Vol.2013-MPS-96 No.24
Vol.2013-BIO-36 No.24
2013/12/12

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[1,2,a)]　徳永 正和[1,b)]　廣瀬 英雄[3,c)]

概要：IRT（項目反応理論）はテストの問題項目特性および受験者能力評価に対して有益な情報を与える．これを不完全マトリクスに適用するEMタイプIRTが提案されている．この方法は，不完全マトリクスの観測された値を用いて空要素の値を予測する．不完全マトリクスの一般的な予測法としてマトリクス分解を用いた方法が挙げられるが，受験者の特性が強く反映されるような不完全マトリクスの場合，EMタイプIRTはより有効に働く可能性がある．ここでは，実際に得られた不完全マトリクスデータに対してEMタイプIRTによる予測を行い，マトリクス分解法の結果と比較することで，その有効性について述べる．また，EMタイプIRTの収束性について検討した結果も述べる．

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 incomplete matrix using EM-type IRT and its application

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

Abstract: The item response theory (IRT) gives us the valuable information about the difficulties of problems as well as the abilities of students. Although, the IRT covers the complete matrix, the EM-type IRT can be applied to also the incomplete matrix. This method predicts the values of the vacant elements using the observed values in the incomplete matrix. The matrix decomposition method is another choice to make up a complete matrix from the incomplete matrix. When the characteristics of users dominate in the matrix, the EM-type IRT has a possibility to work more effective than does the matrix decomposition method. In this paper, we compare the prediction results between the EM-type IRT and the matrix decomposition method using the real case of the incomplete matrix. In addition, we show the results of the convergence for the EM-type IRT.

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

1. はじめに

IRT（項目反応理論）は，テストを受験する受験者の能力を公正に評価する方法として有用である [2], [4], [5], [18]. IRTは受験者能力と同時に，テストの問題項目特性も評価する．IRTによる評価に対し，いくつかのツールが考案されている．[15] はもっとも一般的なツールとして挙げられるが，IRTの専門知識が必要であり扱いが難しいという問題がある．[13] 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の一般的な方法を用いた推定を可能にしている．[13] を用いた研究例として，少問題に対する能力評価の精度および公平性を評価する研究が行われている [8].

IRTを利用することで，適応型オンライン能力評価シス

[1] 九州工業大学大学院情報工学府
〒820–8502 福岡県飯塚市川津 680–4
[2] 日本学術振興会特別研究員 DC2
[3] 九州工業大学大学院情報工学研究院
〒820–8502 福岡県飯塚市川津 680–4
[a)] sakumura@ume98.ces.kyutech.ac.jp
[b)] tokunaga@ume98.ces.kyutech.ac.jp
[c)] hirose@ces.kyutech.ac.jp

Vol.2013-MPS-96 No.24
Vol.2013-BIO-36 No.24
2013/12/12

テムを構築することができる [11], [14]．このシステムは，受験者ごとの能力に合わせたレベルの問題を出題する．このとき，出題される問題は，項目特性が既知の項目群から選ばれる．この項目群を項目バンクという．その回答結果をもとに，受験者の能力は逐次評価される．そのため，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．一方，これに信頼性分野の知見を援用したものとして，ストレス・ストレングスモデルと昇降法を用いた能力評価法 [6] もまた有効なモデルである．

適応型オンライン能力評価システムでは，受験者の能力に合わせたレベルの問題を出題するため，あらかじめ出題される問題の項目特性を知っておく必要がある．そのためには，事前にモニターテストを実施する．モニターテストとは，項目バンクに項目を登録する際に実施される試験であり，その問題項目は，項目バンクに追加する項目群から構成される．モニターテストを受験する受験者集団は，適応型オンライン能力評価システムを受験する集団とは異なる必要がある．モニターテストの回答結果から問題項目特性を推定し，項目バンクに追加する．一般に，項目バンクの項目数が多ければ多いほど，受験者の能力の多様性に柔軟に対応することができる．

適応型オンライン能力評価システムを受験する受験者が多くなれば，その結果をもとにした項目特性のキャリブレーションが必要になる．しかし，このシステムでは，受験者ごとに回答した問題が異なるため，受験者全体および問題項目全体の回答結果は不完全なマトリクスとなる．一般的な IRT では，この不完全マトリクスから項目特性を推定することが困難である．そのため，不完全マトリクスから項目特性および能力評価を行うパラメータ推定法が必要になる．

不完全マトリクスから完全マトリクスを推定する一般的な方法は，マトリクス分解法（MF）が挙げられる [7], [17]．これは，データの背後に確率構造を含まないノンパラメトリックな方法である．しかしながら，能力評価試験のように，受験者の資質や授業の出来，不出来によって，回答パターンがある程度想定される範囲で変動するような場合，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．そこで，本研究では，不完全マトリクスに IRT を適用する方法として，EM タイプ IRT を提案する [9]．この方法は，不完全マトリクスの観測された値を用いて空要素の値を確率的に予測する．受験者の特性が強く反映されるような不完全マトリクスの場合，EM タイプ IRT は MF よりも有効に働く可能性がある．この提案手法を，実際に適応型オンライン能力評価システムによって得られた不完全マトリクスデータに適用し，同時にマトリクス分解法の結果と比較する．さらに，モニターテストの回答結果も合わせて観測値を増加させた場合の不完全マトリクスデータに対しても，両手法を適用し結果を比較する．

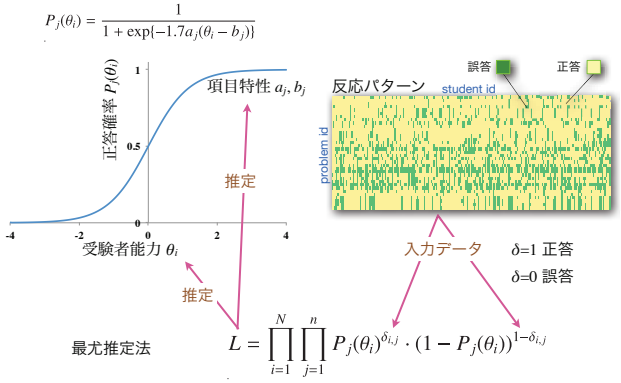


図 1 項目反応理論による推定手順
Fig. 1 Item response theory estimation procedure.

また，EM タイプ IRT の収束性について検討した結果も述べる．

2. IRT

一般的な IRT では，各項目 j に対する受験者 i の評価確率 $P_j(\theta_i)$ が 2 パラメータロジスティック分布に従うと仮定する．このとき，

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

と表される．a_j，b_j は項目 j の識別力と困難度を表し，a_j が大きいほど項目 j の識別力が高くなり，b_j が大きいほど項目 j の困難度が高くなることを意味する．θ_i は受験者 i の能力を表し，θ_i が大きいほど能力が高いことを意味する．

受験者 $i = 1, 2, \ldots, N$ と項目 $j = 1, 2, \ldots, n$ に対する尤度 L は

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

となる．ここで，$\delta_{i,j}$ は，$\delta_{i,j} = (0, 1)$ をとる二値関数を表し，$\delta = 1$ は正答，$\delta = 0$ は誤答を表す．

統計的観点から，式 (1) は確率変数 θ_i に対し，未知パラメータ a_j，b_j を持つロジスティック確率分布であるが，ここで，a_j, b_j, θ_i はすべて未知として扱う．式 (2) で表される尤度 L を最大にすることで最尤推定値を求めることができる．**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 からなるマトリクスを式 (2) の尤度関数に代入し，数値的に a_j, b_j, θ_i を求める．

項目特性値と能力値を同時に求めることは，推定すべき未知パラメータの数が $2 \times n + N$ であるため，簡単ではない．これを解決する方法として，2 つの手法が考案されている．一つは，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] であり，もう一方はマルコフ連鎖モンテカルロ法 (MCMC) を応用した方法 [12] である．[13] のツールでは，両者の手法が組み込まれている．

Vol.2013-MPS-96 No.24
Vol.2013-BIO-36 No.24
2013/12/12

3. 適応型オンライン能力評価システム

適応型オンライン能力評価システムとは，インターネットを通して PC 上で試験を受験することで能力評価を行うシステムを意味する．能力評価および回答ごとの問題の選出には，IRT による特性値を利用できる．

たとえば，受験者がインターネットを通して適応型試験システムを受験したとき，もし最初の問題が正答であれば，次の問題はより難しい問題が選出される．誤答であれば，より易しい問題が選出される．このシステムの運用には，項目バンクと呼ばれるデータベースが必須である．項目バンクとは，IRT によってすでに推定済みの項目特性値を持つ項目群を意味する．これらの項目特性値は，事前にモニターテストなどを実施することで試験データを集め，そこから推定しておく必要がある．

適応型オンライン能力評価システムに既知の項目特性値を用いることで，このシステムの受験者は，一問回答するごとに，式 (2) の尤度関数と項目特性値をもとに，能力値が求められ，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提出することができる．項目バンクに登録された項目群の特性が多様であるほど，能力に適した項目が提出されやすくなり，より少ない項目数で精確な能力評価が可能となる．

4. EM タイプ IRT [14]

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部分（欠損セル）の予測値とする予測手法である．そのために，δ を実数値に拡張し，さらに EM タイプ (expectation-maximization algorithm [3]) のパラメータ推定手順を示す．

4.1 δ の実数値への拡張

式 (2) において，δ は正答のとき $\delta = 1$，誤答のとき $\delta = 0$ を表す二値関数である．ここで，受験者が同じ困難度を持つ異なる項目 m 問中，l 回正答したと考えると，$\delta = l/m$ とみなすことによって，δ に対し，有理数を割り当てることができる．m を無限大とすれば，$\delta \in [0,1]$ の値を持つ実数値になる．

4.2 欠損セルに対する予測

まず，欠損セルに対し，$\delta^0_{i,j} \in [0,1]$ 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．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 などが挙げられる．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．

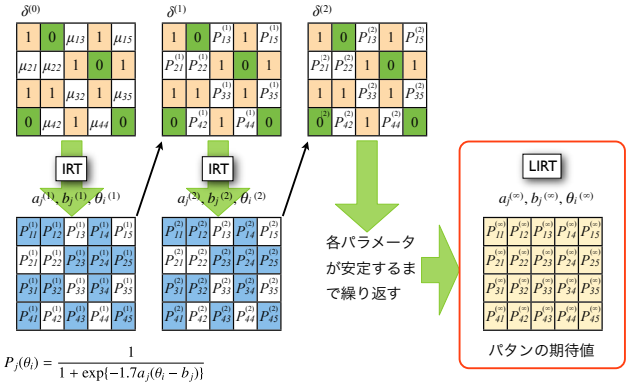


図 2 EM タイプ IRT による予測手順
Fig. 2 EM-type IRT prediction procedure.

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，maximization ステップに対応する．

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，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．この手法は，limiting IRT (LIRT) とも呼ばれる [10]．収束値が常に一意に決定するとは保証されない [16], [19]．しかしながら，経験的には，多くの場合で，更新ステップの初期段階において単調性がない場合があるものの，少なくとも収束することが分かっている [9]．**図 2** に，EM タイプ IRT による予測手順の概要を示す．

また，得られる予測マトリクスの精度評価に，RMSE (root mean squared error) S^k が用いられる．これは，次式で表される観測値とそれに対応する予測値 $\hat{\delta}^k_{i,j}$ の平均二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セルの数を表す．欠損セルの予測値は，S^k に含まれないことに注意する．

5. 適応型オンライン能力評価システムの実施

適応型オンライン能力評価システムでは，項目バンクに登録されている項目が多様であるほど，より精確な能力評価が可能になる．[14] 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

Vol.2013-MPS-96 No.24
Vol.2013-BIO-36 No.24
2013/12/12

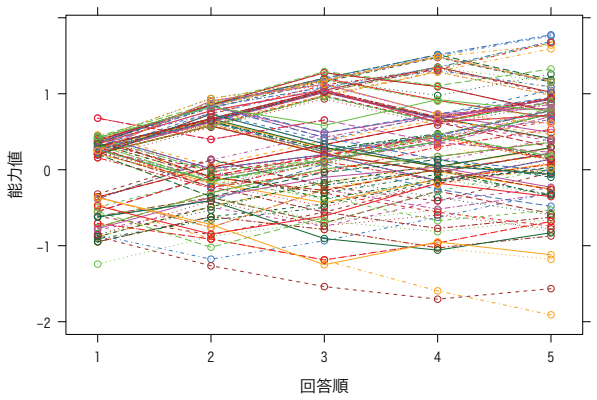


図 3 能力値の遷移
Fig. 3 transition of ability parameter.

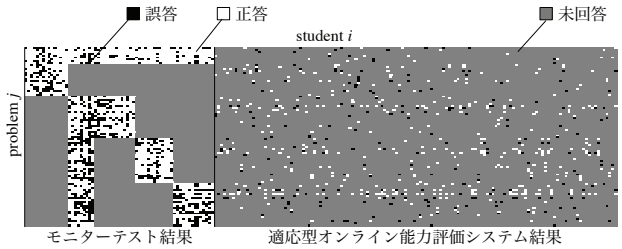


図 4 モニター結果と適応型オンライン能力評価システム結果
Fig. 4 The result of the monitor test and the adaptive online test.

項目バンクを作成し，より高精度な能力評価を可能とするシステムを構築する．

適応型オンライン能力評価システムを高校生を対象に実施する．一人 5 問回答してもらう．受験者数は 138 人であった．**図 3** に，回答順に能力値が遷移している様子を示す．横軸は回答順，縦軸は能力値を示している．各線が受験者を表す．各受験者に対する 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しい項目を選出するように工夫している．能力値にはさまざまな遷移の状況があり，5 問終了時の能力値もほぼ重複しないで評価できていることから，項目バンクの拡充の効果が見える．

各受験者は，それぞれ 5 問回答し，残り 91 問は未回答であるため，最終的に得られる項目および受験者全体のマトリクスは不完全マトリクスになる．その様子を**図 4** に示す．図中の白色のセルが正答を表し，黒色が誤答を表す．灰色のセルは欠損していることを表す．

6. 実データの予測

適応型オンライン能力評価システムによる試験データに対し，EM タイプ IRT および MF による不完全マトリクスの予測を行う．

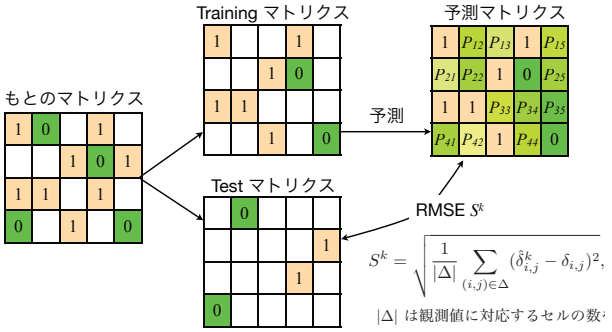


図 5 Training と Test に分けて推定を行う手順
Fig. 5 Estimation procedure using the training data and the test data.

6.1 Training と Test による評価

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 5** に Training と Test に分けて評価を行う手順を示す．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，無作為抽出による．Training と Test を 30 セット作成し，各セットの Training から予測を行い，Test によって RMSE を算出する．得られる 30 個の RMSE の平均および標準偏差を用いて，各手法の評価値とする．

表 1 にその結果を示す．表中のオンライン試験とは，適応型オンライン能力評価システムのマトリクスを示す．オンライン＋モニター試験とは，オンライン試験にモニターテストのマトリクスを合わせたものを示す．ここでは各手法の特徴を見るために，Training における RMSE も記載している．表 1 を見ると，Training の RMSE に対しては，MF のほうが明らかに小さいことが分かる．一方で，Test の RMSE に対しては，EM タイプ IRT のほうが小さい．特に，オンライン+モニター試験データの結果は EM タイプ IRT のほうがより小さく，予測精度が良いことを示している．また，オンライン試験とオンライン+モニター試験の Test の RMSE の変化を見ると，MF に比べて EM タイプ IRT の Test の RMSE のほうが変化が大きい．オンライン試験とオンライン+モニター試験の欠損率（マトリクス全体の要素数に対する欠損セルの割合）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，つまりデータ数の影響を受けていると考えられる．これらのことから，今回扱った不完全マトリクスに対しては，EM タイプ IRT による予測手法のほうがより適していると言える．

Vol.2013-MPS-96 No.24
Vol.2013-BIO-36 No.24
2013/12/12

表 1 9:1 の Training と Test に分けた 30 ケースの RMSE の平均
Table 1 The mean of RMSE for 30 cases using the 90% training data and the 10% test data.

	EM タイプ IRT		MF	
	Training	Test	Training	Test
オンライン試験	0.340	0.521	0.0133	0.523
欠損率 95.1%	0.00440	0.0299	0.00116	0.0295
オンライン+モニター試験	0.380	0.443	0.145	0.501
欠損率 85.7%	0.00162	0.0154	0.00410	0.0178

上は平均，下は標準偏差

7. 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．そこで本節では，今回扱った 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 6(a) にオンライン試験データを用いた場合について，図 6(b) にオンライン＋モニター試験データ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．どちらの場合の RMSE も，30 ケースすべてが単調に減少している．

図 7(a) にオンライン試験データを用いた場合について，図 7(b) にオンライン＋モニター試験データを用いた場合について，30 ケースの Training による対数尤度 $\log L$ の収束の様子を示す．ここで言う対数尤度とは，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，欠損セル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する．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，単調性のない場合もある．しかし，その場合でも，少なくとも収束には至っている．この単調性と収束性については，今後さらに調査が必要である．

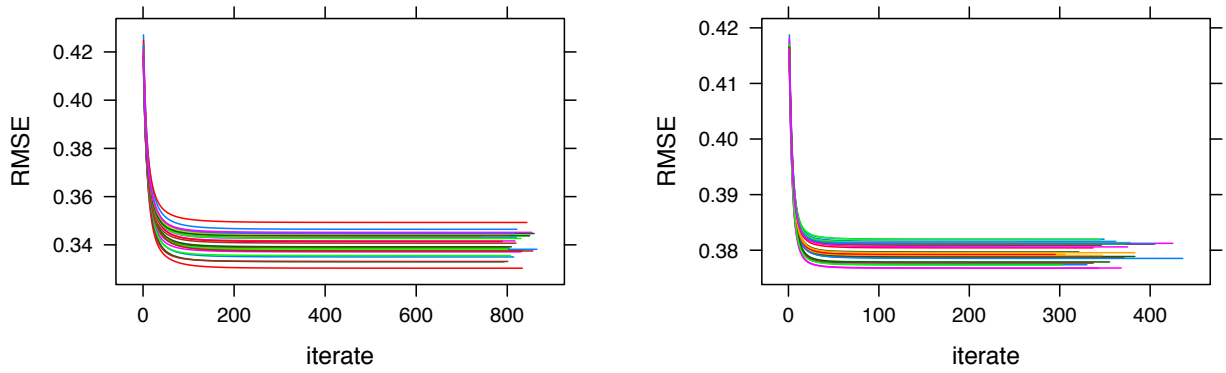
8. まとめ

IRT（項目反応理論）はテストの問題項目特性および受験者能力評価に対して有益な情報を与える．これを不完全マトリクスに適用する EM タイプ IRT が提案する．この方法は，不完全マトリクスの観測された値を用いて空要素の値を予測する．不完全マトリクスの一般的な予測法としてマトリクス分解を用いた方法が挙げられるが，受験者の特性が強く反映されるような不完全マトリクスの場合，EM タイプ IRT はより有効に働く可能性がある．ここでは，新たに得た不完全マトリクスデータに対して EM タイプ IRT による予測を行い，その有効性および収束性について検討した．

今回扱った不完全マトリクスは，適応型オンライン能力評価システムによって得られた試験データである．このデータの予測については，評価に RMSE を用いた場合，マトリクス分解法よりも EM タイプ IRT はより有効に働くことが分かった．また，収束性について検討した結果，すべての場合で収束に至っており，また単調性が見受けられた．この単調性については，今後さらなる調査が必要である．

参考文献

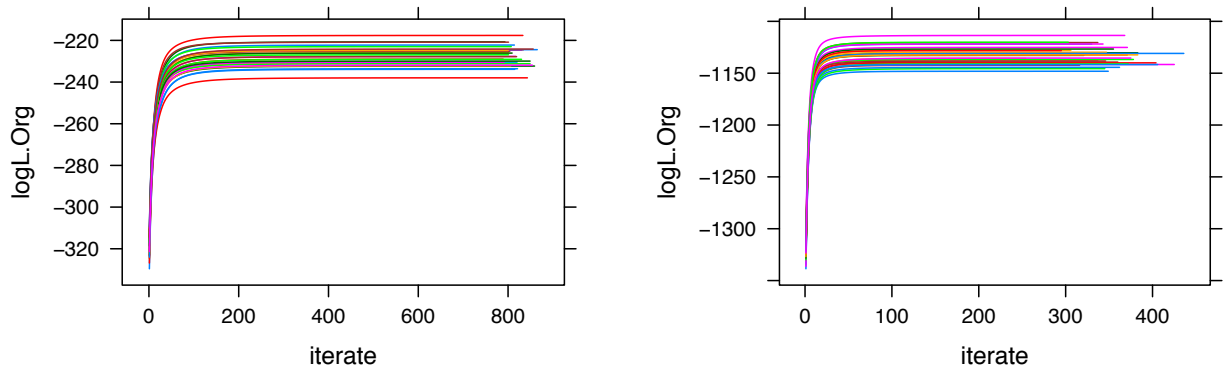
[1] Baker, F. and Kim, S.: *Item response theory: Parameter estimation techniques*, Vol. 176, CRC (2004).
[2] De Ayala, R.: *The theory and practice of item response theory.*, Guilford Press (2009).
[3] Dempster, A., Laird, N. and Rubin, D.: Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 1–38 (1977).
[4] Hambleton, R.: *Fundamentals of item response theory*, Vol. 2, Sage Publications, Incorporated (1991).
[5] Hambleton, R. and Swaminathan, H.: *Item response theory: Principles and applications*, Vol. 7, Springer (1984).
[6] 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).
[7] 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., The Royal Society, pp. 152–156 (2013).
[8] 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.*, ACTA Press, pp. 152–158 (2010).
[9] 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, *Teaching, Assessment and Learning for Engineering (TALE), 2012 IEEE International*



(a) オンライン試験　　(b) モニターテストとオンライン試験

図 6 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 の推移

Fig. 6 RMSE trend for 30 cases using the 90% training data and the 10% test data when using the EM-type IRT.



(a) オンライン試験　　(b) モニターテストとオンライン試験

図 7 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. 7 $\log L$ trend for 30 cases using the 90% training data and the 10% test data when using the EM-type IRT.

Conference on, pp. T1A–6–T1A–10 (2012).

[10] Hirose, H., Sakumura, T. and Ichii, S.: A recommendation algorithm that assumes a probabilistic structure and its application to questionnaire data, *in IPSJ SIG Technical Report.*, Fukuoka, Japan., pp. 1–7 (2011).

[11] Mills, C., Potenza, M., Fremer, J. and Ward, W.: *Computer-based testing: Building the foundation for future assessments*, Lawrence Erlbaum (2002).

[12] 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).

[13] 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).

[14] Sakumura, T., Kuwahata, T. and Hirose, H.: An adaptive online ability evaluation system using the item response theory, *in Education and e-Learning (EeL2011).*, Global Science and Technology Forum (GSTF), pp. 51–54 (2011).

[15] SSL: Bilog-mg, http://www.ssicentral.com/irt/index.html (2005).

[16] Suen, H. and Lee, P.: *Constraint optimization: An alternative perspective of IRT parameter estimation*, chapter 17, pp. 289–300, Norwood, NJ: Ablex. (1994).

[17] 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).

[18] van der Linden, W. and Hambleton, R.: *Handbook of modern item response theory*, Springer (1996).

[19] 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).