K-035

項目反応理論を用いた信頼性のある能力評価システムの研究

Study on reliable evaluation system for achievement level using IRT

尾崎　将範*　　濱本　和彦**　　佐藤　実***　　野須　潔****
Masanori OZAKI*　Kazuhiko HAMAMOTO**　Minoru SATO***　Kiyoshi NOSU****

1．研究背景

インターネットの普及が近年著しくなるにつれ，教育目的に利用される学習コンテンツ，いわゆる e-learning サイトやコンテンツも増加の一途をたどっている．これらのサイト及びコンテンツは，当初は主に受講者が自学自習の為に利用していたが，普及するにつれ，達成度の評価や大学その他の教育機関での単位の認定などにも利用され始めた[1].

被験者の能力及び到達度の測定には，例えば設問ごとに配点を設定しその合計点を算出する方法や，偏差値を用いて被験者グループ内での相対的な能力分布によって評価をする方法などがあるが，問題作成者により配点にバラつきが出る，被験者の能力分布によって評価が大きく異なってしまう，得点の計算に時間と手間を要するなど，いくつかの問題点がある．

そのような問題点を解決するのに，標準化された問題群から項目反応理論（以下 IRT：Item Response theory）により出題する適応型テストの有効性が認められており，TOEIC や GMAT，情報処理技術者試験をはじめ，多くのテストに利用されている[2].

しかしながら，問題の標準化には問題の準備と試行に多くの予算や時間，その他のコストを要する．また標準化は過去の母集団に対して行われるため，受験者の母集団との分布の違いが問題となる．例えば大学の一クラスのような小規模かつ毎年違ったレベルの学生に対して行うような場合や新しい問題を次々と追加している段階の場合，妥当な能力評価や問題の難易度の決定をすることが出来る方法が確立されていないのが現状である．

2．研究目的

本研究では，比較的小規模の集団を対象に妥当な精度での能力測定を行う方法として，問題の標準化と能力判定を平行して行う適応型テストを開発する．この方法は，能力測定を大学等で使用されるＳ，Ａ，Ｂ，Ｃ，Ｆ程度の精度で算出し，あくまでも教育現場において e-learning を使用する際に，低予算かつ短期間で成績判定や単位認定等に有効な評価方法として利用することが目的である．

尚，先行研究では客観的に既に難易度の既知である問題に対する回答データから被験者の測定前評価を設定し，それに従って新たに追加した難易度が不知の問題の難易度の決定及び被験者の能力の測定を行う方法を開発し，既に物理問題によるシミュレーションにより一定の評価を得ている[3]．本研究ではそれをさらに発展させ，本手法による IRT を用いた到達度評価の実際に大学で行われている試験への適用可能性を明らかにすることが主目的となる．

*東海大学連合大学院理工学研究科総合理工学専攻情報理工学コース
**東海大学情報理工学部情報メディア学科
***東海大学理学部基礎教育研究室
****東海大学開発工学部感性デザイン学科

3．理論

IRT とは，従来のテスト利用される評価方法とは異なり，テスト一問ごとに，「仮にこの能力を持った被験者がいた場合，どの程度の確率で正解することが出来るか」ということを定義するものである．テストの困難度と被験者の能力との関係を示す曲線を IRT では項目特性曲線と呼ぶ（図１）．

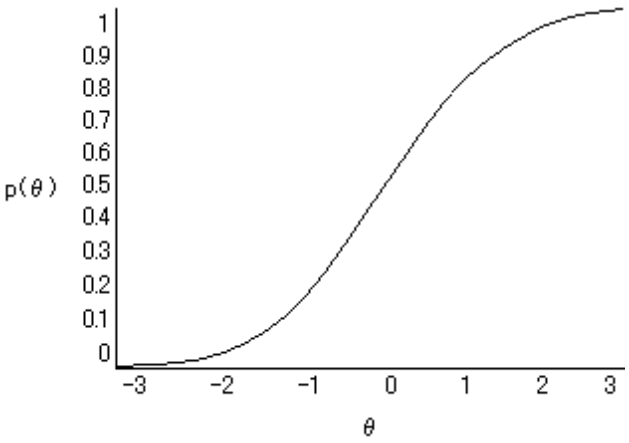

図１．　項目特性曲線

縦軸 p(θ)は，その問題の正解率であり，横軸θは被験者の能力を表す．つまり，能力の低い被験者は正答確率が低く，能力の高い被験者は，問題に正解する確率が高いということである．また，例えば図１のような問題があった場合，θが０の被験者がこの問題に正解する確率は 0.5，つまり 50%ということになる．

また，上記のような曲線は，式１のような数式で定義される．また，式に用いられる変数は，それぞれ以下のように定義される．

$$P_j(\theta)=c_j+(1-c_j)\frac{1}{1+exp(-Da_j(\theta-b_j))}$$

$$P_j(\theta)=c_j+(1-c_j)\frac{1}{1+exp(-Da_j(\theta-b_j))}$$

式１　３パラメータ・ロジスティック・モデルの数式

a：項目弁別力パラメータ

この値が大きくなると曲線の傾きは次第に大きくなり，被験者の能力を弁別する力が大きくなることを表す．逆に a が小さくなるほど曲線はなだらかになり，被験者の能力を弁別する力は小さいということが出来る．

ｂ : 項目難易度

この値が小さくなると曲線は左に，大きくなると右へ寄っていく．被験者の能力による正答確率を直接示す値である．

ｃ : 擬似チャンスレベル

例えば 4 択の場合，被験者が何も考えずに選択したとしても 1/4 の確率で正解することができる．これを避けるため，あらかじめ-3 の能力値を持つ被験者の正答確率を 0.25 に設定しておき，グラフの左端を 0.25 まで引き上げておくような場合に用いられる変数である．

尚，一般的に IRT を使用する場合，どのパラメータを用いるかによって現在までに 3 つのモデルが提案されている．

ｂのみを使用する１パラメータ・ロジスティック・モデル(!PLM)，ａ及びｂを用いる２パラメータ・ロジスティック・モデル(2PLM)，そしてａ，ｂ，ｃの全てを用いる３パラメータ・ロジスティック・モデル(3PLM)である．

今回の実験では，全て記述式の問題である通常の期末テストを用いたため，擬似チャンスレベルを含まず，かつ IRT の中で一番扱いやすいとされる 1PLM を用いることとする．

4. 実験

4.1 実験概要

本研究では，文献[3]で示された，難易度が既知である少数の問題，少数の被験者から，難易度が未知の問題の難易度を妥当な精度で設定し，それにより能力評価が可能であるとした方法を，実際に大学で行われている問題に適用可能か否かを検証する．これが可能であることが実証されれば，大学における e-Learning などにおいて，対面授業やスクーリングを必要としない，e-Learning のみによる信頼性のある成績評価が可能となる．

提案手法の前提条件は，難易度が既知の問題が少数用意されているとこである．よって本研究では，分野として成熟していると考えられ，本学においても既に 6 年間の開講実績がある，「プログラミング実習」を検討対象として取り上げる．

現在東海大学では，S,A,B,C,E の五段階のランク付けによる成績及び合否判定が用いられている．本実験では，まず，対象とするプログラミング実習の定期試験問題の配点による成績評価結果が，正解率を難易度と仮定した IRT により算出した成績と一致することを確認する．それらが一致するということは，当該テストの問題は，既に難易度が既知の問題として取扱うことが可能ということを意味する．

次にこの問題の一部を用いて被験者の測定前能力を計測し，その結果を基に，他の問題(難易度が未知の新たな問題と仮定する)の難易度を算出，それが元々の難易度と一致することを確認する．これが一致するならば，難易度が既知の少数の問題群から難易度が未知の問題の難易度を設定することが可能であり，それによる成績判定も可能となることを意味する．

尚，本実験では，問題群として必要となる問題数についても検討を行う．

4.2 実験方法

本実験には，2006 年 1 月に実施された「プログラミング実習」の期末テストの結果を用いた．

当該講義は過去約 6 年に亘って東海大学電子情報学部（及び旧工学部）にて開講されている．本テストは 2006 年 7 月に当該講義の期末テストとして，受講生 45 名に対して実施された．問題数は 22 問で，配点の合計は 100 点である．なおこのテストは，作成時点では IRT によって結果を考察することは全く考慮されず，例年通りの意図，方法によって作成された．実施結果は下記の通りである（表１）．

表１　被験者の得点分布

評価	合否判定	得点（点）	人数（人）
S	合格	91～100	10
A		80～89	4
B		70～79	10
C		60～69	6
E	不合格	～59	15

次に，テストの実施結果を IRT によって解析した．具体的な手順は以下の通りである．

(1) 回答データを，正解を 1，不正解を 0 とする 2 値データに変換する．

(2) IRT では全問正解者及び全問不正解者のデータは除外する必要があるため，全問正解者の 5 名を除外し，40 名分のデータとする．

(3) IRT を用いて被験者の能力を算出する．尚，この段階で用いる項目難易度については，問題の正答率によって-3～3 を均等な数になるように設定している．

(4) (3)の計算によって設定された被験者の能力値の変化が安定する問題数を測定し，安定した時点での能力値を測定前能力と設定する．

(5) (4)で算出した測定前能力によって，項目難易度の再設定を行う．

(6) 再設定した項目難易度により，被験者の能力値を一人ずつ算出していく．この中で，能力値の変化が安定する人数を測定する．

4.3 実験結果

4.2(3)によって得られた測定前能力と，従来の採点方式によって得られた点数[表 1]の対比は以下の通りである[図 2].

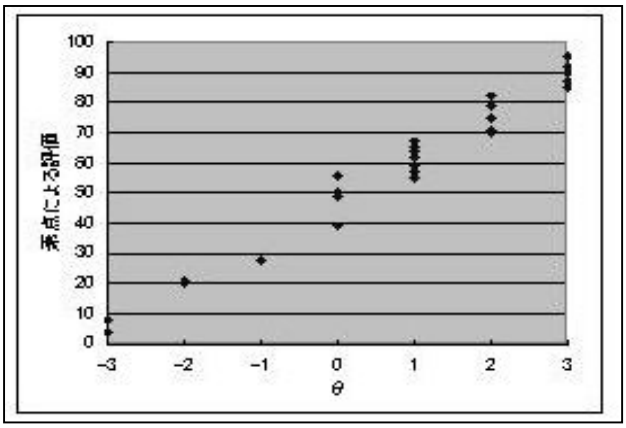

図 2　素点による評価及び IRT による結果の比較

図 2 からもわかるように，100 点満点による評価と測定前能力はほぼ比例している．これにより，従来の評価方法と遜色ない評価が IRT によって可能であると言える．

さらに，この項目難易度を用いて算出した測定前能力の変化の様子が，以下の通りである[図 3].

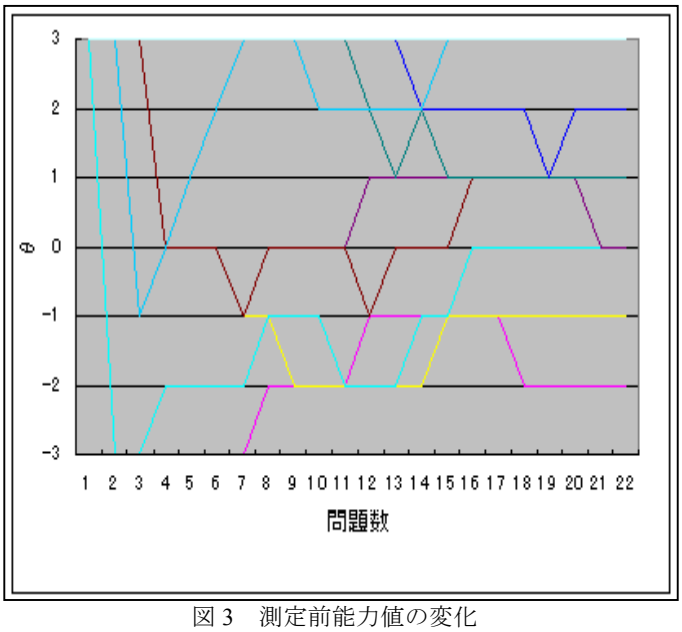

図 3　測定前能力値の変化

測定前能力を推定する場合には，少なくとも各難易度レベルの問題を各一問以上揃えなければならないのは自明である．今回は-3～3 の 7 段階であるため，まず，各難易度レベルから 1 問ずつピックアップした 7 問の問題群を準備し，問題を追加していく毎に能力を推定し，プラスマイナス 1 を許容範囲(成績判定は 4～5 段階であるため)内とし，必要とされる測定前能力を明らかにする．

この図では，それぞれの系統は，被験者ごとの能力値の変化を表している．図を見ると，-3 から 3 を各 1 問ずつ使用して行った段階で大きな変化は無くなり，次に 13～15 問目で小さな変化もかなり少なくなり，その後収束へ向かっていることが明らかである．従って，まず測定前能力を算出する際には，難易度が-3～3 の問題を一問ずつ含んだ 14 問程度の問題が必要であるといえる．尚，この数値は文献[3]の結果と一致する．

次に，この 14 問目の時点での能力値を測定前能力として使用し，項目難易度の設定を行った．結果は以下の通りである[図 4].

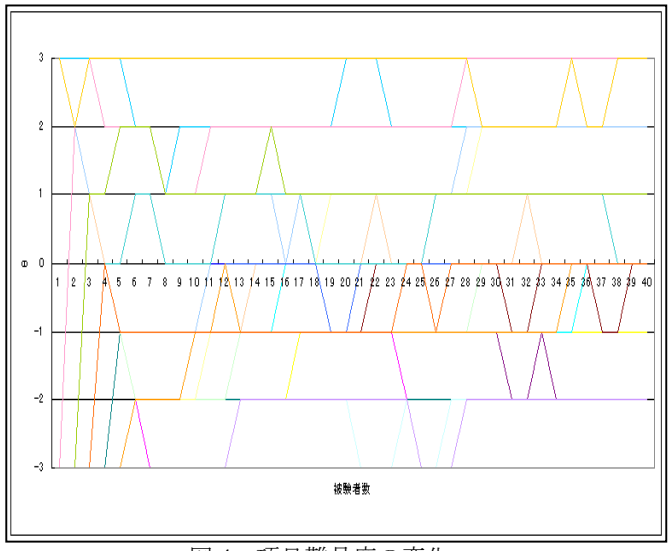

図 4　項目難易度の変化

この図では，それぞれの系統は，各問題の難易度の変化を表している．を見ると，やはり先ほどと同様に，7 人目で大きな変化が終わり，14 名程度の時点でほぼ小さな変化もなくなっている．その後，30 人程度を計算し終わった時点で，ほぼ数字が安定している．従って，難易度の再設定でも，最低でも 14 人程度，可能であれば３０人程度が必要であることが明らかとなった．

さらに，始めに設定した項目難易度と最終的に推定した項目難易度の差異を以下の表に表す．

表 2　当初の項目難易度及び推定した項目難易度の比較

問題番号	1	2	3	4	5	6	7	8	9
定義項目難易度	−2	−2	−1	0	−2	0	−2	0	2
推定項目難易度	−2	−2	−1	0	−2	0	−2	0	2

10	11	12	13	14	15	16	17	18	19	20	21	22
−3	1	2	2	3	−3	1	0	1	2	3	0	0
−2	0	2	2	3	−2	0	0	0	1	3	0	0

この表を見れば明らかなように，正解率によって設定した項目困難度と最終的に推定した項目困難度は，ほぼ一致し，一致していないものも全てプラスマイナス 1 の範囲内に収まっている．以上の結果から，今回提案した手法の有用性は明らかである．

また，今回取り上げたプログラミング実習の講義は，能力別クラス編成等は行っておらず，能力的に完全にランダムな集団である．このような集団にも今回提案した適応型テストを問題なく適用できることが明らかになったことは，本研究の大きな収穫である．

5.まとめ

本研究では，比較的小規模な集団において，多数の問題群を持たずに新たに e-learning システムによって成績評価等を行うような場合に IRT を用いた適応型テストを活用することを提案した．

実験において，７段階のレベル分けにおいては，１４問程度の問題，及び１４人から３０人程度の被験者が存在すれば，信頼性のある評価を実施することが可能であることが発見できた．

今後の課題としては，本実験において利用したプログラミング実習のように，既に数年間の実施による出題ノウハウがあり，かつ知識のみを問う問題の割合が多いテストだけではなく，まだ問題として熟成されていないものを扱う場合や，いくつかの知識を複合した理解力を問うテストでどのように評価をするかということを研究する必要があると思われる．

最後に，昨年度，本研究の基礎となる研究を行い，多くのデータや手法を提供してくれた近藤彰幸君に，多大なる感謝の意を表します．

文　献

[1] 特定非営利活動法人日本イーラーニングコンソシアム，ｅラーニング白書 2006/2007 年度版，東京電機大学出版局，東京都，2006.

[2] 大友賢二，項目応答理論入門，大修館書店，東京都，1996.

[3] 近藤彰幸，濱本和彦，尾崎将範，佐藤実，野須潔，項目反応理論を応用した信頼性のあるオンライン学習システムの開発，信学技報，IEICE Technical Report ET2006-53(2006-11).