熟達度診断のための VELC Test
~ 信頼性と妥当性を検証する ~

靜　哲人
大東文化大学
望月正道
麗澤大学

Abstract
日本人大学生の英語力を可能な限り可視化するための新しい熟達度テストを開発した。リスニングとリーディングの２セクションのそれぞれがさらに３つのパートに分かれ、聴解語彙力（L1）、音声解析力（L2）、聴解での内容把握力（L3）、読解語彙力（R1）、構文解析力 （R2）、読解での内容把握力（R3）を測定する。予備試行の日本人大学生受験者約 5000 人のデータをもとにラッシュモデルによる等化を経て、複数のどのフォームを受験しても同一のスケールでのスコア比較が可能である。信頼性係数は 0.95 を超え、TOEIC スコアとの重相関係数は 0.82 である。結果は、英語力の個人カルテとも言える Web による「e ポートフォリオ」によってフィードバックされる。本テストを定期的に受験することにより、学生は自分の英語力の推移を把握することもできる。これにより、今までの一過性のテストでは実現できなかった様々な方面での利用が期待できる。

I. 開発の目標

従来、日本人大学生の英語力を効率的に測定し結果を迅速にかつ詳細にフィードバックする熟達度テストが存在しなかったと言える。既存のテストには、内容的に大学生にそぐわない、難易度がそぐわない、フィードバックが遅い、などの欠点があった。そのような認識のもとに、それらの欠点を克服することを目標に、新しい熟達度テストの開発に着手した。

目指したのは「日本の大学生の実態にあったプレイスメント・学習効果測定・弱点診断のためのテスト」であったので、VELC (Visualizing English Language Competency) テストと命名した。特に新入生のプレイスメントや在学生の英語力の経時変化の測定に利用することを考え、大学の普通教室で１コマ 90 分間の中で余裕を持って実施でき、かつ結果を数日内にフィードバックできるという物理的な枠を設定してテストデザインを開始した。

II. テスト細目

上記の枠に基づき、まず４技能の中で受容に関わるリスニングとリーディングを多肢選択で測定することとし、この２技能に関わる下位技能で日本人大学生に特に重要と考えられるものに関する検討を重ね、最終的に、表１のようなテスト細目に決定した。

表 1 . VELC Test のテスト細目

パート*	問題形式
リスニング１	日本語の語句を聞いて、それに相当する英単語を、聴覚提示される４つの選択肢から選ぶ。
リスニング２	短い英文を聞いて、指定された位置の語を、視覚提示された４つの選択肢から選ぶ。
リスニング３	ある程度の長さの英文を聞いて、ビープ音によって置換された部分を推測し、視覚提示された４つの選択肢から選ぶ。
リーディング１	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つの選択肢から選ぶ。
リーディング２	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つの選択肢から選ぶ。
リーディング３	ある程度の長さ(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つの選択肢から選ぶ。

注：各パートすべて 20 問、計 120 問

リスニングとリーディングのパート１が測定する構成概念はそれぞれ、聴覚語彙サイズと視覚語彙サイズである。目標語は JACET8000 のレベル１～レベル７から幅広く選定し、問題形式は望月テスト（望月, 1998）をベースにした。リスニングパート２は音声の連続体を意味のあるセグメント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を多肢選択形式にしたものである。リスニングパート３は一種のクローズテスト（Oller, 1979）のリスニング版であり総合的な聴解力が測定できると考えられる。リーディングパート２は一種の invisible-gap filling テスト (Shizuka, 2004)で、長めの文の構造を正しく解析する能力を測定する。リーディングパート３はリスニングと同じくクローズテストの一種であるが空所の補充に当たっては広い文脈が必要となるよう選択肢を工夫するした。

III.開発の過程

語彙リストを利用したパート１を除き、テストに使用する英文素材はすべて研究会の母語話者メンバーが書き下ろした。その際、日本人大学生が必要とする英語力を検討し、理科的、社会的な事物に関するアカデミックな内容を中心とする方針に依った。英文素材を問題に加工したものを研究会内で検討し修正するという作業を経た後、延べ 5000 名を超える日本人大学生を協力者としたトライアル（解答試行）を、次の要領で計３ラウンド実施した。

A 第１次トライアル

複数の協力者集団ごとに、１～３種類の形式の項目候補を試行し、結果をラッシュモデリングのソフトウェアである Winsteps (Linacre, 2005) によって分析し、Infit Mean Squre 等の指標を総合的に判断しながらモデル適合度の高い項目を選定していった。

B 第２次トライアル

第１次トライアルの結果で適合度が良くかつ難易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、それに未実施項目を加えたテストを、新たな協力者集団に試行し、さらに優良項目の数を増やした。第１次と第２次のトライアルの結果、少なくとも同一のパートの中の全項目は不適合項目を除いて同一の難易度尺度上に配置された。

C フォームの組み上げ

暫定的に確定した項目難度値を用いて、フォーム間の難度がほぼ等しくなるように各 120 項目から成る複数フォーム（リスニング、リーディングの全６パートが揃ったもの）に組み上げた。

D 第３次トライアル

組み上げた複数フォームを新たな協力者集団に対して試行して、さらに解答データを蓄積した。

E 最終フォームの決定

計３回のトライアルの解答データの全てを用いて項目難易度を改めて算出し直し、その数値を元に等化した複数のフォームを最終的に決定した。

IV.ベルクテストがフィードバックする結果

本テストは大きく分けて３種類の情報をフィードバックする。

A VELC スコア

本テストのスコアとしては、総合スコア、リスニングスコア、リーディングスコア、リスニング語彙力スコア、リスニング音声解析力スコア、リスニング内容把握力スコア、リーディング語彙力スコア、リーディング文法構文力スコア、リーディング内容把握力スコア、の６種類が算出される。結果の解釈を容易にするため、それぞれのスコアについて、トライアルを受けた日本人大学生全員の平均値が 500、標準偏差が 100 になるように変換したものを用いる。

たとえば今後このテストを受けた受験者のリスニングスコアが 550、リーディングスコアが 450 であったならば、リスニング能力に関しては日本人大学生の平均よりも 0.5×SD だけ高いところ（上から 31%）に位置しているが、逆にリーディングに関しては 0.5×SD だけ低いところ（下から 31%）に位置している、と解釈できる（表２）

表２　ベルクスコアとパーセンタイルランク

ベルクスコア	下からのパーセンタイルランク	上からのパーセンタイルランク
250	1%	99%
300	2%	98%
350	7%	93%
400	16%	84%
425	23%	77%
450	31%	69%
475	40%	60%
500	**50%**	**50%**
525	60%	40%
550	69%	31%
575	77%	23%
600	84%	16%
650	93%	7%
700	98%	2%
750	99%	1%

このような VELC スコアの計算に当たっては Winsteps の UPMEAN コマンド と SCORE FILE コマンドを利用した。

B 知識・スキル別の細分型診断

上述のベルクスコアはスキル分野別の大まかな能力プロファイルを、集団基準準拠的に示すものだが、さらに細かい下位知識・下位技能についてのフィードバ

ックを与えるのが、「知識・スキル別の細分型診断」である。このためにまず各フォームに含まれる全 120 項目を表３に示すようなカテゴリーで分類した。このカテゴリーは複数の研究会メンバーが討議をへて設定したものである。また項目の分類作業は複数の評定者が独立して行ったものをすりあわせ討議によって確定した。最終的な分類表に基づき、フォーム毎に各分類に該当する項目の全協力者の平均正答率を算出した。ベルクテストの受験者は、この分類ごとの自分の正答率と、全国平均の正答率を比較することにより、これらの下位知識・下位知識ごとの自分の位置づけを確認することができる。

表３．知識・スキル別の項目分類の基準

分類略称	分類に該当する項目の種類
高校語彙	正解語が JACET8000 レベル１～２である
大学基礎語彙	正解語が JACET8000 レベル３～４である
大学応用語彙	正解語が JACET8000 レベル５～７である
内容語聞き取り	正解語が内容語である聞き取り項目
機能語弱形	正解語または直前語が機能語の弱形である聞き取り項目
非開放	正解語または直前語に非開放閉鎖音がある聞き取り項目
あいまい	正解語およびその前後の語にあいまい母音シュワを含む聞き取り項目
リンキング	正解語と直前語が C+V でリンクしている聞き取り項目
音素識別	正解語と錯乱肢に類似の音素がある聞き取り項目
長い主語	５語以上の主部を含む
長い目的語	５語以上の目的語を含む
長い前置詞句	５語以上の前置詞句を含む
長い副詞節	５語以上の副詞節含む
関係詞節	関係詞節を含む
後置修飾／説明	５語以上の後置修飾を含む
遠い照応	文境界を超えた、または５語以上離れた照応関係含む
要素合体	and, or で結ばれた５語以上の名詞句、動詞句を含む
文中挿入	挿入的表現を含む
文末付加	文末に付加される表現を含む
文間関係	文と文の関係を理解する必要がある

C 状況別 Can Do レベル

ベルクスコアと知識・スキル別の細分型診断はいずれも集団基準準拠的な情報だが、実際に英語で何がどの程度できそうか、という目標基準準拠的な情報を提供するのが「状況別 Can Do レベル」である。

まず、トライアル２の受験者のうちの約 550 名から表４に示した 10 のリスニング状況、10 のリーディング状況において、自分がどの程度理解できると思うかを、0～10%なら 0、90～100%なら 4 の、0/1/2/3/4 の 5 件法で回答を得た。この回答データに対して、リスニング項目およびリーディング項目に対する解答データと合わせてラッシュモデリングを行うと、それぞれの状況で「成功する難度」(D)がロジット値で算出される。そのロジット値と、各受験者の能力ロジット値（B）を Pr = exp (B-D)/(1+exp(B-D))に代入して得られるのが、その受験者のその状況における成功確率である。表３には、今回のトライアルの協力者の中で平均的な能力（つまり、B=0.0）を持った者の各状況での成功確率を 5%刻みに丸めた数値が示してある。

表３．平均的大学生の状況別 Can Do レベル

中学レベルの教材を読み上げた録音を聞いて...	80%
日本人の先生が授業中に簡単な英語で指示するのを聞いて...	75%
高校レベルの教材を読み上げた録音を聞いて...	50%
ネイティブスピーカーの先生の英語の授業を受けて.....	40%
テンポのゆっくりした英語の歌(バラードなど)を聞いて....	30%
海外旅行中に空港や駅などで出発便やプラット英語のアナウンスを聞いて...	25%
海外旅行中に食事・買い物のために店に入り、あなたの言ったことに店員さんが答えるのを聞いて....	25%
海外で作成されたニュース番組を見て....	10%
英語の映画を字幕なしで見て....	10%
ネイティブスピーカー同士が自然に話しているのを聞いて....	5%
中学レベルの教材として書かれた英文を読ん	90%

で...	
高校レベルの教材として書かれた英文を読んで....	65%
海外の公共交通機関の駅での英語の掲示（切符の買い方・値段など）を見て	50%
自分あてに書かれた簡単な英語の e-メールの文を読んで	50%
学習者用に編纂された英英辞典(Longman など)の定義文を読んで....	35%
海外レストランの英語メニューで料理名と説明を読んで	35%
学習者用に書きなおされた短い小説を読んで...	25%
日本で出版された英語の新聞で国内ニュースを読んで....	25%
海外で出版された英語の新聞でニュースを読んで.	10%
海外で出版された英語のベストセラー小説を読んで....	10%

V. 信頼性と妥当性

(1)信頼性

テスト得点が信頼性を持つためにはまず当該の受験者層に対して項目難易度が適切であることが必要である。項目が易しすぎても難しすぎても結果が生成する情報量は小さくなる。図 1 は最終的にフォーム A（仮称）に採用された項目を受験した 5583 人の協力者（左; #が 15 人を示す）と 120 の項目（右; X が 1 項目を示す）の能力／難度の分布を示す。項目が -3 logits から 1.5 logits 程度まで 4.5 logits という広い幅にほぼまんべんなく広がり、下位層から上位層まで幅広い受験者層に対して対応可能なテストであり、かつ全体として日本人大学生を対象として適切な難易度となっていることが視覚的に確認できる。

次に、フォーム A に含まれる 120 項目の中で 115 項目以上に解答した 226 名のデータに関して信頼性係数を算出すると、素点ベースのクロンバック α に相当する Rasch person reliability は、.95、 さらに Rasch item reliability は .95 であった。この受験者集団に対して非常に高いレベルの信頼性を示していることが確認できる。

(2)基準関連妥当性

次に基準関連妥当性のデータとしてベルクスコアを予測変数、TOEIC を目標変数とした場合の重回帰分析の結果を示す。トライアルの協力者のなかで TOEIC のリスニング、リーディング、総合のスコアをすべて回答した N=375 を分析対象とした。

目標変数は、TOEIC のリスニングおよびリーディング、予測変数をベルクの L1(リスニング語彙力)、L2（音声解析力）、L3（内容把握力）、R1(リーディング語彙力)、R2（文法構文力）、R3（内容把握力）としてステップワイズ重回帰分析を行って最終的に選択されたモデルは以下の通りである。

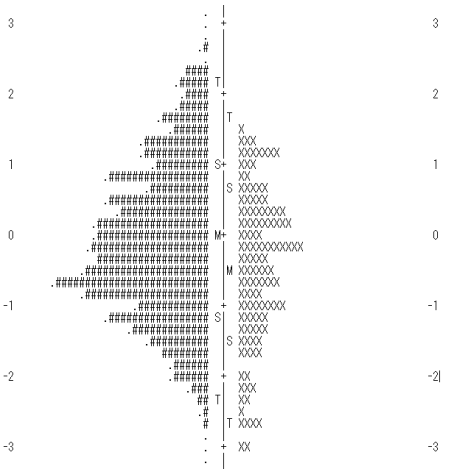

図 1　フォーム A の受験者能力と項目難度の分布

TOEIC L = -74.886 + 0.075*L1 + 0.199*L2
+ 0.248*L3 + 0.119*R3

TOEIC R = -199.599 + 0.075*L1 + 0.079*L2
+ 0.148*L3 + 0.109*R1 + 0.174*R2 + 0.211*R3

これらのモデルの決定係数はそれぞれ 58%と 64%である。また、これらのモデルによる予測値の合計によって TOEIC トータルを予測してみると、決定係数は 68%であり、これは重相関係数として 0.82 にあたる。図 2 に予測値と実測値のプロットを示す。70 分間で実施できる 120 項目のテストで、2 時間かかる TOEIC の結果と 0.82 という高い相関が見られたのは興味深い。決定係数が 68%というのは十分実用になる予測精度と言えよう。

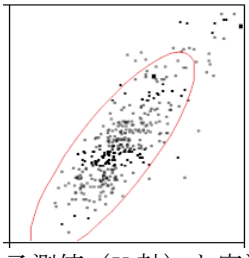

図 2　TOEIC 予測値（X 軸）と実測値（Y 軸）

(3)構成概念妥当性

次に、ベルクテストの 6 パート（L1, L2, L3, R1, R2, R3）と、TOEIC テストの 2 セクション(L と R)の 8 つの

変数を対象にして、因子分析を試みた。SPSS による探索的因子分析の結果抽出した 3 因子の負荷量をもとに Amos によってモデルを構築したのが、図 3 である。理論的な想定通り、ベルクの L2, L3 および TOEIC のリスニングが一つの因子（リスニング）、ベルクの L1 と R1 がひとつの因子（語彙）、ベルクの R2, R3 と TOEIC のリーディングがひとつの因子（リーディング）に関わっているとするモデルが適合度が良かった（GFI=.957, AGFI = .908）。

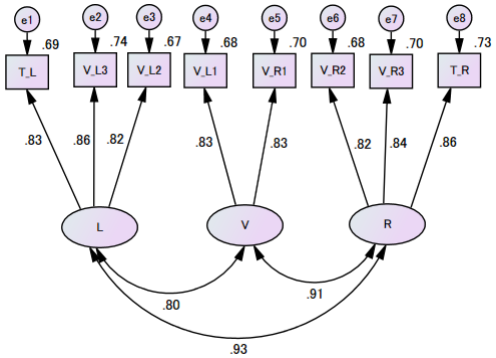

図 3　ベルクテスト 6 パートスコアと TOEIC セクションスコアの因子構造

(4)2012 年度受験データ

次に 2012 年度にベルクテストの Form A を受験した 19 大学の 2327 名の結果データを検討する。

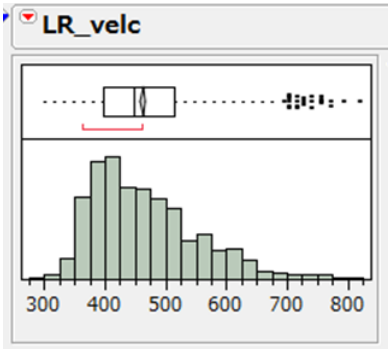

図 3　フォーム A の受験者（N=2327）の得点分布

図 3 に全受験者のスコアのヒストグラムを示す。最低 293 から最高の 820 の広い範囲に分布し、平均値は 462 である。今回のサンプルはやや右に裾野が長い分布になっている。このヒストグラムにより、ベルクテストのスコアは広い範囲に分布することが確認できた。

では大学によってスコアは異なるだろうか。図 4 は、大学別の得点を箱ひげ図にして示したものである。

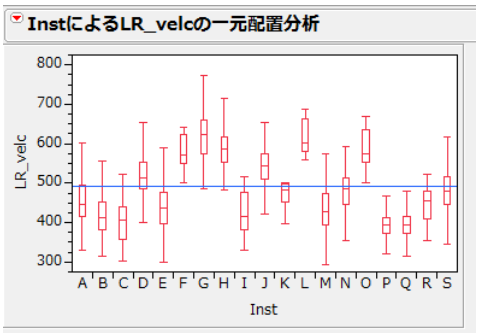

図 4　フォーム A を受験した 19 大学の得点分布

一見して、大学ごとに得点分布が大きく異なっているのがわかる。平均スコアが最も高かったのは G 大学（M=623）、低かったのは Q 大学(M=396)で、分散分析をしてみると全体の差は有意であった(F (18, 2308) = 204.02, p<.000)。これは大学によって異なる英語力がベルクテストによって明らかになっていると解釈してよいだろう。

VI. まとめ

VELC Test は日本人大学生の実態に即して開発された熟達度テストである。難易度的に日本人学生の平均レベルに合わせていることもあり、信頼性係数は非常に高い。また TOEIC スコアを目標基準にした場合の基準関連妥当性も高く、TOEIC の分散の 64%を予測する。また英語力が異なった学生集団が受験するとその英語力を忠実に反映したスコア分布になると考えられる。

70 分で実施できるため年間に複数回実施しても授業への影響は少ない。結果データはウェブで確認できるため、継続的に受験することで英語力の伸びを細かく確認することができる。熟達度に応じたクラス分けや授業効果の測定など幅広い用途に適していると言えよう。

References

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model.* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Linacre, J. M. (2005). *Winsteps* (Version 3.55) [Computer software]. http://www.winsteps.com/

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』第12号. pp.27-53.

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

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

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