

1. はじめに

能力測定の一つの可能性として、コンピュータを利用したテスト実施形態が考えられ、実際米国などにおいては、既にかかなりの部分が実用化されている。一方我が国においては、コンピュータを利用した大規模テストの応用例はまだ少ないが、その必要性は今後増大して行くものと考えられる。このような状況の中、当協会においても数年来、英語能力測定を目的とする適応型テスト（CAT）の研究開発を続けている。ここではコンピュータテストの応用例として、当協会のテストシステムを用いて、モニターを募集して行った実験の分析結果を示すことにより、コンピュータテストの有効性等に関する実験報告を行う。このテストに搭載された項目は、項目応答理論（IRT）に基づき、既知の項目特性値を持つ項目から構築され、これにより共通尺度による比較可能な能力値が推定された。被験者には同時に、同じデ-タベ-ス内から構築された、別問題からなる紙筆型テスト（PPT）の受験が課せられた。ここから同一人についての、適応型テストと紙筆型テスト両方のテストデ-タの収集が可能となった。また一部の被験者には、適応型テストを複数回受験してもらい、各テスト間のスコア変動の調査を行った。これらの実験から、コンピュータによる適応型テストシステムが、どの程度の信頼性を保持するものであるか、あるいは紙筆型テストに比べて、どの程度測定効率が高いものであるかを分析した結果を報告する。

2. 適応型テスト

コンピュータを利用したテストシステムは、広く CBT（Computer-based Testing）と呼ばれるが、今回開発したテストは、その中の適応型テスト（CAT: Computerized Adaptive Testing）と呼ばれるものである。このテストの基本的概念は、テスト施行中にコンピュータが提示項目ごとに被験者の能力を推定し、その都度アイテムバンクと呼ばれる項目集積群から、その被験者に最適な項目を選択して提示しながらテストを進行する方法である。これらの測定方式は、被験者毎に最適な項目を用いてテストを行うことにより、測定精度の向上が期待され、その結果、より少数の項目で能力測定が可能になるなどの利点がある。

当協会の適応型テストは、能力値が全く未知の被験者に対し、1問目に提示する項目として、アイテムバンクに収録されている多数の項目の中から、中程度の困難度の項目を提示する。仮に被験者がこの項目に正答した場合、次に提示する項目として、この項目よりやや難しい項目を提示する。或いは被験者がこの項目に誤答した場合は、次にはやや易しい項目を提示する。具体的には最初の1問の正誤に応じて、次の項目としては、その ± 0.5 程度の困難度の問題を提示する。それ以降は同様に、被験者の正誤反応

により、次に提示すべき項目をコンピュータが判断しながらテストを進行し、被験者の推定能力値がある一定の範囲に安定するまでテストを続ける。被験者の能力レベルに最適な項目を選択してテストを実施することにより、ある能力の被験者にとって易しすぎる項目や、或いは難しすぎる項目を提示することはなくなる。つまり易しすぎる項目や難しすぎる項目からは、その被験者の能力値に関する多くの情報は得られない。このように測定の無駄を排し、効率的に能力測定を行い、その結果精度の高い測定値を得よう、というのが適応型テストの基本的な考え方といえる。

実際に適応型テストを実施する際は、被験者の項目反応パターンと、提示された項目特性値をもとに、瞬時にその被験者の推定能力値を計算し、次に提示すべき項目を選択する必要がある。これら一連の作業には、コンピュータの利用が必須であり、パソコンなどの普及とともに実用化が加速されてきたテスト実施形態といえる。

3. 項目応答モデル

適応型テストを実施する際には、提示する項目の項目特性値（item parameter）が既知であることが前提となるが、さらに項目集積群に格納されている項目のそれぞれの項目特性値は、同一の尺度上で比較可能である必要がある。今回の実験に用いた版には、およそ 3000 の問題項目が搭載されているが、それらはすべて事前に予備テストが実施され（約 $N=2000$ ）、後に述べる項目応答モデルの3パラメ-タ・ロジスティックモデルを用いて最尤推定された項目特性値が付与されたものである。これらの項目特性値を得るために、過去 10 数回の予備テストを実施した。それぞれの予備テストの版は、別々の時期に異なる被験者群に対して実施されたものであるが、等化（equating）の作業を繰り返すことにより、すべての項目が共通尺度上で比較可能なものとなっている。さらに共通尺度上で評価可能になった項目に対する項目反応より、各被験者の能力値も共通尺度上で最尤推定することが可能となっている。

3パラメ-タ・ロジスティックモデルというのは、能力水準の被験者が項目 j に正答できる確率は、その項目が持つ3つのパラメ-タ a_j, b_j, c_j により、次式で与えられるというモデルである。

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

上式の項目特性値 a_j, b_j, c_j は、それぞれ a_j が識別力（discrimination power） $\cdot b_j$ が困難度レベル（difficulty level） $\cdot c_j$ が偶然正答レベル（pseudo-chance level）を示す。さらに式中の D は、ロジスティック曲線を、累積正規分布曲線に近似させるために用いられる定数であり、通常 $D = 1.7$ をとる¹⁾。またここでの最尤推定法は、既知の項目特性値を持つ項目群において、被験者が各項目に与えた解答

パタンのもとで、その出現確率が最大になるような \hat{q} を求める方法である。

いまある被験者が項目 j に対して与えた解答に応じて、それが正答なら $u_j = 1$ 、そうでなければ $u_j = 0$ を取るものとする。このとき、能力水準 θ の被験者が n 個の項目について、 $\mathbf{u} = (u_1, u_2, \dots, u_j, \dots, u_n)$ なる解答パタンを得る確率は、それぞれが独立という前提のもとで式(2)で与えられる。これを局所独立 (local independence) の仮定という。

$$L(\mathbf{q} | \mathbf{u}) = \prod_j P_j(\mathbf{q})^{u_j} Q_j(\mathbf{q})^{1-u_j} \quad (2)$$

これが能力値 θ の人がもつ所与の解答パタンでの尤度関数 (likelihood function) である。ここで、

$$Q_j(\theta) = 1 - P_j(\theta) \text{ である。}$$

上式において θ の値は未知であるが、 θ の取り得る範囲の中で、最大の尤度を持つ $\hat{\theta}$ の値を探し、その値 $\hat{\theta}$ を持つて所与の反応を示した被験者の能力推定値とする方法が最尤推定法 (Maximum Likelihood Estimation) と呼ばれるものである。 $\hat{\theta}$ を求めるには通常、式(2)の対数をとって、 θ について微分したものが 0 となるような解を求める²⁾。すなわち、

$$\ln L(\mathbf{q} | \mathbf{u}) = \sum_j [u_j \ln P_j(\mathbf{q}) + (1-u_j) \ln Q_j(\mathbf{q})] = 0 \quad (3)$$

上式は単純な線形一次式ではないので、通常は Newton-Raphson 法を使用した逐次解法を用いる。ここでは以下により解を求めた。なお式中の r は、試行回数を示すものである。

$$\hat{\theta}_{r+1} = \hat{\theta}_r - \frac{\left[\frac{d \ln L(\mathbf{q} | \mathbf{u})}{d \mathbf{q}} \right]_r}{\left[\frac{d^2 \ln L(\mathbf{q} | \mathbf{u})}{d \mathbf{q}^2} \right]_r} \quad (4)$$

上記の演算を繰り返し、 $\hat{\theta}_{r+1}$ と $\hat{\theta}_r$ との差が、事前の設定値より小さくなった時点で、十分な精度で推定出来たと考え計算を止める。さらに CAT では、能力推定の精度を誤差の標準偏差で求めて、その大きさを元に演算終了の指標とすることもある。ここでの事前の設定値は終了基準 (stopping rule) と呼ばれるものである。なお誤差の標準偏差の算出は、次のステップによる。

- ・被験者の項目 j に対する正答確率 $P_j(\theta)$ を用いて、項目 j の情報量 $I_j(\theta)$ を以下により算出する。

$$I_j(\mathbf{q}) = D^2 a_j^2 (P_j(\mathbf{q}) - c_j)^2 Q_j(\mathbf{q}) / (1 - c_j)^2 P_j(\mathbf{q}) \quad (5)$$

式(5)は項目 j の情報関数 (information function) と呼ばれるもので、その合計をテスト情報関数 (test information function) と呼ぶこともある。

- ・つぎに、最尤法によって推定された $\hat{\theta}$ と θ との誤差分散 $\sigma^2(\hat{\theta} - \theta)$ と、項目情報関数の和 $I_j(\theta)$ の間には、 $\sigma^2(\hat{\theta} - \theta) \approx 1 / \sum I_j(\theta)$ の関係が成り立ち、項目数が増えるにしたがって、それは漸的に項目情報関数の和 (す

なわちテスト情報関数) の逆数に近づくことが知られているので、それを用いて最尤法による推定誤差の標準偏差、つまり測定の精度を推定することができる³⁾。すなわち

$$\sigma(\hat{\theta} - \theta) \approx 1 / \sqrt{\sum I_j(\mathbf{q})} \quad (6)$$

4. CASEC

今回当協会が開発した適応型テストシステムは、「CASEC (Computerized Assessment System for English Communication)」と呼ばれるものである。それは4種類の出題形式から構成され、英語のコミュニケーション能力測定を目的としたものである。各セクションの構成は表1の通りで、また図1にはCASECの画面例を示した。

表1 各セクションの構成

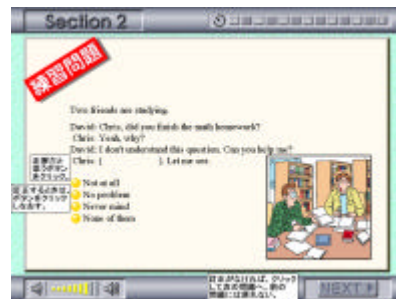
Section	出題内容	解答方式	制限時間
Section 1	語彙力	4肢選択	60秒/問
Section 2	表現力	4肢選択	90秒/問
Section 3	リスニング	4肢選択	60秒/問
Section 4	ディクテーション	書き取り	120秒/問

図1 CASEC 画面例

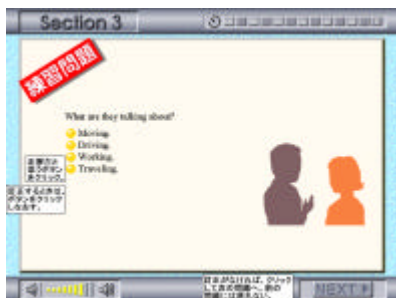
Section 1 語彙



Section 2 表現



Section 3 リスニング



Section 4 テイクション



の小ささが、適応型テストの一つの効果と言える。同じ測定精度を求める場合、適応型テストによる能力推定は、紙筆型テストの約 40%の項目数で推定できることが、今回の実験から認められた。なお適応型テストにおける項目提示の終了基準は、そこまでに実施された項目段階において得られた情報量による標準誤差と、その直前の項目段階において得られた情報量による標準誤差の差が、事前に設定された基準値以下に落ちた時点で、項目の提示を終了するルールによった⁴⁾。今回の実験版における項目提示終了基準は、標準誤差の値が SEM<0.5 かつ誤差変化量が 0.001 以下になった場合、または提示項目数が 30 項目になっても収束しなかった場合に終了させている。この終了基準は、期待する測定精度と、テスト実施時間を勘案して決定されたものである。

表 2 適応型テストと紙筆型テストのスコア間相関行列表

	CATS1	CATS2	CATS3	CATS4	CATT	P&P1	P&P2	P&P3	P&P4	P&PT	MEAN	SD
CATS1	1.000										100.2	18.2
CATS2	0.889	1.000									100.6	12.0
CATS3	0.831	0.807	1.000								107.7	12.2
CATS4	0.744	0.736	0.784	1.000							100.7	11.2
CATT	0.851	0.928	0.826	0.873	1.000						420.0	49.7
P&P1	0.888	0.818	0.804	0.773	0.899	1.000					103.5	10.7
P&P2	0.862	0.758	0.852	0.789	0.890	0.850	1.000				100.7	10.1
P&P3	0.743	0.746	0.823	0.656	0.804	0.706	0.756	1.000			104.5	10.0
P&P4	0.792	0.821	0.738	0.872	0.870	0.870	0.755	0.695	1.000		104.3	9.2
P&PT	0.899	0.865	0.895	0.869	0.956	0.921	0.824	0.869	0.892	1.000	412.5	36.5

5. 検証実験

適応型テストの実用性と有効性を検証するため、CASECを用いて、適応型テストと紙筆型テストの比較実験を実施した。これは、この実験のために 168 名のモニターを募集、データ収集を行い、その結果を分析した結果の報告である。前述の通りこの実験においては、同一のアイテムバンクから構築される、適応型テストと、各自同一形式の紙筆型テストの同一人についてのデータを収集した。またその中の 48 名の被験者に対しては、適応型テストを複数回受験してもらった。これらの収集データをもとに、適応型テストの有効性や信頼性に関する分析を行った。

5.1 適応型テストと紙筆型テストのスコア比較

項目応答モデルの特長の一つに、能力値の推定精度を、各被験者の能力水準毎に評価できることがある。ここでは今回の実験から得られた、同一人に対する適応型テストと紙筆型テストの両データから能力値を最尤推定し、両者の相関を見ることにした。結果は表 2 に示す通りで、両スコアには Section 間で 0.865~0.899 の、また Total 間で 0.958 という、高い相関が認められた。なお推定されたスコアの標準誤差 (SEM: Standard Error of Measurement) は、表 3 が示すように適応型テストは、出題数の固定された紙筆型テスト (各 Section とともに 30 問) よりも少ない問題数にもかかわらず小さな値となっている。この測定誤差

表 3 適応型テストと紙筆型テストの標準誤差の比較

出題内容	P&P SEM平均	P&P出題項目数	CAT SEM平均	CAT項目数平均
Section 1 語彙力	0.629	30	0.489	21.8
Section 2 表現力	0.636	30	0.469	20.2
Section 3 リスニング	0.630	30	0.455	20.0
Section 4 テイクション	0.449	30	0.362	12.9
TOTAL	2.344	120	1.775	74.8

5.2 推定値の収束状況例

前項において、適応型テストにおける項目提示終了基準に関する解説が行われたが、ここでは今回の実験データから、適応型テストにおいてどのように推定値と、その標準誤差が変化して行くかを、実際のデータからグラフ化したものを提示する。実際のデータにおいて、の推定値がどのように変化し、実施項目数に伴い、測定誤差がどのように変化して行くかの様子が、以下のグラフから捉えることができる。

図 2.1 は典型的なケースで、初期値 (1 項目に提示した項目困難度) と被験者の能力値が、近いところから推定が開始された場合で、23 項目で項目提示終了基準に到達している様子がわかる。図 2.2 は能力推定の収束が早かったケースで、識別力の高い項目から構成されるディクテーションのセクションに多く見られる傾向である。式 (5) の情報関数から明らかなように、項目応答モデルにおいては、出題される項目が、被験者の能力値と近いほど (b), また識別力 (a parameter) が高いほど、さらに偶然正答レベル (c parameter) が小さいほど、被験者の能力に開

する情報が多く得られる。ここから適応型テストシステムにおいては、上記のような特性の良問で実施すればするほど測定精度の向上が期待され、少数の項目でも十分精度の高い測定が可能になることを示唆している。図2.3は初期値が被験者の能力値と離れた所から能力推定を開始した場合である。図からも明らかなように、推定能力値近辺に到達するのに8問程度要し、その結果29項目まで提示して初めて終了基準値を満たし、一定値に収束するまでに時間のかかっている様子が分かる。これらのことから、適応型テストにおいては、初期値の取り方が重要であることがわかる。ただし、事前に被験者の能力値に関する情報が獲得されている場合(前回のスコアが既知である場合など)は、その近辺の特性値の項目を最初の項目として提示すれば良いことになる。

図2.1 典型的なケ-ス

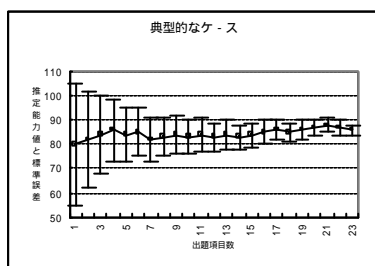


図2.2 収束が早かったケ-ス

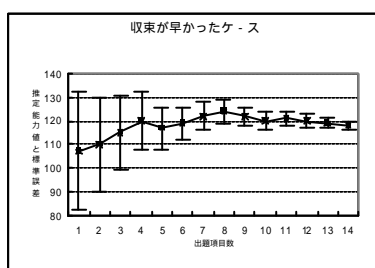
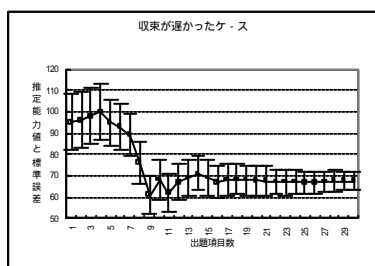


図2.3 収束が遅かったケ-ス



5.3 スコアの信頼性

テストにおいては、信頼性の定義は重要な概念である。テストの信頼性の問題は、同一人が同じテストを受験した

際のスコアの安定性にあるが、今回の実験において、同日に、同一人に複数回の適応型テストを課す実験を行った。今回の実験は、問題は異なるが困難度はほぼ同じテスト項目によるもので、適応型テストを利用した平行テスト法(parallel test method)による信頼性の検証になる⁵⁾。今回の実験においては、実施回毎に異なる項目でテストが実行でき、しかもこのテストが項目応答理論に基づいているため、各回の推定能力値は同一尺度上で比較できた。今回は48名という少人数のサンプルによる実験であったが、それぞれの被験者には、適応型テストをそれぞれ3回受験してもらった。各回毎に得られた推定能力値の平均・標準偏差・間相関係数一覧を表4に示した。このように、各実施回間の推定能力値には、従来型のテスト経験から比較して予想以上の高い一致度が認められた。なおここでの推定値は、一般使用に便利なように、Woodcock(1978)の基準により $9.1 * \text{スコア} + 100$ に概算したものをを用いている⁶⁾。

表4 複数回実施による推定能力値の一致状況

	1回目	2回目	3回目	平均	標準偏差
1回目	1.000			423.2	42.6
2回目	0.975	1.000		425.2	42.5
3回目	0.969	0.964	1.000	423.7	42.9

(N=48)

6. まとめ

CBTの応用事例として、当協会が開発したCATシステムを用いた検証実験の一部を報告した。今回の実験からコンピュータを利用した適応型テストは、測定精度や信頼性に関して、極めて有効な手法であることが確認された。テスト実施後に行った被験者に対するアンケート調査からも、おおむね好意的に受け止められるテスト方式であった。コンピュータテストの有効性は確認されたが、今後の実用化に向けては、現在設定している項目提示終了基準と実際の被験者の受験印象・所用時間の関係などを検討する必要もある。あるいは被験者の反応時間をどのように評価に反映させるかなど、残された検討・解決すべき課題は多いと思われるが、いずれも解決可能な事柄と認識しており、アイテムバンクの構築が可能な暁には有望な方式であると期待されている。

参考文献

- 1) Ronald K. Hambleton, Hariharan Swaminathan: ITEM RESPONSE THEORY Principles and Applications, 3.3.2 36/37, Kluwer・nijhoff Publishing(1985)
- 2) 芝祐順 編 項目反応理論 基礎と応用 A-1 87/89, 東京大学出版会(1991)
- 3) 池田央 著「行動計量学シリ-ズ7 現代テスト理論」, 35 61/63, 朝倉書店(1994)

- 4) Haward Wainer :Computerized Adaptive Testing :A Primer ,5 103/108 ,LEA(1990)
- 5) Robert L.Linn :Educational measurement(3rd ed.) , Macmillan(1989). (池田央, 藤田恵重, 柳井晴夫, 繁榊算男 (編約) “教育測定学 ” ,第3版, みくに出版 (1992)
- 6) Woodcock, R. W. :Development and Standardization of the Woodcock-Johnson Psycho-Educational Battery , Hingham ,MA :Teaching Resources Corporation(1978)