の} \\ \quad \text{等質テストに含まれるとき,} \\ 0\text{:上記以外} \end{cases}
\end{aligned}
\qquad (11)
$$

ここで，λ_i は，互いに独立な [0, 1) の連続一様分布からの乱数であり，本問題が解かれるたびにリサンプリングされる．構成済みの等質テストを $\boldsymbol{C}$ として，O は，重複項目数の上限である．$LB(\theta_l)$ は，情報量の下限であり，$UB(\theta_l)$ が上限である．$LB(\theta_l)$ と $UB(\theta_l)$ は，アイテムバンクに含まれる項目の特性に応じて適切に設定する必要がある．本研究では，項目の情報量の平均を m_{ib} とし，標準偏差を sd_{ib}，等質テストのテストの長さを n としたとき，情報量の上限を $(m_{ib} + sd_{ib})n$

とし，下限を $m_{ib}n$ とした．石井他 (2017) の手法では，整数計画問題を用いた等質テストの構成と局所解へ収束することを回避するためのランダムな削除を繰り返す．

2 段階等質適応型テストでは，事前に石井他 (2017) の手法を用いてアイテムバンクを分割して情報量が等質な項目集合を複数構成する．この項目集合を用いた 2 段階等質適応型テストのアルゴリズムについて詳述する．第 1 段階では以下のアルゴリズムに従って項目選択する．

（1） 項目集合をランダムに割り当てる．

（2） 能力推定値を $\hat{\theta} = 0$ に初期化する．

（3） 能力推定値 $\hat{\theta}$ を所与として項目集合から情報量が最大となる項目を選択して出題する．

（4） 項目への反応データとそれまでの解答履歴から能力推定値 $\hat{\theta}$ を求める．

（5） 手順 (3) と (4) を $\hat{\theta}$ の更新幅がしきい値 ϵ 以下になるまで繰り返す．

次に，第 2 段階では以下のように出題方略が変更される．

（1） 能力推定値 $\hat{\theta}$ を所与としてアイテムバンク全ての項目集合から情報量が最も高い項目を選択する．ハイステークスな試験では，暴露数の上限値が制約として決まっている場合がある．必要に応じて，このステップで暴露数の上限値を制約として組み込む．

（2） 項目への反応データとそれまでの解答履歴から能力推定値 $\hat{\theta}$ を求める．

（3） 手順 (1) と (2) をテストの終了条件まで繰り返す．

従来の適応型テストでは，能力推定値の更新幅がしきい値以下になるときをテスト終了条件として設定していた [2]．しかし，等質適応型テストでは，テスト終了条件を能力推定値の更新幅がしきい値以下とした場合，項目集合に情報量の高い項目がなくなり，能力推定値が能力真値に収束する前にテストが終了することがある．そこで本研究では，van der Linden らの適応型テスト [5] と同様にテストの長さをテスト終了条件とした．本手法では，上記の手順でアイテムバンクの項目をできる限り一様に活用しながら受検者の能力を高精度で推定する．

6. 評 価 実 験

本章では，2 段階等質適応型テストの有効性を確認するため，シミュレーション実験により 2 段階等質適応型テストと従来の適応型テストの性能を比較する．本実験では，2 段階等質適応型テスト（Proposal）と **2.4** で述べた適応型テスト (CAT)，van der Linden らの整数計画問題（Integer Programming Problem）に基づく適応型テスト (IP) [12]，Miyazawa and Ueno の情報量の制約を用いた適応型テスト（MU）[19]，van der Linden and Veldkamp の確率的な項目選択制御を用いた適応型テスト（LV）[5], [16], [17]，Kingsbury and Zara (1989) のアイテムバンクの分割を用いた適応型テスト (KZ) [18] を比較する．

6.1 反応パターンの生成

シミュレーション実験の手順は以下のとおりである．

（1） (a)(b)(c)(d) の特徴をもつアイテムバンクを生成する．項目数は 1000 とする．項目 i のパラメータは，以下の分布からサンプリングする．

（a） $\log a_i \sim N(-0.5, 0.2)$，$b_i \sim N(0, 1.0)$

（b） $\log a_i \sim N(-0.5, 0.2)$，$b_i \sim N(0, 3.0)$

（c） $\log a_i \sim N(-0.75, 0.2)$，$b_i \sim N(0, 1.0)$

（d） $\log a_i \sim N(-0.75, 0.2)$，$b_i \sim N(0, 3.0)$

（2） 受検者の能力真値を $\theta \sim N(0, 1)$ からサンプリングする．

（3） 受検者の能力推定値を $\hat{\theta} = 0$ に初期化する．

（4） CAT を用いてアイテムバンクから項目を選択する．ここでは，能力推定値 $\hat{\theta}$ を所与として情報量基準に基づき項目を決定する．

（5） 項目への反応データを，能力真値と項目パラメータを所与として発生させる．

（6） 項目への反応データとそれまでの解答履歴から能力推定値 $\hat{\theta}$ を求める．

（7） テスト終了条件を満たすまで手順 (4) から (6) を繰り返す．

（8） 手順 (2) から (6) を 1000 回繰り返す．生成された出題パターンと解答履歴を用いて，a）各項目の暴露数，b）能力推定値の真値との RMSE，に関する統計量を求めた．

テスト終了条件については，前章のとおり，能力推定値が能力真値に収束する前に能力推定値の更新幅が小さくなり，テストが終了することが確認された．このため，テスト終了条件はテストの長さとした．本評価実験では，テストの長さを 30 項目と 50 項目とした．2 段階等質適応型テストでは，項目集合サイズと 2 段階目への切り替えの条件を事前に設定する必要がある．2 段階等質適応型テストは，項目集合サイズと 2 段階目への切り替えの条件を適切に決定することで，

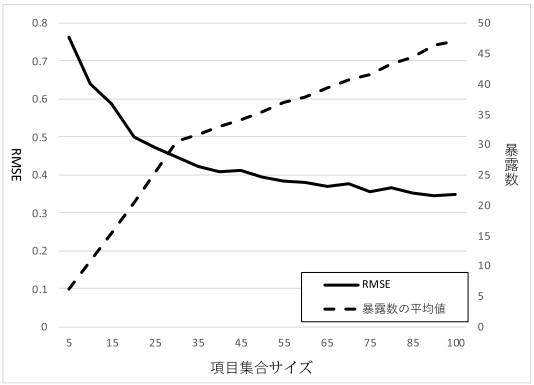

図 1　項目集合サイズに対する RMSE と暴露数
Fig. 1　RMSE and the number of exposure items for each item group size.

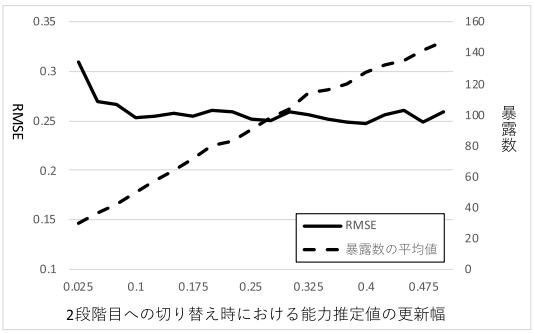

図 2　2 段階目への切り替え時における能力推定値の更新幅に対する RMSE と暴露数
Fig. 2　RMSE and the number of exposure items for each update difference of the examinee's ability estimate when switching to the second stage.

暴露数の減少と測定誤差の増加のトレードオフを制御する．

6.2　2 段階等質適応型テストのパラメータチューニング

本節では，項目集合サイズ（項目数）と 2 段階目への切り替えの条件を決定する．まず，項目集合サイズを求める．ここでは，項目集合サイズを 5 から 100 までステップ数を 5 として前述の手順でシミュレーション実験を実施した．結果を図 1 に示す．図 1 は，横軸が項目集合サイズであり，縦軸に RMSE と暴露数の平均値を示した．項目集合サイズは，図 1 のとおり，RMSE と暴露数に関してトレードオフの関係であることがわかる．しかし，RMSE は，項目数の増加とともに減少幅が小さくなっていることがわかる．そこで，本研究では，RMSE の減少の幅が 0.01 以下のときの項目数を採用することにした．KZ 法の項目集合サイズについては同様の方法で決定している．

次に，2 段階目への切り替えの条件を検討する．ここでは，2 段階目への切り替え時の能力推定値の更新幅（以下，切替時更新幅と呼ぶ）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，横軸が切替時更新幅であり，縦軸に RMSE と暴露数の平均値を示した．切替時更新幅については，項目集合サイズと同様に，RMSE と暴露数に関してトレードオフの関係があり，かつ暴露数は単調増加し，RMSE は収束していることがわかる．このため，RMSE の減少の幅が 0.01 以下のときの切替時更新幅を採用する．

本論文では，上記の手順によって決定された項目集合サイズと切替時更新幅を用いて他の手法と比較する．

6.3　実 験 結 果

実験結果を表 2 に示す．それぞれの表の括弧には標準偏差の値を記した．また，手法の左隣の括弧には，各手法でのパラメータチューニングの値を表す．2 段階等質適応型テストの括弧には，項目集合サイズと切替時更新幅を示す．KZ の括弧は，項目集合サイズを表す．なお，全ての手法で項目選択時間が 1 秒以内であった．

暴露数の平均値と未出題の項目数は，全ての条件で MU の値が小さかった．MU は，テスト情報量の制約内で暴露数が最小の項目を選択するため，暴露数と未出題項目数を最も減少できる．しかし，MU の RMSE の値は大きく，能力真値と能力推定値が大きく乖離している．更に，RMSE の改善の度合いを分析するため，受検者の能力推定値 $\hat{\theta}$ を 100 点満点の試験として点数を算出する．具体的には，受検者の得点分布は平均が 50 点，標準偏差が 20 点の正規分布に従っていると仮定し，受検者の試験得点を $s = 20\hat{\theta} + 50$ とした．このとき，能力推定値 $\hat{\theta}$ の RMSE が 0.1 のときには，100 点満点の試験得点では 2 点の測定誤差があることを意味する．MU は，他の手法と比較して，RMSE に 0.1 以上の差があり，試験得点の測定誤差が 2 点以上であることがわかる．2 段階等質適応型テストは，MU の次に暴露数と未出題項目数が小さかった．2 段階等質適応型テストでは，項目集合が異なる項目から構成されているために受検者ごとに出題される項目が異なり，できる限り項目を一様に活用できた．テストの長さについては，その値が大きくなったとき，暴露数が増加

表2 実験結果
Table 2 Experimental results.

テストの長さ	識別力パラメータ $\log a_i$ の分布	難易度パラメータ b_i の分布	手法	暴露数の平均値		暴露数の最大値	未出題の項目数	RMSE
30	N(-0.5, 0.2)	N(0, 1.0)	CAT	193.55	(201.51)	1000	845	0.25
			LV	90.36	(40.51)	149	668	0.29
			IP	75.38	(36.78)	100	602	0.29
			KZ(20)	127.12	(121.31)	462	764	0.29
			MU	30.00	(0.88)	34	0	0.40
			Proposal(25, 0.15)	74.62	(131.13)	652	598	0.24
		N(0, 3.0)	CAT	169.49	(203.70)	1000	823	0.27
			LV	82.42	(46.77)	152	636	0.33
			IP	143.54	(175.28)	500	791	0.31
			KZ(30)	120.48	(138.14)	608	751	0.30
			MU(35)	30.36	(13.86)	64	12	0.47
			Proposal(35, 0.3)	55.45	(125.95)	765	459	0.27
	N(-0.75, 0.2)	N(0, 1.0)	CAT	238.10	(240.92)	1000	874	0.33
			LV	95.85	(40.43)	158	687	0.36
			IP	76.92	(36.45)	100	610	0.36
			KZ(15)	109.89	(114.99)	437	727	0.36
			MU	30.00	(0.85)	33	0	0.47
			Proposal(25, 0.175)	74.81	(74.81)	812	599	0.31
		N(0, 3.0)	CAT	192.31	(231.09)	1000	844	0.35
			LV	88.24	(46.71)	160	660	0.40
			IP	170.45	(190.78)	500	824	0.37
			KZ(25)	125.00	(147.05)	610	760	0.37
			MU(35)	30.43	(13.38)	68	14	0.57
			Proposal(35, 0.225)	51.37	(122.82)	740	416	0.35
50	N(-0.5, 0.2)	N(0, 1.0)	CAT	216.45	(205.10)	1000	769	0.20
			LV	97.28	(40.00)	162	486	0.24
			IP	97.28	(36.29)	100	463	0.25
			KZ(25)	149.70	(179.61)	794	666	0.21
			MU	50.00	(0.87)	53	0	0.31
			Proposal(25, 0.075)	91.58	(147.46)	669	454	0.20
		N(0, 3.0)	CAT	166.11	(202.77)	1000	699	0.23
			LV	87.72	(47.39)	170	430	0.27
			IP	147.49	(175.73)	500	661	0.26
			KZ(20)	157.23	(180.05)	1000	682	0.24
			MU(35)	35.00	(9.19)	68	0	0.47
			Proposal(35, 0.1)	64.27	(133.26)	768	222	0.23
	N(-0.75, 0.2)	N(0, 1.0)	CAT	248.76	(242.18)	1000	799	0.26
			LV	101.63	(39.63)	167	508	0.30
			IP	93.33	(31.30)	100	450	0.32
			KZ(25)	173.01	(210.94)	846	711	0.27
			MU	50.00	(0.86)	53	0	0.37
			Proposal(25, 0.175)	90.25	(90.25)	875	446	0.25
		N(0, 3.0)	CAT	196.08	(239.94)	1000	745	0.28
			LV	94.52	(44.87)	175	471	0.35
			IP	168.92	(189.73)	500	704	0.33
			KZ(25)	135.14	(192.08)	821	630	0.29
			MU(35)	35.00	(8.75)	72	0	0.57
			Proposal(35, 0.175)	73.64	(158.51)	778	321	0.27

している．特に，KZ は，暴露数が大きく増加している．一方，LV や IP は暴露数が比較的小さいことがわかる．しかし，LV や IP は，RMSE の値が大きく，測定誤差を十分に減少できていない．2 段階等質適応型テストは，暴露数の減少と測定誤差の増加のトレードオフを制御し，与えられたアイテムバンクから効果的なテストを実現できている．

RMSE については，2 段階等質適応型テストと CAT の値が最も小さい．しかし，CAT は暴露数の値が高く，特定の項目群が過度に暴露されている．また，項目の未出題項目についても，CAT の値が最も高く，アイテムバンクを十分に活用できていない．2 段階等質適応

表 3　最大暴露数の制約を用いた実験結果
Table 3　Experimental results using a constraint on the maximum number of item exposures.

テストの長さ	識別力パラメータ $\log a_i$ の分布	難易度パラメータ b_i の分布	手法	暴露数の平均値	暴露数の最大値	未出題の項目数	RMSE
30	N(-0.5, 0.2)	N(0, 1.0)	CAT	76.53 (36.00)	100	608	0.29
			LV	76.34 (36.13)	100	607	0.29
			IP	75.95 (36.41)	100	605	0.29
			KZ(20)	64.10 (39.84)	100	532	0.30
			MU	30.00 (0.88)	34	0	0.38
			Proposal(30,0.25)	57.80 (43.63)	100	481	0.28
		N(0, 3.0)	CAT	74.26 (37.93)	100	596	0.31
			LV	74.07 (38.12)	100	595	0.32
			IP	74.07 (38.12)	100	595	0.32
			KZ(20)	62.89 (40.25)	100	523	0.32
			MU(35)	30.49 (13.96)	64	16	0.42
			Proposal(35,0.075)	42.25 (40.55)	100	290	0.30
	N(-0.75, 0.2)	N(0, 1.0)	CAT	79.79 (34.08)	100	624	0.36
			LV	78.33 (35.39)	100	617	0.36
			IP	77.72 (35.96)	100	614	0.38
			KZ(25)	71.60 (37.36)	100	581	0.37
			MU	30.00 (0.85)	33	0	0.43
			Proposal(30,0.35)	68.64 (41.49)	100	563	0.34
		N(0, 3.0)	CAT	78.33 (35.92)	100	617	0.38
			LV	77.12 (36.23)	100	611	0.39
			IP	76.73 (36.95)	100	609	0.39
			KZ(20)	68.03 (39.02)	100	559	0.40
			MU(30)	30.00 (9.05)	71	0	0.49
			Proposal(30,0.1)	42.98 (42.70)	100	302	0.37
50	N(-0.5, 0.2)	N(0, 1.0)	CAT	79.74 (34.32)	100	373	0.25
			LV	79.62 (34.31)	100	372	0.25
			IP	79.24 (34.84)	100	369	0.25
			KZ(15)	74.63 (35.71)	100	330	0.25
			MU	50.00 (0.85)	53	0	0.37
			Proposal(30,0.1)	63.81 (40.78)	100	217	0.24
		N(0, 3.0)	CAT	80.13 (34.49)	100	376	0.27
			LV	79.24 (35.44)	100	369	0.27
			IP	80.26 (34.37)	100	377	0.27
			KZ(25)	74.96 (37.69)	100	333	0.29
			MU(35)	35.00 (9.19)	68	0	0.43
			Proposal(35,0.175)	73.42 (39.42)	100	319	0.26
	N(-0.75, 0.2)	N(0, 1.0)	CAT	83.75 (30.81)	100	403	0.30
			LV	83.47 (31.38)	100	401	0.30
			IP	83.19 (31.60)	100	399	0.31
			KZ(20)	82.51 (31.37)	100	394	0.31
			MU	50.00 (0.85)	53	0	0.36
			Proposal(25,0.175)	66.14 (42.42)	100	244	0.29
		N(0, 3.0)	CAT	82.24 (32.78)	100	392	0.33
			LV	81.04 (33.64)	100	383	0.33
			IP	81.83 (33.26)	100	389	0.33
			KZ(20)	80.00 (34.32)	100	375	0.35
			MU(35)	35.00 (8.75)	72	0	0.49
			Proposal(35, 0.075)	63.78 (42.71)	100	216	0.32

型テストは，実験条件の一部で CAT よりも RMSE の値が小さい．$b_i \sim N(0,3)$ のアイテムバンクを用いた実験結果は，$b_i \sim N(0,1)$ のアイテムバンクを用いた実験結果と比較し，全ての手法で RMSE が増大した．しかし，2 段階等質適応型テストは，CAT の RMSE の値と同じかそれより小さい値であることが確認された．

2 段階等質適応型テストでは，第 1 段階で過学習を避け，より高速に推定値が真の能力値近傍まで到達し，第 2 段階ではアイテムバンク全体から項目を選択することで高精度なテストが実現されている．アイテムバンクの識別力については，その値が小さいとき，どの条件でも RMSE の値が大きかった．適応型テストは，

表 4　実データのアイテムバンクを用いた実験結果
Table 4　Experimental results using an actual item pool.

最大暴露数の制約	テストの長さ	手法	暴露数の平均値	暴露数の最大値	未出題項目数	RMSE
無	30	CAT	227.27 (227.99)	1000	846	0.24
		LV	95.85 (40.83)	159	665	0.34
		IP	80.86 (33.28)	100	607	0.33
		KZ(20)	131.58 (140.35)	532	750	0.29
		MU	30.67 (1.85)	42	0	0.48
		Proposal(20, 0.225)	80.21 (163.75)	748	604	0.24
	50	CAT	243.90 (233.59)	1000	773	0.20
		LV	104.60 (39.98)	183	500	0.27
		IP	83.61 (31.66)	100	380	0.29
		KZ(25)	165.56 (198.94)	801	676	0.23
		MU	51.12 (2.46)	72	0	0.37
		Proposal(20, 0.075)	69.83 (151.16)	788	284	0.20
有	30	CAT	81.08 (33.19)	100	608	0.33
		LV	79.37 (34.73)	100	600	0.34
		IP	79.58 (34.66)	100	601	0.33
		KZ(25)	73.35 (36.38)	100	569	0.34
		MU	30.67 (1.85)	43	0	0.45
		Proposal(20, 0.35)	64.10 (43.71)	100	510	0.30
	50	CAT	82.78 (32.29)	100	374	0.28
		LV	83.47 (31.60)	100	379	0.28
		IP	82.78 (32.30)	100	374	0.30
		KZ(25)	78.37 (35.87)	100	340	0.29
		MU	51.12 (2.47)	72	0	0.36
		Proposal(30, 0.1)	62.11 (43.60)	100	173	0.28

他の手法と比較して RMSE の増加が小さかった．しかし，暴露数の平均値が高く，特定の項目群が過度に暴露されている．2 段階等質適応型テストは，アイテムバンクの特性に応じて項目集合サイズや切替更新幅によって適切に制御されることによって，暴露数の減少と測定誤差の増加のトレードオフを制御できることがわかる．

暴露数の最大値については，MU の次に IP の値が最も小さかった．しかし，IP は暴露数の最大値を直接制御しており，2 段階等質適応型テストを含め，他の手法に容易に組み込むことができる．暴露数の最大値を制約とした実験の結果を表 3 に示す．暴露数については，全ての手法で暴露数の平均値が減少している．一方，情報量が高い項目が出題されにくくなるので RMSE の値が増加している．2 段階等質適応型テストは，RMSE の値が最も小さかった．更に，暴露数の平均値が小さい．2 段階等質適応型テストは，最大暴露数を制御しながら高精度なテストを実現できた．

7. 実データを用いた評価実験

本章では，実データのアイテムバンクを用いて 2 段階等質適応型テストの有効性を評価する．ここでは，リクルート（株）で開発された SPI [6] のアイテムバンクを用いて，シミュレーション実験と同様の手順で実験した．アイテムバンクの項目数は 978 であった．なお，項目選択時間は，**6.3** の実験結果と同様，全ての手法で 1 秒以内であった．

結果を表 4 に示す．暴露数の平均値と未出題項目数については，2 段階等質適応型テストは MU の次に値が小さかった．しかし，MU は，前述のとおり，真の能力値と能力推定値が大きく乖離している．2 段階等質適応型テストは，他の手法と比較して未出題項目数を小さくできた．アイテムバンクの困難度は，SD が 1.57 であり，受検者の能力値上で最適な項目が広く分布している．CAT と LV，KZ は，測定誤差が大きい項目特性をもつ項目が全く出題されなかったと考えられる．CAT と LV，KZ の項目選択のアルゴリズムでは，アイテムバンクから情報が高い項目が選択される傾向があるため，情報量の低い項目が選択されにくい．2 段階等質適応型テストは，項目集合から受検者の能力値に応じて選択し，項目をできる限り一様に活用できた．

RMSE については，2 段階等質適応型テストと CAT の値が最も小さい．しかし，CAT は，暴露数の値が高く，特定の項目群が過度に暴露されている．実データのアイテムバンクでは，シミュレーションから生成さ

れたアイテムバンクと同様に，暴露数と RMSE についてトレードオフの関係が確認された．2 段階等質適応型テストは，トレードオフを制御し，実データのアイテムバンクでできる限り一様に項目を活用しながら高精度なテストを実現している．

8. むすび

本研究では，従来の等質適応型テストの問題を解決するため，2 段階等質適応型テストを提案した．2 段階等質適応型テストでは，事前にアイテムバンクを分割して情報量が等質な項目集合を複数構成し，テストの前半に項目集合から項目選択し，テストの後半にアイテムバンク全ての項目から項目選択する．アイテムバンクの分割には，Ishii et al. (2014) [28] より構成数が多い，石井他 (2017) [31] を用いた．本論では，シミュレーション実験と実データを用いた実験により，2 段階等質適応型テストの利点として以下の点が確認できた．

（1） 2 段階等質適応型テストは，測定誤差を CAT と同程度にすることができた．

（2） 2 段階等質適応型テストは，項目をできる限り一様に出題でき，項目の暴露数を減少させることができた．

（3） 2 段階等質適応型テストは，アイテムバンクの特性やテストの終了条件にかかわらず上記の特性をもつ．

本手法は，IP や KZ を用いた項目選択と比較し，現在の暴露数をデータベースに保持する必要がない．このため，インターネットへの接続が必要な WAN 方式での実施が困難な場合に活用することができる．今後，並列化を用いた等質テスト構成技術 [46] を活用し，より効果的な適応型テストの実現を検討する．

謝辞 本研究は JSPS 科研費 JP19H05663，JP19K21751，JP21K12170，JP22K19825 の助成を受けたものです．

文　献

[1] M. Ueno, “Web based computerized testing system for distance education,” Educational Technology Research, vol.28, no.1-2, pp.59–69, 2005.

[2] 植野真臣，永岡慶三，e テスティング，培風館，2009.

[3] M. Ueno, K. Fuchimoto, and E. Tsutumi, “e-testing from artificial intelligence approach,” Behaviormetrika, vol.48, no.2, pp.409–424, 2021.

[4] M. Ueno, “Ai based e-testing as a common yardstick for measuring human abilities,” 18th Int. Joint Conf. Computer Science and Software engineering, pp.1–6, IEEE, 2021.

[5] W. van der Linden and C. Glas, eds., Elements of Adaptive Testing (Statistics for Social and Behavioral Sciences), Springer, 2010.

[6] Recruit, “Synthetic personality inventory(spi)”. http://www.spi.recruit.co.jp/

[7] 株式会社ベネッセコーポレーション，“Global test of english communication”. https://www.benesse.co.jp/gtec/

[8] W.D. Way, “Protecting the integrity of computerized testing item pools,” Educational Measurement: Issues and Practice, vol.17, pp.17–27, 1998.

[9] H. Wainer, “Cats: Whither and whence,” Psicológica, vol.21, no.1, pp.121–133, 2000.

[10] L. Swanson and M.L. Stocking, “A model and heuristic for solving very large item selection problems,” Applied Psychological Measurement, vol.17, no.2, pp.151–166, 1993.

[11] Y. Cheng and H. Chang, “The maximum priority index method for severely constrained item selection in computerized adaptive testing,” British Journal of Mathematical and Statistical Psychology, pp.369–383, 2009.

[12] W. van der Linden and L.M. Reese, “A model for optimal constrained adaptive testing,” Applied Psychological Measurement, vol.22, no.3, pp.259–270, 1998.

[13] R.D. Hetter and J.B. Sympson, “Item exposure control in CAT-ASVAB,” pp.141–144, American Psychological Association, 1997.

[14] M.L. Stocking and C. Lewis, “Controlling item exposure conditional on ability in computerized adaptive testing,” J. Educational and Behavioral Statistics, vol.23, pp.57–75, 1998.

[15] M.L. Stocking and C. Lewis, “Methods of controlling the exposure of items in cat,” pp.163–182, Springer, 2000.

[16] W. van der Linden and B.P. Veldkamp, “Constraining item exposure in computerized adaptive testing with shadow tests,” J. Educational and Behavioral Statistics, vol.29, no.3, pp.273–291, 2004.

[17] W. van der Linden and S. Choi, “Improving item-exposure control in adaptive testing,” J. Educational Measurement, vol.57, no.3, pp.405–422, 2019.

[18] G.G. Kingsbury and A.R. Zara, “Procedures for selecting items for computerized adaptive tests,” Applied Measurement in Education, vol.2, no.4, pp.359–375, 1989.

[19] Y. Miyazawa and M. Ueno, “Computerized adaptive testing method using integer programming to minimize item exposure,” Advances in Artificial Intelligence: Selected Papers from the Annual Conference of Japanese Society of Artificial Intelligence (JSAI 2019), pp.105–113, Springer, 2020.

[20] S.W. Choi, S. Lim, and W. van der Linden, “Testdesign: an optimal test design approach to constructing fixed and adaptive tests in r,” Behaviormetrika, vol.49, pp.191–229, 2022.

[21] 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.

[22] 宮澤芳光，宇都雅輝，石井隆稔，植野真臣，“測定精度の偏り軽減のための等質適応型テストの提案，” 信学論（D），vol.J101-D, no.6, pp.909–920, June 2018.

[23] M. Ueno and Y. Miyazawa, “Uniform adaptive testing using max-

imum clique algorithm," 20th International Conference, AIED 2019, pp.482–493, 2019.
[24] K.T. Sun, Y.J. Chen, S.Y. Tsai, and C.F. Cheng, "Creating IRT-based parallel test forms using the genetic algorithm method," Applied measurement in education, vol.21, no.2, pp.141–161, 2008.
[25] P. Songmuang and M. Ueno, "Bees algorithm for construction of multiple test forms in e-testing," IEEE Trans. Learning Technologies, vol.4, no.3, pp.209–221, 2011.
[26] D.I. Belov and R.D. Armstrong, "A constraint programming approach to extract the maximum number of non-overlapping test forms," Computational Optimization and Applications, vol.33, no.2, pp.319–332, 2006.
[27] W. van der Linden, Linear Models for Optimal Test Design, Springer, 2005.
[28] T. Ishii, P. Songmuang, and M. Ueno, "Maximum clique algorithm and its approximation for uniform test form assembly," IEEE Trans. Learning Technologies, vol.7, no.1, pp.83–95, 2014.
[29] T. Ishii and M. Ueno, "Clique algorithm to minimize item exposure for uniform test forms assembly," Artificial Intelligence in Education 2015, Lecture Notes in Computer Science, vol.9112, pp.638–641, 2015.
[30] T. Ishii and M. Ueno, "Algorithm for uniform test assembly using a maximum clique problem and integer programming," Artificial Intelligence in Education 2015, Lecture Notes in Computer Science, vol.10331, pp.102–112, 2017.
[31] 石井隆稔，赤倉貴子，植野真臣，"複数等質テスト構成における整数計画問題を用いた最大クリーク探索の近似法，" 信学論（D），vol.J100-D, no.1, pp.47–59, Jan. 2017.
[32] 仁田善雄，齋藤宣彦，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.
[33] 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，vol.33, no.2, pp.54–57, 2014.
[34] F.M. Lord, Applications of Item Response Theory To Practical Testing Problems, Lawrence Erlbaum Associates, 1980.
[35] F.M. Lord and M.R. Novick, Statistical Theories of Mental Test Scores, Addison-Wesley, 1968.
[36] F.B. Baker and S.-H. Kim, eds., Item Response Theory: Parameter Estimation Techniques, CRC Press, 2004.
[37] W. van der Linden, ed., Handbook of Item Response Theory, Volume One: Models, Chapman and Hall/CRC, 2016.
[38] W. van der Linden, ed., Handbook of Item Response Theory, Volume Two: Statistical Tools, Chapman and Hall/CRC, 2016.
[39] M. Ueno and T. Okamoto, "Item response theory for peer assessment," Proc. IEEE Int. Conf. Advanced Learning Technologies, pp.554–558, 2008.
[40] M. Uto and M. Ueno, "Item response theory for peer assessment," IEEE Trans. Learning Technologies, vol.9, no.2, pp.157–170, 2016.
[41] 宇都雅輝，植野真臣，"ピアアセスメントの低次評価者母数をもつ項目反応理論，" 信学論（D），vol.J98-D, no.1, pp.3–16, Jan. 2015.
[42] A. Birnbaum, "Some latent trait models and their use in inferring an examinee's ability," Statistical theories of mental test scores, eds. by F.M. Lord and M.R. Novick, pp.397–479, Addison-Wesley, 1968.
[43] R.D. Bock and R.J. Mislevy, "Adaptive eap estimation of ability in a microcomputer environment," Applied Psychological Measurement, vol.6, no.4, pp.431–444, 1982.
[44] J. Mislevy, R, "Bayes modal estimation in item response models," Psychometrika, vol.51, no.2, pp.177–195, 1986.
[45] W. van der Linden, "Review of the shadow-test approach to adaptive testing," Behaviormetrika, vol.49, pp.169–190, 2022.
[46] 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，" 信学論（D），vol.J103-D, no.12, pp.881–893, Dec. 2020.

（2021 年 10 月 15 日受付，2022 年 7 月 10 日再受付，
9 月 26 日早期公開）

宮澤　芳光　（正員）

2014 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了．博士（工学），長岡技術科学大学，東京学芸大学を経て，2019 年より大学入試センター助教に着任，現在に至る．e テスティングの研究・開発に従事．

植野　真臣　（正員）

1994 東京工業大学大学院総合理工学研究科了．博士（工学），東京工業大学，千葉大学，長岡技術科学大学を経て，2006 より電気通信大学勤務，同大学教授に着任，現在に至る．人工知能，e テスティング，e ラーニング，ベイズ統計，ベイジアンネットワークなどの研究に従事．