VELC Test® フォーム A の選択肢特性分析

靜　哲人

語学教育研究所 創設 30 周年記念フォーラム

pp. 97-115

大東文化大学語学教育研究所

2015 年 3 月 31 日

VELC Test® フォーム A の選択肢特性分析 [1]

靜　哲人

Analyzing Option Performances of VELC Test® Form A

Tetsuhito Shizuka

Abstract

The VELC Test® is an English proficiency test specifically developed for Japanese EFL university students. The test consists of 60 listening questions and 60 reading questions, with four multiple-choice answers for each question. The score is scaled to an assumed population mean of 500 and a standard deviation of 100. The test has multiple forms that have been equated through Rasch modeling such that the scores obtained on any one form are comparable along one common dimension. A total of 12,857 university students sat for one of the multiple forms of the VELC Test® in the academic year 2013, 1,940 of whom took Form A. Of those who took Form A, 1,797 answered all 120 test items. The purpose of this study was to investigate the performance of each answer-option of every item in Form A, based on the responses of those 1,797 examinees. First, as a quick and crude check of the validity of each option, the option-total correlation, i.e., the point-biserial correlation between the 1/0 matrix representing choice/non-choice of each option and the overall score matrix, was computed. Second, the choice ratio of each option as a function of total-score band was plotted and the trace lines were examined. The results indicated that all 120 correct options were performing effectively and 359 out of the 360 distractors were behaving satisfactorily. In addition, it was found that the test consists of items that differ in the ability bands where they best discriminate. Hence, overall, the study has provided additional evidence in support of the VELC Test's validity.

キーワード：熟達度テスト、多肢選択項目、項目分析、トレースライン

1　VELC Test とは

1.1 テストの概要

VELC Test®とは、筆者の所属する英語能力測定・評価研究会（VELC 研究会）で 2010 年から 2011 年にかけて開発した、リスニングセクションとリーディングセクションからなる日本人大学生向けの英語熟達度テストである。[2]　2012 年度に希望大学を対象に試行実施した後、2013 年度より公式実施を開始しているもので、現在主としてプレイスメントや授業効果の測定などの用途で利用されている（長 2013；眞砂 2014）。

項目の元となる英文素材は VELC 研究会所属の母語話者が書きおろし、それを筆者らが問題項目化した。その予備問題項目を合計 5,000 名を超える日本人大学生集団に試行し、その結果にラッシュモデル（Rasch 1960; Bond & Fox 2007; 靜 2007）を当てはめ、モデル適合度が基準を満たした項目のみを選別した。３ラウンドに渡る試行を経て、難易度が等化された複数フォームが完成しており、どのフォームを受験しても結果の直接比較が可能である。

他の熟達度テストと比べた時の特長のひとつに結果通知の迅速性が挙げられ、事務局に解答データが到着した翌日には「e-ポートフォリオ」上で結果が閲覧可能となる。e-ポートフォリオとは個人別の結果とその分析をウェブ上に表示するもので、複数回受験すればその経時変化もグラフで視覚的に確認できるものである。

1.2 問題形式

テストはリスニング 60 項目とリーディング 60 項目の計 120 項目で構成されており、リスニング、リーディングそれぞれがさらに３パートに別れるので、合計６タイプの問題からなる。その詳細を表１に示す。

L1（リスニングセクションパート１、以下同様）は、日本語の訳語に続いて

音声提示される４つの英単語の中から、日本語に対応するものを選ぶ形式で、聴いて語彙が理解できるかを測定する。L2 は、短い文を聴き、アステリスクで指定された部分の語を答えるもので、連続した音声を正しく単語単位で聴き取れるかを測定する。アステリスクの単語は空欄が始まって４つ目の語が指定される。なお順番の数え間違いによる誤答を防ぐため、誤答選択肢には刺激文音声には使用されていない語のみを使用している。L3 は音声提示される刺激文の最後の語が電子音に置換されており、その電子音の箇所に入るはずの語句を印刷提示された選択肢から選ぶ形式である。文脈を追って正しく聴解できる力を測定する。

表１　VELC Test のパート別問題形式

パート	問題例（正体は印刷されて、斜体は音声で提示）	項目数
L1	*歩く　a) write, b) happen, c) walk, d) love*	20
L2	If fact, ( ) ( ) ( ) (*) ( ) ( ) ( ) ( ) . a) all, b) though, c) must, d) almost *In fact, my sister is almost as tall as me.*	20
L3	*In Japan today, more than 50% of high school students go on to [BEEP].* a) school, b) university, c) books, d) jobs	20
R1	経験　a) society, b) experience, c) notice, d) language	20
R2	Today, people a) can use the Internet b) find it easy to c) communicate with d) each other. [who]は a)～d)のうちどこに入るか。	20
R3	In Japan, high school students often _____, but university students usually do not. a) eat lunch, b) wear uniforms, c) work part-time, d) study English	20

R1（リーディングセクションパート１、以下同様）は L1 と同様の形式を印刷

ベースで提示するもので、見て語彙が理解できるかを測定する。R2 は、文から抜き出された 1 語がどこから抜き出されたかを問うもので、構文理解力および読解力を測定する。筆者が以前から Invisible-gap filling と呼んで到達度テストで使用していた形式（Shizuka, 2004）を多肢選択の熟達度テストに応用したものである。R3 はある程度の長さの文章の空所に入るフレーズ、文を問うもので、行間を読む力を含む読解力を測定するものである。

他の商用熟達度テストでよく見られる、いわゆる統合的項目(integrative items)を意図的に避け、各項目で測定する能力を比較的細かく設定することにより、診断的情報を得ることを狙っている。

1.3 これまでの妥当性検証

VELC Test®の信頼性、妥当性については研究会のメンバーがすでにいくつかの観点から検証を試みた（静 2012b; 2012c; 2013; 望月 2013; 静・望月 2014; 水本・熊澤 2014)。静・望月(2014)では、まず同テストを日本人大学生が受験した時に.94 を超える信頼性指数が得られ、かつ受験者層の最も厚い能力レベルにおいて最も正確な測定がなされることを確認した。また項目難度の幅が広く、難度は安定しているという結果も得た。また試行段階におけるデータでは、TOEIC®の総合スコアとの相関係数が.82 を超えたので、TOEIC®を基準とした場合にも実用上も十分な基準関連妥当性があると思われた。さらに構造方程式モデリングにより本テストの構成概念妥当性に関する複合的な証拠を確認した。

水本・熊澤（2014）は、（１）通年の英語授業で前期には Form A、後期には Form B と、ともに期末テストとして VELC Test®を用いたケース(N = 190)と、（２）１年以上にわたり３回以上受験したケース(N= 353)を分析し、熟達度の変化を VELC Test®がどのように検知したかを調査した。調査（１）の結果としては、受験者能力推定値は前期よりも後期のほうが有意に高くなっていたが、効果量の点ではほとんど実質的な意味はなかった。わずかな能力の伸びを VELC Test®は検知することが確認されたと同時に、どのような授業であっても半年間で得られる能力の伸び幅はわずかであろうという教師の抱く一般的な直観を裏付

ける結果が得られたと言える。調査（２）では、14ヶ月間の個人個人のスコアの変化に対してマルチレベルモデルを当てはめてみたところ初期状態ではVELCスコアが 504 点だったのが、14 ヶ月目には 520 点になったというモデルケースが浮かび上がった。やはり少しずつではあるが授業を通じて熟達度が向上してゆく様子を VELC Test®がとらえたという結果が得られたと言える。

2 本研究の目的

本研究の目的は、VELC Test®の項目の特性を、誤答選択肢（錯乱肢）を含めた選択肢の振る舞いという観点から検証することにより、VELC Test®の妥当性に関するさらなるデータを得ることである。

「正答選択肢は、能力が上がるに従って選択される率が上がっているか」「誤答選択肢のひとつひとつは能力が上がるに従って選択される率が下がっているか」「誤答選択肢のなかに機能していないもの、ほとんど選択されていなものなどはないか」「項目ごとに弁別力が高い能力層が変わっているか」などの疑問に答えることを試みた。

3 手順

3.1 分析対象データ

2013 年度に VELC Test®を受験した大学生および高等専門学校生 12,857 名のなかで Form A を受験したのは 1,940 名であった。その中で 120 問のすべてに解答していた 1,797 名が選んだ選択肢の記号データ（A, B, C, D）を分析の対象とした。データは大学名および受験者個人名を削除した形で VELC Test®運営事務局より提供を受けた。

3.2 分析方法

3.2.1 各選択肢の選択・非選択（1/0）と総合スコアの相関

まずおおまかな状況を把握するために、各選択肢が選択された場合を 1、選択されなかった場合を 0 と置き換えて生成される 1/0 行列と総合スコア行列の相関

係数を確認した。その選択肢が正答である場合は、相関係数は正でかつ有意であることが期待される。総合スコアが高いほど、解答としてその選択肢を選ぶ確率が高くなるはずであるからである。一方選択肢が誤答である場合はその逆で、総合スコアが高いほどその選択肢を避ける傾向が強くなるはずであるので、負で有意な相関があることが期待される。ただし誤答選択肢は３つあるため、誤答である場合の選択率を分け合う傾向がある（いわば「票が割れる」）。そのため選択肢によっては全体の選択率が低くなり、相関が 0.00 に近く有意でなくなる場合も考えられるが、その場合も少なくとも正の値であるべきである。

このような相関係数は、選択された記号データ(A, B, C, D)を Excel®等で 1/0 データに変換して手作業で求めることも可能だが、今回はラッシュモデリングのソフトウェアである Winsteps® の Output files メニューの中の Category/Option/Distractor Files DISFILES= コマンドを利用して一気にアウトプットされる Distractor File を利用した。

Distractor File アウトプットの一部を表２に示す。これは L1_005（リスニングセクションパート１の項目の中の通番 05）の選択肢の振る舞いを示すデータである。CODE コラムが ABCD の選択肢記号である。VALUE コラムの値は、その行の記号が正答(1)であるか誤答(0)であるかを示す。この項目は B の行が１なので B が正答である。USED が各選択肢を選んだ受験者の数を、USED%がその率を示す。90%以上の受験者が正解の B を選んだことが分かる。AVGE MEAS は average measure の略で、各選択肢を選んだ学生の、平均能力値（単位はロジッツ）を示す。B の値が最も高く（この場合はたまたま正の値）、A、C、D の値はより低い。この項目の正解である B を選んだ受験者の平均能力が、誤答である A, C, D をのいずれを選んだ受験者の平均能力よりも高いことが分かる。その右の S.E.MEAS は能力値の標準誤差である。OUTFIT MNSQ コラムの数値はラッシュモデルへの適合度 Outfit Mean Square を示す。概ね値が 0.50 から 1.50 の範囲ならば測定のために有用である。0.50 より小さい、あるいは 1.50 から 2.00 の間なら有用ではないが害はない。2.00 を超えると有害である(Linacre, 2014: 596)。PTMEA は point-measure correlation の略であり、今回の分析の焦点であ

る相関係数を表している。この項目の場合、正解である B のみが.27 と正の値であり、ほかの 3 つは負の値である。この PTMEA の値を中心にすべての項目を点検した。

表 2　Winsteps の Distractor File の出力の一部

CODE	VALUE	USED	USED %	AVGE MEAS	S.E. MEAS	OUTFIT MNSQ	PTMEA	LABEL
A	0	81	4.5	-1.01	0.09	0.94	-0.19	L1_005
B	1	1620	90.2	0.09	0.03	0.97	0.27	L1_005
C	0	66	3.7	-0.78	0.11	1.14	-0.13	L1_005
D	0	30	1.7	-0.93	0.15	0.95	-0.11	L1_005

3.2.2　トレースライン分析

上の相関分析は分析対象のすべての被験者のデータを元にひとつの相関係数を算出するものだが、当該の選択肢が被験者のレベルによってどのように振舞っているのか、すなわち弁別しているのかの情報は得られない。そのような被験者レベル別の弁別の様子に光を当てるのがトレースライン分析（Haladyna, 1999: 177）である。トレースライン分析とは、各選択肢の選択率が受験者能力レベルの変化に応じてどのように変化するかを、横軸に受験者の能力レベルを、縦軸に選択率をとってプロットして視覚的に確かめることを言う。

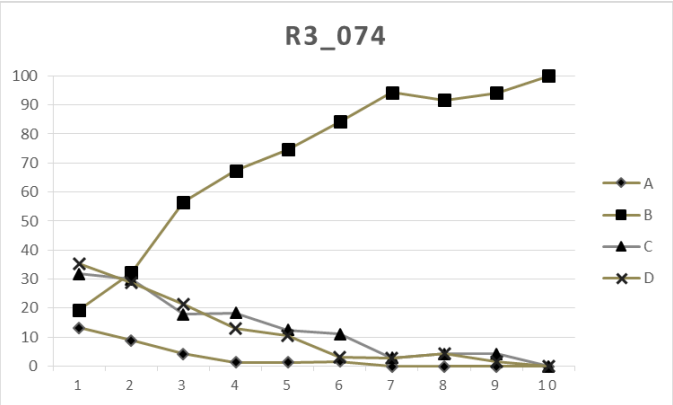


図 1　トレースラインの例。横軸は受験者の総合レベル（1 が最も低く 10 が最も高い）、縦軸はそのレベルの受験者がその選択肢を選んだ率。

サンプルとして図１を示す。４つの折れ線が４つの選択肢の選択率の変化を表している。右肩上がりの折れ線がひとつあり、これが正答選択肢のトレースラインである。受験者の全体的能力レベルが上がるに従ってこの選択肢を選ぶ率が上がっている。すなわちこの選択肢を選ぶかどうかがその受験者の全体的能力を示唆するということであり、正答選択肢として妥当なパターンである。残りの３つの折れ線はいずれも右肩下がりである。これらは誤答選択肢すなわち錯乱肢であり、受験者の全体的能力レベルが下がるに従って、いずれも選択される率が上がっている。すなわち能力の低い受験者ほどこれらの錯乱肢に「錯乱」されるわけで､これらは誤答選択肢として妥当な働きをしていると言える｡ただ選択肢Ａは、ＣとＤに比べると選択率が非常に低く、ほとんど「魅力」がないので、より「魅力的」になるように改善の余地がある。例えばこのような解釈を行うのがトレースライン分析である。

今回の分析で横軸の受験者のグループ分けには､VELC スコアレポートとしてもフィードバックされる 10 のレベル分けを利用した。この 10 レベルは、想定母集団のスコアが、平均が 500、標準偏差 100 の完全な正規分布をすると仮定したときに、人数が 10 等分されるように VELC スコアを切り分けたものである。

Form A を構成する全 120 項目について Excel®によりトレースラインを作成し、視認により選択肢の振る舞いを観察し､Distractor File の数値とも照合しながら、項目の特性を吟味した。

4 結果

VELC Test®の項目は非公開であり、本論文でも項目内容を具体的に示して論ずることはできない。よって以下はその制約の中での記述となる。

4.1 選択肢毎の相関係数分析

表３に、各項目の正答選択肢に関する選択状況と総合スコアの点双列相関係数の記述統計をパート別に示す。リスニングパートの平均値は.32～.38、リーディングパートの平均値は.44～.50 である。いずれもかなり高いがリスニング項目よ

りもリーディング項目の値のほうがコンスタントに高い傾向がある。最小値を比較するとリスニングパートは.14～.18 と、.20 を下回るが、リーディングパートは.25～.31 と、.25 以上である。いずれにしても、すべて正でかつ有意な相関であり、正答選択肢として想定通りの効果的な振る舞いをしていると言える。

表３　正答選択肢と総合スコアの相関係数のパート別記述統計

	平均	標準偏差	最高	最低		平均	標準偏差	最高	最低
L1	.38	0.14	.66	.14	R1	.44	0.08	.56	.29
L2	.32	0.10	.48	.18	R2	.45	0.09	.57	.25
L3	.37	0.09	.56	.18	R3	.50	0.09	.63	.31

注：選択肢数は各パートとも、$k = 20$（20 項目にそれぞれ１つの正答選択肢）

表４に誤答選択肢に関する同様の記述統計を示す。上述のように、誤答選択肢に関してはそれを選択することと総合スコアとの間には負の相関があることが正常な状態であり、正の相関があってはならない。平均値をリスニングパートとリーディングパートで比較すると、正答選択肢の場合と同様、リスニング(-.17～-.13)よりもリーディング（-.22～-23）においてやや関係が強いことが分かる。最高値を見ると、L1, L2, L3 のいずれにおいても小さいながらも正の値であり、何らかの理由により想定外の振る舞いが起こっていることがわかる。そこでひとつひとつの選択肢を点検してみると、６つのリスニング項目において、それぞれひとつの選択肢で小さいながらも正の値があることが判明した。

表４　誤答選択肢と総合スコアの相関係数のパート別記述統計

	平均	標準偏差	最高	最低		平均	標準偏差	最高	最低
L1	-.17	0.10	.06	-.18	R1	-.22	0.08	-.05	-.38
L2	-.13	0.10	.10	-.14	R2	-.22	0.08	-.03	-.39
L3	-.16.	0.09	.05	-.16	R3	-.23	0.08	-.04	-.23

注：選択肢数は各パートとも、$k = 60$（20 項目にそれぞれ３つの誤答選択肢）

4.2　トレースライン分析：誤答に正の相関があった項目

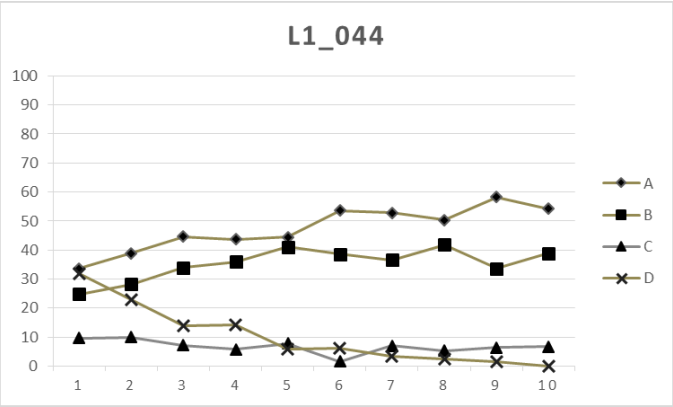
トレースラインは全 120 項目に関して作成したが、そのすべてを提示するのは紙幅の関係上無理である。そこでまず前項で判明した、相関分析で負の値である

べき誤答選択肢に小さいながらも正の値が見出された項目のトレースラインを提示し、状況を解釈することとする。

相関係数が正であった誤答選択肢とは、L1_044 の B (.067), L2_012 の A (.1027), L2_018 の C (.065), L2_026 の D (.0059), L3_054 の D (.0196), L3_095 の B (.0525)の 6 つである。(それぞれ L の後の数字がパートを、アンダーバーの後の 3 桁の数字は識別番号を、その後の記号が当該の選択肢を、括弧内の数値が相関係数を示す。)

＜L1_044 の選択肢 B＞

L1_044 は、リスニングのセクション 1 なので選択肢はすべて語彙である。A、B、C、D といずれも JACET 8000 ではレベル 5 の名詞であった。正解である A のトレースラインは右肩上がり、つまり能力が上がるに従って選択率も上がっているので想定通りである。しかし誤答である B もまたわずかではあるが右肩上がりで能力が上がるに従って選択率も漸増している。ラッシュモデルとの適合度を示す Outfit も 2.0 をわずかに上回り、測定にとってやや「有害」、というレベルである。



CODE	VALUE	USED	USED %	AVGE MEAS	S.E. MEAS	OUTFIT MNSQ	PTMEA	LABEL
A	1	843	46.9	0.16	0.04	1.43	0.14	L1_044
B	0	625	34.8	0.09	0.04	2.05	0.06	L1_044
C	0	129	7.2	-0.17	0.10	1.91	-0.04	L1_044
D	0	200	11.1	-0.91	0.05	0.53	-0.28	L1_044

図 2　L1_044 のトレースラインおよび選択状況データ。

ただし、訳語との対応という点においてBが正解にいくらかでも近いということはまったくない。よって内容的な観点からの出題ミスとは言うことはできない。ではなぜこのようなトレースラインなのだろうか。

まず誤答のCを見ると、どのレベルにおいても選択率が10%以下で、ほとんど「錯乱」力がなかったことがわかる。音節数は、Aが３音節、Bが４音節、Cが１音節、Dが４音節であり、目立ってCだけが短い。Cの除外には、ひとつだけ目立って異なる選択肢はおそらく正解ではない、という判断が働いた可能性がある。もうひとつの誤答Dは、おそらく４つの中で親密度が最も高く、提示された訳語に対応する語彙ではないことがすくなくともレベル５以上の受験者には分かった可能性がある。（なお、４つの語はいずれも横川・他（2009）の「日本人英語学習者の英単語親密度リスト：音声編」には含まれておらず、直観を超えた親密度の査定はできない。）

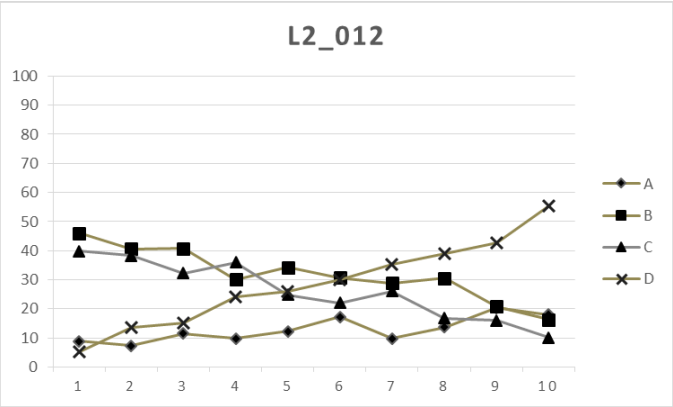
もしそうだとすると、レベル５以上の受験者のほとんどにとっては、４選択肢中の２選択肢は消去され、事実上の２選択問題に近くなったことが考えられる。しかしAとBの間でまったく無作為の２択でなかったのは、正解のAの選択率がBのそれよりもコンスタントに高かったことから分かる。そしてAとBのAVRG MEASを比較すると、Aを選んだ受験者のほうがBを選んだ受験者よりも平均能力が高かった。

AとBのトレースラインがどちらも右肩上がりなのは、能力が高くなるほど「正解はAとBのどちらかである」ことが分かった、ということを示す。つまりAは正解選択肢として妥当な振る舞いをしていたのだが、なぜかBも（内容的には正解ではないにもかかわらず）準正解選択肢のような振る舞いをしてしまったということである。このようなパターンになった原因は、最終的には解釈しきれない部分が残ると言わざるを得ない。

＜L2_012の選択肢A＞

これはリスニングのセクション２の項目なので空欄にあたる語を聞き取って選

ぶものである。図３を見ると、誤答選択肢 A のトレースラインががやや右肩上がりになっている。しかし正答の D のラインとはかなり離れており、選択率はわずか 12.5%である。Outfit 値を見ても 1.99 と 2.00 は超えていないため実質的には問題ないと考えて良いと思われる。



CODE	VALUE	USED	USED %	AVGE MEAS	S.E. MEAS	OUTFIT MNSQ	PTMEA	LABEL
A	0	225	12.5	0.30	0.07	1.99	0.10	L2_012
B	0	585	32.6	-0.30	0.04	1.05	-0.18	L2_012
C	0	489	27.2	-0.41	0.04	0.91	-0.22	L2_012
D	1	498	27.7	0.60	0.05	1.17	0.33	L2_012

図３　L2_012 のトレースラインおよび選択状況データ

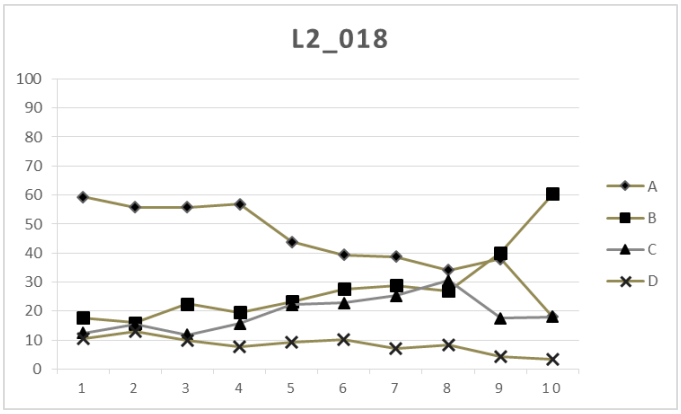
＜L2_018 の選択肢 C＞

L2_018 のトレースラインおよび選択状況データを図４に示す。正答の B と誤答の C は受験生レベル１～８まではほとんどトレースラインが重なっている。しかしレベル９と１０では大きく離れる。つまり誤答選択肢 C は熟達度が高いレベルで弁別するという機能を果たしていると言える。Outfit も 1.81 であり、Linacre (2014)の基準によるならば、積極的に問題があるとまでは言えない。

＜L2_026 の選択肢 D＞

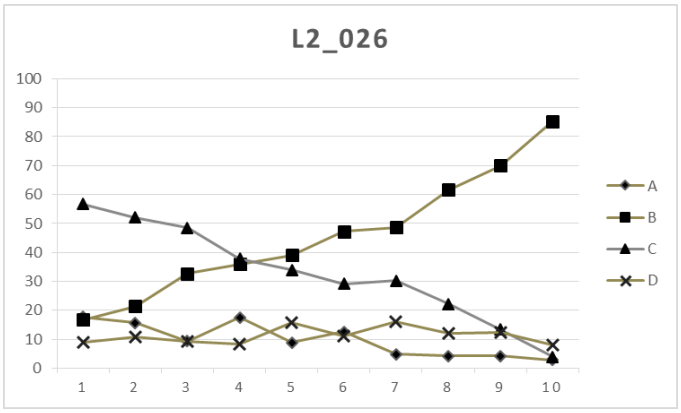
L2_026 のトレースラインおよび選択状況データを図５に示す。相関係数として問題になったのは選択肢 D であるが、トレースラインを一見して問題がないことが分かる。相関係数も正の値とは言え、.01 であり事実上.00 である。Outfit

も 1.62 で問題はない。



CODE	VALUE	USED	USED %	AVGE MEAS	S.E. MEAS	OUTFIT MNSQ	PTMEA	LABEL
A	0	809	45.0	-0.31	0.03	1.02	-0.24	L2_018
B	1	496	27.6	0.49	0.06	1.47	0.27	L2_018
C	0	335	18.6	0.14	0.06	1.81	0.07	L2_018
D	0	157	8.7	-0.37	0.07	0.93	-0.10	L2_018

図 4　L2_018 のトレースラインおよび選択状況データ

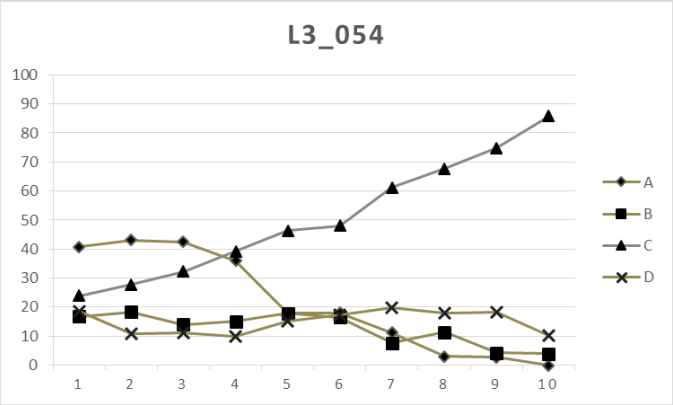


CODE	VALUE	USED	USED %	AVGE MEAS	S.E. MEAS	OUTFIT MNSQ	PTMEA	LABEL
A	0	180	10.0	-0.55	0.07	0.89	-0.16	L2_026
B	1	797	44.4	0.52	0.04	1.10	0.42	L2_026
C	0	619	34.4	-0.54	0.03	0.83	-0.34	L2_026
D	0	201	11.2	0.01	0.07	1.62	0.01	L2_026

図 5　L2_026 のトレースラインおよび選択状況データ

＜L3_054 の選択肢 D＞

リスニングのパート３は、トークの最後の語がビープで置換してあり、そこに当てはまるものを選ぶ、という形式である。L3_054 のトレースラインおよび選択状況データを図６に示す。正解の C だけが明らかな右肩上がりラインである。総関係数が.02 だった D は明らかに事実上平坦なラインで、正解の勾配とはかけ離れている。全体の選択率も 14.2%と低いことからも、積極的に問題があるというよりも、単に魅力のない、弁別力のない選択肢だったことがわかる。Outfit の値の 1.73 もこの解釈を裏付けるものである。



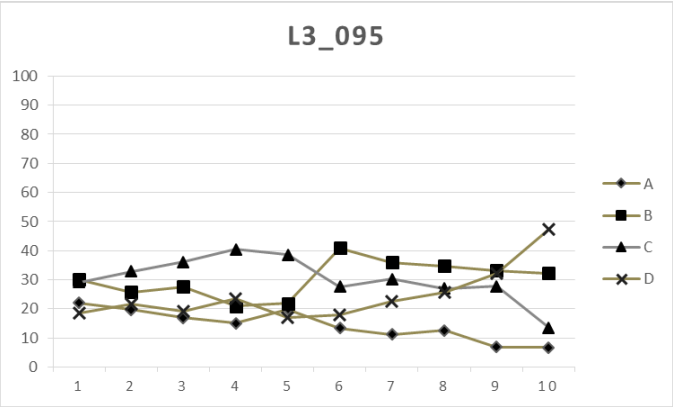
CODE	VALUE	USED	USED %	AVGE MEAS	S.E. MEAS	OUTFIT MNSQ	PTMEA	LABEL
A	0	427	23.8	-0.78	0.03	0.59	-0.38	L3_054
B	0	231	12.9	-0.43	0.06	1.00	-0.15	L3_054
C	1	884	49.2	0.46	0.04	1.11	0.41	L3_054
D	0	255	14.2	0.05	0.06	1.73	0.02	L3_054

図６　L3_054 のトレースラインおよび選択状況データ

＜L3_095 の選択肢 B＞

この項目のトークと選択肢を改めて確認してみても、相関係数がごく小さいながらも正の値（.05）を示した選択肢 B が正解である要素はまったくない。したがって内容的な出題ミスではない。しかし選択肢が入るべき直前の語とのコロケーションだけを見ると、他の誤答選択肢である A と C よりもいくぶんもっともらしいかもしれない。図７を確認すると、レベル６~８の受験者は正解の D よりもこの B を多く選んでいる。しかし注目すべきは、正解 D のレベル９から１０にか

けての勾配が急激に大きくなっている結果、レベル１０では正解の D と誤答の B の選択率が 20 ポイント近く離れているという点である。すなわち、相関係数だけを見ると問題があるように見える選択肢 B は、受験者のトップレベルにおける弁別に寄与していることが分かる。



CODE	VALUE	USED	USED %	AVGE MEAS	S.E. MEAS	OUTFIT MNSQ	PTMEA	LABEL
A	0	265	14.7	-0.36	0.06	0.97	-0.13	L3_095
B	0	533	29.7	0.08	0.05	1.58	0.05	L3_095
C	0	553	30.8	-0.21	0.04	1.03	-0.12	L3_095
D	1	446	24.8	0.34	0.06	1.74	0.18	L3_095

図７　L3_095 のトレースラインおよび選択状況データ

4.3 トレースライン分析：弁別するレベルの違いによる分類

120 項目の正答選択肢のトレースラインは、同じ右肩上がりでも、その勾配の様子から、大きく分けて３つのパターンに分類できた。（１）勾配が下位、中位レベルでは緩やかであるが上位のレベル８～１０で特に急になる項目、（２）勾配がすべてのレベル概ね等しい項目、（３）勾配がレベル１～３などで急だが他のレベルでは緩やかである項目、である。トレースラインの傾斜はすなわち弁別力の高さを表すので、（１）~（３）はそれぞれ、（１）主として上位群を弁別する項目、（２）下位から上位までをほぼまんべんなく弁別する項目、（３）主として下位群を弁別する項目、であると言える。図８、 図９、図１０に３タイプの代表的なトレースラインを示す。

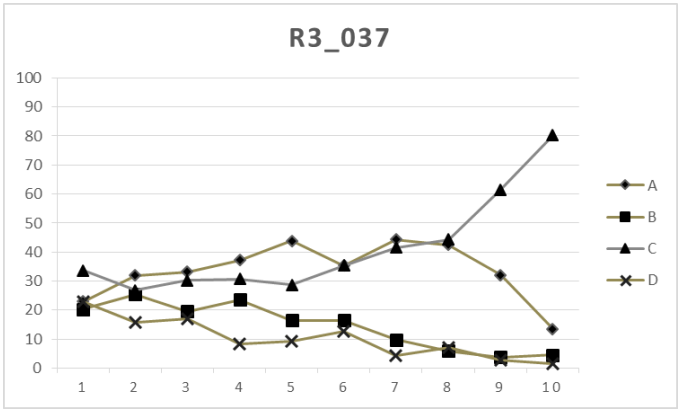


図８　主として上位群を弁別している項目のトレースライン

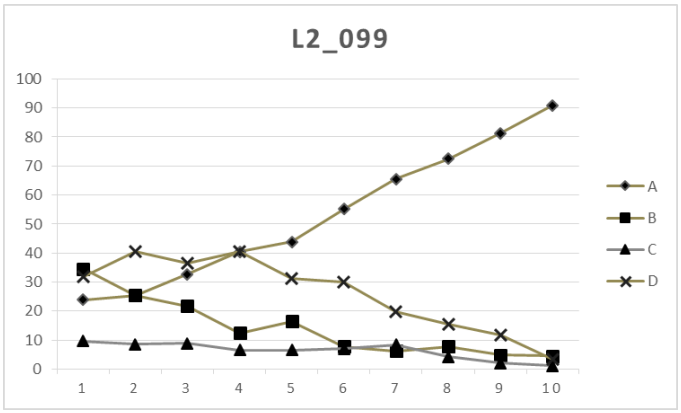


図９　下位群から上位群までまんべんなく弁別している項目のトレースライン

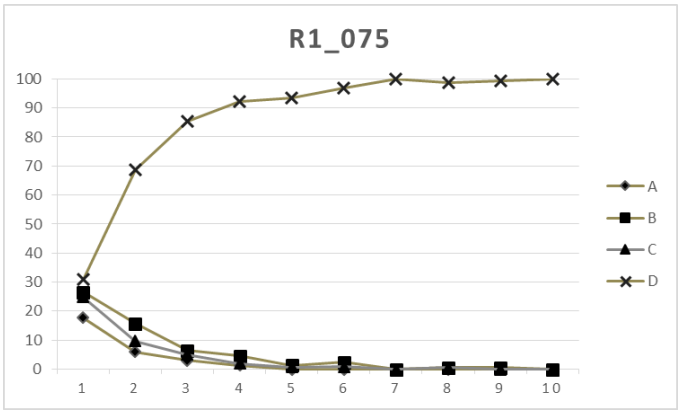


図１０　主として下位群を弁別しているトレースラインの例

それぞれのタイプの割合を６つのパート別にまとめてみると表５のようになる。

タイプの分類は視覚的なおおまかな印象によるので厳密なものではない。しかしある程度の傾向を同定するには十分だと考える。

表５　主として上位を弁別、下位から上位までをまんべんなく弁別、主として下位を弁別する項目の、パート別の割合（パーセンテージ）

	L1	L2	L3	R1	R2	R3	合計
主に上位	20	25	30	15	15	35	23
まんべんなく	55	60	65	40	60	50	55
主に下位	25	15	5	45	25	15	22

すべてのパートを合計すると、主として上位を弁別する項目と主として下位を弁別する項目の割合はほぼ等しく、いずれも 22~23%である。半数以上の項目がすべてのレベルの受験者をまんべんなく弁別するタイプに分類される。パート別に見ると、視認による語彙問題である R1 に下位を弁別する項目の割合が突出して高い（45%)。シンプルな語彙問題が下位層の弁別に役立っているということである。逆に主として上位を弁別する項目の割合が高いのは L3 と R3 である（30% と 35%)。文脈を理解しないと正解できないタイプの問題形式が、上位層の弁別には適していたと解釈できる。

5 考察および結論

VELC Test®は、あらかじめ 5,000 名を超える大学生に試行したデータをもとに、特性（項目難度およびラッシュモデル適合度）が基準に適合した項目のみによって構成されている。よって使用されている項目はすべて優れていると「想定」されている。この想定が実際に事実であるか否かを 2013 年度 Form A の本実施データによって検証すること、が本論文の目的であった。

各選択肢の選択と総合スコアの間の点双列相関係数の結果から見ると、想定通り、調べた 120 項目の正答選択肢はすべて妥当であると言える。つまりどの正答選択肢も、総合的な能力が上がるほど選択される率が上がる、弁別力のある選択肢だと再確認された。

では誤答選択肢はどうであろうか。正答選択肢を選択しない場合は、(無解答の場合を除き) 3 つの誤答選択肢のいずれかを選択する。すると、総合的な能力が上がるほど正答選択肢を選択する率が上がるということは、裏を返せば総合的な能力が下がるほど誤答選択肢のいずれかを選択する率が上がっていることになる。この意味で、各項目の誤答選択肢は 3 つ合わせてひとつのセットとして考えた時には、すべて弁別力のある選択肢セットだと言ってもよい。

しかしそれでは 120 項目に 360 ある誤答選択肢のひとつひとつが、すべて効果的に機能していたのか、といえば今回のデータからは「そうとまでは言えない」という結果が判明した。360 件のうち 1.7%にあたる 6 件の誤答選択肢について事実上ゼロに近いながらも正の値の相関係数を持つものがあったからである。しかしトレースラインを吟味してみたところ、誤答選択肢として場合によっては修正もしくは差し替えが必要かと思われたのは 1 件のみ (L1_044) で、あとの 5 件はラインの形状からも全体選択率からもモデル適合度からも事実上問題はないことが確認された。つまり 360 件の誤答選択肢のうち想定外の動きをしているのがわずか 1 件、0.3%であることが確認されたわけで、VELC Test® フォーム A の項目の質が揃っていることが改めて検証されたと考えて良いと思われる。

また 120 項目の (正答選択肢の) トレースラインの形状を吟味するなかで、項目による弁別「守備」範囲の違いが明らかになった。すなわち主として上位の受験者の弁別に効果的な項目、下位から上位までまんべんなく弁別する項目、主として下位の受験者を弁別する項目の 3 種類が観察された。過半数はまんべんなく弁別する項目群に属するが、それらを上位を弁別する項目群と下位を弁別する項目群が補っている形と言える。全体の正答率が高い項目であっても下位の受験者の弁別に役立っている場合があり、逆に正答率が低い項目であっても上位の受験者の弁別に欠かせない場合がある。弁別の「守備範囲」が異なるさまざまな項目から構成されることで、VELC Test®は比較的幅広い能力層の受験者の弁別を行っていると考えられよう。

注

1 本稿は 2014 年 8 月 28 日に広島大学にて行われた大学英語教育学会(JACET)第 53 回国際大会での口頭発表「VELC Test フォーム A の選択肢分析から見える各アイテムの特性」に加筆修正を加えたものである。

2 VELC Test® についてのより詳細な情報は、ベルク研究会のウェブサイトにある。http://www.velctest.org/

引用文献

長加奈子(2013)「VELC Test の導入と活用法：北九州市立大学国際環境工学部」英語能力測定･評価研究会 VELC Test 公開記念第 2 回研究会［基調講演］(於：研究社英語センター）7 月 28 日.

靜哲人(2007).『基礎から深く理解するラッシュモデリング：項目応答理論とは似て非なる測定のパラダイム』関西大学出版部.

靜哲人(2012a).「大学生のための新しい英語テストの開発」英語能力測定・評価研究会 VELC Test 公開記念第 1 回研究会［基調講演］（於：研究社英語センター）7 月 29 日.

靜哲人(2012b). 「VELC テストによる TOEIC スコアの予測: リスニングとリーディングについて示唆されるもの」日本言語テスト学会第 16 回全国研究大会（於：専修大学生田キャンパス）10 月 27 日.

靜哲人(2012c)「ベルクテストの妥当性を検証する ：2012 年度データにもとづいて」 JACET 関西支部 2012 年度秋季大会（於：京都産業大学）11 月 24 日.

靜哲人(2013). 「実施データに基づく VELC Test の信頼性・妥当性の検証」英語能力測定・評価研究会 VELC Test 公開記念第 2 回研究会［基調講演］（於：研究社英語センター）7 月 28 日.

靜哲人・吉成雄一郎 (2012). 「大学生の英語力『可視化』の試み：熟達度診断のための VELC Test の開発」 JACET 第 51 回国際大会（於：愛知県立大学）9 月 1 日.

靜哲人・望月正道 (2014). 「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」 *JACET Journal, 58,* 121-141.

眞砂薫(2014) 「VELC Test の導入とその活用法：近畿大学薬学部・医学部」英語能力測定・評価研究会第 3 回研究会［基調講演］（於：日本教育会館） 7 月 27 日.

水本篤・熊澤孝昭（2014).「VELC Test による英語能力変化の測定」英語能力測定・評価研究会第 3 回研究会［基調講演］（於：日本教育会館）7 月 27 日.

望月正道(2013).「VELC 語彙問題の分析」英語能力測定・評価研究会 VELC Test 公開記念第 2 回研究会［基調講演］（於：研究社英語センター）7 月 28 日.

横川博一・他(2009).『日本人英語学習者の英単語親密度 音声編』くろしお出版.

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

Haladyna, T. M. (1999). *Developing and validating multiple-choice test items* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Linacre, J. M. (2014). *A user's guide to Winsteps® Ministep Rash-model computer programs. Program manual 3.81.0*. Available from: http://www.winsteps.com/index.htm

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.

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

謝辞

データを整理して提供してくださったベルク研究会の事務局スタッフに感謝します。また本論文の骨子を大学英語教育学会第 53 回国際大会にて発表した際、結果の解釈に関して獨協大学の安間一雄氏より貴重なコメントをいただきました。ここに記して謝意を表します。