プレースメント・テストの統計的処理の試み

三枝　令子

Ⅰ．はじめに

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，その結果によって学生のクラス分けを行っている。従来は，テストの採点結果をクラス分けに使用するだけであったが，1985年より新しいプレースメント・テストの準備がおこなわれ，その結果にもテストとしての良否等について検討が加えられるようになった。今後テストの改良を重ねていく際の参考として，現行のプレースメント・テストについて記録を残しておく必要があると考えられる。以下は，1985年秋より３回に渡って実施したプレースメント・テストについての施行方法と結果，並びに結果の分析である。

Ⅱ．テストの概要

1985年夏頃より現行のプレースメント・テストの準備が行われ，10月に一回目が施行された。テストは「聴解」，「文字」，「語彙」，「文法」，「読解」の五つの下位テストから成り，各50問計250問であった。86年４月のテストでは「文法」に語順の問題が10問追加され，計260問，86年10月のテストでは「聴解」に10問追加があり，計260問であった。配点は各１点で，85年は250点，86年の２回は260点満点である。以下85年10月実施分のテストをＰ85Ａ，86年４月実施分のテストをＰ86Ｓ，86年10月実施分のテストをＰ86Ａと称する。Ｐは Placement，Ａ，Ｓはそれぞれ Autumn，Spring の略である。また「聴解」，「文字」，「語彙」，「文法」，「読解」の下位テストを，時にそれぞれ LIS，LET，WOR，GRA，RED と略す。

１．出題者と出題形式

問題の作成には，主に以下の者が当たった。

ＬＩＳ　：　草薙裕，加納千恵子，池田裕
ＬＥＴ　：　戸田昌幸，山本一枝，畝田谷佳子
ＷＯＲ　：　堀口純子，田辺和子，佐藤政光
ＧＲＡ　：　寺村秀夫，野田尚史，青木直子
ＲＥＤ　：　佐久間まゆみ，大坪一夫，小口叔枝

（順不同）

解答の形式は，「文字」と「聴解」の一部を除いてすべて四肢選択で，出題の内容と配点は以下の通りである。

「聴解」

1．テープの声と同じ単語もしくは文を選ぶ。	1点×10問
2．テープの四つの会話文から最も適当なものを選ぶ。	1×5
例1ただいま。ーいってらっしゃい。	
2ただいま。ーおかえりなさい。	
3ただいま。ーどういたしまして。	
4ただいま。ーいってきます。	
3．テープを聞いて該当する絵を選ぶ。	1×5
4．会話を聞いて，その内容と一致する文を選ぶ。	1×15
5．テープの文（一文か二文）と一致する内容の文を選ぶ。	1×5
6．長文を聞いて質問に答える。	1×10

「文字」

1．ひらがなをカタカナにする。	1×5
2．漢字の読み方を書く。	1×21
3．漢字を書く。	1×24

「語彙」

1．動詞・形容詞・名詞などの穴埋め問題	1×25
例A：吉田先生の＿＿はどちらすか。	
B：土浦です。	
1．おたく　2．おうち　3．ごたく　4．うち	
2．ある単語を最も正しく使っている文を四つの中から選ぶ。	1×25

「文法」

助詞の穴埋め，活用，接続詞，副詞等の問題，問題文はほとんど一文。	1×50

「読解」

1．文（一文か二文）を読んで，その内容と一致する文を選ぶ。	1×25
2．文（三文～九文）を読んで，内容に関する質問に答える。	1×25

以上のほかに，P86Sでは「文法」に語順の問題が10問，P86Aでは「聴解」に日本語の運用能力の測定を意図して作られた問題が10問追加された。更にP86Aでは，P85A，P86Sで明らかに不適当と分析された数問を変更した。これについては，結果のところで述べる。またP86Aでは各下位テスト内の出題順序を問題の形式が変わらない範囲で，P86Sでの正答率の高い順番に並らべかえた。

2．実施方法

実施時間は説明の時間を除いて全体で2時間半である。P85Aは一教科30分の制限時間を設け，

30分経過すると問題文と回答用紙を回収した。P86Aも同じ方法をとった。P86Sでは監督者の間で連絡が行き届かず回収が徹底しなかったので，学生は自分のかけたいだけ前半の問題に時間をかけ，最後になって時間が足りなくなることになった。その結果P86Sでは5つの下位テスト中最後の「読解」の解答率が非常に落ちた。表－1に問題の提出順序と制限時間を示した。

	P85A		P86S		P86A	
1	LIS	30分	LIS	時間は全体で2時間半	LET	20分
2	LET	30分	LET		WOR	30分
3	WOR	30分	WOR		GRA	30分
4	GRA	30分	GRA		LIS	30分
5	RED	30分	RED		RED	40分

表－1　問題の提出順序と制限時間

P86Aで出題順序を変えたのは，「聴解」が最初であると遅刻者がいた場合それまでの問題にまったく手がつけられないからである。また「文字」を20分，「読解」を40分に変更したのは，「読解」が5教科中最もむずかしく，一方「文字」は考えたからできるということはあまりなく時間を余す学生が多いためである。なお「聴解」の時間は実質的にはテープを回している時間になるが，これはほぼ30分である。

3．採点基準

1)　問題はすべてひとつの答を要求しているので，ふたつ以上の答があれば無答とみなした。

2)　数字で回答すべきなのにアルファベットが書いてあったり，その逆の場合にも無答とみなした。

3)　「文字」の漢字の書きについては，採点者によって基準が異なる恐れがあるため，どこまでを正解とするか一覧表を作り，採点の揺れを小さくするよう努めた。なおP85A，P86Sでは漢字を書く問題で送りがなのまちがいは問わなかったが，P86Aでは問題文のあいまいさをなくし，送りがなも採点の対象とした。またカタカナ・ひらがなを書く問題では，正解以外の文字が書いてあれば誤答とした。漢字の書きとりは，完全解答のみが正解で，かなだけ，漢字の一部のみは無答とした。

4．採点方法

採点[1]にはパーソナルコンピュータを用いた。素データの入力には「dBASE Ⅲ」[2]を使用した。入力ミスをできる限り少なくするため，3人が別々に「dBASE Ⅲ」で素データを入力し，その後ベーシックのプログラム「COMPA」[3]によってその3人の入力データを比較し，入力の

まちがいをチェックした。更にベーシックプログラム「MLISTA」[3]「ANSINA」[3]によって，別に入力した正答パターンと突き合わせを行い，合計点を計算した。P85A，P86SではコンピュータによるP採点が試行の段階で，手作業が先行したが，P86Aではテスト実施日の翌日の夕方には，採点ミスが皆無に等しいと思われる結果が出せ，クラス分けに用いられた。

5．受験者

表－2は受験者数と国籍による構成を示している。中国には中国語を使用する点から台湾，中華人民共和国，香港を含めた。PALLは，P85A，P86S，P86Aを合わせた数である。

テスト / 母語	P85A	P86S	P86A	PALL
総数	78 (100)	95 (100)	99 (100)	272 (100)
中国	33 (42)	42 (44)	32 (32)	107 (39)
韓国	23 (29)	24 (25)	24 (24)	71 (26)
その他	22 (28)	30 (32)	43 (43)	95 (35)
国籍延べ数	14	21	28	32

表－2　受験者の構成　（単位：国籍延べ数以外は人数，（ ）内はパーセント）

P85A，P86Sでは中国・韓国・その他の比率がさほど変わっていないのに対して，P86Aでは中国語を母語とする者が減り，その他が増えていることが注目される。

Ⅲ．テストの結果

1．成績

表－3は，テスト3回分の成績を示している。P86SGRAとP86ALISにはそれぞれ10問追加問題があったが，比較のため，そのデータを省いたものも載せた。

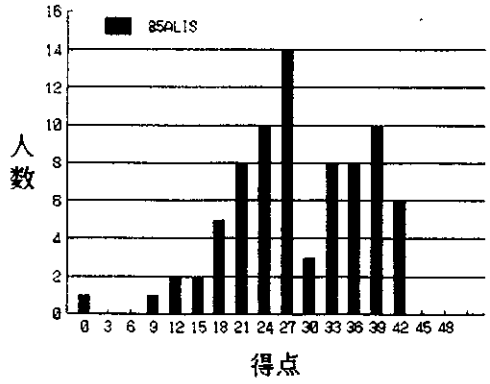
テスト / 得点		LIS	LET	WOR	GRA	RED	TOTAL
P85A	満点	50	50	50	50	50	250
	平均	29	25	27	32	25	138
	標準偏差	9	14	10	10	12	50
	最高値	44	50	47	48	48	225
	最低値	0	0	1	0	0	12
	中央値	28	24	26	33	33	131.5
P86S	満点	50	50	50	50 60	50	250 260
	平均	30	25	28	33 39	20	136 141
	標準偏差	11	16	12	12 14	14	57 60
	最高値	47	50	48	48 59	49	239 247
	最低値	4	0	0	0 0	0	7 7
	中央値	31	29	29	36 41	18	138 144
P86A	満点	50 60	50	50	50	50	250 260
	平均	31 35	22	26	31	37	137 141
	標準偏差	9 10	15	10	10	14	54 55
	最高値	46 53	48	47	47	47	230 238
	最低値	5 5	0	0	6	0	21 21
	中央値	33 36	23	27	34	29	150 152

表－3　テストの結果（単位：点）

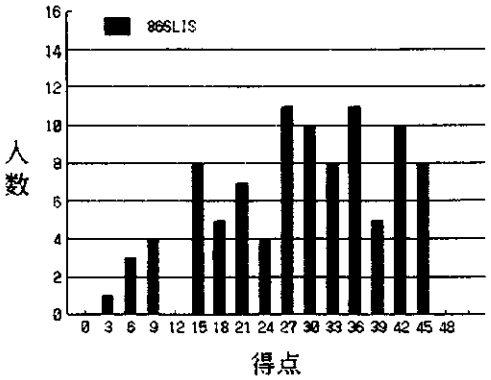
全体の平均点には3回のテストで大差は認められない。下位テストの平均点は，GRA，LIS，WORの順に下がり，それにLETとREDが続く。P86SのREDの平均点が他の2回に較べて低いのは，先に実施方法のところで述べたように時間が足りなかったためと考えられる。P86AのLETの平均点が低い理由としては，一つには受験者の国籍構成，この回は漢字圏以外の国の学生の占める割合が大きいこと，二つには採点基準の変化，この回は送りがなも正しくふれなければ正答にならないことが考えられる。標準偏差は平均点の低いLETで最も大きい。他の下位テストと異なり「文字」は答を書かせるために，書けるか書けないかの二者択一となりできる人とできない人の差がはっきり出て分布が広がると考えられる。逆にLISは得点の分布に散らばりの少ないことが注目されるが，これについては問題の項目分析のところで考える。

次に図－1～7に3回のテストの得点のヒストグラムを示す。X軸が学生の得点，Y軸が人数を表わし，X軸の目盛りは下位テストが3点おき，各回及び3回分の合計の総得点は15点おきである。

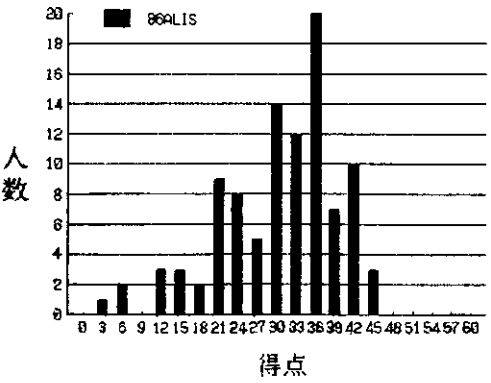
下位テストの場合図のX軸上の0点には0～2点までの学生の人数，3点には得点が3～5点までの学生の人数が目盛られている。以下同様である。なお3回のテストの比較がしやすいように追加問題の得点は含めていない。すなわち下位テストはすべて250点満点での得点である。図は Office-graph [4] によって作成した。



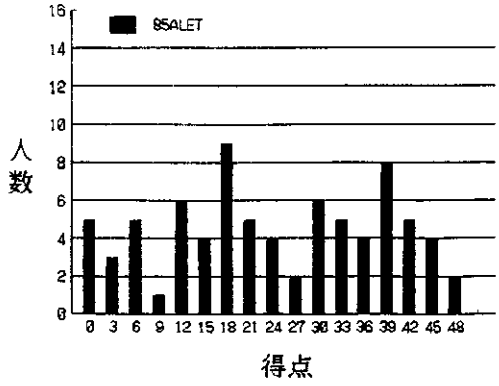
図－1－1
P85A聴解



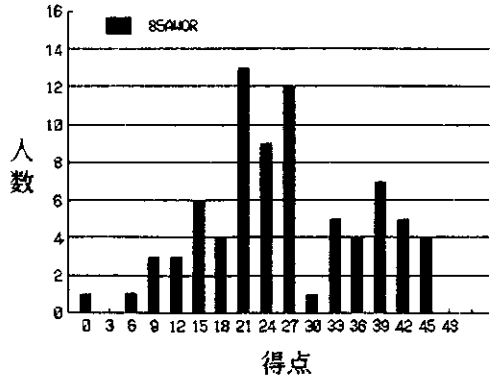
図－1－2
P86S聴解



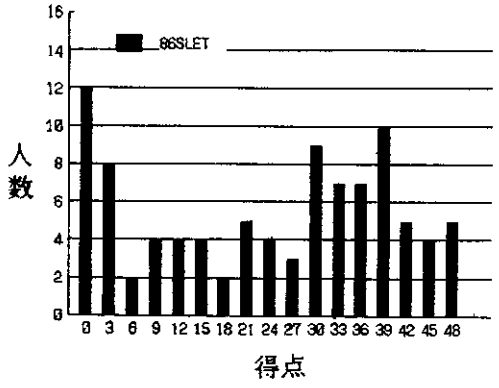
図－1－3
P86A聴解



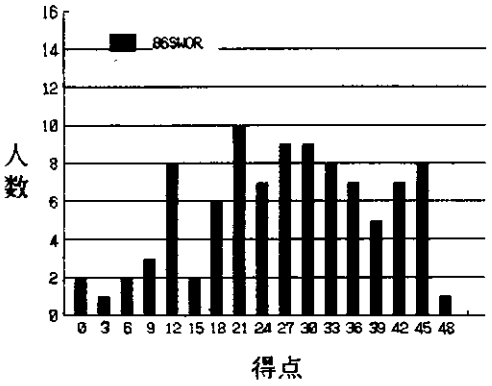
図－２－１
P85A文字



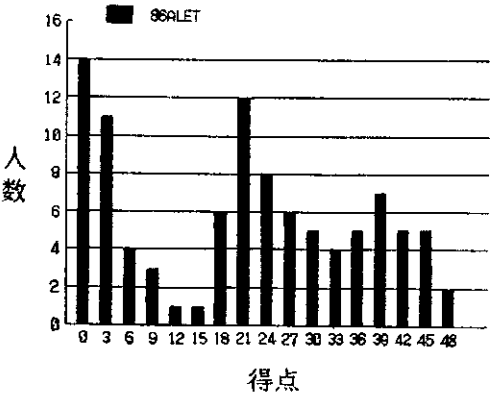
図－３－１
P85A語彙



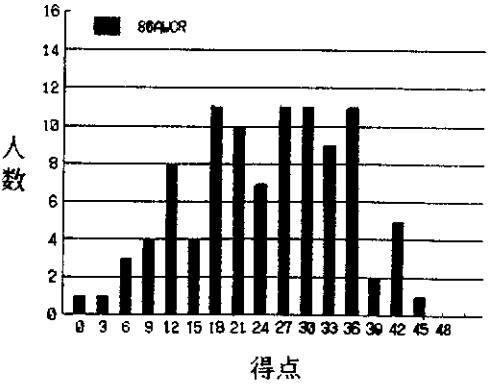
図－２－２
P86S文字



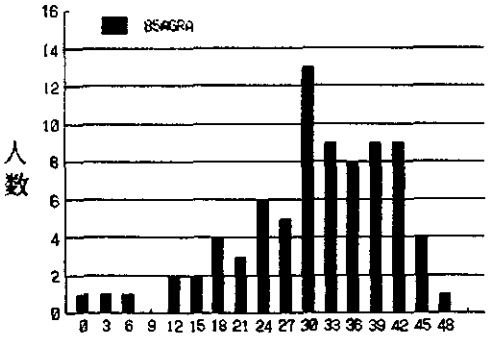
図－３－２
P86S語彙



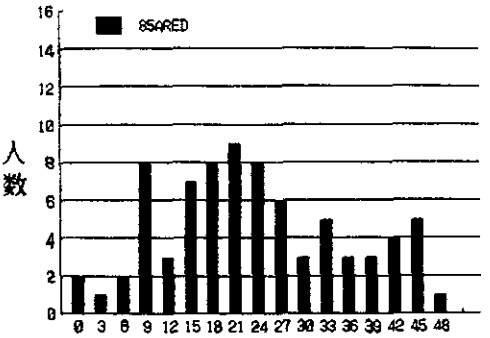
図－２－３
P86A文字



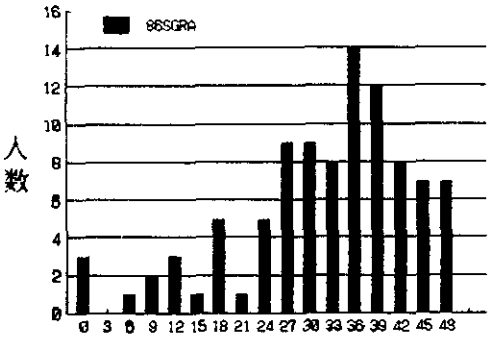
図－３－３
P86A語彙



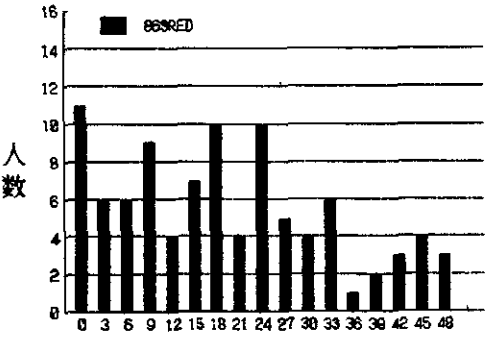
図－4－1
P85A文法



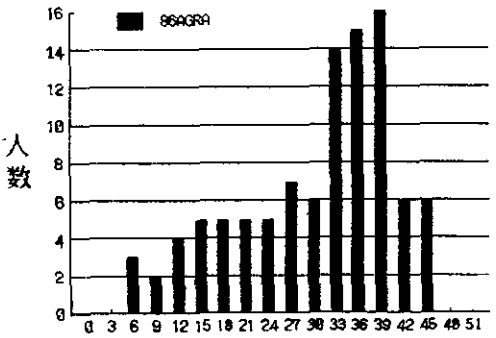
図－5－1
P85A読解



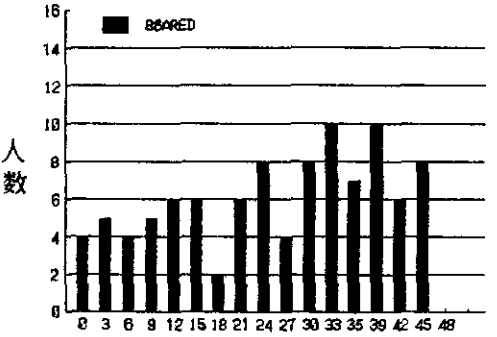
図－4－2
P86S文法



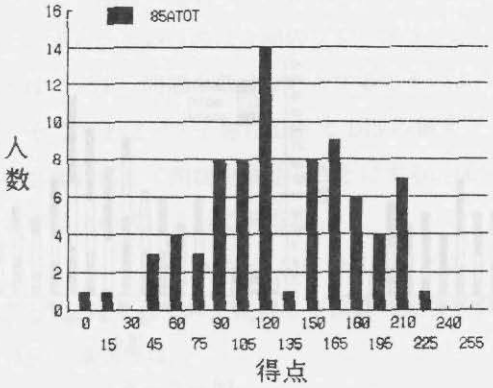
図－5－2
P86S読解



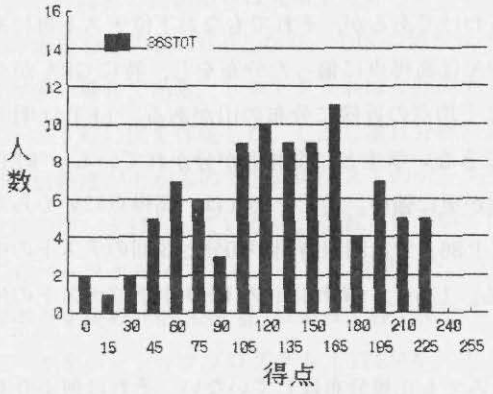
図－4－3
P86A文法



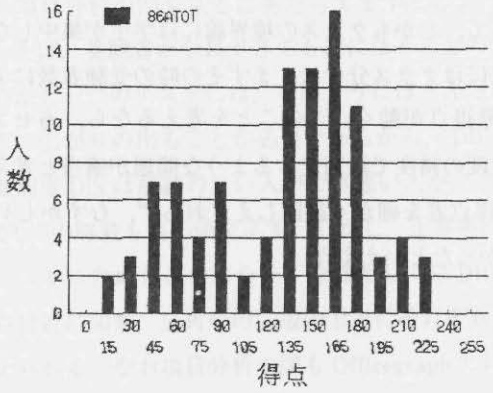
図－5－3
P86A読解



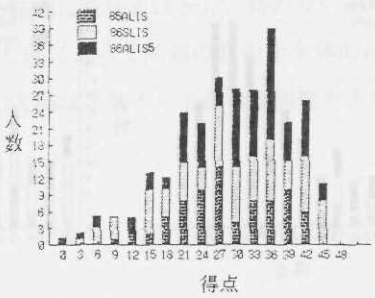
図－６－１
Ｐ85Ａ総得点



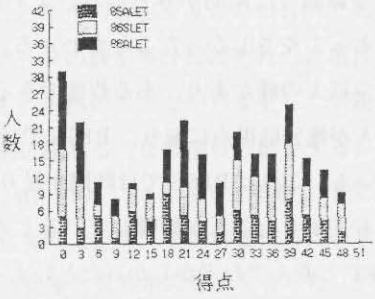
図－６－２
Ｐ86Ｓ総得点



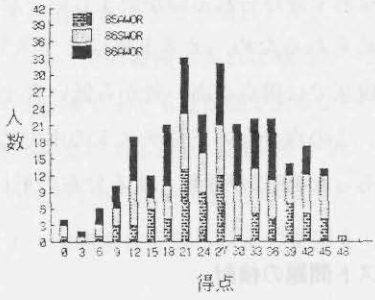
図－６－３
Ｐ86Ａ総得点



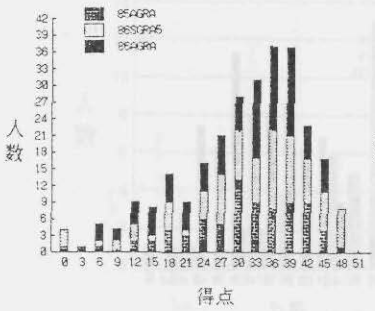
図－７－１
聴解３回分の合計



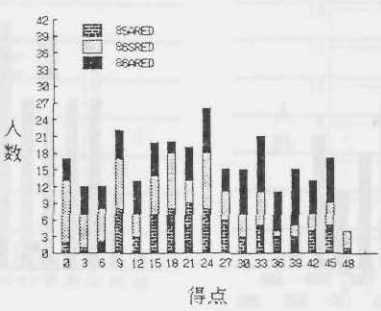
図－７－２
文字３回分の合計



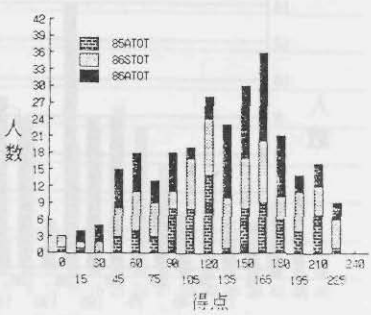
図－７－３
語彙３回分の合計



図－７－４
文法３回分の合計



図－７－５
読解３回分の合計



図－７－６
総得点３回分の合計

受験者の日本語力は各回のテストによって異なるわけであるが，それでもなお下位テスト毎にある程度の似通った傾向がみられる。まずLISとGRAは高得点に偏った分布をし，特にGRAが受験者にとってやさしかったことがわかる。WORは平均点の近傍に分布の山がある。LETは明らかに２つ以上の峰があり，ある程度できる学生とできない学生とに受験者が分かれている。REDはP85Aが幾分低得点に偏り，P86SではSその傾向が更に強い。逆にP86Aは，高得点にいくらか偏っている。これはP86Sでは時間が足りず，逆にP86Aでは制限時間が40分と３回のテストの中では最も長かったことが影響していると考えられる。しかしいずれにしろREDが下位テストの中ではLETと並んでむずかしかったといえる。

ところでこれらの分布は図－７の３回分の合計をみても正規分布はしていない。それは何よりもデータ数が少ないからであるが，ここで問題になるのは我々が理想とする分布のあり方である。現在このプレースメント・テストの結果をもとに大きく３つのクラスに学生を分けている。クラス分けのためには，学生ができれば絶対的な基準によって，しかもクラスの境界線には学生が集中しないような形で分けられるのが望ましい。しかし実際にはクラス分けは，まずその時の受験者数によって決められるため，テスト毎にクラス分けの境界得点が動く。このことを考えるなら，当センターの現状では得点が高い者から低い者までを同程度の精度で識別できるような問題が適当と考えられる。この点今回の下位テストの中でGRAは高得点者を細かく識別しえておらず，むずかしい問題であったREDが理想とする分布の形に最も近かったといえよう。

Ⅳ．テスト問題の検討

１．項目分析

P85A，P86S，P86Aの３回のテストの受験者数を合計すると272名になる。そこで全体をほぼ３分の１ずつに分けそれぞれ上位群，中位群，下位群として項目分析を行い各問題項目の良否を検討した。上位群は，169点以上の学生，下位群は116点以下の学生で各90名である。しかし各問題

についての項目の正答率（DIFと略す）と項目の識別度（DISと略す）の値はここには載せないことにする。というのは今回のプレースメント・テストの問題を今後使用する可能性がまったくないとはいえず，問題を載せないとするとそのような数字の羅列は意味をほとんど持たないからである。そこで下位テスト毎にDIFとDISの値をプロットした図によって問題の良否を全体的に検討し，必要に応じて問題を載せることにする。図－8～12のX軸は正答率，Y軸は識別度を表わす。正答率は次の式によって求める。

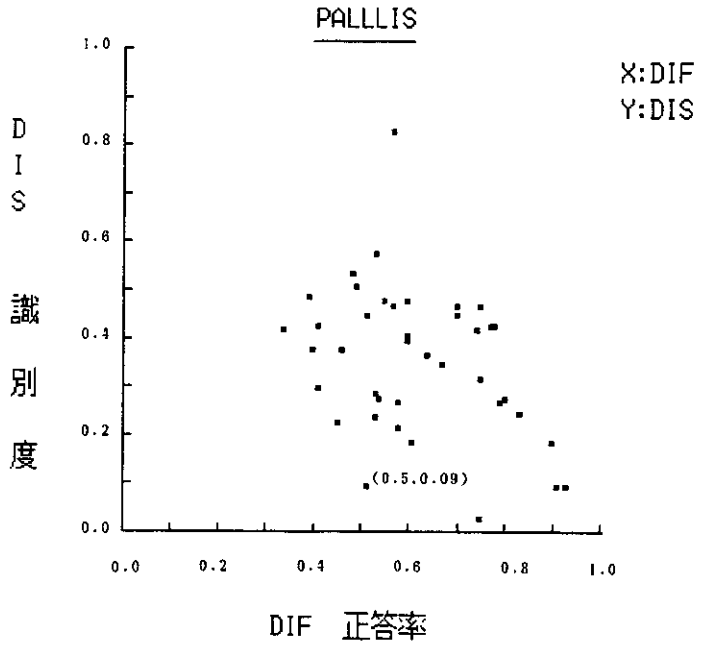
$$DIF = (HC + LC) / 2N$$

HCとLCはそれぞれ上位群と下位群の正答者数，Nは上位群もしくは下位群の人数である。ここでは90。識別度は，

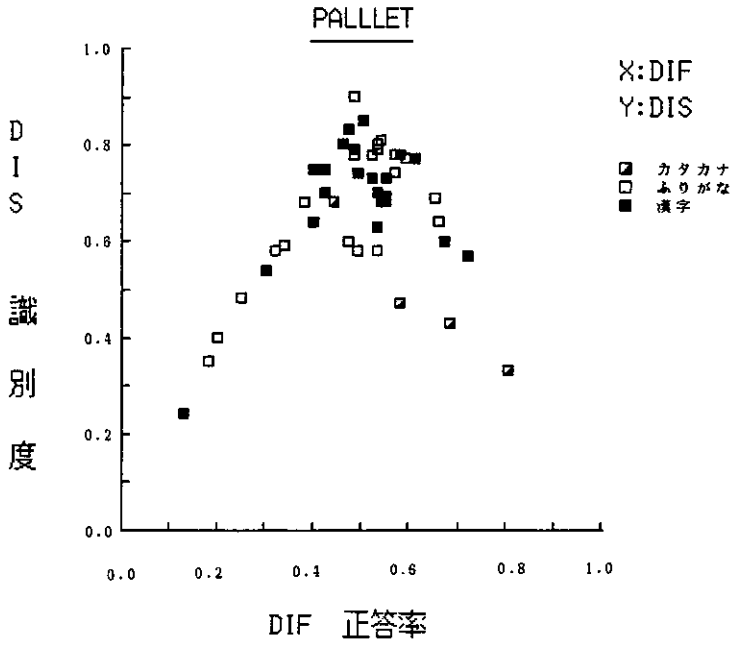
$$DIS = (HC - LC) / N$$

によって表わす。仮に上位群が全員正答し，下位群が全員誤答したとするならDISの値は1.0になり，この時その問題項目は成績上位者と下位者を最もよく識別する。上位者と下位者を全く識別しない場合DISの値は0.0となり，DISの値がマイナスになるのは成績下位群に上位群よりも正答者が多い場合である。ヒストグラムは今回のテストにおける学生の分布をみるのが目的であったからテスト毎に図を作成した。しかし項目分析においてはデータの数が多ければそれだけ項目分析自体の信頼度が上がるので3回のテストのデータを合わせる。当然のことながら変更した問題と追加問題のデータは入っていない。これらについては後に個別に論じる。なお項目分析をするにあたってもパーソナル・コンピュータを使用した。まず先の「MLISTA」によって上位群，下位群の境界得点を決め,「dBASE Ⅲ」によって各下位テストと総得点毎の上位群，下位群を作った。次にこのデータをベーシックプログラム「ITEMA」[3]に入れて識別度と正答率の算出を行った。「ITEMA」では識別度と正答率の算出のほかに，上位群，下位群，母集団全体それぞれの，各選択肢の選択状況とその比率が出せるので結果の検討にはそのデータも使用した。

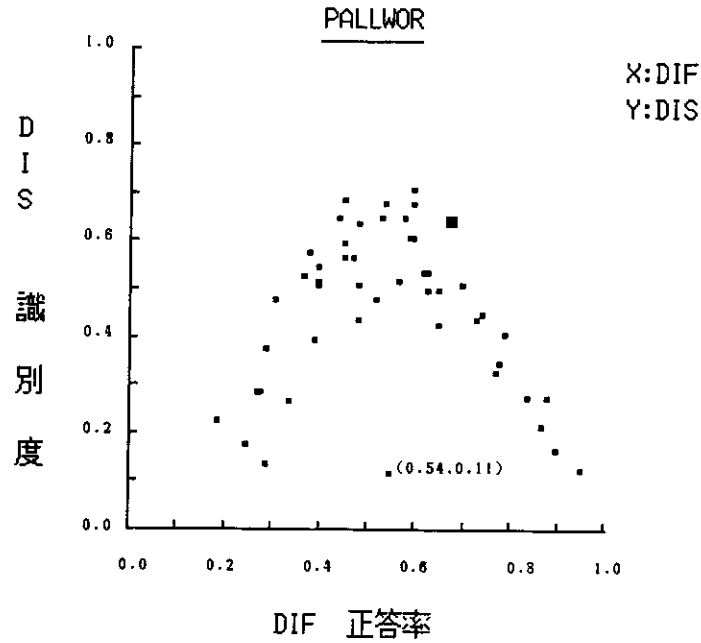
項目分析の図を見るにあたってまずどのような分布が望ましいかを考えたい。正答率についていえば，受験者が全員できても逆にほとんどの受験者ができなくてもテストとしては適当でない。しかし一つのテストには，そのようにほとんどの者ができる項目とできない項目とが入っていて分布に広がりの出ることが必要であるから，DIFの値は0.3～0.7位の範囲にあれば順当といえよう。識別度DISは成績の良い人ができ悪い人ができないという意味でまともな問題であるならば，DIFが0.5の時最も高い値を示す。しかし，正答率にある範囲が必要である以上，識別度が低い項目もテストに含めざるを得ない。全体としてはDIFが0.5，DISが1.0の点を三角形の頂点として，頂点付近に50%，三角形の二辺の上方にそれぞれほぼ25%ずつ問題項目の散らばるのが望ましいと考えられる。なお項目分析の図もOfficegraph[4]による。



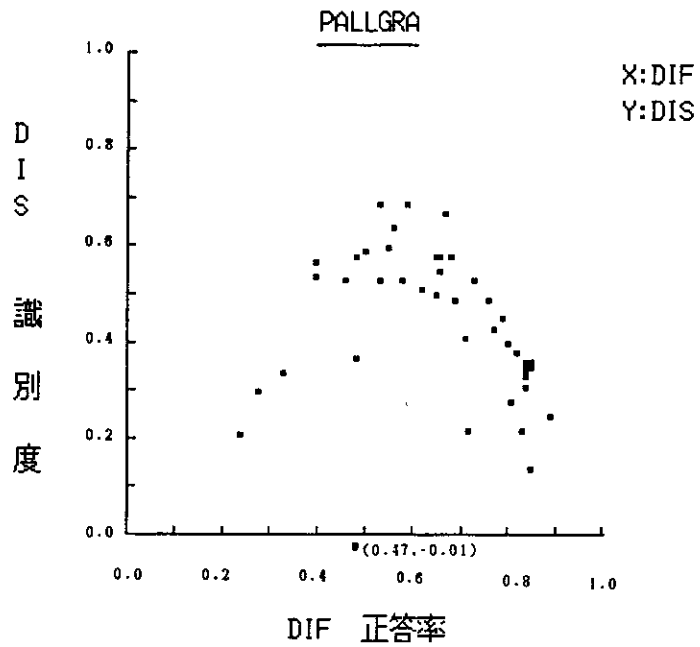
図－8　聴解における DIF と DIS の関係



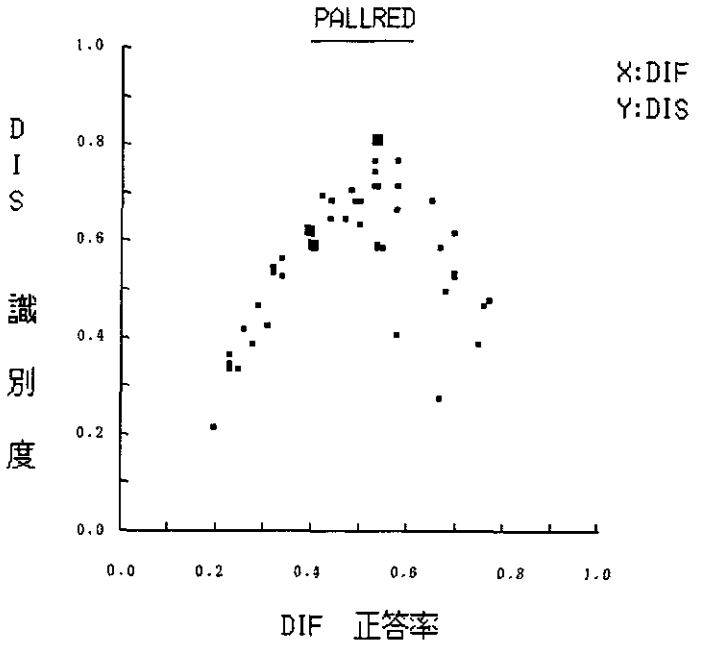
図－9　文字における DIF と DIS の関係



図－10　語彙における DIF と DIS の関係



図－11　文法における DIF と DIS の関係



図－12　読解における DIF と DIS の関係

聴解（図－８）は正答率が全般に高い。識別度は一題を除いては0.6を超える問題がない。DIF が0.5であっても DIS が低いということはこのテストの信頼性が低いということになろう。問題がテープの声で一回限りの繰り返しのきかないものであることや，そもそも聴解が他の教科とは異なった能力をみているのではないかと考えられる。図－８において DIF が0.5であるにもかかわらず DIS が低くなった問題は次のようなものである。

１．きのうから風邪をひいて熱があるんです。
　　それはだめですね。
２．きのうから風邪をひいて熱があるんです。
　　それはよかったですね。
３．きのうから風邪をひいて熱があるんです。
　　それはごくろうさまですね。
４．きのうから風邪をひいて熱があるんです。
　　それはいけませんね。

この４組の会話文の中から一番適当なものを選ぶ。正答は４であるが，４を選択した者と１を選択したものの数にはほとんど差がなく，かつそれぞれの上位群と下位群の数にも差がない。

文字（図－９）は全般に正答率が低い中で，ひらがなをカタカナにする問題は正答率が高く識別度が低い。一方漢字を書かせる問題は識別力の高い問題といえる。図－９において三角形の頂点に位置する項目（DIF 0.47，DIS 0.89）は５つの下位テスト中最も識別度の高い問題であるが，それ

は「悩み」の読みである。

語彙（図－10）は識別度が特に高い問題項目はないものの比較的バランスよく三角形の辺上に問題が散らばっているといえる。この中でひとつやさしすぎもむずかしすぎもしないのに識別度の低い問題（DIF 0.54, DIS 0.11）のあるのが気になる。この問題は次のようなものである。下線部に入る最も適当なことばを１～４の中から選ぶ。

A：お母さまのぐあい，いかがですか。

B：ええ，おかげさまで，だいぶよくなりました。

A：それは，＿＿＿ですね。

1．多幸　2．安全　3．決心　4．安心

そこでこの問題の母語別の反応をみてみると次のようであった。

選択肢	中国語	韓国語	その他
1	28	74	11
2	0	0	7
3	0	1	10
4	72	25	57
無答	0	0	15
計	100	100	100

表－4　母語別の反応　（単位：パーセント）

数字は中国語・韓国語・その他の話者をそれぞれ100として，各選択肢を選んだ人数の割合を示す。この表から全体の26％を占める韓国語話者が正答の選択肢4よりも１を多く選んでいることがわかり，この問題の識別度の低い原因は韓国語の影響によると考えられる。

文法（図－11）はやさしい問題項目が多く，識別度も全般に高くない。識別度がマイナスになった問題（DIF 0.47, DIS －0.01）は，次の問題である。下線部に入ることばを選ぶ。

A：私，大学院に入れるかどうか心配なの。

B：ぼく＿＿受かったんだから，きみなら心配ないよ。

1．なんか　2．でも　3．など　4．ほど

正答選択肢２に次いで１を選択する者が多い。２を選択するものについていえば，P86Aでは上位群と下位群の人数にほとんど差がなく，P85Aでは逆に下位群に正答者が多い。この項目の母語別の反応には識別度を下げるような要因は見られない。筆者にはこの「なんか」と「でも」の基本的な意味・用法が学生に未習であることが識別度を下げた原因ではないかと思われる。初級の学習者は恐らくこの「なんか」の用法は知らないであろうし，中級になっても過小評価するという程度の説明しかなされないことが多い。その点「でも」は，たとえ未習であったとしても「も」の用法

から類推しやすいといえるかもしれない。いずれにしろ更に検証が必要な項目といえる。

読解（図－12）は，語彙以上に問題項目がバランスよく三角形の辺上にのり，識別度も高い。

2．テスト問題の一部変更とその結果

P85A，P86S，P86Aは基本的に同じテスト問題であるが，P86AではP85A，P86Sの項目分析の結果をもとに識別度の極めて悪い問題等に若干の変更を加えた。

まず聴解は，長文のテキストを聞いて10の内容に関する問に答える長文聴解問題のうち2問の識別度が非常に低かったため，テキストにも質問の方にも手を加え，問題のあいまいさをなくした。その結果P86Aでは識別度が上がった。3回のテストは母集団が異なるのであるから，識別度が上がったという言い方は厳密には適当でない。しかし3回のテストの平均はほぼ同一であるので，識別度の変化が母集団の差を無視できるほどに大きければ，それをもって識別度が上下したと言うことにする。

文字はカタカナを書く問題を一問，漢字の読みを書く問題一問，漢字を書く問題三問を変更した。カタカナの問題は元は「こんびゅうたあ」を書かせる問題であったが，語末の長母音がなくてもまちがいではないので，語間に長母音の入った「ちょこれいと」に変更した。しかし識別度に特に変化はなかった。漢字の読みは元は「住民の憩いの場」であったがむずかしすぎて識別が低かったため「今9時11分です」に変更し，その結果識別度が上がった。漢字を書く問題のうち一問は他の項目と重複部分があったために変更し，二題は正答が一字に定まらないものであったために（例「贈られる」，「送られる」のどちらでも正答）変更した。

文法の変更は9題で，うち4題はこそあどの問題である。元の問題は次のようなものであった。

太郎さんが電車にのっていると，花子さんも乗ってきました。花子さんは本を持っています。

太郎「＿a＿は何の本ですか。」

花子「＿b＿はテニスの本です。＿c＿本はとてもいい本なんですよ。」

太郎「そういえば，きのうテレビでテニスの試合がありましたね。見ましたか。」

花子「ええ，見ました。」

太郎「私も見ました。＿d＿試合はとてもおもしろかったですね。」

a：1．これ　2．それ　3．あれ　4．この

b：1．これ　2．それ　3．あれ　4．この

c：1．これ　2．この　3．その　4．あの

d：1．この　2．その　3．それの　4．あの

dを除いた3問はやさしすぎて識別度が非常に低かった。3問はこそあの用法が初級の教科書に出てくるそのままの形であるから，学生はほとんど自動的に正答を選ぶことができたであろう。しかし，実際の状況ではたとえばaの場合，相手の本に直接触れながらの発話であれば「これ」で指示することが可能である。そこで答がその状況設定のもとでは一義的に決まるように問題を作りかえ

た。その結果は興味深いことに識別度が更に下がった。これは問題が改良されたために母語の影響がはっきり出てきたためであるが，この「こそあ」に関する問題は，これ自体一つの大きなテーマとなりうるものなので，ここではこれ以上論じない。[5]

読解の変更は2問である。2問とも問題文にあいまいさがあったために識別度が低かったと考えられ，問題文を変えることによって識別度は上った。図－13は，この識別度の変化を示したものであるが，ちなみに変化の大きかった方の問題（◪，■，□で示したもの）は次のようである。

「人類は森の中で生まれた」といわれるほど，人間と森林は密接な関係にあります。今は，森の中に住んでいなくも，水や酸素は森林から受けとっているわけですから，それがなくなったら生きていけません。木材や燃料も森林の産物であるのはいうまでもありません。

質問1．人間は何がなくなると生きていくことができませんか。

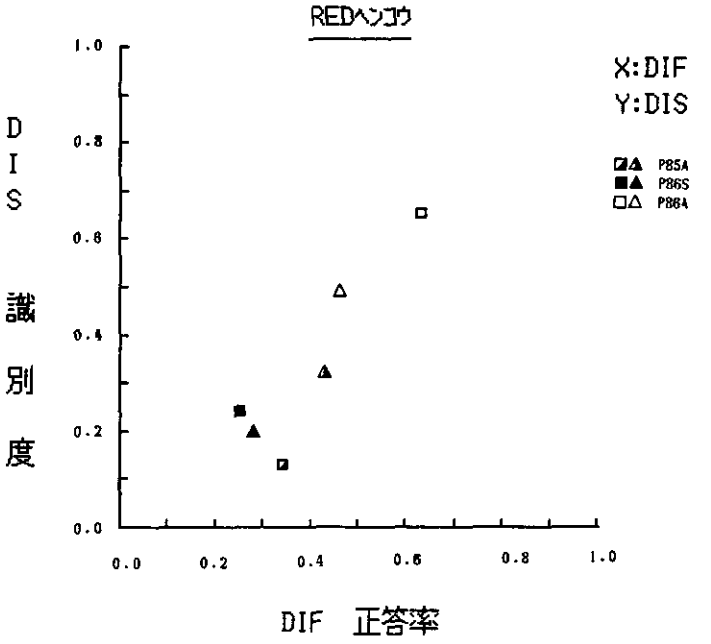
1．森林

2．水か酸素

3．木材

4．燃料

P85A，P86Sでは選択肢2を選択する者の方が1を選択するものより多かった。そこで選択肢2の「水か酸素」を「産物」に変更したものである。



図－13　読解の変更問題のDIFとDISの関係

3．追加問題とその考察

P86Sでは文法問題を10問，P86Aでは聴解の問題を10問追加した。

文法の追加問題は語順に関するもので図－14はその項目分析の結果である。ちなみに表中のDIF 0.56，DIS 0.68の問題は次のようなものである。

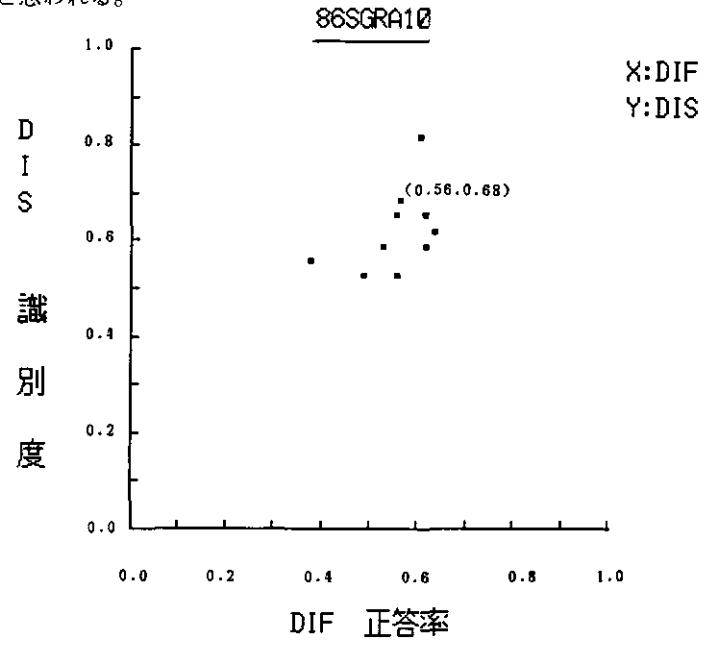
1．あの人はどうして医者に息子をさせたかったのでしょうか。

2．あの人はどうして息子を医者にさせたかったのでしょうか。

3．あの人は医者に息子をどうしてさせたかったのでしょうか。

4．あの人は息子を医者にどうしてさせたかったのでしょうか。

以上の4文から一番よい文を選ぶ。概して語順の問題は識別度が高い。膠着言語である日本語は語の順序にある程度の自由があり，またこの問題は文脈がないので，本当は正答が一つとはいえない。それにもかかわらずこのような語順の問題が日本語の習得度の識別に意味をもつということは注目すべきことと思われる。



図－14　文法追加問題におけるDIFとDISの関係

聴解の追加問題は，問題が音声テープで与えられたため聴解と称しているが，本来は日本語の運用能力を測りたいと考え作成したものである。従来のテストは主に教科書の内容をどの程度習得したかをみるものといえるが，語学を習得するということにはその文化圏での思考方法を身に付けるということも含まれると考えられる。このような前提に立って日本人らしい表現意図の習得度をみる目的で作成した10問であったが，結果は図－15に示されるように識別度が低かった。表中のDIF 0.27，DIS 0.06の問題は次のようなものである。

A：図書館にない本があるんですが……。

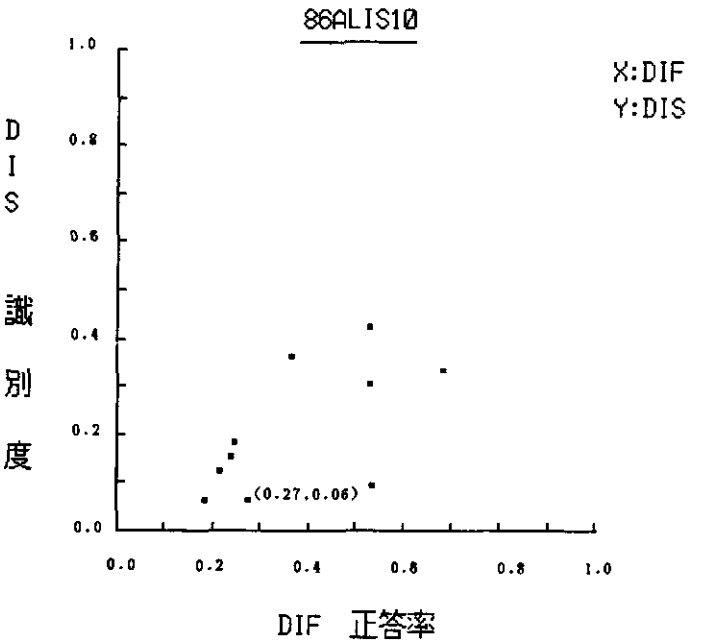
B：1．それは残念ですね。

2．本屋にいけばありますよ。

3．じゃ，私のをおかししましょう。

4．私のうちにはあります。

識別度が低かった原因の一つにはテープの音声が聞きとりにくかったという技術的な問題と，またひとつには日本人でも答がひとつには定まらないという問題のあいまいさが考えられる。人間関係を含むその場の状況を把握しているかどうかを問題にするならば，ビデオを利用するのも一つの方法であろう。



図－15　聴解追加問題における DIF と DIS の関係

4．相関と信頼性

表－5は，追加問題を除いて3回のテストを合わせ，下位テストのα係数と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。なおα係数の算出にはベーシック・プログラム「ALPHAA」[3] を用い，相関はSPSSのディスケット版によって求めた。

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

N＝272

表－5　下位テストと総得点間の相関と信頼性　（ ）内はα係数

下位テスト間の相関では「文字」と「聴解」が最も低いが，これは納得できる結果といえる。逆にもっとも相関が高いのは「語彙」と「文字」であった。「語彙」は総得点との相関も高い。信頼性は「聴解」が他の下位テストに較べて低い。これは日本語能力認定試行試験の結果[6]ともまた国際基督教大学でなされた試験の結果[7]とも一致するものである。当センターの聴解テストは他の下位テストと同じ項目数であったから，聴解の信頼度係数が下がるのは項目数が少ないためではなく，問題の施行方法自体に原因があると考えられる。

V．今後の課題

以上1985年から86年にかけて3回に渡って施行したプレースメント・テストの結果についてみてきた。すでに新しいプレースメント・テストの準備が進んでいるが，現行のテストの施行・採点・結果の分析を通して出てきたいくつかの問題をここにまとめておく。

1．施行方法の統一

すでに述べたように3回のテストは制限時間等に異なる点があり，このことはテストの比較をする際の妨げとなっている。すでに「プレースメント・テストの施行方法」と題するマニュアルができているが，今後はこうしたものをテストの際に参考し，できるだけ同一条件のもとでテストが行われるように努めたい。

2．テスト時間の短縮

2時間半という試験時間はかなり長く，学生の負担が大きすぎるという印象を受ける。今後はテストの信頼度は落さずにテストを短くする方法が考えられなければならない。その場合項目分析の結果をもとに項目数を減らすという方法も考えられようし，また浅野等[8]が試みた下位テストを削減するという方法も考えられよう。

P86Aで「聴解」をテストの4番目にしたことについて，「疲れて集中できないのでもっと早くやってほしい。」という意見が学生から出された。学生がやりやすいように，また正答率が上がるようにとP86Aでは項目を正答率の高い順に並らべたが，この効果は明らかでない。しかしテストを受ける者の立場に立ったこうした配慮は必要であると思われる。

3．採点基準の統一

採点基準は採点者によって揺れが出るものであるが，そもそも問題の作成時点で採点に揺れを生じさせてしまっていることがある。一例をあげるならば，はじめは4肢選択肢の記号にアルファベットの小文字のａｂｃｄを使っていたが，学生の答のａとｄに紛らわしいものがあり判断に迷った。こうしたことについても記録があるので今後参考にしたい。

4．採点方法

1985年秋より従来の手作業によるテストの採点をパーソナルコンピュータによって行うようにした。大型のコンピュータを使わずにパーソナルコンピュータによったのは，外部に依頼する時間のロスを防ぎたいことと，いつでも身近かなところですぐにデータを扱えるようにしておきたいと考えたからである。当初の目標は現在ほぼ達せられたので，今後はパーソナルコンピュータでは算出しにくい因子分析等の利用のため大型コンピュータとの接続も図りたい。

なお採点にあたってミスをできる限り少なくするため，ベーシックプログラム「COMPA」を使って入力した素データをチェックするようにしたところ，手作業で採点していた時には全体のほぼ40％にあたるデータに採点まちがいのあることがわかった。データ量が多いことや途中で採点基準を変えた箇所（主に「文字」）があるにしても自戒すべき数字である。その意味でも「COMPA」導入の意義は大きかった。

5．項目分析

プレースメント・テストの問題としては識別度の高いことは望ましいことであるが，識別度が低いからといて，日本語の問題として悪いとは必ずしもいえない。むしろ識別度の低い問題の中に母語の影響の大きい項目があり，日本語として考えるべき問題を含んでいることはすでにみえてきた。母語別の誤答分析とも合わせて更に項目分析結果の細かい検討を行い，教育の現場に還元したいものである。

6．テスト問題の分類

問題の正答率や各項目における学生の反応分析によって，学生にとってどんな問題が習得しやすいのか，あるいはどんなところがまちがえやすいのかある程度推測することができる。今後この面での分析も更に体系的に行う必要があろう。

7．下位テストの構成

現行のテストは五つの下位テストからなっているが，問題の中にはかなり重なりあったものもある。そもそもクラス分けのテストとしていくつの下位テストが，それぞれどのような内容でどういう比率であるべきか，今後検討していかなければならない。そのためにはまず現行のテストの因子分析等が必要であろう。

8．テストの規準化

教師は今のところ学期がはじまって授業にともかくも出てみなければ学生の日本語力がわからない。同じテストを繰り返して使うのは同一受験者を出すという面では好ましくないが，今後は一部

を再試行するなりしてテストの規準化を図り，プレースメント・テストの結果によって学生の日本語力がある程度推測できるようにしたいと考えている。

注

(1) テスト結果の処理・データの分析は川原裕美，酒井たか子，三枝令子が担当した。

(2) 日本語 dBASE Ⅲ　日本アシュトンテイト株式会社

(3) 三枝紀雄氏の作成による。(本論集『Development of Item Analysis and Related Programs for Personal Computer』参照)

(4) Officegraph　ver 2.0　NEC

(5) 本論集　酒井たか子（誤用研究中間報告)「コソアの用法の研究
—根本原則のキャンセル条件—」参照

(6) 参考文献(4)参照

(7) 参考文献(5)参照

(8) 浅野博・大友賢二・吉江森男，1986　「英語テスト・データ分析の方法ー下位テストの削減ー」外国語教育論集第８号　筑波大学外国語センター

参考文献

(1) 池田央　1978　「テストで能力がわかるか」　日本経済新聞社

(2) 海保博之編著　1985　「心理・教育データの解析法10講　基礎編」　福村出版

(3) 三宅一郎・山本嘉一郎　1985　「SPSS 統計パッケージⅠ基礎編」　東洋経済新聞社

(4) 日本語教育学会　1985　「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」

(5) 石田敏子他　1985　「外国人学習者の日本語学力構造の解明」(昭和57，58，59年度科学研究費補助金研究成果報告書)

(本論文は昭和61年度筑波大学学内プロジェクト研究からの援助に基づくものである)