英語プログラムとプレイスメント・テスト

——2種のテスト結果の分析をもとに——

清　水　裕　子

Ⅰ．は　じ　め　に

教育に関わる者は何らかの形で測定と評価活動に直面しなければならない。最も直接的なものは，授業を通しての担当教員による学習者への評価であるが，さらに，あるプログラム（例えば英語プログラム），学部教育，大学全体というように，より大きな観点からも評価は関わりをもっている。どの段階においても，その主たる関心は，指導効果であり，学習者にどのようなレディネスが備わっていて，いかに学ぶことができるか，そして，いかに学ぶことができたかということにある。

立命館大学の経済・経営両学部のBKC新展開のひとつに，外国語教育の抜本的改革が掲げられ，英語力別のクラス編成を行い，運用能力を育成するための教授-学習システムのもとに初年度がスタートした。旧態依然とした英語の授業が行われている大学が多い中で，本学経済・経営学部のカリキュラムは画期的なものと言える。新カリキュラムの英語プログラム（以下，本稿では経済・経営学部の英語カリキュラムを「英語プログラム」とする）の効果を知るには，学生への質問紙調査や面接を通じて授業に対する意見を調査するのも一方法である。実際に，言語コミュニケーションセンターを通じて質問紙調査が行われたり，教育科学研究所主催のもとに学生からの意見を聴く機会も得ており[1)]，これらについては，外国語教育におけるFD研究プロジェクト・ニューズレターNo. 6において報告されている。

本稿では，入学した学生が英語プログラムの中で関わる言語テストの中からプレイスメント・テストを取り上げ，筆者が手にすることができたデータをもとに，その整合性を考えていく。英語プログラムにおける言語テストや評価のあり方について検討すべき課題は多いが，本稿が説得力のある教育改革につながるためのデータのひとつになることを期待する。

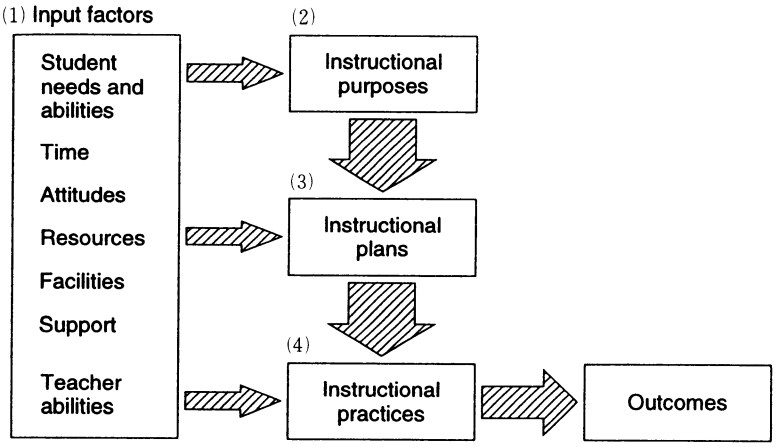
具体的には，まず，教授-学習過程における評価の観点から，テスト法（testing）を概観しながらプレイスメント・テストに焦点を当てていく。次に，プレイスメント・テストの実施状況に関して，本学経済・経営学部の状況を示した後，今年度（1998年）4月及び6月に実施されたテスト結果をもとに分析を行い，英語プログラムにおけるテスト法の今後の検討課題を考察する。

Ⅱ．語学プログラムにおけるテスト法

1．教授-学習過程での評価

評価活動の基本は教室内での活動にある。Genesee & Upshur（1996：15）は次のような図を示し，教授-学習活動における評価を説明している。

図1 The context of classroom-based evaluation



from Genesee & Upshur（1996）（ただし，番号は筆者による）

(1) 学習者要因（学習者のニーズや学力）や学習環境要因（時間，設備等）に応じて，(2)指導目的（why），(3)指導計画（how），(4)実際の授業（what）が展開し，この4つのinput factorsの結果，その所産であるoutcomesに至るわけである。Outcomesは，ことばの習得の形や言語に対する態度変化，文化に対する意識として学習者の中にでてくる。これを我々の英語プログラムに当てはめてみる。(1)の中の学習者要因に関しては，本学の入学試験によって選抜され入学してきた学習者が対象であり，英語力に関しては4月の段階で独自のプレイスメント・テストを実施し，5段階のレベル分けがされている。また，学習環境要因としては，使用教室や設備，授業時間数，担当教員等が決定されている。また，(2)，(3)に関しても，新カリキュラムの準備段階で明確な目標や指導計画，方法が打ち出されている[2)]。そして決められたシラバスのもとに(4)実際の授業が展開され，outcomesとして内的には学習者の中に学習効果が定着し，外的には一つの評価の形として各学習者に履修科目の成績が与えられる。また，TOEFL-ITP（Test of English as a Foreign Language-Institutional Testing Program）が第1セメスター終了直前に実施され，学習者の英語力に関する情報を得ると共に，この得点を後のクラス編成のための資料として活用することになっている。以上の過程には，入学試験，プレイスメント・テスト，TOEFL-ITP，授業内での評価活動という複数の言語テスト（この場合は全て英語）が関わり，種々の決定を行っていることになる。

ところで，Brown（1996）が，

…… the decisions made by administrators and teachers all affect students' lives, some-

times in dramatic ways, involving a great deal of time and money, and at other times in more subtle ways, including psychological and attitudinal factors.

と示しているように，テストによる決定は学習者に多大な影響を与えることが多い。さらに，彼は決定のタイプにより次のようにテストを分類している。（1995：40；1996：1-15）ひとつは，プレイスメント・テストや熟達度テスト（Proficiency test）のように〈プログラム・レベルで決定を行うテスト〉で，もう一つは，シラバスに基づいて作成される到達度テスト（Achievement test）や診断テストのような〈教室レベルでの決定を行うテスト〉である。[3] また，結果が解釈される枠組から捉えると，前者は集団規準準拠テスト（Norm-referenced test, 以下，NRT とする），後者は目標基準準拠テスト（Criterion-referenced test, 以下，CRT とする）であり，本稿で中心に扱っているプレイスメント・テストは NRT のひとつとみなされる。では，そのプレイスメント・テストとはどのようなテストで，実際にどのような下位テストが使用されているのであろう。

2．語学プログラムにおけるプレイスメント・テスト

定　義

ここに，プレイスメント・テストのいくつかの定義がある。

A placement test is designed to sort new students into teaching groups, so that they can start a course at approximately the same level as to other students in the class. It is concerned with the student's present standing, and so relates to general ability rather than specific points of learning.（Andrew Harrison 1983：4）

Placement tests are intended to provide information which will help to place students at the stage（or in the part）of the teaching programme most appropriate to their abilities.（Arthur Huges 1989：14）

Placement tests are designed to assess students' levels of language ability so that they can be placed in the appropriate course or class.（Alderson, Clapham & Wall 1995：292）

これらの定義に共通することは，英語力の上で homogeneity をもつグループを形成するための〈general ability〉の測定が目的になっているということである。ここで再び Brown の分類に

表1　熟達度テストとプレイスメント・テストの比較（Brown 1996）

Test Qualities	Proficiency	Placement
Detail of Information	Very general	General
Focus	Uually, general skills prerequisite to entry	Learning points all levels and skills of program
Purpose of Decision	To compare individual with other groups/ individuals	To find each student's appropriate level
Relationship of Program	Comparisons with other institutions	Comparisons within program
When Administred	Before and sometimes at exit	Beginning of program
Interpretation of Scores	Spread of scores	Spread of scores

戻る。〈General ability〉を測定するテストとして，彼は同じNRTの中に熟達度テストとプレイスメント・テストをあげ，両テストを比較して表1のようにまとめている。（1996：9 Table 1.2より一部抜粋）

〈General ability〉を測るためにどのような情報を提供してくれるテストを用いるかについては，'General'をより大きな枠組みからみたものが熟達度テストであり，あるプログラムや状況との関連性を持たせたものがプレイスメント・テストになる。しかし，明確な境界線は存在しないし，後者の解釈も，実際には次に示すように大きく二つのアプローチに別れるようだ。

アプローチ1

Huges（1989：14）は，プレイスメント・テストというものは，特定の状況に応じて作成されたものが最適であるとし，彼のことばを借りると，既成ではなくtailor-madeであること，つまり，in-houseのテストであることが望ましいとしている。

プレイスメント・テストとしては，プログラムの指導内容と深く関連性をもたせなければならないという立場は，ESP（English for Specific Purpose）の分野で強い。ESPの研究分野では，特に工学系におけるwriting指導に関するニーズ分析が1990年代から積極的に行われてきており，学部での作文指導と卒業後に必要となる作文力との違いから，プレイスメント・テストや学習のあり方の問題が指摘されているが（Winsor 1990；Jenkins, Jordan & Weiland 1992），ここでは，ESPではなく，一般的なESL（English as a Second Language）のプログラムにおけるプレイスメント・テストの具体例を二つ紹介しておく。

まず，カリフォルニア大学ロサンゼルス校では，ESLPE（The English as a Second Language Placement Examination）という独自のテスト・バッテリーを開発し，ESLコースのクラス編成に用いている。この開発に関わっているJulie Thornton氏（1998）によると，ESLPEの開発では，「プレイスメント・テストは特定の状況や指導目的に応じて作成すべきだ」と言う立場にたっており，ESLコースの目的と内容に合致した下位テストの構成および難易度になっている。現行のテストは，作文，聞き取り，読解の3種の下位テストからなり，テスト時間はそれぞれ1時間，合計3時間である。聞き取りテストはEAP（English for Academic Purpose）の観点から，講義形式の英文をノートをとりながら聴いた後に多肢選択形式の問題と真偽問題に答える形をとっている。読解テストも多肢選択形式を用いており，長年用いられていたクローズ・テストは，現在では使用されなくなっている。

もうひとつの例としてハワイ大学マノア校のELI（The English Language Institute）の例を紹介する。このプログラムでは，先に示したBrownの熟達度テストとプレイスメント・テストの区別を明確にしており（表1参照），前者は一般的な英語力を念頭に置き，広範にわたる言語能力を測定するものとし，後者は与えられた言語プログラムの関連から，より狭い範囲の言語能力を測定することを目的とするべきだとしている。つまり，熟達度テストによってどのプログラムが適切か，あるいは，特定のプログラムが適しているか否かを決定し，次の段階でプレイスメント・テストを実施し，プログラム内のどのレベルに学習者を配置するかを決めている。なお，ハワイ大学では，熟達度テストとしてTOEFLを，プレイスメント・テストとしては独自に開発したテストを用いている。（Brown 1995）どちらもskill-basedの下位テストを含む形式であるが，

ELIでは，リスニング，読解及び作文のコースがそれぞれレベル別に履修できるようになっており，プレイスメント・テストも，oral interviewに加えて，この3種の下位テストを開発，実施しているようである。（Brown 1996 : 12, 26）なお，TOEFLの下位テストについては後述する。

アプローチ2

もうひとつの立場にあるのが，Bachman（1990 : 82）が示すように，特定のテストを開発するのではなく，複数の目的のためのテストを開発することによって，より能率的にテストを作成することが可能であるという考えである。ただし，その前提として，それぞれの用途の妥当性が検証されなければならない。拡大解釈すれば，TOEFLあるいは他の既存の標準テストを用いることの妥当性が検証されれば，それをプレイスメント・テストとして活用できることになる。

本学言語教育システム研究室では，その研究課題のひとつに，他大学におけるプレイスメント・テスト実施状況に関する調査を計画している。現段階では，具体的な調査の実施には至っていないが，筆者の知る範囲では，TOEFL, TOEIC（The Test of English for International Communication），CELT（Comprehensive English Language Test）などの複数の下位テストを備えた標準テストや，大学英語教育学会（JACET）の開発による聴解力テスト等の単一のスキル（大抵の場合はリスニングのようである）のテストを活用しているところが多いようである。また，これらの標準テストをプレイスメント・テストとしてではなく，指導の事前・事後テストとして用いて教育効果の検証を行うなど，研究目的での利用も多く見られる。

ところで，本学経済・経営学部の状況を見てみると，プレイスメント・テストとして，入学直後の1回生を対象に独自の開発によるテストを実施し，後にTOEFL-ITPを利用している。但し，前者については，直接プログラムの内容に焦点を当てたものではないという点からは，熟達度テストの性格が強く，後者については，NRTではあるが，結果の解釈においてはプログラム内の集団の比較に用いていることになる。

次に，両学部が実施している2種類のプレイスメント・テストについて，その構成等の説明と実際のテスト結果の分析を行っていく。

Ⅲ．立命館大学経済・経営学部におけるプレイスメント・テスト

両学部の1回生時第1セメスターには，2種類の試験を一斉受験することになっている。ひとつは4月に実施されるプレイスメント・テストで，もうひとつは6月末に実施されるTOEFL-ITPである。

1．テストの構成

プレイスメント・テスト

7年前に，授業の円滑な進行のために，英語を母語とする教員のクラス編成用として経営学部

がプレイスメント・テスト（以下,「プレイスメント」とする）を実施し，その後，経済学部でも実施するようになったいきさつがある。テスト作成にあたっては，難易度にばらつきをもたせるようにし，skill-basedの構成で，リスニング・セクションが65%，読解・語彙セクションが35%の配分のテストが開発されている。テストの構成と時間配分及び項目数は以下の通りである。

リスニング（解答時間　約17分）（合計　40問　65点満点）

Ⅰ　聴こえてくる質問に対する応答を選択　（15問）
Ⅱ　聴こえてくる文の内容と同じものを選択　（15問）
Ⅲ　聴こえてくる対話に関する質問に対する応答を選択　（10問）

語彙・リーディング（解答時間　35分）（合計　35問　35点満点）

Ⅳ　類義語（句）　（20問）
Ⅴ　短文空所補充　（6問）
Ⅵ　短文内容理解　（9問）

現在，このテストは，新入生の入学直後に両学部で一斉に実施され，結果の合計点を中心にして第1セメスターの英語必修科目である「英語1～4」（各1単位）のクラス編成が行われている。なお，プレイスメントを実施する趣旨は，履修要項に以下のように示されている。

「“英語1～4”のクラス編成は，オリエンテーション期間中に実施するプレイスメント・テストにより5段階到達度別に編成します。そのねらいは，いかなる学力レベルから出発しても，学習の動機づけを重視し，学習主体たる学生諸君の充実感・達成感を実現することにあります。前・後期の各セメスター中に実施されるTOEFL-ITPを全員が受験し，そのスコアをもって次年度のクラス編成を行います。」（経済学部履修要項1998 p15より）

TOEFL-ITP

第2セメスターを含む次年度のクラス編成には，6月末に実施されるTOEFL-ITP（以下,「TOEFL」とする）の結果が用いられる。TOEFLの下位テストの構成や項目数，時間，得点は次の通りである。（Institutional Testing Program Manual for Supervisors, ETS 1996より）

		時間	項目数	得点範囲
section 1	Listening Comprehension	35分	50問	20-68
section 2	Structure and Written Expression	25分	40問	20-68
section 3	Reading Comprehension	55分	50問	20-67
合　計		115分	140問	200-677

2．結　果　分　析

本年度4月に実施されたプレイスメントと6月末に実施されたTOEFLの結果を比較，分析し，両テストのプレイスメント・テストとしての妥当性を考えていく。分析に当たっては，経済学部生のみのデータを用いた。なお，プレイスメントに関しては，下位テスト毎の結果が入手不

可能であったため，２種類の下位テスト（リスニング・テストおよび語彙・リーディング・テスト）の得点を合わせた合計点のみを分析に用いた。一方，TOEFL に関しては下位テスト毎の結果も分析に用いた。

プレイスメントの結果の分析

表２　プレイスメントの基礎統計

	平　均	標準偏差	例　数	最小値	最大値
全　体	46.50	11.786	787	16	90
Super Adv.	75.56	6.277	22	60	90
Advanced	60.90	4.491	152	49	71
Upper Int.	49.18	3.162	277	44	55
Intermediate	38.42	3.705	264	31	44
Basic	26.53	3.548	72	16	31

プレイスメントの素点を中心にして，Super Advanced（SA）・Advanced（AD）・Upper Intermediate（UI）・Intermediate（IM）・Basic（BA）の５段階の水準にクラス編成を行っているが，表２は，全体及びレベル毎の基礎統計を示したものである。全体の平均値が46.50点（SD=11.786, n=787）で，±１標準偏差の範囲が58.29から34.71であり，この範囲の者は Advanced から Intermediate の３レベルに配置されていることになる。実際には，±１標準偏差の中にに入る者は，Advanced152名中53名，Upper Intermediate277名全員，Intermediate264名中213名であった。つまり，本学部の英語プログラムの中心となるレベルは，Upper Intermediate と Intermediate となる。

各レベル間に統計的に有意な差がなければ，プレイスメント・テストとしての本来の役割，つまり，受験者の英語力を測定し，適切なレベルに配置するという機能を果たしていないことになる。そこで，有意差を検定するに当たり，各レベルの標本数が22から277までばらつきがあるため，Fisher の PLSD を用いて多重比較を行った。分散分析の結果（表３），条件（レベル）による効果は有意であった。（合計：$F(4, 782)=1721.231$, $p<.0001$）さらに，多重比較によると（表４），どのレベル間においても，有意差が観察され（$p<.0001$），クラス編成方法の整合性が認められたことになる。

ところで，表２の基礎統計の標準偏差を比較した場合，SA レベルの数値が他に比べて大きいことから（SD=6.277），郡内のばらつきの度合いが大きく，同一レベルと見なしてはいるものの，力の差が大きいことになる。実際，箱ひげ図を見たところ（図２），SA の下方にはずれ値が存在しており，これは他の要因によって SA レベルに配置されたとも推察できるが，１レベル下への配置が適切であったとも考えられる。なお，UI 及び IM については安定した分布になっていると言えよう。

表３　分散分析表（プレイスメント）

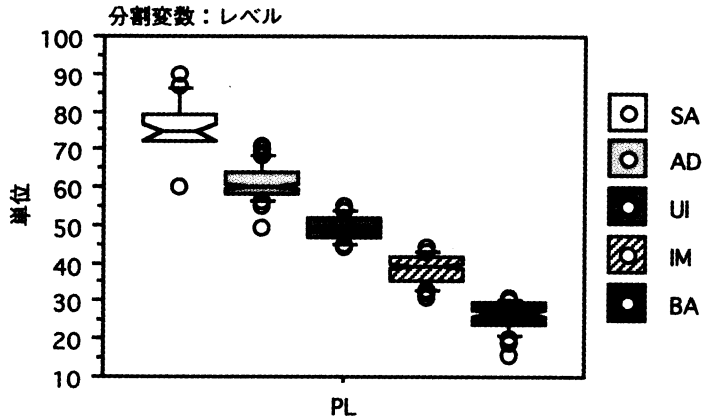
要　因	df	平方和	平均平方	F 値	p 値
レ　ベ　ル	4	98044.687	24511.172	1721.231	<.0001
誤　　差	782	11136.060	14.240		

表4 Fisher の PLSD（プレイスメント）

（効果：レベル 有水準：5％）

	平均値の差	棄却値	p 値
SA vs AD	14.644	1.690	<.0001
AD vs UI	26.365	1.641	<.0001
SA vs IM	37.129	1.644	<.0001
SA vs BA	49.018	1.805	<.0001
AD vs UI	11.721	.748	<.0001
AD vs IM	22.485	.754	<.0001
AD vs BA	34.374	1.060	<.0001
UI vs IM	10.764	.637	<.0001
UI vs BA	22.653	.980	<.0001
IM vs BA	11.889	.985	<.0001

図2 箱ヒゲ図（プレイスメント）



TOEFL の結果の分析

6月末に実施した TOEFL の，合計点および下位テスト毎の基礎統計を，受験者全体およびレベル毎に示したものが表5である。

前年度の同時期に実施された TOEFL の平均値は413点（n=576）で，今年度の数値はそれを約13点上回っているが，標準偏差等の統計資料がないため，検定によって有意性を見ることができない。また，今年度4月に実施された他学部の結果は，次のようになっているが，実施時期の違いやその間の指導効果等の内的妥当性（internal validity）に関わる問題が大きいため，これらの数値を比較することは危険である。ただ，今後，長期的に英語プログラムの効果やプレイスメント・テストのあり方を調査研究していく上では，データを入手し分析していく必要がある。

法 学 部	411.6	(n=430)	産業社会	404.3	(n=581)
国際関係	456.3	(n=234)	政策科学	422.8	(n=261)
文 学	408.5	(n=571)	理 工	381.8	(n=1230)

表5　TOEFLの基礎統計

	レベル	平　均	標準偏差	例　数	最小値	最大値
合計	全　体	**426.32**	**46.488**	**724**	**217**	**570**
	Super Adv.	498.68	39.964	19	380	570
	Advanced	463.04	31.591	139	330	537
	Upper Int.	436.77	30.522	255	327	500
	Intermediate	403.08	41.141	250	290	483
	Basic	371.83	42.718	59	217	447
S1	全　体	**40.63**	**4.433**	**724**	**24**	**64**
	Super Adv.	50.63	5.833	19	42	64
	Advanced	43.45	3.886	139	34	54
	Upper Int.	40.83	3.379	255	32	49
	Intermediate	38.79	3.651	250	28	47
	Basic	37.68	3.967	59	24	43
S2	全　体	**42.99**	**6.295**	**724**	**20**	**61**
	Super Adv.	48.42	5.680	19	31	58
	Advanced	46.69	4.686	139	33	61
	Upper Int.	44.36	4.777	255	27	56
	Intermediate	40.82	6.187	250	24	56
	Basic	36.14	6.717	59	20	50
S3	全　体	**44.27**	**6.159**	**724**	**21**	**59**
	Super Adv.	50.53	3.864	19	41	57
	Advanced	48.76	4.899	139	29	59
	Upper Int.	45.83	4.464	255	29	56
	Intermediate	41.31	5.621	250	278	53
	Basic	37.71	5.718	59	21	50

S1=Listening　S2=Structure and Written Expression　S3=Vocabulary and Readind

(1)　合計点をもとに

表5に示したように，全体の平均値が426.32点（SD=46.488, n=724）であったが，先と同様に，±1標準偏差の範囲内（472.81から379.83）のレベル毎の分布を調べた。すると，Super Advanced 19名中3名，Advanced 139名中97名，Upper Intermediate 255名中227名，Intermediate 250名中181名，Basic 59名中29名がその範囲内にあった。このように，TOEFLの結果において±1標準偏差内におけるレベルの者が，プレイスメントの結果を基にした水準のすべてにまたがって分布していることは，プレイスメントの段階付けとTOEFLの結果になんらかのズレがあることを示唆している。

そこで，プレイスメントを基に設定した5段階の水準が，TOEFLの得点に於いても有効であるか否かを詳しく検証する必要がある。但し，ここで断っておかなければならないのは，プレイスメントとTOEFLの実施の間に，週4コマ，約10週間の指導が行われているため，この指導や学習者の他の要因による影響及び測定道具の一貫性の欠如などの内的妥当性を脅かす要因が多く，単純に統計処理をすることは危険であるということである。しかし，それらの影響を排除す

る方法や補うための他のデータがないため，それを承知の上で敢えて分析を行っている。

まず，TOEFLの得点をもとに分散分析を行ったところ（表6），各レベル間の差は，合計点においても，また下位テストにおいても統計的に有意であった（$p<.0001$）。また，相関関係をみても，両テストの合計点の間には$r=.689$で強い相関がみられた（表7）。つまり，この結果を見る限りでは，TOEFLにおいても，4月のプレイスメントによる設定水準が有効であったことになる。

ところが，レベル毎に相関関係を調べたところ，SAとUIで有意な相関がみられなかった（表7）。このことは，それぞれのテストが測定している内容や構成要素の違い及び難易度の影響による可能性が考えられる。

表6　分散分析表（TOEFL）

	要因	df	平方和	平均平方	F値	p値
合計	レベル	4	624917.894	156229.474	120.398	<.0001
	誤　差	717	930388.438	1298.613		
S1	レベル	4	4371.940	1092.985	79.734	<.0001
	誤　差	717	9828.581	13.708		
S2	レベル	4	6896.212	1724.053	57.347	<.0001
	誤　差	717	21555.589	30.064		
S3	レベル	4	8889.439	2222.360	86.580	<.0001
	誤　差	717	18404.214	25.668		

表7　プレイスメントとTOEFLの相関係数

	相関係数	n
全体	.689 ***	705
SA	.345 ns	19
AD	.304 **	138
UI	.187 ns	252
IM	.395 ***	238
BA	.396 ***	58

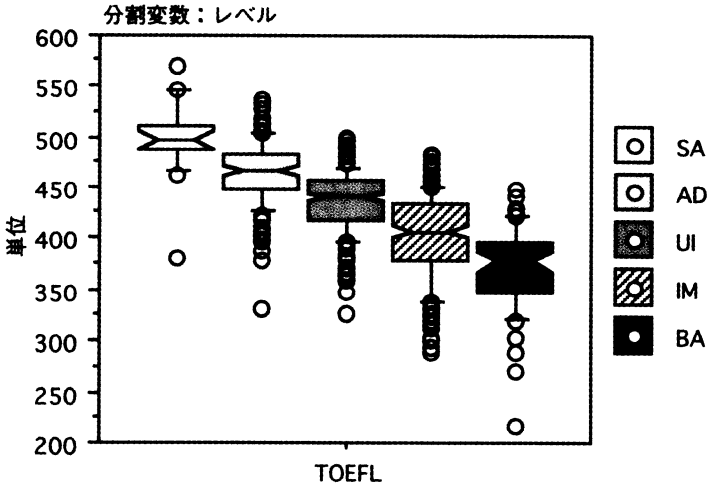
*p<.05, **p<.01, ***p<.001

次に箱ヒゲ図で郡内のばらつきの度合いを見たところ（図3），プレイスメントの結果と同様にSAの下位にはずれ値が存在した。また，他のレベルにおいてもはずれ値が目立った。さらに，標準偏差から判断すると，SAレベル内でのばらつきは小さいが，他のレベルにおいてはばらつきが大きく，プレイスメントの結果よりも安定性が悪い。これは，TOEFLは「高度な英語力を測定するテストなので，400以下は有意なスコアとは認められない」（三修社　1991）と言われているように，より高得点の者（この場合はSAレベル）に対しては，安定し且つ信頼性のある測定道具だと判断できるのかもしれない。

(2)　下位テストをもとに

各下位テストの基礎統計を表5に，分散分析の結果を表6に示してある。この3種の下位テストの結果をもとに，合計点では観察できなかったレベル間の特質を見ていく。

図3　箱ヒゲ図（TOEFL-全体）

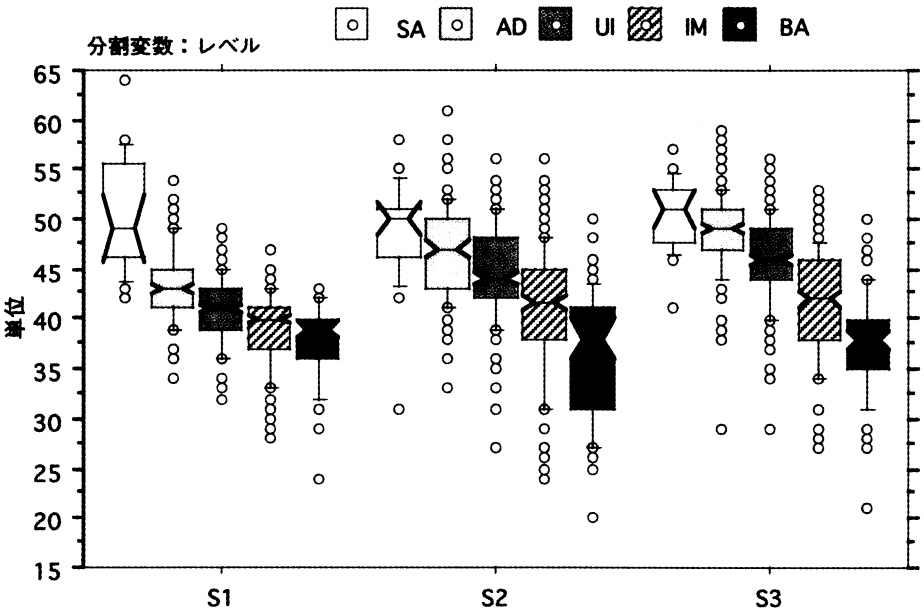


まず，Fisher の PLSD による多重比較によると（p. 374の資料参照），S 1 （Listening）における IM と BA の間，S 2 （Structure & Written expression）および S 3 （Vocabulary & Reading）における SA と AD の間には有意な差が観察されなかった。つまり，聞き取りにおいては，図 4 の箱ヒゲ図からもわかるように IM と BA の間では，レベル間の差が小さく，平均値の差が1.114（p 値＝.0380）（IM：$\overline{X}$＝38.79，SD＝3.651・BA：$\overline{X}$＝37.68，SD＝3.967）で，両レベルをひとつのグループと見なすことができよう。同様に，S 2 と S 3 に関わる言語領域では，SA と AD は等質グループと見なせよう。（S 2 において平均値の差1.730，p 値＝.1974（ns），SA：$\overline{X}$＝48.42，SD＝5.680・AD：$\overline{X}$＝46.69，SD＝4.686）（S 3 において平均値の差1.771，p 値＝.1534（ns），SA：$\overline{X}$＝50.53，SD＝3.864・AD：$\overline{X}$＝48.76，SD＝4.899）プレイスメントでは下位テスト毎の結果がないために，詳細な分析ができないが，もしリスニング・テストの得点分析を行えば，音声言語における英語力の特質がさらに観察できる可能性がある。

箱ヒゲ図を用いると変数の分布を視覚的に比較できるが，特に目立つのは，S 1 における SA レベルの上方向への広がりと，S 2 での BA レベル及び S 3 での IM と BA の下方に占める割合の大きさである。SA については，プレイスメントでも同一レベル内でのばらつきが大きかったが，TOEFL の結果から推測すると，特に聴解力においてその傾向が強い可能性がある。

ところで，このような結果は，今後の英語プログラムでの指導に反映させたいものである。例えば，筆者の速読指導の効果に関する研究（Shimizu 1994, 清水1996）では，読みのスキルの学習を行うことで，読みの活動においてそれらのスキルを活用できるようになり，WPM（words per minute）だけでなく読解力，聴解力も伸長した。ところが，伸びの見られなかった学習者は，文法力などの基本的な言語能力が不十分な者で，スキルの習得にも支障がみられた。このことは，前述の，文字言語を通しての下位テスト（S 2 と S 3 ）での下位レベルのグループ（IM と BA）の結果に当てはまり，この段階の学習者に語彙，文法を含む bottom-up process に関わる英語能力の治療的指導の必要性を示唆していると言えよう。また，SA レベルは，文字言語を通じての下位テスト（S 2 と S 3 ）において AD レベルとの有意性を示さなかったのに対して，音声言語を用

図4 箱ヒゲ図（TOEFL-下位テスト）



いた下位テスト（S1）では，ADとの重なりもなく顕著に秀でていた。このことは，一般に聴解と読解の相関が高いことから，読みの要領を経験させるような読解ストラテジーの指導により，読解力の向上の可能性が期待できよう。

Ⅳ．考察と今後の課題

本稿では，言語プログラムにおけるテストの中から，特にプレイスメント・テストを取り上げ，経済学部1回生の受験した2種類のテスト（独自開発のプレイスメント・テストとTOEFL-ITP）の結果の分析を行った。分析から明らかになったことは，4月に実施したプレイスメントをもとに設定されたレベル間の有意性が，TOEFLにおいても統計的に観察され，受験者全体からみると両テスト間の相関が高かった。このことから，設定水準によるクラス編成の上では，プレイスメントもTOEFLも妥当な測定道具だとみなすことができた。ところが，レベル毎に見たときの相関の問題や，レベル内でのばらつきやはずれ値の出現状況が異なることなどから，両テスト間になんらかの違いがあることが推察される結果となった。

ここで今後の課題として，次の4点をあげておく。

(1) Brown（1995：72）は理想的なNRTは，難易度（facility index）と弁別力（discriminating index）の点から受験者集団に適していなければならないとしているが，今回の経済学部のテスト・データに関しては，個々の受験者の応答結果が入手できないために，項目分析ができなかった。ただ，本学で採点処理しているプレイスメントに関しては，今後，下位テスト別や個々の応

答の分析が可能であろう。項目分析によって難易度や弁別力を知り，それに基づいた項目の改良を加えると共にアイテム・バンクの構築も可能になる。

(2) 社会的に評価を得ているテストの利用が，強い波及効果（washback effect）をもたらすとされているが（Gates 1995），TOEFL をプレイスメント・テストとして用いることが，はたして有効であるのか否かについては，難易度や内容的妥当性の上からも検討の余地がある。特に，TOEFL は低得点の者には床面効果（floor effect）が起こり，あるレベル以下の者を識別できない可能性があるし，独自のプレイスメントについても，床面効果や天井効果（ceiling effect）が起きていないかを検証した上で，対象となる受験者に，より適した測定道具を考えていくことが必要である。

(3) 現在のプレイスメントの採点法に関して，下位テスト間および下位テスト内での配点比重に問題がないかも検討が必要である。

(4) 下位テストの分析により，学習者の英語力の特質を知り，英語プログラムのシラバスに活かしていくことも重要である。

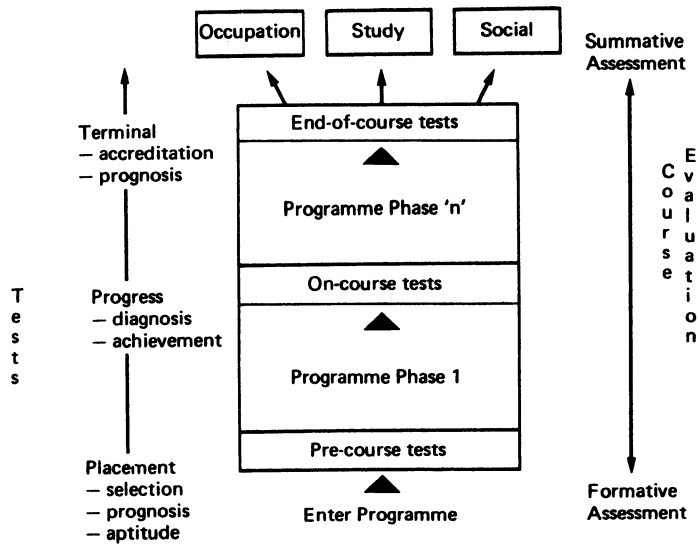
プレイスメントの目的で，既存の標準テストを採用するのか，プログラム内で新規に開発するのか，あるいはプログラムのニーズに合うように現存のものを改良していくのか，いずれの方法をとるにしても，上述のようなテスト結果の分析に加えて，それぞれのテスト項目が測定しようとするスキルやテクスト・タイプ，テスト形式などを含んだ item specification を行い，内容的，構成概念的妥当性等の研究も行っていかなければならない。テスト開発では，Hughes（1989：48-58）が示すように，1 Statement of the problem, 2 Writing specification for the test content, 3 Writing the test, 4 Pretesting のような段階を踏んで行くわけであり，長期的な視点で作業にあたらねばならない。

最近のテスト法の分野での関心は，熟達度テストも CRT の得点を提供するために開発できるかどうかという問題であり（Bachman, L. F. and S. j. Savignon 1986），Bachman（1990：88）は，「……もともと能力あるいは内容領域の目標基準的水準に照らし合わせて作成されたテストの得点を，受験者の適切なグループの言語行為にもとづいて集団準拠的に解釈することが可能である。同様に，既存の目標基準準拠テストにあわせて集団規準準拠テストを尺度化しようとすることもまた時には有用である」としており，テストが複数の目的で相互乗り入れ出来る可能性もあることを示唆している。このような動向にも関心を向けておく必要がある。

最後に，教授-学習とテストについて触れておく。テスト・システムとプログラム及び学習ニーズは三位一体の関係になければならない。次の Carrol and Hall（1985：9）の図 5 はそれを端的に示している。本稿は NRT であるプレイスメント・テストを中心に，administrators の立場で，プログラム・レベルでの決定の観点からテスト結果の分析を行ってきた。つまり，図 5 における pre-course tests について考えてきたわけであるが，実際の教育現場では CRT が中心であり，教室レベルでの決定については，on-course tests 等を通じて個々の教師が最も直接的に責任を負うものである。現在の英語プログラムでは，教室レベルでの決定，つまり学習者への成績の決定については，5 つのレベル毎に傾斜評価が行われているが，その妥当性も検証していかなけ

ればならないであろう。また，統一教材のもとで英語力別クラス編成を行っていることとの関連から，総括的テストや形成的テスト（summative and formative tests）の必要性についても，学習目的やニーズと照らし合わせながら考えていく必要がある。さらに，6月末のTOEFLの結果が，次のセメスター以降のプレイスメントの資料となっているが，第1セメスターの学習効果がTOEFLのスコアに反映しているのか否か，また教室内での評価が後のクラス編成に加味されるのか否か等の検討も必要であろう。

図5 Test-programme Relationship（Carrol and Hall 1985）



テストの目的は，教育目的に加えて研究目的ということもあげられる。研究により解明された情報を活用し，教育現場に還元しながら，よりよい測定道具や方法による評価が行われる環境を作っていきたいものである。

1） 1998年7月2日プロジェクトBV外国語教育におけるFD研究「98年度春期オープンクラス・ウィークを実施して」

2） 立命館大学言語コミュニケーションセンター「外国語学習の手引1998」等参照。

3） Bachman（1981：68）の場合はマクロ的評価とミクロ的評価ということばを用いて説明している。

参 考 文 献

Alderson, C., Clapham, C. & Wall, D.（1995）. *Language Test Construction and Evaluation.* Cambridge University Press.

Backman, L. F.（1981）. Formative evaluation in program development in R. Mackay and J. D. Palmer（eds.）. *Language for Specific Purposes: Program Design and Evaluation.* Newbury House.

Bachman, L. F.（著）池田央，大友賢二（監訳）（1990）「言語テスト法の基礎」みくに出版

Bachman, L. F. and S. j. Savignon.（1986）. The evaluation of communicative language proficiency ; a critique of the ACTFL oral interview. *The Modern Language Journal,* 70, 4, 380-9.

Brown, J. D.（1995）. Developing norm-referenced tests for program-decision making. in Brown, J. D. & Yamashita, S. O.（eds）. *Language Testing in Japan.* The Japan Association for Language

Teacher, pp 40-47.

Brown, J. D.（1996）. *Testing in Language Programs.* Prentice Hall Regents.

Carroll, B. J & Hall, P. J.（1985）. *Make Your Own Language Tests.* Pergamon Press.

Educational Testing Service（ed）.（1996）. *Institutional Testing Program Manual for Supervisors.* ETS.

Gates, S.（1995）. Exploiting washback from standardized tests in Brown, J. D. & Yamashita, S. O.（eds）. *Language Testing in Japan.* The Japan Association for Language Teacher, pp 101-106.

Genesee, F. & Upshur, J. A.（1996）. *Classroom-based Evaluation in Second Language Education.* Cambridge University Press.

Harrison, A.（1983）. *A Language Testing Handbook.* Macmillan Press.

Huges, A.（1989）. *Testing for Language Teachers.* Cambridge University Press.

Jenkins, Jordan & Weiland（1992）. The Role of Writing in Graduate Engineering Education : A Survey of Faculty Beliefs and Practices. *English for Specific Purposes*, vol. 12, 51-67.

木村真治，津村修志，清水裕子（1998）「科学的テスト時代のテストの非科学性――テスト問題形式と配点比重についての考察」大学英語教育学会大37回全国大会における研究発表

三修社編（1991）「英語の資格をとるマガジン，Book '91」三修社

Shimizu, Y.（1994）. A Study on the Effectiveness of Speed Reading Instruction.『近畿大学教養部研究紀要』第25巻，第３号，pp13-24.

清水裕子（1996）「日本人大学生の読解態度と英語力の変化」『近畿大学教養部研究紀要』第28巻，第２号，pp55-67.

Thornton, J.（1998）.（personal communication, July 27 and August 5 1998）

Winsor, D. A.（1990）. Engineering writing / writing engineering. *College Composition and Communication*, 41, 58-70.

資　料

多重比較の結果（Fisher の PLSD，効果：レベル，有意水準：5％）

プレイスメント

	平均値の差	棄却値	p 値
SA vs AD	14.644	1.690	<.0001
AD vs UI	26.365	1.641	<.0001
SA vs IM	37.129	1.644	<.0001
SA vs BA	49.018	1.805	<.0001
AD vs UI	11.721	.748	<.0001
AD vs IM	22.485	.754	<.0001
AD vs BA	34.374	1.060	<.0001
UI vs IM	10.764	.637	<.0001
UI vs BA	22.653	.980	<.0001
IM vs BA	11.889	.985	<.0001

TOEFL-合計

	平均値の差	棄却値	p 値
SA vs AD	35.648	17.298	<.0001
AD vs UI	61.912	16.818	<.0001
SA vs IM	95.604	16.830	<.0001
SA vs BA	126.854	18.655	<.0001
AD vs UI	26.263	7.456	<.0001
AD vs IM	59.956	7.483	<.0001
AD vs BA	91.205	10.989	<.0001
UI vs IM	33.693	6.294	<.0001
UI vs BA	64.942	10.217	<.0001
IM vs BA	31.249	10.236	<.0001

TOEFL-S1（Listening）

	平均値の差	棄却値	p 値
SA vs AD	7.186	1.778	<.0001
AD vs UI	9.800	1.729	<.0001
SA vs IM	11.840	1.730	<.0001
SA vs BA	12.954	1.917	<.0001
AD vs UI	2.615	.766	<.0001
AD vs IM	4.654	.769	<.0001
AD vs BA	5.768	1.129	<.0001
UI vs IM	2.039	.647	<.0001
UI vs BA	3.153	1.050	<.0001
IM vs BA	1.114	1.052	.0380

TOEFL-S2（Structure & Written Expression）

	平均値の差	棄却値	p 値
SA vs AD	1.730	2.633	.1974
AD vs UI	4.060	2.560	.0019
SA vs IM	7.605	2.562	<.0001
SA vs BA	12.285	2.840	<.0001
AD vs UI	2.330	1.135	<.0001
AD vs IM	5.875	1.139	<.0001
AD vs BA	10.555	1.673	<.0001
UI vs IM	3.545	.958	<.0001
UI vs BA	8.225	1.555	<.0001
IM vs BA	4.680	1.558	<.0001

TOEFL-S3（Vocabulary & Reading）

	平均値の差	棄却値	p 値
SA vs AD	1.771	2.433	.1534
AD vs UI	4.699	2.365	.0001
SA vs IM	9.218	2.367	<.0001
SA vs BA	12.814	2.624	<.0001
AD vs UI	2.928	1.049	<.0001
AD vs IM	7.447	1.052	<.0001
AD vs BA	11.044	1.546	<.0001
UI vs IM	4.519	.885	<.0001
UI vs BA	8.116	1.437	<.0001
IM vs BA	3.596	1.440	<.0001