VELC Test® 2012-19 年度実施データの分析および総括

靜　哲人

An Analysis and Summary of the VELC Test® 2012-2019 Data

SHIZUKA Tetsuhito

Abstract

The VELC Test®, a 120-item proficiency test developed specifically for L1 Japanese learners of English at tertiary level, has been administered across the nation for the past eight years. This paper analyzes and summarizes the accumulated data in terms of score distribution, item difficulty invariance, item discrimination, and person reliability/separation. Based on the results, it is argued that the test is reliable enough and well-targeted for the intended population of Japanese university students; its practically invariant and uniformly discriminating items separate an applicant group of any given university into at least three and sometimes four ability levels.

Key words：英語熟達度テスト　VELC Test®　信頼性　項目安定性　受験者分離

1. VELC Test® とは

VELT Test® は英語能力測定・評価研究会が開発し、2012 年度より実施・運用を行ってきている、日本語を母語とする大学生を対象とした英語熟達度テストである（靜・吉成，2012；靜・望月，2014）。リスニングセクション 60 問とリーディングセクション 60 問からなり、各セクションがそれぞれ表 1 に示す 3 つのパートに分かれている。

受験者には 120 項目全体にもとづく VELC スコア、セクションごとのリスニングスコアとリーディングスコア、さらに各パートに基づく L1, L2, L3, R1, R2, R3 スコアが返される。これらのスコアはいずれも素点ではなく、ラッシュモデル（Rasch 1960）で推定された受験者能力値を、理論的には平均 500、標準偏差 100 として分布

表 1　VELC Test® のパート別問題形式と項目数

	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

するように変換された値である。開発段階においてのべ 5,000 名を超える日本人大学生に施行した結果より大学生全体の平均的熟達度を推定し、それが VELC スコア 500 になるよう設定されたものである。

あらかじめ難度が等化された複数のフォーム（問題項目のセット）があり、どのフォームを受験しても同一の尺度で結果の比較ができるため、プレイスメントだけでなくコース前後での熟達度の進捗度合い、経年的な変化の測定などにも利用可能である。

VELC Test® の信頼性、妥当性、項目特性などについてはすでに様々な角度から報告がなされてきた（靜，2012a; 2012b; 2013; 2014; 2015a; 2015b; 2017; Shizuka, 2016; 靜・望月，2014; Kumazawa *et al*. 2016）。しかしそれらはほとんどがその時々の単一年度のデータに基づくものであった。

2. 本研究の目的

本格実施を開始してから 8 年目となる今、これまでのデータを経年的・総括的に分析し、あらためてテストの性能を確認しておくのは利用者、受験者に対する説明責任を果たす上で意義のあることである。

本論文はこれまで蓄積されたデータを、受験者スコアの分布、項目難度の安定性、項目のラッシュモデルへの適合度、信頼性と受験者分離度、の観点から分析することをその目的とする。具体的には以下のリサーチクエスチョンに対する答えを探る。

（1）開発時のデータにもとづく係数により、受験者スコアは平均が 500、標準偏差が

100 になるように設定したが、その想定はどの程度実現されているか。また各年度の受験者スコアは熟達度テストに相応しい散らばり具合を見せているか。

(2) 現在使用されている項目は開発時に難度を推定して同一尺度上に並べ、最終的に固定（anchor）したものである。その固定をいったん外した状態で新たな受験者データで推定し直した時、難度はどの程度安定しているか。

(3) 現在使用されている項目は開発時のデータによりラッシュモデルに適合するもののみを選定したものである。その想定は新たな受験者データでも実現されているか。

(4) テストは全国の受験者をどの程度の信頼性をもってどのくらいの数の能力層に分離することに成功しているか。また個別大学の受験者をどの程度の信頼性でどのくらいの数の能力層に分離することに成功しているか。

3. 分析対象および方法

英語能力測定・評価研究会事務局より提供を受けた 8 年分の解答データを分析対象とした。これらのデータでは受験者個人名は言うまでもなく大学名もコード化されており筆者が知ることはできない。データファイルには各個人が各設問に対して選んだ選択肢の生データおよびそれらにもとづく VELC スコアが含まれている。これらのデータを上記リサーチクエスチョンごとに、以下のように分析した。

(1) 過去 8 年分のデータから、各年度 10,000 名を無作為抽出し、すでに換算されている VELC スコアの分布を確認した。

(2) 同一のフォームを異なる年度に受験した異なる集団 1,000 名ずつを抽出し、あらためて項目難度を推定して比較する、という作業を 2 つのフォームについて行った。

(3) (2) で使用した 4 セットのデータに関して、項目の適合度を代表的な指標である Infit Mean Square によって確認した。

(4) (2) で使用した 4 セットのデータに関して、受験者信頼性／受験者分離の度合いを確認した。また年度を超えて、VELC Test の受験層として代表的な 5 つの大学を選び、各大学内での受験者信頼性／受験者分離の度合いを確認した。

以上の (2)～(4) のデータ分析にはすべてラッシュモデリングのソフトウェアである Winsteps（Linacre, 2005）を利用した。

4. 結果

4.1 VELC スコアの分布

今回の分析のために年度ごとに無作為抽出した 10,000 名の VELC スコア、リスニング（L）スコア、リーディング（R）スコアの平均値と標準偏差を表 2 に示す。

表 2　2012〜2019 年度 VELC スコアの全国平均値および標準偏差

	VELC スコア	*SD*	L スコア	*SD*	R スコア	*SD*
2012	470.0	73.6	473.8	77.6	461.0	85.5
2013	478.6	74.7	481.6	80.1	470.2	86.7
2014	484.8	74.8	485.4	81.0	482.3	84.9
2015	473.6	72.4	476.2	76.1	468.7	82.8
2016	476.8	70.0	479.6	75.8	473.0	79.4
2017	479.1	72.1	482.1	77.2	476.0	81.1
2018	482.5	74.8	485.9	78.8	479.6	85.8
2019	494.3	73.9	495.7	79.1	489.5	86.9
平均	480.0	73.3	482.5	78.2	475.0	84.1

注：各年度とも $N = 10{,}000$（無作為抽出）

VELC スコア、リスニングスコア、リーディングスコアとも、平均値はおおよそ 475〜485 あたりに分布している。開発段階で試行してもらった 5,000 名を超えるサンプルの平均値を「日本人大学生全体の平均値」とし、そのレベルがちょうど 500 になるように VELC スコアを設定したわけであるが、その「全国平均値 = 500」という想定を多少下回っていることがわかる。これは試行段階ではかなり英語力の高い層もふくめて広いレベルからまんべんなく受験してもらったことによるものと考えられる。また 2019 年度のみ他の年度にくらべて数値が高いのは、2019 年度のみ、年度全体ではなく年度当初の 4 月受験のデータしか含まれていなかったためかもしれない。

分布の形をさらに詳しく調べ、視覚的に確認するためすべての年度のヒストグラムも生成して検討した。その結果、分布形状に関して年度による違いはほとんどないことが確認できたため、紙幅の関係でここでは例として 2019 年度のヒストグラムのみ示す（図 1）。

下段の分位点情報から以下のことが読み取れる。(1) VELC スコア（左の VELC-T 2019）は 494 を中央値とし、442〜545 に 50% が、360〜645 に 95% が、327〜702 に 99% が分布している。(2) リスニングスコア（VELC-L 2019）は 493 を中央値とし、442〜543 に 50% が、356〜667 に 95% が、311〜739 に 99% が分布している。(3) リーディングスコア（VELC-R 2019）は 490 を中央値とし、429〜547 に 50% が、335〜

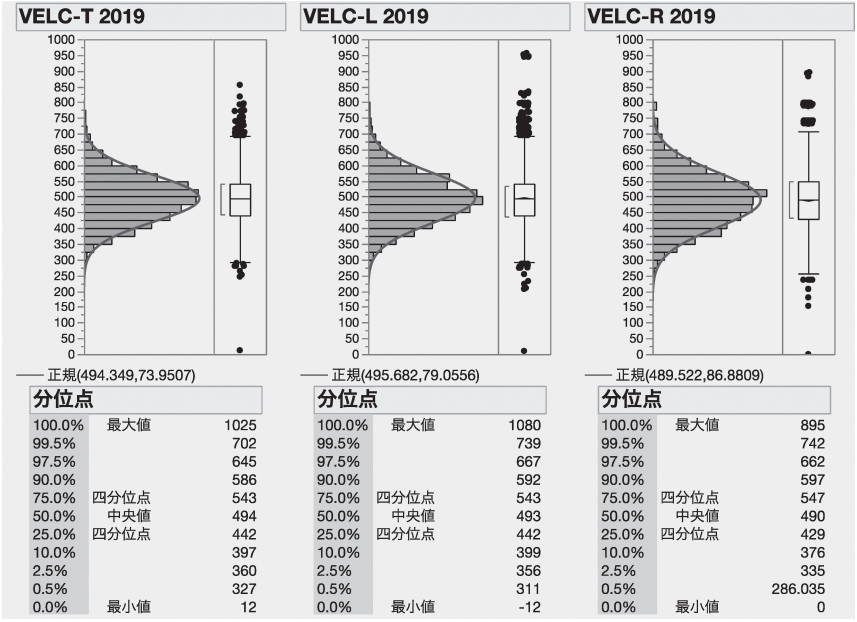

図 1　2019 年度データの VELC スコアの分布状況（*N* = 10,000）

662 に 95% が、286～742 に 99% が分布している。

3 つのヒストグラフには参考までに正規分布曲線を当てはめているが、概ね正規分布に近いと言って差し支えないと思われる。

スコアの分布についてまとめると、現実の受験者のレベルと散らばり具合は理論的な想定の平均 500、標準偏差 100 とはやや異なるものの、日本の平均的な大学生の英語熟達度を測定するという目的に照らして、概ね適切なものであると言えよう。

4.2　項目難度の安定性

受験者の能力を安定的に測定するためには項目の難度が安定していることが不可欠である。項目の難度の安定性とは、受験者サンプルが変わっても項目の難度が測定誤差の範囲を超えて変動することはない、すなわちいわゆる不変性（invariance）があるということにほかならない。VELC Test® ではあらかじめ十分な数の受験者による解答データにもとづいて各項目の難度の値を固定（anchoring）したうえで受験者能力を推定し、スコアに換算している。

しかし、もしそれらの項目の難度値の固定をいったん解き放ち、テストを異なる受験者集団に解答してもらったデータから別々に項目難度を推定した時、どの程度不変性が見られるのかを改めてチェックすることは意義のあることである。そこで複数あ

るフォームの中から 2 つを選び、それぞれを構成する 120 の項目について、年度の異なる受験者集団を用いた不変性の確認をすることとした。

具体的には、あるフォームをある年度に受験した中から無作為抽出した 1,000 名と同じフォームを別の年度に受験した中から無作為抽出した 1,000 名のデータについて別々に Winsteps を走らせて 120 項目の項目難度を独立して推定し、その 2 種類の推定値がどの程度近似しているかを散布図で検証する、という作業を行った。2 つのフォーム（仮にフォーム 1 とフォーム 2 とする）を選び、フォーム 1 については 2018 年度と 2017 年度のデータで、フォーム 2 については 2016 年度と 2015 年度のデータを用いた。項目難度の推定にあたっては UIMean = 0 すなわち当該の分析における項目難度の平均値を 0.0 logits とするという制約を設けた。フォーム 1 の結果を図 2 にフォーム 2 の結果を図 3 に示す。

X 軸にある年度の受験者 N = 1,000、Y 軸に別の年度の受験者 N = 1,000 をとって、当該フォームを構成する 120 項目の難度値をプロットしたものである。どちらのグラフも同じような外観をしており、各項目を示すマーカーが Y = X の直線を中心として集まっている。

不変性という点では Y = X の 1 直線上にすべての点が並ぶのが理想ではあるが、測定誤差の存在もあり、そうはならない。図には項目を示すマーカー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する（Bond & Fox, 2007, p.86）と言われる。2 つの曲線の間の空間の外側に位置する項目数は 120 項目の 5% すなわち 6 項目より明らかに多い。しかし空間の外に位置しているマーカーもその空間の至近にある。

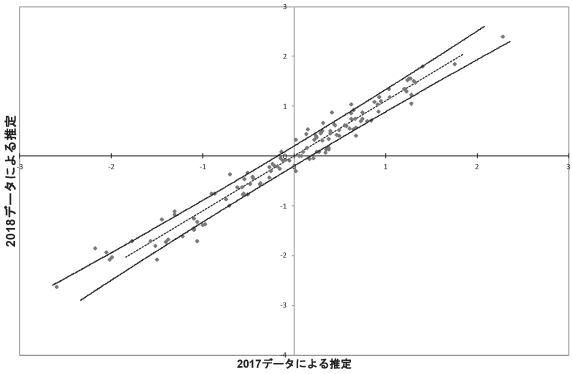

図 2　フォーム 1 の項目難度プロット：
2017 年度データ x 2018 年度データ

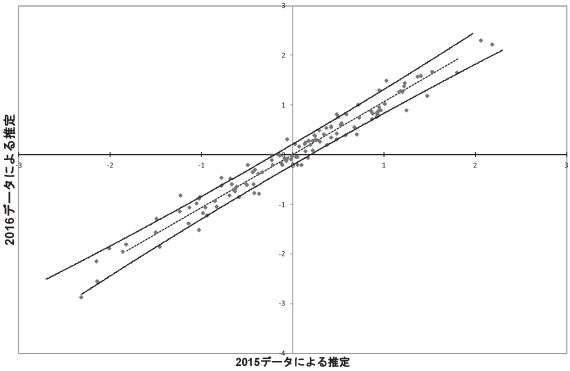

図 3　フォーム 2 の項目難度プロット：
2015 年度データ x 2016 年度データ

異なる集団で推定した項目難度間の相関係数は、フォーム 1 で $r = .980$　フォーム 2 で $r = .981$ であった。この結果は、すべての項目が厳密な不変性を維持しているとまでは言えないまでも実用上は問題ないレベルの準不変性があると解釈できるだろう。

なお再確認しておくと、実際の VELC Test® の結果は項目の難度値を固定（anchor）した状態で受験者能力が計算されているものであり、受験者能力測定の精度はより強い形で担保されていると言える。

4.3　項目のモデル適合度

前項の分析のために Winsteps を 4 回走らせた。(1) フォーム 1 の 2018 年データ、(2) フォーム 1 の 2017 年データ、(3) フォーム 2 の 2016 年データ、(4) フォーム 3 の 2015 年データである。これらの 4 つの分析での項目の Infit Mean Square の分布状況を図 4～図 7 に示す。Infit Mean Square はラッシュモデルの適合度指標の代表的なもので 0.7～1.3 の範囲が適合度の一般的な目安である（Bond & Fox, 2007）。

4 つのデータセットとも全 120 項目が 0.7～1.3 の範囲に収まっており、モデルへの適合は想定どおり問題ないことが確認された。ちなみに Linacre（2005）はより広く範囲をとり、0.5～1.5 の項目が「測定にとって建設的（productive for measurement）である」(p. 197）としている。

4.4　信頼性および受験者分離

テスト得点の信頼性とは当該の受験者集団をどの程度異なるレベルの層に弁別しているか、すなわち受験者分離に成功しているか、ということである。この意味で信頼

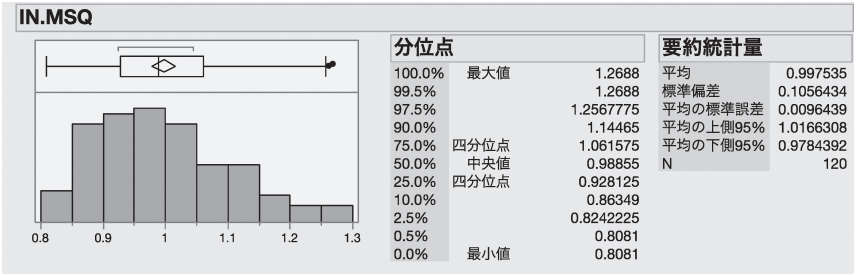

図 4　フォーム 1／2018 年度データでの項目適合度 Infit Mean Square の分布

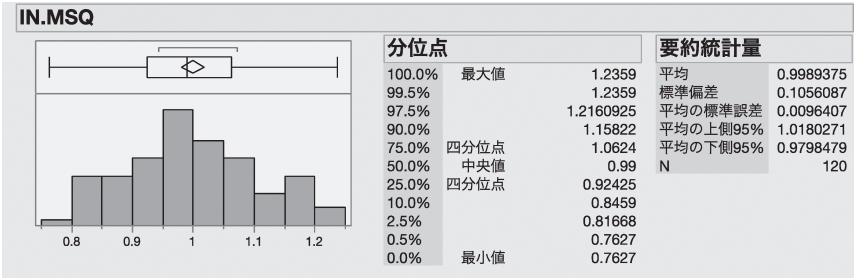

図 5　フォーム 1／2017 年度データでの項目適合度 Infit Mean Square の分布

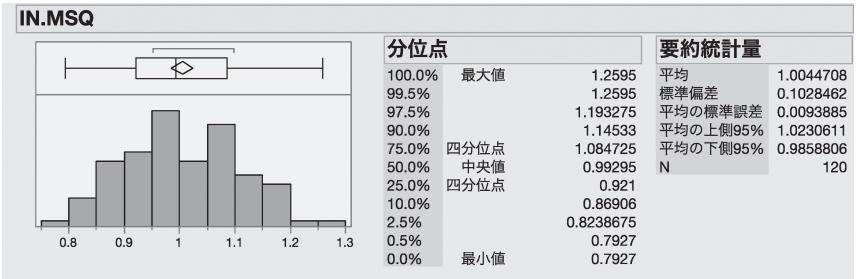

図 6　フォーム 2／2016 年度データでの項目適合度 Infit Mean Square の分布

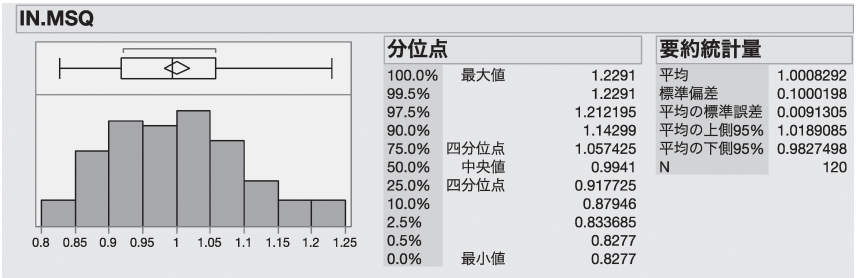

図 7　フォーム 2／2015 年度データでの項目適合度 Infit Mean Square の分布

性とは項目の測定誤差のみの関数ではなく、当該の受験者集団の能力分布の関数でもある。仮に能力が全く等しい受験者のみから構成された受験者集団であれば、いかに精度の高いテストを受けたとしても、信頼性係数は 0.0 となる。標準偏差が 0.0 だからである。

この信頼性・受験者分離に関して Winsteps は 2 つの指標を算出する。信頼性係数（Person Reliability）と受験者分離（Person Separation）である。前者の変動範囲は 0.0 〜1.0 なので天井効果があるが、後者は 0.0 から理論的には無限大の値が可能である。Winsteps のガイドラインによれば、信頼性係数が 0.8 より低い、あるいは受験者分離が 2.0 より低いならば「その測定器具は上位者と下位者を弁別するだけの感度がない（“the instrument may not be not sensitive enough to distinguish between high and low performers”）」。また受験者分離を（4*Separation + 1）/3 に変換した値は受験者階層（Strata）と呼ばれ、「統計的に有意に区別される受験者レベルの数」（“statistically different levels of performance”（Wright, 2001）を表すものである。

前項の分析の 4 つのデータセットの信頼性係数、受験者分離、受験者階層、および素点としての最低点と最高点を表 3 に示す。

表 3　4 つのデータセットの信頼性および受験者分離の程度

	最低素点	最高素点	信頼性	受験者分離	受験者階層
F1/2018	17	117	.94	4.01	5.68
F1/2017	21	118	.93	3.55	5.07
F2/2016	8	117	.92	3.32	4.76
F2/2015	7	114	.93	3.52	5.03

注：F は「フォーム」　素点の満点は 120 点　いずれも $N = 1{,}000$

最低素点と最高素点に着目すると、いずれのデータセットでも最も熟達度の低い受験者でも 0 点はとらず、逆に最も熟達した受験者でも満点はとっていないことがわかる。最高素点と最低素点の差である得点範囲（range）は、100 点、97 点、109 点、107 点と非常に広い。受験者信頼性も .92〜.94 と極めて高い。受験者階層をみるとこれらの 4 つの受験者集団はおおよそ 5 ないし 6 つの異なる能力レベルに識別されたことがわかる。ターゲット集団である日本人大学生全体に対して十分な識別性能を発揮しているといえるだろう。これは VELC Test® が、最も能力の低い受験者でも正解できる項目から、最も能力の高い受験者でも正解できない項目まで幅広い難度で構成されているためと考えられる。

4.5　同一大学内での受験者分離

前項で確認したのは受験者全体から無作為抽出された N = 1,000 という大きな集団のケースである。当然様々な大学からの受験者が混在しており、受験者分離・受験者階層の数値が大きくてもある意味当然である。しかし実際に VELC Test® は、同一大学内での受験者をレベル別に振り分けるプレイスメントに多く使われている。よって同一大学の集団をどの程度分離できているのかを確認することは意味があるだろう。そこで年度別のデータから 5 つの大学を抽出し、それぞれの大学内での受験者分離の状況を調べてみた。5 つの大学は VELC Test® を実際に受験している学力層を、最も低い層から最も高い層までなるべく均等にカバーするよう、VELC スコア平均点が約 50 点刻みで異なる大学を選んだ。5 つのデータセットそれぞれについて改めて Winsteps を走らせて受験者信頼性、受験者分離のアウトプットを確認し、受験者階層数を算出した。表 4 に結果を示す（当該大学での実施担当者を含めて大学名が特定されるのを防ぐため、データを抽出した年度は伏せる）。

VELC スコア平均値は最も高い A 大学が 621.4 である。これは VELC Test® を受験しているなかで最もレベルが高い集団に属するものである。参考までに VELC スコアから予測する TOEIC® L&R の平均は 618.9 である。最もスコアの低い E 大学は VELC スコア平均値が 410.2、TOEIC 予測値の平均が 338.6 である。その間に B 大学、C 大学、D 大学が概ね等間隔で位置していると言ってよいだろう。すなわちこれら 5 大学は VELC Test® を受験している大学のレベルを満遍なくカバーしていると考えられる。

表 4　レベルの異なる 5 集団の得点信頼性および受験者分離の度合い

大学	N	VELC スコア	TOEIC 予測点	最低素点	最高素点	信頼性	受験者分離	受験者階層
A	149	621.4	618.9	56	120	.88	2.74	3.99
B	259	574.0	551.9	13	112	.89	2.86	4.15
C	319	515.3	471.9	24	108	.86	2.46	3.61
D	275	461.9	408.7	23	106	.92	3.31	4.75
E	249	410.2	338.6	24	94	.88	2.69	3.92

120 点満点での素点を見ると A 大学のみ最低が 56 点で他大学を 25 点以上引き離している。また A 大学のみ最高点が満点の 120 点である（この得点は 1 名のみで、次の高得点者は 116 点）。B〜E 大学は、最低素点が 13〜49 であり、最高素点が 94〜112 である。すなわち前項で確認した「最も熟達度の低い受験者でも 0 点はとらず、逆に最も熟達した受験者でも満点はとっていない」という傾向が、A 大学の 1 名のみ

例外として、個別大学データにも当てはまる。

ではA大学は学力が高いために天井効果が起こっているかというと、そういう傾向は見られず、統計的に異なるレベル数を表す「受験者階層」は、A大学でも3.99ある。すなわちA大学の受験者はおよそ4つの階層に分離された。残りの4大学も3.61〜4.75の受験者階層を示している。最も下に位置するE大学でも3.92と、床効果は見られない。下から2番目のD大学は受験者階層が4.75である。すなわちVELC Test® は当該受験者グループの全体的レベルに関わらず、N = 150〜350程度の集団ならば、すくなくとも4レベル程度の異なる下位集団に分離できている、ということである。

4.6　テスト項目の難度と受験者能力の分布

ラッシュモデルの特長として、項目難度と受験者能力を同一の尺度上に配置して視覚的に確認できるということがある。テスト項目難度とA大学の受験者 N = 149の位置関係を示すVariable Mapを図8に、E大学の N = 249の位置関係を示すVariable Mapを図9に示す。難度／能力を表すlogitsを単位とするスケールがグラフ中央に縦に走っており、上に行くほど難度／能力が高く、下に行くほど難度／能力が低い。右手にはテスト項目がXで表され、左手には受験者がXで表されている。

A大学とE大学が受験したフォームは異なるので、ひとつひとつの項目の位置は微妙に異なる。しかし項目群全体としては難度が等化されており、その平均値をMが示す。

図8と図9を比べると、グラフ左半分の受験者を表すXは図8のほうが上に集まっている。A大学の受験者群のほうがE大学の受験者群よりも全体として能力が高いことの現れである。またA大学の受験者群はテスト項目群よりも全体に上方に位置している。受験者能力の平均値をMが示すが、受験者のMは項目のMよりも2 logits近く上にある。これはA大学の多くの受験者にとってテストが比較的容易だったことを表す。一方E大学の受験者群はグラフの下方に位置している。平均を示すMは項目の平均を示すMよりも0.5 logits程度下にある。これは多くの受験者にとってテストがやや難しかったことを示す。

しかしいずれのケースでも受験者は3〜3.5 logitsほどの幅に散らばっており、スコアの幅は広い。受験者階層の数値で確認したように、テストを比較的易しく感じるグループでも、比較的難しく感じるグループでも、スコアは散らばるということである。

5. まとめ

本研究はVELC Test® の本実施データが8年分蓄積したタイミングで改めてテストの性能や特性を確認するため、複数年度データに基づいて、(1) 受験者スコアの分布

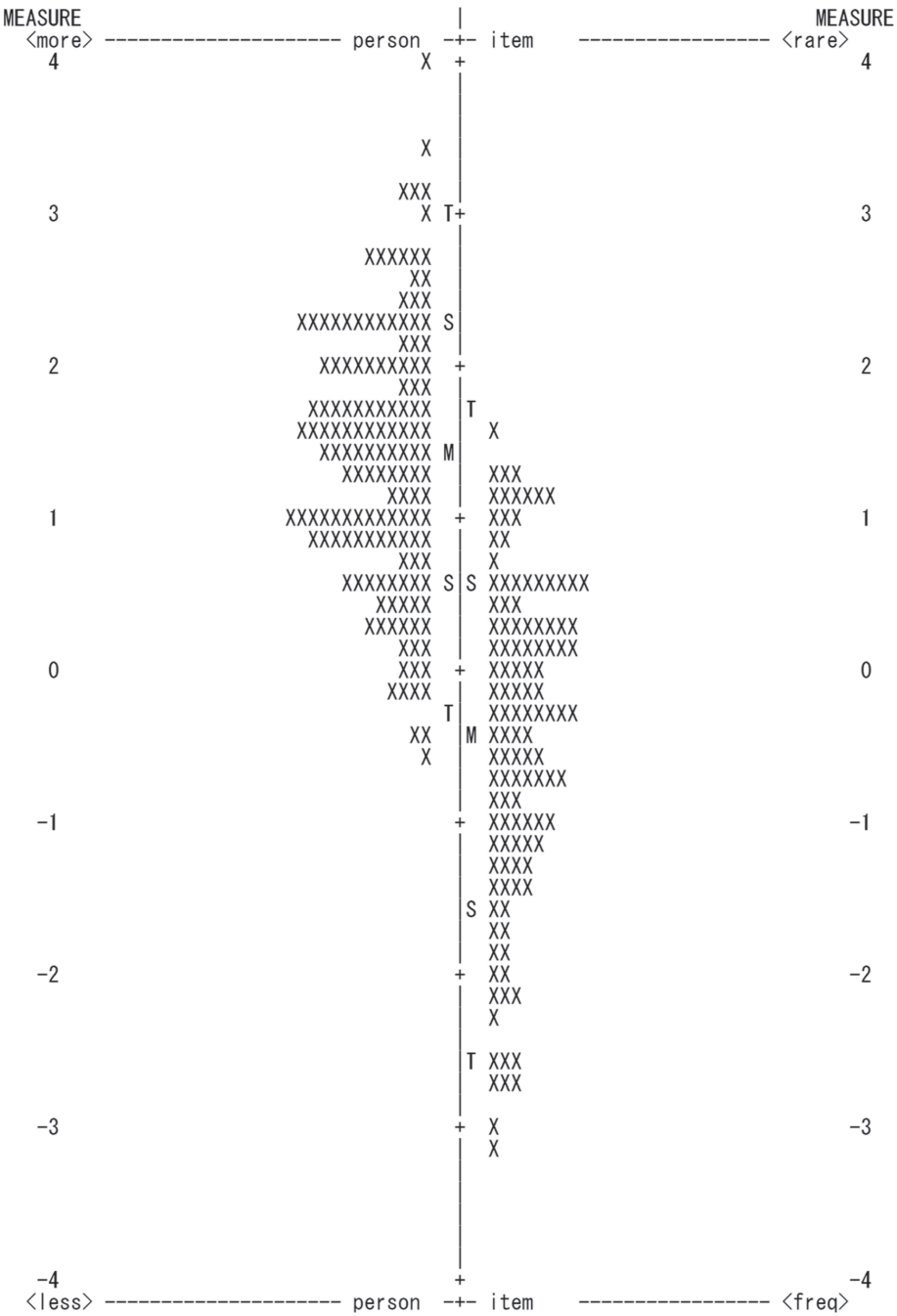

図 8　項目難度（右）と受験者能力（左）の相対的位置関係：A 大学

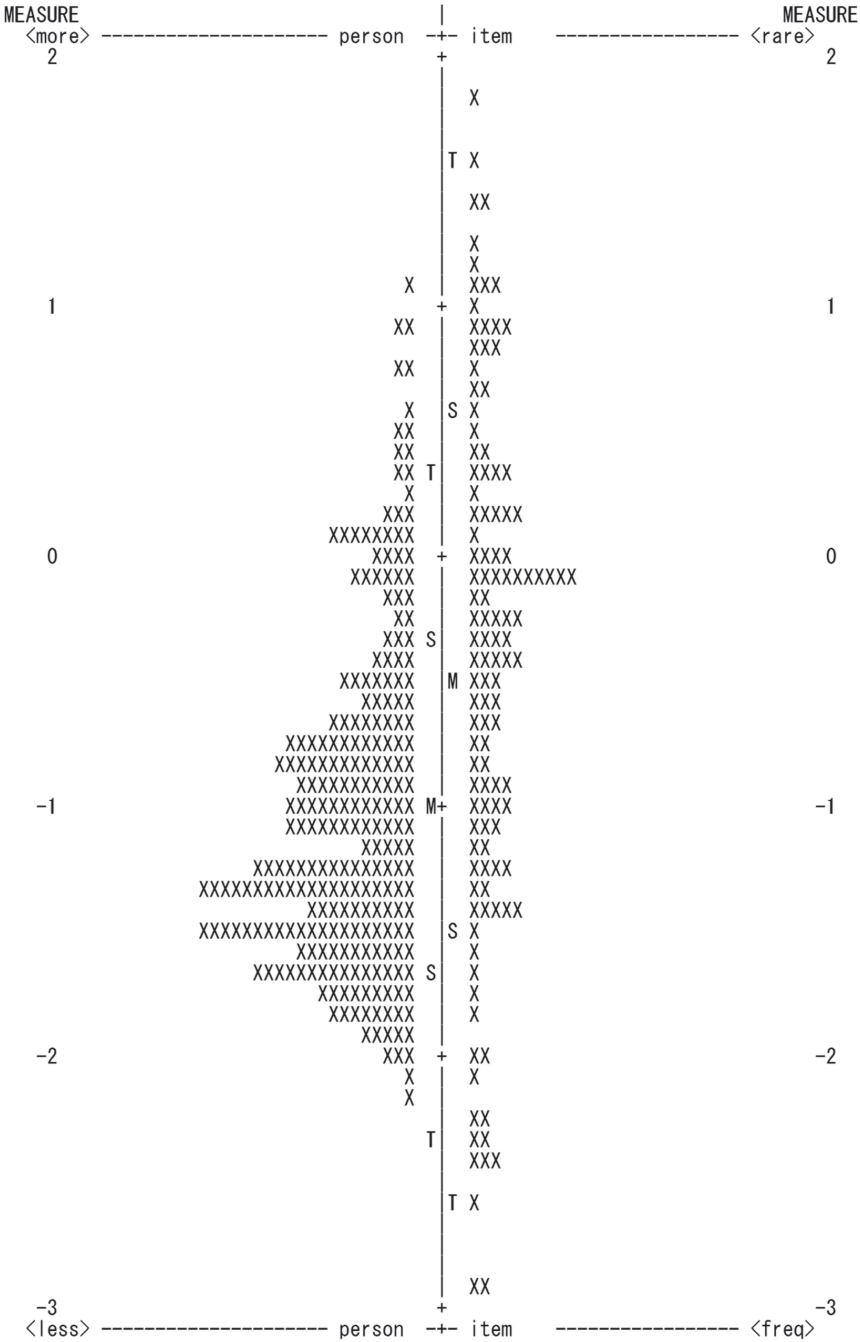

図 9　項目難度と受験者能力の相対的位置関係：E 大学

状況、(2) 項目難度の安定性、(3) 項目のラッシュモデルへの適合度、(4) 信頼性と受験者分離度、を調べたものである。

受験者スコアの分布については、試行段階のサンプルに基づいて設定した「全体平均が 500 で標準偏差が 100」という開発時の想定からはややずれており、本実施の受験者は平均が 475〜485 程度、標準偏差が 70〜80 程度で分布していることが明らかになった。これは開発時には最上位クラスも含めて全国からより幅広い層の受験者の協力を得たことによると考えられる。想定した数値とのズレは、この程度のものであれば、本テストに関して特段の問題を提起するものではない。分布形状についてはおおむね正規分布に近く、ターゲットとしている受験層にたいして適正な難度のテストであると確認されたからである。

項目難度については、固定（anchor）を外した状態で別々の集団によって改めて推定した 2 セットの項目難度の相関係数は r = .980〜.981 と高く、実用上十分な安定性、準不変性が確保されていることが示された。

ラッシュモデルへの適合度については、Infit Mean Square の値が確認したすべての項目について適正な範囲に収まっており、いずれの項目も測定のために重要な役割を果たしていることが明らかになった。すでにトライアル段階でのべ 5,000 名を超える受験者のデータにもとづいて取捨選択された項目であったが、本実施されているデータでも改めて項目の適正さが示されたことは価値がある。

信頼性および受験者分離については、複数大学を合わせた大きな集団（N = 1,000）と、個別大学の比較的小さな集団（N = 149〜319）で検証した。大きな集団で見たときには信頼性は .93 ほどで、全体の受験者を 5 ないし 6 つの統計的に異なるレベルに分離していることがわかった。個別大学で見たときには、信頼性は .86〜.92 ほどで、当該大学の受験者を 4 ないし 5 つの異なるレベルに分離していることがわかった。個別大学については、実際に受験しているなかで最も熟達度の低いレベルの大学（TOEIC® L&R の平均スコアが 340 程度）でも、最も高いレベルの大学（同 620 程度）でも、天井効果や床効果は見られず、おおよそ 4 階層と、学内のプレイスメントを考えたときには実用上十分な程度の受験者分離に成功していることが確認できた。

以上全体として、VELC Test® は日本の大学で学ぶ学習者の英語熟達度を測定するツールとして有効的に機能していることが改めて確認できたと言える。

謝辞

データを管理してくださっている英語能力測定・評価研究会の事務スタッフの方々と、本論文の草稿に有益なコメントを下さった同研究会の望月正道先生と熊澤孝昭先生に、心より御礼申し上げます。

引用文献

静哲人（2012a）「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（2012.10.27）専修大学生田キャンパス.

静哲人（2012b）「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」2012 年度 JACET 関西支部秋季大会（2012.11.24）京都産業大学.

静哲人（2013）「VELC テストの測る英語力構造：確認的因子分析がスコアレポート方式に示唆するもの」大学英語教育学会第 52 回国際大会（2013.8.30）．京都大学吉田キャンパス.

静哲人（2014）「VELC Test（R）フォーム A の選択肢分析から見える各アイテムの特性」大学英語教育学会第 53 回国際大会（2014.8.28）．横浜市立大学.

静哲人（2015a）「VELC Test フォーム A の選択肢特性分析」大東文化大学語学教育研究所創設 30 周年記念フォーラム，97-115.

静哲人（2015b）「ベルクテストの概要とよくある質問：Listening Section Part 2 の作問意図と項目特性」ベルク研究会第 4 回研究会基調講演（2015.9.12）．研究社英語センター.

静哲人（2017）「2017 年度実施 VELC Test® データからみる同一大学内での受験者分離の成功度」日本言語テスト学会第 21 回研究大会（2017.9.10）会津大学.

静哲人・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.

静哲人・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」第 51 回大学英語教育学会研究大会（2012.9.1）愛知県立大学.

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

Fisher, W.（1992）Reliability, separation, strata statistics. Rasch measurement transactions, 6: 3, p. 238. Retrieved https://www.rasch.org/rmt/rmt63i.htm

Kumazawa, T. Shizuka, T. Mochizuki, M., & Mizumoto, A.（2016）Validity argument for the VELC Test® score interpretations and uses. *Language Testing in Asia* 6:2 https://doi.org/10.1186/s40468-015-0023-3

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

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.（2016）Modification of VELC Test® listening section part 2 type multiple-choice 1-blank partial “dictation” items: Effects on distractor discriminations and TOE-IC®-relatedness. 大学英語教育学会第 55 回国際大会（2016.9.3）北星学園大学.

Wright, B. D.（2001）Separation, reliability and skewed distributions: Statistically different sample-independent levels of Performance. Rasch Measurement Transactions, 14: 4, p.786. Retrieved https://www.rasch.org/rmt/rmt144k.htm