

語彙理解力測定における適応型テストの効果

東京大学教育学部教育情報科学研究室 芝 祐 順

東京大学教育学部教育情報科学研究室 柴 山 直

東京大学教育学部教育心理学研究室 鎌 原 雅 彦

東京学芸大学教育心理学研究室 野 口 裕 之

Effectiveness of an Adaptive Testing Procedure for Measuring Verbal Ability

Sukeyori SHIBA, Tadashi SHIBAYAMA, Masahiko KAMBARA and
Hiroyuki NOGUCHI

The purpose of this research is two-fold. Firstly, from a practical point of view, the development of a new computer program for an adaptive testing procedure was necessitated in order to measure the verbal ability more efficiently than the other existing procedures. Secondly, an experiment was designed in order to compare the efficiency of the adaptive testing procedure with that of the conventional testing procedure, analyze the effects of the feedback, investigate the relationship between the verbal ability and the response time, and evaluate the stability of the repeated measurements.

With regard to the first purpose, a satisfactory computer program was created. This program automatically controls the selection of test items during the process of administering the test so that the items administered to each subject should be appropriate in their difficulties. Furthermore, with regard to the second purpose, it was concluded that the results showed the superiority of the adaptive testing procedure over the conventional one in many respects.

I 語彙理解力の測定

I-1 語彙理解力尺度

言葉の意味理解力について測定する場合に語彙調査として個人が獲得している語彙数を推定するという方法が用いられることが多い(例えば, 森岡: 1951, 阪本: 1955)。しかしながら個人のある語に対する意味理解の状態は“理解している”と“理解していない”という2値的な状態にあるのではなく, 両者の中間的な理解の状態も含めた連続的な次元上の1点に位置づけられる。しかも, 個々の語の連続的な次元間で相関が高い場合には全ての語の次元と相関の高い連続的な次元を設定することができる。これは語の意味の一般的理解の水準を表わす次元と考えることができる。芝(1987)ではこのような

考え方に基づいて語の意味の一般的理解の水準を表わす語彙理解尺度を構成した。その際, 尺度を構成するための測定モデルに潜在特性モデルの中の2パラメータ・ロジスティックモデル(例えば Hamblton & Swaminathan 1985)を用いている。そして芝・野口(1982)では追跡的なデータを用いて尺度の等化を行ない, 語彙理解力尺度を改訂した。そして, 全315項目すべてが, 識別力及び困難度パラメータ値とともに公開されている。これらの項目には小学校1年生程度から成人向けのものまで広範囲な困難度のものが含まれている。野口・芝・丹(1983)ではこれらの項目の中から適当なものを選択して小学校1年生を測定対象としたB1版から大学生・成人を測定対象としたB11版まで全部11コの版を用意している。

I-2 適応型テストによる測定

B1版からB11版までの11コの各版では、測定対象とする全ての被験者に対して同じ項目数・項目内容で測定するわけであるが、この場合、全ての被験者に対して等しく精度の高い測定が実現できるとは限らない。例えば長期の海外生活から帰国直後の小学校3年生の児童に対して小学校3年生を測定対象としたB3版を実施したとしてもそれが適切な測定になるとは限らない。全ての被験者について十分な精度を持つ測定を実現しようとするならば、被験者毎に最適な困難度を持つ項目を選択してテストを構成し実施しなければならない。

各被験者の能力水準に応じて実施する項目を調整するテスト方式は適応型テスト, tailored test, adaptive test などと呼ばれるが、本研究では“適応型テスト”という名称を用いる。

適応型テストは、少数の項目を用いて、全ての被験者についてより精度の高い測定が実施できるように、基本的には「直前に呈示された項目に対する当該被験者の反応に基づいて、次に被験者が解答すべき項目を決定し呈示する」という手続を逐次的に繰り返す。この方式の利点には

- (i) 全ての被験者、特に平均値からの偏差が大きい被験者に対して精度の高い測定が可能であること
- (ii) 精度を落とすことなく被験者一人あたりに実施する項目数を減らすことができ、そのため一人あたりに要するテスト時間を短縮できること
- (iii) 難しい項目が続いて被験者にフラストレーションや不安を起こさせたり、易しすぎる項目が続いて飽きさせたりするようなことがない

などがある。

適応型テストの具体的な方式については、例えば、予備テスト部と本テスト部の2つの部分から構成され、被験者は先ず予備テストによっておおよその能力水準が推定され、その結果に基づいて本テスト部に含まれる難易度の異なる複数の版の中から最適な版を指定するという2段階テスト方式、困難度に基づいて分けられた複数の層状のネットワークを予め組んでおき、被験者は項目に対する反応に基づいていくつかの層の間を上下するという多層適応型テスト方式、予め用意された項目プール中からその時点で最大の項目情報量を示す項目を選択して実施する最大情報量方式などがある（詳しくは 野口：印刷中 を参照）。

語彙理解力の測定については、芝・野口・南風原(1978)で多層適応型テスト方式を、芝・野口・大浜(1980)及び芝・大浜・野口(1981)で多層適応型テスト方式を取

り入れている。いずれの場合も適応型テスト方式を用いることによってより精度の高い測定が実施されたことが示されている。

I-3 パーソナルコンピュータの利用

電子技術の進歩により、パーソナルコンピュータシステムが低価格化し、急速に普及しはじめたため、テスト過程の制御にこれを用いることが実用上にも可能となった。大型コンピュータではなく特にパーソナル・コンピュータを用いることの利点として

- ・被験者1名に対してコンピュータを1台用意できるため、被験者の予定に合わせていつでもテストを実施できる
 - ・大型の計算機システムではないので小規模の学校などでも導入することが比較的容易であり、取扱いも簡単である
 - ・フロッピー・ディスクなどを用いることによって、テストの輸送管理が容易である
 - ・テストが改訂された場合に容易に新しい版と交換できる
- などがある。

しかしながら、パーソナル・コンピュータの規模や能力によっては利用上の限界もある。第1には計算時間の問題である。被験者が当該項目に解答したならば、コンピュータはその反応を採点し、次に呈示すべき項目を選択するための基準を計算し、その結果決定された項目の内容を検索し、被験者に呈示するという一連の作業を行わなければならないが、ここで時間がかかってしまうこと、被験者がテストに解答するタイミングをはずしてしまうことになる。第2には記憶容量の問題である。被験者全てについて精度の高い測定を行なうためには広い範囲の困難度から項目を多数用意しておかなければならないが、そうすると全ての項目の内容を主記憶内に収めておくことは不可能になる。外部記憶に収めておいて、必要になる度に主記憶内に呼び出すというのでは時間がかかってしまう。この点について何らかの工夫が必要となる。また、被験者数が増えるに従って、その結果が累積されて来るが、これらのデータをどのような形で保存するかについても工夫が必要になる。

本研究ではパーソナル・コンピュータを利用した最大情報量方式による語彙理解力測定を行なう場合の測定効果について報告する。

II 最大情報量方式の適応型テストによる測定

II-1 テストの概要

最大情報量方式：潜在特性モデルでは各項目毎に項目情報量 $I_j(\theta)$ ($j=1\cdots, n$) が計算される。2パラメータロジスティックモデルの場合、項目 j の特性曲線を $P_j(\theta)$ ($j=1\cdots, n$) とすると、それは

$$I_j(\theta) = 1.7^2 a_j^2 P_j(\theta) \{1 - P_j(\theta)\} \quad (1)$$

で表わされる。ただし、

$$P_j(\theta) = \frac{1}{1 + \exp\{-1.7 a_j(\theta - b_j)\}} \quad (2)$$

であり、 a_j 、 b_j はそれぞれ項目の識別力及び困難度パラメータである。この項目情報量 $I_j(\theta)$ は尺度値 θ に対する項目 j の精度を表わしており、尺度値が θ_i である被験者 i に対して $I_j(\theta_i)$ の値が大きい項目を実施すれば精度の高い測定ができる。従って、常に最大の項目情報量を持つ項目を探して実施すればより精度の高い測定が可能になる。このことを具体化したのが最大情報量方式の適応型テストである。

実際のテスト場面では被験者 i の尺度値 θ_i は未知であるから、その代わりに既に m 項目実施していたならばその結果から推定尺度値 $\theta_i^{(m)}$ を求める。この θ の値に対して情報量が最大となる項目を、項目プールの中の未実施項目から選んで実施する。このような手続を適当な打ち切り基準を満たすまで繰り返すのである。

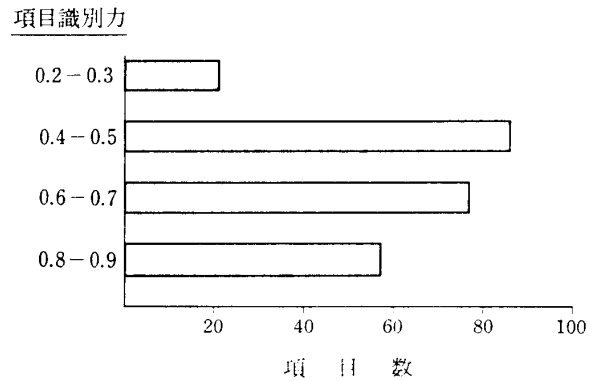
この適応型テスト方式では項目プールの記憶、項目情報量の計算、最適項目の検索などの作業が短時間のうちに行われる必要があり、コンピュータの補助が不可欠である。

項目プール：項目プールには芝・野口（1982）の315項目中から241項目を選んで用いた。項目困難度及び識別力の分布は図1に示した通りである。測定対象はおよそ小学校高学年から大学生までになる。

プログラムの作成：適応型テストを実施するためのプログラムは1985年7月から作成が開始された。言語はBASICによった。

予備調査：上記プログラム中実用上必要な被験者ごとの項目数の上限および下限、特性値を推定するための逐次計算の停止基準や被験者にとって見やすい画面等を調べるための予備調査を行った。その期間は1985年12月から1986年1月、調査対象は都内国立大学附属中学高等学校6学生の生徒約100名であった。

プログラムの流れ：予備調査の結果に基づき決定された



(注) 識別力については小数点第2位以下の端数は存在しない。

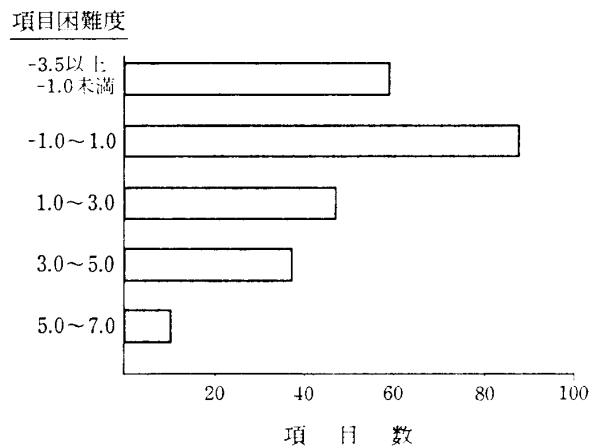


図1 項目特性の度数分布

プログラムの流れはおおよそ以下のようなになる（図2）。入力はすべて被験者がライトペンを使用して画面から行うようになっている。

- ① 被験者の学年を入力する。
- ② テストのやり方を練習するための項目を1問呈示する。
- ③ 被験者の学年に基づいて初期項目を選ぶ。
- ④ 最初の6問目までは、被験者の前問に対する反応の正誤情報のみを使って、正答ならば直前の項目より困難度が+1、誤答ならば-1の項目を選び呈示する。
- ⑤ 6項目分の反応パターンがえられた時点で理解力特性値 θ の推定を行う。
- ⑥ 7問目からは、直前に推定された θ に最も近い困難度を持つ項目の中で、最も高い識別力を持つ項目を選び画面に呈示する。
- ⑦ 呈示された項目に対する被験者の反応とこれまでの反応パターンを加えて新たに θ の推定を行う。
- ⑧ 適当な終了基準に達するまで⑥⑦を繰り返す。
- ⑨ 終了基準に達すればテストを終了し、最後にアン

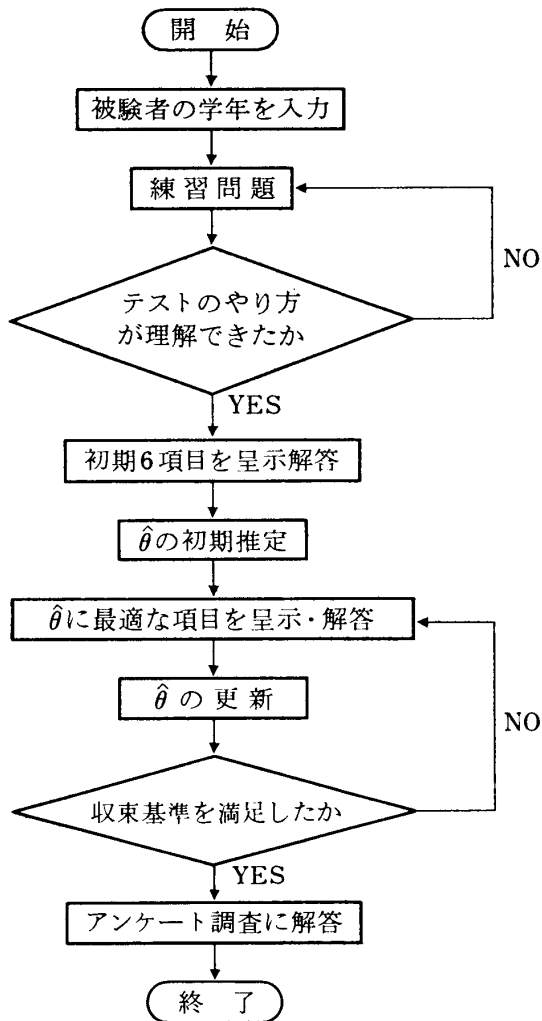


図2 プログラムの流れ

ケート項目を呈示し入力がすめば完了する。

なお、プログラムの構成、問題の画面への出力例、およびアンケート項目の内容についてはそれぞれ図3、4、5を参照のこと。

θ の計算方法：潜在特性理論において項目特性が既知の場合、尤度関数は潜在特性 θ に関する単一の関数で表現され、これを最大とする θ は次の方程式の解として求められる。

$$\sum_{j=1}^n a_j \{p_j(\theta) - \chi_j\} = 0 \quad (3)$$

ただし $\chi_j = \begin{cases} 1, & \text{項目 } j \text{ が正答の場合} \\ 0, & \text{" が誤答の場合} \end{cases}$

(3)式を解く計算アルゴリズムには確実に解の求まる2分法と終盤までの収束が速いセカント法を連結した方法 (Forsythe, G, etc., 1977) を採用した。なお θ の範囲は全問正答、又は全問誤答の場合の発散を防ぐため、 $-7 \leq \theta \leq 7$ とした。

終了基準：テストの終了基準を構成する条件は以下の5個である。

- A 反応項目数が20個未満でないこと
- B 反応項目数が60個以上であること
- C $I(\theta)^{-1/2} \leq 0.2$
- D 最新5項目に関し、変化の大きさを

$$d = \frac{1}{5} \sum_{j=1}^5 |\theta^{(k-j+1)} - \theta^{(k-j)}|$$

で定義し、これが $d \leq 0.1$ であること。ただし、 k は現在まで呈示した項目数、 $\theta^{(k)}$ は k 問目終了時の θ の推定値とする。

- E 変化の方向を

$$\text{sgn}[\theta^{(k)} - \theta^{(k-1)}]$$

で定義し、これが過去6項目までさかのぼって5回の変化の方向に正負がそれぞれ少なくとも、1回ある場合。ただし、sgn は [] 中の式の値の正負を与える操作子である。

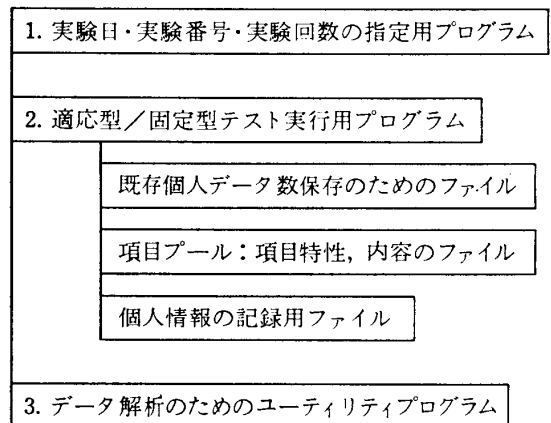


図3 プログラムの構成

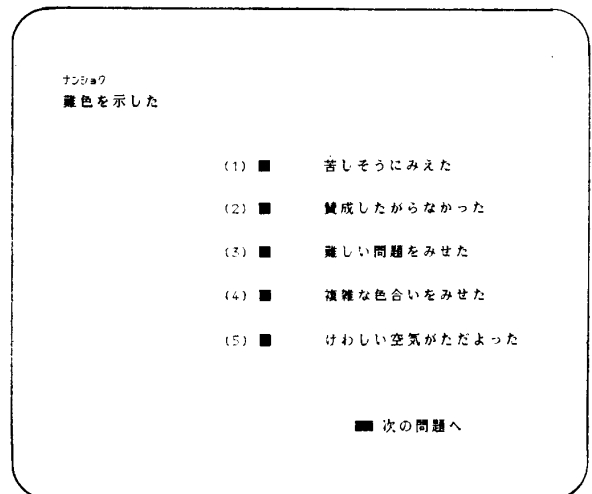


図4 問題項目の画面例

| | |
|----|---|
| 1. | このコンピュータを使ったテストは、やりやすかったですか？ 非常にやりやすかった どちらかというとやりやすかった 何ともいえない どちらかというとやりにくかった 非常にやりにくかった |
| 2. | 画面の文字は読みやすかったですか？ 非常に読みやすかった どちらかという読みやすかった 何ともいえない どちらかというと読みにくかった 非常に読みにくかった |
| 3. | 答える時に使ったペンは、使いやすかったですか？ 非常に使いやすかった どちらかという使いやすかった 何ともいえない どちらかというと使いにくかった 非常に使いにくかった |

図 5 アンケート項目

表 1 固定型テストの項目パラメタの平均と標準偏差

| | 項目識別力 | 項目困難度 |
|------|-------|-------|
| 平均 | 0.53 | 0.85 |
| 標準偏差 | 0.15 | 0.58 |

終了基準としてはこれらの条件を組み合わせて、

$$[A \wedge \{C \vee (D \wedge E)\}] \vee B$$

を採用し、これが真となったときテストを終了する。

固定型テスト：以上で説明した適応型テストとの比較検討のため、表1に示すような平均と標準偏差の項目特性値を持った30個の項目から構成された固定型テストも準備された。このテストは理解力特性値が1である付近で最大のテスト情報量を持ち、1を中心にして-1から3までの特性値を持つ被験者群を測定することを前提に作成された。

なお、適応型および固定型テストとも反応直後にその正誤情報がフィードバックされるものとされないものの2つの条件で実施された。

表 2 実験計画

| 実験の目的 | 実験番号 | 学 年 | 1 回 | 2 回 |
|----------------------------------|------|-----|------------------|---------|
| (1) 適応型と固定型の比較 | 1 | 中 3 | Tn : 54 | Fn : 54 |
| | 2 | 中 3 | Tf : — | Ff : — |
| | 3 | 高 1 | Tn : 34 | Fn : 33 |
| | 4 | 高 1 | Tf : 26 | Ff : 33 |
| (2) 正誤情報のフィードバックの効果 | 5 | 高 3 | Tf : 48 | Tn : — |
| | 6 | 高 3 | Tn : 40 | Tf : 38 |
| (3) 応答時間と語理解力の関係 | | | 1 回目の全データ | |
| (4) 繰り返し測定に対する理解力推定値の安定性 | 7 | 高 1 | Tf : 29 | Tf : — |
| | 8 | 高 2 | Tn : 49 | Tn : 40 |
| | 9 | 高 3 | Tf : 50 | Tf : — |
| (5) スタートレベルに対する頑健性 | 10 | 中 2 | Tn : 43 | Tn : — |
| | 11 | 高 3 | Tn : 36 | Tn : 23 |
| (6) その他のデータ (実験番号と学年が一致しなかった被験者) | | 中 2 | | Fn : 1 |
| | | 中 3 | Tn : 1 | |
| | | 高 1 | Tn : 1 | |
| | | 高 2 | Tf : 2 Tn : 1 | Fn : 1 |
| 合 計 | | | 414 | 223 |

(注) 表中1回、2回の欄で、T、Fはそれぞれ使われたテストが適応型、固定型テストであること、f、nはそれぞれ正誤情報のフィードバックの有無を示す。また数字は実際に集められた被験者の数である。また学年表示は1986年4月現在のもの、—は被験者のない実験。

II-2 実験計画

本研究で当初予定していた目的は、

- (1)適応型テストと固定型テストとの比較
- (2)正誤情報のフィードバックの効果
- (3)応答時間と語彙理解力との関係
- (4)繰り返し測定に対する理解力推定値の安定性
- (5)学年と対応しない困難度、すなわち被験者の学年から予想して難し(易し)すぎる困難度を持つ項目を初期項目にしたときの適応型テストの頑健性を調べることにあった。この目的に沿った上で、広範囲にわたる学年から被験者を集めることも考慮して表2に示したような実験計画を組んだ。ただし実際には行えなかった実験(たとえば実験2など)やプログラムミスのために成功しなかった実験11などがあるため、特に目的(5)に関しては本研究の分析から省くこととした。

II-3 実験方法

実験装置としてパーソナル・コンピュータ(PC-9801 VM 2)とライトペン、プリンターを組み合わせたもの2セットを用意し、被験者の都合の良い時間帯に適宜実施された。被験者は画面の指示にしたがってテストを行い、終了後はノートに被験者の氏名、学年、クラス、出席番号、性別、生年月日、およびコンピュータから示される被験者識別番号を記入した。

II-4 実験対象・期日・場所

実験対象となった被験者は秋田市内の私立女子中学校及び同高等学校の中学2年生から高校3年生までの計5学年、のべ637名である。学年ごとの人数の内分けは表3に示した通りである。また実験期日は1986年3月から同年7月の間である。実験場所は校舎内教育研修室の一角であった。

表3 学年別の被験者数

| 学 年 | 実験1回目 | 実験2回目 | 合 計 |
|-------|-------|-------|-----|
| 中学2年生 | 43 | 1 | 44 |
| 〃 3年生 | 55 | 54 | 109 |
| 高校1年生 | 90 | 66 | 156 |
| 〃 2年生 | 52 | 41 | 93 |
| 〃 3年生 | 174 | 61 | 235 |
| 合 計 | 414 | 223 | 637 |

III 結果と考察

III-1 語彙理解尺度値と反応時間の分布

本研究では語彙理解について広範囲から一様に被験者を得る必要があり、これを実現するために中学校2年生から高等学校3年生までの5学年を対象として実験を実施した(II 2参照)。これらの被験者の語彙理解力の分布は表4に示した通りである。ここで高校2年生の平均値が高校3年生の平均値を上まわっているが、これは本研究で用いる各学年の被験者が協力学校の当該学年のランダム・サンプルではないため、一部の学年について被験者の語彙理解力に偏りが出たためである。しかしながら、本研究の目的に照らすと広範囲に語彙理解力が分布する被験者集団が得られればよく、学年別に比較する必要はないのであるからこのことは特に問題にする必要はない。

各被験者毎に解答した反応時間の平均値を求め、これ

表4 語彙理解尺度推定値の学年別単純集計

| 学 年 | 人 数 | 平 均 | 標準偏差 |
|-------|-----|------|------|
| 中学2年生 | 43 | -.40 | 1.51 |
| 〃 3年生 | 55 | .38 | 1.68 |
| 高校1年生 | 90 | .81 | 1.54 |
| 〃 2年生 | 52 | 2.19 | 1.21 |
| 〃 3年生 | 174 | 2.09 | 1.40 |

表5 反応時間とフィードバックの有無

| | フィードバック無 | フィードバック有 | 全 体 |
|-------------|----------|----------|------|
| 人 数 (人) | 259 | 155 | 414 |
| 平 均 (秒) | 12.4 | 13.6 | 12.9 |
| 標 準 偏 差 (〃) | 4.5 | 5.7 | 5.0 |
| 0以上~3未満 (〃) | 1 | 0 | 1 |
| 3〃~6〃 (〃) | 11 | 1 | 12 |
| 6〃~9〃 (〃) | 44 | 29 | 73 |
| 9〃~12〃 (〃) | 71 | 41 | 112 |
| 12〃~15〃 (〃) | 68 | 34 | 102 |
| 15〃~18〃 (〃) | 33 | 24 | 57 |
| 18〃~21〃 (〃) | 20 | 12 | 32 |
| 21〃~24〃 (〃) | 7 | 5 | 12 |
| 24〃~27〃 (〃) | 2 | 6 | 8 |
| 27〃~30〃 (〃) | 2 | 0 | 2 |
| 30〃~33〃 (〃) | 0 | 0 | 0 |
| 33〃~36〃 (〃) | 0 | 1 | 1 |
| 36〃~39〃 (〃) | 0 | 2 | 2 |

と解答に対する正誤のフィードバックの有無との関係を示したのが表5である。フィードバックの無い場合には被験者259名の平均反応時間が12.4秒、標準偏差が4.5秒であるのに対して、フィードバックの有る場合には被験者155名の平均反応時間が13.6秒、標準偏差が5.7秒であった。フィードバックの有る場合の方が反応時間が若干長く、散らばりも若干大きい傾向が見られるがその違いは小さく、被験者が異なることも考えると特に問題にする必要はない。両方を合わせると、平均反応時間は12.9秒、標準偏差は5.7秒であった。

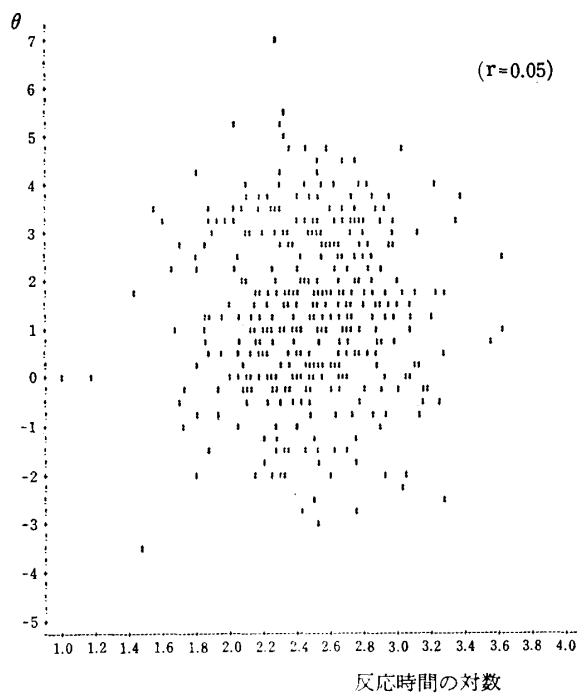
III-2 尺度値と反応時間の関係

理解力 θ の大きさ、あるいは θ と平均項目困難度 b の差の大きさと反応時間の長さとの相関関係を調べるため、反応時間を対数変換したものと、 θ あるいは $\theta-b$ との散布図をそれぞれ図6-1、6-2に示した。なおここで平均項目困難度 b とは、ある被験者が受けた全項目の困難度の平均値のことをいう。データは1回目の適応型テストを受けた全員の結果を使用した。この際、いわゆる部分集団（本データの場合は学年集団）の合併による相関関係の歪みはない。図6からも明らかのように、 θ および $\theta-b$ と反応時間との間には、曲線相関も含めていずれの相関関係も見出せない。したがって θ および $\theta-b$ と反応時間とは無相関であると結論できる。

III-3 適応型テストの機能

本研究で用いられた適応型テストが正しく機能していることを確認するために、以下の4点からの評価を行った。

a θ と項目困難度の対応：ここでは、与えられた項目に対する反応の正誤情報及び項目特性から各問終了ごとに計算される θ に最適な項目が選ばれていることを確かめる。そのために停止基準の各条件をすべて充たして終了した被験者のうち代表的な2名を選んで、 θ の推定値と呈示される項目の困難度の変化を描いたものを図7に示す。図7-1の被験者は全部で20個の項目に反応し、 θ は-1.39、同様に図7-2の被験者は26項目で0.91の特性値をもつ被験者のものである。図7には呈示項目の困難度を破線で、その都度推定された θ を実線で表示してある。図7より、推定された θ に応じて、その値に近接した困難度を持つ項目が呈示されてゆき、テストの終盤では θ の最終推定値の辺りでよりきめ細かい項目の呈示がなされていることがわかる。ここには載せなかったが適応型テストを受けた他の被験者についても同様の結果が得られている。このことより、適応型テストの所期



(注) カッコ内は相関係数(以下同じ)

図 6-1 θ と反応時間の散布図

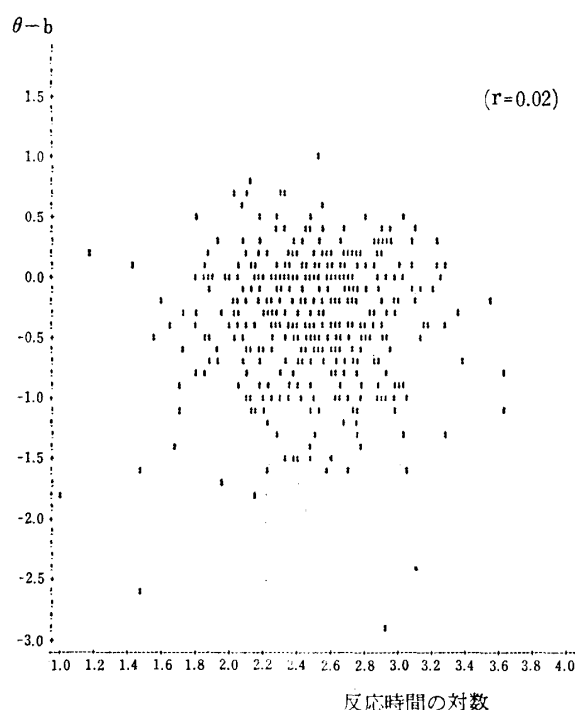


図 6-2 $\theta-b$ と反応時間の散布図

の機能のうち、最適な項目を与えるアルゴリズムは正しく働いているといえる。

b θ の位置とテスト情報量の大きさ：前述したようにこの適応型テストでは最大情報量方式を採用している。

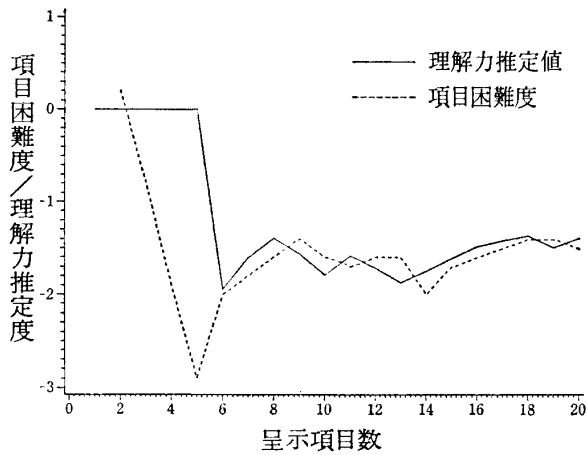


図 7-1 θ と項目困難度の対応 (1)

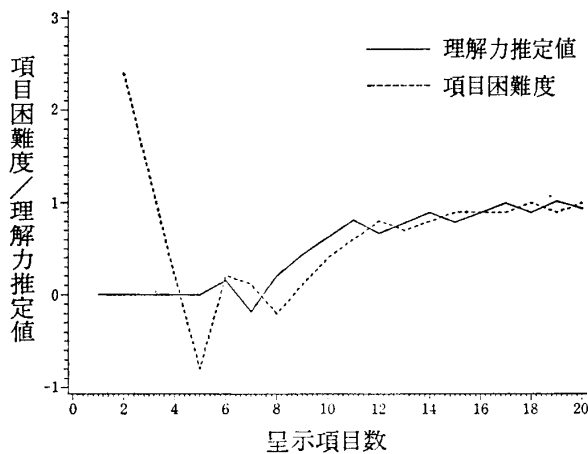


図 7-2 θ と項目困難度の対応 (2)

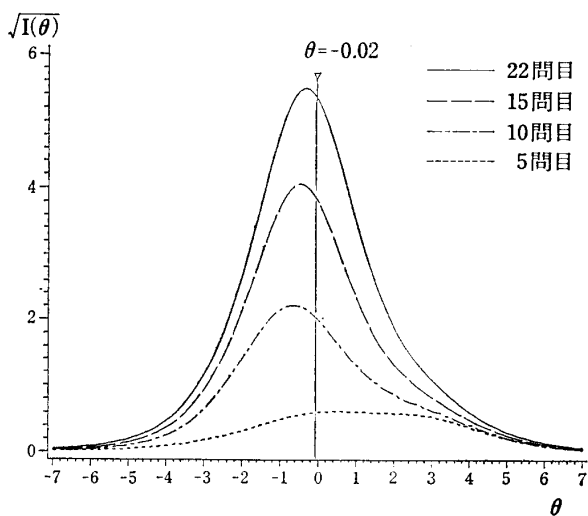


図 8 θ の位置とテスト情報量の大きさ

したがって最終的に推定された理解力 θ は、その被験者が反応したすべての項目を使って求められるテスト情報量が最大となる付近に位置しているはずである。このことを確認するため、適応型テストを受けた被験者の中から任意に選んだ被験者のテスト情報曲線を図8に示した。この被験者は計22項目について反応し、理解力の最終推定値は -0.02 であった。また図8の4本の曲線は下から順にそれぞれ6, 10, 15, 22問目までの項目を使って描かれたテスト情報曲線である。ただしテスト情報量の平方根をとって描いてある。項目数が増加するにつれて、上に凸の曲線が切り立っていくことが図からわかる。そして最終推定値は22問すべてを使って求めたテスト情報曲線がピークを示す付近に存在している。以上のことから、本テストに採用された最大情報量方式は正しく作動していると判断してよい。

c 途中経過の概要：次に理解力 θ の推定過程で次々と与えられていく項目の困難度が θ の推定に応じて適切に分離していくかどうかを、適応型テストを受けた1回目の被験者計414名のデータに基づいて調べる。そのためにも、 θ の最終推定値 θ の大小によって4群に分類し、各群に全被験者の約25%がそれぞれ所属するよう範囲を区切る。具体的には、第1群には特性値が0.12未満の、第2群には0.12以上1.30未満、第3群には1.30以上2.68未満、第4群には2.68以上の被験者が入っている。これら4群が5, 10, 15, 20問目の項目困難度及び θ の分布状況をまとめたものが表6である。どの群でも項目数が増加するごとに、困難度の平均は θ の平均に近づき、標準偏差の値は小さくなっている。これは、各群の θ の分布に対応して、与えられる項目の困難度の大きさが分離していることを意味する。すなわち、項目数が増加するに従って理解力に応じた困難度を持つ適切な項目が与えられていると考えてよい。

表 6 項目困難度の途中経過の平均と標準偏差

| | 第1群 | 第2群 | 第3群 | 第4群 |
|----------|-----------------|----------------|----------------|----------------|
| 5 問 目 | 0.02 (1.33) | 1.29 (1.17) | 2.13 (1.28) | 3.81 (3.4) |
| 10 問 目 | -0.74 (1.19) | 1.02 (0.91) | 2.11 (0.92) | 3.78 (1.06) |
| 15 問 目 | -0.75 (1.01) | 0.99 (0.68) | 2.04 (0.78) | 3.68 (0.76) |
| 20 問 目 | -0.71 (0.87) | 0.94 (0.57) | 2.02 (0.67) | 3.69 (0.73) |
| θ | -0.83 (0.89) | 0.74 (0.38) | 1.85 (0.39) | 3.55 (0.70) |

(注) カッコ内の数値は標準偏差

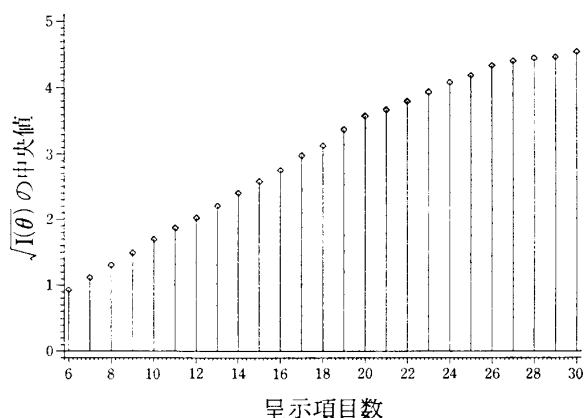


図9 項目数増加と $\sqrt{I(\theta)}$ の中央値の変化

d 項目数増加と $\sqrt{I(\theta)}$ の変化：図9は第1回目の適応型テストを受けた全被験者のテスト情報量の平方根の中央値を6問目から30問目までの各問ごとにプロットしたものである。最高で40問目、すなわち40項目について反応した被験者も2名いたが、31問目以降は被験者数が39名以下に減少したので省略した。図9より20問目まではテスト情報量が直線的に増加し、20問目から26問目まではややその増加率が減少し、26問目以降は頭打ちの状態となっていることがわかる。これは本研究で利用した語彙理解力テストを構成する項目群においては、最大情報量方式によって最適な項目を被験者に与えていけば、せいぜい26問程度で十分に高い精度の推定値が得られることを示している。

III-4 適応型テストと固定型テストの比較

適応型テストが固定型テストよりも効率（同じ精度の推定値をえるために必要な項目数）の点で優れている例として実験1と3の結果を取り上げる。実験1,3はそれぞれ中学3年生と高校1年生を対象に正誤情報を被験者に与えることなく、1回目に適応型テスト、2回目に固定型テストを実施したものである。二つのテストを両方とも受けた被験者は実験1が50名、実験3が28名であった。2つの実験について、適応型、固定型テストによ

表7 適応型テストと固定型テストの比較

| | 実験1 (中学3年生50名) | | 実験3 (高校1年生28名) | |
|-----|----------------|----------------|----------------|----------------|
| | 適応型 | 固定型 | 適応型 | 固定型 |
| 情報量 | 4.91 (0.88) | 4.36 (1.40) | 4.75 (0.60) | 4.04 (1.07) |
| 問題数 | 24.9 (3.8) | 30 (—) | 25.9 (4.6) | 30 (—) |

る θ の推定値の情報量と使用した項目数の平均と標準偏差は表7に示す通りである。両実験とも、適応型テストの方が、問題数が少ないにもかかわらず、 $\sqrt{I(\theta)}$ の平均値が高く、標準偏差が小さくなっている。すなわち、この2つの実験においては、固定型テストと比較して適応型テストの方がより高精度の推定を、より少ない問題数で行えるという意味で効率の良いテストとなっていると言えよう。

なお実験4については1回目、正誤情報のフィードバックを行ったため学習効果が存在し、2回目で実施された固定型テストの結果を歪めていると考えられるので分析対象とはしなかった。

III-5 フィードバックの影響

予備調査段階において、各項目への反応直後にその反応の正誤情報をフィードバックすることによって、被験者のテストへの動機づけが著しく向上することが観察された。そこで本実験では反応の正誤をフィードバックする適応型テストとしない適応型テストを実施し、フィードバックの有無によって反応時間やテストのやりやすさのアンケート項目の結果に差が出るかどうかを調べた。対象となるデータは実験5-1と6-1の高校3年生それぞれ48名と40名を利用した。

反応時間はその分布が正規分布に近づくように対数変換した後、フィードバックの有無で平均と標準偏差に求めた(表8)。両者の平均差の大きさが、それぞれの標準偏差よりもかなり小さな値であることから、フィードバックの有無が反応時間に影響しているとはいえない。

次に正誤情報をフィードバックされたグループとされなかったグループの「このコンピュータを使ったテスト

表8 正誤情報と反応時間の対数の平均と標準偏差

| | 人数 | 平均 | 標準偏差 |
|-----------|----|------|------|
| フィードバック有り | 48 | 2.44 | 0.29 |
| 無し | 40 | 2.50 | 0.37 |

表9 正誤情報とテストのやりやすさ

| | 非常にやりやすい | やややりやすい | 何ともいえない | なりにくい | 非常になりにくい | 計 |
|-----------|--------------|--------------|--------------|-------------|------------|----|
| フィードバック有り | 14 (29.2) | 24 (50.0) | 9 (18.8) | 0 (0.0) | 1 (2.1) | 48 |
| フィードバック無し | 12 (30.0) | 16 (30.0) | 12 (40.0) | 0 (30.0) | 0 (0.0) | 40 |

(注) カッコ内はパーセント

はやりやすかったですか」とのアンケート項目に対する結果は表9のようになった。「非常にやりやすかった」, 「どちらかというやりやすかった」を合わせると、フィードバック有りのグループは79.2%, なしのグループは70%で、フィードバック有りのグループの方が「やりやすい」と答えている人がやや多い。ただし、この質問内容がテストへの動機付けを直接問うているとは言い難いため、上記の結果からフィードバックの存在がテストへの動機付けを促進したと断定することはできない。

III-6 くり返し測定の安定性

同一被験者に対してくり返し測定を行った場合の理解力推定における適応型テストの頑健性について調べるため実験8の結果を分析した。実験8は同一の被験者群に対して、前後2回のテストが約1カ月の期間をおいて実施された。1回目、2回目ともに参加した被験者は31名であった。各回における理解力推定値の平均と標準偏差は表10、1回目と2回目の理解力推定値の散布図及び相関係数の値は図10に示す通りである。各回の平均の差の大きさと、各回の標準偏差の大きさ、さらに散布図の形状から判断して、適応型テストによるくり返し測定の結果は非常に安定しているとは言い難い。実際個人ごとに分析していくと1回目で理解力推定値が1.91であったにもかかわらず2回目では-0.85となったケースも存在する。このようなケースが生じた原因をはじめとして、

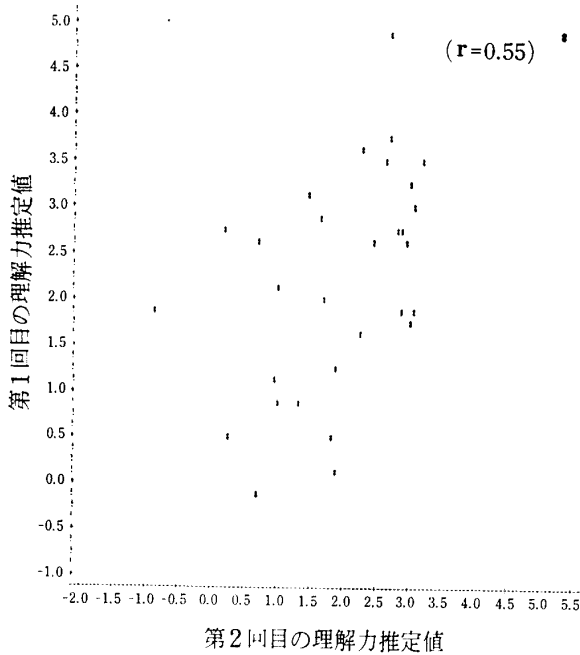


図10 くり返し測定における1回目と2回目の理解力推定値の散布図

表10 くり返し測定の安定性

| | 実験8の1回目 | 実験8の2回目 |
|------|---------|---------|
| 平均 | 2.89 | 2.02 |
| 標準偏差 | 1.27 | 1.20 |

適応型テストによるくり返し測定の安定性に関してはより詳しい議論が必要であろう。

III-7 アンケートの結果

本研究ではテスト実施直後に図5に示すような質問で調査を実施した。調査はテストと同じくパーソナルコンピュータのディスプレイ上に提示し、被験者にはライトペンを用いて解答させた。

まず、「このコンピュータを使ったテストはやりやすかったですか?」に対しては表11に示す結果が得られた。高校3年生で他の学年に比べてやり易いとした者の数がやや多く、平均値が1.98とやや小さくなっているが、特に問題にする程の差ではない。全体としては「どちらかと言えば易しかった」「非常に易しかった」という反応が合わせて60%を占め、「どちらかと言えばやりにくかった」「非常にやりにくかった」という反応は合わせて14%と少ない。従って、本研究のように新しい装置を利用したテストに対して被験者が特に抵抗感を持つようなことはないと言える。

表11 コンピュータを用いたテストのやり易さについて

| | 人数 | 1 | 2 | 3 | 4 | 5 | 平均 | 標準偏差 |
|---|-----|-----|-----|----|----|----|------|------|
| 中 | 2 | 43 | 5 | 10 | 15 | 7 | 2.98 | 1.20 |
| | 3 | 55 | 14 | 15 | 9 | 9 | 2.67 | 1.40 |
| 高 | 1 | 90 | 23 | 32 | 19 | 13 | 2.34 | 1.11 |
| | 2 | 52 | 13 | 12 | 14 | 9 | 2.60 | 1.26 |
| | 3 | 174 | 58 | 70 | 39 | 5 | 1.98 | 0.88 |
| 計 | 414 | 113 | 139 | 96 | 43 | 23 | 2.33 | 1.15 |

表12 画面の文字の読み易さについて

| 人数 | 1 | 2 | 3 | 4 | 5 | 無答 | 平均 | 標準偏差 |
|-----|-----|-----|----|----|---|----|------|------|
| 414 | 225 | 129 | 37 | 14 | 6 | 3 | 1.65 | 0.89 |

表13 ライト・ペンの使い易さについて

| 人数 | 1 | 2 | 3 | 4 | 5 | 無答 | 平均 | 標準偏差 |
|-----|-----|-----|----|----|----|----|------|------|
| 414 | 210 | 134 | 38 | 21 | 10 | 1 | 1.76 | 0.98 |

次に「画面の文字は読みやすかったですか」に対しては表12に示した結果が得られた。質問1で学年間に差が見られなかったため、ここでは全学年をまとめて示した。その結果，“非常に読みやすかった”及び“どちらかというと読みやすかった”とする者が全体の過半数を占め，“どちらかという読みにくかった”及び“非常に読みにくかった”とする者は全体の5%未満であった。

さらに「答える時に使ったペンは使いやすかったですか？」に対しては表13に示した結果が得られた。これについても全学年をまとめて示した。その結果“非常に使いやすかった”及び“どちらかという使いやすかった”とする者が80%を超え，“どちらかという使いにくかった”及び“非常に使いにくかった”とする者は約7%と少数であった。従って、大部分の被験者は画面の文字に対しても、ライトペンに対しても、全く抵抗を示していないと考えられる。

以上3つの結果を合わせると、本研究で用いた装置によるテストは今後実用化の可能性が十分にあると言える。

謝辞：本研究をすすめるにあたって、聖霊女子短期大学付属中学校・高等学校の先生方、生徒の皆さんの御協力を得たことに心から感謝いたします。

プログラム作成に際しては、教育情報科学研究室の藤田倫子さんに御協力いただいたことを感謝します。

本研究には文部省科学研究費（昭和60. 61年度試験研究2）の補助を受けました。本研究の結果の分析等には東京大学大型計算機センターを利用しました。

注

Forsythe, G., Malcolm, M. A. and Moler, C. B. Computer methods for mathematical computations, Prentice-Hall. (森 正武訳 計算機のための数値計算法, 日本コンピュータ協会, 1978)

Hambleton, R.K. & Swaminathan, H. 1985 Item Response Theory—Principles and Applications—, Kluwar Nijhoff Publishing.

森岡健二 1951 義務教育終了者に対する語彙調査の試み., 国研年報2, 秀英出版。

野口裕之・芝 祐順・丹 直利 1983 語彙理解力尺度の研究Ⅱ——項目固定版と適応形テストによる測定——, 東京学芸大学紀要第1部門教育科学, 34, 101-114。

野口裕之（印刷中）心理・教育テストとパーソナルコンピュータ, 心理学評論。

阪本一郎 1955 読みと作文の心理, 牧書店。

芝 祐順 1978 語彙理解尺度作成の試み, 東京大学教育学部紀要, 17, 47-58。

芝 祐順・野口裕之・南岡原朝和 1978 語彙理解力のための多層適応形テスト, 教育心理学研究, 26, 229-238。

芝 祐順・野口裕之・大浜幾久子 1980 多層適応形テストによる語彙理解力予備測定の効果, 東京大学教育学部紀要, 19, 27-34。

芝 祐順・大浜幾久子・野口裕之 1981 在外日本人児童の日本語彙理解力に関する調査, 東京大学教育学部紀要, 20, 111-128。

芝 祐順・野口裕之 1982 語彙理解力尺度の研究Ⅰ——追跡データによる等化——, 東京大学教育学部紀要, 22, 31-42。