多層適応形テストによる語彙理解力予備測定の効果

東京大学教育学部教育心理学研究室　芝　祐順
東京学芸大学教育学部教育心理学教室　野口裕之
東京大学教育学部教育心理学研究室　大浜幾久子

The Efficiency of Preliminary Measurement by Means of a Short Form of the Stratified Adaptive Test in Measuring Verbal Ability

Sukeyori SHIBA, Hiroyuki NOGUCHI and Kikuko OHAMA

The semi-adaptive testing procedure, which consists of a program test for preliminary measurement and a conventional group test for accurate measurement, is proposed. The program test is a short form of the stratified adaptive test with eight pages, each of which contains seven items of diverse difficulty. Each subject is to be tested individually, at first, by the program test and then, to be given an edition of conventional test chosen among nine editions from the easier to the more difficult one. The obtained scores of the program test are used in deciding the appropriate level of difficulty for each edition.

Sixty pupils from the first graders to the sixth graders are tested by this semi-adaptive testing procedure, and the efficiency of the procedure is analyzed in terms of various aspects. It is demonstrated that this four or five minutes program test works very well as a prelimiary measurement. This testing procedure is found to be fairly efficient to measure the abilities of subjects which distribute in a broad range.

I　問　題

われわれは語彙理解力を測定するための尺度を構成し最も効果的な測定方式について検討してきた（芝，1978；芝・野口・南風原，1978）。一般に30分ないし1時間程度の時間で全問に解答させる方式の語彙検査では，個々の版の測定可能範囲をそれ程広くすることはできない。これまでに作成された11の版についてその情報関数によって各版の可能な測定範囲の吟味をしているが，それによると各版とも十分な精度をもって測定できるのは当該学年の理解力分布のせいぜい2標準偏差程度の範囲にすぎず，その範囲外の理解力を持つ被験者に対しては別の版のテストを実施した方が精度の高い語彙理解力推定値が得られる。そのため一般にテストを実施する際には年齢や学年その他の事前情報によって被験者に適した測定可能範囲をもつテストを選択している。つまり事前情報を利用することによってそれだけ測定の効率を上げている。しかし，事前情報があまり有効に機能しない事態では被験者の理解力と実施したテストの測定可能範囲とがかならずしもうまく一致するとは限らない。このような場合に全ての被験者を測定可能範囲に含むテストを構成するためには一つのテスト中に膨大な数の項目が必要となる。これは，被験者に心理的及び時間的な負担を強いることでもあり，測定の効率という観点からも望ましいことではない。そこで事前情報が乏しい場合でも広範囲にわたって精度の高い測定値が得られる適応形テスト方式の開発が進められてきた。これはプログラムテストとか逐次テストとか Tailored Test などと呼ばれ，手続の原理はテスト内の項目に対する当該被験者の反応に基づいて逐次に実施する項目を決定するというものである（芝・野口・南風原，1978；野口，1979参照）。この方式の利点としては，

全ての被験者に対して精度の高い測定が可能であること，

精度を落すことなく被験者に実施する項目数を減ら

本研究は昭和54年度科学研究費補助金を受けた。
また，計算は全て東京大学大型計算機センターの HITAC 8800/8700 システムを使用した。

すことができ時間も短縮できること，

項目が難しすぎたり易しすぎたりして被験者のモチベーションを低めるようなことがないこと，

などがある。

具体的な適応形テストの構成法を大別すると(1)2段階テスト(2)項目固定型多段階テスト(3)項目可変型多段階テストがある。芝・野口・南風原(1978)では語彙理解力を測定するために多層適応形テストを用いているが，そこでは逐次の最適項目の選択にコンピュータを用いることなく適応形テストの利点を生かす工夫がされている。すなわち被験者は渡されたテスト冊子の各頁に印刷された8項目中の1項目に解答する。そして各項目の選択枝の末尾には次に解答すべき項目の番号が印刷されていて被験者は自分の選択した選択枝によって次の頁で解答すべき項目が指定される。被験者は最終的には全部で35項目に解答するが，その項目はかならずしも他の被験者と同じであるとは限らない。そして実際に実施した結果，各被験者の理解力に応じた項目を実行し精度の高い測定値が得られている。

ところで本研究では対象を幼児・児童とするために，順次に解答すべき項目への移行を被験者自身に任せてしまう前述の方式を採用することには困難がある。また，個人テストとして実施し，検査者が被験者に次に解答すべき項目を呈示してやることによってこの困難は除去できるが，比較的長時間かかる個人テストを多人数に対して実施しようとすることは実際上困難である。この難点を解決するために本研究ではプログラムテストを用いた予備テストと精密測定のための本テストとを組合せて実施する測定方式を試みた。テストの実施方法としては，まず予備テストとして簡単なプログラムテストを個人的に実施し，その結果にもとづいて，本テストで使用すべき標準版として困難度を異にする9版の中から最適の版を選ぶ。つづいて，その標準版を実施し，語彙理解力の推定値を算出する。

本研究ではこのような方式にもとづく多層適応形テスト（L版）について測定上の効果を検討する。

Ⅱ　予備テスト用のプログラムテストの作成

予備テストの冊子は8ページからなる。各ページには困難度水準の低い項目から高い項目まで，あわせて7項目ずつが印刷されている(図1参照)。検査者はその被験者の語彙理解力に適すると思われる項目を第1ページの7項目の中から一つ選んで，被験者に解かせる。被験者が解答として選んだ選択枝に対応して，次のページで被験者に与えるべき項目が決定される。原理的には，その被験者の選択が正答で，彼の理解力の高いことを示唆するときには，次のページにおいては困難度のより高い項目を与え，そうでないときには，困難度の同じ，あるいはより低い項目を与えることにする。このような指定はすべて解答記録用紙(図2参照)に示されているので，検査者はこれに従って各ページで与えるべき項目を指示していけばよい。この方式はすでに発表された多層適応形テストと同じものであるので参考までに両者の比較を表1に示す。

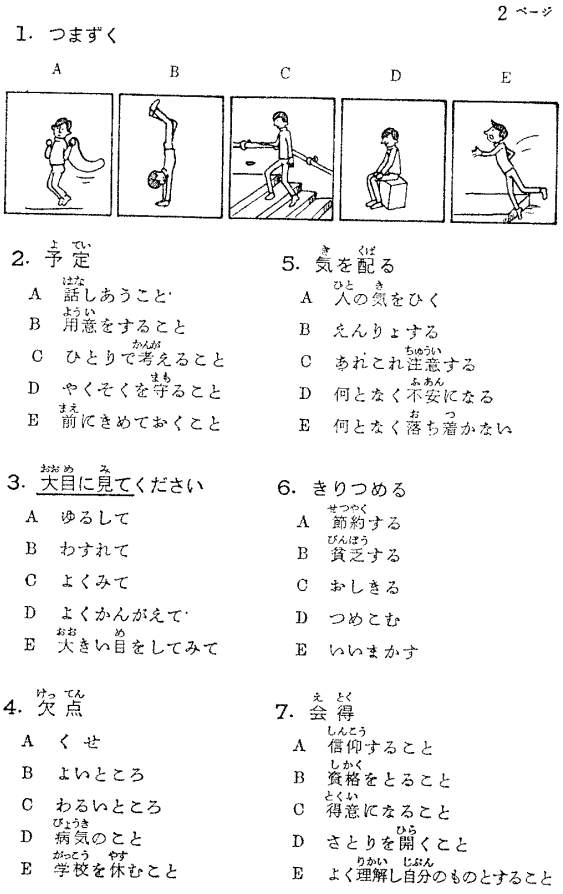

図1　予備テストのページ例

1.　使用した項目の特性

項目はすべてこれまでに項目分析をおこないその特性のわかっているものから選ばれた。その困難度は表2に示す通りである。このテストの測定範囲として，6歳児より中学1年生程度を予定しているので，この範囲の語彙理解力の分布を参考にして，項目困難度の層別が決められた。

2.　進行の指定

各被験者に対して，その理解力に適した困難度の項目をわりあてることによって，測定の効果を高めようとするのが適応形テストの狙いである。このため各項目の選

1 ページ	2 ページ	3 ページ	4 ページ
1 A－1 B－1 C－1 D－2 E－1	1 A－1 B－1 C－1 D－1 E－2	1 A－1 B－1 C－1 D－1 E－2	1 A－1 B－1 C－1 D－2 E－1
2 A－2 B－1 C－1 D－3 E－1	2 A－1 B－2 C－2 D－1 E－3	2 A－2 B－1 C－1 D－1 E－3	2 A－1 B－3 C－1 D－1 E－2
3 A－3 B－2 C－2 (D)－4 E－3	3 A－4 B－2 C－2 D－3 E－2	3 A－2 B－3 C－3 D－4 E－3	3 A－3 B－2 C－2 D－4 E－3
4 A－4 B－3 C－3 D－3 E－5	4 (A)－4 B－3 C－5 D－3 E－3	4 A－4 (B)－3 C－3 D－5 E－4	4 A－3 B－5 C－3 D－3 E－4

矢印は本文中の例の進行を示す。

図2　解答記録用紙（一部分）

表1　さきに発表された多層適応形テスト（芝・野口・南風原，1978）と本研究の予備テストとの比較

	多層適応形テスト	本研究の予備テスト
使用項目数	280	56
層別数	8	7
ページ数	35	8
実施の形態	集団検査	個人検査
進行の指定	テスト冊子の各選択枝に付記して被験者に指示	解答記録用紙に付記し検査者に指示
検査時間	30分程度	5分程度
対象	中学生以上	小学生以上

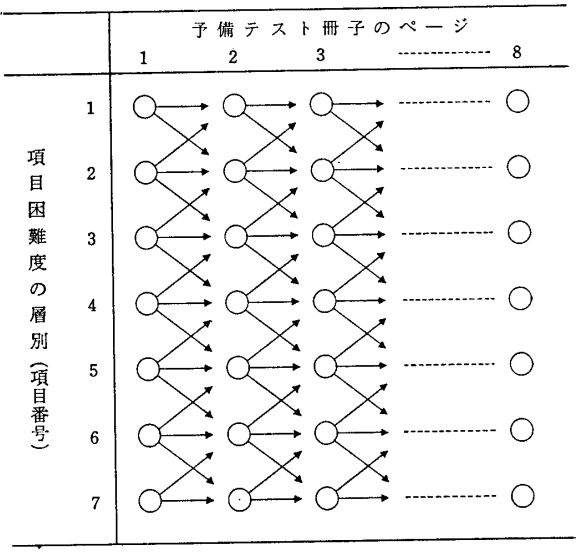

図3　予備テストの進行

表2　予備テストにもちいられた項目の困難度*

	予備テスト冊子のページ							
	1	2	3	4	5	6	7	8
項目困難度の層別（項目番号）1	−5.0	−5.0	−5.0	−5.0	−5.4	−5.3	−4.4	−4.3
2	−3.9	−3.6	−4.1	−4.0	−4.2	−4.0	−4.1	−4.1
3	−3.3	−3.1	−3.0	−2.8	−2.8	−3.3	−3.3	−3.4
4	−2.6	−2.6	−2.4	−2.3	−2.2	−2.2	−2.2	−2.2
5	−1.1	−1.4	−1.1	−1.1	−1.1	−1.5	−1.2	−2.1
6	−0.8	−0.7	−0.5	−0.5	−0.4	−0.4	−0.2	−0.5
7	0.3	0.4	0.3	0.2	0.5	0.3	0.6	0.6

項目識別力は平均 .67，標準偏差 .19

＊困難度の単位については芝（1978）参照

表3　予備テストの結果による標準版のわりあて

簡易スコア	9-13	14-17	18-22	23-28	29-34	35-41	42-48	49-55	56-63
標準版	AP1	AP2	B1	B2	B3	B4	B5	A6	J1

択枝ごとの選択者の理解力分布が利用された。すなわち，これまでの研究（芝，1978）でえられている項目分析の資料によって，あらかじめ各項目について選択枝ごとの理解力分布をもとめておく。その分布の位置の高い選択枝を，この予備テストで選んだ被験者には，困難度のより高い項目を与えるように，逆に分布の位置の低い選択枝を選んだ被験者にはより易しい項目を与えるように，プログラムテストの進行を指定する。この原理によると，正答の選択枝を選んだものは，そのつぎに一層だけ困難度の高い項目に進むよう指示される。残りの4枝については，それぞれの選択者のデータにより，理解力分布の位置の高い選択枝からは同じ困難度の層の項目へ，位置の低い選択枝からは困難度の低い層の項目へと進行させる。4つの誤答の選択枝の間で，理解力分布の位置にあまり差のないときには，4つの選択枝を選んだものをすべて同一の項目へ進行させることもある。また，当然のことであるが図3にみられるとおり，第1層と第7層の項目からの進行は変則的になっている。

図2の解答記録用紙に示された被験者の例では，まず，はじめに第1ページの問3が指定され，被験者がこれに対して選択枝Dを選んだ。解答記録紙の該当するDの右側に4と指定されているので，第2ページでは問4へ進む(図2の矢印参照)。そして，この問では選択枝Aを選んだため，解答記録用紙の指定に従って第3ページでは問4を与えることになる。以下同様の手続きによって第8ページの問まで行い，予備テストを終了する。

3.　採点法

予備テストの結果によって，本テストで使用する標準版の困難度が決められる。そのために，本テストをはじめるまえに予備テストの採点をしなければならない。採点の方法は2通りある。その一つは被験者に指定された項目に対する応答パタン（正答と誤答によるパタン）をもちいて，その被験者の理解力の最尤推定値（$\hat{\theta}_p$）をもとめるものである[1]。この方法では，コンピュータを必要とするが，第2の方法はコンピューターを必要としない簡易採点法である。簡易採点によるスコア（xであらわす）は，予備テストでその被験者が解くように指示された項目の番号をすべて加えあわせることによってえられる。ただし，第8ページの項目に対する応答によって，実際には存在しない第9ページの項目の番号が指定される。これは第8ページの項目への応答からえられる情報を利用するための工夫で，簡易スコアをもとめるときには仮空の第9ページの指定された項目番号も加算することとする。こうして，各被験者は最低9と最高63の間の簡易スコアを得ることになる。

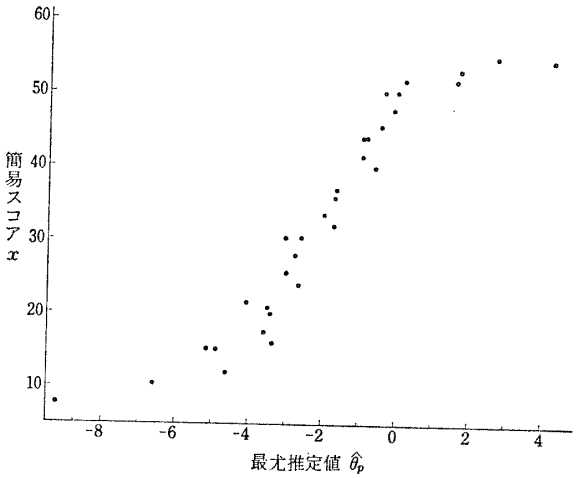

図4　簡易スコアと最尤推定値

簡易スコアxと最尤推定値$\hat{\theta}_p$との関係は，条件つき確率$f(\hat{\theta}_p|x)$によって理論的にもとめることもできるが，ここでは，簡単なシミュレーションによって，いくつかの典型的な進行パターンに対する簡易スコアと推定値$\hat{\theta}_p$との相関関係をもとめた。その結果は図4に示した通りである。予備テストの結果による標準版の対応は，この図と予備実験[2]の結果とを参考にして，表3のように決定された。

なお，予備テストと組み合わせて使用される標準版9版の項目特性は表4に示す通りである。

1) 理解力の最尤推定値のもとめ方については芝(1978)参照。実際には計算上の必要から若干の修正が施されている。

2) 本研究の予備テスト試作版は，1979年3月にスイス・ジュネーブ日本語（補習）学校において，5歳児から高校生まで計94名を対象に実施された。

Ⅲ　テストの実施

被験者は小学校1年生から6年生まで各学年男女5名ずつ，計60名である。これら被験者は各学年毎に担任の教官にできるだけ学力等がかたよらないように選んでもらった。

テストは6名の検査者によって1979年7月9日から7月11日の間に実施された。テスト時間は個人差はあるが，予備テスト部が5分程度，本テスト部が20分程度で

あった。

まず，予備テストを個別に実施し，すぐに簡易スコアを求め，その結果より各児童のための標準版を決定した。この標準版は集団的に実施された。

Ⅳ　結果の分析と考察

1.　簡易スコアの分布

予備テストにおける被験者の簡易スコア分布状況を確認するために，各学年毎の簡易スコアの分布を図5に示した。図の横軸は簡易スコアを，実線で各学年で起こりうる簡易スコアの範囲を，その上の白丸は各被験者をそれぞれ表わしている。ここでいう各学年で起こりうる簡易スコアの範囲とは次のようにして決まるものである。例えば，3年生ならば第1ページでは項目番号3から出発して，その被験者の反応に応じて分岐する。いまその被験者が全ての項目に誤答した場合を考えてみると，順次与えられる項目の番号は3, 2, 1, 1, 1, 1, 1, 1, 1となり簡易スコアは12である。

一方，その被験者が全ての項目に正答した場合を考えてみると，順次与えられる項目の番号は3, 4, 5, 6, 7, 7, 7, 7, 7となり簡易スコアは53である。これらがそれぞれ3年生のとりうる簡易スコアの最小値・最大値となるから，12から53が3年生のとりうる簡易スコアの範囲ということになる。

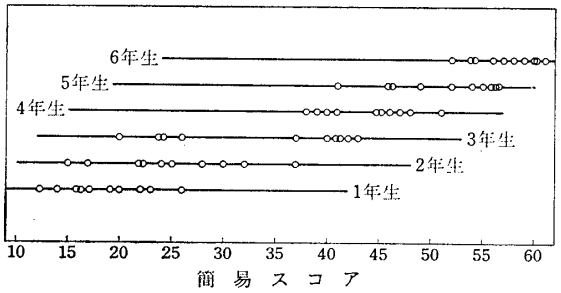

図5　簡易スコアの分布

図5によると最低12・最高61の間に全被験者が比較的均等に分布している様子がうかがわれる。特定のスコアに被験者が集中しなかったということから予備テストにおける被験者の簡易スコアの分布状況は望ましい結果を示していると言えよう。細かく見ると34, 35, 36付近に該当する被験者数が少ないのでこの点についてはデータ数が増加した時点で再検討すべきである。

2.　簡易スコア，予備テストの推定値，本テストの推定値の相関

予備テストの分岐が実際に有効に機能していることを確認するために，簡易スコアx・予備テストの推定値$\hat{\theta}_p$及び本テストの推定値$\hat{\theta}$の同時分布が図6-a, b, cに示されている。

図6-aによると簡易スコアxと予備テストの推定値$\hat{\theta}_p$の相関係数は.93とかなり高い値を示しており，続いて実施すべき標準版の決定に簡易スコアを用いても十分にその目的を果たしうることが確認された。

図6-bによると簡易スコアxと本テストの推定値$\hat{\theta}$の相関係数は.87でありこれもかなり高い値である。このことは予備テストの分岐機能が有効に働いていて，実施すべき標準版の決定という目的を十分に果たしていることを示している。

図6-cによると予備テストの推定値$\hat{\theta}_p$と本テストの推定値$\hat{\theta}$の相関係数も.88とかなり高く，プログラムテストの有効性を示している。

以上，図6-a, b, cについて全て高い相関関係を表わしていることを述べたが，各図中に数名だけ全体の傾向と

表4　プログラムテストと組み合わせて使用される標準版語彙テストの特性

版　名	項目数	項目困難度	項目識別力	適用程度
AP1	30	−5.37 (0.66)	0.83 (0.18)	年少組
AP2	30	−4.87 (0.38)	0.82 (0.22)	年長組
B1	34	−4.23 (0.41)	0.83 (0.21)	小　1
B2	36	−3.54 (0.67)	0.76 (0.23)	小　2
B3	38	−3.02 (0.63)	0.68 (0.22)	小　3
B4	40	−2.36 (0.58)	0.65 (0.17)	小　4
B5	45	−1.27 (0.56)	0.65 (0.16)	小　5
A6	56	−1.14 (0.82)	0.61 (0.19)	小　6
J1	56	−0.14 (1.25)	0.54 (0.17)	中　1

B1からB5までの版は芝（1978）におけるA1からA5までの版の改訂版である。項目困難度は芝の仮尺度のbの値の平均値をあらわし，カッコ内はその標準偏差を示す。項目識別力も同様のaの値の平均値をあらわし，カッコ内はその標準偏差を示す。

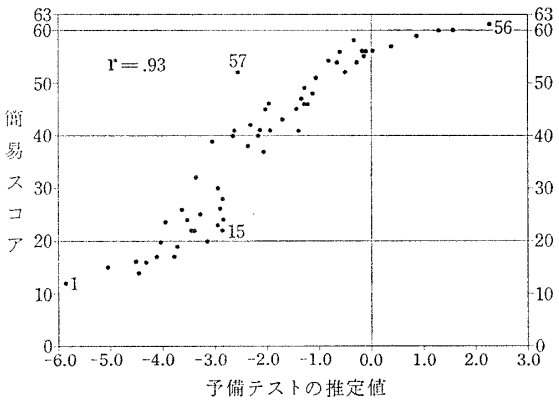

図6-a　簡易スコアと予備テストの推定値の相関

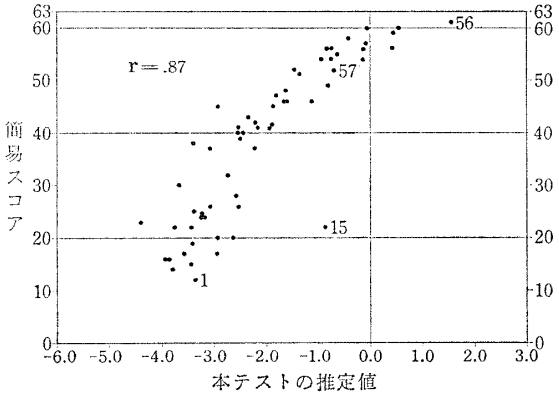

図6-b　簡易スコアと本テストの推定値の相関

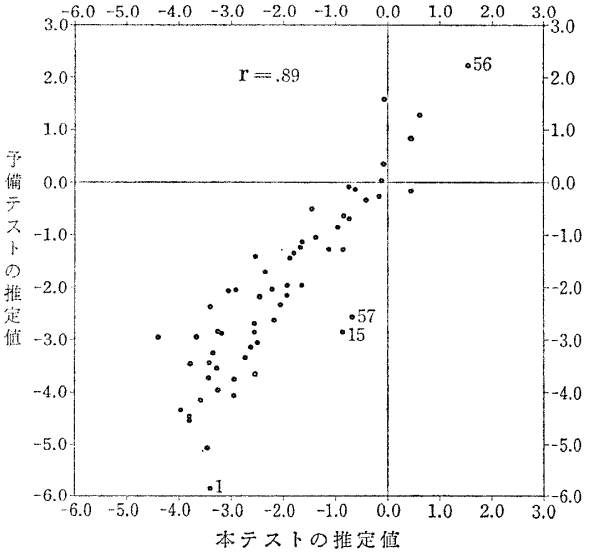

図6-c　予備テストの推定値と本テストの推定値の相関

はずれた状況を示している被験者がいる。特に顕著な事例については図中に被験者番号を示した。これら4名の被験者については本研究で使用している推定法の性質から推定値 $\hat{\theta}_p$ または $\hat{\theta}$ の値の信頼度が低いためである。

次に予備テストの各ページ実施後の推定値と精密測定部の推定値 $\hat{\theta}$ の相関は第1ページ実施後が.52，第2ページ実施後が.58，以下.58，.70，.74，.83，.86，.88とページを進むごとに順次高くなってくるが，このことは第1ページ以降の分岐が全体的に見て適切であったことを示している。第1ページ実施後に既に.52と比較的高い相関が得られているのは学年により第1ページで実施する項目を決定するというように事前情報が活かされているためである。

3.　情報曲線の変化

プログラムテストの評価で重要なのは，個々の被験者にその理解力に応じた最適な困難度の項目を与えて測定の精度を高めるという目的がうまく果たされているかという点である。このことを確かめるためには，特定の被験者が実際に進んだ分岐の道筋の項目群に対する情報量 $I(\theta)$ の大きさで評価できる。また，本テストとして実施した版が適当であったか否かについて情報量 $I(\theta)$ の大きさで評価できる。(情報量については芝 (1978)，Birnbaum, A. (1968) などを参照。)

ただし，現在のところ情報量 $I(\theta)$ の大きさの評価基準が確立していないので，指定された標準版における当該被験者の推定値付近の情報量については情報曲線の中の相対的な評価にとどまる。

まず，予備テストによる推定値 $\hat{\theta}_p$ と本テストによる推定値 $\hat{\theta}$ とが比較的一致している望ましい事例として被験者55について検討する。

被験者55：簡易スコアが54で，A6版が指定され，推定値が −.96である。予備テストの推定値は −.87である。本テストの推定値と予備テストの推定値が非常によく一致しているのが特徴で，予備テストで適切な分岐が行なわれたために8項目という少数の項目で正確な推定ができたものと思われる。実際に予備テストで8項目中4項目に正答し，A6版でも56項目中27項目に正答し，いずれも正答数にかたよりがない。この点からも分岐が適切に行なわれたことが確認できる。図7で見ると確かに予備テストで1項目実施する度に，この被験者の推定値付近で高い情報が得られて行く様子がうかがわれ，本テストでは点線で示したようにまさにこの被験者の推定値付近でA6版の最大情報量が得られることがわかる。この被験者はプログラムテストが最も成功した例と言えよう。

次に，前にも述べたように本研究では最尤推定法により個々の被験者の理解力の推定値を求めているが，この方法はその特徴から被験者の反応が正答または誤答にかたよりすぎると推定の信頼度が低くなる。このため，予備テストで正答または誤答が2つ以下の被験者，本テストで正答または誤答が5つ以下の被験者の推定値 ($\hat{\theta}_p$ ま

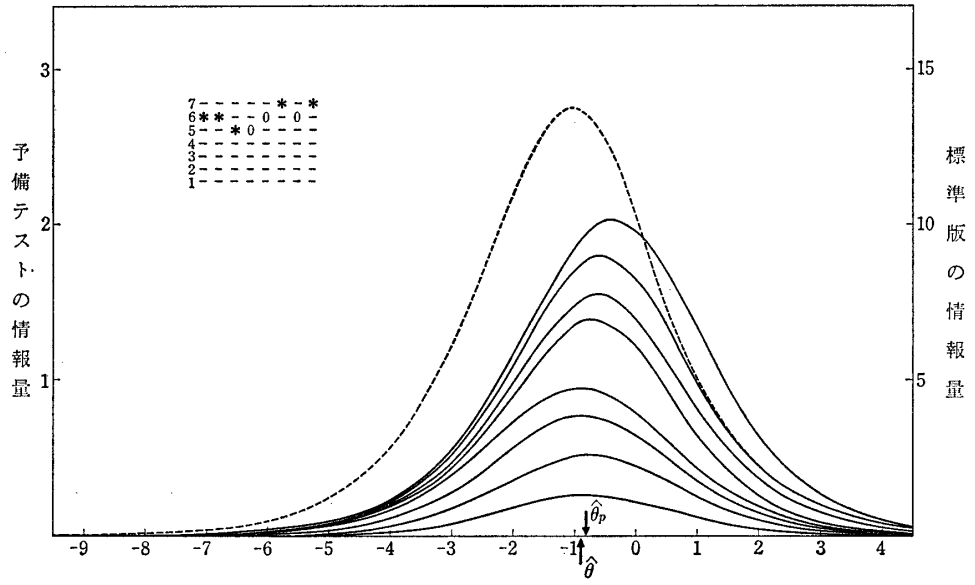

図7　被験者55の予備テストへの応答パタンの情報曲線と，この被験者に実施された標準版の情報曲線

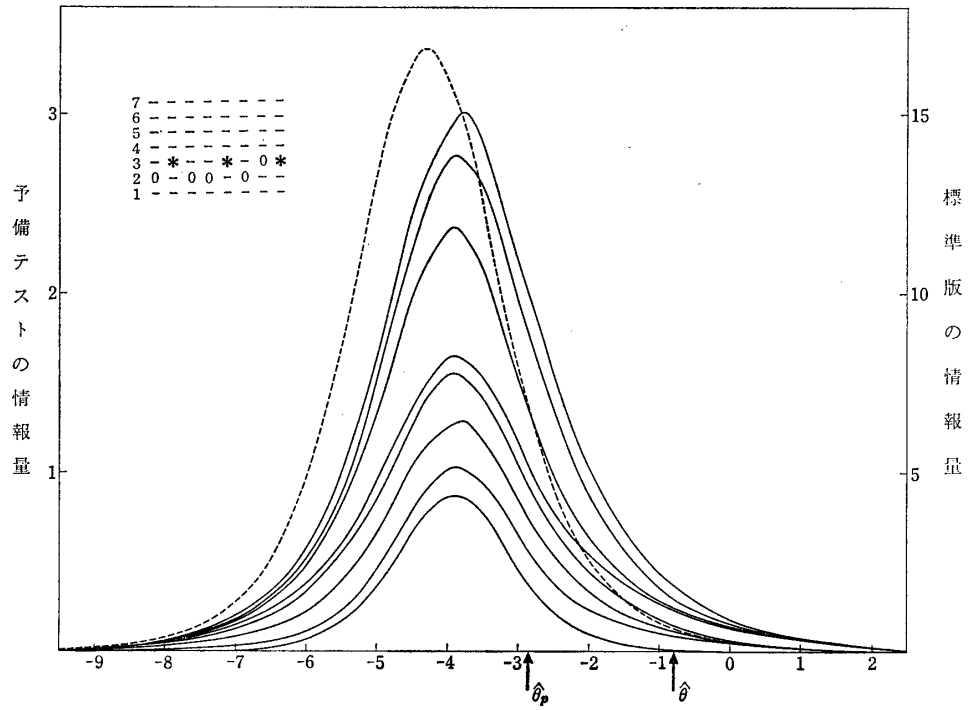

図8　被験者15の予備テストへの応答パタンの情報曲線とこの被験者に実施された標準版の情報曲線

たは $\hat{\theta}$) は信頼度が低くなる。このような結果となった事例として被験者15について検討する。

被験者15：簡易スコアが22で，B 1版が指定され，推定値が−.86である。予備テストの推定値は −2.88 である。図8を見ると予備テストで情報曲線が1項目実施する度に上昇して行く様子がうかがわれる。しかしながら $\hat{\theta}_p$ も $\hat{\theta}$ も情報曲線のピークからはずれている。特に $\hat{\theta}$ での情報量 $I(\hat{\theta})$ は8項目実施後でピークの情報量の1/7以下という低い状態である。さらに本テストのB 1版について見ると図8に点線で示したように情報量 $I(\hat{\theta})$ が .73でありB 1版の最大情報量の 1/20 以下である。このことについては，B 1版で34項目中33項目に正答しているために推定法の制約から正確な推定値が得られなかったということとB 1版ではなくそれより難しい標準版に

分岐すべきだったという2つの原因が考えられる。簡易スコア22をB2版以上の版に分岐させることは他の事例とも合わせて検討する必要がある。

4. まとめ

本研究で予備テストとして用いた多層適応形テスト（L版）は一般に測定の効率を上昇させるのに有効な方式であることが確認された。個々の被験者について考えると本テストで使用する標準版が易しすぎるような場合があるが，これはむしろ簡易スコアにもとづいて標準版を指定する際の問題で，どの程度のスコアをどの版に対応づけるかについては今後データ数を増やしつつ検討する必要がある。この場合，簡易スコアではなくて予備テストの推定値を用いるならば，より精度の高い対応づけが可能となろうが，そのためにはコンピータなどの補助手段が使用可能であることが前提となる。測定場面へのマイクロコンピュータなどの導入という問題は今後の検討に値するものと思われる。さらに残された問題に，どの程度の情報量 $I(\theta)$ が得られたならば測定の情報として満足できるかという情報量 $I(\theta)$ の評価基準の問題がある。このように，いろいろと残された問題をかかえてはいるが，本研究ではプログラムテストと標準版を併用した方式に対してその有効性が示された。

付 記

本研究にあたって，東京学芸大学附属世田谷小学校（校長：八野正男教授）およびスイスのジュネーブ日本語学校（小根山茂子校長）に御協力いただきました。御関係の諸先生，生徒の皆さんに心から感謝致します。

文 献

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In Lord, F.M. & Novick, M.R. Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley 1968

野口裕之 ベイズ的逐次テストと項目固定型テストの比較 東京学芸大学紀要 第1部門 教育科学 第30集 pp. 47-56 1979

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

芝 祐順・野口裕之・南風原朝和 語彙理解力測定のための多層適応形テスト 教育心理学研究 第26巻 第4号 pp. 229-238 1978