英語多肢選択テストの項目分析[1)]

水　野　康　一　（経済学部）

In order to develop an English proficiency test that can be used as a reliable measure of students' performance in English classes, the question items of a test used by the author (EP Test) were analyzed in terms of item difficulty, discrimination power, and actual number of options. Using the information gained by the item analysis, a revised test (EP-R Test), which consists of question items selected to suit students' proficiency levels and adopts a unique scoring system, was administered to students. The test scores produced by EP-R Test were more reliable and consistent with the scores of TOEIC Test. It was also found that the new scoring system, which utilizes the information of test takers' choice of options, constantly improve the reliability and correlation coefficients.

はじめに

大学教育のアカウンタビリティを高めるため，外国語科目においては到達目標の指標として外部英語能力試験の結果を採用しようとする動きが全国的に見られる[2)]。本学でもすでに実用英語能力検定（英検）やTOEFL，TOEICの成績を以って英語の単位認定を行っているが，今後，全学の英語クラスの合格基準に上記英語試験の得点を参照することについても検討が始まっている。取り組みの進んでいる大学では，すでにプレースメント（習熟度別クラス分け）や教育効果測定のために，TOEIC-IP（団体受験）を実施し，学生全員に受験させているところもあり，英語教育プログラムの評価にTOEICを中心とした外部英語能力テストの団体受験結果を持ち出してくる大学も今後さらに増加することは間違いないであろう。

一方でTOEICなどの外部英語能力テストを全学生に受けさせるには，解決すべき実際的な問題も多い。公開テストで6,300円，IPテストで3,850円という受験料の負担問題，IPテストを学内で一斉実施してプレースメントを行おうとする場合，過密な年度当初のスケジュールにおいて3時間近いテスト時間を確保し，教室や監督者を手配する困難さなど，実施に向けては綿密な議論を重ねる必要があるであろう。もし，正式な成績証明を必要とせず，英語授業におけるプレースメント，教育効果測定，および成績評価にスコアを用いようとするのであれば，TOEIC Bridgeテスト（TOEICスコア450点以下の学習者のために作られた縮小改訂版TOEIC）などを利用することもできるが，コスト（受験料負担）の問題は依然残る。外部試験の「平行テスト（模擬テスト）」を独自に作成するという方法も考えられるが，TOEFL，TOEICに関しては，項目応答理論（IRT）に基づいた特殊なスコア計算方法が一切公開されていないため，独自テストでは信頼性，

再現性の高いスコア算出方法も開発せねばならない。

以上のようなことから，本論では受験者の英語能力を正確に測定でき（信頼性の高い），TOEICなどの外部標準テストのスコアと出来るだけ相関が高く（基準連関妥当性の高い），項目数を減らして実施時間を短くした（実用性の高い）英語能力テストの開発可能性を検討する。信頼性，妥当性，実用性の要求を同時に満たすために，筆者が作成し現在まで用いてきた英語能力テストの項目分析を行い，その結果から採点および得点集計について新たな試案を提出したい。

1．英語能力（EP）テスト

1. 1. テストの概要

英語授業における指導において，学習者の英語能力レベルをある程度把握しておくことは重要である。筆者は教養英語の初回授業において，授業の概要を説明した後に残る時間を利用して，受講者に対して筆者が用意した英語能力テスト（以後EPテストと称す）を過去10年以上にわたって実施してきた。テスト実施の目的は，クラス全体の英語能力レベルを把握することのほか，英語能力レベルが高すぎる，あるいは低すぎるなど，一斉授業ではケアが困難な受講生をあらかじめ特定することである。授業ガイダンスの残りという限られた時間に実施することから，テストの内容は多肢選択式の文完成問題（multiple-choice sentence completion task）30項目のみに限定している。英語能力の中でも主として文法力や語彙力を測定するこのテスト形式を採用しているのは，筆者の授業が英文読解を中心的活動としているためである。多肢選択式にしている理由は，より多くの項目をテストに含めることができるからである。筆者の担当する学習者の幅広い英語能力レベルに対応するため，問題は英検２級から３級，およびTOEICの問題集からそれぞれ10問ずつ選び，それらを編集したものである[3)]。制限時間は25分と設定しているが，これは受験者のほぼ全員が解答を終える時間である。

表１は学部，学年，入学年度，コースの異なるいくつかのクラスで過去に実施したEPテストの結果をまとめて集計したものである。

表１　EPテストの基本統計量（$n=246$）

項目数	30
平均正答数（%）	17.69（58.97）
標準偏差	4.22
分散	17.80
最高値	25
最低値	6
中央値	18
最頻値	18
Spearman Brown（折半法）	0.695
KR-20	0.698
英検３級問題の正答率	0.811
英検２級問題の正答率	0.450
TOEIC問題の正答率	0.480
最易項目の正答率	0.943
最難項目の正答率	0.211

一部のクラスで実施したTOEICスコア（Total Score）との相関係数を求めてみたところ，0.382（標本数44）であった。EPテストと同形式のTOEIC Reading Section Part V（40問）との相関も0.438に留まったことから，EPテストには改善余地が大きいことが分かった。

1. 2. 信頼性

まず，EPテストの信頼性の問題について考える。信頼性はテストスコアの再現性を示すものである。信頼性は高い（信頼性係数が1に近い）のが理想であるが，今回のEPテストでは，Spearman Brown 信頼性係数，KR-20信頼性係数のいずれも0.7程度であった（表1）。信頼性係数はテスト項目数を増やすことによって大幅に改善できるが，PEテストは実用性という観点から試験時間を限定しているので，問題形式を変更しない限り，項目数を大幅に増やすことは難しい。ところがもし外部基準テストで採用されていない形式のテスト問題を採用すると，基準テストとの相関が損なわれる危険が大きく，またその場合，困難な構成概念妥当性の検証が必要となってくる。したがって，今回はEPテストの改良にあたり，項目数，問題形式を変更しないことを大前提とし，項目の内容についての検討（項目分析）を中心に進めていく。

1. 3. 項目難易度

項目数以外にテストの信頼性を下げる大きな要因は項目難易度である。受験者の英語能力レベルに対して項目難易度が極端に低い場合は，受験者のほぼ全員が正解してしまうため，そのテスト項目は弁別力を持たない。すなわちテスト項目として最初から含まれていなかったのと同じことであり，実質的なテスト項目数はその分だけ少ないこととなる。項目難易度が極端に高い場合も，逆の意味で同じことであるが，とりわけ多肢選択テストでは，当て推量で正解を得られる可能性があるので，そのような項目では英語能力が低いグループが高いグループよりも正答率が高くなるケースもある。このような項目は信頼性にとって大きな攪乱要素である。

EPテストについて考えると，特に英検3級の問題を中心に正答率が高く，正答率が90%を超える項目も4問含まれていた。また，逆に最も難易度の高い項目では正答率が21.1%であった。古典的テスト理論による四者択一問題の理想的な項目困難度は0.625（項目正答率62.5%）であるため，EPテストには受験者のレベルの合わないテスト項目が多く含まれていたことが判明した。図1はEPテスト成績の度数分布をグラフにしたものであるが，75%のデータが15点から23点の比較的狭い範囲に分布しており，このことが弁別力の高い項目が少なく，テストの実質項目数が限られていたことを表している。

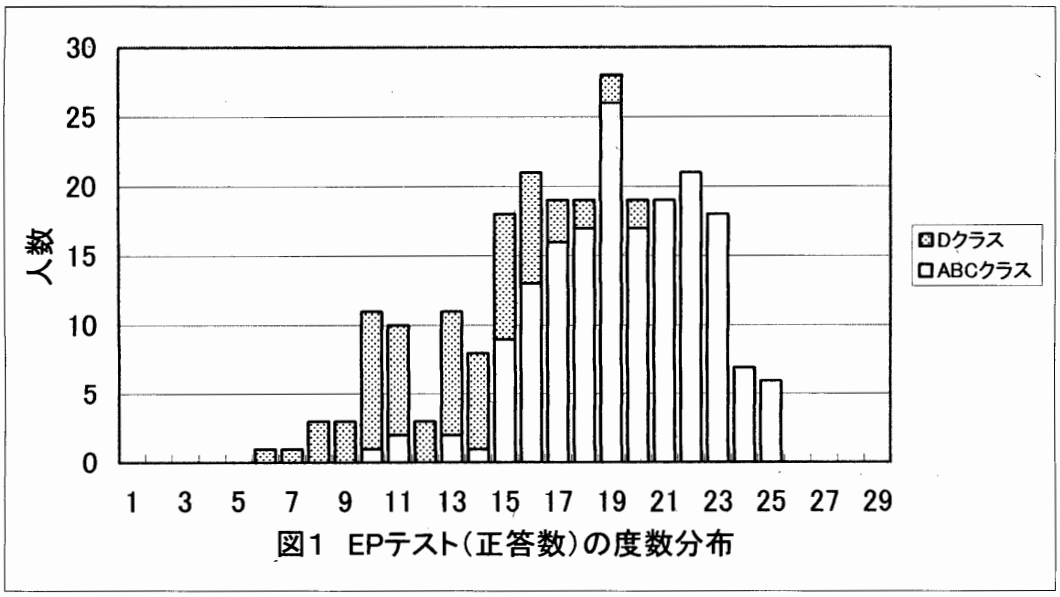

図1 EPテスト(正答数)の度数分布

ところで，受験生の英語能力レベルに問題を移して図1を見ると，度数分布が全体的にいびつな形になっていることに気づく。データ集計に用いた4クラスは，A）社会科学系学部教養課程(2003年度入学)，B）社会科学系学部専門課程（2000-2002年度入学)，C）文系学部（合同）教養課程（1997年入学，テストは1997年に実施)，D）社会科学系学部の社会人コース（2002-2003年入学）であった。学部，学年，入学年度，コースの異なるクラスのデータを抽出したのは，クラス集団間の英語能力レベル差を検討する目的のほか，幅広い受験者の英語能力レベルに対してEPテストがどれほどの英語能力判定能力を有しているかを検証するためである。クラスごとの集計データをまとめたものが表2である。

表2 EPテストのクラス別統計量

クラス	A	B	C	D
人数	44	61	69	72
平均正答数	20.36	19.23	19.39	13.11
標準偏差	2.93	2.85	3.02	3.17
分散	5.87	8.11	9.12	10.07
最高値	25	25	25	20
最低値	11	10	11	6

一元配置分散分析および多重比較の結果，DクラスはA，B，Cのいずれのクラスとも平均値の差に有意差が認められ（$p<.001$)，A，B，Cクラスの間にはどの組み合わせにも有意差は見られなかった。Dクラスは社会人を対象としたコースで，高等学校卒業以来，何年も英語学習から遠ざかっていたという人が多く，大学教養課程の英語授業には困難を感じていると漏らす学生も多い。EPテストは本来このような就学上の困難を抱える学生を早期に見つけ出し，指導上の適切な対応をとるためのものであった。実際，Dクラスの分布（図1の棒グラフのやや暗い部分）をA～Cのクラスの昼間学部生の分布（白い部分）と比較すると，15点未満の学生は大学教養レベル

の授業に困難をおぼえるであろうということが容易に判断できるのである。

EPテストのいびつな成績分布は，大きく英語能力が異なる2つのグループ（クラス）のデータを混ぜ合わせた結果であることが，図1より視覚的に見て取れる。クラス内の英語能力レベルは接近しているので，それぞれのクラスのレベルに応じたテストを用意すれば，1つのテストに分別力の高い有効項目をより多く含ませることができ，結果的に信頼性の高いテストが作れるはずである。本学の文系学部については学部や学年，および入学年度に関わらず，EPテストの結果については統計的有意差が認められなかったことから，本研究においてはとりあえず同レベルのクラスとして扱い，まずA～Cクラスを対象とした英語能力テストの開発を試みる。

2．EPテストの項目分析

2. 1. 項目応答理論と古典的テスト理論

受験者のレベルを考慮したテストの開発にあたっては，初段階において項目応答理論による方法と古典的テスト理論による方法のいずれかを選択をすることができる。項目応答理論に基づくテストの利点は，受験者および母集団の能力レベルに関わらず，正確な能力測定が行えることである[4]。偏差値など，従来の標準化された得点を用いる方法は，テスト受験生全体のレベルが変化すると正確な能力判定は不可能になるが，この理論を採用すると複数のテストでたとえ受験生の母集団が変わっても，安定した能力スコアを得ることが可能になる。これはTOEICのようなスコアの信頼性を要求される大規模な英語能力検定テストに向いており，そして実際に採用されているテスト理論である。もう一つのいわゆる古典的テスト理論とは，項目応答理論登場以前のテスト方法論全体を指す。本論では後者の古典的テスト理論による項目分析および能力測定（スコア変換）を行うが，その理由は前項で示したように，テスト開発の対象とする受験生（本学の文系学部生）の英語能力レベルが比較的接近しており母集団のレベルが変わらないこと，同じ受験者が何度も受験する異なる複数のテストの開発を考えていないこと，そして最大の理由は項目応答理論の理解や複雑な計算が必要とされず，本論の結果が広く実践者に利用されやすいこと，である。

2. 2. 項目困難度，項目弁別力，実質的選択肢数

いわゆる古典的テスト理論による項目分析に使われる指標である項目困難度，項目弁別力，実質的選択肢数を項目ごとにまとめたものが表3である。なお，項目1から10が英検3級レベル，項目11から20までが英検2級レベル，項目21から30がTOEIC問題であった。

表3　EPテストの項目分析結果

項目	項目困難度	項目弁別力	実質的選択肢数
1	0.927	0.379	1.362
2	0.943	0.344	1.314
3	0.854	0.520	1.719
4	0.911	0.176	1.482
5	0.923	0.357	1.423
6	0.890	0.424	1.586
7	0.793	0.504	1.972
8	0.687	0.515	2.197
9	0.878	0.414	1.573
10	0.435	0.510	2.985
11	0.467	0.313	3.530
12	0.211	0.340	3.669
13	0.711	0.610	2.425
14	0.419	0.334	3.224
15	0.402	0.291	3.554
16	0.589	0.375	3.058
17	0.382	0.263	3.724
18	0.398	0.413	3.366
19	0.553	0.280	3.218
20	0.435	0.114	3.652
21	0.516	0.354	2.719
22	0.789	0.313	2.055
23	0.675	0.315	2.649
24	0.354	0.287	3.710
25	0.610	0.367	2.886
26	0.305	−0.032	3.473
27	0.297	0.111	3.833
28	0.419	0.317	3.507
29	0.549	0.405	2.913
30	0.366	−0.044	3.545

項目困難度，実質的選択肢数には強い相関関係（$r=0.969$）が見られた。項目困難度が上がるほど（数値が小さいほど）実質的選択肢数が増えているということは，すなわち受験者の解答が均等にばらついているということである。この相関関係には例外的な傾向を示す項目が見当たらず，このことはテスト項目に準備された正答以外の選択肢（錯乱肢）が極めて効果的に機能していたということを表している。ゆえに個々の項目の適切さを求めるにあたっては，項目困難度と項目弁別力から判断すればよいことになる。

前述したように項目困難度は理想値0.625との差の絶対値が小さいほど，受験生の解答から多くの情報が得られる事になり，テスト項目としては優れている。適切な困難度の項目は弁別力も強いと仮定すれば，項目の困難度と弁別力の分布図を描くと，項目弁別力は項目困難度0.625をピークとして山型の弧に近くなるはずである。実際の分布（図2）ではこの傾向から外れた項目がいくつか見られた。そこでそれらの項目について個別に検討してみる。

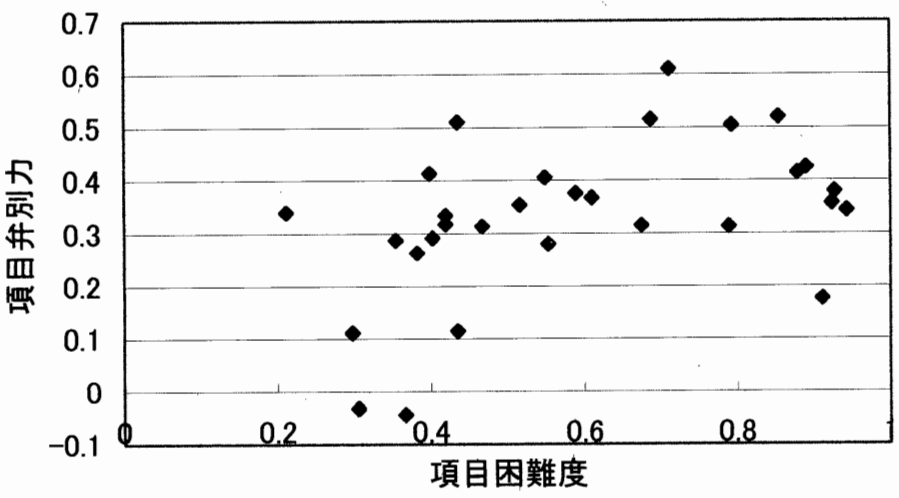

図2 項目困難度と項目弁別力の相関関係

2. 3. 項目弁別力がマイナス値である項目

(26) Deborah could not ( ) from laughing at the funny way the hotel clerk pronounced her last name.

1. keep 2. prevent 3. avoid 4. refuse

(30) The convenience store was completely ( ) of apple juice, so I bought some orange juice instead.

1. gone 2. empty 3. short 4. out

項目弁別力指数がマイナス値になるということは，能力レベルと正答率が負の相関関係にあるということである。すなわち，能力の高い人のほうが，低い人よりも正答率が低かったということを意味する。項目困難度があまりに高かったため，当て推量で解答を選んだ下位群の方が確率的に正解率が高かったという多肢選択テスト特有の効果によるものと考えられるが，実際に確認のためテスト受験者全員を能力別にグループ分けし，正答率および解答パターンを調べてみる。グループ分けは便宜的にEPテストのスコア20点以上を上位群（n=90），15点以上19点以下を中位群(n=105)，14点以下を下位群（n=51）とした。図1から上位群は文系学部クラス上位群，中位群は文系学部クラス下位群に相当すると考えられる。

表4は各能力グループ別に見た項目26，30の解答選択率である。英語能力下位群の正答率が上位群より高いことが確認できるが，下位群の正答率がいずれも40%近いことから，下位群が全くの当て推量で4つの選択肢のうち一つを選んだわけではないことが分かる。むしろ英語能力上位群において正答よりも好まれる攪乱肢（項目26のprevent，項目30のshort）があり，それによっ

て正答を選んだ人が相対的に少なくなったことが原因であろう。例えば項目30において，上位群の多くはshort of（不足した）を知っていたが，out of（売り切れた）の意味を知らなかった，あるいは誤解していたことが分かる。一方，下位群ではshort ofの用法を知らない者が多く，shortはその「短い」という意味から，早い段階で解答候補から省かれてしまったのであろう。もし仮に項目30において選択肢3も正答と処理していれば，一転して非常に弁別力の優れた項目となったはずである。

表4　能力レベル別解答選択率（項目26，30）

項目	選択肢	上位群	中位群	下位群
26	1（正答）	26.7	29.5	39.2
	2	46.7	39.0	31.4
	3	24.4	23.8	15.7
	4	2.2	7.6	13.7
30	1	6.7	7.62	9.8
	2	11.1	23.8	43.1
	3	45.6	10.5	3.9
	4（正答）	36.7	33.3	43.1

2. 4. 困難度が高いにもかかわらず分別力も高い項目

(12) The number of students who were late for school got (　　) during examination week.

1. fewer　2. slighter　3. lesser　4. smaller

表5　能力レベル別解答選択率（項目12）

項目	選択肢	上位群	中位群	下位群
12	1	35.6	47.6	27.5
	2	5.6	7.6	31.4
	3	17.8	4.8	35.3
	4（正答）	41.1	12.4	3.9

項目困難度が高くなればなるほど能力レベルの低い受験生は当て推量による正答が増え，項目分別力も下がるのが通常だとすると，項目12は全くの例外である。解答選択率を見ると成績下位群が正答を真っ先に解答候補から外していることが分かり，設問および選択肢設定において非常に優れた良問だったといえる。ただし，本研究の対象となる受験生レベルには項目難易度が高すぎるので，項目分析によるテスト改良では除外候補となってしまう。もし，選択肢1（fewer）も正答と処理すると，項目弁別力は若干下がるものの（0.34→0.28），項目困難度は大幅に上がり（0.211→0.595），本研究で対象とする受験生のレベルに適応した項目となる。

3．EPテストの改善

3. 1. 多肢選択テストの問題

多肢選択式テストは正答が１つ設定され，それ以外の解答は誤答と処理されるのが通常である。EPテストの項目12のfewerや項目30のshort（いずれも誤答）を選択する受験者は，その他の誤答を選択する受験者よりも英語能力が高いことが統計的にも，経験的にも分かっているのだが，誤答の持つ情報はスコアに反映されることはない。しかし，特に本研究のように，項目数を増やさずに信頼性の高い多肢選択テストを開発しようとする状況で，答案の持っている情報の一部を捨てることはできれば避けたいものである。多肢選択試験における誤答は所詮誤答以外の何物でもないので，それらを正答と同様に扱うことには何の理論的根拠も妥当性も見当たらない（ゆえに理論家や実践者の賛同は得られにくいと思われる）が，限られた情報からできるだけ受験者の能力を正確に測るためには，受験者の解答が項目ごとの正答／誤答という2値データに変換される過程で失われる情報は捨てがたい価値を持つものである。

3. 2. 誤答情報の利用（試案）

ではこのような多肢選択式テストの誤答が持つ情報は，能力レベル判定に対してどれほどの重みを持っているのであろうか。このことをEPテストで検証するために，次の手順に従いテストの信頼性係数および妥当性係数の再計算を行ってみた。

① 正答以外で上位群が最も多く選んだ選択肢（誤答）も正答とみなした場合の項目困難度，項目分別力を再計算。
② 項目困難度が理想値（0.625）に近づき，かつ項目弁別力がプラス値となる項目を選択。
③ 選ばれた項目中，①で正答とみなした選択肢（誤答）一つにつき，0.5ポイントをスコアに加算。

EPテストの場合，①で選ばれた項目数は10（項目11, 12, 14, 15, 17, 18, 20, 24, 26, 30）であった。誤答加算後の修正スコアを再計算した結果，信頼性係数および妥当性係数の両方で改善が確認できた（表６）。誤答情報の利用方法については改善の余地があり，最適な方法が見つかれば信頼性，妥当性はさらに向上が見込める可能性もある。

表６　EPテストの信頼性・基準関連妥当性

検証内容	方　　法	スコア修正前	スコア修正後
信頼性	Spearman Brown（折半法）	0.695	0.751
基準関連妥当性（相関係数）ただし$n=44$	TOEIC　Section V	0.438	0.483
	TOEIC　Reading Section	0.478	0.509
	TOEIC　Listening Section	0.091	0.131
	TOEIC　Total	0.382	0.422

上記のようなスコア修正手順においてどの項目のどの選択肢（誤答）について加点を行うかは，すでにあるテストデータに基づき事後に検討したアポステリオリな情報であるが，受験者の母集団が変わってもグループ間の能力分布が近似している場合は，同じ情報をアプリオリに適用しても，信頼性向上の効果が得られるはずである。以下ではこの仮説の検証も含めて，作成したEPテストの全面改訂版（EP－Rテスト）について報告する。

4．改訂版EPテスト（EP-Rテスト）

EPテストはすでに10年近く使用してきたために，一部の文章がすでに時代遅れな内容になってしまっていること（fax machines have greatly accelerated…など）に加えて，本調査で外部基準テスト（TOEIC）との相関がそれほど高くないこと（r=0.4〜0.5）が判明したため，今回全面的に改訂を行った。本研究から得られた情報を基に，EP-Rテストは，1）英語能力レベルが接近している文系学部学生専用に用いる目的で，彼らのレベルに合わせて最適化する，2）外部基準テストとの相関，すなわち基準関連妥当性を高めるため，全問TOEIC問題を採用する，3）実用性を犠牲にしないため，試験形式は多肢選択式分完成問題，選択項目数30のままとする，4）テストの信頼性を向上させるため，誤答情報をアプリオリに適用した修正スコアを利用する，といった方針で作成した。具体的には，Bクラス（TOEIC演習）で使用したTOEIC模試および期末試験から得られていた文完成問題（TOEIC Reading Test, Part V）50問のデータについて項目分析を行い，困難度0.3〜0.9の項目の中から，弁別力上位30問を選択し，項目困難度の低い問題順に並べて編集した。前述の誤答情報処理手順によって，誤答加点（0.5）を行う項目および選択肢を決定（7項目が該当）し，表計算ソフトで採点用テンプレートも用意した。このようにして作成したテスト（EP-Rテスト）を2003年12月にAクラスの受講生を対象に実施した結果が表7，表8である。

表7　EP-Rテストの基本統計量（n=44）

	スコア修正前	スコア修正後
平均値	17.115	18.841
標準偏差	4.07	3.85
分散	16.58	14.82
最高値	26	26.5
最低値	8	9.0
中央値	18	19.5
最頻値	17	17.0
最易項目の正答率	0.818	0.818
最難項目の正答率	0.341	0.341

表８　EP-Rテストの信頼性・基準関連妥当性

検証内容	方法	スコア修正前	スコア修正後
信頼性	Spearman Brown（折半法）	0.651	0.684
基準関連妥当性（相関係数）ただしn=41	TOEIC　Section V	0.610	0.631
	TOEIC　Reading Section	0.640	0.652
	TOEIC　Listening Section	0.206	0.260
	TOEIC　Total	0.549	0.582

テスト改訂前後の信頼性係数の変化について調べるために，改訂前のEPテストのAクラスだけ（n＝44）の信頼性係数を求めたところ，スコア修正前と修正後でそれぞれ0.376，0.431であった。EP-Rの信頼性係数は0.651，0.684（表８）であるので，かなりの改善が見られたといえる。テスト項目のレベル（困難度，弁別力）調整により，実質的項目数が増えたことが主な理由であると考えられる。受験者の標本数がさらに増えれば，表６（n＝246）の0.695，0.751を上回ることはほぼ間違いないと思われる。

外部基準テストとの相関（基準関連妥当性）についても，改訂版のテストにおいて係数の上昇が見られた。すべての項目をTOEICテストの問題に差し替えた効果もあるかもしれないが，改訂前のテストと基本的な出題形式が大きく変わっているわけではないことから，テスト自体の信頼性の向上が，TOEICとの相関にも表れていると考えられる。統計学的には２つのテストスコアの相関係数はそれぞれのテストの信頼性の積の平方根を超えることがないということが分かっている。すなわちテストの信頼性が高くなければ，テスト間の高い相関係数は得られないということである。EP-Rテストが外部基準テストの結果を予測する簡便な手法となるためには，さらに一層の信頼性向上を目指す必要がある。

誤答情報を利用した得点修正の効果については，設定した項目数（7）や修正割合（0.5）が少なかったものの，信頼性係数，基準妥当性係数のいずれにおいても向上という形で確認された。項目数や修正割合の変更，さらに今回は試していない減点法や加重方式など，誤答情報をスコアに反映させる最適な方法が見つかれば，さらに信頼性向上が期待できる。この点で受験者の誤答分析およびそのデータ利用は今後も検討する価値が十分にあると思われる。

５．結　語

テストの信頼性向上のために，本論で提案した誤答情報の利用法（加点方式）は，その妥当性に理論的基盤を持たない，あくまで事後の統計処理手法であるが，EP-Rテストでも，信頼性，基準関連妥当性の改善に貢献していることが確認できた。特に今回は，事前に蓄積された情報を新しい受験者に対してアプリオリに使用してもその効果が見られたということは大きな収穫である。同じ能力レベルのグループに対してのみ有効という条件はあるが，手続きや計算の容易さ，理解のしやすさでは，項目反応理論によるスコア算出より優れているからである。多肢選択式英語能力テストにおける誤答情報の利用についてはまだ改良の余地が大きいが，さらなるデータ収集やテスト開発を通して洗練されたシステムにしていくことを今後の課題としたい。

【注】

1）本研究は香川大学経済学部地域社会システム学科新特別研究費の援助を受けて実施したものである。

2）4年制大学および短期大学における習熟度別クラス編成，到達目標設定のための英語統一テストの全国的な実態調査は杉森（2003）に詳しく報告されている。

3）本研究で用いたEPテストおよびEP-Rテストはいずれも市販教材の項目をそのまま利用しているため，本稿ではテスト問題の転載は差し控えさせていただく。なお，テストの詳細についての問い合わせは筆者まで（mizuno@ec.kagawa-u.ac.jp）。

4）項目応答理論の詳しい解説については大友（1996）を参照のこと。さらに入門者向けに解説した参考書には高橋（2002）がある。

【参考文献】

Brown, J. D. (1996). *Testing in Language Programs*. Upper Saddle River, NJ: Prentice Hall.

大友賢二.（1996).『項目応答理論入門』大修館書店.

清水裕子，木村真治・杉野直樹，山川健一，大場浩正，中野美知子.（2003).「英語文法能力標準テストの妥当性・信頼性の検証と新文法能力テストMeasure of English Grammar（MEG)」『政策科学』10巻3号，pp.59-68.

杉森幹彦.（2003).「英語統一テスト・習熟度別クラス編成・到達目標の設定および測定に関する実態調査の報告」『政策科学』10巻3号，pp.3-26.

高橋正視.（2002).『項目応答理論入門～新しい絶対評価』イデア出版局.

TOEIC運営委員会.（2000).『TOEIC公式ガイド&問題集』（財）国際ビジネスコミュニケーション協会.