情報処理学会論文誌　数理モデル化と応用　Vol.7　No.2　17–26 (Nov. 2014)

EMタイプIRTによる不完全マトリクスの完全化とその応用

作村 建紀[1,a)]　徳永 正和[2,b)]　廣瀬 英雄[3,c)]

受付日 2013年11月13日，再受付日 2013年12月31日，
採録日 2014年3月3日

概要：項目反応理論（IRT）は，あるテストの問題を受験者が解答したときの解答パターンのマトリクスから，テストの問題項目特性値および受験者能力評価値を推定することができる．IRT は通常，すべての問題に全員が解答している完全マトリクスを用いるが，ここでは未解答部分を含む不完全マトリクスに IRT が適用できる EM タイプ IRT を提案する．まず，テストの問題項目特性および受験者能力が分かっているときに不完全マトリクスを模擬したデータを用いて，提案法が元のパラメータを再現できることを確認し，次に実際に行ったテストに対して提案法による予測を行った．さらに，提案法を不完全マトリクスの予測法の１つであるマトリクス分解法と比較した．その結果，提案法はマトリクス分解法よりも良い予測精度を出すことが分かった．

キーワード：項目反応理論，不完全マトリクス，EM タイプ IRT，適応型試験，マトリクス分解法，キャリブレーション

Making up the Complete Matrix from the Incomplete Matrix Using the EM-type IRT and Its Application

Takenori Sakumura[1,a)]　Masakazu Tokunaga[2,b)]　Hideo Hirose[3,c)]

Received: November 13, 2013, Revised: December 31, 2013,
Accepted: March 3, 2014

Abstract: The item response theory (IRT) can estimate the parameters for the items of problems and the abilities of students by using the matrix of answering pattern. The conventional IRT is applied to the complete matrix in which all the students answered to all the problems. In this paper, we have proposed the EM-type IRT method which allows the vacant elements in the matrix, i.e., the incomplete matrix. First, we showed the validity of the proposed method by estimating the parameters using the incomplete matrix in which all the parameters are known in advance. Then, we applied the method to the real data case, and compared the estimated results with those by the matrix decomposition method (MD), where the MD is, in general, applied to the prediction for the incomplete matrix. The proposed method provides the better prediction accuracy than the MD does.

Keywords: item response theory, incomplete matrix, EM-type IRT, adaptive test, matrix decomposition method, calibration

1. はじめに

項目反応理論（IRT）は，あるテストの問題を受験者が解答したときの解答パターンのマトリクスから，テストの問題項目特性および受験者能力評価を推定することができる [4], [6], [7], [22]．通常はすべての問題に全員が解答している完全マトリクスを用いる．IRT にはいくつかのツールが考案されている．最も一般的なツールに文献 [19] があるが，IRT の専門知識が必要であり取り扱いが難しいという

[1] 中央大学
Chuo University, Bunkyo, Tokyo 112–8551, Japan
[2] 株式会社 KIS
KIS Co. Ltd., Minato, Tokyo 108–0074, Japan
[3] 九州工業大学
Kyushu Institute of Technology, Iizuka, Fukuoka 820–8502, Japan
[a)] sakumura@indsys.chuo-u.ac.jp
[b)] m-tokunaga@kis.co.jp
[c)] hirose@ces.kyutech.ac.jp

問題があった．そこで，文献 [17] では，0/1 表記で表したテスト結果の EXCEL ファイルを単にドラッグ・アンド・ドロップすることで，IRT のパラメータ推定を可能にするシステムを開発した．また，少ない問題数を用いたときの能力評価の精度についての研究も行われている [10]．

問題のパラメータがあらかじめ分かっていれば，IRT を利用した適応型オンライン能力評価システム（以下，適応型システムと呼ぶ）を構築することができる [13], [18]．これは，受験者の能力に合わせたレベルの問題を自動的に出題するシステムである．出題される問題の項目群を項目バンクという．受験者の解答がシステムに取り込まれるごとにその受験者の能力を逐次評価することができる．出題する問題は受験者の能力に最適に合わせることができるため，より少ない問題で能力評価の精度を高めることができる．適応型システムには昇降法やストレス・ストレングスモデルと昇降法を組み合わせた方法 [8] もあるが，いずれも出題される問題の項目特性は分かっている必要があった．そのためには，事前に予備テストを実施するのが一般的である．予備テストとは，項目バンクに項目を登録する際に実施される試験である．

予備テストを受験する受験者集団は適応型システムを受験する集団とは異なるため，適応型システムを受験する受験者が多くなれば，予備テストによって準備していた項目特性が適応型システムを受験する受験者から得られる項目特性とずれてくる可能性があるので，項目特性の再構成（キャリブレーション）が必要になってくる．このとき，通常の適応型テストでは受験した問題だけで受験者の能力推定が可能であったものが，項目特性まで推定しなければならなくなるため，そのままでは不完全マトリクスに対応できない推定法を変更する必要が出てくる．本論文では，このように不完全マトリクスの解答パターンから項目特性を推定する新しい方法として EM タイプ IRT を提案する．この方法は，データの背後にロジスティックモデルの確率構造を仮定し，不完全マトリクスでの観測された要素の値を用いて観測されていない空要素の値を確率的に予測するものである．

不完全マトリクスから完全マトリクスを推定する方法の 1 つに，マトリクス分解法（MD; matrix decomposition method）がある [9], [21]．これは，データの背後に確率構造を仮定しないノンパラメトリックな方法であり，推薦システムのようなものに用いられている．しかしながら，受験者の資質がある程度予測できて，問題の解答パターンもある程度想定される確率分布の下で変動すると仮定できるような場合，背後に確率分布の構造を仮定した方が推定精度が良くなることも考えられる．つまり，能力評価試験のような受験者の特性が強く反映される不完全マトリクスの場合，EM タイプ IRT は MD よりも有効に働く可能性がある．

ここでは，まず，テストの問題項目特性および受験者能力が分かっているときに不完全マトリクスを模擬したデータを用いて，提案法が元のパラメータを再現できることを確認し，次に実際に行ったテストに対して提案法による予測を行った．推定されたパラメータに現実性を与えるため，ここでは実際に大学内で行った数学のテストから得られた解答パターンを利用した．さらに，提案法を不完全マトリクスの予測法の 1 つであるマトリクス分解法と比較した．

2. IRT

IRT では，各項目 j に対する受験者 i の評価確率 $P_j(\theta_i; a_j, b_j)$ が 2 パラメータロジスティック分布に従っていると仮定する．このとき，

$$P_j(\theta_i; a_j, b_j) = \frac{1}{1 + \exp\{-1.7a_j(\theta_i - b_j)\}} \tag{1}$$

と表現される．a_j，b_j は項目 j の識別力と困難度を，θ_i は受験者 i の能力を表す．受験者 $i = 1, 2, \ldots, N$ と項目 $j = 1, 2, \ldots, n$ に対する尤度 L は

$$L = \prod_{i=1}^{N} \prod_{j=1}^{n} P_j(\theta_i; a_j, b_j)^{\delta_{i,j}} (1 - P_j(\theta_i; a_j, b_j))^{1-\delta_{i,j}} \tag{2}$$

となる．ここで，$\delta_{i,j}$ は，$\delta_{i,j} = (0, 1)$ をとる 2 値関数を表し，$\delta = 1$ は正答，$\delta = 0$ は誤答を表す．

通常のロジスティック分布におけるパラメータ推定では，式 (1) では θ_i が確率変数で，この観測値が得られると，a_j，b_j を未知パラメータとして推定する．しかし，IRT では a_j，b_j，θ_i すべてが未知数になるところが問題を困難にしている．**図 1** に IRT におけるパラメータ推定手順の概要を示す．誤答 0 と正答 1 からなる $\delta_{i,j}$ を式 (2) の尤度関数に代入し，それを最大にするような a_j，b_j，θ_i を同時に求めることになる．

パラメータ推定法としては，周辺最尤法とベイズ理論を用いた 2 段階アルゴリズム [1] や，マルコフ連鎖モンテカルロ法（MCMC）を応用した方法 [16] が利用できる．文献 [17] のツールには両者の手法が組み込まれ，オンラインで

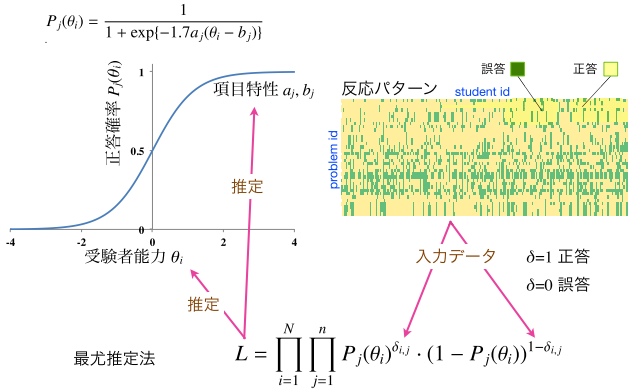


図 1　項目反応理論による推定手順
Fig. 1　Item response theory estimation procedure.

利用が可能である（http://ume98.ces.kyutech.ac.jp/score-service/）．

3. 適応型オンライン能力評価システム

適応型システムは受験者の能力に合わせたレベルの問題を自動的に出題するシステムである．受験者の解答がシステムに取り込まれるごとにその受験者の能力を逐次評価することができるため，出題する問題を受験者の能力に最適に合わせることができる．

たとえば，受験者がインターネットを通して適応型システムを受験したとき，もし受験者が最初の問題に正答すれば，システムは受験者に次により難しい問題を与える．誤答であれば，より易しい問題を与える．

既知の項目特性値を用いれば，適応型システムは受験者が 1 問解答するごとに能力値を求め，そのときの受験者の能力値レベルに最も一致する項目を次の項目として選ぶことができる．項目バンクに登録された項目群の特性が多様であればあるほど，能力に最も適した項目が選ばれやすくなり，より少ない項目数で正確な能力評価が可能となる．

4. マトリクス予測法

4.1 EM タイプ IRT [18]

EM タイプ IRT は，項目 j の項目特性値 a_j，b_j および受験者 i の能力値 θ_i を不完全マトリクスから推定することにより，正答確率 $P_j(\theta_i)$ によって不完全マトリクスの欠損要素を予測する方法である．そのために，δ を有理数まで拡張し，次に EM タイプ（expectation-maximization algorithm [5]）を用いる．以下にこのパラメータ推定手順を示す．

4.1.1 δ の有理数への拡張

式 (2) において，δ は正答のとき $\delta = 1$，誤答のとき $\delta = 0$ を表す 2 値関数であった．ここではこれを，受験者が同じ困難度を持つ異なる項目 m 問中で l 回正答したと考え $\delta = l/m$ と見なすことによって，δ に対し有理数を割り当てることができるように拡張する．

4.1.2 欠損要素に対する予測

まず，解答パターンのマトリクスで解答されていない要素に対し，$\delta^0_{i,j} \in [0,1]$ を満たす任意の初期値を与え，$\delta_{i,j} = 0, 1$ の観測値はそのまま残す．このとき得られる初期マトリクスは，$0 \le \delta^0_{i,j} \le 1$ を満たす．$\delta^0_{i,j}$ の初期値としては，項目 j の平均正答率 μ_j や，受験者 i の平均正答率 μ_i などがあげられる．各パラメータの初期値を a^0_j，b^0_j，θ^0_i とし，初期尤度 L^0 を式 (2) で定義する．

初期マトリクス $\{\delta^0_{i,j}\}$ を用いて，式 (2) の尤度 L を最大にするパラメータ a^1_j，b^1_j，θ^1_i を推定し，尤度 L^1 を得る．このときのパラメータ推定法は，2 段階アルゴリズムまたは MCMC のどちらかを用いることができる．この手順は，EM アルゴリズムの maximization ステップに対応する．

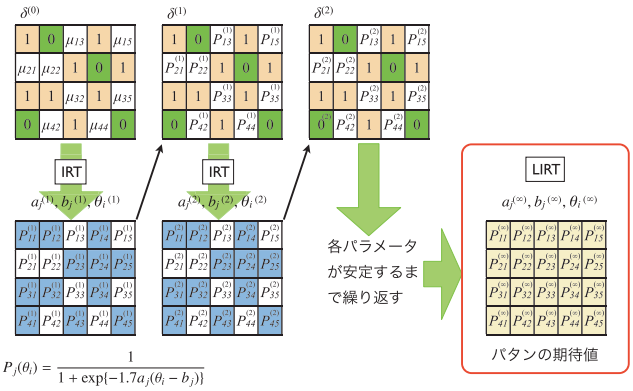


図 2　EM タイプ IRT による予測手順
Fig. 2　EM-type IRT prediction procedure.

次に，得られたパラメータを用いて式 (1) から正答確率 $P_j(\theta_i) \in [0,1]$ が算出できる．ここで，$\hat{\delta}_{i,j} = P_j(\theta_i)$ の関係が成り立つことから，観測値および $P_j(\theta_i)$ によって，$\delta^1_{i,j}$ を得る．この手順は，EM アルゴリズムの expectation ステップに対応する．

この 2 ステップの手順を繰り返し，L^k，$\delta^k_{i,j}$，a^k_j, b^k_j，θ^k_i $(k = 0, \ldots)$ を得る．$k \to \infty$ とすれば，期待される収束値 L^∞，$\delta^\infty_{i,j}$，a^∞_j，b^∞_j，θ^∞_i を得る．ただし，通常の EM アルゴリズムのような単調増加性は期待されないが，EM アルゴリズムとよく似た収束性を持つので，ここでは EM タイプ IRT と呼ぶ．この手法は，limiting IRT（LIRT）とも呼ばれる [12]．収束値がつねに一意になるとは保証されない [20], [23]．しかしながら，経験的には，多くの場合で同じ値に収束することが分かっている [11]．**図 2** に，EM タイプ IRT による予測手順の概要を示す．

得られる予測マトリクスの精度評価には下記の S^k を用いる．これは，次式で表される観測値とそれに対応する予測値 $\hat{\delta}^k_{i,j}$ の平均 2 乗誤差の平方根である．

$$S^k = \sqrt{\frac{1}{|\Delta|} \sum_{(i,j)\in\Delta} (\hat{\delta}^k_{i,j} - \delta_{i,j})^2} \qquad (3)$$

ここで，$|\Delta|$ は観測値に対応する要素の数を表す．欠損要素の予測値は S^k に含まれないことに注意しておく．収束判定として，次式を満たした場合に収束と見なす．

$$|S^k - S^{k-1}| < 1.0 \times 10^{-8}$$

ここでは δ の大きさが $[0,1]$ であることを配慮して，収束条件を相対誤差ではなく絶対誤差を用いてもよいと考えたからである．

4.2 マトリクス分解法（MD）[21]

マトリクス分解法（matrix decomposition method; MD）は，推薦システムにおいてよく利用されている方法であり [2], [14], [15]，これを受験者と項目から作られるマトリクスにも適用できる．マトリクス分解法では，不完全マト

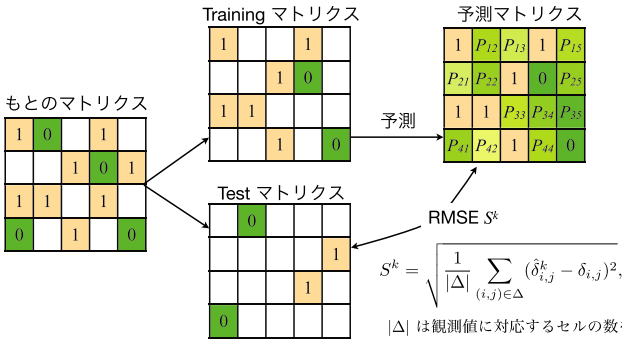


図 3　Training と Test に分けて推定を行う手順

Fig. 3　Estimation procedure using the training data and the test data.

リクス V が 2 つの未知マトリクス U と M の積 P で表されるような U，M を最小 2 乗法によって探索する [21]．最小 2 乗法では次のように 2 乗誤差に罰則項をつけた最小化を行う．

$$E = \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n} I(i,j)\{V(i,j)-P(i,j)\}^2 + \frac{k_u}{2}\sum_{i=1}^{m}||U_i||^2 + \frac{k_m}{2}\sum_{j=1}^{n}||M_j||^2 \quad (4)$$

ここで，$I(i,j)$ は観測値のインデックス関数，k_u と k_m は過学習を防ぐための正則化係数である．大規模マトリクスに対する最適化においては確率的勾配法が利用されるが，ここでは次式で表される通常の勾配法を利用している．

$$U^{(t+1)} \leftarrow U^{(t)} + \mu\sum_{i=1}^{m}\frac{\partial E}{\partial U_i} \quad (5)$$

$$M^{(t+1)} \leftarrow M^{(t)} + \mu\sum_{j=1}^{n}\frac{\partial E}{\partial M_j} \quad (6)$$

ここで，μ は学習率である．

5.　予測値の評価

予測値の評価を行うために，もとの不完全マトリクスデータを Training および Test に分け，Training で予測を行い，Test でその精度を評価する．評価式は，式 (3) に示す $RMSE$ を用いる．**図 3** に Training と Test に分けて評価を行う手順を示す．Training と Test は T セット作成する．

6.　実データの収集

EM タイプ IRT と MD の予測精度を評価するために，ここでは実際に能力評価試験を行うことで得られるマトリクスデータを用いる．能力評価試験として，従来の筆記試験と適応型システムによる試験を対象とする．

6.1　筆記試験による完全マトリクス

完全マトリクスとして，大学の学部生に実施した筆記試験による数学の試験問題とその結果（以下，データ A）を利用する [24]．このときの受験者数は 87 人，試験の問題項目数は 40 問である．このときの解答結果を**図 4** に示す．行方向は各問題項目を表しており，列方向は各受験者を表している．また，行方向は IRT によって推定される問題項目の困難度を昇順で，列方向は同じく IRT によって推定される受験者の能力値を降順でソートしている．そのため，マトリクスの左上部は，能力の高い受験者で，かつ易しい問題項目になるため，正答（白色）が多くなっている．逆に，マトリクスの右下部は，能力の低い受験者かつ難しい問題項目になるため，誤答（黒色）が多くなっている．

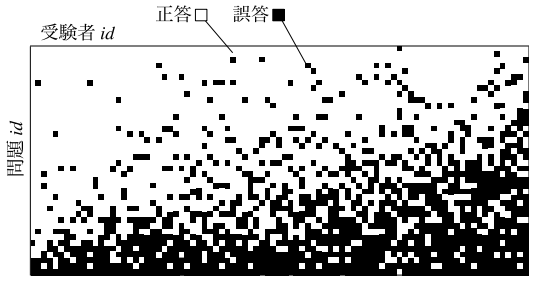


図 4　数学の試験（データ A）の解答結果（87 人 × 40 問）

Fig. 4　The result of mathematics examination (data A).

6.2　適応型システムによる不完全マトリクス

適応型システムでは，項目バンクに登録されている項目が多様であればあるほど，より正確な能力評価が可能になる．文献 [18] では，高校数学の能力評価を対象とし，項目バンクの項目数は 30 問であった．そこで，今回新たに高校数学レベルの項目を 66 問を追加し，計 96 問の項目バンクを作成し，能力をより精度良く評価できるシステムを目指した．このシステムを実際に受験者に適用し，新たな不完全マトリクスを得た．

6.2.1　項目バンクの拡充

項目バンクに項目を追加するためには，追加したい項目群（追加項目群）から構成される予備テストを適応型システムの受験者集団とは別の集団に受験してもらい，その解答結果から項目特性を推定する必要がある．ただし，予備テストの項目群と項目バンクにすでに登録してある項目群（旧項目群）とでは，項目特性の尺度が異なっているため，共通の尺度に揃える操作が必要になる．これを等化という．一般的な等化では，予備テストを構成する項目群に旧項目群の項目を数問混ぜておくことで，旧項目群のもとの項目特性と予備テストの結果から推定される項目特性とを共通尺度上に変換している．尺度変換には平均シグマ法 [3] などが利用されている．ただし，予備テストの項目数が多すぎると，予備テスト受験者に対する負荷が大きくなり，正確な試験結果が得られにくくなるため，予備テスト項目数を少なくする工夫が必要である．

それを解決する 1 つの方法に分冊法がある．分冊法は，

予備テストを複数の小冊子に分割しそれぞれに共通の項目群を付加して試験を行う等化法である．各小冊子の解答結果から得られるそれぞれの項目特性値は，共通項目の特性値をもとに共通尺度に等化される．**図 5** に分冊法による等化法の概要を示す．

ここではまず，旧項目群 30 問に追加する 66 問を 3 つの小冊子に分割し，共通項目として旧項目群から 10 問選んだ．各小冊子の項目数はそれぞれ 30 問程度とした．予備テストは受験者 51 人を対象に行い，得られた解答結果から IRT によって項目特性値を推定した．**図 6** に分冊法による項目困難度の等化結果を示す．図中の点線はそれぞれの解答結果から求めた項目困難度の密度分布，実線は等化後の項目困難度の密度分布を示している．ここでは小冊子 2 の項目特性値を基準に等化した．そのため，小冊子 2 の項目困難度は等化前後で変化はない．また，全体的に等化前後での変化は小さいことから，小冊子間において項目困難度にそれほどの差はなかったと考えられる．

6.2.2　適応型オンライン能力評価システムの実施

ここでは，実際に実施した適応型システムについて説明する．1 人 5 問の設問で 1 問 1 問に逐次解答するようになっている．作成した項目バンクに登録されている問題は合計 96 問で，すべて高校数学レベルである．受験者数は 138 人であった．このとき得られる不完全マトリクスデータ（以下，データ B）を**図 7** に示す．図中の左側が予備テストの結果，右側が適応型システムの結果を表す．図中の白色の要素は正答を表し黒色は誤答を表す．灰色の要素は欠損していることを表す．また，**図 8** では，解答するごとに受験者の能力値が変化している様子が分かる．横軸は解答の順，縦軸は能力値を示しており，曲線 1 本 1 本がそれぞれの受験者を表す．各受験者に対して，1 問目はまわりの受験者と重複しないように項目バンクの中から無作為抽出している．その際，極端に難しい，あるいは極端に易しい問題が出題されないように，平均的な問題からやや易

図 5　分冊法による等化

Fig. 5　Equalization using the booklet method.

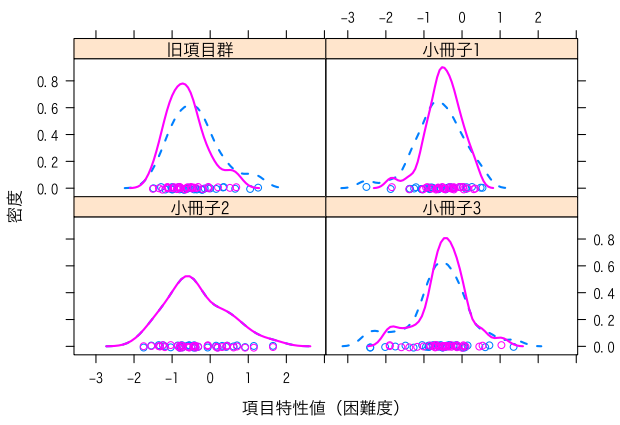


図 6　項目特性値（困難度）の等化結果

Fig. 6　Difficulty parameter after the equalization.

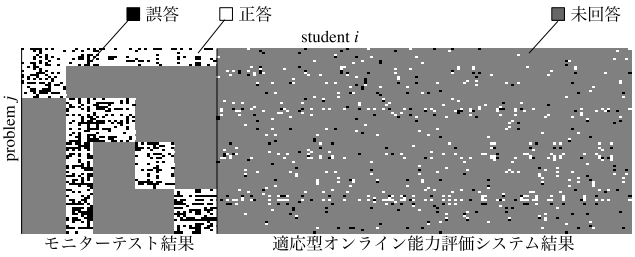


図 7　予備テスト結果と適応型システム結果

Fig. 7　The result of the monitor test and the adaptive online test.

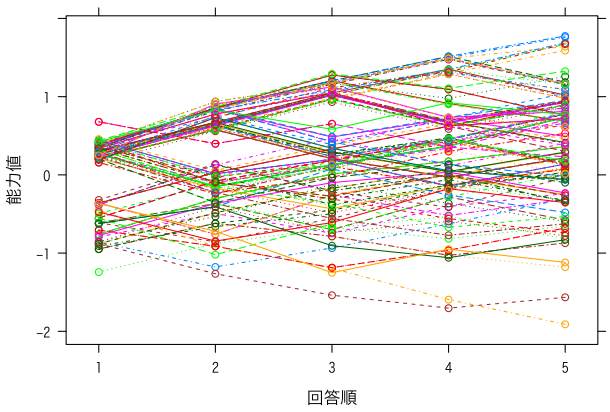


図 8　能力値の変化

Fig. 8　Transition of the ability parameter.

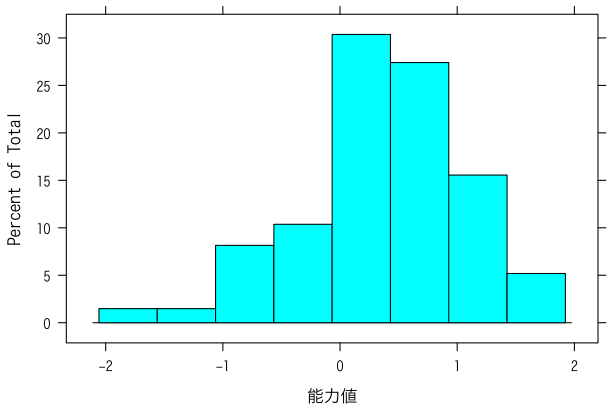


図 9　5 問終了時の能力値のヒストグラム

Fig. 9　Histogram of the ability parameter after solving five questions.

しい項目の間から選ばれるように工夫している．図 **9** に 5 問終了時の能力値のヒストグラムを示す．能力値のモードは 0 よりやや右側にあり，左側に裾が長い分布になっており，出題の問題レベルよりも受験者の能力値が高かったことが分かる．

7. シミュレーション

ここでは，EM タイプ IRT が正しく機能していることをコンピュータシミュレーションにより検証する．検証手順を以下に示す．

(1) データ A から，IRT によって，能力パラメータ θ_i $(i=1,\ldots,N)$，項目パラメータ a_j，b_j $(j=1,\ldots,n)$，すべてのマトリクスの要素の正答確率 $p_{i,j}$ を計算する．

(2) θ_i，a_j，b_j から，データ A を模倣したマトリクス A_m $(m=1,2,\ldots,M)$ をモンテカルロ法によって生成する．

(3) M 個のマトリクスのそれぞれについて，以下を行う．

(a) A_m を Training（K 個の要素）と Test（k 個の要素）に無作為に分ける．

(b) Training から，EM タイプ IRT によって，能力パラメータ $\hat{\theta}_i$，項目パラメータ $\hat{a}_j$，$\hat{b}_j$，および欠損値（Test に相当する部分）の正答確率 $t_{i,j}$ を求める．

(c) $\hat{\theta}_i$，$\hat{a}_j$，$\hat{b}_j$，$t_{i,j}$ のそれぞれについて，$bias$ と var を計算する．

$$bias=\frac{1}{n_x}\sum(\hat{x}-\tilde{x}),\quad var=\frac{1}{n_x-1}\sum(\hat{x}-\tilde{x})^2$$

$$(\tilde{x},\hat{x},n_x)\in\{(p_{i,j},t_{i,j},k),(\theta_i,\hat{\theta}_i,N),(a_j,\hat{a}_j,n),(b_j,\hat{b}_j,n)\}$$

モンテカルロ法によるマトリクス生成法は文献 [25] によった．また，本研究では $M=10$ とした．$K+k=N\times n=87\times 40=3480$ であり，$K=2784$（全体の 80%）の場合と $K=696$（全体の 20%）の場合で計算を行った．このときの計算結果を**表 1**，**表 2**，**表 3**，**表 4** に示す．提案法の予測精度を評価するため，欠損要素に $[0,1]$ からランダムに選んだ数値を入れた場合の結果と比較した．表ではそれを random と表している．

表 1 および表 2 から，EM タイプ IRT による予測値は，ランダムに欠損値を埋めた場合に比べ，$bias$，var ともに小さい値を示していることが分かる．また，表 3 および表 4 から，どのパラメータも $bias$，var ともに小さい値を示しており，元のパラメータを再現していることが分かる．

表 1　シミュレーションによる欠損値の $bias$ と var（$K=2784$）
Table 1　The $bias$ and var for the missing values by the simulation ($K=2784$).

$K=2784$	EM タイプ IRT		random	
m	$bias$	var	$bias$	var
1	2.42×10^{-3}	7.22×10^{-3}	-1.54×10^{-1}	1.94×10^{-1}
2	1.53×10^{-3}	8.93×10^{-3}	-1.85×10^{-1}	2.06×10^{-1}
3	-6.18×10^{-4}	9.35×10^{-3}	-1.48×10^{-1}	2.09×10^{-1}
4	8.62×10^{-5}	8.24×10^{-3}	-1.66×10^{-1}	2.12×10^{-1}
5	-3.21×10^{-5}	8.24×10^{-3}	-1.77×10^{-1}	2.01×10^{-1}
6	-1.95×10^{-3}	8.33×10^{-3}	-1.74×10^{-1}	2.13×10^{-1}
7	3.97×10^{-3}	9.99×10^{-3}	-1.73×10^{-1}	2.02×10^{-1}
8	8.83×10^{-3}	7.50×10^{-3}	-1.84×10^{-1}	2.03×10^{-1}
9	1.56×10^{-3}	6.18×10^{-3}	-1.66×10^{-1}	2.00×10^{-1}
10	1.08×10^{-3}	6.95×10^{-3}	-1.70×10^{-1}	2.01×10^{-1}
平均	1.69×10^{-3}	8.09×10^{-3}	-1.70×10^{-1}	2.04×10^{-1}

表 2　シミュレーションによる欠損値の $bias$ と var（$K=696$）
Table 2　The $bias$ and var for the missing values by the simulation ($K=696$).

$K=696$	EM タイプ IRT		random	
m	$bias$	var	$bias$	var
1	1.37×10^{-2}	2.71×10^{-2}	-1.74×10^{-1}	2.00×10^{-1}
2	-4.16×10^{-3}	2.32×10^{-2}	-1.71×10^{-1}	1.96×10^{-1}
3	8.96×10^{-3}	2.09×10^{-2}	-1.69×10^{-1}	2.04×10^{-1}
4	-5.13×10^{-3}	1.97×10^{-2}	-1.81×10^{-1}	1.95×10^{-1}
5	1.45×10^{-2}	2.42×10^{-2}	-1.74×10^{-1}	2.04×10^{-1}
6	7.87×10^{-3}	2.29×10^{-2}	-1.71×10^{-1}	2.01×10^{-1}
7	1.47×10^{-2}	2.39×10^{-2}	-1.66×10^{-1}	1.97×10^{-1}
8	1.11×10^{-2}	2.85×10^{-2}	-1.69×10^{-1}	2.05×10^{-1}
9	-4.31×10^{-3}	2.05×10^{-2}	-1.67×10^{-1}	2.07×10^{-1}
10	3.35×10^{-2}	2.63×10^{-2}	-1.76×10^{-1}	2.06×10^{-1}
平均	9.08×10^{-3}	2.37×10^{-2}	-1.72×10^{-1}	2.02×10^{-1}

8. 実データの予測

データ A およびデータ B に対して，EM タイプ IRT および MD によるマトリクス予測手法を適用し，その予測精度を比較する．

8.1 データ A の予測

ここでは，完全マトリクスであるデータ A から一部を欠損値として扱い，欠損部分の答えが分かっているケースとして予測精度の比較を行う．ただし，マトリクスの各行および各列について，解答結果が少なくとも 1 つは存在するようにする．欠損値の割合については，マトリクス全体の 20%および 80%が欠損している 2 つのケースを考える．欠損値の生成法については，無作為抽出による．

各ケースにおいて，もとの完全マトリクスから欠損させた部分を Test データ，不完全マトリクスとして残った部分を Training データとして扱う．Training データから欠損部分のマトリクス予測を行い，その答えである Test データとの $RMSE$ を式 (3) によって求める．これを 30 パターン行い，得られる 30 個の $RMSE$ の平均および標準偏差によって予測精度とする．

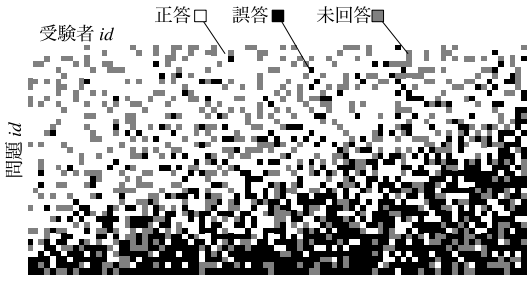
データ A から無作為に 20%を欠損させた場合の一例を**図 10** (a) に示す．同じく，無作為に 80%を欠損させた場合

表 3　シミュレーションによる θ_i，a_j，b_j の $bias$ と var（$K = 2784$）
Table 3　The $bias$ and var for θ_i, a_j, and b_j by the simulation ($K = 2784$).

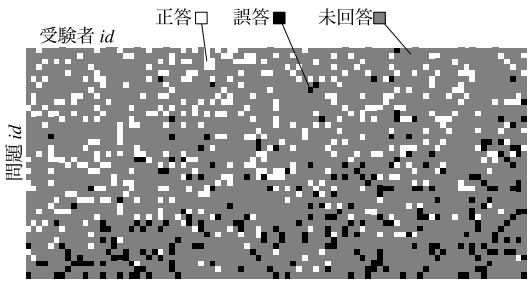
$K = 2784$	θ		a		b	
m	$bias$	var	$bias$	var	$bias$	var
1	1.42×10^{-2}	8.62×10^{-2}	5.23×10^{-2}	9.65×10^{-2}	-1.14×10^{-1}	1.17×10^{-1}
2	1.23×10^{-2}	1.33×10^{-1}	1.64×10^{-2}	5.75×10^{-2}	-1.07×10^{-2}	1.02×10^{-1}
3	1.70×10^{-2}	1.24×10^{-1}	5.37×10^{-2}	7.03×10^{-2}	-9.43×10^{-2}	9.99×10^{-2}
4	1.08×10^{-3}	1.05×10^{-1}	3.35×10^{-2}	8.28×10^{-2}	-7.09×10^{-2}	1.13×10^{-1}
5	2.51×10^{-2}	1.41×10^{-1}	3.56×10^{-2}	7.55×10^{-2}	-2.45×10^{-2}	1.00×10^{-1}
6	-7.16×10^{-3}	1.20×10^{-1}	3.21×10^{-2}	7.47×10^{-2}	-2.05×10^{-2}	8.86×10^{-2}
7	5.21×10^{-2}	1.28×10^{-1}	7.13×10^{-2}	6.84×10^{-2}	-7.06×10^{-2}	1.04×10^{-1}
8	2.97×10^{-2}	1.28×10^{-1}	2.94×10^{-2}	6.48×10^{-2}	-3.21×10^{-2}	7.54×10^{-2}
9	1.67×10^{-2}	8.32×10^{-2}	6.80×10^{-2}	6.47×10^{-2}	-3.84×10^{-2}	1.52×10^{-1}
10	8.63×10^{-3}	8.14×10^{-2}	2.03×10^{-2}	7.21×10^{-2}	-1.02×10^{-1}	1.40×10^{-1}
mean	1.70×10^{-2}	1.13×10^{-1}	4.13×10^{-2}	7.27×10^{-2}	-5.78×10^{-2}	1.09×10^{-1}

表 4　シミュレーションによる θ_i，a_j，b_j の $bias$ と var（$K = 686$）
Table 4　The $bias$ and var for θ_i, a_j, and b_j by the simulation ($K = 686$).

$K = 686$	θ		a		b	
m	$bias$	var	$bias$	var	$bias$	var
1	1.58×10^{-2}	2.74×10^{-1}	7.80×10^{-2}	8.66×10^{-2}	-1.01×10^{-1}	3.97×10^{-1}
2	5.87×10^{-3}	2.42×10^{-1}	3.56×10^{-2}	8.38×10^{-2}	-4.08×10^{-2}	2.55×10^{-1}
3	5.81×10^{-3}	3.45×10^{-1}	7.67×10^{-2}	8.56×10^{-2}	-1.20×10^{-1}	2.04×10^{-1}
4	1.09×10^{-2}	2.82×10^{-1}	7.80×10^{-3}	8.41×10^{-2}	-8.66×10^{-2}	2.27×10^{-1}
5	7.51×10^{-3}	3.07×10^{-1}	7.42×10^{-2}	7.49×10^{-2}	-1.36×10^{-1}	2.12×10^{-1}
6	1.72×10^{-2}	3.55×10^{-1}	8.13×10^{-2}	6.99×10^{-2}	-1.50×10^{-1}	2.05×10^{-1}
7	1.58×10^{-2}	3.10×10^{-1}	9.80×10^{-2}	9.48×10^{-2}	-1.71×10^{-1}	3.19×10^{-1}
8	2.21×10^{-2}	4.15×10^{-1}	8.17×10^{-2}	7.53×10^{-2}	-1.97×10^{-1}	3.34×10^{-1}
9	-1.23×10^{-2}	2.75×10^{-1}	5.90×10^{-2}	6.54×10^{-2}	-8.32×10^{-2}	2.28×10^{-1}
10	2.98×10^{-2}	3.65×10^{-1}	3.61×10^{-2}	6.62×10^{-2}	-2.33×10^{-1}	2.99×10^{-1}
平均	1.18×10^{-2}	3.17×10^{-1}	6.28×10^{-2}	7.87×10^{-2}	-1.32×10^{-1}	2.68×10^{-1}



(a) 20%欠損



(b) 80%欠損

図 10　データ A から生成した不完全マトリクス
Fig. 10　The incomplete matrix generated from data A.

表 5　データ A から生成した不完全マトリクスの 30 ケースの $RMSE$ の平均
Table 5　The mean of $RMSE$ by using 30 cases of the incomplete matrix generated from data A.

	EM-type IRT		MD	
	Training	Test	Training	Test
Training 80%	0.3578	0.3818	0.2147	0.4542
Test 20%	0.0022	0.0087	0.0046	0.0138
Training 20%	0.3270	0.4037	0.0299	0.4518
Test 80%	0.0116	0.0062	0.0025	0.0073

上は平均，下は標準偏差

の一例を図 10 (b) に示す．このときの予測結果を**表 5** に示す．Test の $RMSE$ の平均値を見ると，EM タイプ IRT の $RMSE$ が小さく予測精度が良いことを示している．

8.2　適応型システム結果（データ B）の予測

データ B に対し，EM タイプ IRT および MD による不完全マトリクスの予測を行う．ここでは，もとのデータを Training と Test に 9 対 1 に分けて評価を行う．分け方は，

表 6　データ B を 9:1 の Training と Test に分けた 30 ケースの $RMSE$ の平均

Table 6　The mean of $RMSE$ by using 30 cases of the 90% training data and the 10% test data (data B).

	EM タイプ IRT		MD	
	Training	Test	Training	Test
データ B1	0.340	0.521	0.0133	0.523
	0.00440	0.0299	0.00116	0.0295
データ B2	0.380	0.443	0.145	0.501
	0.00162	0.0154	0.00410	0.0178

上は平均，下は標準偏差

無作為抽出による．データ A の場合と同様に，Training と Test を 30 セット作成する．また，データ B には，予備テストの試験結果が含まれているため，適応型システムの結果のみをデータ B1，予備テストも合わせたものをデータ B2 と再定義する．

表 6 にこのときの $RMSE$ を示す．表 6 を見ると，Test の $RMSE$ に対しては，データ B1，B2 ともに EM タイプ IRT の方が小さいことが分かる．特に，データ B2 の結果は EM タイプ IRT の方がより小さく，予測精度が良いことを示している．また，データ B1 とデータ B2 の Test の $RMSE$ を比較すると，MD に比べて EM タイプ IRT の Test の $RMSE$ の方が変化が大きい．データ B1 とデータ B2 の欠損率（マトリクス全体の要素数に対する欠損要素の割合）はそれぞれ 95.1%と 85.7%であり，EM タイプ IRT は欠損率が小さくなることによって精度が良くなっておりデータ数の影響を受けていると考えられる．

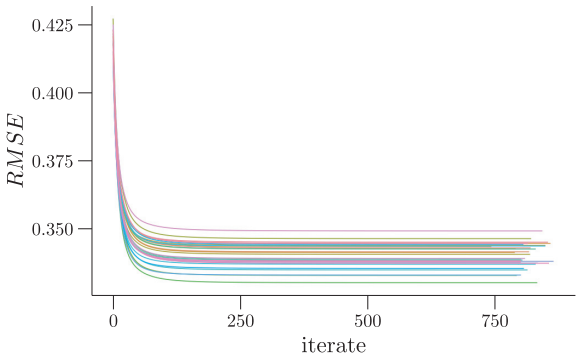
9.　収束性の検討

EM タイプ IRT は，その予測手順の性質から，予測値が収束に至るまでの $RMSE$ の挙動は単調ではないと考えられる．そこで本章では，データ B における 30 ケースの Training に対する予測値の収束を見ることで，その収束性について検討する．

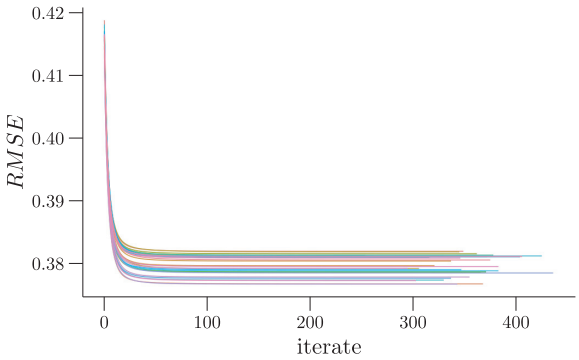
図 11 (a) にデータ B1 を用いた場合について，図 11 (b) にデータ B2 を用いた場合について，30 ケースの Training による $RMSE$ の収束の様子を示す．どちらの場合も，30 ケースすべてが単調に減少している．

次に，30 ケースの Training による対数尤度 $\log L$ の収束の様子を，データ B1 については**図 12** (a) に，データ B2 については図 12 (b) に示す．ここでいう対数尤度とは，観測された値に対応した観測値による尤度である．欠損値を予測した値は含まれない．どちらの場合の $\log L$ も，30 ケースすべてが単調に増加している．

EM タイプ IRT では，計算の過程で，欠損要素を予測値で置き換える操作を繰り返し，そこで得られる観測値と予測値を組み合わせたマトリクスを次の初期値としている．つまり，計算の更新ごとに，扱うデータが異なっていることになる．そのため，計算の更新ごとに得られる $RMSE$ および $\log L$ の値は，単調に減少または増加するとは限らない．実際に文献 [11] の場合には単調性は見られなかった．しかし，今回のように経験的には多くの場合で同じ値に収束している．



(a) データ B1



(b) データ B2

図 11　9:1 の Training と Test に分けた 30 ケースに EM タイプ IRT を適用したときの Training に対する $RMSE$ の変化

Fig. 11　RMSE trend by using 30 cases of the 90% training data and the 10% test data via the EM-type IRT.

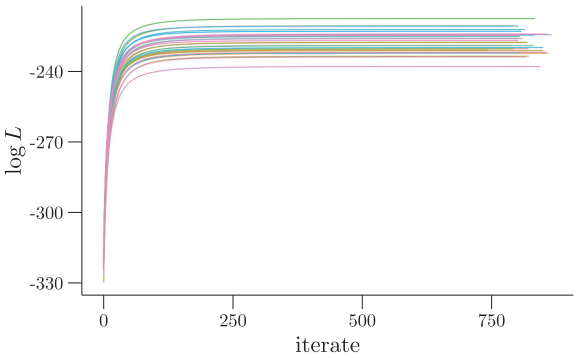
10.　考察

10.1　制約付き MD について

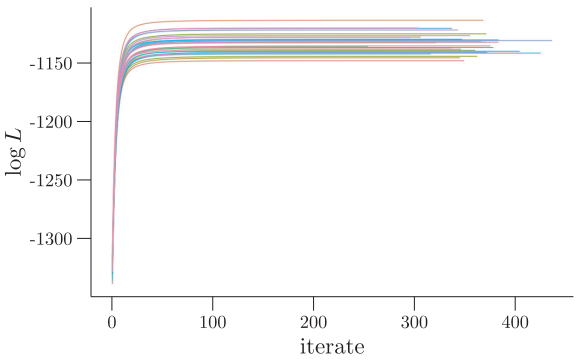
本研究では，不完全マトリクスに対する予測手法として，EM タイプ IRT と MD の比較を $RMSE$ による評価で行った．EM タイプ IRT では，予測値 T は確率値として得られるため，$0 \leq T \leq 1$ を満たす．一方，MD では，予測値は 1 以上の値や 0 未満の値も含まれる．そこで，MD では，$0 \leq T \leq 1$ を満たすように，予測値 T が 0 未満の場合は 0 に，1 以上の場合は 1 に制約を付ける操作が行われる．ここで，制約なしの場合を考える．**表 7** を得る．制約なしの場合，明らかに予測精度は悪くなっている．不完全マトリクスの欠損部分を予測する場合，0/1 の範囲に制約をつける操作は自然である．

10.2　誤分類率による評価

もとのマトリクスデータは 0/1 の 2 値データであるた



(a) データ B1



(b) データ B2

図 12　9:1 の Training と Test に分けた 30 ケースに EM タイプ IRT を適用したときの Training に対する $\log L$ の推移

Fig. 12　$\log L$ trend by using 30 cases of the 90% training data and the 10% test data via the EM-type IRT.

表 7　制約がある場合とない場合の MD における $RMSE$ の平均

Table 7　The mean of $RMSE$ by using 30 cases of the constrained MD and unconstrained MD.

	制約付き MD	制約なし MD
データ B1	0.523	0.557
	0.0295	0.0287
データ B2	0.501	0.575
	0.0178	0.0178

上は平均，下は標準偏差

め，予測マトリクスもまた 2 値に分類する問題ととらえることができる．そこで，得られる予測マトリクスを 2 値データに分類したときの誤分類率について考察する．簡単のため，観測値 $x_{i,j}$ に対する予測値 $\hat{x}_{i,j}$ を改めて

- $(x_{i,j}=0)\cap(0\le\hat{x}_{i,j}<0.5)\ \Rightarrow\ T=0$
- $(x_{i,j}=0)\cap(0.5\le\hat{x}_{i,j}\le 1)\ \Rightarrow\ T=1$
- $(x_{i,j}=1)\cap(0\le\hat{x}_{i,j}<0.5)\ \Rightarrow\ T=1$
- $(x_{i,j}=1)\cap(0.5\le\hat{x}_{i,j}\le 1)\ \Rightarrow\ T=0$

と分類した場合を考える．T は 2 値関数であり，$T=1$ のとき誤分類，$T=0$ のとき正分類を表す．このときの 30 ケースの Test に対する誤分類率の平均 $\sum_{i,j}T/\#T$ は，**表 8** となる．この結果を見ると，MD のほうが誤分類率が低いことが分かる．両手法とも，データ数の増加によって

表 8　Test データの誤分類率の平均

Table 8　The mean of misclassification rates by the test data results.

	EM タイプ IRT	MD
データ B1	0.390	0.337
	0.0608	0.0491
データ B2	0.287	0.259
	0.0247	0.0239

上は平均，下は標準偏差

誤分類率は低下しており，データによる誤分類率のばらつきも小さいことが分かる．

以上のことから，両手法の予測値を連続値として評価すると EM タイプ IRT のほうが優れているが，0/1 の離散値にすると MD が優れていることが分かった．

11.　まとめ

項目反応理論（IRT）は，あるテストの問題を受験者が解答したときの解答パターンのマトリクスから，テストの問題項目特性値および受験者能力評価値を推定することができる．IRT は通常，すべての問題に全員が解答している完全マトリクスを用いるが，ここでは未解答部分を含む不完全マトリクスに IRT が適用できる EM タイプ IRT を提案した．まず，テストの問題項目特性および受験者能力が分かっているときに不完全マトリクスを模擬したデータを用いて，提案法が元のパラメータを再現できることを確認した．次に実際に行ったテストに対して提案法による予測を行った．さらに，提案法を不完全マトリクスの予測法の 1 つであるマトリクス分解法と比較した．その結果，提案法はマトリクス分解法よりも良い予測精度を出すことが分かった．

謝辞　本研究に協力いただいた月原由紀，山本諭，河郷久史，桑幡隆行，野口和久さんらに感謝する．本研究の一部は科学研究費（挑戦的萌芽研究 23650543，特研奨励 No.25・5474）によった．

参考文献

[1] Baker, F. and Kim, S.: *Item response theory: Parameter estimation techniques*, Vol.176, CRC (2004).
[2] Bell, R., Bennett, J., Koren, Y. and Volinsky, C.: The million dollar programming prize, *Spectrum*, Vol.46, No.5, pp.28–33, IEEE (2009).
[3] Cook, L.L. and Eignor, D.R.: IRT equating methods, *Educational Measurement*, Vol.10, No.3, pp.37–45 (1991).
[4] De Ayala, R.: *The theory and practice of item response theory*, Guilford Press (2009).
[5] Dempster, A., Laird, N. and Rubin, D.: Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society, Series B* (*Methodological*), pp.1–38 (1977).
[6] Hambleton, R.: *Fundamentals of item response theory*, Vol.2, Sage Publications, Incorporated (1991).

[7] Hambleton, R. and Swaminathan, H.: *Item response theory: Principles and applications*, Vol.7, Springer (1984).
[8] Hirose, H.: An optimal test design to evaluate the ability of an examinee by using the stress–strength model, *Journal of Statistical Computation And Simulation*, Vol.81, No.1, pp.79–87 (2011).
[9] Hirose, H., Nakazono, T., Tokunaga, M., Sakumura, T., Sumi, S. and Sulaiman, J.: Seasonal infectious disease spread prediction using matrix decomposition method, *4th International Conference on Intelligent Systems, Modelling and Simulation, ISMS 2013*, Bangkok, Thailand, pp.152–156, The Royal Society (2013).
[10] Hirose, H. and Sakumura, T.: An Accurate Ability Evaluation Method for Every Student with Small Problem Items using the Item Response Theory, *Computers and Advanced Technology in Education, CATE 2010*, pp.152–158, ACTA Press (2010).
[11] Hirose, H. and Sakumura, T.: Item response prediction for incomplete response matrix using the EM-type item response theory with application to adaptive online ability evaluation system, *2012 IEEE International Conference on Teaching, Assessment and Learning for Engineering* (*TALE*), pp.T1A–6–T1A–10 (2012).
[12] Hirose, H., Sakumura, T. and Ichii, S.: A recommendation algorithm that assumes a probabilistic structure and its application to questionnaire data, IPSJ SIG Technical Report, Fukuoka, Japan, pp.1–7 (2011).
[13] Mills, C., Potenza, M., Fremer, J. and Ward, W.: *Computer-based testing: Building the foundation for future assessments*, Lawrence Erlbaum (2002).
[14] Netflix: Netflix prize, available from ⟨http://www.netflixprize.com/⟩.
[15] Netflix: Netflix Update: Try This at Home., available from ⟨http://sifter.org/˜simon/journal/20061211.html⟩.
[16] Patz, R. and Junker, B.: Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses, *Journal of educational and behavioral statistics*, Vol.24, No.4, pp.342–366 (1999).
[17] Sakumura, T. and Hirose, H.: Test Evaluation System via the Web using the Item Response Theory, *Information*, Vol.13, No.3, pp.647–656 (2010).
[18] Sakumura, T., Kuwahata, T. and Hirose, H.: An adaptive online ability evaluation system using the item response theory, *Education and e-Learning* (*EeL2011*), Global Science and Technology Forum (GSTF), pp.51–54 (2011).
[19] SSL: Bilog-mg (2005), available from ⟨http://www.ssicentral.com/irt/index.html⟩.
[20] Suen, H. and Lee, P.: *Constraint optimization: An alternative perspective of IRT parameter estimation*, chapter 17, pp.289–300, Norwood, NJ: Ablex. (1994).
[21] Takimoto, S. and Hirose, H.: Recommendation Systems and Their Preference Prediction Algorithms in a Large-Scale Database, *Information*, Vol.12, No.5, pp.1165–1182 (2009).
[22] van der Linden, W. and Hambleton, R.: *Handbook of modern item response theory*, Springer (1996).
[23] Yen, W., Burket, G. and Sykes, R.: Nonunique solutions to the likelihood equation for the three-parameter logistic model, *Psychometrika*, Vol.56, No.1, pp.39–54 (1991).
[24] 月原由紀，作村建紀，廣瀬英雄：IRT を用いた解析学試験評価と e-learning 支援の試み，PC Conference 論文集，pp.123–124 (2010).
[25] 月原由紀，鈴木敬一，廣瀬英雄：項目反応理論による評価を加味した数学テストと e-learning システムへの実装の試み，コンピュータ&エデュケーション（CIEC），Vol.24, pp.70–76 (2008).

作村 建紀

2014 年九州工業大学大学院博士後期課程修了，同年から中央大学理工学部助教．博士（情報工学）．日本統計学会，日本計算機統計学会，日本 OR 学会，IEEE 各会員．

徳永 正和

2014 年九州工業大学大学院博士前期課修了，同年から株式会社 KIS．日本 OR 学会会員．

廣瀬 英雄　（正会員）

1977 年（株）高岳製作所入社，技術本部数値情報センター長，首席研究員後，1988 年技術本部副技師長，1989 年スタンフォード大学統計学科研究員，1995 年広島市立大学情報科学部教授，1998 年九州工業大学情報工学部教授．工学博士．電気学会，応用統計学会，日本計算機統計学会，日本統計学会，日本応用数理学会，電子情報通信学会，日本数学会，日本 OR 学会，統計科学研究会，日本信頼性学会，コンピュータ利用教育学会，日本工学教育協会，行動計量学会，大学教育学会，IEEE（PES; DEIS; Reliability; Computer; Communications; Information Theory; Signal Processing; Systems, Man, and Cybernetics），ASA，IMS，MPS，SIAM，AMS，ACM 等の会員，IEC/TC 国際エキスパート．