広島工業大学紀要研究編
第52巻（2018）103–108
論　　文

アダプティブテスティングにおけるIRT困難度の推定：

LCTの結果を用いた支援推定法

廣瀬　英雄*

（平成29年9月15日受付）

Estimation of IRT Difficulties in Adaptive Testing:

An LCT Assisted Estimation Method

Hideo HIROSE

(Received Sep. 15, 2017)

Abstract

It seems difficult to estimate the difficulties of the problem items in the adaptive testing even if we use the EM-type IRT (item response theory) method because the response matrix made by the adaptive testing becomes sparse which is not enough to estimate the item difficulties and examinees' abilities. In this paper, we propose to use the known ability values from other testing results if the examinees are the same. In the follow-up program systems, we have been using three kinds of testing, the LCT (learning check testing), the CWT (collaborative work testing), and the FPT (follow-up program testing), in which the LCT provide accurate estimates for ability values. We use the LCT's ability values in the CWT estimation. For stable estimation, we have used one-parameter estimation method (Rasch model) instead of two-parameter estimation method which is used in common. The estimation is performed well. In addition, comparing the method of 2017 CWT difficulty estimation with that of the 2016, we have found that in estimating the difficulty parameters using artificial response matrix where 0 values are imposed in the vacant elements the estimates of difficulty parameters are heavily biased to difficult problem side.

Key Words: collaborative work testing, learning check testing, item response theory, dually adaptive IRT, EM-type IRT

1　はじめに

「愛あるって」[1–4]のようなアダプティブテスティングでは，通常，何らかの方法で事前に得られた問題の困難度を用いて受験者の能力あるいは習熟度を推定している。CWT（collaborative work testing）[5]もアダプティブテスティングであるが，そこでは受験者と問題がデータベースの中で増加していっても自動的にチューニングをしながら受験者の習熟度を問題の困難度を求める設計（dually adaptive IRT）[6]になっている。

ある程度の大きさの応答マトリクスにある程度のアクセス数があれば設計どおりの推定は可能であるが，システム稼働初期には推定に困難が伴うので，初期には「愛あるって」と同様，問題の困難度に適切な初期値を与えている。データベースへのアクセス数が増えると自動的なパラメータの推定は可能になると思われるが，次の節で示すように，

* 広島工業大学データサイエンス研究センター＆環境学部建築デザイン学科

実際には，稼働からある程度の時間が経っても応答マトリクスが密にならないことがある。このとき，問題の困難度の初期設定値から自動推定による設定値に移行する間の暫定的な措置として，アクセスの多かった問題に対してだけでも問題の困難度を推定してデータベースにセットしておくことは，以降の習熟度推定値の推定精度を上げる意味では意味のあることである。

2016年度には，CWT のアクセス履歴から，受験者と問題が集中した応答マトリクスの部分を両方ソーティングして，アクセス密度の高い部分マトリクスを抜き出し，問題の一部の困難度を部分的に求めていた[5]。アクセスのあった827問の問題数に対して推定できた問題数はそれほど多くなく167問だけであった。非常に疎な（スパースな）マトリクスを用いてのパラメータ推定は，マトリクスの大きさが Netflix のコンペ[7] で示されたようなマトリクスの要素数が10億というように巨大な場合には推定は可能であるが，CWT では1000人が2000問にアクセスするという程度で（小さくもないが）それほど巨大ではない。中途半端で推定は困難のままである。

そこで，今回は2017年度のアクセス履歴（約21000アクセスで2016年度は53000アクセスであった）を用いて昨年度より多くの問題の困難度を推定することを試みる。

2　CWT の応答マトリクス

2017年度前期（2017年 4 月から 7 月）に CWT にアクセスされた応答マトリクスを示す。縦方向は受験者 id，横方向は問題 id を表す。図では応答があった場合には黄色から赤色に色付けしているが，多くの要素への応答数は小さいので黄色部分だけが多く見える。図 2 は図 1 の青く囲った部分を拡大してものである。応答マトリクスが疎であることがわかる。

このような場合，EM タイプ IRT[8,9] を用いても推定は非常に困難である。

3　LCT の結果を利用した困難度の推定法

アダプティブテスティングのデータベースが単独に存在しているのであれば2016年度に推定したような方法になるが，幸いなことに，フォローアッププログラム[10] のシステムでは，3 つのオンラインテスティングが同時に動いておりアクセスする学生は共通である。そこで，他のデータベースから求められた受験生の習熟度を既知のものとして CWT の問題の困難度を推定する利用することを考える。

ここで，上で示した方法が何を表しているかを明確にするために，IRT の理論的な背景について振り返ろう[11–17]。IRT では，各問題 j に対する受験者 i の評価確率 $P_j(\theta_i)$ がロジスティック分布，

図 1　2017年度前期の CWT 応答マトリクス（ほぼ全体）
（縦方向は受験者 id，横方向は問題 id を表す）

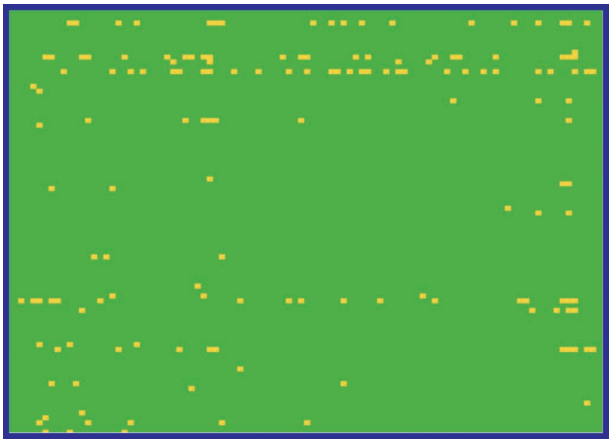

図 2　2017年度前期の CWT 応答マトリクス（図 1 の拡大）
（縦方向は受験者 id，横方向は問題 id を表す）

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

に従っていると仮定している。a_j，b_j，c_j は問題 j の識別力（簡単にいうと，問題の良し悪しを表す），困難度（文字どおり，問題の難易度を表す），当て推量（偶然に正答する確率を表す），θ_i は受験者 i の学習習熟度（ability）を表している。1.7（もっと正確には1.702）は分布が標準正規分布に近くなるように調整された定数である。受験者 $i=1,\cdots,N$ が項目 $j=1,\cdots,n$ に対して取り組んだ結果，その解答が正答なら，$\delta_{ij}=1$，誤答なら $\delta_{ij}=0$ と書き表すと，すべての受験者がすべての問題に挑戦した結果（これを反応パターンという）の確率は，独立事象を仮定すれば，$c_j=0$ と仮定した場合，

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

と表された。これを尤度関数といった。誤答 0 と正答 1 からなる δ_{ij} を上の尤度関数 L に代入し，それを最大にするような a_j，b_j，θ_i を同時に求めるのが IRT による基本的な評価法であった。このときには，最初に a_j，b_j を求める際には θ_i の周辺分布を用い，次に求められた a_j，b_j を用いて今度は θ_i を推定するというように 2 段階の推定法を用いていた。

さて，$\delta_{i,j}$ は 1 回だけのアクセスなら 1 もしくは 0 になるが，何度も同じ問題にアクセスすることもあるため，尤度関数を

$$L=\prod_{i=1}^{N}\prod_{j=1}^{n}P_j(\theta_i)^{s_{i,j}}(1-P_j(\theta_i))^{f_{i,j}}$$

のように書き改める。ここに，$s_{i,j}$ は $f_{i,j}$ はそれぞれ，正答した回数，誤答した回数であり，数値は非負の整数と考えてよい。また，θ_i は LCT の結果から得られた推定値を用いる。もう少し詳しく説明すると，15回の授業で毎回実施される LCT の結果から15回分（実際には13とか14回になることもある）の応答を同時に用いた応答マトリクスを作り，通常の IRT を用いて問題の困難度と習熟度を同時推定した結果を用いるということである。LCT 応答マトリクスは要素に空欄がない完全マトリクスであるため，IRT の計算は容易である。

θ_i を既知として a_j，b_j を推定することは比較的容易である。ただし，$s_{i,j}\cdot f_{i,j}=0$ のときには推定が困難になるため，ここではそのような問題については推定を断念している。また，推定をより安定的に行うため，ここではて $a_j=1$ のように仮定して 1 パラメータ推定を行なっている。このようにしても CWT を実施する際の受験生への問題の（難易度の）誘導にはそれほどの問題は起こらない。

図 3 に，LCT（解析基礎 A）から求めた受験生の習熟度の分布を示す。図 3 ではところどころ −2.5の値が固まって見えるが，LCT の受験に参加していない学生が主である。その内容は，授業欠席の場合と2015年度以前でフォローアップのメンバーになっていない学生である。そのような LCT に参加していない学生の情報は CWT にも関係が薄いので大きな影響は与えない。

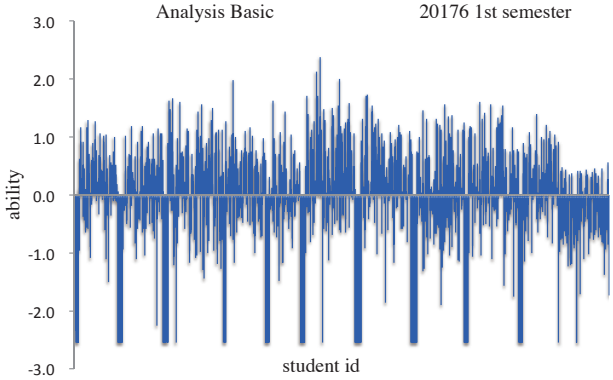

図 3　2017年度前期の LCT での受験生の ability 値
（解析基礎 A）

このように LCT から得られた受験生の習熟度を援用して求めた問題の困難度（b）の値を図 4 に示す。横軸は問題の id を表す。正答率が0.7程度であることを考えると平均的に b の値に負の値が多いことは自然である。

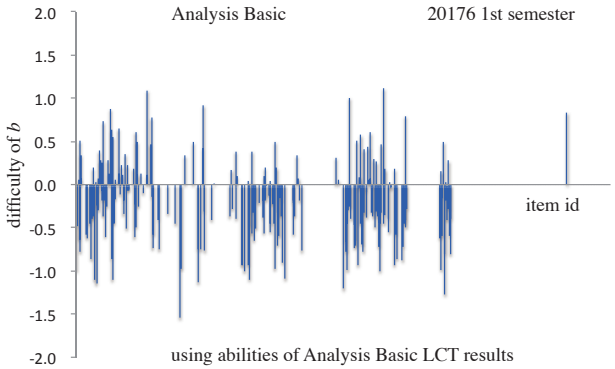

図 4　2017年度前期の CWT 問題の困難度（b）の値
（解析基礎 A に解析基礎 A の LCT 結果を利用）

また，同様な方法で求めた線形代数での受験生の習熟度の分布と推定された問題の困難度を図 5，6 に示す。

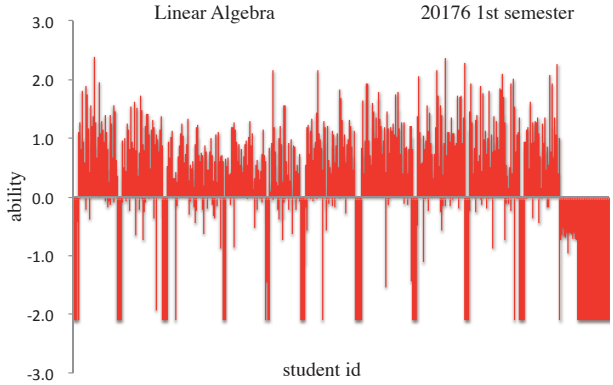

図 5　2017年度前期の LCT での受験生の ability 値
（線形代数 A）

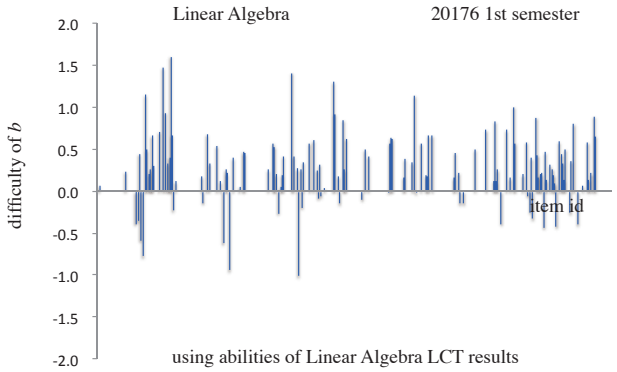

図 6　2017年度前期の CWT 問題の困難度（b）の値
（線形代数 A に線形代数 A の LCT 結果を利用）

4　考　察

4.1　援用する習熟度のカテゴリーにおける注意点

図 4 と 6 を比較すると，図 4 では比較的困難度の値が小さく，図 6 では比較的大きいことが分かる。図 3 と 5 の受験生の習熟度を比較すると，解析基礎 A よりも線形代数 A の方が高いことがおおよそわかる。実際に両者を比較したものが図 7 であるが，相関はあるものの線形代数 A の問題の方がよく解けていることがわかる。

では，線形代数 A の問題の困難度も低めに計算されてもよさそうな気がするが，ここで注意することがある。LCT 受験生は新入生全員（ほぼ1100人）で，全員受験しているため推定値に偏りはない。しかし，CWT は主にフォローアッププログラムに出席しながら演習を行なうということもあり，受験者の層が LCT とは異なり偏っている可能性がある。LCT から得られた能力値はある程度の参考にはなりそうであるが，その程度は見極めておく必要がある。

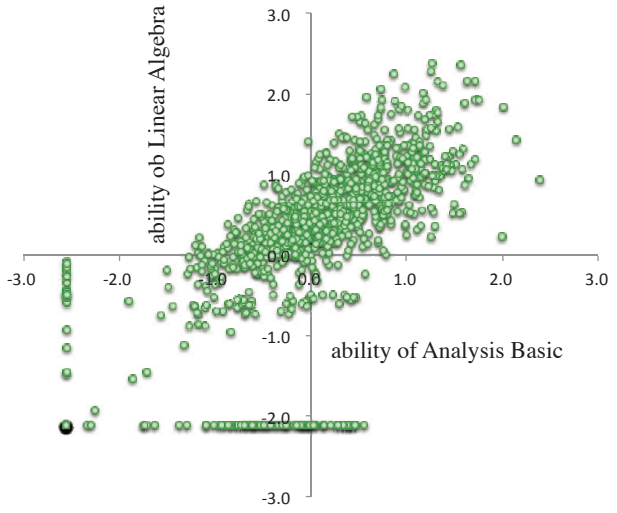

図 7　LCT での受験生の ability 値の比較
（解析基礎 A と線形代数 A）

そこで，２つの試みを行った。１つは，解析基礎 A の CWT の困難度推定に線形代数 A の LCT の結果から得られた能力値を用いること。もう１つは，能力値情報として何も与えないこと，つまり，すべての b を 0 として計算を行うことである。

図 8，９に解析基礎 A の CWT の困難度推定に，線形代数 A の LCT の結果から得られた能力値を用いた困難度（b）の推定値と，能力値情報として何も与えないときの困難度（b）の推定値を示す。また，図10，11に線形代数 A の CWT の困難度推定に，解析基礎 A の LCT の結果から得られた能力値を用いた困難度（b）の推定値と，能力値情報として何も与えないときの困難度（b）の推定値を示す。

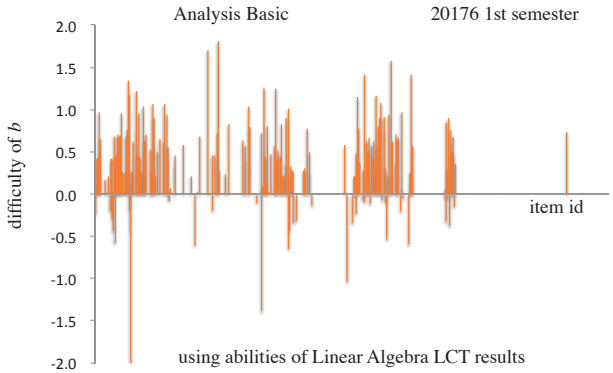

図 8　2017年度前期解析基礎 A の CWT 問題の困難度（b）の値
（解析基礎 A に線形代数 A の LCT 結果を利用）

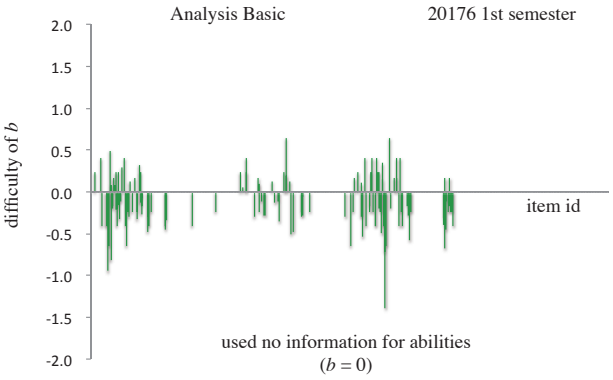

図 9　2017年度前期解析基礎 A の CWT 問題の困難度（b）の値
（解析基礎 A の困難度の情報なし）

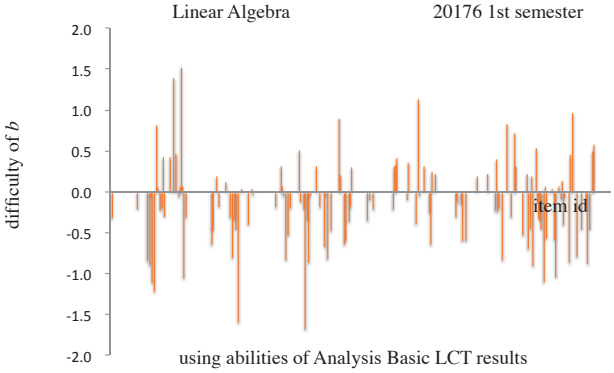

図10　2017年度前期線形代数 A の CWT 問題の困難度（b）の値
（線形代数 A に解析基礎 A の LCT 結果を利用）

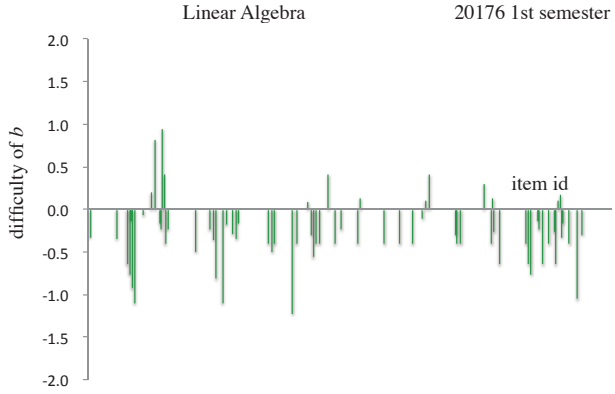

図11　2017年度前期線形代数 A の CWT 問題の困難度（b）の値
（線形代数 A の困難度の情報なし）

このように，LCT での困難度を与える際には LCT の問題の性格もある程度把握しておく必要があることを示唆している。もう少し説明しよう。

線形代数の LCT では易しい問題が多かったので ability の値は平均的に正の方向に推定されている（図５）。

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

での (θ_i-b_j) で正答か誤答かの確率が決まるので，b_j の推定値は θ_i の集団の影響を受けることになる。この集団の ability の平均は不明であるが，CWT の正答率は平均で0.7なので CWT 受験者の層はある程度高かったと思われる。従って，LCT での問題の難易度と CWT の問題の難易度を比較すると CWT の方が高かったとも考えられる。解析基礎 A ではこの違いがあまり出ていなかったのかもしれない。つまり，解析基礎 A と線形代数 A の LCT の問題の困難度の違いが影響していることも考えられる。

そこで，事前情報を与えずに困難度を推定すると 0 付近に引き寄せられてマイルドな推定値が出てきているためこの情報を使うというのも 1 つの方法と考えられる。

これまで，マトリクス（user, item）の大きさが（1100, 100）の（完全）応答マトリクスから ability を推定し，その値を利用して大きさが（1100, 1000）のスパースな応答マトリクスでの difficulty の推定を試みた。上で示したように 2 段階の推定では近似値を求める意味ではある程度妥当な気がするが，もう少し精密には等価（equating）を行なう必要もあるのだろうと考える。しかしながら，中規模でのスパースな応答マトリクスで通常の等価は使えない。EM タイプ IRT の方法も使えない。そうすると，暫定的な一方法として利用することはあながち間違いでもないように考える。

4.2　IRT 計算における注意点

この計算法は簡便に見えるが非常に有効な推定法である。さて，アダプティブテスティングでは通常の IRT のように応答マトリクスのすべての要素に応答値（通常 1 か 0）が入ってはいない。代わりに，ある問題には何回のアクセスがあって，アクセス者の ability もおよそわかっているとき，この問題への解答の正否の数（自然数）を用いて計算する。しかし，IRT の計算法をそのまま踏襲してアダプティブな計算を行なうのは正しくない。その場合には，応答マトリクスの空欄のところに強制的に数値（普通 0）を入れることがあるが，これは誤った結果をもたらす。0 という情報は解答しなかった（アクセスしなかった）のではなく，解けなかったという意味だからである。

例えば，スパースなマトリクスからソーティングによって密な部分マトリクスを作り，そこでの EM タイプ IRT を用いて推定値を求めたとする。密な部分が多いのでアクセスのなかったところに 0 を入れて計算しても大きな変化はないものだろうと思われるが，きちんと全体に対する空欄の割合を求めてからそのような計算を行わなければならない。このようにして求めた問題の困難度（b）の値を図12に示す。全体的に見ると平均的に b の値に正の値が多い。この理由は次のとおりであると思われる。

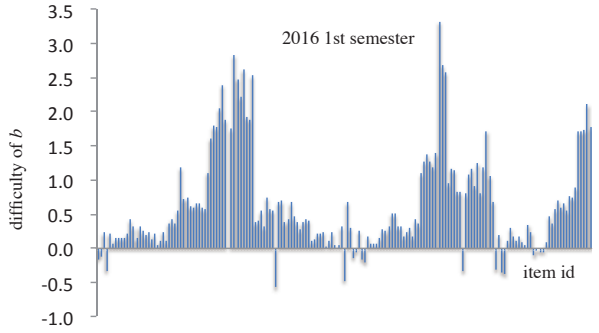

図12　2016年度前期の CWT 問題の困難度（b）の値

図13に，空欄がどの程度あったかを示す。縦軸は全体（密なマトリクスの部分の問題数 =167問）に対する空欄の比を表す。密な部分を取り出したようであるが，空欄のままで計算せずに 0 入れて計算したために問題の困難度が大きく推定されてしまっていることがわかる。これは全くの推定誤りである。

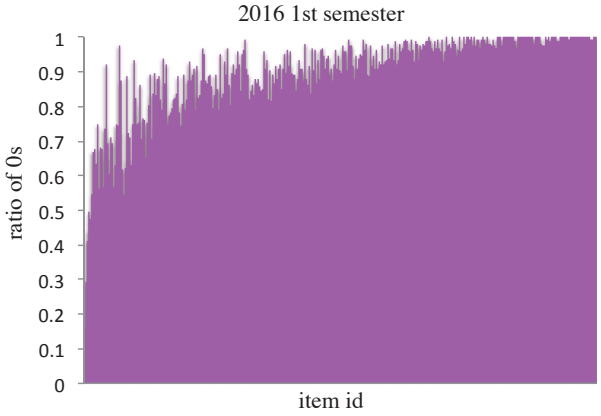

図13　密な部分マトリクスでの問題毎の 0 値の比率

5　まとめ

アダプティブテスティングを用いて受験者の習熟度を推定するときには，通常，既知である問題の困難度を用いている。困難度が不明な場合でも推定が可能になる dually adaptive IRT システムを用いればこの困難は回避されるが，それには応答マトリクスが相当大きくなければ安定に推定ができないという面がある。そこで，ある程度大きいが巨大ではない応答マトリクスでも，他のテスティング結果からの受験者の習熟度を援用して問題の困難度が推定できるようにできないか試みた。その結果，１パラメータのラッシュモデルを使えばかなり安定的に比較的妥当な推定値が得られることが分かった。この推定値は次にアダプティブテスティングを使う際の問題の困難度に用いることができる。

援用しようとするテスティング結果を用いる際には，受験者の性格を把握しておくことも重要であることを示唆した。また，アダプティブテスティングと通常の IRT の計算とを区別しなければ誤った計算を行なう可能性があることを，スパースな応答マトリクスをソーティングによって密な部分マトリクスを形成し EM タイプ IRT を用いて計算した問題の困難度は難しい方に偏って推定されることにも注意した。

文　献

1）大澤，桑田，田中，藤原，廣瀬，小田部，理工系学生のための基礎物理学 -Web アシスト演習付，培風館，2017.

2）桂，岡崎，岡山，齋藤，佐藤，田上，廣門，廣瀬，理工系学生のための微分積分 -Web アシスト演習付，培風館，2017.

3）廣瀬，藤野，確率と統計 -Web アシスト演習付，培風館，2015.

4）桂，池田，佐藤，廣瀬，著理工系学生のための線形代数－Web アシスト演習付，培風館，2015.

5）廣瀬，ラーニングアナリティクス：フォローアップ演習（CWT）の場合，広島工業大学紀要教育編，pp. 149-155, Vol. 51, 2017.

6）H. Hirose, Y. Aizawa, Automatically Growing Dually Adaptive Online IRT Testing System, IEEE International Conference on Teaching, Assessment, and Learning for Engineering 2014（TALE 2014）, 5C_5, pp. 528-533, 2014.

7）S. Takimoto, H. Hirose, Recommendation systems and their preference prediction algorithms in a large-scale database, Information, Vol. 12, No. 5, pp. 1165-1182, 2009.

8）H. Hirose, T. Sakumura, Item Response Prediction for Incomplete Response Matrix Using the EM-type Item Response Theory with Application to Adaptive Online Ability Evaluation System, IEEE International Conference on Teaching, Assessment, and Learning for Engineering 2012, pp. 8-12, August 20-23, 2012.

9）作村，徳永，廣瀬，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 , 情報処理学会論文誌，数理モデル化と応用 Vol. 7, No. 2, pp. 17-26, 2014.

10）廣瀬，新入生全員を対象としたオンラインテストの実際，広島工業大学紀要教育編，pp. 27-35, Vol. 16, 2017.

11）R. K. Hambleton and H. Swaminathan, Item Response Theory: Principles and Applications. Springer, 1984.

12）R. Hambleton, H. Swaminathan, and H. J. Rogers, Fundamentals of Item Response Theory. Sage Publications, 1991.

13）W. J. D. Linden and R. K. Hambleton, Handbook of Modern Item Response Theory. Springer, 1996.

14）月原，鈴木，廣瀬：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.

15）作村，徳永，廣瀬，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，情報処理学会論文誌，数理モデル化と応用 Vol. 7, No. 2, pp. 17-26, 2014.

16）H. Hirose, T. Sakumura, Item Response Prediction for Incomplete Response Matrix Using the EM-type Item Response Theory with Application to Adaptive Online Ability Evaluation System, IEEE International Conference on Teaching, Assessment, and Learning for Engineering 2012, pp. 8-12, August 20-23, 2012.

17）H. Hirose and T. Sakumura, An Accurate Ability Evaluation Method for Every Student with Small Problem Items Using The Item Response Theory, Proceedings of the International Conference on Computer and Advanced Technology in Education（CATE 2010）, pp. 152-158, August 23-25 2010.