短時間で能力ランクを判定するための Moodle プラグインの開発

秋山 實[*1]
Email: akiyama@ei.tohoku.ac.jp

*1：東北大学大学院　教育情報学教育部

◎Key Words 潜在ランク理論，コンピュータ適応型テスト，Moodle

1. はじめに

近年，ｅラーニングやインストラクショナルデザインなど新しい教育のツールや方法論が普及してきており，適切なタイミングで学習者の理解度や能力を効率的かつ短い時間で測定することが必要になっている．

1.1 テストの長さと精度のトレードオフ

全ての受験者が全ての問題を受験する従来のテスト（以後，リニアテストと呼ぶ）の測定精度は，テストを構成する問題の数が多いほど高くなる．しかし，問題が多いテストは時間がかかる．

表１　サンプル数とモデル適合度指数(RFI)

	受験者数		
ランク	160	80	40
3	0.773	0.776	0.659
4	0.814	0.822	0.728
5	0.827	0.838	0.725
6	0.842	0.849	0.760
7	0.844	0.856	0.763
8	0.847	0.860	0.767
9	0.844	0.858	0.757
10	0.845	0.851	0.768
11	0.839	0.852	0.761
12	0.831	0.847	0.750

1.2 コンピュータ適応型テストとその特徴

コンピュータ適応型テスト（Computerized Adaptive Test：CAT）は受験者の回答を基にその時点の受験者の能力を動的に推定し，次に出題する問題を選択・出題するため，受験者の能力から大きくかけ離れた易しい，あるいは，難しい問題は出題されない．したがって，リニアテストに比べ，半分以下の問題でリニアテストと同程度の精度で受験者の能力を測定できると言われている．

CAT を採用すれば測定精度を落とさずに短時間でテストを実施できるが，CAT には問題バンク（特性パラメータの値が予め推定された問題のデータベース）が必要であり，予備テストを実施しておく必要がある．多くの CAT が項目応答理論（Item Response Theory, IRT）に基づいて設計されており，適切に特性パラメータを推定するために必要なサンプル数は，最もシンプルなモデルである１パラメータロジスティックモデルを採用した場合でも 200 名，20 問以上の回答データが必要と言われている[2]．

1.3 潜在ランク理論の特長

潜在ランク理論[3]（Latent Rank Theory，LRT，後述）は，IRT に比べモデルの制約が少ないため，少ないサンプル数で特性パラメータを適切に推定でき，IRT を適用できないサンプル数が少ない状況にも適用することができる[4]．

表１は「日本語を読むための語彙量テスト」[5]（後述）の回答データを受験者数 160 名からランダムに削除して，80 名，40 名と減らした時の各ランクにおけるモデル適合度指標の一つである RFI[6]（Relative Fit Index）を EXAMETRIKA[7]によって算出したものである．RFI は 0 から 1 あるいはそれ以上の値をとり，大きければ大きいほどモデル適合が良いと判断できる．本研究ではデル適合度が良好であると判断する基準を RFI が 0.7 以上（アンダーラインで示す）とする．表１をみると広い範囲でモデル適合度は良いといえる．IRT は適用できないような，受験者数 40 名，問題数 150 問であっても LRT の場合，特性パラメータの推定が可能である．

問題数についても同様のことが言える[4]．問題数 50 問，受験者数 40 名でもランク数 5 から 15 までの範囲で上記と同様のモデル適合度が得られる．

2. 潜在ランク理論

LRT は，発表の当初はニューラルテスト理論[10]（Neural Test Theory，NTT）と呼ばれていた．「テストの解像度は連続値で表わすほど高くない」[3]と考え，LRT では受験者の能力を 5 から 20 程度のランクという離散値で表わす．これは順序尺度である．順序尺度は順序にだけ意味があり，ランクの加減乗除などの演算はできない．これに対して IRT では受験者の能力を連続値で表わす．これは間隔尺度である．LRT において推定精度を定義する場合，順序尺度では差に意味がないので誤差の指標である RMSE（Root Mean Square of Error）を計算することはできないので，真値のランクと推定値のランクが一致しない率を使う．

受験者の能力を離散値で表すことで測定が粗くなるというよりは，問題の特性を表わすアイテム参照プロファイル（Item Reference Profile：IRP，図 1）や受験者の各ランクへの所属確率を表わすランクメンバーシッププロファイル（Rank Membership Profile：RMP，図 2）という概念の導入により，IRT よりも特性パラメータを豊かに表現できるという特長を持っている．

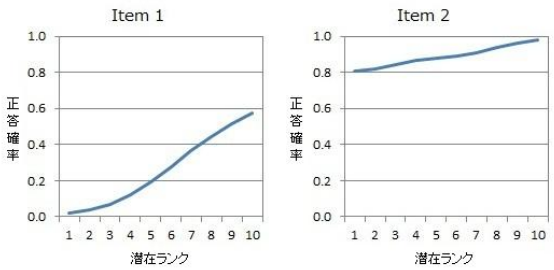

図 1　ランク数 10 の場合の IRP の例

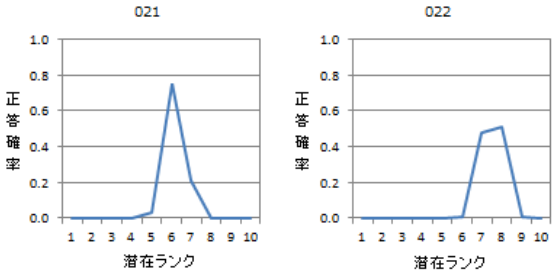

図 2　ランク数 10 の場合の RMP の例

2.1　IRP

IRP は，IRT の項目特性曲線に相当するもので，ランク数と同じ数の要素を持つベクトルで，それぞれのランクの要素は，そのランクの能力を持つ受験者がその問題を受験した場合に正答する確率を表している．

図 1 の Item1 は隣接するランクの正答確率の差が大きく，能力差を識別する力があり，難しい問題である．Item2 は隣接するランクの正答確率の差が小さいので，受験者の能力を識別する力が弱く、全体的に正答率が高いので易しい問題といえる．

2.2　RMP

IRT では能力値が一つの値で表されるが，LRT ではランク数と同じ数の要素を持つベクトルで表され，それぞれのランクの要素は，そのランクにその受験者が所属する確率を表している．

図 2 の受験者０２１は，ランク 6 の所属確率が最も大きく，隣接するランクの所属確率との差が大きいので能力ランク 6 であるが，受験者０２２は，ランク 8 の所属確率が最も大きいが，ランク 7 も次に大きく，ランク 7 からランク 8 に能力が変化していると見ることができる．（その逆の可能性もある）

2.3　ランク

受験者の能力を表すランクは，RMP の各ランクのうち最も大きい所属確率を持つランクとして定義されている．

受験者の能力をいくつのランクに区分するかは，IRT のモデルのパラメータ数と同じく，LRT で用いるモデルを決定することと同様の意味がある．異なるランク数の特性パラメータは比較することも，同じか否かを判定することもできない．

3.　評価

評価は，「日本語を読むための語彙量テスト」の 182 人の回答データとそれをもとに EXAMETRIKA で分析して得た EXCEL ファイル（IRP，RMP など特性パラメータと問題の名前を含む）を Moodle にアップロードして，シミュレーションを行った．

3.1　使用したテストと回答データ

「日本語を読むための語彙量テスト」は「日本語を読むための語彙データベース[12]」から 100 語につき 1 語の割合でサンプリングした 150 語を使って問題を作成し，15, 000 語までの語彙量を測定することができる．1，000 語ごとに五種（和語，漢語，外来語，混種語）と品種の割合を統制したサンプリングを行っている．

2010 年 5 月から 10 月にかけて日本，オーストラリア，ニュージーランドの 5 大学・機関の日本語プログラムで実施して得た回答データのうち，日本語学習歴のない中国語母語話者を除いた 182 名からランダムに抽出した 160 名のデータを使用している．

3.2　シミュレーションによる評価

LRT-CAT の性能を知るには，シミュレーションが役に立つ．シミュレーションでは，受験者の能力の真値がわかっているので，推定誤差を知ることができるからである．LRT-CAT に付加したシミュレーション機能を用いて，小さなアイテムバンクを用いた場合も含め，シミュレーションによる評価を行った．

サンプル数の小さい回答データは，必要なサイズになるよう受験者または問題を間引いて作成した．「日本語を読むための語彙量テスト」の問題は 1，000 語毎の語彙量で並んでおり，語彙の種類もバランスを取ってあるので，3 問おきに削除して 100 問，2 問おきに削除して 50 問の回答データを作成した．受験者の方はランダムなキーを付加してソートし前半を削除する方法で 80 名，40 名の回答データを作成した．これらの回答データをそれぞれランク数 4 から 14 までの 11 通りについて EXAMETRIKA を用いて特性パラメータを推定し，これを LRT-CAT の問題バンクとしてアップロードした．

評価の際のシミュレーション条件は，終了条件については，ΔRMP を 0.01 とした．

推定されたランクの一致率は概ね 85%を超えている部分をアンダーラインで示す．良好とまでは言えないが精度に関しては概ね有効であることを確認できた．受験問題数は受験者の 95%を含む範囲（平均受験問題数＋2 * 標準偏差を小数点以下を切り上げて 95%上限値として算出した）では，70%以下になるケースをアンダーラインで示す．（表 3）．

3.3　実地テストによる評価

LRT-CAT を使用して CAT を実施した結果と全問解答するリニアテストを同じ受験者に受験させ，推定された能力ランクと受験問題数を比較した．

東京の私立大学の留学生 129 名が，CAT，リニアテストの順で受験した結果を表 4 に示す．

実施条件は，シミュレーションと異なり，ΔRMP を

表 2. シミュレーションのランク一致率

		ランク数 4	5	6	7
160 名	150 問	0.98	0.93	0.89	0.83
	100 問	0.95	0.90	0.81	0.77
	50 問	0.90	0.82	0.74	0.67
80 名	150 問	0.98	0.95	0.91	0.87
	100 問	0.94	0.87	0.84	0.78
	50 問	0.82	0.85	0.78	0.74
40 名	150 問	0.93	0.82	0.79	0.77
	100 問	0.97	0.90	0.88	0.81
	50 問	0.87	0.77	0.82	0.65

表 3 シミュレーションの受験問題数(95.987 上限値)

		ランク数 4	5	6	7
160 名	150 問	66	92	106	105
	100 問	72	88	93	94
	50 問	50	50	50	50
80 名	150 問	57	85	101	103
	100 問	69	60	85	91
	50 問	49	48	49	50
40 名	150 問	66	77	90	95
	100 問	56	70	59	57
	50 問	45	44	48	48

表 4 実地テストの結果

	全体	ランク 1	2	3	4	5
一致率	0.72.1	NA	0.22	0.364	0.387	0.962
問題数	39	NA	28.2	24.8	21.5	14.4
受験者	129	0	9	11	31	78

表 5 リアルデータシミュレーションの結果

	全体	ランク 1	2	3	4	5
一致率	0.798	1.000	1.000	0.545	0.419	0.987
問題数	30	30	30	30	30	30
受験者	129	2	7	7	19	94

0.03 とした．受験者の母語は，76.9%が中国語，12.3%がベトナム語，3.1%がミャンマー語，7.7%がその他の言語であった．

実地テストの結果は，ランク 1 から 4 でランク一致率が悪い．そこで，リニアテストの回答データを用いてリアルデータシミュレーション（乱数を使って回答の正誤を決めるのではなく，受験者が実際に回答したリニアテストの回答を使って LRT-CAT のシミュレーションを行う）を最低受験項目数を 30 問，最大受験項目も 30 問に指定して実施したところ，各ランク毎のランク一致率が，1.000，1.000，0.545，0.419，0.987 と全体的に良い結果となった．特に，最低ランクと最高ランクで大幅に改善した．根本原因が LRT-CAT のアルゴリズムにある可能性は否定できないが，全員に 30 問受験させることで，150 問のテストの約 80%の精度で 30 問（リニアテストの 20%）という極めて少ない問題数でテストが実行できることが分かった．

テストの精度や受験問題数は，個々のテストの問題の質に依存している部分が大きく，適用するテストごとにシミュレーションを行って，問題を選択するアルゴリズム，能力を推定するアルゴリズム，終了条件などを最適に設定する必要がある．LRT-CAT には，シミュレーション機能があるので最適条件で実施することができる．

この LRT-CAT の利用例としては，以下のような手順で，1 回だけリニアテストを実施すれば，2 回目以降は，LRT-CAT で半分程度の問題数で同程度の精度のテストが実施でき，テストの短時間化が実現できると思われる．

1) リニアテストを作成し，実施する
2) EXAMETRIKA で分析し，目的やモデル適合度などでランク数を決定する
3) 結果の EXCEL ファイルを LRT-CAT にアップロードして，シミュレーションを実施し，最適条件を決定する
4) LRT-CAT で短時間でリニアテストと同じテストを実施する

現在は，2)と 3)の手順で EXAMETRIKA を使っているが，今後，LRT-CAT に分析機能を取り込むことで手順が簡略化でき，さらに実施しやすくなると思われ，改善を予定している．

Moodle のプラグインとして実装され，Moodle2.3 の環境で標準的な活動モジュールとしてインストールでき，コースバックアップ・リストアなどが可能である．配布は， http://moodle2x.info で行われている．

4. おわりに

本研究では，オープンソースソフトウェアの e ラーニングシステムとして日本の大学でも普及している Moodle のプラグインとして動作し，潜在ランク理論に基づくコンピュータ適応型テストを実施できるプラグイン LRT-CAT を開発した．この LRT-CAT の応用範囲は広いと思われる．たとえば，Moodle 上のコースで，毎回の授業終了間際の数分間で学習内容の理解度を測るテストを LRT-CAT で実施し，次回の授業でフォローするなどして落ちこぼれをなくす対策が実施できる．

謝辞

本研究に使用した「日本語を読むための語彙量テスト」の利用を快諾して下さった東京大学教養学部の松下先生に深く感謝します．

参考文献

(1) Dougiamas, M.: “Moodle”, <http://moodle.org>, (2012 年 12 月 6 日閲覧).

(2) リン, ロバート（池田央他編）: “教育測定学ハンドブック 第 3 版（上）”, pp.246-247, C.S.L.学習評価研究所, （1992）.
(3) 植野真臣, 荘島宏二郎 : “学習評価の新潮流”, pp.83-111, 朝倉書店, （2010）.
(4) 秋山實 : ” 潜在ランク理論のパラメータ推定に必要な問題数と受験者数 “, 日本テスト学会第 10 回大会発表論文抄録集, pp.178-179, （2012）.
(5) 松下達彦 :“「日本語を読むための語彙量テスト」の開発”, 2012 年日本語教育国際研究大会予稿集, 第一分冊, pp.310, (2012)
(6) 豊田秀樹編 : “共分散構造分析（AMOS 編）“, pp.245, 東京図書, （2008）.
(7) Shojima, K. : “EXAMETRIA 5.3”, (2008-2012), <http://www.rd.dnc.ac.jp/~shojima/exmk/>, (2012 年 12 月 6 日閲覧).
(8) 秋山實 : ” LRT モデルに基づく CAT の開発とシミュレーションによる特性解析”, 日本テスト学会第 9 回大会発表論文抄録集, pp.146-147, （2011）.
(9) 木村哲夫 : ”潜在ランク理論に基づくコンピュータアダプティブテストーアルゴリズムの提案と検証ー”, 日本テスト学会誌, 8, pp.70-84, （2012）.
(10) Shojima, K. : “Neural Test Theory”, DNC Research Note, 07-02（2007）, <http://www.rd.dnc.ac.jp/~shojima/Shojima2007RN07-02.pdf>, (2012 年 12 月 6 日閲覧).
(11) Shojima, K. : “Maximum Likelihood estimation of latent rank under neural test model”, DNC Research Note, 07-04（2007）, <http://www.rd.dnc.ac.jp/~shojima/ntt/Shojima2007RN07-04.pdf>, (2012 年 12 月 6 日閲覧).
(12) 松下達彦 : “日本語を読むための語彙データベース（総合版）”, Ver.4.0, < http://www.geocities.jp/tatsum2003/ >