プレイスメント・テストの結果分析

——今後の課題にむけて——

清　水　裕　子

1．はじめに

教育に携わる中で，多くの場合，評価活動を避けて通ることは不可能であり，そのための情報を得る手段として，テスト等を用いての測定が必要になってくる。そして教授-学習過程で展開される測定と評価に基づいて，様々な決定を学習者に与えることになる。それは，授業活動の中で課されたタスクに対する達成度や理解度を示すものであるかもしれないし，入学や進路に関わるものかもしれない。Brown（1996）は，それらを教室レベルにおける決定とプログラムレベルにおける決定に分類し，前者を目標規準準拠テスト（CRT：Criterion-Referenced Test），後者を集団基準準拠テスト（NRT：Norm-Referenced Test）と関連づけている。Brown（1996：9）および和田（1999：12）をもとに，決定の種類とテストを分類すると図１のようになる。

最近のわが国の教育，特に初等・中等教育界では絶対評価が注目を浴びてきているが，勿論，大学においても，目標到達度の検証や診断的情報の提供からも，教授-学習過程に即した測定と評価は重要な意味を持っている。これについてはカリキュラムおよび指導目標との関連が強いことから，稿を改めることにし，本稿においては，立命館大学経済・経営学部の英語プログラムにおける測定，とりわけ，そこで実施しているプレイスメント・テストに焦点を絞ることにする。

ところで，教授－学習過程では，①テスト方法，②プログラムおよび③ニーズの３要素が有機的な関係をもって，より効果的な学びの場が提供され展開していくわけであるが，現段階の我々の英語プログラムは，そのような理想的な状況には未だ達していない感が強い。そこで，本稿では，現在用いている英語プレイスメント・テストの結果の分析に基づき，その問題点と今後のあり方について検討していく。特に，プレイスメント・テストが我々の英語プログラムにどのように機能しているのかを見，もし効果的に作用していない点があるとすれば，その対処法を講じるための資料を提示することにする。

なお，本研究は，立命館大学における2001年度学術研究助成（特定研究１：研究代表者・清水裕子，共同研究者・野澤和典）を受けて行った調査研究の一部を紹介するものである。

図１　プログラムにおけるテストの種類

決定の種類とテスト			テストする情報	実施時期
教室レベル	目標規準準拠テスト（CRT）	診断テスト	非常に具体的 ↕ 非常に一般的	コースのはじめまたは中間
		達成度判定テスト		コースの終わり
プログラムレベル	集団基準準拠テスト（NRT）	配置クラス決定テスト		コース/プログラムのはじめ
		熟達度判定テスト		入学前や卒業時

2．ＢＫＣ経済・経営学部における外国語プログラム

2.1　プレイスメント・テストの実施について

本学ＢＫＣキャンパスでは，経済・経営両学部が同じプログラムに従って外国語教育を進めている。これは，1998年度の本学外国語教育改革の際に構築されたカリキュラムに則ったものである。英語に関しては，１回生の第１セメスターにおいて必修になっており（図２参照），レベル別クラス編成を行ない，統一教材およびシラバスで授業を進めている。これらの受講対象になる新入生の中には，英語を受験科目として受けていない者も存在する。また，受験科目として受けていても，本学の入試では，共通テストや平行テストの実施を行っていないため，同じ尺度による英語力の測定が行われていない。そこで，新入生に対して，独自に開発した英語テストを入学直後に実施し，その結果に基づき英語科目のクラス編成を行っている。

独自のテストによるクラス編成の歴史は，十数年前にさかのぼり，当初は，英語を母語とする教員の授業進行の効率化を図ることを目的にしていたと聞く。しかし，ここ数年の内に項目分析等を通してのテストの見直しを行い，現在では，後述するテスト形式・構成による英語プレイスメント・テストが開発され，その結果をもとに英語科目のクラス編成に活用している。

2.2　ＢＫＣ経済・経営学部における外国語プログラムの概要

図２は，経済・経営学部の外国語プログラムをまとめたものである。第１セメスターにおける「英語１～英語４」〔各１単位〕の４種が必修科目であり，入学直後のオリエンテーション期間中に実施する英語プレイスメント・テスト（図内 Test 1）に基づき，５つの水準に分けたクラス編成を行っている(2002年現在)。第１セメスター終了時には，各自の興味や目的に応じて，第２・３セメスターに履修するプログラムを決定するが，次の３つのコースが用意されており，第３セメスター終了時点で，外国語科目12単位の取得が完了することになる。

① 英語専修コース：各セメスターで英語を４単位ずつ履修。

② 英語・初修２言語履修コース：各セメスターで英語２単位，初修外国語２単位を履修。（初修外国語はドイツ語，フランス語，中国語，スペイン語から選択する。）

③ 初修重視コース：各セメスターで初修外国語を４単位ずつ履修。（初修外国語はドイツ語，フランス語，中国語から選択する。）

さらに，上記コースの①あるいは②の場合，つま

図２　経済・経営学部における外国語プログラム

	第１セメスター			第２セメスター ⟶	第３セメスター
入学試験	Test 1	英語１ 英語２ 英語３ 英語４ （各１単位）	Test 2	英語専修コース	英語５ 英語６ 英語７ 英語８ （各１単位）	英語９ 英語10 英語11 英語12 （各１単位）
				２言語履修コース	英語５ 英語６ （各１単位） 初修・基礎（２単位）	英語９ 英語10 （各１単位） 初修・展開（２単位）
				初コース修重視	初修・基礎Ⅰ（２単位） 初修・基礎Ⅱ（２単位）	初修・展開Ⅰ（２単位） 初修・展開Ⅱ（２単位）
	４単位			４単位	４単位	

り英語科目を履修する場合，第２・３セメスターの英語クラスは「リベラルアーツプログラム」・「ビジネスキャリアアプログラム」・「アカデミックキャリアプログラム(英語専修コースのみ)」から選択することになっている。なお，これらのプログラムにおけるクラス編成は，2000年度までは６月に実施するTOEFL-ITP試験のスコアをもとに，コース別およびプログラム別にレベル別クラス編成を実施していたが，2001年度からは定期試験期間中にTest 2を実施し，Test 2の結果及びTOEIC-IPやTOEFL-ITPの得点をもとに配置を行っている。

2.3　プレイスメント・テストの必要性

大学入試においては，特に多くの私立大学に見られるように，複数の入試日が設定されており，入学者全員が同じ試験を受けて入学してくることは少ない。また，入試形態も多様化しており，いわゆるpaper-pencilタイプのテストを受けずに入学する場合もある。立命館大学においても，入学者の受験するテストや選抜形態は多岐にわたる。また，同じ形態のテストであっても，入試日によって異なるテストが実施されるのは当然のことである。例えば，英語を例にとった場合，基本となるテスト形式はほぼ統一されているものの，難易度や弁別力の点で完全に平行テストであるとは言えない。このことは本学に限ったことではなく，日本の入試体制の中では，予備テストを行い，その結果分析等を行い，厳密な意味での平行テストを作成，実施することは不可能である。つまり，入学者の英語力は，共通かつ唯一の測定道具により測定されているのではなく，異なる道具を用いて測定していることになる。勿論，異なるテスト間では，得点の比較をすることはできないし，合否判定に関しては，異種のテスト間の得点を比較することはない。つまり，入学してきた学生の英語力に関しては，受験したテストにおける英語の得点は入手可能であるが，テスト種毎に異なった基準をもっており（合計点も異なる場合がある），単純に比較することはできないのである。そこで，入学者全体の英語力に関する特質を把握したり，またレベル別にクラス編成を行う場合には，対象となる集団に対して，共通のものさしによる測定を行うことが必要になってくる。

2.4　ＢＫＣプレイスメント・テストの概要

ＢＫＣでは，4月の新入生対象のオリエンテーション期間中に，クラス配置を目的に英語テストを実施してきているが，2001年度に実施したテストに関する情報は以下のとおりである。

表１　テストの構成

<table>
<tr><th>下位テスト</th><th>主なタスク</th><th colspan="2">項目数</th><th colspan="2">時間配分（分）</th><th>配点</th></tr>
<tr><td rowspan="3">Ⅰ
Listening</td><td>応答問題</td><td>15</td><td rowspan="3">30</td><td colspan="2" rowspan="3">約20</td><td rowspan="3">60</td></tr>
<tr><td>対話問題</td><td>10</td></tr>
<tr><td>ミニ・レクチャー</td><td>5</td></tr>
<tr><td rowspan="2">Ⅱ
Grammar</td><td>空所補充</td><td>20</td><td rowspan="2">30</td><td colspan="2" rowspan="2">15</td><td rowspan="2">60</td></tr>
<tr><td>誤文認識</td><td>10</td></tr>
<tr><td rowspan="2">Ⅲ
Reading</td><td>クローズ文
（英文数１）</td><td>15</td><td rowspan="2">30</td><td>8</td><td rowspan="2">33</td><td rowspan="2">60</td></tr>
<tr><td>・スキャンニング問題
（英文数１）
・主題探し問題
（英文数２）
・推測問題（英文数２）</td><td>15</td><td>25</td></tr>
<tr><td colspan="2">合計</td><td colspan="2">90項目</td><td colspan="2">68分</td><td>180点</td></tr>
</table>

受 験 者：経済・経営・理工学部新入生
試験時期：4月第1週のオリエンテーション期間中
テストの指示：日本語
リスニングについてはカセットテープを使用。
配 布 物：問題冊子および解答用マークシート。
回答方法：マークシートに記入。項目はすべて4者択一形式。
採点方法：機械採点。(試験当日に処理を行い，次の日に配置決定。2日後に学生に提示し，授業開始。)
テストの構成：(表1参照)

3．テスト結果の分析と問題点

2001年度4月に実施したプレイスメント・テストに関して，経済学部のデータをもとに結果を分析したが，ここではその結果とクラス編成における問題点を見ていく。

我々のプログラムでは，現在までのところ，テスト結果をもとに5つの水準（SA: Super Advanced, AD: Advanced, UI: Upper Intermediate, IM: Intermediate, PI: Pre-Intermediate）のクラスを設定している。レベル分けおよびクラス分けに際しては，3種の下位テスト（Listening, Grammar, Reading：各60点満点）の得点を合計した得点（180点満点）を用いている。これは，試験実施直後（通常，試験翌日）にクラス分け作業を行い，その翌朝にクラス編成を掲示し，授業が開始されるため，解答に用いたマークシートの読み取り作業を行う処理作業の時間の関係から，下位テスト毎の得点が即座に入手できないためである。そこで，合計点というひとつのパラメータのみをもとにレベルの決定を行うわけであるが，このことによる問題点については「3．3　問題の所在」で示すことにする。

3．1　結果分析：合計点を基に

5水準別の基礎統計を表2に示したが，180点満点中，最上位群（SA）の平均値は138.06点（標準偏差：7.357），最下位群（PI）は47.90点（標準偏差：7.732）であり，その差は90点以上もあり，同一プログラム内の学習者の英語力の格差がわかる。また，全体でみると最高点が152点であるのに対して，最低点が23点，標準偏差が22.501であることからも，同じプログラム内の学習者を，英語力の面で均質とみなすには無理があり，レベル分けにより授業を進めていくことの必要性や整合性が理解できよう。

表2　英語プレイスメント・テストの基本統計
（合計点・180点満点）

水準（人数）	平均値	標準偏差	最高点	最低点
SA (32)	138.06	7.357	152	131
AD (243)	115.74	6.958	130	105
UI (280)	97.48	4.782	106	88
IM (273)	78.24	8.961	91	58
PI (60)	47.90	7.732	57	23
全体 (888)	93.75	22.501	152	23

なお，この5水準のレベル設定に関しては，合計点を基にした水準間の差の有意性は統計的にも検証されている。また，各水準内の等質性については，図1のボックス・プロットが示している。つまり，本テストの合計点に基づく限り，各水準内では，それぞれを一つの水準グループと見なすことに問題はないと解釈できる。ただし，最下位レベル（PI）において，はずれ値が観察されているが（図3のPIに

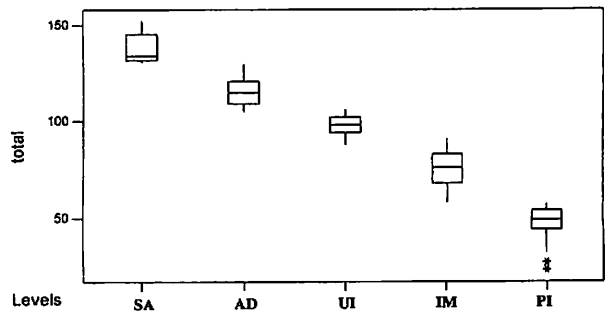

図3　合計点をもとにした水準毎のボックス・プロット

おける＊印），さらにPI水準以下のクラスを設定することは，プログラム全体からみても非効率的と判断する。

3．2　結果分析：下位テストの得点を基に

クラス編成作業で入学者を5水準に分ける際には，合計点のみに基づくのではなく，各下位テスト毎の得点を参考にしながらクラス分けをすることを意図していたが，新学年の開始時であり，非常に厳しいスケジュールの中でそのような処理を実施することは不可能であるとの回答をデータ処理側から得，やむを得ず合計点のみを基に配置を行ってきている。このように，処理作業を行う者と，クラス配置の作業を行う者が同一でないことは，分業の点では効率が良いが，一方の要求が受け入れられずにそれぞれの処理が行われてしまい，より望ましい配置を阻むことになっているのが現実である。テスト実施から1ヶ月余経った時期に，下位テストの得点を入手することができ，その結果を分析したところ，興味深い結果が見られた。

表3は下位テスト毎の基本統計を示したものである。単純に素点を比較してみると，最上位群(SA)と最下位群（PI）ではListeningとGrammarが逆転しているが，それ以外の3群ではListening＜Grammar＜Readingの順で点数が上がっている。また，全体での得点分布の幅はReadingがもっとも広く（標準偏差：10.376），Listeningがもっとも狭い（標準偏差：7.880）状況であった。

次に，各水準間での有意性を検証するために，Tukey-Kramer法による分析を行ったところ，表4に示すように，水準間の平均点の差と棄却値から，どの水準間においても有意水準5％で統計的に有意であることが観察された。つまり，合計点をもとにした水準設定は，下位テスト毎に見ても特に問題がなかったことが証明されたことになる。

ところが，表3の中で，最高点と最低点を水準間で比べてみると，明らかに数字に重なりがあることがわかる。たとえば，AD水準のListeningの平均値

表3　下位テスト（各60点満点）の基本統計

水準（人数）	下位テスト	平均値	標準偏差	最高点	最低点
SA (32)	Listening	43.88	7.448	56	28
	Grammar	43.44	5.309	54	28
	Reading	50.75	4.429	57	40
AD (243)	Listening	31.56	5.627	52	16
	Grammar	38.49	5.001	52	24
	Reading	45.68	5.346	57	27
UI (280)	Listening	26.40	5.234	46	8
	Grammar	32.93	5.518	46	18
	Reading	38.15	5.793	54	22
IM (273)	Listening	21.21	4.922	32	6
	Grammar	25.04	5.926	40	10
	Reading	28.99	6.580	47	14
PI (60)	Listening	15.67	4.152	26	4
	Grammar	14.70	5.113	26	4
	Reading	18.53	5.947	34	5
全体 (888)	Listening	26.12	7.880	56	4
	Grammar	31.17	8.995	54	4
	Reading	36.46	10.376	57	5

表4　Tukey-Kramerの法による下位テストにおける各水準間の差

比較する水準	Listening		Grammar		Reading	
	平均値の差	棄却値	平均値の差	棄却値	平均値の差	棄却値
AD-IM	10.355	1.271	13.454	1.319	16.694	1.421
AD-PI	15.897	2.078	23.794	2.156	28.150	2.322
AD-SA	−12.311	2.711	−4.944	2.813	−5.067	3.029
AD-UI	5.164	1.264	5.565	1.311	7.530	1.412
IM-PI	5.542	2.055	10.340	2.133	11.456	2.297
IM-SA	−22.666	2.694	−18.397	2.795	−21.761	3.010
IM-UI	−5.191	1.226	-7.888	1.272	-9.165	1.370
PI-SA	−28.208	3.156	−28.738	3.274	−33.217	3.526
PI-UI	−10.733	2.051	−18.229	2.128	−20.620	2.291
SA-UI	17.475	2.690	10.509	2.791	12.596	3.006

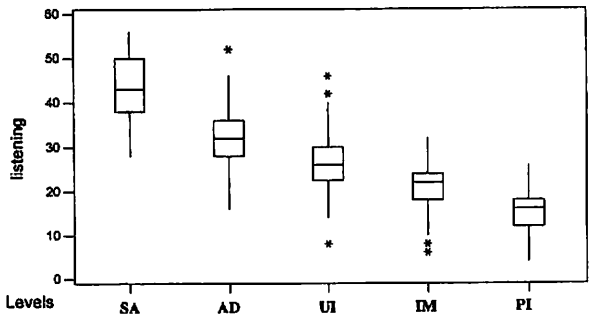

図4　Listeningセクションの水準毎のボックス・プロット

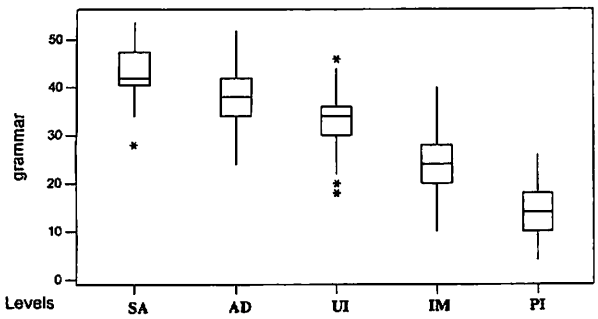

図5　Grammarセクションの水準毎のボックス・プロット

は31.56点であるが，その水準内の者の最高点は52点であり，これは最上位水準（SA）の平均値を大きく上回っていることになる。また最低点は16点であり，これは最下位水準（PI）の平均値とほぼ同じである。そこで，各群内の分布状況を観察するために，下位テスト毎にボックス・プロットを作成してみたところ，図4～6のようになった。

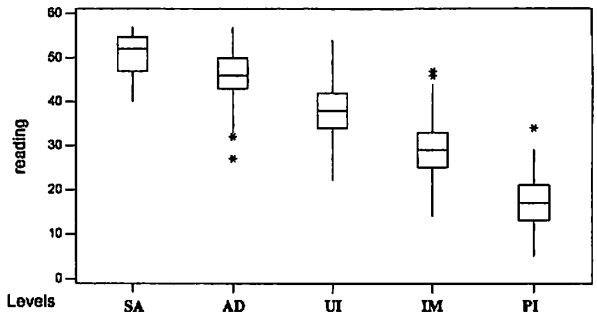

図6　Readingセクションの水準毎のボックス・プロット

3．3　問題の所在

これらの図から明らかなように，各水準内にはずれ値（グラフ内の＊印）が存在し，これらの値をとった者は，その下位テストにおいては，実はその水準に属していない学習者だということになる。ここで，合計点のみに頼ったレベル分けの問題点が明確になってきている。本学の元常勤講師であるDianne Lim氏が実施したエスノグラフィー研究においても，学習者および英語を母語とする教員の面接調査及び質問紙調査の分析結果の中に，配置されたレベルの不適合の一面が伺える報告が行なわれていたが（Lim & Shimizu 2001），これは今回の分析結果を補う情報であると言える。

例えば，岐阜大学地域科学部では，TOEFL-ITPの下位テストの得点を基に，各学習者の英語力の特質を活かしたクラス配置を試みているが（Sugino et al. 2002），詳細な分析に基づく配置決定は，小規

模のプログラムでは可能なのかもしれない。しかし，本学経済・経営両学部のように1500名を越える学生規模で試験を実施し，その直後に分析を行いクラス配置を実現させることは困難である。また，我々のプログラムの英語１～英語４の４種の科目それぞれが焦点を当てているスキルや学習目標を鑑みて，科目毎にクラス編成を変えることは，時間割の構成上，不可能なことである。

しかしながら，今回の分析で下位テストにおいてはずれ値の存在が観察された以上，合計点のみによるクラス配置を継続することは，用いるテストによって学習者の英語力の特質に関する情報が得られるにも関わらず，その情報から目をそむけてしまっていることになる。合計点をひとつの基準値とし，下位テストの得点をも考慮した水準設定とクラス配置が実現することで，習熟度別クラス編成による学習効果がより拡大するのは確かであり，少なくとも，学習者自信が納得できるクラス配置が実現できると言えよう。

４．今後の課題

今回の分析結果も含めた上で，今後のプレイスメント・テストに纏わる環境に関して３つの課題を提示しておく。

４.１　下位テスト毎の得点とその活用

Multiple Measures（Ardovino et al. 2000）において，ひとつの測定が学習者の学習状況やプログラムの効果を十分に反映するかどうかと言う質問に対する答えとして，Supovitzは，".....any single testing method has its own particular set of blinders and that *bias is intrinsic in any type of test.*"と述べている。つまり，如何なるテスト法においても，何らかの限界があり，複数の方法でアセスメントを行い，適切な評価を与えていかなければならないことになる。今回の分析においても，一見，整合性があるように思われた配置も，厳密には見直すべき点があることが観察された。ただ，時間的および人的資源の欠如から，より適切な配置作業が阻まれているのは事実であり，プログラムの主体となる学習者に対して，より良い学びの環境を提供していく方法を考えていかなければならない。次年度に向けては，少なくとも下位テスト毎の得点の入手を依頼し，その実現が期待できる状態になってきているが，さらには，それらの数値が，学習者へのフィードバックや診断的情報として提供でき，それが学習へのモティベーションにつながっていくようなシステムが構築できることを期待する。

４.２　カリキュラムとテストの関連性とアイテム・バンクの構築

本研究では，プレイスメント・テストというものに焦点を当てたが，このようにカリキュラムやプログラムの内容との関連が深いテストでは，実際の指導およびアセスメントを一直線上に置いて考えなければならない。本テストは，実際には図１におけるNRTとして実施しているものではあるが，CRT的な要素と言う意味合いも含んでいる。つまり，テスト開発の際には，プログラムにおける指導目標を考慮しながら項目作成を行ってきたのである。ただ，教室で実際に指導されることや使用している教科書の学習の達成度を，直接測定するための道具ではない。今後，学習者に対してより具体的，診断的情報の提供のための測定道具の開発も，カリキュラムとの関連から考えて行かねばならないであろう。何れにしても，信頼性および妥当性があり，公平な測定道具としてのテストを確立し，提供していくには，今後も，必要なデータの収集と分析を行い，テスト項目のデータベースとしてのアイテム・バンクの構築や，項目の見直しと改訂によるさらなる精錬を重ねていかなければならない。

４.３　外部テストの有効活用

Ardovino等（2000：9）が，"some schools see standardized NRTing as an unnecessary chore, with no connection to the "real" job of teaching and learning."と述べているように，独自開発のテストではなく社会的にも認められている標準テスト等を用いる場合には，実際のカリキュラムとの関係を考慮し，慎重な態度で実施する必要があろう。しかし，わが国でも広く利用されているTOEICやTOEFLに

関しては，それらの得点を上げることのみが学習目的になってしまう危険性があるというマイナス面と共に，信頼性が高いテストであり，その得点が社会的にも認められるているという事実から，学習へのモティベーションにつながるという肯定的な意見も多く存在する（Shimizu 2002：242)。本学の体制としては，TOEFLやTOEICの受験を推進する方向にあるようだが，それらのテストが備える特徴と目的を理解し，英語力と受験目的との整合性を，受験者自身が見いだせる方向で活用されることを願う。

5．最後に

本稿で扱った英語プレイスメント・テストは，大学側が積極的に受験を勧めているTOEICやTOEFLのテスト構成概念を踏襲する形で開発したものである。これらのテストは構造言語学の影響を受けた言語能力感を基本にしており，つまり，言語の熟達度は文法や語彙などの要素や，聞き取り，読解などの技能から成り立つというものである。しかし，その後，1990年代にはいり，Bachman（1990）のcommunicative language abilityという概念が登場し，言語能力というものを，状況の文脈と実世界の知識に連動させる能力と規定されていった。今後，このような言語感が，英語を外国語（第2言語としてではなく）として学ぶ環境にあるわが国において，テスト理論や実際のテスト開発にどのような影響を与えていくのかを追っていく必要がある。

最後に，言語テストの研究分野では，高度な統計処理と共に理論面での展開も進んでおり，それらを取り入れ活用することで，本プログラムにおけるプレイスメント・テストの開発と実施およびその分析による教育現場への還元が大いに期待できる。今後も，言語テスト分野だけでなく，カリキュラムや教授法研究における理論と実践が足並みを揃え，より望ましい英語プログラムを展開できることを望む。

参考文献

Ardovino, J, Hollingsworth, J. & Ybarra S. (2000). *Multiple Measures.* Corwin Press.

Bachman, L.F. (1990). *Fundamental Considerations in Language Testing.* Oxford University Press.

Bachman, L.F. and Palmer, A.S. (1996). *Language Testing in Practice.* Oxford University Press.

Bailey, Kathleen M. (1998). *Learning about Language Assessment.* Newbury House.

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

Carroll, J.B. (1961). Fundamental considerations in testing for English language proficiency of foreign students. *Testing the English Proficiency of Foreign Students.* Washington, DC: Center for Applied Linguistics:30-40.

Henning, G. (1987). *A Guide to Language Testing.* Newbury House.

Lim, D. & Shimizu, Y. (2001). The Effectiveness of Placement Tests and Participants' Response. (Presented at JACET 40th Annual Convention at Fuji Women's College, Sapporo)

和田稔訳（1999)．『言語テストの基礎知識』．大修館書店.

大友賢二，ランドルフ・スラッシャー監訳（2000)．『〈実践〉言語テスト作成法』．大修館書店.

Shimizu, Y. (2002). Survey Research on the Use of Placement Test at Four-Year Universities in Japan. *Ritsumeikan Studies in Language and Culture. Vol. 14, No. 1, 231-243.*

Sugino, N. et al. (2002). English Language Proficiency of the 1st-Year FOREST Students (1). *Bulletin of Faculty of Regional Studies, Gifu University, Vol. 10, 137-146.*