プレースメント用リスニングテスト改善報告

宮内　俊慈
平田　裕
小山　揚子

要旨

05 年に開発したプレースメント用リスニングテストを 05 年秋学期と 06 年春学期に実施し、留学生の日本語レベルの識別に利用できるかどうか検証した。その結果、機械的に適用するには無理があるが、参考データとして、特に会話力が弱いと思われる学生をどのレベルにするか判断する上で有効であることが分かった。検証作業では「古典的テスト理論」だけでなく「項目応答理論」に基づく分析を行い、テストの信頼性と妥当性を高めるための改善点を洗い出した。本稿では改善版の開発報告と試行結果の評価も合わせて行う。

【キーワード】　プレースメントテスト、リスニング、選択肢数、項目応答理論、
項目困難度

1. はじめに

05 年の本学紀要 15 号で報告したように、関西外国語大学の日本語プログラムでは新規留学生のクラス分けのためのリスニングテストを開発した。通常のリスニングテストと違う条件としては、1) ごく初級の学生から上級の学生まで合計 300 名以上という多数の新規留学生に対して、同じ問題で一斉に行う、2) 他の種類のプレースメントテストに影響を出さないように、リスニングテストの説明から回収までの全工程を 20 分ほどの短時間で行う、というものである。前記の報告は試行までであったが、05 年秋学期、06 年春学期の 2 回、本番として実施した。そして、その結果を検証し、問題のある項目を改善し、新たな試行テストを作成した。混乱をさけるため、本番で使用したテストを旧テスト、新たに作成した試行テストを新テストと呼ぶことにする。本稿で、旧テストの結果を振り返るとともに、改善に向けての新テストの開発及びその試行結果について報

告する。

2. 旧テスト本番結果の検証

2.1.1 基本統計値と信頼性係数

表 1 に 05 年秋と 06 年春の本番データにおける基本統計値と信頼性係数としてクロンバックのα係数を示す。

表 1　旧テストにおける 05 年秋・06 年春の基本統計値

	05 年秋	06 年春
受験者数	318	184
項目数	15	15
最低点	0	0
最高点	15	14
平均点	5.544	5.000
標準偏差	3.066	3.410
α係数	0.718	0.800

ここで、05 年秋と 06 年春とで受験者数に大きな開きがあるのは、プレースメントテストは新規の学生のみに実施しており、春学期から新規にスタートする学生が少ないためである。また、全く日本語学習のバックグラウンドがない学生はプレースメントテストを受験しないので、この受験者数と留学生別科の学生数とは一致しない。

信頼性係数とは、テストの信頼性を示す係数である。ここで、「テストの信頼性」というのは、他の言葉で言えば「テスト得点の安定性がある」ということである。つまり、テストの信頼性が高いということは、能力を測定した時にその測定誤差があまりないということを意味する（大友 1996）。この信頼性係数に関しては、06 年春のデータではかろうじて 0.8 を示したものの、05 年秋のデータでは 0.8 を大きく割り込んでいる。これは、本論集第 12 号（坂井、宮内 2002）で報告した漢字プレースメントテストの信頼性係数が 0.9 を大きく越えているのと比較するとかなり低い数値となっている。0.718 という数字は、先ほどの測定誤差の観点から言えば、ある学生の取った点数の 30%近くは測定誤差（偶然によるもの）であり、70%ほどしか信用できないということである。このリスニングテストによってのみ、学生のプレースメントを決めるわけではないためプレースメントテスト全体としては、大きな問題とはならないかもしれないが、プレースメントの判断材料の一つになることを考えれば、信頼性が高いことに越したことはない。

この数値が低くなった主な原因としては、問題数が 15 問と少ないことが考えられ、今回のプレースメントテスト改善に着手した最大の理由となっている。

2.1.2 レベル別平均点の比較

旧テストがプレースメントテストとして機能しているかどうかを評価する一つの方法として、最終的に決まった学生のレベル別平均点の比較を行った。本学のプレースメントは主に文法力のテストの点数によって決められており、レベルが高ければ聞き取り能力が高いとは必ずしも言い切れない。しかし、授業開始後２週間は調整期間となっており、会話力を含む学生の実力とレベル分けがうまく合っていない場合は、関係するレベルの教員が総合的に判断し、必要であれば再テストを行って違うレベルにクラス分けし直している。そのため、最終的に決定されたレベルと学生の総合的日本語力との相関性はかなりの程度高いものと言える。従って、本テストがプレースメントテストとして正しく機能していることの一つの条件としてこのレベル毎の平均点の比較は、大いに意味があると言える。

表２　旧テストにおける 05 年秋と 06 年春のレベル別平均点

レベル	度数		平均値		標準偏差		標準誤差	
	05 年秋	06 年春	05 年秋	06 年春	05 年秋	06 年春	05 年秋	06 年春
1.7	37	20	12.32	9.80	7.142	7.281	1.174	1.628
2.0	78	49	16.77	14.86	8.105	8.944	0.918	1.278
3.0	90	54	24.09	20.37	8.873	8.450	0.935	1.150
4.0	77	31	28.36	30.06	9.823	8.996	1.119	1.616
5.0	9	7	43.56	40.00	10.853	5.164	3.618	1.952
6.0	7	5	49.14	51.20	11.936	5.215	4.512	2.332
合計	298	166	22.99	21.04	11.776	12.393	0.682	0.962

表 2 に 05 年秋と 06 年春のレベル別の平均点を示す。このレベル別平均点の比較においては、前節の素点データとは異なり、便宜上各項目 4 点として 60 点満点でスコアの計算がされている。また、度数の合計が前節の受験者数と異なるが、この表には受験者であるにもかかわらずレベル 1（日本語学習経験なし）にプレースメントされた学生、および、アカデミックジャパニーズ（日本人と同様のクラスを履修）になった学生を含めていないからである。

このレベル別の平均点をプロットしてグラフにしたものが、図 1 である。05 年秋のデ

ータにおいても 06 年春のデータにおいてもレベルが上がるにつれて平均点が上がっていることが見て取れる。これらの平均点の差が有意なものであるかどうかを検証するた

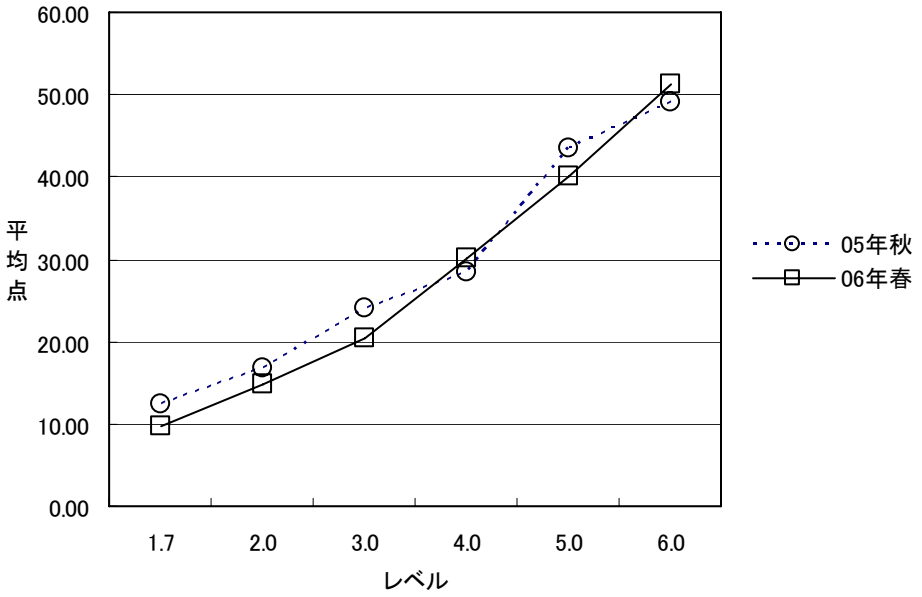

図１　本番テストのレベル別平均点

めに、一元配置の分散分析とその後の多重比較（Tukey の HSD）を行った。多重比較の結果を添付資料として示す。

レベル 2 と 3 の間、レベル 3 と 4 の間では 05 年秋、06 年春ともに有意な差が見られた。レベル 1.7 と 2 の間とレベル 5 と 6 の間はいずれのデータでも有意な差は見られなかった。レベル 4 と 5 の間は、05 年秋では有意な差が見られたが、06 年春では 10%水準では有意な差があるものの、5%水準では有意な差が認められなかった。

この結果から、本プレースメントテストがレベル 2、3、4 のプレースメントには有効に機能していることが示唆される。レベル 1.7 と 2 の差は有意なものではなく、このテストを用いてこの二つのレベルのプレースメントを決めることには問題があると言えるが、元々このテストは低いレベルの学生の弁別を目的として開発されたものではなく、また、レベル 1.7 というのは内部的に区別されているレベルであり表面的にはレベル 2 として扱われているためこのテストの欠点とはならない。また、一番上のレベルであるレベル 5 と 6 の間の差も有意でなく、このリスニングテストが上級レベル内の識別に有効でないことが示唆されたが、これらのレベルに入るには教師のインタビューが義務づけられており、リスニング力や文法力だけでなく総合的に判断しているので、この点で

も問題はない。一方、レベル 4 と 5 の差を明瞭に示すことは、この聞き取りプレースメントテスト導入当初からの目的であったことから、このレベル間の差が 06 年春のデータでは明確に示せなかったことは、このテストの改善の必要性を示唆している。もう少し難しめの問題を増やすことが必要になると言えそうである。

2.1.3 項目分析

今回のリスニングプレースメントテストの改善においては、従来の古典的統計理論に基づく項目困難度、項目弁別力、実質選択肢数だけでなく、項目応答理論に基づく項目困難度パラメータおよびモデルとの適合度の統計値も参照しつつテスト問題の項目分析を行った。表 3 に 05 年秋と 06 年春の本番結果のそれぞれの値を示す。この表における数値は全て Test Data Analysis Program (TDAP) Ver. 2.0（『テストで言語能力は測れるか』（大友賢二監修／中村洋一著）添付プログラム）を使用し出力されたものである。このプログラムでは、1 パラメータ・ロジスティック・モデル(The Rash Model)を採用しており、パラメータの推定においては、簡易推定法として用いられる PROX 法（approximation procedure）を用いて計算が行われている（中村 2004）。

2.1.3.1 項目弁別力指数

項目弁別力指数については、本論集第 12 号の漢字プレースメントテストの改良の際に詳述したので、ここでは詳しく述べないが、簡単に言えば、ある項目が能力の高い受験者とそうでない受験者を弁別することができる度合のことであり、-1.000 から+1.000 の範囲で示される。+1.000 に近いほど項目弁別力は高いと言える。どれぐらいの数値であれば適切であるかについては、「上位・下位項目弁別力指数（ULD）」では 0.400 以上、「点双列相関係数による項目弁別力指数（DISC）」では 0.300 以上であればよい項目であるとされている（中村 2004）。

この 2 種類の項目弁別力指数、ULD と DISC で不適切と判断される項目は、それぞれで違った結果が得られた。ULD で不適切と判断されるのは、項目 11、12、13、15 の 4 問（表 3 参照）、DISC で不適切と判断されるのは項目 12 のみである（06 年春に 0.262 で一般的に適切であると考えられる 0.300 未満）。より厳しい結果を採用ということで、今回の改善に着手する際には、ULD に基づき改訂項目の絞り込みを行った。

具体的な ULD の数値であるが、項目 11、12、13、15 の 4 問は 05 年秋、06 年春共に 0.4 を下回り、それぞれ、項目 11（05 年秋: 0.291、06 年春: 0.320）、項目 12（05 年秋: 0.360、

表 3. ０５年秋と０６年春の本番項目分析統計値

項目 NO.	上位・下位項目弁別力指数(ULD)		点双列相関係数による項目弁別力指数(DISC)		実質選択肢数(AENO)		項目困難度(DIFF)		項目困難度適切度(ADIF)		項目困難度パラメータ(Final Calib.)		モデルとの適合度(t)	
	05 秋	06 春	05 秋	06 春	05 秋	06 春	05 秋	06 春	05 秋	06 春	05 秋	06 春	05 秋	06 春
1	0.558	0.500	0.521	0.513	2.977	2.879	0.286	0.228	0.322	0.207	0.334	0.560	-2.504	-0.825
2	0.628	0.600	0.490	0.462	2.420	2.293	0.535	0.505	0.819	0.761	-0.984	-1.141	-0.130	2.034
3	0.686	0.760	0.519	0.604	2.543	2.582	0.465	0.418	0.681	0.587	-0.635	-0.646	-1.197	-0.572
4	0.698	0.840	0.505	0.609	2.661	2.543	0.541	0.511	0.832	0.772	-1.016	-1.173	-0.428	-2.031
5	0.384	0.720	0.371	0.514	2.180	2.090	0.619	0.598	0.989	0.946	-1.427	-1.695	1.721	0.985
6	0.570	0.500	0.513	0.486	2.516	2.527	0.318	0.255	0.385	0.261	0.148	0.362	-1.510	2.779
7	0.523	0.600	0.436	0.513	2.773	2.693	0.563	0.554	0.876	0.859	-1.129	-1.428	0.224	0.311
8	0.616	0.680	0.484	0.590	2.665	2.986	0.522	0.424	0.794	0.598	-0.920	-0.677	-0.118	-1.566
9	0.523	0.620	0.499	0.657	2.739	2.696	0.264	0.207	0.278	0.163	0.473	0.730	-0.821	-5.038
10	0.547	0.600	0.449	0.554	2.927	2.986	0.352	0.299	0.454	0.348	-0.046	0.070	-0.592	0.691
11	0.291	0.320	0.344	0.502	3.107	2.999	0.145	0.136	0.039	0.022	1.425	1.393	0.774	0.009
12	0.360	0.220	0.393	0.262	2.795	2.651	0.198	0.158	0.146	0.065	0.943	1.165	-0.291	5.350
13	0.395	0.380	0.402	0.410	2.626	2.485	0.274	0.234	0.297	0.217	0.413	0.519	1.673	0.158
14	0.465	0.720	0.415	0.576	3.320	3.051	0.403	0.413	0.555	0.576	-0.314	-0.614	1.916	1.523
15	0.198	0.220	0.436	0.450	3.297	3.164	0.060	0.060	-0.131	-0.130	2.736	2.573	-7.575	-6.250

06 年春: 0.220)、項目 13（05 年秋: 0.395、06 年春: 0.380)、項目 15（05 年秋: 0.198、06 年春: 0.220）となっている。

項目 12 から 15 まで（セクション III）は、比較的長いダイアログを聞いてからその内容に当てはまる正解を一つ選ぶといった典型的なリスニング問題形式であり、項目 12 と 13 が 1 つのダイアログに対する問題、項目 14 と 15 がもう一つのダイアログに対する問題という構成になっていた。この問題形式では、問題の始まりから解答が終了するまでの時間がかかる割に問題数が増やせないという点と、上に述べたように項目弁別力指数が低いという 2 つの点から、この形式の問題を全てやめ、項目 1 から 6 のやりとりが 1 回のみの形式の問題（セクション I）と項目 7 から 11 のある発話に対する最も適切な応答を選ぶ形式の問題（セクション II）を増やすことにした。

2.1.3.2 実質選択肢数（AENO）

多肢選択形式の問題においてどのような選択肢を作成するかは、テスト作成の際に非常に重要な問題となる。特に錯乱肢は、中村（2004）が述べているように「能力の低い受験者に正答だと思わせる魅力があり」、かつ、「能力の高い受験者には誤答であると分かるものでなくてはならない」。受験者の誰もが選ばないような選択肢がテスト項目の中に含まれている場合には、改善の必要がある。実質選択肢数（AENO）とは、提供された選択肢が実質的にいくつの選択肢として作用していたかを示す数値となり、今回のテストのように選択肢が 4 つの場合、0.000 から 4.000 までの数値を取る。2.000 となった場合には、選択肢が 4 つにも関わらず、実質的には 2 つしかないのと同じ意味だということになる。

そういう意味では、今回のテストでは、全廃する項目 12 から項目 15 を除けば、項目 2（05 年秋：2.420、06 年春：2.293)、項目 5（05 年秋：2.180、06 年春：2.090）が改善の対象となる。

2.1.3.3 項目困難度（DIFF）と項目困難度適切度（ADIF）

項目困難度（DIFF）は、正答率とも呼ばれるが、0.000 から+1.000 の数値をとり、1.000 に近ければ易しい問題、0.000 に近ければ難しい問題だと言える。プレースメントテストのような集団基準準拠テストでは、DIFF は 0.500 が理想的だとされる（中村 2004)。これは、正答率が 1.000（全員が正解）あるいは、0.000（全員が不正解）の場合、相対的な位置関係に関する情報が何も得られず、逆に、正答率と誤答率が拮抗する 0.5000 の

時に、相対的位置関係に関する情報が一番多く得られるということと関係している。多肢選択式テストの場合は、当て推量による選択の可能性を含めて考慮しなければならず、今回のテストのように選択肢が 4 つの場合、最適困難度は、0.5 + 0.5 x 1/4 = 0.625 となる。さらに、DIFF と最適困難度を比較して、その DIFF がどのくらい適切なのかを求めたものが、項目困難度適切度（ADIF）となる。「集団基準準拠テストにおいては、項目困難度適切度が高くなれば、より多くの情報を集めることができる」（中村 2004）。

2 回に渡る本番テストの結果を見ると、廃止予定の項目 15 が DIFF が最も高く（05 年秋：0.060、06 年春：0.060）正答率が 10%を切っており、ADIF も唯一マイナスの数値を示している（05 年秋：-0.131、06 年春：-0.130）。続いて、項目 11 の DIFF が高く（05 年秋：0.145、06 年春：0.136）、しかも、この項目も同じく ADIF が低い（05 年秋：0.039、06 年春：0.022）。今後のテストで再使用するかどうか、どのように改善するかは検討すべき項目となる。

2.1.3.4 項目困難度パラメータ（Final Calib.）とモデルとの適合度（t）

項目困難度パラメータ（Final Calib.）は、通常-3.000 から+3.000 の間の値を取り、0 は平均的な困難度であることを意味し、負の値は平均的な困難度より易しく、正の値は平均的な困難度より難しいことを意味している。一般的なプレースメントテストでは、受験者の能力を正確に測定したいわけであるから、-2.000 や-3.000 と言った項目困難度パラメータを持つ易しい項目から 2.000 や 3.000 といった項目困難度パラメータを持つ難しい項目まで万遍なく含まれていることが望まれる。しかし、今回の聞き取りプレースメントテストにおいては、本学におけるレベル 3 からレベル 5 までの学生の聞き取り能力を区別することが目標であり、また、試験時間の制約から問題数にも制限があるため、-2.000 を越える項目困難度パラメータを持つ易しい項目は必要としない。また、3.000 程度の項目困難度パラメータを持つ最高難度の問題を入れることも妥当ではない。

05 年秋と 06 年春の本番データの項目困難度パラメータを項目順にプロットしたものが図 2 である。これを見ると極端に易しい項目、あるいは難しい項目はなく、しかも、平均的な困難度より難しい項目が 7 項目、逆に平均的な困難度より易しい項目が 7 項目、そして、項目困難度パラメータが 0 に非常に近い項目が 1 項目とバランスの取れた構成になっているように思われる。ただ、廃止することにした項目の 1 つである項目 15 が 2 回の本番データで 2.5 を越えており（05 年秋: 2.736、06 年春: 2.573）、やはり、難しすぎる項目になっているようである。また、-1.000 を下回る項目が 4 項目あり（項目 2、項

目 4、項目 5、項目 7)、易しい問題が若干多い傾向が見られる。

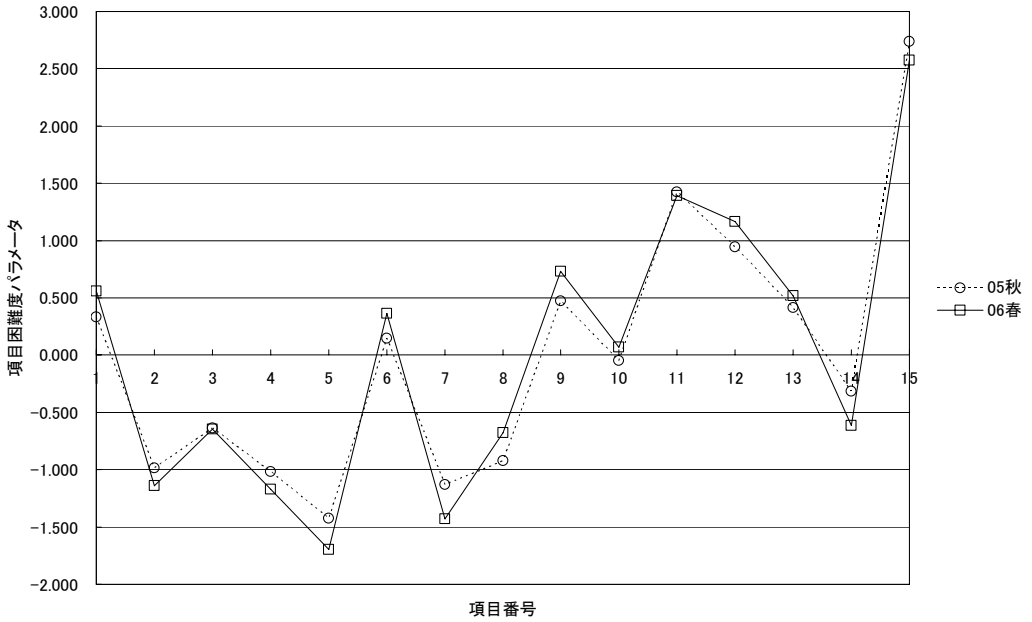

図２　項目困難度パラメータの分布

項目応答理論は、「テスト項目に依存せず受験者の能力値を算出できるモデル」を提供していることが特徴である。別の言い方をすれば、このモデルは、A というテスト項目を受けて求められたある受験者の能力が、B というテスト項目を受けた場合でも同じであるように作られている。このモデルに対して、適合範囲外となるテスト項目と受験者を除いていくことによって、テスト項目と受験者グループを理想状態に近づけることができる。

今回使用した１パラメータ・ロジスティック・モデル（1PLM）では、受験者の能力を求めるモデルとして、

$$P = 1 / (1 + \exp(-(\theta - b)))$$

というモデルが使用される。ここで、P は項目の正答確率、b は項目困難度パラメータ、θ は受験者の能力パラメータである。このモデルでは、b および θ は通常-3.000 から +3.000 までの幅で示され、b が-3.000 の場合、その項目は最も易しく、+3.000 の場合は、最も難しいことを意味する。そして、b=0.000 の場合が、難しくも易しくもない中程度の項目ということになる。また、θ は、-3.000 の場合に最も能力が低く、+3.000 の場合に最も能力が高いことを意味し、0.000 の場合に、中程度の能力の受験者であることを

意味する。さらに、項目困難度（b）が-1.000 で、受験者の能力（θ）が-1.000 であるときに、その項目に対する正答確率（P）が 0.500、つまり 50%となるように設定されている。

「受験者の能力値の算出がテスト項目に依存しない」ということを具体例で見てみよう。例えば、ある能力を持った受験者が、易しい項目Ａ（b=-2.000）と難しい項目Ｂ（b=2.000）を受験したと仮定する。このモデルにおいては、項目ＡとＢに対する正答確率（P）はそれぞれ、P=0.953、P=0.269 と設定されている。従って、そこから導き出される受験者の能力（θ）は、項目Ａを使用した場合も項目Ｂを使用した場合も 1.000 の近似値となる。このモデルでは、項目困難度、正当確率、受験者の能力というパラメーターが理論上そうなるように設定されている。

このモデルに対して著しく不適合のテスト項目とは、特定の受験者グループに対して著しく易しい、または難しく、充分な弁別機能を有していないということである。また、著しく不適合の受験者とは、特定のテスト項目グループに対して、著しく能力が欠如している、または能力が優れていて、そのテストでは弁別できないということである。このようなデータを検証し改善していくことによって、テストの質を向上することができる。今回用いた TDAP の中では、その「モデルとの適合度」の度合が t の値として算出され、この値が+2.000 以上の場合にモデルには適合しない（ミスフィット）と考えられている。

今回の本番テストのデータにおいてテスト項目のミスフィットと判断されるものは、06 年春における項目 6（t = 2.779）と項目 15（t = 5.350）である。項目 15 に関しては、テスト項目から削除することにしたので、項目 6 が問題改善の対象となる。

2.2 プレースメント用リスニングテストの利用法検証

前節では旧テストの本番結果を統計的に評価したが、本節ではプレースメント時の実際の利用方法について検証する。

05 年の開発報告（宮内、平田、小山 2005）で述べたように、プレースメントに用いる他の種類のテスト同様、リスニングテストにおいてもレベル分けの基準点を以下のように設定した：（レベル 2　6 点）（レベル 3　10 点）（レベル 4　18 点）（レベル 5　34 点）（レベル 6　38 点）。[(1)]　この基準点は、文法知識のテストで合格していてもリスニングの結果で低いレベルに入る学生が多くなりすぎないように、既に 1 問分（4 点）は低く設定していたものである。しかし、テストフォーマットに対する不慣れなど、リス

ニングテストの結果と学生の実力が一致していないケースも様々な理由で考えられるので（宮内、平田、小山 2005)、リスニングの基準点を機械的に適用するプログラミングはレベル決めを行う表計算ソフト上で行わなかった。

実際に本番のプレースメントテストでのリスニングテストの結果を見たところ、機械的に上記の基準点を適用すると低いレベルにまわさなくてはいけない学生がまだ多すぎるレベルもあった。05 年秋を例にとると、文法テストの結果ではレベル 4 に入る学生のうち、リスニングの基準点 18 点を適用すると 14 人もレベル 3 に廻らなくてはいけなくなる状態であった。

下のレベルに抑えておく学生が多くなると、実際に授業が始まってから上のレベルへの移動のためのクラス見学と再プレースメントテストを希望する学生が増加することが予想され、学期始めの忙しい時期に教師側の仕事の負荷が著しく増大するという問題が起きる。リスニングテストの結果に基づいて学生を下のレベルにまわすにしても、現実的には人数的な要件も考える必要がある。

そこで、あらかじめ予見していたことではあるが、リスニングの点数はあくまでレベル決めの参考データにすることとし、極端に点数の低い学生についてのみ、一人ひとり全てのテスト結果や日本滞在予定、学習歴、使用教科書などから、本学ではどのレベルで学習した方がいいのかを総合的に判断した。

リスニングテストはレベル 3、4、5 の識別に役立てる意図で開発したが、レベル 5、6 に入る学生は全てインタビューを行ってから判断するので、本稿ではレベル 3、4 のプレースメントについてのみ検証する。

05 年秋学期では、文法テストの結果で見るとレベル 3 であるのに、リスニングの基準点に達していなかった学生は 2 名であった。初級の最後のレベルであるレベル 3 までは、教科書の文法項目を既習であるかどうかという要素が強いので、リスニングが弱いからといってレベル 2 にするのは学生のモチベーション上からも難しい。この 2 名の最終成績はB+とCで、特にいい成績を取れた訳ではないが、レベル 3 の授業についていけないことはなかった。(2)　この点から、レベル 3 においてはリスニングの点数が悪くても文法の知識レベルや学習に対する姿勢で判断してよいと考えられる。

レベル 4 では上記のように 14 名がリスニングの基準点に達していなかったが、リスニングの結果を基にレベル 3 に下げたのは、15 問中 1 問しかできなかった学生が 1 名、2 問しかできなかった学生が 2 名、合計 3 名のみである。このうちの 1 名は、文法テストでもレベル 4 にぎりぎりの合格点だったので、レベル 3 に振り分けた。授業開始後も

特に学生からの不満はなく、レベル 3 での最終成績も B で、このレベルで適切だったと思われる。

残りの 2 名は同じ大学からの留学生で、文法テストの結果は『げんきII』の範囲まではほぼ完璧であり、文法知識と会話力のバランスが取れていない極端な例である。リスニングが極端に悪いので、一旦レベル 3 にプレースメントしたが、既習の『げんきII』をもう一度やることになるので、授業が始まってからレベル 4 への移動をこの 2 名が希望してきた。同じ大学から留学してきている他の学生がレベル 4 にプレースメントされているというのも理由であったと考えられる。教師側の判断としては、一学期のみの滞在予定でもあるし、『げんきII』の内容を知識としてはほぼ完璧にマスターしていたことから、レベル 4 への移動を認めた。(3)

担当した教員の話しでは、学生とのコンサルテーションの過程で、「リスニングの力が極端に弱い」「会話力が極端に弱い」という問題を説明し、レベル 3 で既習内容の反復練習をし、会話力を高めることを勧めたので、文法テストの結果のみですんなりレベル 4 で学習することになるよりも、会話力に対する問題意識が学生の中で高くなったという報告があった。この例では、リスニングテストの結果がレベル決定に直接結びついてはいないが、その検討過程においては有効利用できるデータだと言えるだろう。

06 年春学期のプレースメントにおいても、リスニングテストの結果の利用方法は 05 年秋学期と同様に、基準点を機械的に適用せず、参考データとするということで行った。レベル 3 と 4 の切り分けで話しをすると、リスニングの結果がプレースメントに大きく関係したのは 2 名である。文法テストでレベル 4 に入る基準点をパスしているのに、リスニングの結果が 15 問中 2 問しかできていなかったのでレベル 3 に抑えた学生が 1 名、レベル 3 に入っていた学生でレベル 4 のクラス見学と再プレースメントを希望したが、リスニングテストの結果を考慮してレベル 3 での学習を説得した学生が 1 名である。双方とも 2 学期間の滞在予定だったため、1 学期目はレベル 3、2 学期目はレベル 4 という組み合わせが適切だろうというのも大きな判断理由の 1 つである。学期後の担当教員の評価では、上記の学生 2 名の最終成績は A-と B であり、プレースメントとしてレベル 3 での学習が妥当だったとのことであった。

上記 2 名の学生以外でも、レベル 3 の学生でリスニングが 1〜3 問しか正答できなかった者が 8 名、レベル 4 で 4 問しかできなかった者が 4 名いた。リスニングの基準点としてはレベル 3 が 10 点、レベル 4 は 18 点なので、3 問正答したレベル 3 の 3 名は一応リスニングの基準点をパスしていることになる。しかし、同レベルの他の学生と比較す

るとリスニングの点数としてはかなり悪い。

これらの 12 名の学期の最終成績は、D-が 1 名、C/C+が 3 名、B-/Bが 7 名、A-が 1 名で、Fの成績はついておらず、学期後の担当教員のコメントとしても、プレースメントとして妥当だったとのことであった。この場合も、リスニングの基準点を機械的に適用しないという方針でよかったということである。(4)

その一方で、我々教師の中にはこれまでの経験から、日本語の学習を大きなトラブルなしで継続していくには、各レベルで B から B+以上の成績が望ましいという感覚がある。上述の実例と見ると、文法知識とリスニング力のバランスがとれていない学習者は、一学期のみの留学をこなせても、やはり上のレベルに進むにつれて学習のスムーズな継続が難しくなってくることが考えられる。

ここまで、実際のプレースメントにおけるリスニングテストの利用方法を、05 年秋学期、06 年春学期の 2 回の本番を通して見てきた。その結果、リスニングテストの成績をプレースメントの参考データとして使うのは有効であると言えよう。

しかし、本学の日本語プログラムのように、大学のコースワークとして日本語のクラスを提供している限り、文法知識と会話力のバランスが取れていない学生、各レベルをぎりぎりパスして進級するような学生に対応するのもやむを得ないところである。実力（会話力）が伴っていないからといって、既習の内容を学習するレベルに留まらせるのは現実的にはかなり難しい。その点から言えば、リスニングの基準点を機械的に適用するような運用は、将来的にも無理かもしれない。このような制限はあるにせよ、運用上の目安になるので、基準点の信頼性や精度を上げるのは意義があると言える。

3. 問題の改善

本番テスト結果の検証を通して見えてきた改善点は、1) 0.8 を大きく割り込んだ信頼性係数を上げるために問題数を増やす、2) レベル 4 とレベル 5 の差を明瞭に示すためにはもう少し難しめの問題を増やす、3) 項目 12 から 15 の比較的長いダイアログを聞いてその内容に当てはまる正解を選ぶ形式は、項目弁別力指数では 4 問中 3 問が 0.4 を下回っており、その他にも実質選択肢数や項目困難度などにも問題が多く、全廃する、4) 実質選択肢数に問題のあった項目 2 と 5 の選択肢を見直す、5) テスト項目のミスフィットと判断された項目 6 を再検討する（15 は廃止なので対象外）などである。

3.1 問題数２０問への改訂

05 年開発したプレースメント用リスニングテストの大きな目的は、本学のレベル 3, 4, 5 にプレースメントされる学生の会話力をリスニングテストの結果から推し量りたいというものである。05 年秋、06 年春の 2 回の本番実施において、リスニングテストの結果をプレースメントの参考データとして使い、文法の知識と会話力のバランスが極端に取れていない学生の識別には大方のところ有効であったと言える。しかし、その精度を上げるには、合計 15 問という問題数がどうしても大きな制約になっていると思われた。改訂の 1 つのポイントとしては、テスト時間を延ばさずに問題数を増やす、そして、なるべくレベル 3, 4, 5 近辺の学生の振りわけをターゲットにするということである。

ここで 05 年版の問題のタイプを見てみると、3 つのセクションからなり、セクション I はダイアログにおいて 1 回のやり取りという短い会話から簡単な情報を把握する多肢選択問題、セクション II は会話の一部であるキューに対して適切なレスポンスを選択する問題、セクション III が通常のリスニングによく使われるタイプで、少し長めのダイアログの内容理解の多肢選択問題である。問題のタイプについて、詳しくは 05 年の紀要を参照されたい（宮内、平田、小山 2005）。

実施結果を見ると、セクション I は初級者でもそれなりに答えられる問題も混ぜてあり、幅広い学習歴の受験者を対象に有効な問題形式であると言える。セクション II はこのリスニングの特徴的な問題形式であるが、会話力を推し量ろうという意図からは重要である。情報の理解力を測るセクション I で高得点を取っていても、セクション II で全く点数を取れない受験者もおり、大雑把な傾向ではあるが、そういう場合は得てして会話に慣れていないというケースである。

上記 2 つのセクションに対し、セクション III は比較的長いダイアログの中に情報も構文も詰め込まれているので、短期記憶力の問題もあるであろうが、聞く回数が 1 回だけでは日本語の実力としては上級でないと問題として有効に機能しないようである。これは、もともとセクション III は上級辺りをターゲットに開発したからでもあった。また、問題の効率性から見ると、セクション III は問題の実施に必要な時間の割りには問題数を多くできない。

これらの点を踏まえ、改訂版ではセクション III をなくす方針にした。旧テストではセクション I がダイアログ 3 つ、各 2 問で 6 問、セクション II が 5 問、セクション III がダイアログ 2 つ、各 2 問で 4 問、合計 15 問であったが、新テストではセクション I のダイアログを 4 つ、問題数としては 8 問増やして 14 問、セクション II は 2 問増やし

て 6 問、合計で 20 問になるように計画した。追加する問題は、当初の目的の一つであったレベル 4 とレベル 5 を明瞭に区別できるよう、少し難しめのものを考えた。また、選択肢に問題のあった項目 2 との 5 の選択肢も変更した。ミスフィットと判断された項目 6 は今回手をつけず、試行テストの結果を見て再度検討することとした。

リスニングテストの実施方法としては、解答時間の空白も含めてコンピューターに録音したものを止めることなく聞かせるだけであり、旧テストの録音時間は 10 分弱だったので、問題数を増やしても 10 分前後に抑えたかった。実際に問題を作って録音・編集すると 11 分ほどで収まったので、プレースメントテスト全体に対する時間的な影響は旧テストと同じと言っていいだろう。

3.2 新テストの試行結果検証

3.2.1 基本統計値と信頼性係数

表 4 に 06 年春学期の終了直前に行った新聞き取りプレースメントテストの試行結果における基本統計値と信頼性係数としてクロンバックのα係数を示す。

表 4　新聞き取りプレースメントテストの試行結果

	06 春試行
受験者数	340
項目数	20
最低点	0
最高点	18
平均点	10.197
標準偏差	4.295
α係数	0.816

問題数を 20 問に増やしたことで、信頼性係数は 05 年秋の 0.718、06 年春の 0.800 に比べ 0.816 と若干の改善が見られるが、20 問という制限がある状況では、これ以上の信頼性の向上を求めることは非常に難しいものと思われる。

3.2.2 レベル別平均点の比較

新テストの試行結果についても、05 年秋、06 年春の旧テストと同様にレベル別の平均点の比較を行った。その結果を、表 5 に示す。ここでは、テスト項目が 15 から 20 に増えているため、点数も 80 点満点となっている。また、新テストは学期の終了直前に

実施しているため、旧テストでは対象になっていなかったレベル 1 が存在する。そしてこのレベルの得点が次ぎの学期のレベル 1.7 あるいは 2 のプレースメントを決定する際の参考得点となる。同様に順次レベルが 1 つ上のプレースメント決定の参考得点となっていく。そういう訳で、本番データでアカデミック・クラスのデータが除外されているのと同じ理由で、この表ではレベル 6 のデータは除外されている。

表 5　新テストのレベル別平均点

レベル	度数	平均値	標準偏差	標準誤差	最小値	最大値
1.0	41	18.34	8.988	1.404	0	36
1.7	21	23.43	8.698	1.898	8	40
2.0	61	31.15	10.227	1.309	8	60
3.0	90	39.82	12.081	1.273	12	64
4.0	79	54.99	10.157	1.143	32	72
5.0	48	58.25	10.303	1.487	36	72
合計	340	40.79	17.205	0.933	0	72

新テストのレベル別平均点に関しても、本番データと同様にその平均点をプロットする（図 3）とともに、平均点の差の有意性の検証も実施した（添付資料）。その結果、

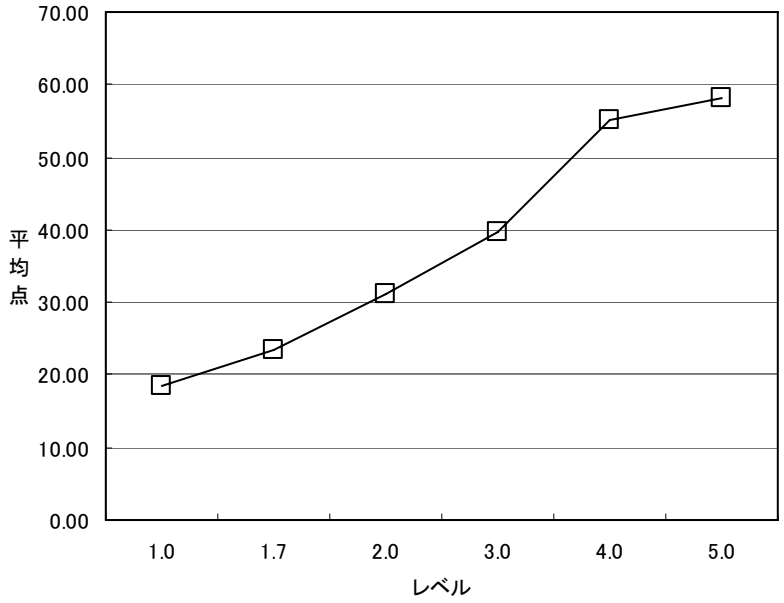

図３　新テストのレベル別平均点

レベル 1 と 1.7、レベル 4 と 5 以外のレベル間では、全て有意な差が認められた。

レベル 1 と 1.7 の差が見られなかった点に関しては、本番テストの際に述べたように本プレースメントテストの主な対象とはしていないため、大きな問題ではない。また、レベル 4 と 5 でも差が見られなかったが、これらのレベルの学生は本番時にはレベル 5 と 6 にそれぞれ進級する学生であるので、改訂前のテストでレベル 5 と 6 の間の差が見られなかったのと同様だとみなすことができる。そして、レベル 3 と 4 の間で差が見られたことについては、本番時ではレベル 4 と 5 の識別ということになり、今回の改訂の大きな目標であったため、その点に関してはクリアできたように思われる。

3.2.3 項目分析

新テストの結果についても本番データと同様に TDAP Ver. 2.0 により処理を行った。その結果を表 6 に示す。

表 6　新テストの項目分析データ

項目 NO.	上位・下位項目弁別指数（ULD）	点双列相関係数による項目弁別力指数 (DISC)	実質選択肢数 (AENO)	項目困難度 (DIFF)	項目困難度適切度 (ADIF)	項目困難度パラメータ (Final Calib.)	モデルとの適合度 (t)
1	0.761	0.590	2.622	0.579	0.909	-0.208	-2.859
2	0.587	0.541	2.048	0.724	0.803	-1.091	-3.026
3	0.674	0.572	2.745	0.476	0.703	0.328	-1.598
4	0.446	0.439	1.936	0.735	0.779	-1.195	-1.148
5	0.587	0.520	2.086	0.679	0.891	-0.779	-1.207
6	0.753	0.514	2.074	0.753	0.744	-1.305	-3.357
7	0.598	0.490	3.516	0.374	0.497	0.845	-0.216
8	0.609	0.490	2.043	0.629	0.991	-0.477	-0.411
9	0.565	0.502	2.484	0.303	0.356	1.366	-1.460
10	0.587	0.528	2.339	0.315	0.379	1.267	-2.582
11	0.413	0.367	1.801	0.750	0.750	-1.283	0.249
12	0.424	0.379	2.643	0.209	0.168	2.094	-0.168
13	0.022	0.067	2.139	0.018	-0.215	N/A	N/A
14	0.598	0.463	2.978	0.468	0.685	0.458	2.513
15	0.424	0.339	3.174	0.453	0.656	0.442	3.906
16	0.326	0.305	1.846	0.794	0.662	-1.668	0.028
17	0.522	0.446	2.051	0.715	0.821	-1.030	-1.990
18	0.663	0.523	2.673	0.456	0.662	0.491	-0.952
19	0.652	0.574	2.797	0.403	0.556	0.776	-1.320
20	0.674	0.583	3.257	0.365	0.479	0.968	-2.478

3.2.3.1 項目弁別力指数（ULD と DISC）

ULD が 0.4 未満になったものは、項目 13（新規項目）と項目 16（旧テストの項目 7）となった。一方、DISC が 0.3 未満になったものは、項目 13 となった。

3.2.3.2 実質選択肢数（AENO）

項目 2（新規項目）、項目 4（旧の項目 2）、項目 5（旧の項目 3）、項目 6（旧の項目 4）、項目 8（新規）、項目 9（新規）、項目 10（新規）、項目 11（旧の項目 5）、項目 13、項目 16（旧の項目 7）、項目 17（旧の項目 8）において、AENO の値が 2.5 未満になっている。特に、項目 4、項目 11、項目 16 が 2.0 未満で非常に低いものとなっている。

3.2.3.3 項目困難度（DIFF）と項目困難度適切度（ADIF）

新規項目の項目 13 が、DIFF で最も低い値を示し最も難しい項目となっており（DIFF = 0.018）、また、ADIF も負の値を示しており、項目弁別力の低さと併せて考えても、この項目は改善が必要と思われる。

3.2.3.4 項目困難度パラメータ（Final Calib.）とモデルとの適合度（t）

新テストでは、「項目困難度パラメータ」をより正確に推定するため、TDAP で元のデータを処理した後に、ミスフィットとなる 2.000 を越える受験者を除外し、再計算を実施した。その結果、項目 13 に関しては、全受験者が不正解となり、受験者の能力を測定できない不適切な問題であることが判明した。また、再計算後も、ミスフィットとなった項目は、新規に作成した項目、項目 14（t = 2.513）、項目 15（t = 3.906）で、これらについても、改善の必要性が明らかとなった。

全体の難易度のバランスとしては、再計算の際に適用除外になった項目 13 を除けば、-2.000 – -1.000 以下の項目が 6 項目、-1.000 – 0.000 の項目が 3 項目、0.000 – 1.000 の項目が 7 項目、1.000 – 2.000 の項目が 2 項目、2.000 – 3.000 の項目が 1 項目となった（図 4）。

改訂前より、難しめの問題が増えたことは事実だが、まだ、-1.000 以下の問題が全体の 30%を占めており、これが、レベル 4 と 5 の差が有意なものとならなかったことと関連しているものと思われる。

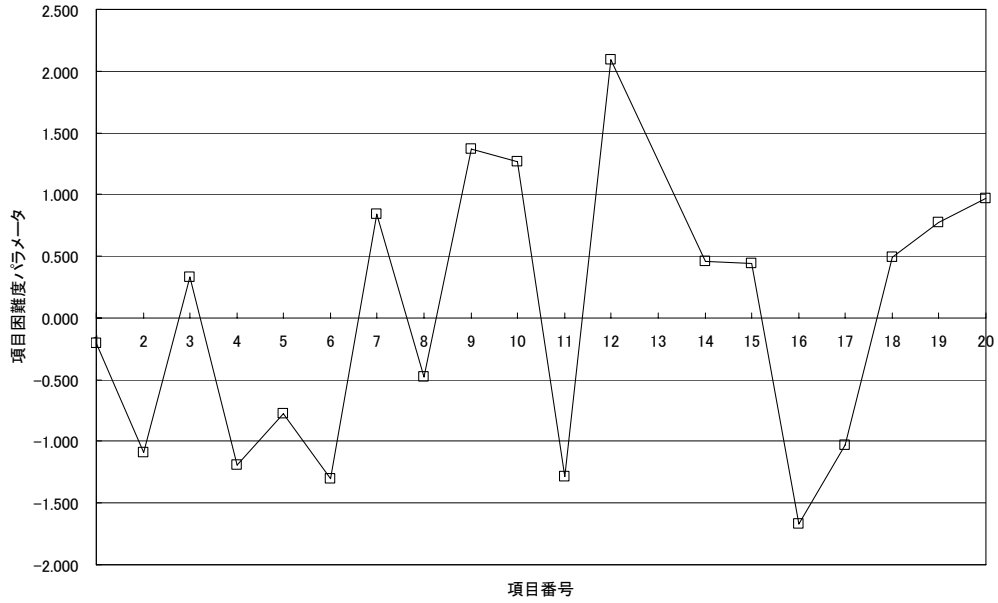

図４　項目困難度パラメータの分布

4. 今後の課題

本稿では、旧テストの結果と利用法を検証し、リスニングテストの基準点を機械的に適用してレベルを識別するには限界があるが、参考データとしては十分活用できることがわかった。そしてより精度をあげるべく問題の改善を行い、それを新テストとして試行し、問題点を洗い出してみた。今後の課題としては、洗い出された問題点にそって設問を吟味し、さらに各問題の質を高めることが求められている。

注

⑴ 配点は 1 問 4 点であるが、パスしているかどうかを明確にするため、基準点は 4 点の倍数から外している。

⑵ 日本語プログラムでの成績は以下のようになっている：A 93-100; A- 90-92; B+ 87-89; B 83-86; B- 80-82; C+ 77-79; C 73-76; C- 70-72; D+ 67-69; D 63-66; D- 60-62; F 0-59

⑶ 上級のレベル 5 はかなりの会話力が求められるので、二学期間の留学ということであれば今回の留学ではレベル 4 までの学習が適当だろうという風に、別の判断の可能性もある。

⑷ 学期当初のリスニングの点数がよくても最終成績が F になるケースも当然ある。

参考文献

大友賢二（1996）『項目応答理論入門』大修館書店

中村洋一（2004）『テストで言語能力は測れるか』桐原書店

宮内俊慈・坂井美恵子（2002）「統計分析に基づく漢字プレースメントテストの妥当性検討」『関西外国語大学留学生別科日本語教育論集』第 12 号 pp.67-82.

宮内俊慈・平田裕・小山揚子（2005）「プレースメント用リスニングテストの開発報告」『関西外国語大学留学生別科日本語教育論集』第 15 号 pp.87-106.

(smiyauc@kansaigaidai.ac.jp)

(hirata@kansaigaidai.ac.jp)

(koyama@kansaigaidai.ac.jp)

レベル別平均値の多重比較結果（旧テスト2005年秋本番）

レベル	レベル	平均値の差		標準誤差	有意確率
1.7	2.0	-4.44		1.773	0.125
	3.0	-11.76	*	1.734	0.000
	4.0	-16.04	*	1.777	0.000
	5.0	-31.23	*	3.301	0.000
	6.0	-36.82	*	3.661	0.000
2.0	1.7	4.44		1.773	0.125
	3.0	-7.32	*	1.374	0.000
	4.0	-11.59	*	1.427	0.000
	5.0	-26.79	*	3.127	0.000
	6.0	-32.37	*	3.504	0.000
3.0	1.7	11.76	*	1.734	0.000
	2.0	7.32	*	1.374	0.000
	4.0	-4.27	*	1.379	0.026
	5.0	-19.47	*	3.105	0.000
	6.0	-25.05	*	3.485	0.000
4.0	1.7	16.04	*	1.777	0.000
	2.0	11.59	*	1.427	0.000
	3.0	4.27	*	1.379	0.026
	5.0	-15.19	*	3.129	0.000
	6.0	-20.78	*	3.506	0.000
5.0	1.7	31.23	*	3.301	0.000
	2.0	26.79	*	3.127	0.000
	3.0	19.47	*	3.105	0.000
	4.0	15.19	*	3.129	0.000
	6.0	-5.59		4.476	0.812
6.0	1.7	36.82	*	3.661	0.000
	2.0	32.37	*	3.504	0.000
	3.0	25.05	*	3.485	0.000
	4.0	20.78	*	3.506	0.000
	5.0	5.59		4.476	0.812

レベル別平均値の多重比較結果（旧テスト2006年春本番）

レベル	レベル	平均値の差		標準誤差	有意確率
1.7	2.0	-5.06		2.233	0.215
	3.0	-10.57	*	2.203	0.000
	4.0	-20.26	*	2.413	0.000
	5.0	-30.20	*	3.695	0.000
	6.0	-41.40	*	4.207	0.000
2.0	1.7	5.06		2.233	0.215
	3.0	-5.51	*	1.660	0.014
	4.0	-15.21	*	1.931	0.000
	5.0	-25.14	*	3.400	0.000
	6.0	-36.34	*	3.950	0.000
3.0	1.7	10.57	*	2.203	0.000
	2.0	5.51	*	1.660	0.014
	4.0	-9.69	*	1.896	0.000
	5.0	-19.63	*	3.380	0.000
	6.0	-30.83	*	3.933	0.000
4.0	1.7	20.26	*	2.413	0.000
	2.0	15.21	*	1.931	0.000
	3.0	9.69	*	1.896	0.000
	5.0	-9.94		3.521	0.059
	6.0	-21.14	*	4.055	0.000
5.0	1.7	30.20	*	3.695	0.000
	2.0	25.14	*	3.400	0.000
	3.0	19.63	*	3.380	0.000
	4.0	9.94		3.521	0.059
	6.0	-11.20		4.927	0.211
6.0	1.7	41.40	*	4.207	0.000
	2.0	36.34	*	3.950	0.000
	3.0	30.83	*	3.933	0.000
	4.0	21.14	*	4.055	0.000
	5.0	11.20		4.927	0.211

レベル別平均値の多重比較結果（新テスト試行）

レベル	レベル	平均値の差		標準誤差	有意確率
1.0	1.7	-5.09		2.826	0.467
	2.0	-12.81	*	2.127	0.000
	3.0	-21.48	*	1.984	0.000
	4.0	-36.65	*	2.027	0.000
	5.0	-39.91	*	2.240	0.000
1.7	1.0	5.09		2.826	0.467
	2.0	-7.72	*	2.665	0.046
	3.0	-16.39	*	2.552	0.000
	4.0	-31.56	*	2.586	0.000
	5.0	-34.82	*	2.755	0.000
2.0	1.0	12.81	*	2.127	0.000
	1.7	7.72	*	2.665	0.046
	3.0	-8.67	*	1.747	0.000
	4.0	-23.84	*	1.795	0.000
	5.0	-27.10	*	2.032	0.000
3.0	1.0	21.48	*	1.984	0.000
	1.7	16.39	*	2.552	0.000
	2.0	8.67	*	1.747	0.000
	4.0	-15.17	*	1.624	0.000
	5.0	-18.43	*	1.882	0.000
4.0	1.0	36.65	*	2.027	0.000
	1.7	31.56	*	2.586	0.000
	2.0	23.84	*	1.795	0.000
	3.0	15.17	*	1.624	0.000
	5.0	-3.26		1.927	0.537
5.0	1.0	39.91	*	2.240	0.000
	1.7	34.82	*	2.755	0.000
	2.0	27.10	*	2.032	0.000
	3.0	18.43	*	1.882	0.000
	4.0	3.26		1.927	0.537