Shizuka, T. & Mochizuki, M. (2014).
The development and validation of a standardized placement test for Japanese university students.
JACET Journal, 58, 121-141.

大学英語教育学会　紀要第58号抜刷
(2014年3月)

The Development and Validation of a Standardized Placement Test for Japanese University Students

SHIZUKA, Tetsuhito
Daito Bunka University
MOCHIZUKI, Masamichi
Reitaku University

一般社団法人 大学英語教育学会

JACET Journal 58 (2014) 121–141

日本人大学生のための標準プレイスメント・テスト開発と妥当性の検証

The Development and Validation of a Standardized Placement Test for Japanese University Students

SHIZUKA, Tetsuhito
Daito Bunka University
MOCHIZUKI, Masamichi
Reitaku University

Abstract

This paper reports on the development and validation of a new English placement test specifically designed for Japanese university students. After discussing problems with EFL placement tests currently used at tertiary institutions in Japan, the paper describes the concepts, specifications, and development process of the new test. It then presents the results of validation attempts based on the response data collected from a sample of 1,800 Japanese EFL learners. The main findings are: (a) the test exhibits a very high Rasch person reliability of .94, with its items well targeted for the intended population; (b) the test consists of items with quite a wide range of difficulty and those difficulties are invariant across sub-samples; (c) the test produced substantially different mean scores for the 18 institutions at which the participants were enrolled; (d) the test exhibits a high criterion-related validity against the TOEIC® test, predicting 68% of its total score variance; and (e) the test data show the best fit to a three-factor correlated model. The paper concludes with the interpretation that the evidence gathered illustrates the validity of the new test.

Keywords: プレイスメント・テスト, テスト開発, ラッシュ・モデリング, 妥当性, 構造方程式モデリング (SEM)

はじめに

本論文は日本人大学生の英語授業におけるプレイスメントを第一の用途とした新しい熟達度テストの開発過程とその妥当性検証の結果を報告するものである。[1] まず (1) 大学英語教育におけるプレイスメント・テストの現状を概観し，つぎに (2) 新テストのコンセプトおよび開発の過程を記述し，その上で (3) 試験的実施によって集積された解答データを分析した結果を示してテストの妥当性を検討する。

背景

プレイスメントの必要性

近年，効果的な英語教育を行うためにプレイスメント・テストによる習熟度別クラス編成を行う大学が増えてきた。これには2つの原因が考えられよう。ひとつは，少子化や入

試形態の多様化により学力差のある学生が同じ大学に入学するようになったことである（市原・井上・佐藤・鈴木・長谷川・丸本・水口，2007；土肥，2006；土肥・柳瀬，2009；藤森，2012）。従来通りの一般入試で学力による選抜を受けて入学する学生がいる一方で，推薦入試のように高校の内申書で一定の学力は保証されて入学するもの，さらにはAO入試のようにまったく学力の保証なしに入学するものもいる。藤森（2012）は2011年度の新入生全員に受験させたTOEIC®で最高865点から最低75点という800点近い幅が観察されたと報告する。吉永（2004）は新潟大学で実施したTOEIC®得点を考察して，400点以上の差がある学生を同一クラスで教えることはできないとしている。

第2の原因は，使える英語力の育成がこれまで以上に大学英語教育に要請されてきているためである。文部科学省は2002年に「『英語が使える日本人』育成のための戦略構想」を打ち出した。それを受けて大学でも英語教育にも数値目標を導入する例が見られるようになっている。土肥（2006）によると，千葉大学では中期計画に外部試験による目標設定を組み込む取り組みを始めている。吉永（2004）は，TOEIC®を利用した数値目標設定により学生の自学自習の具体的目標を与え，かつ教員の恣意的な授業を防ぐことができるとしている。使える英語力の要請は，教養としての英語ではなく実用的な英語の育成という形で大学カリキュラムに影響を与えている（丸山，2011）。TOEIC®やTOEFL®を意識した大学英語教科書が増加していることにもそれは見てとれよう。そしてTOEIC®やTOEFL®のような客観的外部試験の得点で英語教育の成果を示すためには，学生の能力にあったレベルで授業を受けさせることが重要である。こうして社会の要請に応えた効果的な英語教育を行う必要からも，習熟度別のクラス分けが広まってきたと言えよう。

それではどのようなプレイスメント・テストが用いられているだろうか。小泉（2011a）は，大学で実施されているプレイスメント・テストを大学独自のテスト，外部試験，外部試験を改編したものの3種類に分類している。

大学独自に作成するプレイスメント・テスト

個別の大学が独自に作成するプレイスメント・テストでは，当該大学の英語教育の目的や実態に即した試験作成が可能で，出題内容や試験時間も自由に設定することができる。出題内容は，読解問題のみ（Otomo, Niinuma & Maki, 2001），語彙・リスニング（Gorringe & Mazzarelli, 2011），文法・語彙（市原・井上・佐藤・鈴木・長谷川・丸本・水口，2007），文法・語彙・リスニング（James, 2009），文法・読解・リスニング（Grogan, 2011），文法・語彙・クローズ問題・読解（Nakamura, 2010）のように多様ではあるが，文法と語彙は英語の基礎となる知識として組み込まれることが多い。

大学独自のプレイスメント・テストには，それぞれの教育課程に合ったクラス分けが可能であるという利点はあるが，同時に問題点もある。まず妥当性と信頼性の問題がある。大学の教育課程にふさわしい能力をどの程度持っているかによってクラス分けすることを意図したとしても，実際にその能力を測定するテストでなければ，妥当なプレイスメントはできない。またテストの信頼性が低ければ，同じ学力を持った学生が別のクラスに振り分けられる等の事態も起こる。第2に，学力の経時比較が難しい。経時比較には同じ仕様の平行版テストを使用することが望ましいが，実際に個別大学で複数の平行テストを準備するのは容易なことではない。

外部試験の種類

外部試験としては，TOEIC®（三浦，2006；土肥・柳瀬，2009；鈴木・美濃部・シーハン・

杉浦，2010），TOEIC Bridge®（石原，2011；鈴木・美濃部・シーハン・杉浦，2010），G-TELP®（佐藤・中川・山名，2003；丸山，2011），CASEC®（小泉，2011a）などがプレイスメント・テストとして利用されていることが報告されている。

TOEIC®は，Educational Testing Service（ETS）が開発したテストで，身近な内容からビジネスまで幅広くどれだけ英語でコミュニケーションできるかを測定することを目的とする。受験者はリスニング100問（45分），リーディング100問（75分）を2時間で解答し，10～990点まで5点きざみでの評価を受ける。

TOEIC Bridge®は，TOEIC®と同じETSによる，より易しいテストである。テスト時間と問題数はTOEIC®の半分で，リスニング50問（25分）とリーディング50問（35分）が計1時間で実施される。スコアは20点～180点の2点刻みで表される。

G-TELP®は，International Testing Service Center（ITSC）による，英語非母語話者が実際の場面でどれだけ英語でコミュニケーションできるかを測定することを目的にしたテストである。G-TELP®には，ビジネスなどの日常生活の場面でネイティブと同様に複雑なコミュニケーションができるというレベル1から，形式的な表現方法を用いてネイティブと簡単なコミュニケーションができるというレベル4までがある。レベル1以外は，文法，リスニング，読解・語彙の3つの分野で出題される。試験時間は105分（レベル1，90問）から65分（レベル4，65問）である。

CASEC®は，教育測定研究所が実施するオンライン受験の英語コミュニケーション能力判定テストである。出題内容は，語彙（16問），空所補充（16問），リスニング（17問），ディクテーション（11問）の60問で，40分～50分で受験する。

外部試験の利点と問題点

外部テストのプレイスメントのための使用には利点と問題点とがある。第1の利点はテストの作成・採点を大学教員が行う必要がないことである。解答用紙をテスト実施団体に送付すれば，数日のうちに受験者ごとの成績データが送られてくる。大学ではそのデータを元にクラス分けを行うだけである。第2の利点としては，テストが標準化されていていくつかの版があるために，同一受験者の成績の経年変化や年度の異なる入学者の英語力の変化をみることができることである。

問題点としては，第1に費用がかかることである。TOEIC®は団体試験4,040円，公開試験5,565円，TOEIC Bridge®は団体試験2,835円，公開試験4,200円である。G-TELP®は個人受験の場合レベル1が6,300円，レベル2が5,250円，レベル3が4,200円，レベル4が3,150円である。CASEC®は，団体受験で500人以上ならば1人2,205円，個人受験では3,500円である。

外部試験をプレイスメントに利用することの第2の問題点は，幅広い能力の受験者に適した短時間のテストがないことである。同一大学で英語の能力差が大きい場合，G-TELP®のように能力に合わせて異なるレベルのテストを受験させるか，またはTOEIC®のように能力差が大きい受験生にも対応できるテストを受験させることになる。前者の場合，学生の申告に基づきレベルを設定することになるだろうが，これは極めて煩雑な作業である。反面，単一のテストで全受験生に対応すると，2時間という試験時間が長すぎたりレベルが難しすぎたりで，途中で受験を放棄する場合もある（藤森，2012）。

小泉（2011a）による分類の第3である，外部試験の一部をプレイスメントのために利用する例は，著者らが知る限り報告されていない。そのような形の利用には著作権の問題が発生するため，少なくとも公然と実施するのは難しいはずある。

新テストの開発

コンセプトとデザイン

このような現状を踏まえ，新たな「外部テスト」として，日本の大学での英語授業の習熟度別編成を的確にかつ効率的に行うことができるテストを開発することを計画した。新テストは厳密に等化された複数の平行フォームを持ち，同一の学生の英語力の経時変化を測定できるものでなくてはならないと考えた。また従来の外部テストよりも幅広い層の受験者に対応しながら試験時間を短くすることで試験監督者および受験者への負担を軽減するものとした。具体的には普通教室で1コマ90分間内に実施でき，結果が数日内にフィードバックできるという実用性上の条件を設定した。この条件を満たすのはリスニングとリーディングを多肢選択形式で出題し紙ベースで解答するテストのみである。

リスニングは，音声を知覚し，単語を特定し，文構造を解析し，文や発話の意味を解釈する過程である。これは語彙的，音韻的，形態論的，統語的，談話的，語用論的知識だけでなく，発話内容に関する背景知識が複雑に自動的に統合されて可能になる（Buck, 2001; Field, 1998; Flowerdew & Miller, 2005; Kelly, 1991）。同様に，リーディングも，テキストから得られる語彙的，形態論的，統語的，談話的，語用論的知識と読み手が持つテキスト内容に関する背景知識が自動的に融合して可能になる（Grabe, 2009; Jiang & Grabe, 2007; Mondria & Wit-de Boer, 1991; Pritchard, 1990; Segalowitz & Segalowitz, 1983）。テストで用いられるタスクタイプには，リスニング，リーディングともに，多肢選択式，空所補充式，真偽判定式，短答式などがある（小泉, 2011b; 中川, 2011）が，このうち標準テストで最も多いのは多肢選択式である。

表1
新テストの細目

	問題形式	項目数
L1	日本語の語句を聞き，それに相当する英単語を，聴覚提示される4選択肢から選ぶ。	20
L2	短い英文を聞き，指定された位置の語を，視覚提示された4選択肢から選ぶ。	20
L3	ある程度長い英文を聞き，ビープ音で置換された語を，視覚提示された4選択肢から選ぶ。	20
R1	日本語の語句を見て，それに相当する英単語を，視覚提示された4選択肢から選ぶ。	20
R2	1語が欠けた非文を読み，指定された1語を文中のどの位置に戻せば正文となるかを，4選択肢から選ぶ。	20
R3	30～80語程度の英文に設けられた空所に補充すべき語句を，4選択肢から選ぶ。	20

以上を踏まえ，表1に示すような3種類のリスニング問題形式と3種類のリーディング問題形式を持つテストを開発することを決定した。テストの名称は，英語力を可視化する（Visualizing English Language Competency）という目標を込め，VELC Test® とした。リスニングパート1（以下L1）とリーディングパート1（以下R1）では，それぞれ聴覚語彙サイズと視覚語彙サイズを測定することにした。聴覚と視覚では語彙サイズが異なる可能性があると考えたからである。目標語はJACET8000のレベル1～7から幅広く選定し，問題形式は望月テスト（望月, 1998; 望月他, 2003）をベースにした。リスニングパート2（以下L2）は音声の連続体を意味のある塊に切り分ける能力を測ることを目標にした，いわば部分ディクテーションの多肢選択形式版である。リスニングパート3(以下L3)は一種のクロー

ズテスト（Oller, 1979）のリスニング版（靜, 2011）であり総合的な聴解力の測定を狙いとした。リーディングパート2（以下R2）は一種の不可視空所補充（invisible-gap filling）テスト（Shizuka, 2004; 靜, 2008）で，長めの文の構造を正しく解析する能力を測定することを狙った。リーディングパート3（以下R3）はL3と同じくクローズテストの一種である。

開発に使用した測定モデルと結果のフィードバック形式

複数の等化フォームの開発には項目応答理論モデル群のうちのいずれかが必要であるが，本テストではラッシュ・モデル（Rasch, 1960; Wright & Stone, 1979; Bond & Fox, 2007; Smith & Smith, 2004）を選択した。まず理想の測定モデルを設定しそれに適合する項目を集めるラッシュ・モデリングのアプローチ（Wilson, 2004; 2005）が新テストの開発にはふさわしいと判断したからである。ラッシュ・モデリングは受験者能力値としてロジット（logits）を単位とする値を算出するが，成績のフィードバックには平明さを優先してロジット値を一次変換した整数（VELCスコアと呼ぶ）を用いることとした。具体的には，下で述べる試行時のデータの全国平均が500，標準偏差が100になるように調整した値を，L1からR3までの6つのVELCパートスコア，リスニング（以下L）とリーディング（以下R）の2つのVELCセクションスコア，そして120項目すべてにもとづくVELCトータルスコア（以下TTL）の9つの数値を算出するものとした。ラッシュ・モデリングでは項目難度が予め判明しているテストなら事前に素点／VELCスコア換算表を準備しておくことができるため，この方法で迅速なフィードバックを行うこととした。

開発の過程および解答データの集積

テストの開発は概ね以下のような手順で行い，その過程で解答データが集積された。

項目案の作成　L1，R1の項目は第二著者が作成し，第一著者がチェックした。目標語はJACET8000のレベル1～レベル8から選び，選択肢は当該目標語のレベルと同等以下の語彙から選んだ。L2，L3，R2，R3に関しては英文素材は母語話者（日本在住の大学英語教員4名）に書き下ろしを依頼し，その英文を主として第一著者が問題化し第二著者がチェックした。いったん作成したL2，L3，R2，R3の項目案に対して，日本人大学英語教員4名および母語話者英語教員4名にフィードバックをもらい，適宜修正を加えた。

第1次試行　日本人大学生（N = 2,861）に，例えばL1とL2のみなど特定パートごとに項目案を試行し（2010年9～10月），結果をラッシュ・モデリングのソフトウェアであるWinsteps 3.75.0（以下バージョン情報は略）によって分析し，項目難度，Infit Mean Square等の指標で総合的に判断しながらモデル適合度が高くかつ幅広い難度を持つ項目群を選定した。

第2次試行　第1次試行の結果から適合度が良く難度が安定していると判断された項目群をリンクとして選び，それに未実施項目案および第1次試行での項目分析に基づいて修正した項目案を，別の大学生集団（N = 1,101）に試行した（2011年4～5月）。結果を再びWinstepsで分析し，モデル適合度がよく難度幅の広い項目群を選定した。

暫定フォームの作成　第1次と第2次の試行の結果，同一パート内の全項目は同一の難度尺度上に配置されたので，その項目難度値を用いて，各120項目から成る難度の等しい複数フォームに暫定的に組み上げた。

第3次トライアル　組み上げた暫定的な複数フォームをさらに別の集団（N = 1,620）に対して試行した（2011年9～10月）。

最終フォームの確定　計3回の試行解答データ（合計N = 5,580）の全てに対して改めて

ラッシュ・モデリングを行い，その結果で等化した複数フォームを，項目内容のバランス，リスニングセクション所要時間の平準化等にも配慮しながら，最終的に確定した。

完成フォームの試験的実施　完成した複数フォームのうちForm A（およびごく一部ではForm B）は，2012年の1月から12月にかけて計32の大学（高等専門学校含む）で試験的に実施され，その解答データ（N = 12,366）が集積された。

テストの妥当性検証

テストが妥当性を持つためには信頼性が高いことが必要条件である（Linn & Gronlund, 2000; 渡部, 2003など）。そこでまず（1）受験者能力推定の信頼性を検討する。つぎに（2）項目難度の安定性（不変性）という観点からの項目難度の信頼性を確認する。その上で妥当性に関しては，まず（3）異なる大学間で異なる熟達度が検出されたかという観点での基準関連妥当性を検討する。つぎに（4）TOEIC® スコアがどの程度予測できるかという観点での基準関連妥当性を検討する。最後に（5）共分散構造分析によって本テストが測定する構成概念の妥当性を検討する。以上5つのアプローチの中で，(4)以外については完成フォームの試験的実施で集積したデータを使用し，（4）については開発途中の試行時に集積したデータを使用する。これはTOEIC® データが収集できたのが試行時の受験者集団のみであったためである。なお，ラッシュ・モデリング，構造方程式モデリング以外の分析には統計ソフトウェアJMP® 10.0.2を使用した。

受験者能力推定の信頼性

分析対象データ　2012年度内に本テストのForm Aを受験したのは32大学（高等専門学校含む）に在籍するのべ12,366名であったが，この中には同Formを2度あるいは3度（数ヶ月の間隔をおいて）受験したグループを一部含んでいた。また大学によって受験者数は19名～8,100名までと大きくばらついていた。そこで同Formの初回受験者が100名以上いた18大学からそのような受験者を無作為に100名ずつ抽出した合計1,800名（100名×18大学）による解答データを対象として，Winstepsにより分析した。

テスト全体としての得点の信頼性　スコアの信頼性とは観測スコアの分散に占める真のスコアの分散の割合である。素点に関して通常報告されるクロンバックのアルファ係数は，ラッシュ・モデリングによる場合は，Winstepsのアウトプットにある受験者信頼性（person reliability）値がそれにあたる。これは受験者能力の推定値の分散に占める真の能力値の分散の割合（の推定値）である。受験者信頼性値は1.00が上限であるため見かけ上の天井効果が出やすいので，それを改善した受験者分離（person separation）も報告される。この指標の値は信頼性をもって分離できる受験者層の数を表す。また受験者に関する信頼性とパラレルな概念として，項目に関しての信頼性の指標である項目信頼性（item reliability）と項目分離（item separation）も報告される。以上の指標を表2に示す。

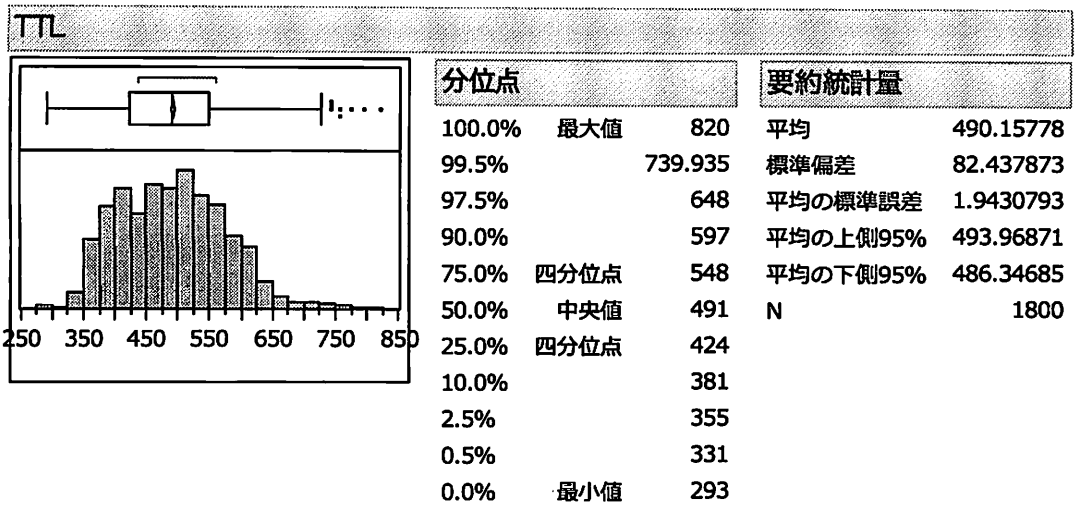
表2
受験者能力値の信頼性，項目難度の信頼性に関わる指標の値

	項目数	*N*	PR	PS	IR	IS
TTL	120	1800	0.94	4.04	1.00	18.13
L	60	1800	0.86	2.53	1.00	18.05
R	60	1800	0.91	3.26	1.00	17.42
L1	20	1800	0.68	1.46	1.00	19.29
L2	20	1800	0.68	1.40	1.00	18.54
L3	20	1800	0.72	1.59	0.99	12.72
R1	20	1800	0.75	1.72	1.00	17.92
R2	20	1800	0.76	1.77	0.99	13.26
R3	20	1800	0.78	1.90	0.99	13.20

注　PR：受験者信頼性　PS：受験者分離　IR：項目信頼性　IS：項目分離

受験者能力値に関しては，120項目に対する解答で推定されるTTLの信頼性が当然最も高く.94であり，これは受験者分離に換算すると4.04である。すなわちTTLに関しては約4段階の受験者レベルを分離することができている。項目難度に関しては受験者数が多いこともあって推定誤差が非常に小さい。項目信頼性はほぼ1.00であり，項目分離は18.0を超えている。これはすなわち，信頼性をもって難度の異なる項目がテストを構成していて，最も易しい項目から最も難しい項目まで，全体で18段階の項目レベルが分離される，ということを意味する。本テスト全体では非常に幅広い範囲の難度をもつ項目によって構成されており，かつ十分に高い信頼性をもって受験者能力の分離ができている，と言える。

受験者能力別の得点信頼性　古典的テスト理論と比較した時のラッシュ・モデリングの強みのひとつは，得点全体の信頼性を単一の数値で要約した指標である受験者信頼性や受験者分離に加えて，受験者の能力バンドごとの得点信頼性を標準推定誤差という形で示せることである。能力値の標準推定誤差は当該能力値とその推定に使用された項目群の難度との関係で決まる。すなわち任意の項目群の平均難度に近い位置にいる受験者の能力は，そこから遠い位置にいる受験者の能力よりも高い精度をもって推定される。よって，受験者の能力が当該テストの難度に適合していればいるほど推定精度は高く，当該テストが易し過ぎるもしくは難し過ぎる受験者の推定精度は相対的に低くなる。



分位点		
100.0%	最大値	820
99.5%		739.935
97.5%		648
90.0%		597
75.0%	四分位点	548
50.0%	中央値	491
25.0%	四分位点	424
10.0%		381
2.5%		355
0.5%		331
0.0%	最小値	293

要約統計量	
平均	490.15778
標準偏差	82.437873
平均の標準誤差	1.9430793
平均の上側95%	493.96871
平均の下側95%	486.34685
N	1800

図1　TTLの得点分布のヒストグラム，箱ひげ図，分位点，および要約統計量

この意味で，分析対象とした1,800名の下位者から上位者までの異なる能力層に対してそれぞれどの程度の誤差が見込まれるか，言い換えればどの程度の信頼性が見込まれるかを明らかにしておくことは意義がある。そこでTTLの得点分布（図1）と，得点の関数としての標準推定誤差（図2）をつきあわせてみる。[2]

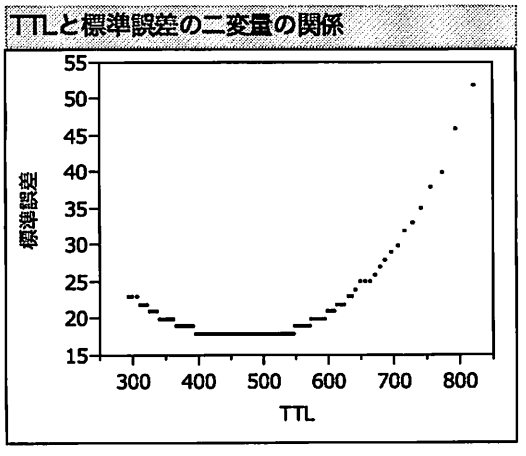


図2　TTL得点の関数としての標準推定誤差

図1の分位点情報から，受験者の95%が355～648の範囲に，最も中核的な50%が424～548の範囲に存在したことがわかる。そして図2はTTLが400～550の範囲において標準誤差が18と最も推定精度が高く，それよりも得点が低くあるいは高くなるにつれて徐々に精度が低くなることを示している。

図1と図2の情報を統合するならば，受験者の中核的な50%（424～548）が最も小さな18～19の誤差で，最も能力が低い2.5%と最も能力が高い2.5%を除いた残りの95%（355～648）は18～25という小さな誤差で推定されていることがわかる。TTLが650を超えると誤差は徐々に大きくなり，700では約30,800では約50になる（図2）。しかしその範囲に存

在する受験者は非常に限られている（図1）。すなわち本テストは，本テストを受験する日本人大学生の中核的な層に適切に項目難度を合わせており，それによって信頼性のある測定が可能になっていると言える。

項目難度の不変性

テストが信頼性を持つための必要条件には項目難度の不変性も含まれる。この場合不変性とは受験者集団が変わっても項目同士の相対的難度が変わらないことを意味する。テストを構成する項目の難度が安定していてはじめてテストが受験者能力を安定的に推定できるからである。本テストは項目難度の値を試行段階のデータ（N = 5,580）に基づいて推定された値に固定（anchoring）して受験者能力を推定する。それは試行段階の解答データによって十分安定した項目難度が得られていると考えられるからである。

しかし試みにその固定を一時外し難度が未決定の状態に戻した上で，異なる2つの受験者集団のデータによって別々に推定した一対の項目難度のセットを比較してみるのは価値あることである。その状態でほぼ変わらない項目難度が得られれば，項目難度の安定性のひとつの証拠となるからである。そこで，2012年度データから上述の手順で抽出した1,800名をランダムに900名ずつA群，B群に分けWinstepsによって改めて項目難度を推定した。なおその際，UPMEAN=0コマンドにより，各分析での受験者能力値の平均を0.0とする制約を設けた。図3に，別々に推定した難度値の散布図を示す。

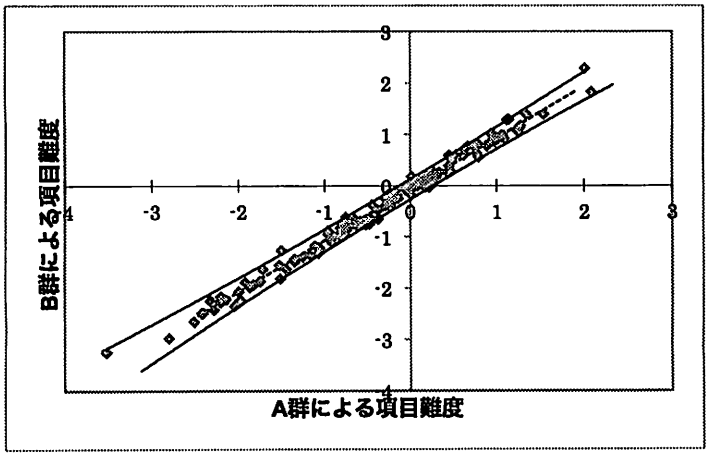


図3　ランダム分割900名ずつのサンプルで別々に推定した項目難度

真の項目難度が等しくても，推定には誤差が必ず伴うので，完全に$y = x$の1直線上に並ぶことはない。しかし推定誤差の範囲内でその直線の周辺に位置するはずである。図を見るとほぼすべての項目が$y = x$の直線周辺に分布していることが見て取れる。相関係数はr = .994である。図には項目を示す点群を左上と右下から挟むように2本の曲線が描かれているが，これは95%信頼性区間の境界を示す。2つの集団にとって項目群が同一の難度を持っていれば少なくとも95%の項目がこの2つの曲線の間の空間に分布し，かつ項目を示す点の近似直線がX軸とY軸の交点を通る（Bond & Fox, 2007, p.86）はずである。図3では全120項目中，曲線の外側にわずかだが完全に出ているのが1項目，ほぼ曲線上が3項目程度であり，2つの集団にとっての項目難度は等しいと考えて差し支えないと言える。念のため，これら2セットの難度行列と実際のフィードバックに用いる値を固定した難度行列

の相関係数も確認したところ，いずれも $r = .971$ であった。以上により本テストの項目難度は受験者集団に左右されることなく安定しており，この意味でのテストの信頼性も高いと言える。

異なる大学間の差の検討

次に，受験者が所属する大学によってどの程度スコアが異なっていたかを検討する。今回分析した 18 大学には北海道から九州までの国立，公立，私立が含まれており，学部も文系，理系さまざまである。当然英語力もさまざまであると考えられる。これらの大学ごとに明らかに異なるスコアが得られていれば，非常に間接的ではあるがテストの基準関連的妥当性を示唆していると考えてよいだろう。表 3 に TTL の平均値と標準偏差を大学別に示す。U03, U04 などは大学を示す無作為のコードである。

平均値の最高は U17 の 612.7 で，最低は U27 の 393.5 であり，2 つの平均値の間には 219.2 という大きな差があった（$d = 3.98$，効果量大）。分布の状況を大学別にヒストグラムで表したのが図 4 である。明らかにいくつかの大学間には有意でかつ大きな差がありそうである。念のために分散分析を行うこととし，まず等分散性の検定のためレーベン検定を行ったところ，等分散性であるという帰無仮説は棄却された，$F\ (17,\ 1782) = 7.94$，$p < .0001$。そこで等分散を仮定しないウェルチ分散分析を行ったところ，$F\ (17,\ 663.28) = 179.9$，$p < .0001$ で有意であることが確認された。つまり視認からも明らかなように，18 大学の平均値には有意な差が存在した。

表 3
大学別の TTL の記述統計

	U03	U04	U07	U08	U09	U11	U12	U13	U17
Mean	415.9	577.4	430.0	513.2	480.3	500.9	542.0	437.0	612.3
SD	50.5	45.9	61.6	51.7	47.2	46.7	54.5	51.6	70.6
	U18	U20	U22	U23	U24	U26	U27	U28	U30
Mean	531.6	568.6	434.8	459.9	481.4	396.5	393.5	520.4	527.1
SD	47.8	63.5	57.7	57.2	56.9	32.1	32.6	59.1	68.5

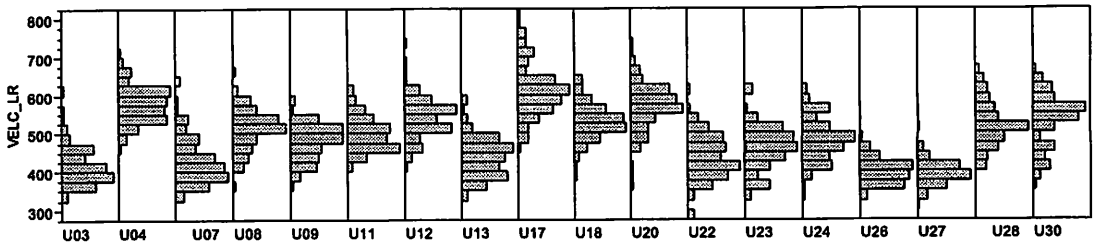


図 4　大学別の TTL のヒストグラム

つぎにどの大学とどの大学の間の差が有意であるかを調べるために，Tukey-Kramer HSD 検定を行った。結果を表 4 に示す。B，C などの文字が縦に並んでいるが，これは同じ文字で結ばれた大学間は有意な差があるとは言えないことを表す。例えば，B で結ばれた U04 と U20 の間には有意な差があるとは言えないが，U17 と U04 は同じ文字で結ばれていないので有意な差がある，のように読み取る。すると U17 のみは他のどの大学よりも平均点が

有意に高かったことがわかる。他の大学に関しては，同じ文字で結ばれている大学群がひとつの能力帯を表すと考えると，これらの18大学は，間に重複はあるがおおよそ11のグループ（A～K）に能力が分かれたことがわかる。これは本テストが，参加者集団の間にある英語力の違いを検知したと解釈するのが順当であろう。すなわちテストの妥当性が非常に間接的ながら示唆された。

表4
18大学のTTL平均値の多重比較の結果

Level											Mean
U17	A										612.3
U04		B									577.4
U20		B	C								568.6
U12			C	D							542.0
U18				D	E						531.6
U30				D	E	F					527.1
U28				D	E	F					520.4
U08					E	F					513.2
U11						F	G				500.9
U24							G	H			481.4
U09							G	H			480.3
U23								H	I		459.9
U13									I	J	437.0
U22									I	J	434.9
U07										J	430.0
U03										J	K	416.0
U26											K	396.5
U27											K	393.5

TOEIC® データとの基準関連妥当性

分析対象データ　上述の理由により，この分析に関してのみ最終フォームを確定する以前の第3次試行で収集されたデータを用いる。第3次試行は120項目から成る完全フォームによって行ったので全受験者1,620名についてL1～R3の6つのパートスコアが算出された。[3] この中でTOEIC®のトータルスコア（TOEIC_TTL），リスニングスコア（TOEIC_L），リーディングスコア（TOEIC_R）が判明した375名を分析対象とした。

変数についての予備的確認　L1，L2，L3，R1，R2，R3を予測変数としてTOEIC_LとTOEIC_Rそれぞれを目標変数とする重回帰分析に先立ち，変数の記述統計を確認した。

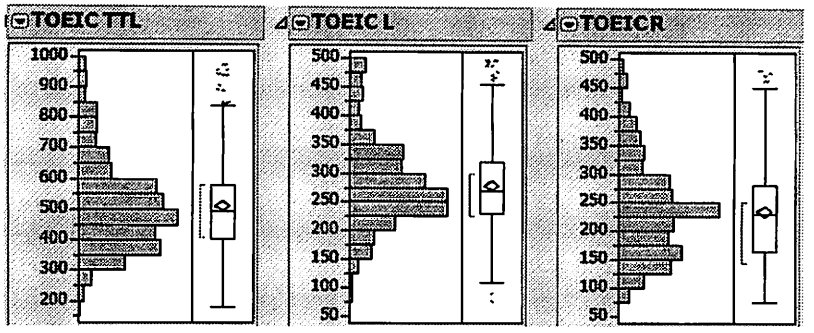


図5　目標変数のTOEIC® スコアの分布

まず収集されたTOEIC® スコアの分布を図5に示す。いずれも重回帰分析の適用を不適とするほどの極端な歪み等はないことが視認できる。次にこれらのスコアの平均値と，公開されているTOEIC® の全国平均値を表5に示す。

表5
TOEIC® 平均点に関する本研究データと全国データの比較

	TOEIC_TTL	TOEIC_L	TOEIC_R
本研究のデータ（N = 375）	514	279	235
2011公開テスト：大学生（N = 301,996）	574	304	250
2011 IPテスト：大学生（N = 411,085）	445	250	197

本研究のデータの数値はTOEIC_TTL，TOEIC_L，TOEIC_Rいずれも大学生の公開テスト平均よりも低く，IPテスト平均よりも高い。本研究のデータの中の公開テストスコアとIPテストスコアの比率は不明であったので確かなことは言えないものの，大学生サンプルとして全国の平均から大きく外れていると考える理由はないと思われる。

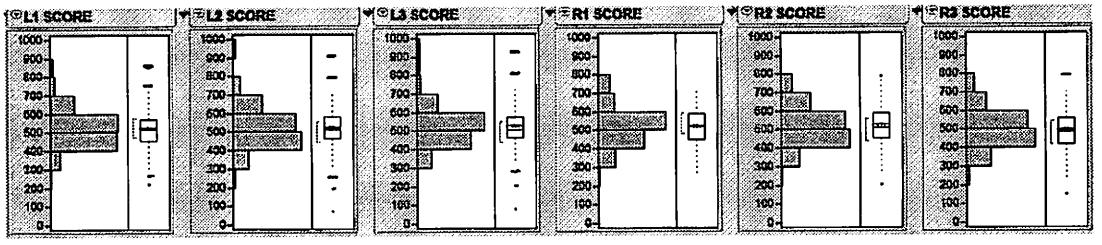


図6　TOEIC® スコアの予測に用いた予測変数であるVELCパートスコアの分布

次に予測変数VELCパートスコア（L1 ～ R3）の分布を図6に示す。正規分布からの過度の逸脱はないことが視認できる。

表6
VELCパートスコア間の相関係数

	L1	L2	L3	R1	R2	R3
L1	1.000					
L2	.569	1.000				
L3	.609	.693	1.000			
R1	.717	.503	.537	1.000		
R2	.672	.636	.667	.685	1.000	
R3	.647	.617	.654	.660	.740	1.000

次に予測変数間の相関係数を表6に示す。係数は.503（R1-L2）から.740（R2-R3）までである。すべて有意な正の相関であり，かつ高すぎるものはない。よって多重共線性の心配はなく，重回帰モデルの予測変数に投入する候補として問題はない。

回帰分析の結果　目標変数および予測変数の適切性が確認できたのでステップワイズ法によって重回帰分析を行った。Cp統計量を目安にしてTOEIC_Lを予測するために最終的に採用したモデルを表7に示す。自由度調整済み決定係数は.578である。

表7
本テストでTOEICリスニングを予測する重回帰モデル

変数	係数	標準誤差	*t*値	*p*値
Intercept	-74.886	16.893	-4.43	<.0001
L1	0.075	0.038	1.97	.0498
L2	0.199	0.037	5.38	<.0001
L3	0.248	0.037	6.70	<.0001
R3	0.119	0.039	3.04	.0026

同様にTOEIC_Rを予測するための最終モデルを表8に示す。

表8
本テストでTOEICリーディングを予測する重回帰モデル

変数	係数	標準誤差	*t*値	*p*値
Intercept	-199.599	18.090	-11.03	<.0001
L1	0.075	0.042	1.76	.0794
L2	0.079	0.038	2.10	.0365
L3	0.148	0.038	3.89	.0001
R1	0.109	0.045	2.40	.0170
R2	0.174	0.045	3.90	.0001
R3	0.212	0.044	4.85	<.0001

このモデルの自由度調整済み決定係数は .633 である。(L1 の p 値が .05 を上回っているが，第二種の過誤を避けるために保持した。) つまり本テストによって TOEIC_L の分散の約 58% を，TOEIC_R の分散の約 63% を説明することができた。TOEIC_TTL は，TOEIC_L と TOEIC_R の単純な和なので，これらのモデルによる TOEIC_L 予測値と TOEIC_R 予測値の和を求めて，実際の TOEIC_TTL との相関を計算してみると，r = .825 であった。つまり本テストのデータにより TOEIC_TTL の分散の 68% を説明することができたことになる。予測値と実際の TOEIC_TTL の散布図と，残差プロットを図 7 に示す。残差パタンには異常がないことが視認できる。以上を総合すると，TOEIC® を外的基準とした場合の本テストの妥当性は十分に高いと言える。

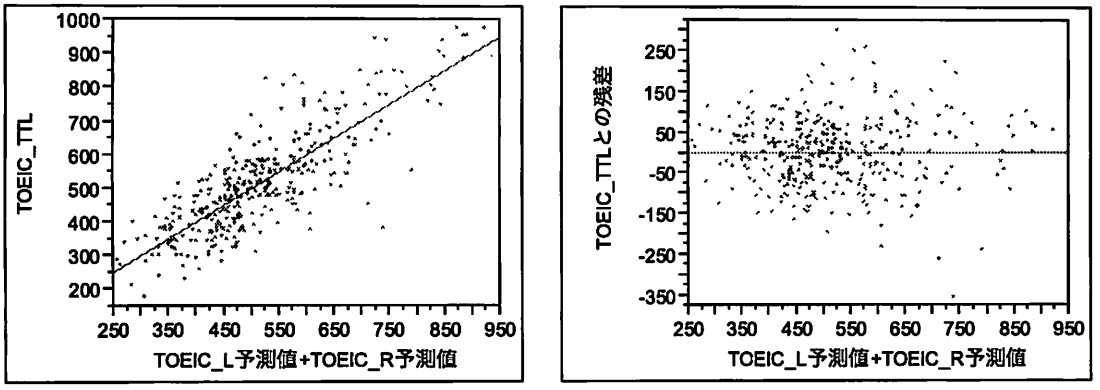


図 7　TOEIC_TTL 予測値／実測値プロット（左）　TOEIC_TTL 予測値／残差プロット（右）

構成概念妥当性

最後に本テストが測っている構成概念を因子構造の面から探るため，上述の試験的実施データ（N = 1,800）の L1，L2，L3，R1，R2，R3 を観測変数として，IBM® SPSS® Amos 20.0.0 を用いた構造方程式モデリング（SEM）を行った。SEM においては単一のモデルを検証するのではなく，複数の競合モデルを立てそれらの適合度を比較するべきとされている (Thompson, 2004, p.115)。そこで図 8 に示すような，理論的に想定される 6 つのモデルをデータに当てはめ適合度を比較した結果を記す。

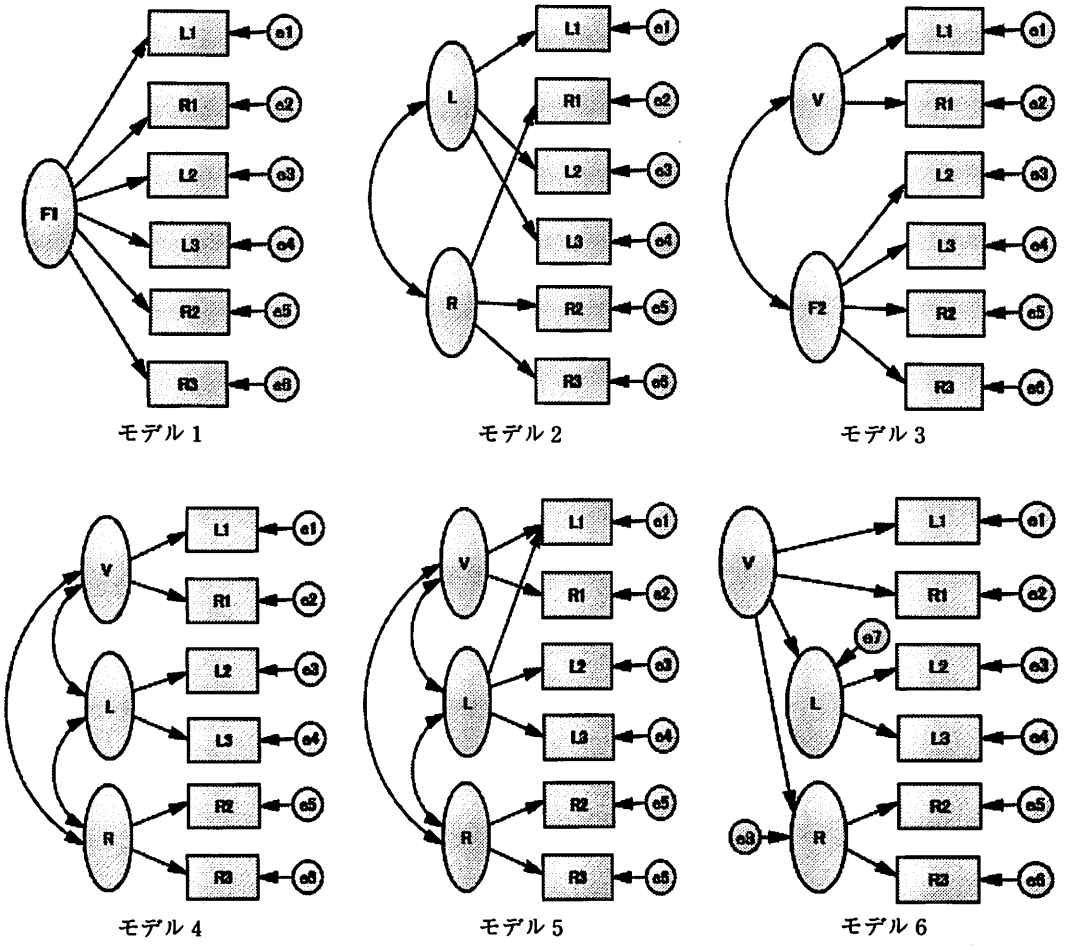


図 8　適合度を比較した 6 つの競合モデル

モデル 1 は 1 因子モデルである。6 つのスコア全体が単一の英語力に影響されていると想定する。全 120 項目が何らかの英語力を測る以上，このモデルがある程度適合せねばならないはずである。モデル 2 は L1，L2，L3 がリスニング因子に，R1，R2，R3 がリーディング因子に影響されていると想定したモデルである。リスニングとリーディングは密接に関わるが分離可能な構成概念だと考えられる（In'nami & Koizumi, 2012）のでこのモデルもある程度適合すると期待された。モデル 3 はもうひとつの 2 因子モデルである。文脈なしで語彙について問う L1，R1 とそれ以外のパートは別の因子による可能性もあると考えた。モデル 4 は確認的因子分析（CFA）に先立って行った探索的因子分析（EFA）での 3 因子解の因子負荷量にもとづいて設定したものである。L1 と R1 が語彙因子に，L2 と L3 がリスニング因子に，R2 と R3 がリーディング因子に影響されると想定した。モデル 5 はモデル 4 に加えてリスニング因子から L1 へのパスも引いたものである。[4] モデル 6 はやはり 3 因子モデルだが，語彙因子が L1 と R1 に影響するだけでなく，リスニング因子とリーディング因子を通じてすべての変数に影響すると想定したモデルである。語彙はリーディングにせよリスニングにせよあらゆる言語活動のもと（building block）であるため，このモデルも

妥当であると考えた。なお因子が複数あるモデルにおいては因子間の相関を想定した。[5]

表9
6モデルの適合度指標

	Chi-sq.	*df*	*p*	GFI	AGFI	CFI	RMSEA	SRMR	AIC
モデル1	435.782	9	.000	.923	.820	.942	.162	.042	459.782
モデル2	403.601	8	.000	.926	.805	.947	.166	.042	429.601
モデル3	223.053	8	.000	.957	.888	.971	.122	.030	249.053
モデル4	58.393	6	.000	.990	.963	.993	.070	.014	88.393
モデル5	7.744	5	.171	.999	.994	1.000	.017	.005	39.744
モデル6	241.263	7	.000	.958	.874	.968	.136	.031	269.263

適合度の指標を表9にまとめて示す。カイ2乗値が有意であることはモデルがデータと有意に異なることを表すが，サンプル数が多い場合には重視されない。ただ特筆すべきはモデル5のみ有意でない（つまりモデルがデータに適合している）ことである。GFIはすべてのモデルで.90以上なのでいずれの場合も「説明力がある」（豊田，2007，p. 18）パス図であると判断される。AGFIが.95を超えているのはモデル4と5である。CFIについては.95を超えているのはモデル3，4，5，6である。RMSEAは.05以下が「当てはまりが良く」，.10以上が「当てはまりが良くなく」，その間が「グレーゾーン」と判断される（豊田，p.18）ので，モデル5が当てはまりが良く，モデル4がグレーゾーンにあり，モデル1，2，3，6は当てはまりが良くない，と解釈される。SRMRは.08以下ならば「当てはまりがよい」（竹内・水本，2012，p.198）と判断される。よってSRMRによればすべてのモデルの当てはまりは良く，なかでも最も良いのはモデル5で，次がモデル4である。AICは絶対的大きさでなく数値の差に意味があるので，モデル5が最も適合し，次にモデル4の当てはまりが良いと解釈できる。

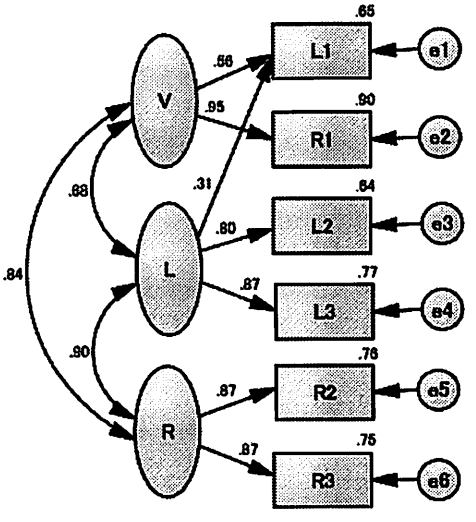


図9　モデル5のパス係数

以上を総合すると，これらの中で最も適合度がよいのはモデル5で，以下，モデル5＞モデル4＞モデル3＞モデル6＞モデル2＞モデル1であると解釈できる。ただしこの中での適合度が最も低いモデル1でもGFIが.90を超え，SRMRも.08を下回っているので，これら6つのモデルはいずれも「ある程度（以上）の説明力はある」と解釈できる。本研究での最良モデルとして採用するモデル5のパス係数等を図9に示す。

考察

以上の分析の結果は以下のようにまとめることができる。

(1) 本テストは日本人大学生を対象としたとき，全体として十分に高い信頼性を示した。受験者能力の弁別に関わる指標である受験者信頼性係数はテスト全体で.94，リスニングのみで.86，リーディングのみで.91であった。また受験者能力別にみると，受験者層の最も厚い能力帯において最も正確な測定がなされるようにテストの難度が設定されていることが確認された。

(2) 本テストは易しい項目から難しい項目までの幅が広く，その難度は安定している。まず項目難度の弁別に関わる項目信頼性係数はテスト全体でもパート毎でも1.00もしくは0.99であった。項目分離係数に換算すると12.71～18.13であり項目難度の範囲が非常に広い。これは能力のかなり低い受験者であっても能力推定ができないほどの低得点はとりにくく，逆に能力のかなり高い受験者であって能力推定ができないほどの高得点はとりにくいことを示唆する。また無作為に分割した2サンプル間で改めて推定した項目難度の相関は極めて高く（$r = .994$），推定誤差の範囲内において項目難度が不変であるといってよい。よって信頼性のある測定のための必要条件を備えていると言える。

(3) 本テストは大学生の間に存在する英語力の個人差を検知することができると解釈できる結果が得られた。分析した18大学の平均スコアには有意で大きな差があり，重なりのある11の能力バンド，重なりのない6つの能力バンドが明らかになった。この結論は18大学の受験者集団の能力は同一ではなかったという前提に立ったものであり，より確かな結論を得るためには当該受験者の英語力に関する別の指標との比較が必要であることは言うまでもない。しかし大学受験時の予備校模試などで明らかになっている最低合格者の偏差値の差の存在などと考え併せると，サンプル間の能力差を本テストが検知したという解釈を疑う合理的な理由はないように思われる。

(4) 本テストはTOEIC®を基準とした時にも十分に高い基準関連妥当性を示した。本テストは特別にTOEIC®を模して開発されたものではなく，項目数もTOEIC®の200に対して120，所要時間もTOEIC®の120分に対して70分と相対的にはかなり短い。それでもTOEIC®総合スコアとの相関が.82を超え，分散の68%を説明したことで，本テストはTOEIC®スコアを予測するための代替テストとしても十分な実用性を備えていることが示されたと言える。逆に説明されなかった32%はTOEIC®と本テストが異なっている部分（それぞれの誤差分散を含む）であり，両テストの項目タイプ・内容の違いを考えれば当然であろう。

(5) 最後に，本テストの観測変数間の関係，潜在因子間の関係に，理論上のいくつかの想定を支持するパタンが観察されたことにより，テストの構成概念妥当性に関する複合的な証拠が得られた。第1に，本テストは全体として英語のテストであり，大きなくくりで「英語力」という一つの構成概念を測定しているはずであるが，この想定はモデル1がある程度は適合したことで支持されている。すなわちTTLという単一の値でのフィードバックには一定の妥当性があるとの裏付けが得られた。第2に，リスニングセクションとリーディ

ングセクションは，関連しながらも分離可能であるふたつの構成概念を測定しているはずだが，この想定はモデル2の適合度も悪くないことにより裏付けられる。これによりLとRというそれぞれ単一の値でのフィードバックにも一定の根拠があることとなる。第3に，語彙は知識であり，技能であるリスニング，リーディングとはやや次元が違うものであるとも考えられるが，この想定は，モデル3の適合度で支持されている。本テストの語彙パートは文脈のない語彙問題形式なので，その特徴が現れたと考えられる。第4に前述のリスニングとリーディングの分離，知識と技能の分離の両方を組み合わせるならば，語彙知識に関わるパート（L1，R1），リスニング技能に関わるパート（L2，L3），リーディング技能に関わるパート（R2，R3）はそれぞれ別の因子に関わるはずだが，モデル4の適合度が非常に良いことが，この想定を支持する。第5に，リスニング形式の語彙項目であるL1は，主として語彙因子に関わりながらそこにリスニング因子もわずかに関わると考えられるが，それを表現したモデル5の適合度が極めて良好であったことは，この想定を裏付けるものである。第6に，語彙知識はすべての英語技能の基礎であり，リスニング力にもリーディング力にも影響を与えるはずであると想定される。モデル6がある程度の適合度を示していることで，この想定も裏付けられたと言える。

最後に，本研究でモデル4と5の適合度が最も良かったということは，本テストのスコアレポート方法について示唆するものがある。それは，現行のLとRの2種類のセクションスコアに加え，V（語彙）というセクションスコアを設定してフィードバックすることも妥当であろう，という点である。その場合，モデル4に基づくならば，VはL1とR1の合計40項目，LはL2とL3の合計40項目，RはR2とR3の合計40項目によって算出することになろうし，モデル5に基づくならば，VはL1とR1の合計40項目，LはL1，L2，L3の合計60項目，RはR2とR3の合計40項目によって算出することになろう。いずれにせよ，Vとして独立したセクションスコアを提示することは，受験者の今後の学習に対するヒントとしても有効と思われる。

なお，TOEIC®には，リスニングとリーディングの2因子モデルの適合度が最も良かったと報告されている（In'nami & Koizumi, 2012）が，3因子モデルの適合度が最も良かった本テストとの違いは，両テストの問題形式を考えれば納得できよう。TOEICにも本テストにもリスニングセクションとリーディングセクションがあるが，文脈から切り離した純粋な語彙問題があるのは本テストだけだからである。

本研究の限界に基づく今後への展望

本研究では全体としてVELC Test®の妥当性をある程度まで確認する結果とはなったが，以下に述べる限界を克服したさらなる研究が必要である。まず受験者の英語力に関するより正確な外部基準をもちいた基準関連妥当性の検証が必要である。本研究では1,800名のスコアを分析して大学別に平均点の差を検出したが，これだけではテストの妥当性を示す強い証拠とは言えない。またTOEIC®を予測するために用いたVELCスコアは試行時データから算出した値であって，完成フォームを受験した学生から得られたものではない。また入手できたTOEIC®データの正確性は受験時期等も含め厳密には担保されていない。実際に本テストを受験した学生の直近の別テストスコア，授業担当教員からの評価など，より直接的な外部基準が望まれる。また受験者の性別，学年，専攻等の属性情報が利用できれば属性別の差異項目機能（DIF）の有無なども検証できる。

また本研究でのSEMは1,800名サンプル全体に対するもののみであったが，In'nami and Koizumi (2012) のようにサンプルを2分割しての2母集団同時分析を行えば，本テストの

因子構造についてさらに確かな知見が得られると思われる。

次に本研究で行った因子分析は，L1，L2，L3，R1，R2，R3という6つのパートスコアに基づくものであったが，120の項目それぞれの応答データに基づく分析も将来的には必要である。項目レベルのデータからも本研究で採択されたモデルと同じか近いモデルが導きだされれば，VELC Test®の構成概念妥当性はより確かなものになるはずである。

総括すれば，より正確な外部基準に基づく基準関連妥当性の検討，より詳しい受験者属性情報にもとづくDIFの検証，多母集団分析，項目レベルの因子分析にもとづく構成概念妥当性のさらなる確認などの方向に研究を進めていく必要があろう。

謝辞

VELC Test®のeポートフォリオ・システムを設計された吉成雄一郎氏（東京電機大学），テストのドラフトに対してフィードバックをくださった杉森直樹（立命館大学），竹内理（関西大学），松本佳穂子（東海大学）の各氏，テストの実施を支えたVELC研究会の実務スタッフ，本論文に関して貴重なアドバイスをくださった水本篤氏（関西大学）と熊澤孝昭氏（関東学院大学），そして査読者の方々に心より感謝し，御礼申し上げます。

注

1. 本論文で報告しているテスト開発過程の一部および2012年度実施データの一部の分析は，未公刊論文（靜・望月，2013）としてウェブに掲載している。
2. 得点ごとの標準推定誤差は，Winstepsの Output files メニューの “Score file SCFILE=” で出力される。
3. 試行時のフォームは後の最終フォームとは含まれる項目の組み合わせが異なるが，項目難度は等化されているためパートスコアは完成版によるものと比較可能である。
4. 同様にさらにリーディング因子からR1へのパスを加えたモデルは識別されなかった。
5. 高次因子分析モデルも当てはめてみたが，不適解を生じた。

引用文献

石原知英（2011）「愛知大学名古屋校舎2010年度入学生の英語力の推移―TOEICクラスの運営を中心に」『愛知大学言語と文化』25，1–16.

市原一裕・井上昭雄・佐藤克彦・鈴木章能・長谷川哲子・丸本嘉彦・水口和野（2007）「これまでのプレイスメント・テスト実施を振り返る」『大阪産業大学論集・社会科学編』117，31–75.

小泉利恵（2011a）「プレイスメント・テストの有効性：2種類のテストの比較と学生の反応から」『常磐国際紀要』15，1–15.

小泉利恵（2011b）「リスニングの測定・評価」石川祥一・西田正・斉田智里（編）『英語教育学体系第13巻・テスティングと評価：4技能の測定から大学入試まで』（pp.173–187）．東京：大修館書店.

佐藤敏子・中川武・山名豊美（2003）「習熟度別クラス編成とプレイスメント・テスト」『つくば国際大学研究紀要』9，11–22.

靜哲人（2008）「テスト等による評価」小寺茂明・吉田晴世（編）『スペシャリストによる英語教育の理論と応用』（pp.161–176）．東京：松柏社.

靜哲人（2011）「リスニングテストの作成とその評価」冨田かおる・小栗裕子・河内千栄子（編）『英語教育学体系第9巻・リスニングとスピーキングの理論と実践：効果的な授

業を目指して』(pp.127–145). 東京：大修館書店.
靜哲人・望月正道 (2013)「熟達度診断のための VELC Test：信頼性と妥当性を検証する」http://www.velctest.org/contact/VELCTestdatosei_paper.pdf
鈴木元子・美濃部京子・マーク・シーハン・杉浦香織 (2010)「TOEIC Bridge テストの活用 - 導入結果を踏まえて」『静岡文化芸術大学研究紀要』11, 31–42.
竹内理・水本篤 (編著) (2012)『外国語教育研究ハンドブック：研究手法のより良い理解のために』東京：松柏社.
豊田秀樹 (2007)『共分散構造分析 [Amos 編]- 構造方程式モデリング -』東京：東京図書.
土肥充 (2006)「TOEIC IP による千葉大生の英語力の現状分析」『千葉大学人文と教育』2, 15–29.
土肥充・柳瀬弘美 (2009)「千葉大学における TOEIC IP スコアの包括的分析」『言語文化論叢』3, 31–45.
中川知佳子 (2011)「リーディングの測定・評価」石川祥一・西田正・斉田智里 (編)『英語教育学体系第 13 巻・テスティングと評価：4 技能の測定から大学入試まで』(pp. 222–236). 東京：大修館書店.
藤森吉之 (2012)「習熟度別クラス編成により顕在化された初級レベル大学生の基礎英語力」『白鴎大学論集』26, 237–268.
丸山真純 (2011)「長崎大学経済学部生の英語習熟度 (1) - 二つの英語試験と TOEIC 得点の観点から」『經營と經濟』91, 93–113.
三浦笙子 (2006)「TOEIC 導入の成果を考察する：東京海洋大学科学部 TOEIC クラスにおける教授法と追跡調査をもとに」『東京海洋大学研究報告』2, 57–62.
望月正道 (1998)「日本人英語学習者のための語彙サイズテスト」『語学教育研究所紀要』12, 27–53.
望月正道・相澤一美・投野由紀夫 (2003)『英語語彙の指導マニュアル』東京：大修館書店.
吉永契一郎 (2004)「学生の意識調査と TOEIC の試行結果からみる新潟大学の英語教育」『大学教育学会誌』26, 89–94.
渡部良典 (2003)「妥当性と信頼性」小池生夫・井手祥子・河野守夫・鈴木博・田中春美・田辺洋二・水谷修 (編)『応用言語学事典』(p.771). 東京：研究社.
Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.
Buck, G. (2001). *Assessing listening*. Cambridge: Cambridge University Press.
Field, J. (1998). Skills and strategies: Toward a new methodology for listening. *ELT Journal, 52*, 110–118.
Flowerdew, J., & Miller, L. (2005). *Second language listening: Theory and practice*. Cambridge: Cambridge University Press.
Gorringe, A., & Mazzarelli, S. (2011). Piloting a listening test for placement purposes.『活水論文集英語学科編』54, 15–20.
Grabe, W. (2009). *Reading in a second language: Moving from theory to practice*. Cambridge: Cambridge University Press.
Grogan, M. (2011). An examination of the language centre placement test.『国際文化論集』44, 49–62.
In'nami, Y., & Koizumi, R. (2012). Factor structure of the revised TOEIC® test: A multiple-sample analysis. *Language Testing, 29*, 131–152. doi: 10.1177/0265532211413444

James, R. G. (2009). Placement testing for three university departments.『上武大学経営情報学部紀要』34, 53–60.

Jiang, X., & Grabe, W. (2007). Graphic organizers in reading instruction: Research fingings and issues. *Reading in a Foreign Language, 19*, 34–55.

Kelly, P.(1991) Lexical ignorance: The main obstacle to listening comprehension with advanced foreign language learners. *IRAL, 29*, 135–149.

Linn, R. L., & Gronlund, N. E. (2000). *Measurement and assessment in teaching* (8th ed.). Upper Saddle River, NJ: Prentice-Hall.

Mondria, J. A., & Wit-de Boer, M. (1991). The effects of contextual richness on the guessability and the retention of words in a foreign language. *Applied Linguistics, 12*, 249–267.

Nakamura, Y. (2010). Constructing a large-scale placement test for measuring students' English proficiency.『慶應義塾大学日吉紀要言語・文化・コミュニケーション』42, 1–19.

Oller, J. W., Jr. (1979). *Language tests at school.* London: Longman.

Otomo, A., Niinuma, F., & Maki, H. (2001). The Minimal English Test as a placement test : A preliminary study.『盛岡大学紀要』28, 1–8.

Pritchard, R. H. (1990). The effects of cultural schemata on reading processing strategies. *Reading Research Quarterly, 25*, 273–295.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Denmark Institute for Educational Research (Expanded edition, 1980). Chicago, IL: University of Chicago.

Segalowitz, N. S., & Segalowitz, S. J. (1983). Skilled performance, practice, and the differentiation of speed-up from automatization effects: evidence from second language word recognition. *Applied Psycholinguistics, 14*, 369–385.

Shizuka, T. (2004). Reliability and validity of "invisible gap filling" items. *JLTA Journal, 6*, 108–127.

Smith, E.V., Jr., & Smith, R. M. (Eds.). (2004). *Introduction to Rasch measurement.* Maple Grove, MN: JAM Press.

Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications.* Washington, DC: American Psychological Association.

Wilson, M. (2004). On choosing a model for measurement. In E.V. Smith, Jr. & R. M. Smith (Eds.), *Introduction to Rasch measurement.* (pp. 123–142). Maple Grove, MN: JAM Press.

Wilson, M. (2005). *Constructing measures: An item response modeling approach.* Mahwah, NJ: Lawrence Erlbaum.

Wright, B. D., & Stone, M. H. (1979). *Best test design: Rasch measurement.* Chicago, IL: MESA Press.