プレースメント・テストの妥当性と今後の展望

三枝　令子

要　　旨

筑波大学留学生教育センターでは、毎学期はじめに、プレースメント・テストを行い、その結果をもとにクラス分けを実施している。本稿ではまずプレースメント・テスト第二版の結果とその分析、並びに第一版との比較を行う。さらにクラス分けテストとしての妥当性を検討し、妥当性を高めるためにセンターの教育の到達目標、また日本語能力という概念の明確化の必要性が論じられている。最後にセンター以外の五つの日本語教育機関におけるこのテストの使用状況が報告される。

＜キーワード＞　プレースメント・テスト、項目分析、テストの妥当性
到達目標、日本語能力

1．はじめに

先に「プレースメントテストの統計的処理の試み」と題して『日本語論集』第二号（1982）筑波大学留学生教育センターで、1985年秋より三回に渡って実施したプレースメント・テストの結果並びに結果の分析を報告した。これを第一版と称する。本稿ではまず、この第一版に引続き作成した第二版のテストの1987年春より三回に渡る実施の結果並びに分析の結果を報告する。今回の版は、前回の改良という側面があるので、第一版との比較も行う。

あるテストがよいテストであるか否かはテスト内部の検討と同時に、そのテストが測定しようとしているものを測定し得ているかどうかという妥当性の面からも評価されなければならない。改良がなされてテスト自体が信頼できるものになったとしても、それはこのプレースメント・テストが役に立つということを意味しない。もちろん測定したいものを完全に測定できるテストというのは存在しないだろうし、また妥当性という概念自体一義的に定まるものではない。しかし妥当性の検討を行うことによってより妥当性の高いテスト作りが可能になると考えられる。ここでは妥当性を従来行われているように三つのタイプ、内容的妥当性、基準関連妥当性、構成概念妥当性に分けて、それぞれの面から今回のテストを検討することにする。最後に希望により一版二版とも他の日本語教育機関において使用されたので、その実態についても言及し、このテストの一般化の可能性を検討する。

2．第二版テストのねらい

第一版は1985年秋より三回用いられた。しかし第一に、いくつかの改良すべき点が明らかになっ

たこと、第二に、少数ながら同じ学生が同一試験を受験しているため、新たな版の作成が必要になった。そこで1986年秋より第二版作成の準備が進められた。第二版も「語彙」「聴解」「読解」「文法」「文字」という五つの下位テストは変えないことにした。ただ第一版が全体として二時間半かかり、学生の負担が大きかったので、各下位テストを50問から30問に減らし、全体で2時間以内に納めることを目標とした。以下、1987年4月実施分のテストを P87S、87年10月実施分のテストを P87A、88年4月実施分のテストを P88S と称する。Pは Placement、A, S はそれぞれ Autumn、Spring の略である。また「語彙」「聴解」「読解」「文法」「文字」という下位テストを、時にそれぞれ WOR、LIS、RED、GRA、LET と略す。問題の作成には主に以下の者が当たった。

語彙　：　大坪一夫、堀口純子、酒井たか子
聴解　：　西村よしみ、加納千恵子
読解　：P87S　　佐久間まゆみ、小口叔枝、川原裕美、竹中弘子
　　P87A/P88S　酒井たか子　三枝令子
文法　：　野田尚史、山本一枝、三枝令子
文字　：　戸田昌幸、加納千恵子、田中幸子　　（順不同）

問題作成に当たって、問題数の減少の他に第一版の分析結果及び反省から次のようなことをあらかじめ各担当者に依頼した。

【語彙】　第一版が文の穴埋め問題や、正しい語の使い方をしている文の選択問題であったため、文法問題との境界がはっきりしなかった。そこでできるだけ独立に語彙力だけを見る問題にする。

【聴解】　日本人の話が聞けるということを目標にして、音の聞き分けではなく内容が取れるか否かを見るようにする。

【読解】　第一版では半分は文法問題であった。読解という以上長い文を読んで、内容がどれだけ取れるかということで、測るようにする。そのためには課題文はある程度少なくなっても構わない。

【文法】　第一版では他の下位テストに比べてやさしかった。難易に幅を持たせて識別力を上げるようにする。

3．テストの内容

各下位テストは30問よりなり、配点は各一点で全体で150点満点である。解答の形式は、「文字」に記述式の問題がある他は、ほとんど三肢もしくは四肢選択である。ただし「読解」については、P87Sのみ真偽式（○×式）を用いている。[1]　出題の内容と問題数は以下の通りである。

【語彙】

1　例のふたつの語彙の関係に合わせて、課題語彙の反意語、助数詞等を選ぶ。＜1点×6問＞

例1　1000円：10円＝たかい：（　　）

1 すくない　2 ひくい　3 やすい　4 ちいさい

例2　茶：粗茶＝自宅：（　　）

1 お宅　2 拙宅　3 私邸　4 貴邸

2　文脈に合う語彙を選ぶ。＜1点×24問＞

例3　二階から庭に（　）、けがをした。

1 ふって　2 おちて　3 ころんで　4 たおれて

【聴解】　問題はテープレコーダーによって与えられる。

1　ミニマルペア　テープの会話を聞いて、会話文中の語もしくは文と同じものを選ぶ。四肢選択。＜1点×4問＞

例4　いつ事故にあったんですか。

通学の時です。

1　ちゅうがく　2　すうがく

3　しゅうがく　4　つうがく

2　テープの文を二度聞いて、内容と一致する絵を選ぶ。三肢選択。＜1点×5問＞

3　ニュース（教師作成）を聞いて、質問に答える。三肢選択。質問のみ記述されている。＜1点×5問＞

4　ふたつの会話を聞いて、登場人物の人間関係や話し手の意図についての質問に答える。二肢から四肢。＜1点×5問＞

5　三つの会話を聞いて、登場人物の人間関係や感情についての質問に答える。二肢から四肢。＜1点×6問＞

6　長い会話文を聞いて、内容についての質問に答える。四肢。＜1点×5問＞

【読解】

読解は87年春とそれ以後の二回とでは問題が全く異なる。この理由については後の項目分析のところで述べる。

P87S

与えられた文が課題文の内容に合うかどうかを答える問題。課題文は八つ。一題は4文から6文からなる。各課題文に3～5問の問題がある。＜1点×30問＞

例5　宮崎県の幸島のサルは、砂だらけのいもを水につけ、手で洗って砂を落としてから食べることを思いついた。次に、水より海水につけて食べる方がおいしいと気付いた。今では、この島の子ザルはみな、いもを海水につけて食べている。この知識は、初めは子ザルが遊んでいるうちに覚えたことが、しだいに母ザルにも広がっていった。しかし、なぜか、父ザルには広がっていかなかった。

1（　）幸島のサルはいもをしばらく水につけたあとで食べる。

2（　）幸島のサルはいもの砂を手で落として食べる。

3（　）幸島の子ザルはいもを海水につけて、砂を落としてから食べる。

4（　）子ザルから母ザルにいもを洗うことが伝わった。

5（　）父ザルはいもを洗って食べることを知っていた。

P87A／P88S

1　文の内容と合っている文を選ぶ。四肢選択。＜1点×5問＞

2　文（１文）と同じ内容の文を選ぶ。三ないし四肢選択。＜1点×15問＞

3　課題文（5文から10文）についての質問に答える。四肢選択。＜1点×10問＞

【文法】

1　文脈に合う助詞、副詞、指示詞等を選ぶ。＜1点×25問＞

2　4文の中からもっとも適当なものを選ぶ。＜1点×5問＞

【文字】

1　漢字の共通部分を選ぶ。記述式。＜1点×4問＞

例6　語　話　読　課→　　（正答は「言」）

2　漢字の読み方を書く。記述式 。 ＜1点×10問＞

3　正しい漢語の切り方を選ぶ。三もしくは四肢。＜1点×5問＞

例7　天気予報：1　天／気予報　2　天気／予報　3　天気予／報

4　漢字を書く。＜1点×8問＞

5　かたかなを書く。＜1点×3問＞

4. 実施方法

4.1 実施時間

実施時間は全体で2時間10分である。テストの開始前に各下位項目の所要時間を言っておき、ひとつの下位テストが終わると直ちに問題文を回収した。表1に問題の提出順序と制限時間を示す。

	P87S/P87A	P88S
1	WOR 20	WOR 20
2	LIS 45	LIS 45
3	RED 30	GRA 30
4	GRA 20	RED 25
5	LET 15	LET 10
計	130	130

表1 問題の提出順序と制限時間 (単位：分)

「聴解」は疲労度の少ないテスト開始直後が望ましいが、遅刻者がいた場合やり直しがきかないので、2番目に行い、また「読解」は疲れるため最後は避けた。当初2時間以内に全体を納めたいと考えていたが、「聴解」が予定より長くなり、2時間を越えた。

4.2 採点基準

第一版と変更はない。「文字」の漢字の書きについては、今回は送り仮名も書くように問題文で指示したので、それも採点の対象にした。また前回同様書かれた文字をどこまで正解とするかという基準を作り、それによって採点の統一を図った。

4.3 採点方法

前回同様素データの入力には「dBASE Ⅲ」[2]を使用した。[3]その後の処理は全く前回と同じである。

なおP87Aより学生にもテストの結果を知らせることにした。プレースメントテストの結果が、学生の日本語力そのものを反映しているわけではないが、1　テストを受けた人の中で自分の力がどのぐらいかを知ること、　2　五つの下位テストのなかで自分はどこが弱いかを知ること　は学生にとっても意味のあることと考え、名前を伏せて成績の一覧表を張り出し、自分がどの成績かは教師に聞いてわかるようにした。

5．受験者概要

表2は、受験者の母語別の人数を示す。PALL2　は、P87S、P87A、P88S　の三回を合わせた人数である。PALL1　は、第一版のテストの人数の合計である。

テスト / 母語	P87S	P87A	P88S	PALL2	PALL1
総数	107(100)	94(100)	93(100)	294(100)	272(100)
中国語	48(45)	36(34)	44(47)	128(44)	107(39)
韓国語	23(22)	8(9)	20(22)	51(17)	71(26)
その他	36(34)	50(47)	29(31)	115(39)	95(35)
国籍延べ数	22	18	18	31	32

表2　受験者の母語別構成　　単位：国籍延べ数以外は人数
（　）内はパーセント

表2からわかるように、各回とも約100人の受験者である。この数は第一版の時からほとんど変化していない。母語別について見ると、P87A では「韓国語話者」が極端に少なく、一方「その他」が半数を越えている。また「中国語話者」においても、表3に見られるように中華民国がこの回は少なく、逆に中華人民共和国と香港が多いという例年と随分異なった構成になっている。「その他」には、アメリカ、タイ、インドネシアが上位三カ国として含まれており、これが63％を占める。

テスト / 国籍	PALL2	P87S	P87A	P88S
総数	294	79	94	93
中華人民共和国	70	22	26	22
大韓民国	51	23	8	20
中華民国	48	24	4	20
アメリカ	21	5	14	2
タイ	20	5	11	4
インドネシア	13	3	7	3
ブラジル	9	4	3	2
フィリピン	9	2	3	1
香港	8	1	5	2
スリランカ	7	2	3	2
オーストラリア	4	2		2
フランス	3	1	1	1
デンマーク	3		1	2
ヨルダン	3	1	1	1
バングラデシュ	3	2		1

メキシコ	2			2
イギリス	2	1	1	
カナダ	2	1	1	
ニカラグア	2	1	1	
ポーランド	2	1	1	
コロンビア	2	1	1	
エジプト	2	1		1
イタリア	1	1		
イラク	1	1		
シンガポール	1	1		
ノルウェー	1			1
ニュージーランド	1		1	
ペルー	1		1	
アフガニスタン	1	1		
ベネズエラ	1			1

表3　受験者の国別構成　（ 単位：人数 ）

6．テストの結果[4)]

6.1 成績

テスト / 下位テスト		P87S	P87A	P88S
WOR	満点	30	30	30
	平均	16	15	16
	標準偏差	5	6	6
	最高値	26	29	29
	最低値	3	2	2
LIS	満点	30	30	30
	平均	12	17	19
	標準偏差	5	6	6
	最高値	26	30	30
	最低値	0	1	2
RED	満点	30	30	30
	平均	20	15	17
	標準偏差	6	5	6
	最高値	30	28	28
	最低値	1	0	3
GRA	満点	30	30	30
	平均	14	13	14
	標準偏差	5	6	6
	最高値	24	27	27
	最低値	2	1	2
LET	満点	30	30	30
	平均	16	15	17
	標準偏差	6	7	7
	最高値	29	28	28

	最低値	4	0	0
TOT	満点	150	150	150
	平均	78	75	83
	標準偏差	23	27	26
	最高値	126	130	132
	最低値	14	11	30

表4　テストの結果（単位：点）

表4は、テスト三回分の成績を示している。TOTは、五つの下位テストの合計点を示す。（以下同様）全体の平均点は、P87A、P87S、P88Sの順に高くなっている。第一版が全体で最大2点の開きしかなかったことと比べると、第二版のテストの受験者は回によってかなり日本語力に差があった可能性がある。「語彙」「文法」「文字」は各回とも安定していて平均ではほとんど差は認められない。「聴解」は、P87Sが極端に悪い。逆に「読解」は、P87Sがよい。この結果は問題の有り様をそのまま反映していると言える。すなわちP87Sでは、「聴解」は問題の指示が非常にわかりにくく、多くの受験者が試験場でパニックを起こし、試験官の再三の説明にもかかわらず、どう解答すべきかわからないまま問題が進んでいくような状況であった。一方「読解」は、P87Sでは問題が真偽式であったため、正答率が高くなった。このため　P87A 以降は、「聴解」は問題はほとんど変えずに、問題用紙やテープの指示の与え方、また問題文を聞く回数を増やす等の変更を行った。一方「読解」は、問題を変えずに解答形式だけを変更するのは困難なため、全面的に改訂して解答を四肢選択の形式にした。この点で　P87Sとそれ以降の二回のテストとは厳密な比較はできないことになった。P87AとP88Sの二回に限って言えば、五つの下位テスト全てにおいてP88Sの方が平均値が高く、P88Sの受験者の日本語力が高かったと言える。標準偏差は、第一版と同様「文字」で最も大きい。五つの下位テストの中で部分的ではあるが唯一記述式の解答形式を取っているため、力の差が大きく出て、分布が広がったと考えられる。

図1～6に第二版の3回のテストの得点のヒストグラムを示す。[5]　X軸が学生の得点、Y軸が人数を表す。X軸の目盛りは下位テストが3点おき、合計の得点は15点おきである。また下位テストの場合、X軸上の0点には0～2点までの学生の人数、3点には3～5点までの学生の人数が目盛られている。以下合計の図も含めて同様である。

今回は第一版ほど、下位テスト毎の共通した傾向というものが見られない。これには　第一に問題の変更、第二に受験者の能力差が考えられる。能力の差という点については、P88S の受験者自身の能力の高さと P87A の学生の国籍の多様性が原因のひとつとして指摘できる。

【語彙】

P87A では低得点から高得点にかけて平均的に分布しているが、P87S は、低得点に、P88S は高得点に幾分分布が偏っている。

【聴解】

P87Sでは分布が低得点に偏っている。この原因については既に述べたとおりである。しかしP87AとP88Sも同じ問題でありながらかなり異なった分布をしている。すなわち P87A は平均の近傍に分布が集まり、P88S は明らかに高得点に分布の山がある。

【読解】

P87S では分布が高得点に偏っている。P88S は分布の山が二つあるが、全体としては高得点に分布が偏り、P87Aは低得点に偏っている。

【文法】

三回のテストで問題にほとんど変更がない。全体に低得点に偏っている。第一版はやさしすぎたので今回は難しい問題を作ろうとした作成者の意図は達せられたと言える。

【文字】

P87S と P88S には分布に二つの山がある。P87A では低得点から中得点にかけてほぼ均一な分布がみられる。この回は中国語話者、特に中華民国の学生が少なかったことが影響しているのかもしれない。全体として五つの下位テストの中では「文字」がプレースメントテストの分布のあり方として最も適当と言えそうである。

【全体（TOT）】

いずれのテストも、90～105点の間に分布の山があり、同じような分布をしている。下位テストのそれぞれには問題として不備な点があっても、問題が数集まることによって多少の不備は補われるとみてよいだろう。

さてプレースメント・テストとしては各下位テストの合計点をもとにクラス分けが行われるわけで、その全体の分布がクラス分けのためにどれだけ妥当性を持っているかということが一番問題なところである。先に２号（1986）で述べたように、センターの現状では得点が高いものから低いものまでを同程度の精度で識別できるような問題が一応適当と考えられる。その点で今回のテストは分布が平均の近傍に集まりすぎ、特に P87A の学生には全体に問題がむずかしかったと言える。ただテストの真の妥当性はこのテストによるクラス分けがコース運営上適当であったかどうかということで判断されるべきものであり、分布からだけでは何とも言えないところが大きい。テストの妥当性については後に再度検討することにする。

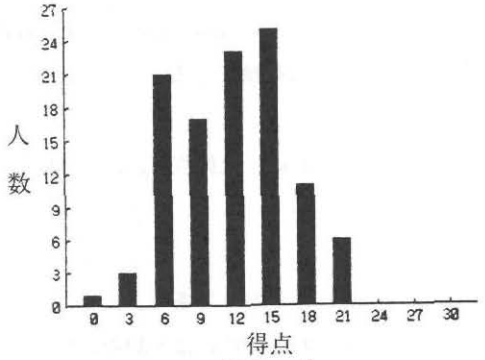

図1－1

P87S語彙

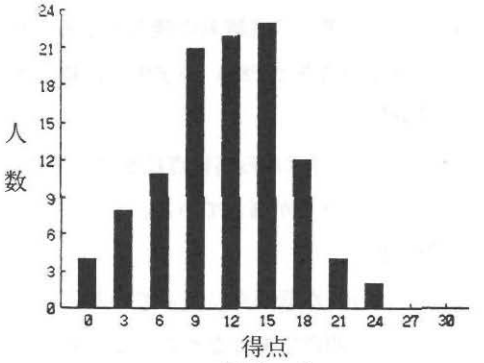

図2－1

P87S聴解

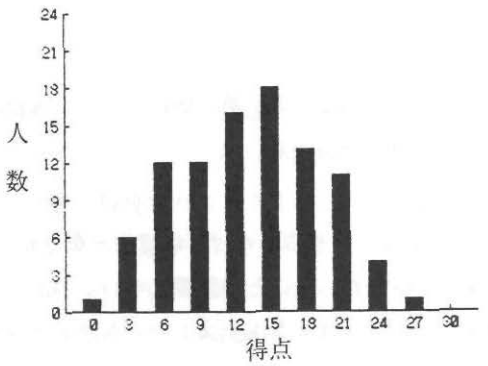

図1－2

P87A語彙

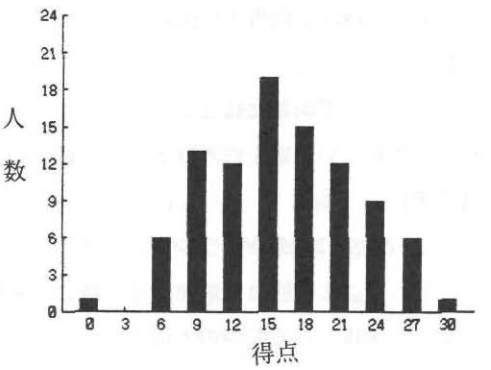

図2－2

P87A聴解

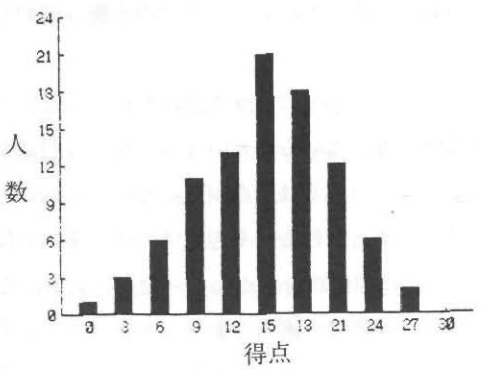

図1－3

P88S語彙

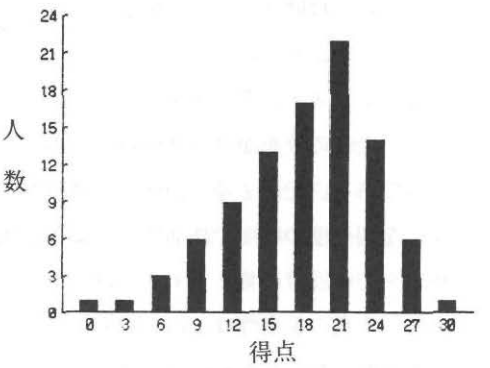

図2－3

P88S聴解

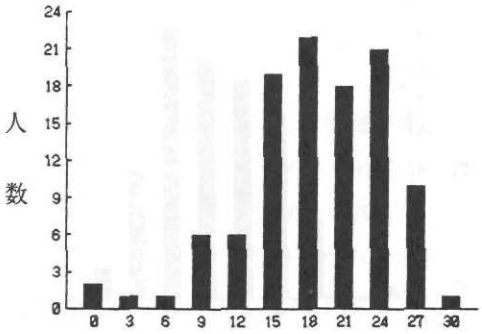

図３－１

P87S読解

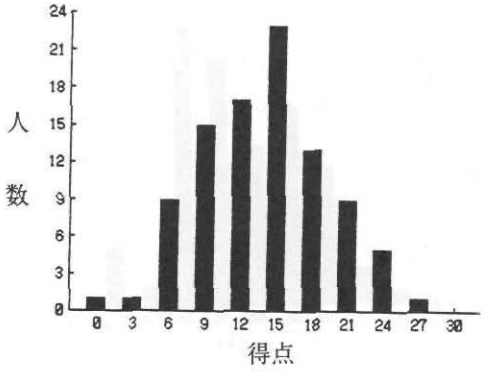

図３－２

P87A読解

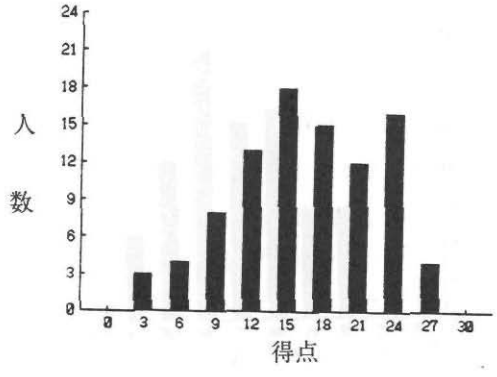

図３－３

P88S読解

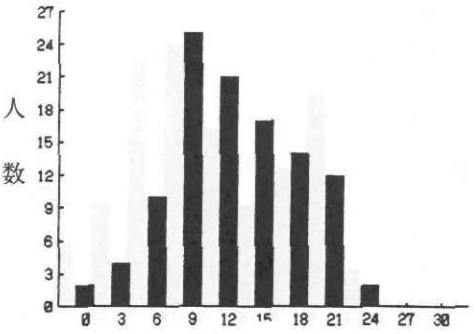

図４－１

P87S文法

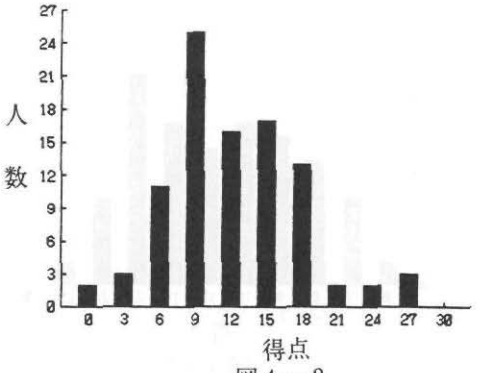

図４－２

P87A文法

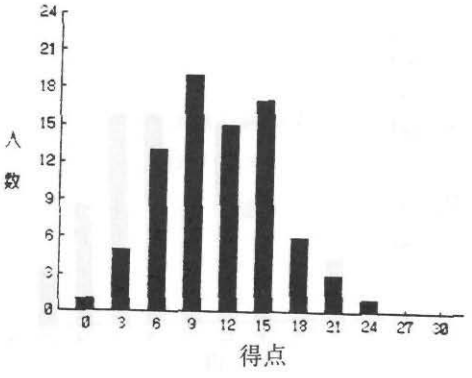

図４－３

P88S文法

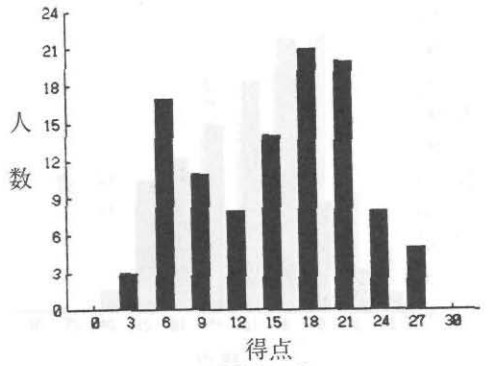

図５－１

P87S文字

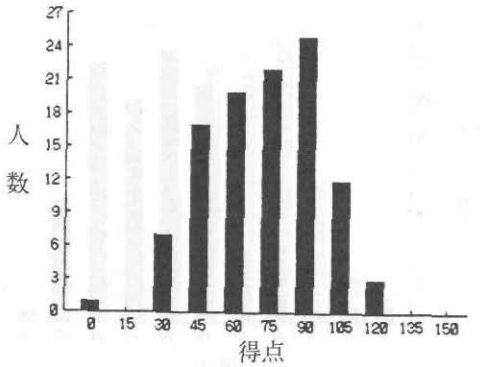

図６－１

P87S総点

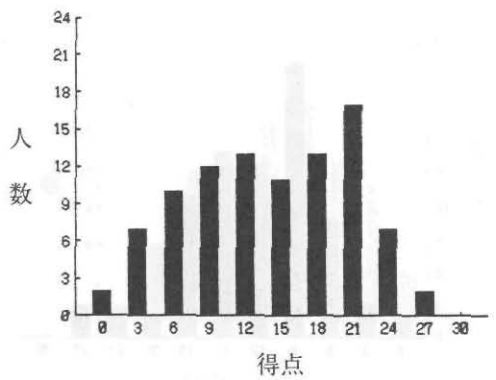

図５－２

P87A文字

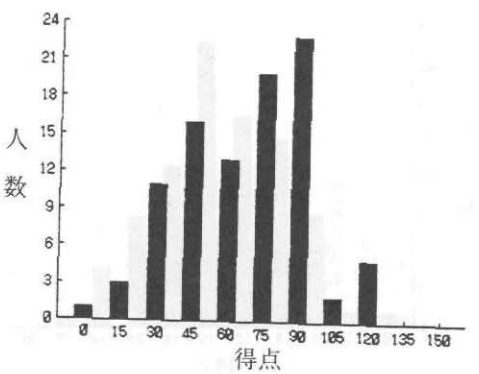

図６－２

P87A総点

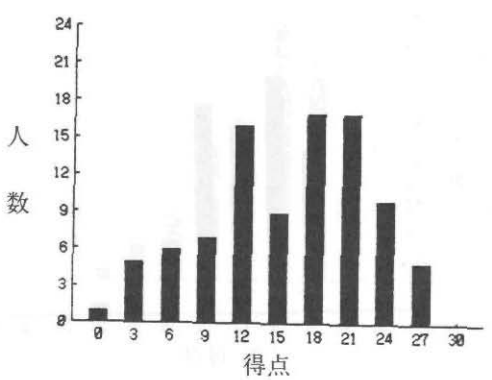

図５－３

P88S文字

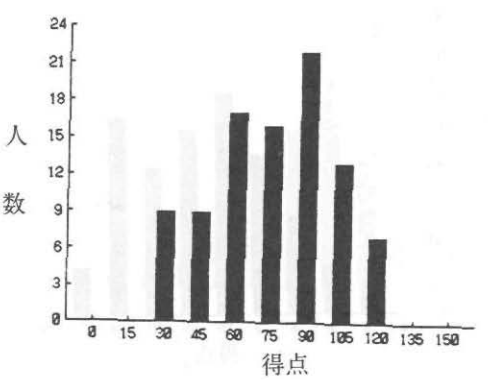

図６－３

P88S総点

7．テスト問題の検討

7.1 項目分析

今回は P87S と後の二回とで問題が異なるため、第一版の時のように全てのデータを合わせることができない。そこでそれぞれの回について項目分析を行ない、また共通問題によるP87AとP88Sについては、項目分析の統計的な精度を高めるために両者を合わせた項目分析も行った。これをPALL2-2と称する。[6] 但し、この二つのテストにも若干問題を変更した箇所があり、それらの問題は当然分析から省いた。図７～１[15] は、P87S と PALL2-2　のそれぞれの下位テストの正答率と識別度をプロットしたものである。P87A、P88S それぞれのの項目分析の図は省略するが、その結果については必要に応じて言及する。図のＸ軸は正答率（時にDIFと略す）、Ｙ軸は識別度（時にDISと略す）を表す。それぞれの算出方法、また数値の意味するところは先の２号に記したので参照されたい。簡単に言えば、ここで言う正答率は受験者全体を3群に分けた場合の成績上位者と下位者における正答者の割合で、この値が低ければその問題がむずかしいこと、高ければ問題がやさしいことを意味する。また識別度は、その問題が総得点での成績上位者と成績下位者とをどれだけよく識別するかの指標で、この値が高ければ、識別力のある問題と言える。目安として DIFの値は 0.3～0.7位、DISの値は 0.4以上が適当と言える。今回は一回毎の項目分析を行って比較したので、同じ問題であっても受験者の異なりから違う値を示した。問題によって値の違うものと違わないものがあるのは興味深い。DISで 0.3以上の開きのある問題もあった。ここでは、各下位テストにおいて問題が全体としてバランスよく分布しているかどうかということと、三回のテストを通じて特に識別度が悪い問題を明らかにすることを主な目的として分析する。なお図中の白抜きの四角は、同じ値に二つの問題がプロットされていることを示している。

【語彙】

図７－２のPALL2-2の結果をみると、問題がかなりきれいに三角形の辺上に乗っていることがわかる。P87Sは、識別度が0.4以上の問題が40％と少なく、三角形が小さい。PALL2-2で識別度が0.4以下の項目をみてみると、11題中3題が正答率が0.4～0.6で、これらの問題の識別度の低さは、問題がやさしすぎるためでもむずかしすぎるためでもないことがわかる。図７－２で、三角形の内側に入っている問題がこれに当たる。この原因については酒井[7] 参照。P87S で識別度がマイナスになった問題は次の問題であった。

「交渉は（　　）にさしかかった。」

1壁　2頂点　3最高　4やま場

(DIF　0.13　　DIS　-0.06)

そこで P87A 以降この問題を差し替えた。一方、P87S、PALL2-2 とも識別度が最も高かった問題は次の問題である。

「おばあさんは子供に（　　）。」
1　かわいい　2　つらい　3　好きだ　4　あまい

選択肢の 4 と並んで 3 を選ぶものが、特に成績下位者に多かった。

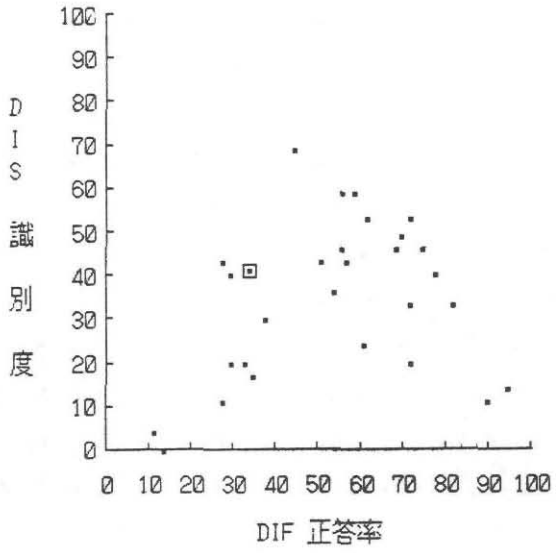

図７－１　P87S語彙におけるＤＩＦとＤＩＳの関係

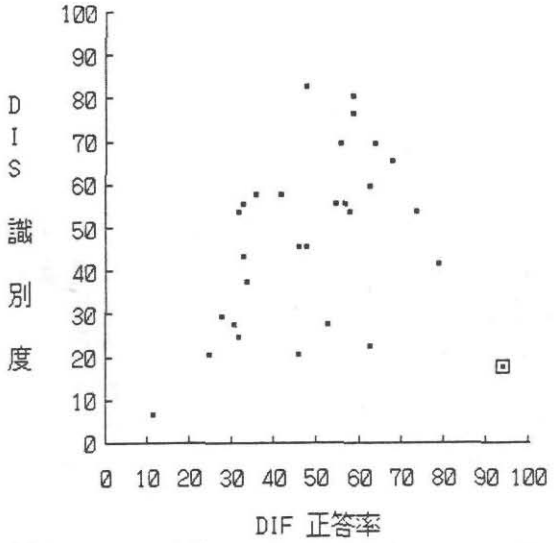

図７－２　PALL語彙におけるＤＩＦとＤＩＳの関係

【聴解】

「聴解」は、テープの問題文自体には変更がないが、問題用紙における指示の仕方を大幅に変えている。こうした指示の変更が受験者の反応にどのように影響するかということについては、この次のプレースメント・テスト第三版の結果と合わせて、聴解問題作成者が報告する予定であるということなので、ここではその点について詳細な分析は行わない。本稿は、プレースメント・テスト第二版の報告という意味あいが大きいので、その範囲で問題用紙の変更とその結果表れた目立だった変化について述べることにする。表5が三回のテストの指示及びテープの変更箇所である。表5にも記したが、指示及びテープの変更の他に、問題5の6題のうち4題は選択肢の文がP87Sで正答が一義的でない等の理由で不適当と判断されたため、P87A以降変更改良した。問題文に手を加えたのはこの箇所だけである。

テスト／問題	P87S	P87A	P88S
1	テープの会話文を聞いて、次の1.2.3.4.の中から答えの文と同じものを1つ選びなさい。問題は一度しか読みません。	テープの会話文を聞いて、次の1.2.3.4.の中から答えの文と同じものを1つ選んで解答用紙にその番号を書きなさい。問題は一度読みます。	テープの会話文を聞いて、次の1.2.3.4.の中から答えの文の中にあることばと、同じものを一つ選んで解答用紙にその番号を書きなさい。問題は一度しか読みません。問題を始める前に例を一つ聞いてみます。
	記述された例題なし。	記述された例題なし。	記述された例題が一題
2		問題用紙の絵の書き直しと絵の大きさ、割付の変更。	P87Aに同じ。
			例題の問題文も記述。
3	質問のみ記述。	質問と答の選択肢も記述。	質問のみ記述。

	テープ１回。	テープ２回。	テープ２回。
4	記述された課題文は○×式。	記述された選択文は三肢選択式。	P87Aに同じ。
	テープ１回。	テープ２回。	テープ２回。
5	選択文のみ記述。	選択文のみ記述。但し6題中4題の問題文が幅に改訂。	P87Aに同じ。
	テープ１回。	テープ２回。	テープ２回。
6	質問のみ記述。	質問と答の選択文も記述。 問題用紙の図の拡大。	質問と答の選択文も記述。さらに選択文の地名等の漢字をひらがなにする。 P87Aに同じ。
	テープ２回。	テープ２回。	テープ２回。

表５　「聴解」の問題変更点

問題１と２については試験場において次のような問題が生じた。各テストとも問題用紙はＢ４版の大きさで、P87Sでは問題１と問題２は、間にほとんどスペースがなく続いていた。そのため学生は問題用紙を受け取った時から問題２の絵の方に目が行き、問題１のテープの指示が頭に入らず混乱した。次に２の絵の選択の問題では、問題毎の絵の切れ目が瞬間的には見て取れない学生が多かった。そこでP87Aでは一題毎に絵をＢ５版の中に納めるようにした。それでも問題１で生じる混乱は変わらなかったので、更にP88Sでは問題１の例題の問題文と答の文を問題用紙に記入した。しかしこれでも混乱は生じた。この版を使った他の日本語教育機関においても同じ様な状況だったようで、「聴解の最初の１番がパニックを引き起こした。絵を見てそちらに注意を取られて混乱をおこした人が多かったようである。」「学生達は１など特にあわててしまったと言っていた。」等の報告が寄せられた。聴解は学生の緊張度も高いものであるから、問題内容以前にこうした指示の与え

方にも細かい配慮が必要だということがわかる。

図８－１にみるように、P87Sでは正答率も識別度も低い値に集まっている。P87Sでは正答率が50％以下の項目は約80％である。それに対して、図８－２にあるようにPALL2-2では逆に50％以上の項目が80％を占め、また正答率が0.3以下の問題はなく、全体に正答率は決して低くない。しかし識別度は必ずしも高くなく、聴解の不安定さを示していると言える。問題１は、P87Sでは識別度がすべて0.3以下で、問題２でも最高が0.35である。図８－１の識別度がマイナスの2題は、問題１と問題２にある。PALLでは、事態はかなり改善されたが、それでも9題中3題が0.4以下の識別度である。

問題３ではP87Aは答も問題用紙に書いてあったが、問題のみ記述したテストと結果は変わらない。正答率ではむしろ P87A より、答が記述されていなかった P88S の方が上回っている。P88S で最も識別度の高かった問題のひとつはこの中にあった。

問題４と５では会話文を聞いて、会話の登場人物の人間関係や話し手の意図・感情を推測させることを作成者は狙った。識別度は高くない。P87S では、成績下位者のうち35％以上が、問題４と５のすべての問題において無答で、問題がむずかしすぎたことがわかる。問題文に改良が加えられたP87A以降も、識別度はさほど高くなっていない。図８－２で識別度が０にプロットされている問題は、正解が二つあると考えられる。これは別としても、正答率が高くも低くもないのに識別度の悪い問題がいくつかある。たとえば次にあげる問題がその一つである。

テープ　W：ねえ、太郎のことなんだけど。最近帰りが遅いのよ。心配だわ。
M：そうだね・・・・・・。うん、・・・・・・。
W：ねっ。聞いてんの。
M：太郎のことだろう。ちょっと、静かにしてくれよテレビが聞こえないじゃないか。ほっといても大丈夫だよ。
W：いつも、こうなんだから。だから、あの子が、うちが面白くないっていってるのよ。
M：分かったよ。聞きゃいいんだろ。

質問（テスト用紙に記述される）

1　女の人の話を、男の人はよく聞いています。
2　男の人は、女の人が言っていることをほとんど聞いていません。
3　男の人は、女の人に話を聞いてほしいと思っています。
4　男の人と女の人は、太郎のことを話し合っています。

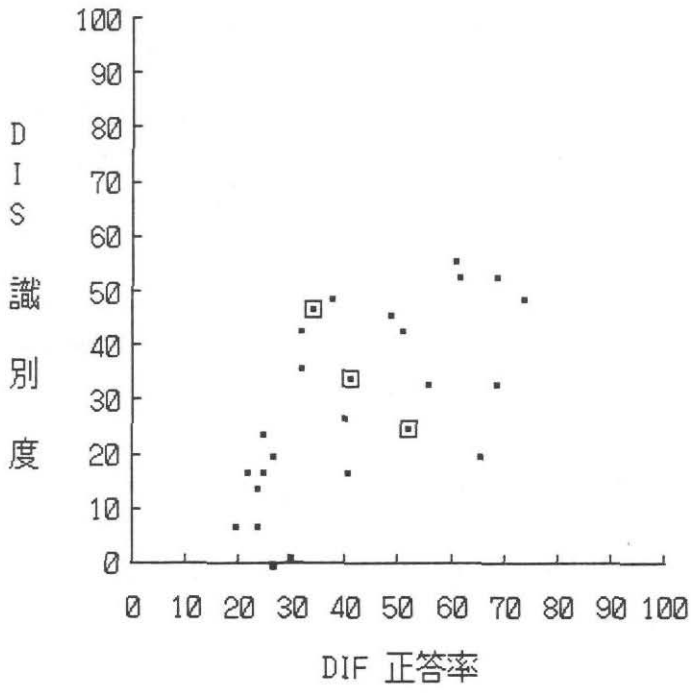

図８－１　P87S聴解におけるＤＩＦとＤＩＳの関係

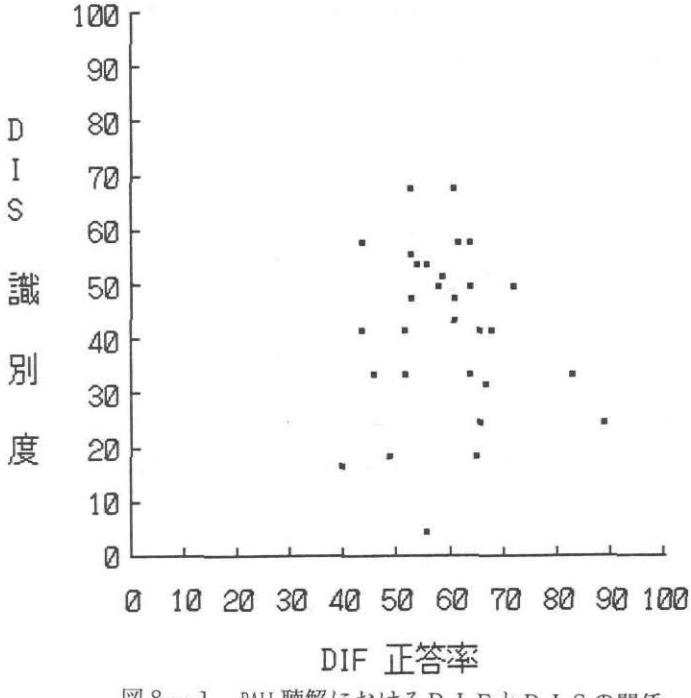

図８－１　PALL聴解におけるＤＩＦとＤＩＳの関係

表６は、PALL2-2における上位群と下位群の各選択肢の選択状況である。N.A.は無答の意。

選択肢／群	1	2	3	4	N.A.	計
上位群	0	33	0	16	0	49
下位群	2	17	6	19	5	49

表６　各選択肢の選択状況　　（単位：人数）

この問題の正解は２だが、上位群下位群ともに４を選択するものがかなりいて、それが識別度を下げていることがわかる。４を選択した理由としては、まず第一に、会話のテーマという点ではこれが正解でもあること（その点でこの選択肢は他の選択肢と内容が異なっている）、第二に、会話の一方が聞く気をほとんど持たないにしても、それを話し合いと取るか否かは主観的な問題と考えられることである。聴解力を会話の内容ではなく、機能的に聞く能力でみようという試みは興味深いが、この問題は、選択肢の作成の段階で識別度を下げる原因を作ってしまった。

P87S で最も識別度が高く、また PALL2-2 で識別度の最も高かった問題2題の内1題は、どちらも問題６の同じ問題で、桜前線についてのかなり長い会話を聞かせた上で、会話のトピックを問うものであった。

【読解】

図を見て明らかなように、P87S は全ての下位テストの中で最も分布が小さくまとまり、しかも正答率は全体の３分の２が、0.6か0.7代である。これは問題が二者択一であったためである。しかし下位群には無答の者がかなりいるので、○×式だからといって、学生がでたらめに解答していたとは言えない。しかしでたらめに反応したとしても５０％の正答率は得られるわけであるから、この解答形式はクラス分けテストには向かないと言える。そこで P87A では急遽問題を全面的に差し替え、四肢選択形式にした。ただし準備の時間が限られていたので、ほとんどの問題は既成の問題集より取り、しかも前半は１文から３文までの短文でその内容を問い、後半の10問を５文以上の読解にした。その結果、PALL2-2では、比較的三角形に近い分布が得られた。

P87Aで識別度がマイナスになった問題は次の問題である。

「東京に台風が来るとは聞いていたが、そんなに被害が大きくなるとは思わなかった。」

１　この人は東京に住んでいるが、被害の大きいことを知らなかった。

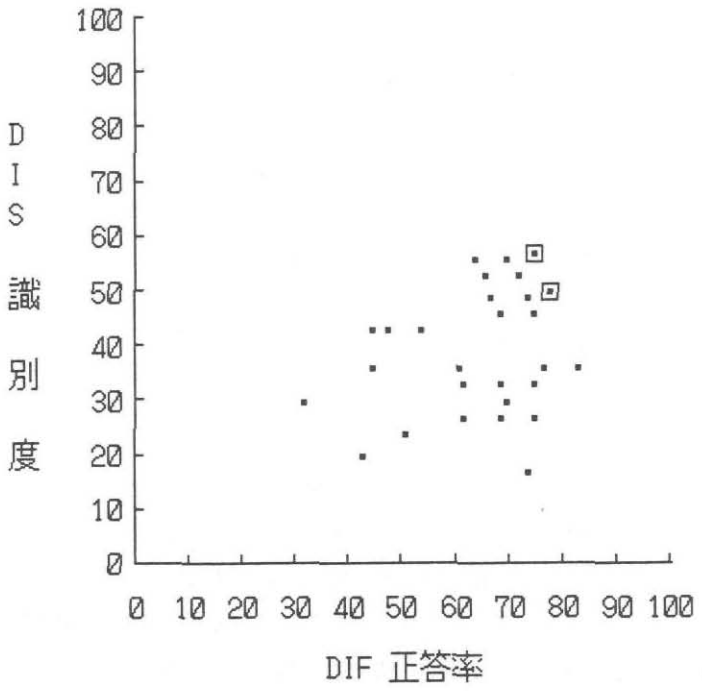

図９－１　P87S読解におけるＤＩＦとＤＩＳの関係

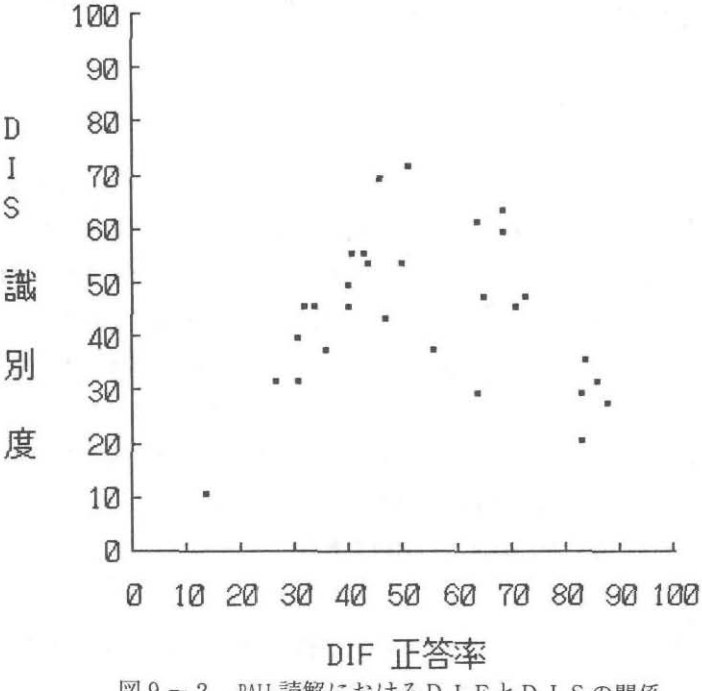

図９－２　PALL読解におけるＤＩＦとＤＩＳの関係

2　この人は東京に住んでいないので、被害の大きいことを知らなかった。
3　この人は東京に住んでいるので、被害の大きいことを知っている。
4　この人は東京に住んでいないのに、被害の大きいことを知っている。
(DIF 0.31　DIS－0.03)

この問題は指示詞の「そんなに」によって話者は東京にいあわせなかったことがわかるかをみるものであった。しかも台風の来る時点では、被害の大きくなることを予想していなかったのであるから正解を2とした。学生全体の反応は　1 28人（30%）　2 32人（34%）　3 17人（18%）　4 13人（14%）　5 無答　4人（4%）で 2 についで 1 を選択した者が多い。これは学生が「そんなに」を英語の”such”のように程度を強調する副詞と取っていたためと考えられる。この項目は P88S では他の問題に差し替えた。また識別度が P87A でマイナス、PALL2-2 でも最も低かった問題は次のようなものである。

「あの人は来るものとばかり思っていた。」
1 あの人は来たばかりだ。
2 あの人は来ない。
3 あの人は来ないはずだ。
4 あの人は来るつもりらしい。

選択肢の4を選ぶ学生が多い。

以上の二問を見てもわかるように、これらの問題は読解とはいいがたい。前半短文を入れたのは30問すべてをいわゆる読解問題にすると学生の負担が大きくなると考えた結果の妥協策であった。ただ実際には最も識別度の高い問題は P87A、P88S とも長文の問題であった。値は PALL2-2 のものである。

「入口のレジのところに、公衆電話があった。ちょうどポケットの中に、十円玉があったので、高木さんに電話をかけようと思った。ところが、彼の電話番号を書いておいた手帳がない。手帳は、部屋のテーブルの上だということに気付いた。仕方がないので、コーヒーを飲みながら、３０分以上も彼が来るのを待っていた。」

この人はどうして高木さんに電話をかけようとしたのか。最も適当なものをひとつ選びなさい。
1　公衆電話があったから。
2　十円玉を持っていたから。
3　部屋に手帳を忘れたから。
4　約束の時間より早く来てしまったから。　(DIF 0.50　DIS 0.71)

問題文が素直で、文法的にもそれほどむずかしくない。それにもかかわらず識別度が高いというのは、一文レベルではないいわゆる読解問題でもいい問題が作れるということであろう。
後半の長文問題には無答が多い。特に終わりになるにしたがって増えており、PALL2-2 の最後の2題の無答率は75％である。時間がなくなって手が付けられなくなったか、むずかしくて手を付けなかったためと考えられる。

【文法】

全体に正答率が左に偏っている。すなわちむずかしい問題が多い。このことは後半問題が進むにつれて無答率が高くなっていくことにもうかがわれる。特にP87Aでは70％の問題が正答率50％以下である。識別度が高かった問題は、「やりもらい」、「とともに」「に関する」等の後置詞の類であった。P87S、PALLを通して最も識別度の高かった問題は次の「やりもらい」の問題である。

「きょうは母の日なので私は母に「今日は私に夕食を＿＿＿＿＿。」と言いました。」
1　作ってもらいます　　2　作ってください
3　作ってあげます　　4　作らせてください

成績下位者のほとんどが選択肢の3を選んだ。逆にPALLで識別度がマイナスになった問題は次の問題で、4より3を選んだ学生の方が多かった。

「＿＿＿＿＿＿＿＿、１０円玉持っていない。」
1　電話をかけたいから　　2　電話を書けたいんだから
3　電話をかけたいけど　　4　電話をかけたいんだけど

しかしこの問題は、正答率はPALLで0.43、P87Sで0.40と特に低いわけではない。文法の項目内容としてもむずかしすぎるとは決して言えないと思うが、実際に日本語を使った経験が少なく、本だけで勉強してきた学習者には抜け落ちている項目かもしれない。問題がむずかしくもやさしくもないのに識別度が低い問題はこの他に1題あった。それは「コソアド」に関するものなので、母語の影響が表れたためと考えられる。図１０－１、１０－２で図の右にプロットされている問題は、正答率が高いために識別度が落ちているわけだが、次の問題である。

「兄はいつも学校のプール＿＿泳ぎます。」
1　へ　　2　が　　3　に　　4　で

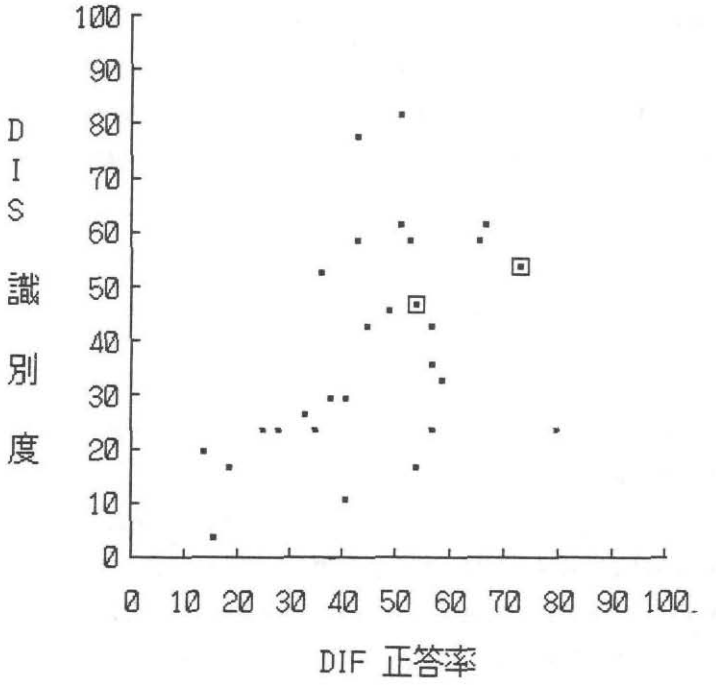

図10－1　P87S文法におけるDIFとDISの関係

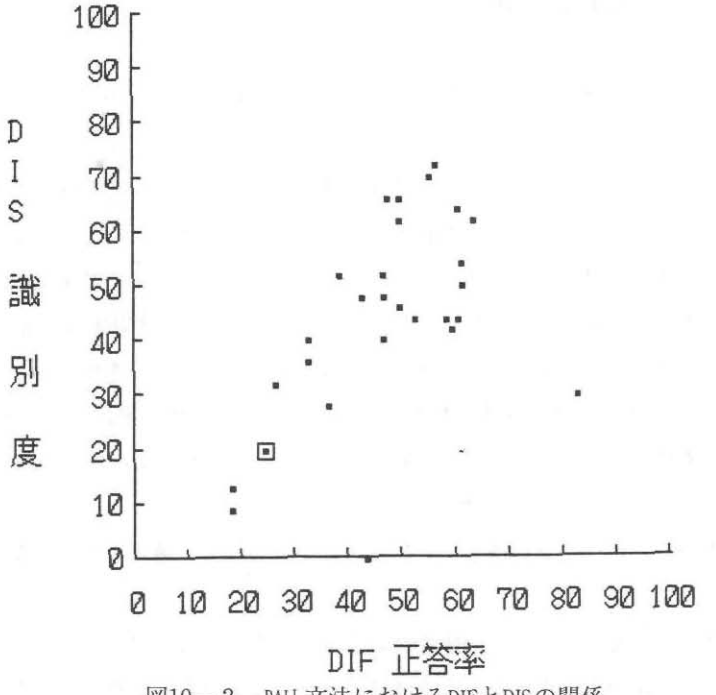

図10－2　PALL文法におけるDIFとDISの関係

【文字】

問題が上に高い三角形の辺上に均等に分布しており、理想的な分布と言ってもいいだろう。積み上げ式の要素が濃い「文字」だからここまで出来たという面は否定できないが、今後ともこうしたテスト作りを目指したいものである。一版同様今回も全てのテスト問題の中で識別度の最も高い問題は「文字」の問題であった。それはP87S、PALLとも次の二問で、いずれも漢字の読み（下線部分）である。

与えられた仕事をする。

水中から姿を現した。

漢字の読みと書きとでは、書きの方が読みより正答率のばらつきが大きかった。最も正答率の低かった問題（図１１－２の右下方にプロットされている）は、「なげく」の書きで、日本では一般に使われない漢字を書いた受験者が多かった。

今回「文字」には、問題形式に従来の漢字の読み書きの他にいくつかの新しい試みがみられた。まず熟語の切れ目を見つける問題は、全体にやさしく識別度が上がらない。特に先に「テストの内容」のところで例を挙げた天気予報の問題は3回ともDISの値が約0.1であった。またP87Sでは一題正答が二つあるとも考えられる問題があったため、P87A以降他の問題に差し替えた。しかし識別度にはほとんど変化は見られなかった。

次に漢字の部首を見つける問題では、先の熟語の切れ目を見つける問題以上に識別度が落ちた。P87Sでマイナスとなった問題は次の問題である。

集　李　校　茶　→　　(DIS　0.26　DIF　－0.19)

これは問題自体がやさしすぎることと、正答は共通の部分がないということでばつとするべきところを「茶」の下の部分を木ととって、木を部首として挙げる誤答が多かったことによる。そこでP87A以降は「茶」を「盃」に替えて、その点であいまいさをなくした。その結果PALLでは　(DIS　0.55　DIF　0.26）と識別度が多少あがった。部首を見つける問題は、いずれも正答率が高いために識別度が低く、プレースメント・テストの問題としては不適当と言わざるを得ない。

今回「文字」の中には3題カタカナを書く問題があった。これも一版の単にカタカナを書く問題とは異なったねらいを持っていた。P87Sでは問題に次のような指示が与えられていた。

「次の文の中から、かたかなで書くことばをみっつ見つけて、適当なかたかなに直しなさい。

あきはばらのでんきやでやすいすてれおをみつけた。へっどほんがついてななまんごせんえんだった。　ふつうのみせでかうよりにじゅうごぱあせんともやすかった。後省略　」

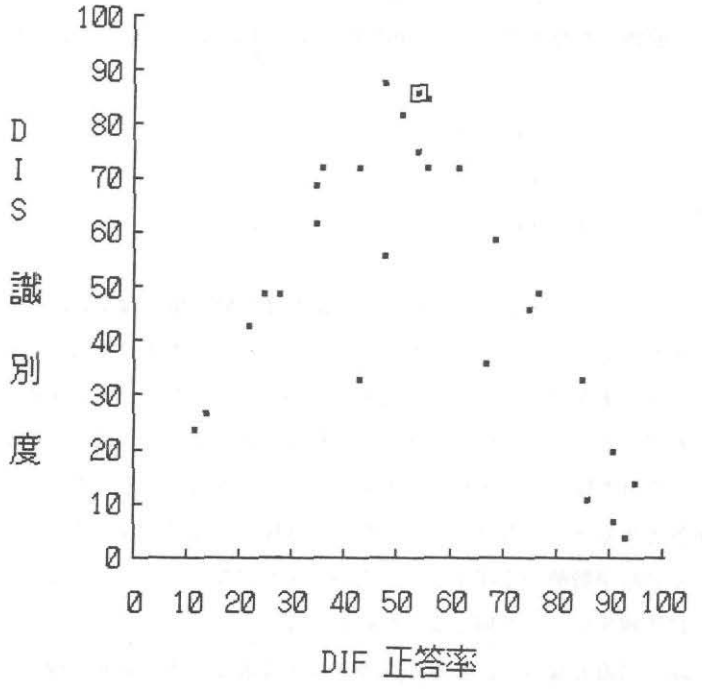

図11－1　P87S文法におけるDIFとDISの関係

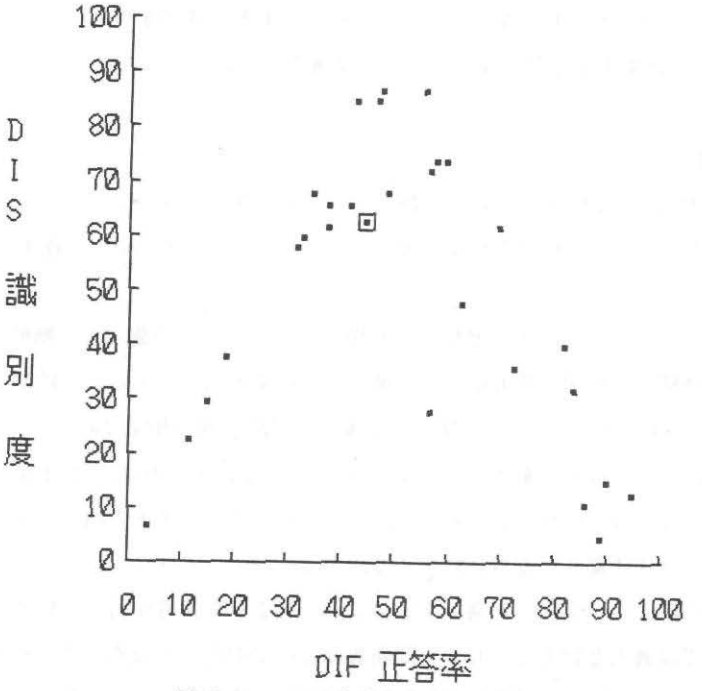

図11－2　PALL文法におけるDIFとDISの関係

この問題のねらいは、カタカナを書く能力を見ると同時に、日本語の中でどの言葉をカタカナにすべきかという正書法の問題をも含ませている。P87Sの結果は次のようなものであった。

ステレオ	(DIF 0.68　DIS 0.58)
ヘッドホン	(DIF 0.47　DIS 0.55)
パーセント	(DIF 0.42　DIS 0.32)

この項目分析の結果を見ると、「パーセント」を除けば識別度は低い値ではない。しかし採点上は非常に問題があった。それは受験者が冒頭の「あきはばら」からカタカナにすることが多かったためである。カタカナで書く言葉を見つけてという指示がわからなかったことと、実際に「あきはばら」がカタカナで書かれていることも多いためと考えられる。例えば「あきはばら」をカタカナで書き、ステレオも正しく書けた場合、ステレオは本来ヘッドホンが書かれるべき箇所に書かれてしまうが、ステレオを正答とすると「あきはばら」の間違いは採点上は「ヘッドホン」の誤答として処理することになる。その結果数値の意味するところが不明瞭になった。P87Aでは指示文中の「かたかな」と「みっつ」に下線を引いたが同じような結果になった。そこでP88Sでは「次の文中の＿＿＿のことばをかたかなに直しなさい。」としてカタカナで書くべき言葉を明確に示し、かつ「ぱあせんと」については、長母音を書き間違えることのないように、％という記号で表した。以上のことは識別度の問題とは関係がないけれども、問題作成に当たって注意するべきことであった。一版では「レストラン」「レポート」等のカタカナがやさしすぎて識別度が落ちたが、今回のレベルのカタカナは、P88Sの結果を見る限り難易度の点では適当な問題であった。

7.2　相関[4]と信頼性[8]

表7は、P87S及びPALL2-2の下位テストのα係数、下位テスト間および下位テストと総得点間の相関係数を示したものである。比較のため第一版のPALL（3回分の合計、PALL1と称する）も載せる。

今回のPALL2-2をみると、「文字」と「聴解」の相関が低く、次に「語彙」と「聴解」の相関が低い。「文字」と「聴解」の相関が最も低い点は第一版の結果も同じである。「聴解」は総点との相関が常に低く、P87Sは特に低い。一方「文字」「語彙」「読解」間の相関は高い。「語彙」と「総点」との相関は、個別にみれば常に最も高いというわけではないが、PALL2-2では五つの下位テストの中で最も高い。もしテストの時間を短縮するために五つの下位テストの内一つを除くとすれば、総点との相関が最も高い「語彙」が妥当ということになる。

α係数は「聴解」が低く、「文字」が高い。全体に第一版より信頼性が落ちているが、0.8以上あるのでテストとしては適当と言える。しかし問題数はこの30問辺りが限界と考えられる。なお、「読解」は、P87Sでは解答方法に問題があったものの信頼性は落ちていない。ここでいう信頼性、

	WOR	LIS	RED	GRA	LET
WOR	(0.80)				
LIS	0.40	(0.79)			
RED	0.71	0.53	(0.85)		
GRA	0.75	0.52	0.66	(0.79)	
LET	0.78	0.47	0.69	0.78	(0.90)
TOTAL	0.86	0.69	0.86	0.88	0.90

N=107

表７－１　P87S下位テストと総得点間の相関と信頼性係数
（　）内はα係数

	WOR	LIS	RED	GRA	LET
WOR	(0.88)				
LIS	0.67	(0.84)			
RED	0.81	0.72	(0.86)		
GRA	0.78	0.71	0.78	(0.81)	
LET	0.80	0.62	0.75	0.75	(0.91)
TOTAL	0.92	0.83	0.91	0.90	0.90

N＝187

表７－２　第二版PALL2-2の下位テストと総得点間の相関と信頼性係数
（　）内はα係数

	WOR	LIS	RED	GRA	LET
WOR	(0.93)				
LIS	0.73	(0.89)			
RED	0.76	0.74	(0.96)		
GRA	0.84	0.72	0.72	(0.92)	
LET	0.86	0.68	0.73	0.77	(0.97)
TOTAL	0.93	0.85	0.89	0.90	0.91

N＝272

表７－３　第一版PALL1の下位テストと総得点間の相関と信頼性係数
（　）内はα係数

α係数というのはテスト内部の同質性を見るものであるため、○×式によって正答率が同じような値になりプレースメント・テストとしては問題があるにもかかわらず、その点で信頼性が高くなる結果となった。

8．テストの妥当性

8.1 内容的妥当性（content validity）

Carmines & Zeller(1979) によれば、内容的妥当性は「基本的には経験的な測定が内容の特殊な領域をどの程度反映しているかに依存している」と言う。[9] プレースメント・テストの場合には、テストの問題が本来あるべきテストの範囲をどの程度網羅しているかどうかが評価されるべき点になろう。この際二つの問題が考えられる。一つは、現行の五つの下位テスト数が適当であるかどうかということ、もう一つは五つの下位テストそれぞれにおいて項目内容が適当であるかどうかという点である。前者については、広く日本語能力の測定という面から考えると、話すことと書くことのテストが欠けている点が指摘できる。対面テストにかかる時間、一貫した採点基準の維持、また記述式テストの採点の手間を考え導入をためらっている。しかし、記述式テストについては、採点基準の設定を入念に行えば不可能ではないので今後検討したい。項目内容の適切さについては「聴

解」と「読解」が特に問題になろう。これからの能力は、「漢字」「語彙」「文法」ほど積み上げ式の能力でないため範囲の設定が困難である。たとえば「聴解」の場合、一版にはミニマルペアーの聞き取り問題があったが、二版では音の聞き分けだけでは聴解能力を示していないということで聞き取るべき単語は会話文の中に文脈をもって提示され、選択肢の単語もすべて有意味語になった。またテープの内容も、学生が出合うであろう場面を想定してニュースや講義調のものを取り入れるなど工夫がされている。「語解」については限られた文章で読解能力を測定することの是非、すなわち主題が既知であるか否か、語彙が既習であるか否か等に常に問題にされるところである。ただ5章の項目分析のところで取り上げた電話の例のように、やさしい素直な文章でも識別度の高い問題は作れるわけで、こうした問題作りを今後目指す必要がある。また、「文法」は積み上げ式の能力であると述べたが、確かに初級の日本語文法には単純な表現から複雑な表現へという段階が一応設定されているように見受けられる。しかし、問題作成者が受験者にとってむずかしいだろうと考えることと、受験者が実際にむずかしいと思うこととが一致していないのはプレースメント・テストの結果にも現れている。また終助詞の「ね」と「よ」の使い分けに係わる問題は、今回識別度が高かった。しかし終助詞の使い分けは従来の日本語教育ではどちらかと言えばなおざりにされてきた部分である。こうしたことは網羅すべき文法の範囲というものが既成の教科書の文法の範囲にのみ求めるべきでないことを示唆しているように思われる。問題作成者には、常に柔軟な頭で日本語を見直しその構造をより深く理解することと、それをよりよい項目の作成につなげる技術が求められていると言えそうである。

ところでクラス分けテストというものは、そもそもはセンターの教育の到達目標というものが立てられ、その目標に向けて用意された複数のクラスに学生を分けるためのものであろう。その意味でテストの内容的妥当性を高めるためには、センターの到達目標がまず設定される必要がある。基本的な目標は、大学院での研究活動に支障のない日本語能力と言える。それは大きく「読む」「書く」「聞く」「話す」能力に分けられる。表8はそれぞれを五つのレベルに分けた到達目標基準の試案である。学生にとっては専門の研究が支障なく行われることがより重要と考えられるので、中上級のレベルでは、さらに一般と専門に目標の設定を分けた。レベル3は、 500時間程度学習した学習者の日本語レベルを想定している。レベル5が最終目標であるこの基準を現実に使えるものにするためには、それぞれのレベルの能力について更に詳細に内容網目を検討しなければならない。現在日本語能力試験が施行され、そこに日本語学習の一つの到達目標の姿を見ることができるが、筑波には筑波の学生に合った到達目標が必要であり、今後こうした基準の検討と、それに立脚したクラス運営というものが考えられていく必要がある。

レベル / 能力	1	2	3		4		5	
			一　般	専　門	一　般	専　門	一　般	専　門
読むこと	ひらがなの読み	漢字200字位 メニュー・駅名等自分に必要な情報がとれる	新聞の投書・コラム欄，大衆小説が読める	論文要旨が読める（要する時間は問わない）	説明書から必要な情報がとれる新聞の報道記事が読める	学会の論文集から自分に必要な論文が選択できる	新聞の社説が読める	論文・専門書が読める
書くこと	ひらがなの書き	季節のあいさつ状	礼状	不正確さは残るが意味はとれる程度に専門について説明が可能	新聞の投書が書ける	ゼミのレジメが書ける	一般的な分野で翻訳が可能	レポート・論文が書ける
聞くこと	答が明瞭であればわかる	コントロールされた話し方での日常レベルの会話がわかる	話し方が明瞭かつ早くなく，話題が既知の場合理解が可能	対人レベルで指導教官の話やゼミでの話が理解できる	コントロールされた話し方でなくても日常生活レベルであれば理解可能	自分の興味・関心のあるトッピクであれば講義・ゼミの聴解が可能	ラジオ・テレビニュース，現代のテレビドラマがわかる	講義・ゼミの討議が理解できる
話すこと	あいさつ ごく簡単な質問文	感情表現等の主観の表明	文法的に間違いはあるが意志の伝達は可能	対人レベルで専門について内容が伝えられる	一般的な話題では文法的な間違いがほとんどなく，また間違えても訂正が可能	対人レベルで質疑応答が可能	日本人と全ての話題において適当な待遇表現を用いて意見が交わせる	学会発表が可能

表８　留学生教育センター到達目標基準試案

8.2 基準関連妥当性 (criterion-related validity)

予測的妥当性とも呼ばれる基準関連妥当性は、何らかの外在基準との関連によって示される。プレースメント・テストの外在基準の一つとしてクラスでの成績が考えられる。ほとんどの教師はコース中にクイズ・テスト等を行っており、またコース終了後各教師にアンケートによって学生の評価、コメントを記入してもらうことが続けられ記録として保存されている。ただ終了時まで出席する学生の数が非常に少ないこと、各科目によって出席している学生が異なること、そして評価基準が教師によって異なることから、プレースメント・テストの結果とクラス内での成績の相関を見ることはほとんど不可能に近い。酒井（1986）[10]の報告は対象を何人かの学生に限定したケーススタディであるが貴重な資料と言える。

そこで主観的ではあるが、プレースメント・テストによるクラス分けが授業の運営にあたって支障がなかったかどうかと言う点から妥当性を考えることにする。

まず第二版のテストの結果によってどのようにクラス分けが行われたかを次に見る。

テスト / クラス	P87S		P87A		P88S	
総数	107		94		93	
	人数	点の範囲	人数	点の範囲	人数	点の範囲
D	46	86～130	35	87～130	31	99～130
C	33	63～85	36	55～86	24	78～98
B	21	60以下	21	35～54	20	60～77
T	6		1		18	
A			1			

表9　第二版によるクラス分け

表中のA～Dというのはクラス名で、Aは国費の研究留学生のためのクラスである。通常は予算上別枠のためと、未習者がほとんどのためプレースメント・テストの対象にはならない。[11] B～D

クラスは、大ざっぱにそれぞれ初中級、中級、中上級に該当する。Tは教員研修生のためのクラスで、これだけが1年のコースである。表7のクラスの横の欄にある初めの数字は、そのクラスに分けられた学生の人数で、次の数字はプレースメント・テストにおける総点の範囲を示している。表にみるように、コースによってクラスの人数、点の範囲に差がある。これは第一に、受験者数は毎回異なるが、実質的にBCDの3クラスしかないこと、第二に、初級にはできるだけ人数を少なくし、その分中上級には学生が多くなってもやむを得ないという判断がクラス分けに際して働くためである。さらにB以下の得点の受験者については、同じテキストや同一クラスを再度繰り返すことのないように配慮するため、必ずしもテストの得点によらないクラス分けがなされる。

ここで問題になるのは、このテストの総点による上記のクラス分けが授業の運営にあたり少なくとも教師にとって適当だったかどうかということである。しかしこの問題に答を出すのはむずかしい。というのは、現実にはコース開始時に教師と学生の間で話し合いによって、クラス間の学生の移動が行われ、クラス分けの不適切な点、学生の不満はその時点で解消されるからである。点数によるクラス分けにこうした人的配慮が加わることは必要なことであるし、そうしたものが加味されれば、現在のプレースメント・テストは一応満足すべきものと言える。ただコース開始時に学生、教師双方から指摘される問題については、検討しておく必要がある。それは大きく次の三つに分けられる。

一つは、上のクラスに変わりたいという学生の希望である。大学院の入試を控える学生が多いだけに、あせりの気持ちがあるのだろう。この点については教師は経験的に能力に見合ったクラスで授業を受けるのが能力を伸ばすことにつながると感じているので、学生がクラスの境界点に近ければ許可するが、そうでなければテストの結果をもとに決められたクラスで授業を受けるように説得する。二つ目の問題も学生側からのもので、自分は専門については語彙も豊富で、漢字もよく知っているのに、センターのテストの内容は一般的で、学生にとって必要な能力による評価がなされていないというものである。これについては、テストの性格上個別的な問題にすることはできない。ただ実際の授業内容を考える上では、考慮するべき面を含んでいる。三つ目は、教師の側からおもに出てくるもので、下位テストは五つあるが総点をもとにクラス分けを行うため、学生の能力のばらつきがクラス分けに考慮されていないという点である。筆者としては、現行のプレースメント・テストに改善すべき問題点があることは否定しないが、それ以前にこれはセンターのコース運営上の問題が大きいと思われる。中上級でもひとクラス30～40名という人数は、良心的な語学授業の出来る範囲を超えている。まして初級で20名という数では、各学生の弱点に応じて伸ばす教育をするのはむずかしい。この点での改善がまず何にもまして必要と思われる。

8.3 構成概念妥当性 (construct validity)

Carmines & Zeller(1979) によれば、「基本的には、構成概念妥当性は、その測定と、測定される諸概念（または構成概念）に関して理論的に導かれる仮説に合致する他の測定とがどの程度関係

しているか、ということに関している」と言う。[9] ここでは日本語能力というものの構成概念が問題になりその理論が必要とされている。先に内容的妥当性のところでテスト範囲の設定の問題を取り上げたが、それと同時に「聴解」「読解」というのがそもそもどのような能力なのか、日本語能力はどのような構造を持つものなのかを知ることがよいテスト作りに不可欠と言える。今回の「聴解」の会話を聞いて人と人の関係やどんな気持ちで話しているかを考えさせる問題には、話された言葉を理解するだけでなく、話し手の意図まで理解できることが「聴解」には含まれるべきだという作成者の考えがある。

Oller (1979) は、言語能力が複数のテスト要素に分けて測定できるものかどうかという疑問から、次の三つの言語能力についての仮説を提案した。

1．Divisibility Hypothesis

2．Indivisibility / Unitary Competence Hypothesis

3．Partial Divisibility Hypothesis

1の仮説は、異なった構成要素のテスト間には基本的に共通の変数はないと考えるもので、読む、書く、聞く、話すといった下位能力にテストを分けることが可能だと考える。2の仮説は、言語能力の構成は、部分に分けられる機械のような統一体であると考えるものである。この仮説では、たとえば語彙のテストで測定される語彙の知識は文法の知識と区別することはできないとする。3の仮説は、1と2の中間的なもので、ある下位能力を測定するテストは全ての語学テストに共通する中心的なもの（a general factor）と、それに加えて幾分か他の下位能力を測定するテストとは共通しない特有なもの（specific factors）を含んでいると考える。Oller (1979) は、いくつかの実際に行れたテストの結果をもとに因子分析によってこの仮説を検証している。1979年の時点では、2の仮説が多くの場合支援されたが、その後の研究によりBachman and Palmer (1982)、青木・萬谷等（1985）によれば現在では3の仮説が支持されているという。言語能力をどう考えるかによってテストの形式が変わってくることは言うまでもない。青木・萬谷等（1985）は、米国における外国語テストの歴史とその根底にある言語能力観の変貌を簡潔に解説しているが、その中で紹介されているHiggs & Cliffordの語彙、文法、発音、流暢さ、社会言語学的要因のそれぞれの言語要素が学習過程のレベルによって異なった重みで機能していると言う仮説や、Cooperの場面、相手に応じた言葉の使い分けが能力の重要な要素と考える仮説などは大変示唆に富んでいる。既成の枠に捕らわれず広い視野に立って言語能力を考えそれを問題作りにも反映させていくことが今後の課題の一つである。

9．他機関での使用状況[12]

第一版のうち、P86Aの版は、静岡大学、長崎大学と（社団法人）国際交流サービスに、第二版のP87Aの版は、先の三機関に加えて、麗澤大学、豊橋技術大学の計五機関の使用に供された。表１０

は、それらの大学における使用状況の一覧である。[13]

表１０の使用状況及びテストの結果から次のようなことが言える。

1　技術研修生を対象にする機関は、学習者の能力と目的が大学とはかなり異なり、センターのテストは参考程度の役にしか立たない。各大学でのテスト結果は、成績上位者から下位者までを一様に含んでおり筑波の難易度のテストがそのまま他の大学でも利用できると言える。

2　実際にクラス分けに利用しているのは一校だけだが、他の大学は人数の問題でクラス分けが出来ないという事情がある。たとえクラス分けに利用しなくても学習者の未習時の能力の把握という点で、どの機関もこうしたテストを必要としている。

3　国立大学の場合、その多くはひとりの専任が孤軍奮闘しているという状況のようで、テストを共用することができればその労力が節約できる。そういった意味でこうしたことがもっと組織的に行われていく必要がある。ただしその段階になったら、各大学のプライバシーの問題もあり、それはもはやセンターでするべきことではなく、独立した機関の仕事と考えられる。
（筑波大学にそのような組織を作ることも可能かもしれないが。）

4　試験の時間については、第三版でさらに短縮する予定だが、大学のひとコマの時間内に納めるのは、テストの今の信頼性を維持しようとすれば無理である。また選択式のみの問題でできていることについては、今後記述式テストの導入を考えてみたい。

１０　今後の課題

1　第二版の目的の一つは、学生にとってより負担の少ないテストを作ることであった。細部に問題は多々あるが、150問でも信頼性のあるテストができたと言える。ただテスト時間についてはさらに短縮が図られる必要がある。現在45分かかる「聴解」の改良がなされれば、十分可能なことであろう。また「読解」は、今回文法に関する問題が入ったが、次回はその点でも改良したい。なお成績の発表は今後も続けていきたい。

2　現在プレースメント・テストには二つの版ができ、三版の作成も進んでいる。こうして異なったテストが実施される場合、学生それぞれのテスト結果を共通の尺度でみるためにテストを等化する必要が生じてくる。共通問題を用意するのが適当と考えられるが今後その具体的な方策についても検討したい。

3　日本語能力の構造を明らかにすることやセンターの教育到達目標を明確に具体化し、それに沿った教育がなされることがテストの妥当性を高めることにつながる。今までどう測定するかというテストの信頼性の面に力を注ぐことが多かったが、今後何を測定するのかという妥当性の面にももっと目を向けて行きたい。また到達目標基準の設定とともにそれに準拠した筑波の能力テス

		静　岡	長　崎	国際交流サービス	麗　澤	豊　橋
受験者総数	第一版	17	12	12	0	0
	第二版	17	21		20	27
使用状況		第二版 LIS GRA 9人 WOR RED LET8人	全下位テスト	第一版LISの内5問 第二版GRA	全下位テスト	LET, LIS, GRAの3下位テストのみ
対象者		学部一年	研究生と学部生	技術研修生	学部一年	研究生・学部生・大学院生
使用目的		能力の把握 外部向け資料	クラス分け （この他に面接の結果と漢字圏・非漢字圏等の条件を加味）	クラス分け （この他にIMJのテキストをもとに聴解問題を17問作成追加，また文字が読めるか否かの条件を加味）	能力の把握 （異なった入試を受けているため）	能力の把握
今後の使用希望		肯	肯	肯否	肯否	肯
上の理由		専任はひとりで作成の余裕がない	静岡に同じ。 またひとりでバリエーションをつけた問題作成は困難	日系人が多いため文字を読む能力が反映するテストでないと使えない。	本来は自前のテストが望ましい。今回は余裕がなかった。	静岡に同じ。
コメント		下位テストを二つに分け異なる学生に実施したのは，ひとコマ内にテストが収まらないためふたコマに分け，学生には両方に出席するよう指示したにもかかわらず，学生側の都合でそれが出来なかったため。	1. もっとコンパクトなテストにならないか。長時間の拘束で怒って，午後の面接に出ずに帰った学生が3人いた。 2. 選択式のみでは測定できない能力がある。 3. テストの結果にカンニング等の外的条件が影響している	全体で一ケ月のコースのため，テストに長時間かけられない。		時間的条件から，三つの下位テストに限定した。

表10　他大学での使用状況

トというものも開発していきたい。

4　他の大学でも筑波のプレースメント・テストが活用されていることがわかった。今後も希望があれば積極的に利用に供したい。外部からの意見が、よりよいテスト作成のための刺激にもなっているが、こうした実質的な面での大学間の横のつながりも強化したいものである。なお、現在テスト問題はアスキー形式のファイルに保存されており、データベース化が可能な状態である。Henning（1987）は、カリフォルニア大学の外国人のためのProficiency Examination の問題のデータベース化を紹介している。そこでは使用月日、識別度、点双列相関係数等の19項目が挙げられており、我々の参考になる。

5　現在のテストは、我々の語学学習法を是とし、その尺度に学生を合わせているという面は否めない。下位テストの種類、四肢選択というテストの形式、総点によるクラス分け、それら全てに教師のテスト観や語学学習の方法論が反映されている。しかし学生がこれまで受けてきた語学教育とはどのようなものだったのだろうか。よりよい教育活動のために学習者自身の持っている枠組みにも目を向けていきたい。

注

1）読解は、当初配点の重み付けがある多肢選択の問題が作成されたが、他の下位テストの問題形式との一貫性を保つため、その問題は採用しなかった。そこで時間的な制約から○×式の問題を使用することになったという経緯がある。
2）日本語ｄＢＡＳＥ日本アシュトンテイト株式会社を使用した。
3）テスト結果の入力については大学外部に依頼したほか、筑波大学大学院の虎尾、坪山の両氏他に協力していただいた。
4）記述統計、相関の算出には、ＳＰＳＳのディスケット版を使用した。
5）ヒストグラムと項目分析の図の作成には Office Graph ver. 2.0 ＮＥＣを使用した。
6）PALL2-2は第二版のテストのうちの二つのテスト（P87AとP88S）を併せたことを意味する。多少煩雑な略称という感があるが、本論集中の酒井論文との統一を図るためである。
7）本論集　酒井たか子「プレースメント・テストの母語群別分析」参照
8）α係数の算出には三枝紀雄作成のベーシックプログラム「ＡＬＰＨＡ３」を使用した。
9）この日本語訳は参考文献２から取ったものである。
１０）酒井たか子（1986）「プレースメント・テスト複数回受験者の得点推移と習得タイプ」『日本語論集』第二号　筑波大学留学生教育センター

１１）P87Aの内一名は、もともとＡクラスの学生だったが、既習者のためテストを受け、結果的にＡが適当と判断された。

１２）静岡大学の佐々木倫子氏、長崎大学の福島邦夫氏、（社）国際交流サービスの森由紀氏、麗澤大学の戸田昌幸氏、豊橋技術大学の吉村弓子氏には、当センターのプレースメント・テストを使用され、貴重なデータをお送りいただきました。さらにテストについての御助言をいただきましたことと合わせて心よりお礼を申し上げます。

１３）使用状況については三枝が各機関の担当者に電話で話を聞いた。

参考文献

1 高見沢孟（1977）「テストと評価－米国国務省日本語研修所の場合－」『日本語教育』３２号

2 Oller, J. (1979) Language Tests at School Longman.

3 Bachman, L. F. and Palmer, A.S. (1982) "The Construct Validation of Some Components of Communicative Proficiency" TESOL QUARTERLY vol.16, No.4

4 デイビット・ハリス著 大友賢二訳注（1983）『英語の測定と評価』ＥＬＥＣ

5 E.G.Carmines & R.A.Zeller 著 水野・野嶋訳（1983）『テストの信頼性と妥当性』人間科学の統計学７ 朝倉書店

6 日本語教育学会（1984）「外国人のための日本語能力認定試験に関する調査研究の経過報告Ｖ」

7 青木・萬谷等（1985）『英語の評価論』英語教育学モノグラフ・シリーズ 大修館書店

8 三枝令子（1986）「プレースメント・テストの統計的処理の試み」『日本語論集』第二号 筑波大学留学生教育センター

9 ＡＣＴＦＵＬ試験マニュアル

10 Grant Henning(1987) A Guide to LANGUAGE TESTING NEWBURY HOUSE