測定精度の偏り軽減のための等質適応型テストの提案

宮澤　芳光[†a)]　宇都　雅輝[††b)]　石井　隆稔[†††c)]　植野　真臣[††d)]

A Proposal of Uniform Adaptive Testing for Reducing Measurement Error Bias

Yoshimitsu MIYAZAWA[†a)], Masaki UTO[††b)], Takatoshi ISHII[†††c)], and Maomi UENO[††d)]

あらまし　適応型テストとは，受検者の能力を逐次的に推定し，その能力に応じて測定精度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．この手法では，易しすぎる項目や難しすぎる項目の出題が減少するため，受検者の測定精度を減少させずに受検時間や項目数を軽減できる．しかし，従来の適応型テストでは，能力が同等な受検者には全く同じ項目群が出題される可能性が高く，実際に適応型テストを導入しているSPIやGTECの重要な問題になっている．本研究では，能力が同等な受検者であっても異なる項目を同一の測定精度を保ちつつ適応的に出題できる等質適応型テストを提案する．具体的には，提案手法では次のように項目出題を行う．1）2017年時点で最先端の複数等質テスト構成手法を用いて，異なる項目で構成されるが測定精度が等質になるような等質テストを多数構成する．2）受検者ごとに異なる等質テストを一つ割り当て，そのテスト内の項目集合をアイテムバンクとみなして適応型テストを実施する．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提案手法の有効性を示す．

キーワード　適応型テスト，eテスティング，複数等質テスト，項目反応理論

1. まえがき

近年，eテスティングと呼ばれる，Web上でテストを受検するCBT（Computer based testing）の実用化が注目されている．eテスティングには，任意のタイミングで何度でも同一の測定精度のテストを受検できる利点がある．そのため，テストの結果が受検者に大きな影響を及ぼすハイ・ステークステストを含む様々なテスト場面において，その導入が進みつつある[1]．

一方，CBTの出題方式の一つとして適応型テストが知られている[2]．適応型テストでは，受検者の解答のたびに能力を推定し，その能力推定値に対して情報量が最も高い項目を出題する．このように受検者の能力に応じて項目を逐次的に出題することで，受検者の測定精度を減少させることなく，出題項目数や受検時間を軽減できる利点がある．また，テスト終了基準を適切に設定することで，全ての受検者を同程度の精度で測定することも可能である[3]．しかし，従来の適応型テストには，以下の二つの問題がある．

（1）　同一の受検者が複数回受検した場合には，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SPIやGTECの重要な問題になっている．

（2）　能力が同等な受検者には同一の項目群が出題される可能性が高いため，アイテムバンクの全ての項目を広く一様に活用することができない．特定の項目群の過度な露出は，受検者への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得る[4]．

適応型テストを運用するためには，難易度や識別力などの特性が既知のテスト項目集合（アイテムバンクと呼ばれる）を事前に構築する必要があり，特にハイ・ステークステストにおいては，項目の作成に膨大な経済的・時間的コストを要するため，テスト実施におい

† 東京学芸大学，小金井市
Tokyo Gakugei University, 4–1–1 Nukuikita-machi, Koganei-shi, 184–8501 Japan
†† 電気通信大学，調布市
The University of Electro-Communications, 1–5–1 Chofugaoka, Chofu-shi, 182–8585 Japan
††† 東京理科大学，東京都
Tokyo University of Science, 1–3 Kagurazaka, Shinjuku-ku, Tokyo, 162–8601 Japan
a) E-mail: miyazawa@u-gakugei.ac.jp
b) E-mail: uto@is.uec.ac.jp
c) E-mail: t.ishii@rs.tus.ac.jp
d) E-mail: ueno@ai.is.uec.ac.jp
DOI:10.14923/transinfj.2017LEP0028

てはアイテムバンク内の全ての項目をできる限り活用することが望ましい．本研究では，これらの問題を解決する新しい適応型テストの枠組みを提案する．

これまでに，同一項目群出題の問題を解決するため，アイテムバンクから複数の等質テストを事前に生成する手法が提案されている [5], [6]．等質テストとは，1）テストの長さと 2）テスト間の測定精度が等質であるにもかかわらず，3）異なる項目から構成されている項目集合である．この等質テストの構成手法は，既に e テスティング技術の基幹技術として多数の研究がなされている [5]〜[9]．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が，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, [5], [6]），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されつつある [10], [11]．しかし，等質テスト構成手法を適応型テストへ適応した研究は見当たらず，それぞれの研究が独立に取り組まれている．

他方，項目露出数の偏りの問題を解決する適応型テスト手法として，制約付き適応型テストが提案されている [2], [12]〜[17]．代表的な制約付き適応型テスト手法である van der Linden ら [2], [14] の手法では，項目選択のたびに，項目露出数や回答所要時間などが所望の条件を満たすような項目集合を構成し，その項目集合の中から情報量が最も高い項目を出題する．また，露出数の偏りを軽減させる別のアプローチとして，アイテムバンク分割法（Item-Pool Partitioning）と呼ばれる手法が提案されている [18], [19]．この手法では，アイテムバンクを複数のグループに分割し，最も項目の露出数が少ないグループから情報量が高い項目を選択する．しかし，これらの手法では，露出数を制御し，受検者ごとに異なる項目を出題できるが，出題された項目に測定精度の偏りがあるためにテストの長さや測定精度について受検者間で大きな差が生じる．

この問題を解決するために，本研究では，能力が同等な受検者であっても，同一の測定精度を保ちつつ，異なる項目を適応的に出題できる等質適応型テストを提案する．具体的には，1）テストの長さと 2）テスト間の測定精度が等質であるにもかかわらず，3）異なる項目を出題できる適応型テストを提案する．

本研究の主なアイデアは，等質テスト構成技術の最近の発達により，多数の等質テストが生成できるようになったことを利用して，新しい適応型テストの枠組みを提案するというものである．提案手法の具体的なアプローチは以下のとおりである．1）2017 年時点で最大数の複数等質テストを構成できる Ishii et al. [6] の手法を用いて複数等質テストを構成する．2）受検者ごとに異なる等質テストを割り当て，そのテスト内の項目集合をアイテムバンクとみなして適応型テストを実施する．

提案手法では，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る．また，これにより出題される項目出題の多様性が向上するため，アイテムバンク内の項目を満遍なく利用でき，露出数の偏りも軽減されると期待できる．提案手法では，受検者ごとにできる限り異なる等質テストを割り当てる必要があるため，2017 年時点に現存する手法で最大の複数等質テストを構成できる Ishii et al. [6] の手法を用いることが主な提案である．

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提案手法の有効性を示す．

2. 項目反応理論

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．

項目反応理論は，コンピュータ・テスティングの普及とともに，近年様々な評価場面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 [20]〜[24]．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 [25]〜[27]．

（1） 測定精度の低い異質項目の影響を小さくして受検者の能力を推定できる．

（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，適応型テストや等質テスト自動構成といった現在のテスト運用の基礎をなす理論であり，情報処理技術者試験の一つである IT パスポート試験 [28] 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 [29] をはじめとする様々な評価場面で広く活用されている．一般に項目反応理論は，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 母数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．本研究でも 2PLM を用いるため，次節ではこのモデルについて詳細を示す．

2.1 2 母数ロジスティックモデル

2PLM では，能力値 $\theta \in (-\infty, \infty)$ をもつ受検者がテスト項目 $i \in \{1, \ldots, I\}$ に正答する確率を以下の式で表現する．

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

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 をとる変数を表す．また，$a_i \in [0, \infty)$ は項目 i の識別力パラメータ，$b_i \in (\infty, \infty)$ は項目 i の難易度パラメータと呼ばれる．

ここで，これらのパラメータの解釈を示すために，特性の異なる複数の項目に対する 2PLM の項目反応関数（item response function：IRF）を図 1 と図 2 に示した．図では，横軸が受検者の能力値，縦軸が項目への正答確率を表す．

図 1 では，識別力パラメータ a_i を三つの値に変えた場合の IRF を示した．識別力のパラメータ a_i が低い項目 1 は，IRF の傾きが小さく，能力値の変化に伴う正答確率の変化が少ないことが分かる．これは項目への正誤が能力値に依存しないことを意味しており，能力測定には不適切な項目と解釈できる．一方で，識別力パラメータ a_i が高い項目 3 では，能力 $\theta = 0$ 付近で正答確率が大きく変動していることが分かる．これは，この項目が，能力 $\theta = 0$ 付近の受検者の能力を精度良く識別できることを意味する．

また，図 2 には，難易度パラメータ b_i を三つの値に変えた場合の IRF を示した．難易度パラメータ b_i が高い項目 3 は，項目 1，2 より IRF が右にシフトしていることが分かる．その結果，能力値全域において正答確率が低なっており，正答が難しいという特性が表現されている．また，難易度パラメータ b_i は，能力値と等しいとき，すなわち，$b_j = \theta$ のとき，正答確率が 0.5 となり，その付近で項目反応関数の勾配が最も急になる．このことは，$\theta = b_i$ となる受検者の能力を精度良く評価できることを意味している．

2PLM を用いることにより，これらの項目特性を考慮して受検者の能力 θ を推定できる．

2.2 フィッシャー情報量

項目反応理論における能力推定の標準誤差は，フィッシャー情報量の逆数に漸近的に一致することが知られている [20]．そのため，項目反応理論では，測定精度を表す指標としてフィッシャー情報量が一般に利用される．

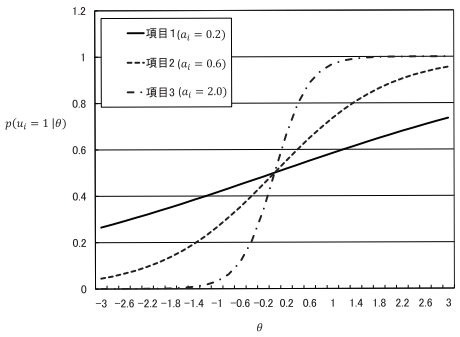

図 1 異なる識別力パラメータ a_i の項目に対する項目反応関数

Fig. 1 Item response function for item with different discrimination parameter a_i.

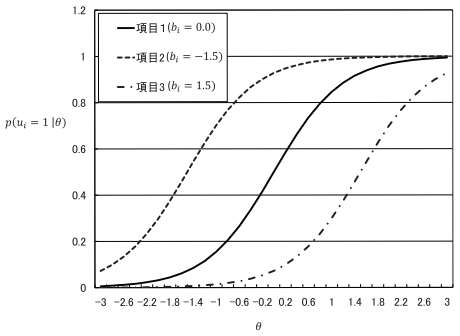

図 2 異なる難易度パラメータ b_i の項目に対する項目反応関数

Fig. 2 Item response function for item with different difficulty parameter b_i.

2PLM では，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 [30]．

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

ここで，

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

フィッシャー情報量 $I_i(\theta)$ が高い項目は，能力値 θ 付近で，受検者の能力をよく識別することを意味する．したがって，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．この考え方に基づき，フィッシャー情報量が高い項目を逐次的に出題するコンピュータテスティングの出題形式が適応型テストである．なお，受検者に出題したテストに含まれ

る項目群のフィッシャー情報量の総和をテスト情報量と呼び，テストの測定精度を表す．

2.3 適応型テスト

項目反応理論を用いた適応型テスト（CAT: Computer-Adaptive Testing）では，項目パラメータが既知の項目集合（以降，アイテムバンクと呼ぶ）を所与として，以下のような手順で項目の出題を行う．

（1） 受検者の能力値を初期化する．

（2） 能力値を所与として，フィッシャー情報量が最大となる項目をアイテムバンクから選択し，出題する．

（3） 項目に対する正誤データから受検者の能力推定値を更新する．

（4） 上記の手順 2 と 3 を，受検者の能力推定値の更新幅が一定値 ϵ 以下になるまで繰り返す．

このように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，項目を固定したテストと比べて，少ない出題項目数で高精度な能力推定が可能となる．しかし，従来の適応型テストでは，同一の能力をもつ受検者に対しては同一の項目が出題される傾向があり，同一受検者が複数回受検可能なテストでは実用化が難しい．更に，能力値は標準正規分布に従うため，平均値 $\theta = 0$ 付近に対して情報量が高い項目は高頻度で出題され，露出数が増加する．露出数の多い項目は，経年による運用において受検者に暴露される可能性が高まり，テストの信頼性低下要因となり得る [4]．

これらの問題を解決するために，本研究では，適応型テスト間でできる限り異なる項目を出題できる新たな適応型テスト手法を提案する．

3. 提案手法

提案手法の主なアイディアは，複数等質テスト構成手法を用いて，情報量や回答所要時間等の統計的性質は等質であるが，異なる項目から構成された等質テストを複数生成し，それらの中から一つの等質テストを各受検者に割り当て，テスト内の項目集合をアイテムバンクとみなして適応型テストを行うというものである．以降では，提案手法の詳細を説明する．

3.1 複数等質テスト構成手法

複数等質テストとは，情報量や所要時間等のテストの統計的性質が等質なテスト群を意味する．資格試験等では各回の性質が等質となることが求められるため，複数等質テストの自動構成手法は e テスティングの主要技術として広く研究が進められてきた [5]〜[9]．2017 年時点で最大数の複数等質テストを生成できる手法としては，Ishii et al. [6] が知られている．

Ishii et al. [6] の手法では，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）構造を抽出することで複数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構成可能な，重複条件以外の全てのテスト構成条件を満たす，可能テストを頂点とする．

エッジ：二つの可能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（テスト）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等質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，このグラフの最大クリークを抽出することで，最大数の複数等質テストを生成できる．

本研究では，受検者ごとにできる限り異なる等質テストを割り当てるため，非常に多くの複数等質テストを構成する必要がある．そこで，本研究では，等質テスト構成手法として Ishii et al. [6] の手法を用いる．

3.2 適応的項目出題

提案手法では，Ishii et al. [6] の手法により生成された複数等質テストの中から，各受検者に異なるテストを一つ割り当て，それに含まれる項目集合をアイテムバンクとみなして適応型テストを行う．このとき，受検者へのテストの割り当てはランダムに行い，一度受検者に割り当てたテストは候補から削除し，次の受検者には残りの候補からランダムに割り当てを行う．候補が空になったら，再度全てのテストを候補とする．また，適応型テストは **2.3** と同様の手順で行い，能力推定には EAP（expected a posteriori）[22] を用いる．

提案手法では，受検者ごとに異なる項目集合から項目を選択するため，能力が同等な受検者であっても異なる項目が出題できると期待できる．更に，出題される項目の多様性が向上するため，アイテムバンク内の項目を満遍なく利用することができ，露出数の偏りも軽減されると予想できる．

4. シミュレーション実験

本章では，提案手法の有効性を確認するために，シミュレーション実験による評価を行う．ここでは，通

常の適応型テスト（CAT），制約付き適応型テスト（CCAT と呼ぶ）[2]，アイテムバンク分割法（IPP と呼ぶ）[18] との比較を行う．また，複数等質テストの構成数が実験結果に与える影響を分析するために，複数等質テスト構成手法として van der Linden et al. が提案した Big-Shadow-Test（BST）[2] を用いた適応型テスト（BCAT と呼ぶ）とも比較する．

4.1 実験手順

シミュレーション実験の手順は以下のとおりである．

（1） 500，1000，2000 項目で構成されるアイテムバンクを生成する．これらの項目数は，本研究の実データ実験で使用する SPI のアイテムバンクの項目数を基準に決定した．具体的には，SPI の項目数が 978 であったことから項目数 1000 を基準とし，項目数の影響を評価するために半分の 500 と，2 倍の 2000 を採用した．このとき，各項目のパラメータ真値は $a_i \sim U(0,1)$，$b_i \sim N(0,1)$ からランダムに設定した．

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

（3） 受検者の能力推定値の $\hat{\theta} = 0$ に初期化した．

（4） 各手法を用いてアイテムバンクから項目を選択し，その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た．

（5） 回答履歴データから能力推定値 $\hat{\theta}$ を EAP 法により推定した．

（6） 手順 4 と 5 を，推定値の更新幅が ϵ 以下になるまで繰り返した．ϵ には，現実の適応型テストで一般に利用される 0.05 と 0.01 を用いた [2]．

（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回答履歴を用いて，次の指標に関する統計量を求めた．a）適応型テストの長さ，b）各項目の露出数，c）異なる受検者に同一でない項目を出題した割合（以降では異なる項目の割合と呼ぶ），d）測定精度の等質性

測定精度の等質性については，能力推定値の漸近的な標準誤差を受検者ごとに算出し，それらの値に関する標準偏差の逆数として評価する．なお，標準誤差の平均については，終了条件が全て同じ精度に設定してあるため，各手法の推定誤差はほぼ等しくなっている．また，本実験では，提案手法における等質テスト構成の制約として情報量の上限・下限を表 1 のように定めた．更に，等質テスト内の項目数の影響を分析するために，この値を 50，100 と変化させて実験を行った．ここで，等質テストの項目数 50 と 100 は，SPI において等質テストの項目数に 40〜50 程度が採用されていることと [31]，適応型テストではアイテムバンクの項目数として 100 項目程度以上が必要とされていること [3] を基準に決定した．また，複数等質テスト構成手法では，等質テスト間に多くの重複項目を許すほど生成できる等質テスト数が増加する [6]．そこで，この要因の影響を分析するために，複数等質テスト構成手法における重複項目数を変えて実験を行った．重複項目数は，アイテムバンクの項目数と等質テストの項目数に応じて設定し，表 2 のとおりにした．また，CCAT における露出数の最大値は，できる限り数多くの項目が利用される値を上記のシミュレーション実験で探索して決定した．具体的には，$\epsilon = 0.05$ では，アイテムバンクの項目数が 500 のとき 60，項目数 1000 のとき 30，項目数 2000 のとき 15 とし，$\epsilon = 0.01$ では，アイテムバンクの項目数が 500 のとき 150，項目数 1000 のとき 80，項目数 2000 のとき 50 とした．IPP の分割数は，受検者間でできる限り異なる項目が出題されるように設定した．具体的には，項目分割後の各項目グループの項目数が 5，10，20，30，40 となるように分割しながら，本節と同様のシミュレーション実験を行い，受検者に出題したテスト間で異なる項目の数の平均を算出した．ここでは，終了条件を $\epsilon = 0.05$，アイテムバンクの項目数を 1000 とした．結果を図 3 に示す．図から，各項目グループの項目数を 20 としたときに，この値が最大となったことが分かる．そこで，本実験では IPP の分割数を，分割後の各項目グループの項目数が 20 となるように決定した．

表 1 情報量の上限・下限
Table 1 Upper and lower bound of the Fisher information.

情報量関数（下限/上限）				
$\theta = -2.0$	$\theta = -2.0$	$\theta = -2.0$	$\theta = -2.0$	$\theta = -2.0$
2 / 2.4	3.2 / 3.6	3.2 / 3.6	3.2 / 3.6	2 / 2.4

表 2 重複項目数
Table 2 The numbers of overlapping items.

アイテムバンクの項目数		500	1000	2000
等質テストの項目数	50	0,10	0,10	0,5
	100	0,50	0,20	0,10

4.2 実験結果

適応型テストの終了条件 $\epsilon = 0.05$ としたときの実験結果を表 3 に，$\epsilon = 0.01$ としたときの結果を表 4 に示す．「適応型テスト手法」列の「提案手法 (x)」と「$BCAT(x)$」は，等質テストの項目数を x としたときの結果を表す．

表 3　シミュレーションから生成したアイテムバンクを用いた結果（$\epsilon = 0.05$）
Table 3　Experimental results using simulated item pools ($\epsilon = 0.05$).

アイテムバンクの項目数	適応型テスト手法	重複項目数	等質テストの構成数	テストの長さ		測定精度の等質性	異なる項目の割合		露出数		
				平均	標準偏差		平均	標準偏差	最大	平均	標準偏差
500	CAT	-	-	19.9	2.17	31.3	0.599	0.291	1000	39.8	120.2
	CCAT	-	-	24.1	6.46	5.2	0.942	0.158	60	48.3	22.9
	IPP	-	-	14.5	3.87	13.2	0.849	0.106	135	28.9	40.8
	BCAT(50)	0	8	13.7	3.12	20.0	0.890	0.265	167	27.4	47.6
		10	9	11.8	1.62	18.2	0.795	0.200	525	23.5	64.2
	提案手法 (50)	0	5	13.4	2.23	17.5	0.858	0.300	200	26.8	55.6
		10	136	13.4	2.43	16.9	0.928	0.073	168	26.8	35.2
	BCAT(100)	0	4	16.0	2.91	23.3	0.814	0.310	334	31.9	69.6
		50	4	14.6	2.03	20.8	0.720	0.227	751	29.2	84.6
	提案手法 (100)	0	3	14.7	2.73	15.4	0.787	0.335	334	29.4	72.9
		50	999	15.6	2.48	18.5	0.892	0.093	237	31.2	49.0
1000	CAT	-	-	21.4	2.28	29.4	0.663	0.282	1000	21.4	82.0
	CCAT	-	-	28.4	8.60	1.1	0.971	0.114	30	28.4	6.7
	IPP	-	-	15.7	4.07	11.5	0.921	0.076	79	15.7	22.3
	BCAT(50)	0	15	14.0	3.93	15.4	0.941	0.204	91	14.0	25.6
		10	19	11.8	2.06	18.9	0.892	0.165	372	11.8	33.8
	提案手法 (50)	0	9	14.0	2.61	13.9	0.921	0.235	112	14.0	30.4
		10	8758	13.9	2.53	16.9	0.963	0.052	129	13.9	18.5
	BCAT(100)	0	7	16.6	4.20	10.5	0.899	0.256	167	16.6	37.6
		20	9	13.5	2.15	20.0	0.864	0.192	439	13.5	40.3
	提案手法 (100)	0	4	15.6	2.42	10.4	0.873	0.281	200	15.6	41.5
		20	7092	15.7	2.45	19.2	0.941	0.066	165	15.7	26.2
2000	CAT	-	-	22.3	2.12	43.5	0.694	0.269	1000	11.20	57.6
	CCAT	-	-	28.6	11.51	0.4	0.986	0.079	15	14.30	3.20
	IPP	-	-	15.7	4.03	11.2	0.961	0.062	45	7.80	11.1
	BCAT(50)	0	32	13.8	2.91	31.2	0.975	0.143	32	6.90	11.4
		5	39	12.5	2.35	32.2	0.952	0.136	153	6.26	16.2
	提案手法 (50)	0	12	14.7	2.52	31.6	0.930	0.236	84	7.33	21.5
		5	4272	14.7	2.32	30.8	0.979	0.038	42	7.34	10.3
	BCAT(100)	0	16	16.3	2.90	27.9	0.953	0.189	63	8.15	18.0
		10	19	15.1	2.16	32.2	0.921	0.180	201	7.53	23.3
	提案手法 (100)	0	0	17.5	2.40	42.8	0.805	0.348	250	8.75	40.3
		10	760	16.7	2.30	39.5	0.962	0.049	69	8.34	15.8

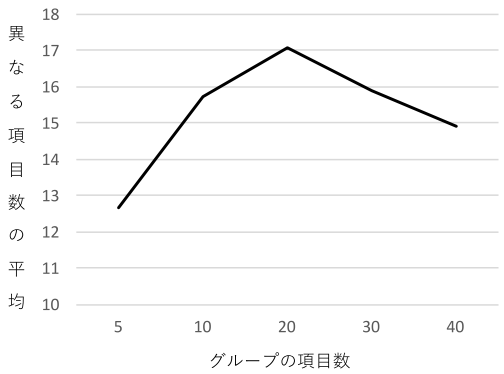

図 3　IPP における異なる項目数
Fig. 3　The number of different items in IPP.

まず，提案手法と BCAT の等質テストの構成数を比較すると，テスト間に重複項目を許したとき，Ishii et al. [6] を採用している提案手法が大幅に多くの等質テストを構成できていることが分かる．

テストの長さについては，CAT，CCAT，IPP と比較して，提案手法が短かったことが確認できる．θ の初期値が受検者の真の能力値から遠い場合，初期値に対して情報量が高い項目を出題するより，情報量が低い項目を出題した方が能力推定の収束が早くなることが知られている [2]．提案手法は，項目数を等質条件で制限するため，ある推定値だけに極端に高い情報量を示す項目を出題しない性質をもつ．この性質により，テストの長さが短縮され，結果として項目の露出数を減少できる．テストの長さの標準偏差も，CAT と CCAT と IPP と比較して，提案手法が小さいことが分かる．提案手法では，情報量が等質である項目集合から項目を選択しているため，能力推定値の収束にかかる項目数を等質にすることができる．BCAT では，一部の条件下で提案手法よりテスト長が短くなっていることが確認できる．しかし，このような場合，BCAT では，項目露出の最大値が極端に大きくなって

表 4　シミュレーションから生成したアイテムバンクを用いた結果（$\epsilon = 0.01$）
Table 4　Experimental results using simulated item pools ($\epsilon = 0.01$).

アイテムバンクの項目数	適応型テスト手法	重複項目数	等質テストの構成数	テストの長さ		測定精度の等質性	異なる項目の割合		露出数		
				平均	標準偏差		平均	標準偏差	最大	平均	標準偏差
500	CAT	-	-	65.1	7.29	45.5	0.482	0.304	1000	130.1	223.5
	CCAT	-	-	73.2	16.07	1.7	0.852	0.257	150	146.4	19.4
	IPP	-	-	47.4	10.14	29.4	0.630	0.162	371	94.8	117.9
	BCAT(50)	0	8	29.2	9.18	35.1	0.855	0.331	167	58.5	71.0
		10	9	22.9	5.02	32.5	0.758	0.231	670	45.8	94.3
	提案手法 (50)	0	5	33.9	7.82	33.28	0.824	0.357	200	67.9	84.7
		10	136	34.3	8.28	30.94	0.907	0.051	227	68.6	40.1
	BCAT(100)	0	4	40.5	9.65	32.5	0.743	0.379	334	81.0	119.8
		50	4	29.3	4.58	30.29	0.658	0.249	915	58.7	128.9
	提案手法 (100)	0	3	39.5	7.35	21.32	0.726	0.398	334	79.1	123.1
		50	999	40.2	6.86	31.72	0.842	0.063	326	80.5	78.9
1000	CAT	-	-	70.4	7.11	45.5	0.568	0.319	1000	70.4	158.1
	CCAT	-	-	79.5	26.93	0.1	0.922	0.198	80	79.5	6.2
	IPP	-	-	56.6	10.74	27.8	0.765	0.120	246	56.6	74.4
	BCAT(50)	0	15	31.8	10.06	29.24	0.919	0.260	91	31.8	39.7
		10	19	25.1	4.54	45.9	0.828	0.199	588	25.1	60.5
	提案手法 (50)	0	9	31.9	7.13	26.52	0.900	0.286	112	31.9	46.3
		10	8758	35.1	8.19	39.61	0.951	0.038	154	35.1	21.8
	BCAT(100)	0	7	39.2	14.49	17.17	0.868	0.310	167	39.2	59.9
		20	9	31.8	4.71	39.61	0.791	0.222	555	31.8	74.9
	提案手法 (100)	0	4	39.4	9.44	15.81	0.831	0.345	200	39.4	70.5
		20	7092	42.8	8.01	39.61	0.916	0.045	226	42.8	41.8
2000	CAT	-	-	80.2	6.98	71.4	0.657	0.313	1000	40.1	109.3
	CCAT	-	-	100	10.08	3.3	0.951	0.149	46	50.0	11.7
	IPP	-	-	58.1	13.08	28.6	0.870	0.082	137	29.1	39.0
	BCAT(50)	0	32	36.6	9.81	124.9	0.968	0.169	32	18.3	14.1
		5	39	30.6	8.31	107.7	0.882	0.160	267	15.3	29.7
	提案手法 (50)	0	12	36.2	6.70	54.6	0.945	0.218	84	18.1	32.7
		5	4272	36.5	6.42	53.4	0.977	0.027	43	18.2	10.1
	BCAT(100)	0	16	48.0	14.41	84.8	0.944	0.209	63	24.0	27.9
		10	19	37.8	6.11	107.0	0.848	0.179	369	18.9	41.9
	提案手法 (100)	0	4	45.4	3.80	145.8	0.901	0.270	250	22.7	69.2
		10	760	45.7	5.60	79.8	0.955	0.036	70	22.8	23.0

おり，特定の項目が過度に出題されている．特定の項目の過度な露出は，受検者への項目内容の暴露につながり，テストの信頼性の低下が知られているため [4]，これを回避することが本研究の目的の一つである．提案手法では，多数の等質テストを構成することによって，従来の適応型テストよりも短いテスト長で，項目の過度な露出を軽減できたことが分かる．

測定精度の等質性は，情報量が高い一部の項目群を繰り返し出題している CAT が最も高かった．一方，CCAT や IPP は，露出数を制御し，受検者ごとに異なる項目を出題しているが，出題された項目に測定精度の偏りがあるため，測定精度の等質性が低い結果となった．提案手法と BCAT は，測定精度が等質な項目集合から項目を選択しているため，測定精度の等質性は CAT に次いで高く，同一の測定精度を保てたことが分かる．

受検者に出題したテスト間で同一ではない項目の割合を表す「異なる項目の割合」を比較すると，CCAT が最も高い値を示したことが分かる．**4.** で述べたように，CCAT ではアイテムバンク内の項目をできる限り多く使うように露出数を調整したため，このような結果となったと解釈できる．一方，提案手法と BCAT は，等質テストの構成時に重複項目を設定するため，CCAT に次いで異なる項目の割合が高くなっている．具体的には，8 割以上の異なる項目を出題していることが分かる．提案手法と BCAT を比較すると，重複項目を許した提案手法が BCAT より異なる項目の割合が高いことが分かる．これは，提案手法がより多くの複数等質テストを構成できたため，受検者ごとに異なる項目を出題できたことを意味しており，Ishii et al. [6] の手法を採用することの優位性を示している．また，異なる項目の割合の標準偏差についても，重複項目を許した提案手法が最も低く，どの受検者にも，同等の割合で異なる項目を出題できている．以上から，

表 5　実データを用いた結果（$\epsilon = 0.05$）
Table 5　Experimental results using an actual item pool ($\epsilon = 0.05$).

シミュレーション回数	適応型テスト手法	重複項目数	等質テストの構成数	テストの長さ		測定精度の等質性	異なる項目の割合		露出数		
				平均	標準偏差		平均	標準偏差	最大	平均	標準偏差
1000	CAT	-	-	18.8	3.44	20.8	0.564	0.302	1000	15.2	78.0
	CCAT	-	-	26.4	6.38	5.1	0.971	0.109	30	27.0	8.6
	IPP	-	-	17.2	4.93	7.5	0.955	0.072	65	10.9	15.8
	BCAT(50)	0	17	16.2	3.76	19.6	0.958	0.175	59	16.6	20.5
		10	18	14.9	3.29	19.6	0.901	0.170	314	15.3	35.6
	提案手法 (50)	0	7	15.7	3.38	20.0	0.900	0.259	143	16.1	36.6
		10	8669	15.7	3.21	20.0	0.957	0.053	130	16.1	20.9
	BCAT(100)	0	8	17.3	2.99	25.6	0.921	0.226	125	17.7	32.9
		10	8	16.7	3.40	22.2	0.852	0.213	347	17.0	46.6
	提案手法 (100)	0	2	17.8	2.99	24.4	0.688	0.359	500	18.2	73.0
		10	7088	17.2	3.16	23.8	0.930	0.066	177	17.6	30.4
5000	CAT	-	-	14.9	3.43	22.2	0.555	0.300	5000	76.3	392.7
	CCAT	-	-	26.7	6.14	10.1	0.963	0.122	190	136.4	83.3
	IPP	-	-	17.5	5.00	7.5	0.902	0.085	747	89.6	190.0
	BCAT(50)	0	17	16.3	3.66	18.9	0.958	0.176	295	83.3	101.8
		10	18	16.3	3.67	18.9	0.958	0.176	1649	103.3	102.1
	提案手法 (50)	0	7	15.6	3.38	18.4	0.900	0.258	715	79.6	180.8
		10	8669	15.5	3.39	18.4	0.960	0.159	715	79.5	181.4
	BCAT(100)	0	8	17.4	3.05	24.3	0.921	0.227	625	89.0	164.3
		10	8	17.3	3.00	24.3	0.921	0.227	625	88.7	164.3
	提案手法 (100)	0	2	17.6	2.96	23.3	0.689	0.358	2500	90.2	362.5
		10	7088	17.7	2.93	22.6	0.969	0.258	830	90.3	363.1
10000	CAT	-	-	15.0	3.47	21.3	0.557	0.300	10000	153.2	785.7
	CCAT	-	-	26.5	6.15	10.2	0.962	0.122	382	271.2	168.5
	IPP	-	-	17.5	5.05	7.6	0.901	0.085	1509	179.3	380.5
	BCAT(50)	0	17	16.4	3.68	19.2	0.958	0.176	589	167.2	203.4
		10	18	16.3	3.67	19.1	0.958	0.176	3400	167.0	203.6
	提案手法 (50)	0	7	16.5	3.33	18.2	0.900	0.258	1429	158.9	361.5
		10	8669	16.6	3.36	20.6	0.950	0.158	1429	159.2	361.4
	BCAT(100)	0	8	17.3	2.99	23.8	0.922	0.226	1250	177.3	326.0
		10	8	17.4	2.97	24.0	0.921	0.226	1250	177.7	327.4
	提案手法 (100)	0	2	16.6	2.89	22.1	0.690	0.358	5000	180.3	724.9
		10	7088	16.7	2.97	22.3	0.960	0.158	1870	180.6	724.6

提案手法では，等質テスト構成における重複項目数を許し，等質テストの構成数を増加させることで，アイテムバンク内のより多くの項目を出題することができ，受検者に出題される項目のパターンも増加することが示された．

露出数については，CCAT を用いた場合に最大値が最も少なかった．異なる項目の割合と同様に，CCAT は露出数の最大値を直接制約できる手法であり，本実験ではアイテムバンク内の項目をできる限り多く使うように露出数の設定を行ったため，このような結果となったと解釈できる．また，提案手法と BCAT を比較すると，上述のように，BCAT では項目重複を許したときに露出数の最大値が極端に大きくなるのに対し，提案手法ではこれを減少できている．提案手法では，多数の等質テストを構成できたため，BCAT と比べて項目露出の偏りを減少できたことが分かる．一方，露出数の平均値は，提案手法，BCAT，IPP が同程度に最も低いことが確認できる．CCAT は，露出数の最大値のみを直接制約しており，その露出数の偏りは制御できていない．提案手法，BCAT，IPP は，アイテムバンク内の項目を広く一様に活用し，露出数の偏りを減らしていることが分かる．

以上の実験結果から，提案手法の性能について次の点が明らかとなった．提案手法では，Ishii et al. [6] の手法による等質テスト構成を用い，更に，項目重複を許して等質テストの構成数を増加させることで，受検者に出題される項目のパターンを大幅に増加させることができ，同時にアイテムバンク内の項目をより広く一様に利用できる．加えて，提案手法は，受検者に異なる項目を出題しているにもかかわらず，テストの長さや測定精度を等質にすることができた．

5. 実データを用いた評価実験

本章では，実データを用いて提案手法の有効性を

表 6　実データを用いた結果（$\epsilon = 0.01$）
Table 6　Experimental results using an actual item pool ($\epsilon = 0.01$).

シミュレーション回数	適応型テスト手法	重複項目数	等質テストの構成数	テストの長さ		測定精度の等質性	異なる項目の割合		露出数		
				平均	標準偏差		平均	標準偏差	最大	平均	標準偏差
1000	CAT	-	-	65.5	10.72	62.5	0.512	0.298	1000	67.0	163.2
	CCAT	-	-	87.9	18.22	3.1	0.911	0.195	90	89.9	1.2
	IPP	-	-	61.7	16.05	19.2	0.852	0.108	179	39.7	54.4
	BCAT(50)	0	17	41.2	5.11	25.0	0.947	0.216	59	42.1	22.6
		10	18	37.3	6.42	23.8	0.886	0.199	331	38.2	54.2
	提案手法 (50)	0	7	41.2	4.51	23.3	0.870	0.320	143	42.1	60.9
		10	8669	40.7	4.70	23.3	0.950	0.034	136	41.6	19.3
	BCAT(100)	0	8	55.5	8.36	31.3	0.899	0.273	125	56.8	50.3
		10	8	50.3	11.23	30.3	0.825	0.251	443	51.4	79.0
	提案手法 (100)	0	2	54.8	8.84	30.3	0.597	0.417	500	56.0	139.3
		10	7088	54.3	7.67	32.3	0.912	0.040	179	55.5	42.6
5000	CAT	-	-	65.5	10.8	62.4	0.519	0.300	5000	334.6	819.8
	CCAT	-	-	87.9	15.27	28.4	0.885	0.224	766	449.6	349.3
	IPP	-	-	62.7	13.57	21.0	0.719	0.117	2225	320.7	593.1
	BCAT(50)	0	17	41.2	4.94	26.2	0.846	0.218	295	210.7	113.2
		10	18	41.3	4.97	26.1	0.846	0.218	1721	211.1	112.9
	提案手法 (50)	0	7	41.2	4.45	26.3	0.869	0.322	715	210.5	305.0
		10	8669	41.2	4.45	26.4	0.869	0.222	715	210.8	305.0
	BCAT(100)	0	8	55.3	8.30	34.4	0.798	0.275	625	283.0	252.9
		10	8	55.5	8.39	34.2	0.798	0.275	625	284.0	253.3
	提案手法 (100)	0	2	54.3	8.53	34.5	0.595	0.418	2500	277.4	694.3
		10	7088	54.4	8.54	34.4	0.874	0.319	670	278.2	695.9
10000	CAT	-	-	65.3	10.65	61.0	0.516	0.300	10000	667.3	1642.7
	CCAT	-	-	87.9	15.43	28.3	0.895	0.224	1533	898.7	700.1
	IPP	-	-	62.7	13.60	20.6	0.720	0.118	4406	640.7	1184.0
	BCAT(50)	0	17	41.6	4.93	27.8	0.846	0.218	589	421.7	226.3
		10	18	41.5	4.92	27.8	0.846	0.218	3401	421.9	226.3
	提案手法 (50)	0	7	41.3	4.41	28.3	0.869	0.322	1429	421.9	610.6
		10	8669	41.3	4.42	28.4	0.869	0.322	1429	422.2	610.7
	BCAT(100)	0	8	55.4	8.36	35.0	0.798	0.275	1250	566.1	506.1
		10	8	55.4	8.32	34.9	0.798	0.275	1250	566.7	506.3
	提案手法 (100)	0	2	54.4	8.46	35.7	0.595	0.418	5000	556.2	1391.1
		10	7088	54.2	8.49	35.5	0.875	0.048	1690	554.6	1388.4

評価する．ここでは，リクルート（株）で開発された SPI [31] 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を行った．アイテムバンクの項目数は 978 であった．また，ここでは，本実験において受検者数に対応する「シミュレーションの繰り返し数」の影響を分析するために，繰り返し数を変えながら実験を行った．具体的には，表 3，表 4 から，アイテムバンクの項目数が 1000 程度のときに Ishii et al. [6] の手法で構成される複数等質テストの構成数の最大が約 10000 であったことから，繰り返しの最大数を 10000 回とし，5000 回，1000 回についても評価を行った．

適応型テストの終了条件 $\epsilon = 0.05$ としたときの実験結果を表 5 に，$\epsilon = 0.01$ としたときの結果を表 6 に示す．また，図 4 に，$\epsilon = 0.01$ として CAT と CCAT，IPP，BCAT(50)，提案手法 (50) における真の能力値とテストの長さの散布図を示す．

表 5，表 6，図 4 から，以下の特徴が示された．1）複数等質テストの構成数は，テスト間に重複項目を許したとき，BCAT よりも提案手法の方が多い．また，複数等質テストの構成数は重複項目数を増やすほど増加する．2）適応型テストの終了条件 ϵ が 0.05 のときは，提案手法と BCAT，IPP，CAT が同程度に最もテストの長さが短く，0.01 のときは，提案手法と BCAT が同程度に最もテストの長さが短い．また，テストの長さの標準偏差は，提案手法と BCAT が最も低く，受検者へのテストの長さのばらつきが少ないことが分かる．図 4 の真の能力値とテストの長さの散布図からも同様の傾向が読み取れる．3）測定精度の等質性については，適応型テストに次いで提案手法と BCAT が最も高く，同一の測定精度を保てたことが分かる．4）異なる項目の割合は，CCAT が最も高い値を示しており，重複項目を設定した提案手法と BCAT も次いで高い性能を示している．また，提案手法と BCAT を

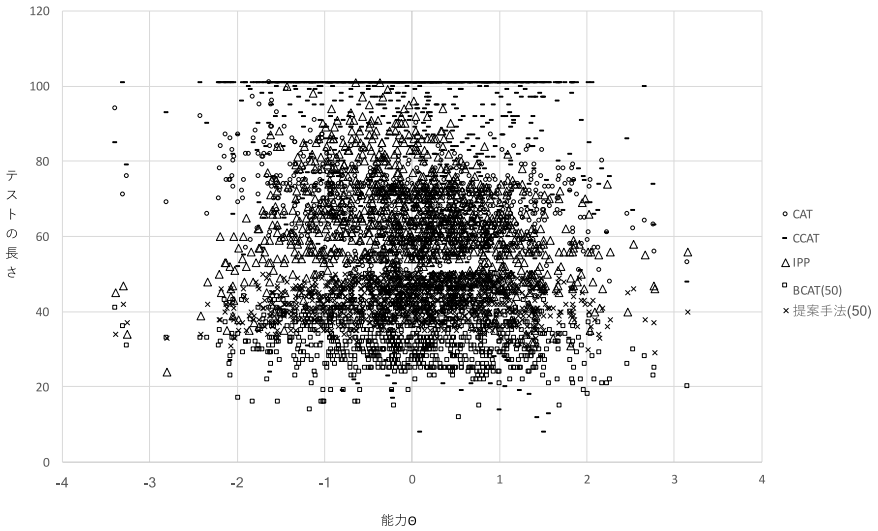

図 4　真の能力値とテストの長さの散布図
Fig. 4　Scatter plot for true ability values and test lengths.

比較すると，提案手法の方が BCAT より異なる項目の割合が高い傾向が読み取れる．これは，提案手法では，Ishii et al. [6] を用いたことでより多くの等質テストを構成でき，受検者に異なる項目集合を割り当てることができたためと解釈できる．また，提案手法は標準偏差が低いことから安定して異なる項目を出題していることも分かる．5）露出数の最大値については CCAT が最も少なく，平均は，IPP や BCAT，提案手法が同程度に少ない傾向が確認できる．BCAT は，項目重複が少ないとき等質テストの数が少ないため，特定の項目が過度に露出する傾向がある．これに対し，提案手法では，多数の等質テストを構成することにより過度な露出項目を減少させることができる．6）以上の傾向は，シミュレーションの繰り返し数に依存しないことが確認できる．上述のとおり，シミュレーションの繰り返し数は受検者数に対応するため，上記の傾向が受検者数に依らない結果であることが分かる．

6. むすび

本研究では，能力が同等な受検者であっても異なる項目を同一の測定精度を保ちつつ適応的に出題できる適応型テストを実現するために，等質適応型テストを提案した．具体的には，最新の複数等質テスト構成技術を用いて等質テストを多数構成し，各受検者に異なるテストを一つ割り当て，そのテストに含まれる項目集合をアイテムバンクとみなして適応型テストを行う手法を提案した．更に，シミュレーション実験と実データ実験により，提案手法の利点として以下の点が確認できた．

（1） テストの長さや測定精度のばらつきを減少させることができ，テスト間の等質性を保てた．

（2） 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，受検者ごとに異なる項目群を出題できる．特に，2017 年時点で最大数の複数等質テストを構成できる Ishii et al. [6] を用いて大規模な数の等質テストを構成することで，この優位性がより顕著になることが示された．

（3） 項目露出数を直接制約できる CCAT には劣るものの，提案手法では出題される項目の多様性が向上するため，通常の CAT や IPP，BCAT に比べてアイテムバンク内のより多くの項目を出題でき，露出数の偏りも軽減できた．

今後は，アイテムバンクの有効活用の観点において CCAT に劣る問題を解決するために，提案手法にも露出数を制約できるアプローチを統合することも検討する．

謝辞　本研究は科研費（15K21007），及び，東京学芸大学「日本における次世代対応型教育モデルの研究開発」〔文部科学省特別経費（プロジェクト分）〕における「OECD との共同による次世代指導モデルの研究開発プロジェクト」の助成を受けたものである．

文　献

[1] 植野真臣，永岡慶三，e テスティング，培風館，2009.

[2] W.J. van der Linden and C.A.W. Glas, eds., Elements of Adaptive Testing, Springer, 2010.

[3] 池田　央，柳井晴夫，藤田恵璽，繁桝算男，教育測定学〈下巻〉，学習評価研究所，1992.

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

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] W.J. van der Linden, Linear Models for Optimal Test Design, Springer, 2005.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] W.J. 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.

[15] J.B. Sympson and R.D. Hetter, “Controlling item-exposure rates in computerized adaptive testing,” Proc. 27th Annual Meeting of the Military Testing Association, vol.17, pp.973–977, 1985.

[16] 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.

[17] M.L. Stocking and C. Lewis, Methods of controlling the exposure of items in CAT, pp.163–182, Springer Netherlands, 2000.

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

[20] F.M. Lord, Applications of Item Response Theory to Practical Testing Problems, Lawrence Erlbaum Associates, 1980.

[21] F.M. Lord and M.R. Novick, Statistical Theories of Mental Test Scores, Addison-Wesley, 1968.

[22] F.B. Baker and S.H. Kim, eds., Item Response Theory: Parameter Estimation Techniques, CRC Press, 2004.

[23] W.J. van der Linden, ed., Handbook of Item Response Theory, Volume One: Models, Chapman and Hall/CRC, 2016.

[24] W.J. van der Linden, ed., Handbook of Item Response Theory, Volume Two: Statistical Tools, Chapman and Hall/CRC, 2016.

[25] M. Ueno and T. Okamoto, “Item response theory for peer assessment,” Proc. IEEE International Conference on Advanced Learning Technologies, pp.554–558, 2008.

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

[27] 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.

[28] 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/.

[29] 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 13 版（平成 27 年度）,” http://www.cato.umin.jp/e-book/13/index.html.

[30] 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.

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

（平成 29 年 9 月 29 日受付，30 年 1 月 14 日再受付，
3 月 6 日早期公開）

宮澤　芳光　（正員）

2014 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 年より東京学芸大学助教に着任，現在に至る．e テスティングの研究・開発に従事．

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．

石井　隆稔　（正員）

2008 年電気通信大学・情報通信工卒．2011 年電気通信大学大学院情報システム学研究科博士前期課程修了．2014 年同大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）．2014 年首都大東京システムデザイン学部特任助教．2016 年東京理科大学工学部助教，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 年東京工業大学大学院総合理工学研究科修了．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 年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．