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

月原由紀・鈴木敬一・廣瀬英雄

抄録

e-learning システムは授業を支援するツールとして普及してきたが，学生の能力を適切に評価できると考えられている項目反応理論（IRT）を用いてテスト評価まで組み込んだシステムは緒についたばかりで，特に大学数学については適用例が少ない。本稿では，1）大学低学年での数学についての学生の能力を適性に判断できるように，テスト結果を IRT 評価と素点評価について比較検討した結果，IRT 評価が優れていることを確認し，2）IRT 評価システムを e-learning システムの中に組み込んで，学生の習熟度を正確に把握し，またそのことによって学生の学習意欲を向上させるためのシステム構築を図る準備を行っていることを報告する。

◎Key Words 項目反応理論, e-learning, 大学数学テスト

A small implementation case of the mathematics tests with the Item Response Theory evaluation into an e-learning system.

Yuki Tsukihara, Keiichi Suzuki, Hideo Hirose

Abstract

Although e-learning systems have widely been used to assist the university lectures, those systems with tests evaluation using the IRT (item response theory) have just begun, and such a case is not so well known, in particular, for mathematics education in universities. In this paper, 1) we describe the superiority of the IRT evaluation results over the raw score evaluation by comparing these two methods, and 2) we show a small e-learning system in which the IRT evaluation system is included to grasp the abilities of students toward mathematics and to enhance the interest to learn mathematics.

Keywords: IRT, e-learning, tests of mathematics in universities

1. はじめに

大学低学年で習得する基礎教育としての数学は，主に学内の数学担当教員によって行われているが，小規模な大学や単科大学などでは非常勤に頼ることも多い。その場合，各科目の内容が統一的になるように詳細な検討を行い，テストによる公正な評価法についても十分議論することが望ましいが，時間的な制約もあって実際にはあまりなされていないように思われる。個々の担当教員だけによる評価はまちまちになりがちで，そのため，学生には数学能力が適正に評価されている実感が得られず，学習意欲の低下にもつながっている。数学能力を適正に評価するしくみが求められている。

連絡先：Contact to：廣瀬英雄

更に，入試の多様化や少子化のため，学生の学習意欲や能力の個人差も年々大きくなっているように感じられる。講義の理解や単位の取得には一定以上の学力水準が求められていることには変わりがない訳で，学力や意欲の向上を支援していくためには，学生の能力に応じた個別のケアも必要になってきている。

学習能力を適正に評価できる方法として提案されている項目反応理論（IRT）[1-6]は，英語能力判定として TOEFL や TOEIC テストに使われ，定着している。しかし，大学数学などでは，正解不正解だけの 2 値応答の問題を作りにくいこともあってか，IRT を積極的に使って学生の能力を測ろうとしている例はまだ実用的なレベルでは少ないようである。ここでは，IRT 評価法を数学のテストにも適用することによって，学生の数学能力の

評価を適正に行い，学生の学習意欲を高めるため，IRTをe-learningに組み込むシステム構築を小規模ではあるが行ったので報告する。

第2節では，ここで使っているIRTの理論的な背景を簡単に述べ，続く第3節で，新入生対象に数学一斉テストを行った結果を用いながらIRT評価が単純な素点評価よりも優れていることを示し，第4節では，基礎的な大学数学の問題を，最近流通しているe-learningシステム（Moodle [11]）に組み込んだテスト評価を行った結果について述べる。

2. IRT（Item Response Theory）

従来の古典的テスト評価法は，問題数は複数であっても各問題にあらかじめ配点を与えておき，総合的な評価は各問題における得点を合計したもので与えられるため，個々の問題の特性による評価を加味した評価法とは言えない。つまり，難しい問題に解答しても易しい問題に解答しても評価は配点だけで決められため，学生の習熟度と問題の難易度とを両方勘案した評価法にはなっていない。IRTは，問題の特性（難易度）と学生の能力とを切り離して考えることによって，個々の問題に対する正解率をこれら両方から計算しているので，問題の特性と学生の能力とを同時に取り扱うことができる総合的な評価法になっている。従って，古典的評価法では同点であっても，難問に解答した割合が多い場合にはIRT評価値は高くなる傾向にあり，学生にとっては自分が適正に評価されたという実感が得られやすくなる。

2.1 IRTによるモデルについて

テストは一般に複数の問題から構成される。ここでは個々の問題を項目と呼ぶ。

学生iが能力値θ_iを持ち，テストのj番目の項目に正解できる確率$P_j(\theta_i)$は，(1)式の2パラメータロジスティックモデルで近似できると考える[1-6]。

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

a_jは学生の能力が正答確率にどの程度敏感に反映するかを表すと考えられるので，項目の識別力と呼ばれる。通常は正の値を示す。この値が大きいほど，能力値が高い被験者と低い被験者との正解率に大きな差が生じる。b_jは，困難度を表す指標になる。値が大きいほど難しい。定数1.7はロジスティック分布が標準正規分布に近くなるようにするための係数である。横軸を能力値θに，縦軸を正答確率$P(\theta)$としたグラフを項目特性曲線と呼ぶ。

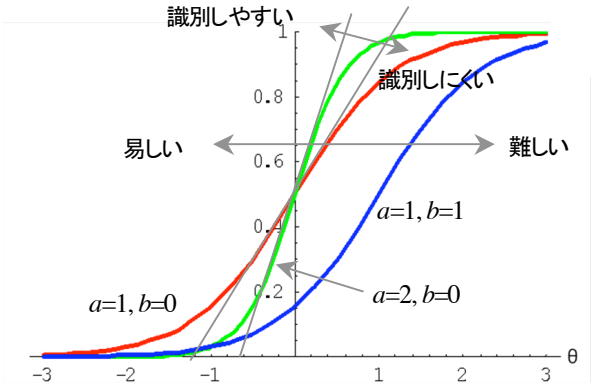


Fig. 1. 項目特性曲線の例

Fig.1に，aとbとを変えた場合の項目特性曲線のいくつかを示す。

2.2 パラメータの推定法

(1)式はある問題とある学生とが与えられたときの正答率を表すので，すべての問題とすべての学生の正答率それぞれが独立に得られたと仮定すると，尤度（L，likelihood function）(2)が得られる。ここにδ_{ij}はインジケータ関数で，学生iがテスト問題jに正解すれば1をそうでなければ0を与える。未知パラメータは最尤推定法により求めることができる。

$$L = \prod_i \prod_j P_j(\theta_i)^{\delta_{ij}} \cdot ((1 - P_j(\theta_i))^{\delta_{ij}} \qquad (2)$$

項目パラメータと被験者パラメータがともに未知である場合，未知数であるパラメータ数が多いため，(2)式のLを最大化するようなパラメータを単純に推定することは一般に困難になる。そこで，項目パラメータa_j，b_jと，能力値θ_iとを以下に示すような方法によって求める。

2.3 項目パラメータの推定法

個々の学生の能力パラメータが未知のままでは，サンプルが増えるだけ未知パラメータも増えることになり，安定して項目パラメータ推定を行うことができない。そこで，(3)式のLを最大にするように個々の学生の能力値を積分消去して，a_jとb_jとを求める周辺最尤推定法によりパラメータ推定を行う。

$$L(U \mid a,b) = \prod_{i=1}^{N} [\int_{-\infty}^{+\infty} g(\theta) \prod_{j=1}^{n} L(u_{ij} \mid a_j, b_j) d\theta] \qquad (3)$$

ここに，$g(\theta)$はすべての学生に共通する能力値分布として標準正規分布を仮定し，Uはu_{ij}を要素に持つ学生全員の正誤パターン行列とする。(3)式中には積分が入っているため直接解析的に求めることが難しい。そこで，区分

求積法を用いて能力尺度をいくつかの離散的な求積点に分け，そこでの観測頻度で重み付けを合計することによって近似的に周辺尤度を求める方法をとる。データ数が多ければ求積点の数を増やすことによって近似の精度を高めていくことができる。このような計算手法として，EM アルゴリズム[2, 3]を用いる。EM アルゴリズムは，欠損値をあたかも観測されているものとして扱いパラメータを推定する手法である。ここでは，積分消去された学生の能力値パラメータ θ が観測されているものとして，(3)式を最大化するように項目パラメータ a_j，b_j を推定する。

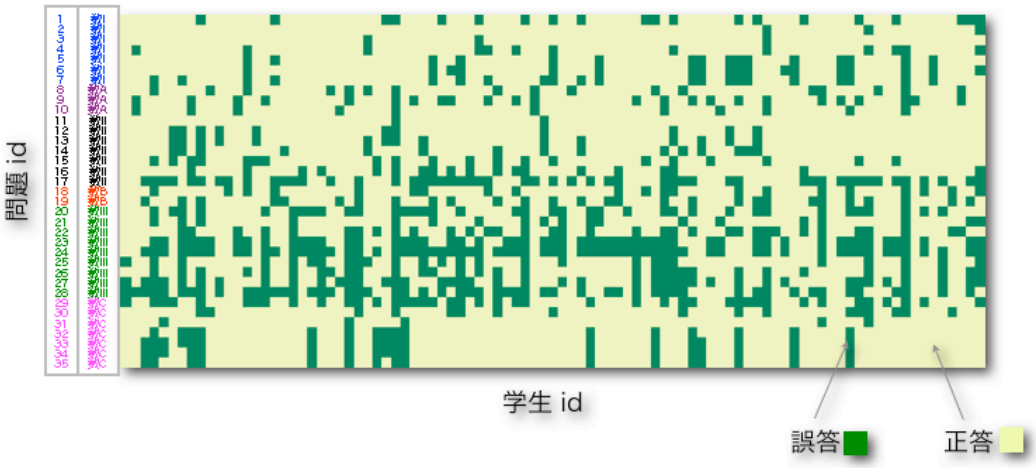


Fig. 2.　ある学科での正答･誤答の反応パターン

2.4　能力値パラメータの推定法

学生が満点(あるいは0点)をとった場合，最尤推定法では正しい推定が行えない。そこで，これを解決する推定法としてベイズ推定法[3-5]がしばしば用いられる。ベイズの定理から，項目反応 δ が得られた時の θ の事後分布 $f(\theta \mid \delta)$ は(4)式のようになり，この事後分布のモードにおける θ を求める推定値とする。

$$f(\theta_i \mid \delta_i) \propto g(\theta_i) L(\delta_i \mid \theta_i) \qquad (4)$$

ここで，$g(\theta_i)$ は i 番目の学生の能力値の事前分布であり，標準正規分布に従うと仮定する。$L(\delta_i \mid \theta_i)$ は i 番目の学生の反応パターンより得られる尤度関数である。学生の能力値はベイズ法を用いて推定する。

3.　IRT 評価と素点評価との比較[7,8]

3.1　数学一斉テスト

九州工業大学情報工学部に入学したばかりの新入生 432 名を対象に，高校で習得すべき基本的な数学の知識を問うテストを一斉に行った。テストには，数 I，II，III，数 A，B，C の中から問題をバランスよく 35 問を選んでいる。Fig.2 に，ある学科での正答・誤答の反応パターンを示す。図では上から順に，数 I，数 A，数 II，数 B，数 III，数 C と並んでおり，数 III の難易度が高かったことが直感的に見て取れる。

3.2　項目特性曲線

テストは被験者が紙に答案を書く一般的な記述式で行い，正答と誤答との 2 値情報に分けた結果をもとにして，学生の能力パラメータと項目パラメータとを推定した。推定の一部には Easy Estimation[12]を援用した。推定された項目パラメータから得られた項目特性曲線を Fig.3 に示す。推定された項目特性曲線からも，数 III の難易度が高く，問題の識別力も高いことが示されている。

3.3　IRT による能力値と素点結果

Fig.4 に，IRT により推定された能力値と素点とを比較した結果を示す。素点が高ければ能力値は高く推定されるが，同じ素点を取っても能力値には幅が観測されるこ

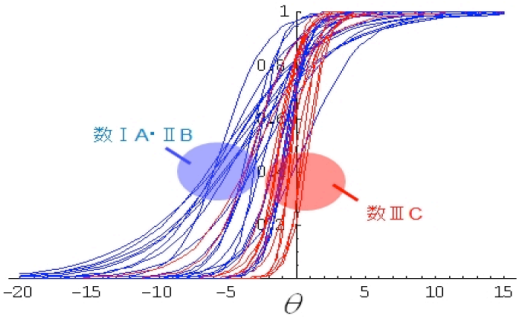


Fig. 3.　推定された項目推定曲線

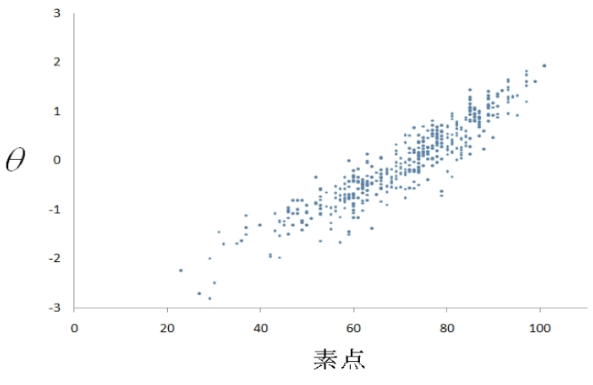


Fig. 4.　IRT による能力推定値と素点との関係

とが分かる。例えば素点60点では，θは0から−1.5までばらついている。$\theta = 0$でも，素点は60点から80点までばらついているので，能力と素点とが逆転することも十分考えられる。ただし，1回だけのテストでは誤差が含まれているので決定的な判断はできない。このような傾向がどの程度確率的に観察されるかを検証しておく必要がある。

そこで次に，真の能力$\tilde{\theta}$を持った学生がここで紹介したテストを受けたとして，IRTによる評価結果と従来の素点評価とが真の能力値にどのように反映されるかをシミュレーションによって確認した。項目パラメータ，能力パラメータは上で推定した値を用い，学生432名に35問を与え，同様のテストをm回行ったものと仮定した。反応パターンは付録に示すような方法により与え，学生の能力の推定値と素点とを求め，真の学生の能力値$\tilde{\theta}_i$と比較した。IRTの評価値の場合にはそのまま推定された$\hat{\theta}_{ik}$を用い，素点の評価値の場合には標準正規分布に変換した$\hat{\rho}_{ik}$を用い，真値との整合性を次の平均2乗平方根誤差（RMSE）で比較した。

$$RMSE(\theta) = \sqrt{\frac{1}{Nm}\sum_{i=1}^{N}\sum_{k=1}^{m}(\hat{\theta}_{ik} - \tilde{\theta}_i)^2} \qquad (5a)$$

$$RMSE(\rho) = \sqrt{\frac{1}{Nm}\sum_{i=1}^{N}\sum_{k=1}^{m}(\hat{\rho}_{ik} - \tilde{\theta}_i)^2} \qquad (5b)$$

$m = 10$の場合についてシミュレーションを行った結果，

$$RMSE(\theta) = 0.40 \qquad (6a)$$

$$RMSE(\rho) = 0.49 \qquad (6b)$$

が得られた。この結果は，IRTの能力推定値を素点の結果から同程度の精度で得るには，同じようなテストの回数が，$(0.49/0.40)^2 = 1.5$倍必要になることを示している。ここで用いたテスト問題は，学生の能力に必ずしも近い問題ばかりではないが，学生の能力に合わせ適切な問題が出題されるアダプティブテスト問題の場合，更に大きな違いが出てくると考えられる。従って，学生の能力を正確に，また効率的に推定するにはIRTは有効な方法であることが示された。

4. IRTのe-learningシステムへの実装[9,10]

TOEFLは，同じ場所で一斉にテストを行い得られたマークシート結果をバッチ的に処理するペーパーテスト方式から，最近ではコンピュータのサーバに個人的にアクセスするCBTのスタイルや，インターネットを介してテストを受けるiBTへと移行してきている。解答結果を直接入力できる利便性や効率性に加えて，個々の受験者の能力に合った問題を問題プールから自動的にダイナミックに選びながらテストを進めていく可能性があるこの方式では，能力推定の精度が上昇する効果がある。

そこで，大学で利用しているe-learningシステムの中にIRTを使ったテストシステムを組み込むことで，便利で効率的，効果的なシステム構築を目指すことを考えた。期末テストのような一斉テストだけでなく，日常の小テストや家庭での課題などにもこれを利用すれば，学生は自分の習熟度を確認しながら意欲的に学習を行うことも期待される。九州工業大学情報工学部では，学生は入学時からMoodleを利用しているため，ここではMoodleの中にテストシステムを実装する試みを行った。

4.1 テストシステム構築の流れ

難易度が分かっているテスト問題を多くプールしてダイナミックに問題を与えることができるシステムを作るため，Fig.5のようなテストシステムを構成したい。図で，①-⑤は，

① 項目分析：学生の反応パターンから項目の識別力と難易度，被験者の能力を推定

② 項目データベースの蓄積：項目データベースに項目とその識別力・難易度を登録

③ テストの構成

④ テストの実施：学生の能力を判定してフィードバック

⑤ パラメータの検証

を表す。

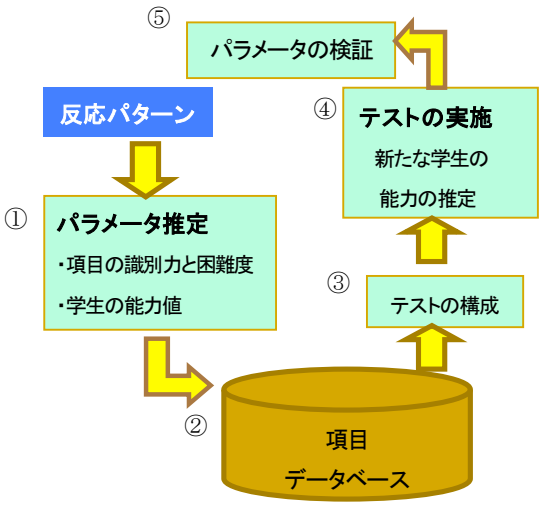


Fig. 5. IRTを適用したテストの流れ

4.2　小さいケースの IRT 評価システムの実装

大学低学年向けの解析学の問題を，四肢選択で答えられる，極限の問題，微分の問題，積分の問題の 3 つの分野に分け，Moodle を介して解答できる小さなシステムを構築した。答案は Moodle 内で自動採点され，1, 0 の正誤反応パターンを作成した後，あらかじめ登録されている項目パラメータの値を用いて，学生の分野別の能力値や総合的な能力値を推定するシステムである。能力値が何を表すか分からない学生のために，Fig.6 のように S, A, B, C の表示も同時に提示している。これが解答直後に提示されるため，学生には習熟度が即座に分かって学習意欲が向上するようである。

Fig.7 に Moodle を介したプロセスを示す。分野を分けることで, 分野毎の得意･不得意がその場で分かるため，

Fig.6　Moodle 上での IRT 評価結果の表示画面例

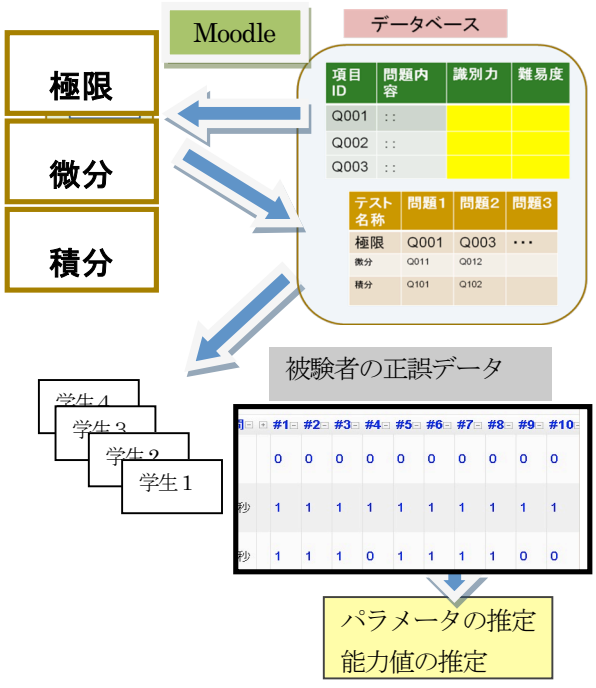


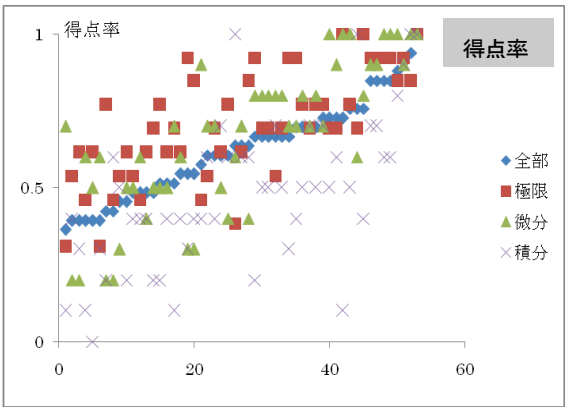
Fig.7　パラメータ推定のためのテストの実施

自分の弱点を把握しやすく，期末試験等への準備や心構えもできるようになる。

4.3　IRT の能力値と困難度の高い問題の正答傾向

Fig.8 では，3 つの分野すべてを受験した学生の得点と IRT による能力値の総合値を小さい順に並べたものと，分野毎のそれらとを比較した結果を示している。素点による評価では，比較的簡単な「極限」のテストで得点率が高い所に集中し，難しい傾向にある「積分」のテストでは低い所に集中している。同じ得点率の被験者も多い。一方 IRT による評価では，問題の組み合わせにより，テストに難易度の偏りがあった場合にも公平な評価になっていると考えられる。

そこで，同じ素点で評価された学生でも，項目の困難度 b が高い問題を多く正答した場合に IRT による評価が高いかどうかを調べてみた。同じ素点を持つ学生に対して，学生が正答した項目の困難度 b の平均と，推定された能力値 $\hat{\theta}$ との関係を Fig. 9 に示す。図では，先の新入生対象一斉テストにおける場合も併記している。図から，新入生対象一斉テストでは $\hat{\theta}$ と困難度の平均の間には明瞭な正の相関関係がみられることが分かる。一斉テスト



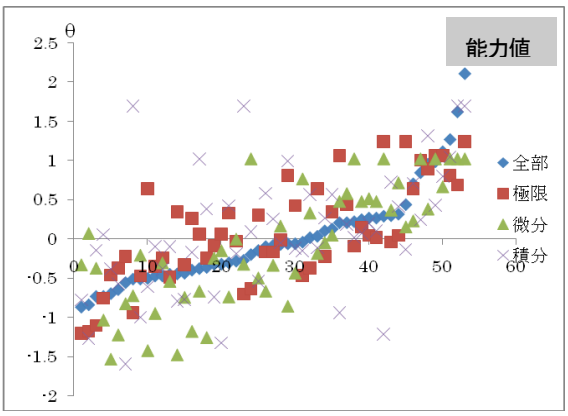


Fig. 8　得点率とIRTによる能力値比較

では監視がいるところで一定時間内に行っていることから，この傾向は信頼できるものと考える。つまり，同じ素点を持っていても困難な問題を解いた学生が能力を高く評価されることが分かった。一方，e-learning を利用した場合，「極限」の分野については上に述べた傾向は見られるものの，「積分」の分野についてはそれが見られない。「極限」の問題は比較的短時間で解きやすい問題であったため，まじめに取り組んだ傾向があるが，「積分」は困難度が高いためと，四肢選択にしているためか，時間をかけて解く前に当て推量で解答したのではないかと推察される。

5. まとめ

大学低学年での数学についての学生の能力を適性に判断できるように，従来素点評価されていたテスト結果を IRT 評価と素点評価とを比較検討することにより，IRT 評価が優れていることを確認した。また，この IRT 評価システムを学内で利用している e-learning システムの中に組み込み，学生の習熟度を正確に把握しながら学生の学習意欲を向上させるシステム構築を試みた。困難度の高い問題を多く解いた学生は IRT での能力値評価が高くなる傾向があり，IRT がより素点評価よりも適正に評価結果を出すことが分かった。また，e-learning システムに IRT を組み込むことで入力の手間が省けるだけでなく，学生へのレスポンスも早くなるため学習意欲の向上が観察された。今後，アダプティブなテストが可能になる程度の多くの問題が蓄積されていけば，学生の能力推定についてもより精度の高いシステムが構築できると期待される。能力を精度よく推定するためには多くの問題を解かせる必要があるが，数学の問題を数多く解くことは学生にとっても教員にとっても負担が大きい。IRT は効率的に能力評価を行ってくれるため，このシステムが限られた教員による教育効果を上げてくれるものと期待している。

本研究には，岡崎悦明教授、古賀雅伸准教授、乃美正哉助教、馬場晶子さん、後藤正人さんにご協力いただいた。ここに謝意を表する。

参考文献

[1] Baker, Item Response Theory: Parameter Estimation Technique, Marcel Dekker (1992)
[2] Linden, Hambleton (ed) : Handbook of Modern Item Response Theory, Springer (1996)
[3] 芝 裕順：項目反応理論 -基礎と応用，東京大学出版会（1991）
[4] 豊田 秀樹：項目反応理論・入門編 -テストと測定の科学，朝倉書店（2002）
[5] 豊田 秀樹：項目反応理論・理論編―テストの数理，朝倉書店（2005）
[6] 豊田 秀樹：項目反応理論・事例編―新しい心理テストの構成法，朝倉書店（2002）
[7] 鈴木，月原，廣瀬：IRT を用いた数学テストの評価，2007 年統計関連学会，p. 280（2007）
[8] 鈴木，月原，廣瀬：IRT を用いた数学の一評価，日

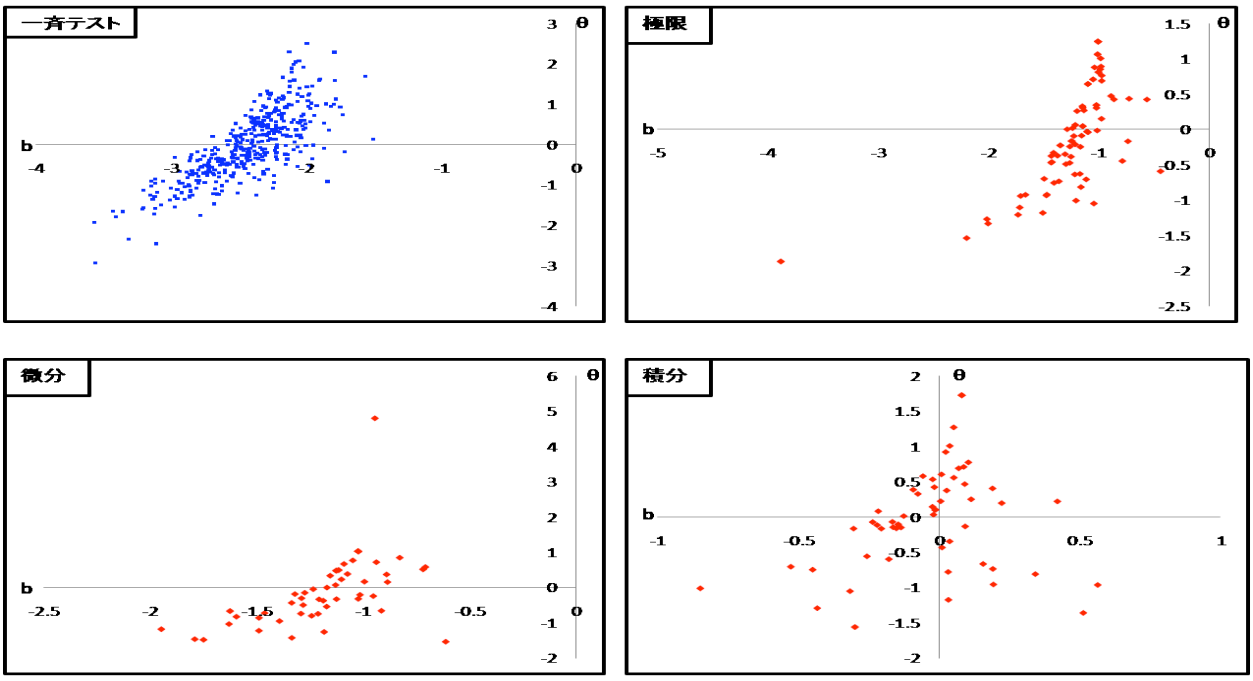


Fig. 9 θ と正答したbの平均値との関係

本行動計量学会第35回大会，pp. 99-100（2007）
[9] 月原，鈴木，廣瀬：IRTを援用したe-learningシステムへの試み：大学数学の基礎教育，電気関係学会九州支部第60回連合大会，10-2A-06, 2007, p. 371（2007）
[10] 月原，鈴木，廣瀬：IRTを用いた数学テストのe-learningシステムへの実装：分野別問題への適用，電子情報通信学会2008年総合大会，D-15-43（2008）
[11] http://moodle.org/
[12] http://irtanalysis.main.jp/

付録

IRT 評価と素点評価の比較のためのシミュレーション

数学一斉テストの条件と得られた推定結果とを初期設定とする。つまり，被験者数を432名，問題数を35問，項目パラメータ，被験者パラメータ，能力値のパラメータは先に推定した値を真値として用いる。

1） 反応パターンの作成

シミュレーションを行うために反応パターンを以下のように作成する。

学生iが能力値θ_iを持ち，テストのj番目の項目に正解できる確率$P_j(\theta_i)$としたので，一様乱数Rを発生させ，

$P > R$ならば$\delta = 1$

$P < R$ならば$\delta = 0$

とし，この学生の反応パターンを作る。

2） 標準化

次に，反応パターンをもとに素点評価とθ_iの推定を行う。このとき，2つの異なる評価方法を同等に扱うために標準化を行う。432名の推定値θ_i（あるいは素点評価 ）の値をρ_i，その平均と標準偏差をそれぞれμ，σとするとき，標準化後の値z_iは，

$$z_i = \frac{x_i - \mu}{\sigma}$$

で与える。

3） RMSE

(5) 式に従ってRMSEをIRTの能力値，素点それぞれについて算出する。被験者432人，推定回数10回であるので，$N = 432$，$m = 10$に設定する。

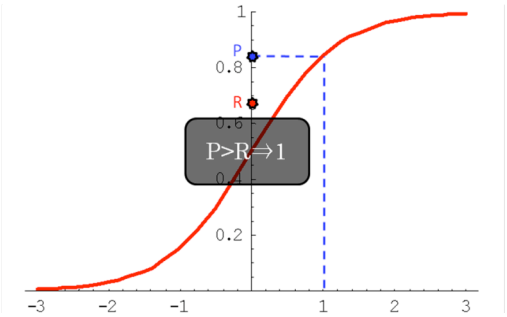


Fig. A PとRの関係

著者略歴

月原由紀（つきはらゆき）
現在の所属：九州工業大学情報工学部技術部
◎専門分野：データ科学

鈴木敬一（すずきけいいち）
◎現在の所属：日本IBM
◎専門分野：データ科学

廣瀬英雄（ひろせひでお）
◎現在の所属：九州工業大学情報工学部教授
◎専門分野：データ科学