

英語プレイスメントテスト分析

——言語テストの観点から——

吉 田 弘 子

要旨

近年、英語プレイスメントテストを実施して習熟度別クラス編成を行い、大学英語教育の充実を図る大学が増加している。しかし、プレイスメントテストの目的や方法については、まだ十分周知されているとは言い難い。本稿では、まず言語テストを集団基準準拠テストと目標規準準拠テストに大別し、その特性をテスト範囲、目的、プログラムとの関係、実施時期、スコアの解釈、テスト問題の予想の有無、得点の分布の各項目について論じる。さらに、日本の大学で汎用されている代表的な英語プレイスメントテスト（G-TELP, TOEIC Bridge, ACE Placement, CASEC）を言語テストの妥当性、信頼性、実用性の観点から分析する。また、プレイスメントテストを習熟度テストとして使用する際の留意点及びテストの等化についても論じる。本稿が、英語プレイスメントテストを選択・運営する際の理解の一助となることを期待したい。

キーワード：英語プレイスメントテスト、集団基準準拠テスト、目標規準準拠テスト、言語テスト

I は じ め に

平成15年3月に文部科学省が策定した「英語が使える日本人」の育成のための行動計画¹⁾が、大学生が卒業時に求められる英語力を「卒業したら仕事で英語が使える」と明確に設定し、社会で通用する英語力を求めたことは、大学英語教育界に大きな影響を与えた。この計画では、大学における英語教育の達成目標の設定状況などの調査・公表を求め、大学における英語教育の改善のための一層の取組を促したため、これを契機に英語教育プログラムの改善に取り組み始めた大学も少なくない。その一つとして、大学英語教育においてプレイスメントテストを用いて、習熟度別クラス編成を実施し、より一層充実した英語教育を提供しようとする大学の増加が挙げられる。しかしながら、英語プレイスメントテストの目的や方法については、まだ十分周知されておらず、その利用にあたっては混乱が見られる例も少なくない。本稿では、言語テスト研究の観点からまず言語テストの種類と特性について論じ、次に日本で使用されている代表的な英語プレイスメントテストを言語テストの条件に基づき分析する。

1) 詳細は http://www.mext.go.jp/b_menu/houdou/15/03/030318a.htm 参照。

II 言語テストの種類と特性

言語テストは、その目的から集団基準準拠テスト (norm-referenced test: NRT) と目標基準準拠テスト (criterion-referenced test: CRT)²⁾ に大別することができる。集団基準準拠テストは、受験者の総合的な習熟度を測定することが目的であり、各受験者のテストスコアは受験者全体のスコアと相対的に解釈される (Brown, 2005, p. 2)。Test of English for International Communication: TOEIC (ETS, 2006) は、リスニングとリーディングの2つの下位テストに分かれているが、「英語によるコミュニケーション能力を総合的に評価することを目的」(ETS, 2009) とした 集団基準準拠テストの一例である。一方、目標基準準拠テストは、明瞭な学習事項を各受験者がどの程度習得できたかを測ることを目的としたテストである。例えば、スピーキングのクラスで、(1) 挨拶をする (2) 自己紹介をする (3) 道を尋ねるなどの具体的な学習目標を、受験生がどのくらいのレベルに到達したかを測定するインタビューテストは目標基準準拠テストである。目標基準準拠テストのスコアは、他の受験生のスコアと比較されることがなく、目的をどの程度達成することができたかどうかについて絶対的に判断されることがその特徴である。したがって、仮にクラス全員がコース目標を100%達成したのであれば、全員のスコアが100点であるということもありうる。定期試験などのクラス内テストも、目標基準準拠テストである。Brown (1996) はさらに、テスト問題の予測性についても集団基準準拠テストと目標基準準拠テストに明確な相違があると指摘している。集団基準準拠テストは一般的な能力を測定するという目的から、多種多様な内容の問題が使用され、どのような言語材料が実際にテストされるか予測することは困難である。しかし、目標基準準拠テストでは、コースの授業目標の明確な設定、コース目標に準拠した授業運営など、授業がコースの目標に則り、適切に行われているなら受験生はテスト問題を比較的容易に予測できると考えられる。

言語テストはさらに、テストで何を決定するのかという観点から、習熟度テスト、プレイスメントテスト、到達度テスト、診断テストの4つのタイプに分類できる。習熟度テストとは、受験者の一般的な英語力を測定するものであり、プレイスメントテストは学習者をプログラムにおいて適切なレベルに配置することを主な目的とする。到達度テストは、学習者がプログラムの中でどの程度学んだか (到達度) をコースの到達目標と照らし合わせて判定するものがある。診断テストは、通常コース開始時や途中で学習者の課題を明らかにするために使用される試験である。習熟度テストとプレイスメントテストは集団基準準拠テストであり、到達度テストと診断テストは目標基準準拠テストである。これらの4つの言語テストについてまとめたものを表1に記す。

このようにこれらの4つの言語テストは異なる特性をもっており、使用目的に応じたテスト利用することが求められる。教育現場では、ひとつのテストが複数の目的で使用され

2) 「基準」と「規準」の使用については諸説あるが (皆見, 2008), 本書では和田 (1999) の訳に従った。

ることもあるが、その際には十分な検証をおこない、テスト結果がその目的にあった機能を果たしているかどうか検討しなければならない。例えば、習熟度テストをプレイスメントテストに用いるのには注意が必要である。アメリカの大学・大学院へ入学する際に必要とされる Test of English as a Foreign Language (TOEFL) (ETS, 2009) は、習熟度テストであるが、これをプレイスメントテストとして用いることは、多くの英語プログラムでは適切ではない。なぜなら、TOEFL は非常に多様な受験者集団の幅広い英語能力を測ることを目的としているため、TOEFL が測定対象とする受験生の習熟度ほど幅広くない特定の大学（あるいは、特定の学部）に入学する学生用のプレイスメントテストとして使用しても、クラス分けが適切に行えるほど受験生を分布させることができない可能性が高い。これは、TOEIC をプレイスメントテストに使用した場合も同様のことがいえる。

表1 テストの種類と特性

	集団基準準拠テスト (NRT)		目標基準準拠テスト (CRT)	
	習熟度テスト	プレイスメントテスト	到達度テスト	診断テスト
対象テスト範囲	通常、入学などに必要な総合的言語能力	言語プログラムの全レベルにかかわる学習項目	コースまたはプログラムの到達目標	コースの到達目標
目的	個々の受験者の全体的な学力を他の受験生と比較する	受験者をプログラムの適切なレベルに配置する	進級や卒業のためにどの程度学習が達したかを判定する	受験生やクラス担当者にさらに学習の必要な項目を知らせる
プログラムとの関係	プログラムと関係なし	プログラム内で比較	プログラム（コース）の目標と直接に関係	プログラム（コース）の目標と直接に関係
実施時期	入学前、またはプログラム終了時	プログラムのはじめ	コースの終わり	コースのはじめまたは途中
スコアの解釈	相対的	相対的	絶対的	絶対的
テスト問題の予想	予想ほぼ不可	予想ほぼ不可	予想可能	予想可能
得点の分布	正規分布	正規分布	正規分布しない	正規分布しない

注：Brown, J. D. (2005). *Testing in language programs: A comprehensive guide to English language assessment* (pp. 3 & 7) New York. McGraw-Hill College より著者の許可を得て筆者訳、一部加筆修正

Ⅲ. 英語プレイスメントテスト分析

1. テストの妥当性、信頼性、実用性

次に日本で汎用されている英語プレイスメントテストを言語テスト条件の見地から検証

する。現在、日本では G-TELP (International Testing Services Center: ITSC, 1984), TOEIC Bridge (Educational Testing Service; ETS, 2001), ACE Placement (英語運用能力評価協会: ELPA, 2002), CASEC (教育測定研究所, 2000) が大学などで英語のプレイスメントテストとして比較的によく用いられている。各テストの基本情報を表 2 に示す。

表 2 英語プレイスメントテスト基本情報

テスト	G-TELP	TOEIC Bridge	ACE Placement	CASEC
ウェブ サイト	http://www.g-telp.jp/	http://www.toEIC.or.jp/bridge/	http://www.kirihara-kyoiku.net/BACE-ACE/placement.html	http://casec.evidus.com/
運営会社	G-TELP 日本事務局	財団法人国際ビジネスコミュニケーション協会 TOEIC 運営委員会	特定非営利活動法人 (NPO) 英語運用能力評価協会 (ELPA)	株式会社教育測定研究所 (JIEM)
開発年	1984年	2001年	2002年	2000年
主管・ 管理	International Testing Services Center (ITSC)	財団法人国際ビジネスコミュニケーション協会 TOEIC 運営委員会	特定非営利活動法人 (NPO) 英語運用能力評価協会 (ELPA)	財団法人英語検定協会が基礎開発 (株)教育測定研究所が開発・運営

言語テストを検証するには、まずテストの妥当性と信頼性に注目しなければならない。妥当性とは、テストが測定しようとしているものが、意図しているものを測定しているかどうかを示す。例えば、英語の話す能力を測定することを目的としたテストならば、テストは英語を「読む」「書く」ではなく、「話す」能力を測定していなければ妥当性があるとは言えない。妥当性はさらにいくつかのタイプに分類できるが³⁾、それぞれの検証方法は Bachman (2004) 及び Brown (2005) に詳しい。ここでその詳細を述べることは控えるが、例えば専門家による判断 (expert judgment) も妥当性検討の一方法である。ここでいう専門家とは、英語のテストであれば英語教育やそれに関連する専門領域の専門家をさす。表 3 は 4 種類の英語プレイスメントテストの妥当性と信頼性に関連する項目をまとめたものである。まず、テスト目的は、各テストが測定しようとする内容である。これらのテスト目的は、対象とする受験生の測定すべき英語力と一致していなくてはならない。また、テストの受験レベルが受験生に適合しているかどうかの検討も必要である。どのような素晴らしいテストであっても、プレイスメントテストの目的である「受験生をテスト結果によ

3) Brown (1996) は、プレイスメントテストなどの集団基準準拠テストの妥当性は、内容性妥当性、構成概念妥当性、目標関連妥当性を主に検証すべき方法として挙げている。

表3 英語プレイスメントテストの妥当性と信頼性

		G-TELP	TOEIC Bridge	ACE Placement	CASEC
妥当性	テスト目的	英語を母語としない人たち（中略）が、実際の場面でどれだけ英語でコミュニケーションがとれるか、その力を測るテスト	TOEIC の特長（英語によるコミュニケーション能力を測定する）を備えつつ初・中級レベルの英語能力測定に照準を合わせて設計されたテスト	高校の英語の基礎学力から高度な運用能力まで、幅広い英語力をカバーし、大学生を対象とした入学時クラス分けテスト	英語コミュニケーション能力判定テスト
	テスト問題と問題数	文法：20問 リスニング：20問 リーディング&語彙：20問	リスニング：50問 リーディング：50問	語彙・文法：14問 リスニング：16問 リーディング：16問	語彙知識：15問 表現知識：15問 リスニング（大意把握）：15問 リスニング（ディクテーション）：10問
	受験レベル	4 レベル	レベル分けなし	レベル分けなし	レベル分けなし
信頼性	信頼性係数	0.84 （ α 使用） *注1	0.85 （KR-20 使用）	0.88 （ α 使用）	0.96－0.98 *注2
	測定の標準誤差（SEM）	5.8 （100点満点）	8 （180点満点）	15 （300点満点）	（～500点） 35～45点 （500～800点） 30～40点 （800点～） 35～45点

注1：レベル4テストの場合

注2：林（2001）による。

り幅広く分布させる」ことができれば、そのテストは測定すべき内容を測定していないといえる。受験レベルが設定されている場合（例 G-TELP）、どのレベルのテストを使用するか事前に十分に吟味する必要がある。また、テスト目的は、観念的に表現されることも多い。あるテストのテスト目的が「コミュニケーション能力」と表されていれば、その「コミュニケーション能力」がどのような概念をさすのかを判断する必要がある。そのためには、実際にテスト問題を吟味してそのテスト目的と問題が整合しているか検討することが求められる。一例をあげれば、表3ではCASECにリスニング問題の中でディクテーションが用いられている。CASECはコンピュータを用いて受験するため、ディクテーションでは受験生は英語を聞いてそのままキーボードに入力していくことが求められる。確かにディクテーションはリスニング力を測る手段であるが、「タイプで英語を入力する」という行為は、英語力とは異なるスキルであり、仮に英語のリスニング力があっても英文

をタイプ入力することに不慣れな受験生はスコアが伸びない可能性があるので要注意である。受験生が英文をタイプ入力することに精通している場合はもちろんこの懸念は無用であるが、大学新入生など受験生のコンピュータ・リタラシに差があると予想される場合は、テストの妥当性として考慮すべき点であろう。

次にテストの信頼性についてであるが、信頼性とはテスト結果に一貫性・安定性があるかどうかを表す。集団基準準拠テストの信頼性は、再テスト法、等価形式法、内部一貫性法などで測定でき (Brown, 1996, p. 223), 信頼性係数 ($r_{xx'}$) を算出することで判断される。完全に信頼できるテストは信頼性係数が1となる (Brown, 2005)。これらのうち内部一貫性法は、ひとつの形式のテストを1度施行するだけで測定できるため、最もよく使用されている。内部一貫性法による信頼性の算出には、クロンバック・アルファ、キューダー・リチャードソン公式が用いられる。それぞれの計算式は次のとおりである (Brown, 2005)。

クロンバック・アルファ (Cronbach α)

$$\alpha = 2 \left(1 - \frac{S_{\text{odd}}^2 + S_{\text{even}}^2}{S_t^2} \right)$$

S_{odd} = 奇数番号のテスト項目の標準偏差

S_{even} = 偶数番号のテスト項目の標準偏差

S_t = テストの総得点の標準偏差

キューダー・リチャードソン公式21 (K-R21)

$$\text{K-R21} = \frac{k}{k-1} \left(1 - \frac{M(k-M)}{kS^2} \right)$$

k = テスト項目数

M = テスト得点の平均値

S = テスト得点の標準偏差

キューダー・リチャードソン公式20 (K-R20)

$$\text{K-R20} = \frac{k}{k-1} \left(1 - \frac{\sum S_i^2}{S_t^2} \right)$$

k = テスト項目数

S_i^2 = 項目分散

S_t^2 = テスト得点の分散

先述した4種類のテストで、ホームページ上に公開されている資料に信頼性係数が明記されていたのは、TOEIC Bridge⁴⁾とCASECである。G-TELPとACE Placementは、筆者が運営会社に問い合わせた結果、信頼性係数が開示された。プレイスメントテストはその結果に

4) 英語版 HP からのみアクセス可能。

基づき受験者のクラスを決めるものであり、受験者に非常に大きな影響を与える試験 (high-stake test) であることを考えると、テストの信頼性はテスト作成者によって積極的に公表されるべき項目であると思われる。

信頼性係数はテストの信頼性の度合いを示したものであるが、同一の能力を有する受験者がテストを繰り返し受験した場合、その受験者の得点が属する領域を示す測定の標準誤差 (standard error of measurement: SEM) もテストの信頼性を論じるときには留意すべき点である。測定の標準誤差が狭ければ狭いほど予想できる変動の幅が狭い、つまり受験者の能力を一貫して測定できるテストであるといえる。測定の標準誤差は次の計算式で算出できる (Brown, 2005, p. 235)。

$$SEM = S\sqrt{1 - r_{XX'}}$$

S = テストの標準偏差

$r_{XX'}$ = テストの信頼性推定値

例えば公開されている TOEIC Bridge の測定の標準誤差は、リスニングとリーディングセクションにおいて各 4 点である。これは、TOEIC Bridge のリスニング・セクションで 60 点をとった受験者は、その実力が変わらない限り毎回のテストで 56～64 点の範囲のスコアを得る、ということを示している⁵⁾。この SEM については、4 種類のテストの中で、ホームページ上に公開されているのは、TOEIC Bridge のみ⁶⁾であったが、この項目もテスト作成者によって積極的に公表されるべきであろう。

なお、プレイスメントテストは、クラス分けを目的とした集団基準準拠テストであり、様々な英語習熟度の学生を幅広く分布させるように作成されている⁷⁾。しかし、学生の英語習熟度がプレイスメントテストの想定受験者から大きく逸脱している場合は正規分布を得られないこともありうる。このような場合は分析結果を正しく解釈することは難しく、有効なクラス分けが困難になる場合も考えられるので、テスト実施者は、少なくとも得点のグラフなどを確認して、正規分布の特徴であるベル・カーブが得られているかを常に確かめる必要がある (Brown, 2005, p. 132)⁸⁾

妥当性と信頼性に加えて、テストを大学などで実際に実施し、習熟度別クラス分けとして用いるためには、テストの実用性も考慮されるべきである。どれほど妥当性と信頼性の

5) より詳しく述べると、「テストを繰り返し受けるとその回数の 68% は 1 SEM プラス (60+4=64) と 1 SEM マイナス (60-4=56) の領域内得点がこの SEM から結論づけられる。

6) 英語版 HP のみからのみアクセス可能。

7) 通常、プレイスメントテストの作成では、種々のパイロットテストに基づき項目の難易度の調整などが行われ、幅広い受験生の習熟度に対応できる項目が採用される。

8) Bachman (2004), Brown (2005) は、尖度 (kurtosis) と歪度 (skewness) が -2 から +2 の範囲である場合、あるいは、尖度と歪度がその標準誤差の 2 倍より小さい場合は、正規分布であるとしている。尖度は、分布の尖り具合、歪度は分布の非対称性を表す統計量で、エクセルや spss の計算式で簡単に求められる。正規分布では、ともに 0 である。

あるテストであっても費用が高すぎたり、テストの結果を得るのに日数がかかりすぎたりすると現実には使用することは難しい。プレイスメントテストを新学期の限られた日程の中で実施し、それに基づいてクラス分けを行うには、成績結果の受領が短期間で行われることが望ましい。また、教室などの施設がテストの実施に支障がないかどうかも重要な項目である。新学期の多忙なスケジュールの中で大人数の学生にプレイスメントテストを受験させる場合は、テスト所要時間も考慮に入れておく必要がある。さらに、テスト受験にコンピュータの使用が必要な場合は、コンピュータの台数が受験者数だけ確保できるかどうかも考慮に入れる必要がある。試験監督が実施する側で十分用意できない場合は、派遣が可能かどうか、その際の費用の見積もりも考えておかななくてはならないだろう。

また、テストの実用性から考えると、どのような形で結果レポートが返却されるのかも確認しておく必要があるだろう。学生の成績一覧はクラス分けの資料となるものであり最も重視されなければならないが、それに加えて個々の学生のテスト結果は、学習者にとり貴重なフィードバックとなるので、個人成績表などで学生に返却されるかということも確認しておきたい。

2. 言語テストの等化

プレイスメントテストを用いる際のテストの条件は1に述べたが、現実的にはプレイスメントテストはクラス分けだけではなく、学力の伸びを測定する目的に使われることも多い。このように英語プレイスメントテストをクラス分け以外の目的で利用する場合は、別の留意が必要である。

例えば、大学一年生入学時（4月）と終了時（2月）にそれぞれ1年生クラスと2年生クラスの配置を決めるプレイスメントテストを実施した場合、クラス分けの判断で使用するならば、先述したテストの条件、つまり妥当性、信頼性、実用性、そして得点の正規性を考慮するだけでよい。しかし、クラス分けで使用する以外に、2回のテスト結果で1年間の学習成果を検証する、あるいは前年の受験生の得点と比較するなどの場合は、同一テストを使用しない限り、バージョンごとのテストの難易度に注意しなくてはならない。もし難易度が異なるテストを使用した場合、仮に受験生の平均点に高低が認められたとしても、それが本当に受験生の習熟度の差なのか、あるいはテストの難易度差なのか結論づけることはできない。難易度が異なるテストで得られた受験生の得点を単純に比較することはできないからである。しかしながら、テストのバージョンが異なってもそれらの難易度が統制されていれば、得点の比較は可能である。その難易度の調整（テストの等化）は、項目応答理論（Item response Theory: IRT）の項目プール（item bank）を用いてできることが知られている。項目応答理論を用いたテストでは、正解した問題×配点（いわゆる素点）で得点を算出するのではなく、受験者の能力値（ロジット値）という絶対尺度で表す。また項目難易度を絶対的尺度で推定することも可能である。これは項目応答理論では、受験者の能力値や項目の難易度という各々のパラメータ（parameter）を同一尺度上で推定することができるからである⁹⁾。その結果、絶対的な難易度を推定した項目を収集して

項目プールを作成することが可能になる。項目応答理論を使用して作成されたテストは、難易度を推定した数百、あるいは数千の項目をあらかじめ用意しておき、テスト目的に応じてそれらの項目が選びだされて作成される。さらに、テストの等化のためには、項目プールに複数のテストで互いに比較可能な項目、つまり全く同じ項目（共通項目）が必要であり、豊田（2002）はこの点を次のように説明している。

これは、たとえば50個の項目から構成される10個のテストを実施し、項目母数を推定しても500個の項目から成る項目プールは作れないということである。この方法では、テスト間の項目母数は比較可能でなく、10個の小さな別々の項目プール（水たまり）ができるだけである。（p. 89）

これらの共通項目を利用した等化計画は共通項目計画（common item design）あるいは係留テスト計画（anchor-test design）と呼ばれる。このようなプロセスを経て作成されて初めて、テストの等化が可能になるのである。

先述の4つのプレイスメントテストを検証すると、IRTを使用しているのは、TOEIC Bridge、ACE Placement、CASECの各テストである。また、テストの等化のため必要な共通項目は、通常非公開が原則であり、このためにはテスト冊子はすべて受験終了後に回収されなくてはならない（Saida & Hattori, 2008）¹⁰⁾。この条件を満たすのは、G-TELP及びTOEIC BridgeとCASECである。したがって、習熟度テストとして、同一受験生の習熟度の変化や異なる年度の受験生の得点を比較するために必要なテストの等化を満たしているのは、TOEIC BridgeとCASECであるとの結論に達することができる。

表4 テスト等化の条件

	G-TELP	TOEIC Bridge	ACE Placement	CASEC
得点	素点	IRT	IRT	IRT
テスト冊子取扱	要返送	要返送	学生持ち帰り可	WEB か CD-Rom で受験

IV ま と め

本稿では、言語テストの種々の観点から英語プレイスメントテストの検証を試みた。本稿の目的は特定の言語テストを擁護したり、あるいは非難したりすることではない。テストを選択し、実施する者（多くは英語教員であろう）にとって、最も大切なのは「何を最優先するのか」という点を明確にすることである。理想を言えば、言語テストの条件がすべて満たされているテストを選択したいところであるが、実際のプレイスメントテスト

9) 項目応答理論の詳細な説明は豊田（2002）を参照されたい。

10) Saida and Hattori (2008) は、“(i)n Japan, tests usually are developed with the assumption that all test items must be disclosed after administration.”と日本ではテスト問題の開示することが一般的であると指摘している。

の運用に当たっては、多くの制約があることは否めない。例えば、プレイスメントテストの予算が非常に厳しい場合は、何よりもテストの実用面、つまり最もコストの安いプレイスメントテストを選ぶことを最優先せざるえないであろう。その際には、その選択したテストについて弱点を含めてどのような特徴があるのかを把握した上で、運用することを考えることが必要になってくる。また、プレイスメントテストを習熟度テストとしても使用する場合は、難易度の等化されたテストの使用を最優先すべきである。しかし、その場合でもテストの妥当性や信頼性、そして実用性を忘れてはならない。プレイスメントテストを実施すれば、自動的にすべての情報が手に入るかのような、テストにオールマイティの機能を求めるのは、現実的とはいえない。完璧なテストは存在しないと同様に、すべての目的を達成する完璧なプレイスメントテストも存在しないことを心にとめるべきである。

最後に、プレイスメントテストを含めた言語テストのカリキュラム計画全体での役割についても触れておきたい。プレイスメントテストの役割とその意義は先述したが、英語教育のプログラムは決して、適切なプレイスメントテストを選択・実施するだけでは完結しない。Brown (1996; 2005) は言語カリキュラムの体系的枠組みを (1) ニーズ・アナリシス (2) 目標と目的 (3) 言語テスト (4) 教材 (5) ティーチングの要素が相互関係にあり、すべての構成要素がプログラム評価と相互に結びついていると論じている。つまり、プログラムは、上記の5要素の情報を集め、分析して統合することによって改善されるのである。ニーズ・アナリシス、目標と目的、言語テスト、教材、ティーチングにおいて相互のフィードバックがなくては、カリキュラムは即座に柔軟性を失うだろう。カリキュラムの体系的枠組みを活かすには、絶えまなく「発展プロセス」を求めること、つまり言語テストをふくむすべての要素は常に見直しの対象としていかななくてはならない。英語プレイスメントも例外でなく、カリキュラムとの整合性を常に検証していくことが求められるのである。

本稿が、言語テストの専門家以外においても、英語プレイスメントテストが正しく理解され、英語教育関係者がプレイスメントテストの選択・運営する際の理解の一助となることを期待したい。

参 考 文 献

- Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge, Cambridge University Press.
- Brown, J. D. (1996). *Testing in language program*. Upper Saddle River, NJ: Prentice-Hall.
- Brown, J. D. (2005). *Testing in language programs: A comprehensive guide to English language assessment*, New York. McGraw-Hill College.
- ELPA (2002). *ACE Placement*. 東京, 英語運用能力評価協会.
- ETS. (2009). *Test of English as a Foreign Language*. Princeton, NJ: Educational Testing Service.
- ETS. (2001). *TOEIC Bridge*. Princeton, NJ: Educational Testing Service.
- ETS. (2006). *Test of English for International Communication*. Princeton, NJ: Educational Testing Service.

- ITSC (1984). *G-TELP*. San Diego, CA: International Testing Services Center.
- JIEM (2000). *Computerized assessment system for English communication*. 東京, 教育測定研究所.
- Saida, C. & T. Hattori (2008). "Post-hoc IRT equating of previously administered English tests for comparison of test scores." *Language Testing* 25: 187-210.
- 林規生 (2001). "英語能力測定における CAT の適応例と効果測定." 計測と制御 40: 572-575.
- 豊田秀樹 (2002). 項目反応理論 入門編——テストと測定の科学. 東京, 朝倉書店.
- 皆見英代 (2008) "「規準」と「基準」・‘criterion’ と ‘standard’ の区別と和英照合 教育評価の専門用語和訳に戸惑う." 国立教育政策研究所紀要 第137集: 273-281.
- 和田稔 (1999). 言語テストの基礎知識 (訳) (*Testing in language program*). 東京, 大修館書店.

謝辞

英語プレイスメントテストの基礎資料収集にあたり, ご尽力いただいた大阪経済大学の下田直美氏に感謝の意を表したい。