VELC Test® Online と VELC Test® P&P の等価性を検証する[1]

英語能力測定・評価研究会（VELC 研究会）

1　はじめに

VELC Test®は 2012 年度から実施されてきている日本語を母語とする大学生のための英語力熟達度テストである（VELC は Visualizing English Language Competency のアクロニムである）。リスニングセクションとリーディングセクション合わせて 120 項目によって聞く技能と読む技能からみた熟達度を測定するもので、主としてプレイスメントや授業効果の測定のために利用されている。その信頼性、妥当性、項目特性などについては繰り返し検証がなされてきている(靜, 2012a; 2012b; 2013; 2014; 2015a; 2015b; 2017; Shizuka, 2016; 靜・望月, 2014; Kumazawa *et al.* 2016)が、直近では靜(2020)が、実施開始から 8 年間のデータについて総括的に分析し、VELC Test®が日本の大学生の英語力測定のために有効に機能してきていることを確認した。

ところが 2020 年初頭から全国的に新型コロナウイルスの感染拡大があり、予定されていた VELC Test®実施の多くが取り止めとなった。そこでこの状況下でも実施できる VELC Test® Online を急遽開発し、2020 年 7 月から運用を開始した。

2　VELC Test® Online

VELC Test® Online（以下では OL 版とする）は、従来のペーパー版の VELC Test®（以下では PP 版とする）をシンプルにそのままオンラインテスト化したものである。PP 版と項目セットは同一、レイアウトもページを横にめくる紙冊子と下にスクロールする PC スクリーンの違いを除いては同一、試験時間も同一であり、受験者の応答によって提示される項目が枝分かれしてゆく適応型テストではない。PP 版と OL 版の差はシンプルに紙媒体の問題冊子を読んで紙媒体のマークシートに解答するか PC スクリーン上で解答するかである。音声については PP 版では試験教室のスピーカーから一斉に聞こえるが OL 版では個人の PC のスピーカーあるいはヘッドセットから聞こえる、という違いはあるが、本質的な差異とは考えがたい。唯一、試験監督がいない状況で受験することが多い OL 版は潜在的には辞書使用などの不正行為が不可能ではないが、まっとうに受験する限り、PP 版と OL 版は等価でなくなる要因は思い当たらない。しかし OL 版実施の初年度に、その想定が実際のテストデータによって裏付けられるかを検証することは意味のあることである。

3　本研究

3.1 目的

本研究は、理論上等価であるはずの OL 版が実際に PP 版と等価であることを、受験データによって検証することを目的とする。

3.2 限界

PP 版と OL 版の等価性に関する「物的証拠」を集めるためには、かなりの程度の数の受験者に PP 版と OL 版を同日に、あるいは少なくともあまり日にちをあけずに受験してもらい、同一受験者の PP 版と OL 版のデータを比較することが必要である。しかし今般の状況下ではそのようなデザインは不可能であるし、教育的に望まし

[1]言語教育エキスポ（2020.10.25)で発表した「VELC Test® Online と VELC Test® P&P の等価性を検証する・その 1」と日本言語テスト学会(2020.12.12)で発表した「同・その 2」の内容を統合した未公刊論文である（2021.1.10)。

いとも考えられない。今回は、別々の受験者集団が PP 版と OL 版を受験したデータを、もしくは同一の受験者集団が別々の時点で受験した PP 版と OL 版のデータを比較した結果を「状況証拠」として、総合的に推測することしかできない。本研究はそのような限界内における報告であることを断っておく。

3.3 分析対象データ

2020 年度に OL 版を受験したのは 3 大学である。これらの大学はいずれも 2019 年度には PP 版を受験している。受験時期は OL 版も PP 版もそれぞれの年度の 7 ～ 9 月すなわちおおよそ年度の中頃である。本報告ではその 3 大学（A 大学、B 大学、C 大学）のべ 3,235 名のデータを分析対象とする。大学別、カテゴリー別の人数を表 1 ～5 に示す。2019 年度に PP 版を受験した学生群と 2020 年度に OL 版を受験した学生群には一部重なりがある。

表 1　A 大学の 2019 年度 PP 版と 2020 年度 OL 版の受験者数

2019_PP			2020_OL		
n = 370			*n* = 377		
1 年	2 年	学年不明	1 年	2 年	学年不明
n = 185	*n* = 178	*n* = 7	*n* = 195	*n* = 173	n = 9

表 2　B 大学の 2019 年度 PP 版と 2020 年度 OL 版の受験者数

2019_PP			2020_OL		
n = 164			*n* = 140		
1 年	2 年	学年不明	1 年	2 年	学年不明
n = 77	*n* = 86	*n* = 1	*n* = 69	*n* = 69	*n* = 2

表 3　C 大学 S 学部の 2019 年度 PP 版と 2020 年度 OL 版の受験者数

2019_PP			2020_OL		
n = 217			*n* = 342		
1 年	2 年	学年不明	1 年	2 年	学年不明
n = 217	*n* = 0	*n* = 0	*n* = 342	*n* = 0	*n* = 0

表 4　C 大学 T 学部の 2019 年度 PP 版と 2020 年度 OL 版の受験者数

2019_PP			2020_OL		
n = 288			n = 239		
1 年	2 年	学年不明	1 年	2 年	学年不明
n = 152	*n* = 136	*n* = 0	*n* = 239	*n* = 0	*n* = 0

表 5　C 大学全学部の 2020 年度 OL 版の受験者数

			2020_OL		
			n = 1679		
1 年	2 年	学年不明	1 年	2 年	学年不明
					n = 1,679

注：S 学部と T 学部の OL 版受験者を内数として含む

3.4 分析方法

分析は以下の手順によった。まず（1）2019 年以前の PP 版のスコア分布と 2020 年の OL 版の学年別、パート別の分布を比較した。次に PP 版と OL 版の、（2）信頼性および受験者分離、（3）同一フォームでの選択肢の選択状況、（4）同一フォームでの項目正答率を比較し、最後に、（5） OL 版の Rasch モデル適合度を確認した。（1）に関しては、ベルクスコアデータ（パート別スコア、セクション別スコア含む）を利用した。（2）～（5）に関しては、生データ（各受験者が各項目に対して選んだ選択肢データ）を利用した。なお生データには解答者の学年情報が含まれていない仕様であるため、学年を問わず PP 版データ全体、OL 版データ全体として分析した。

4 分析結果

4.1　学年ごとのベルクスコア比較

4.1.1　A 大学

まず A 大学についてのベルクスコアを、2019 年度に PP を受けた 1 年生、2020 年度に OL を受けた 1 年生、2019 年度に PP を受けた 2 年生、2020 年度に OL を受けた 2 年生に分けて一覧にした（表 6）。いくつかの観察ができる。

（A 1）いずれの学年でも、総合スコアにおいて OL 版のほうが PP 版よりも高い傾向が顕著である。パート別にみると、この傾向はリスニングにのみ当てはまる。

（A 2）同一年度（＝同一モード）で比較すると、 1 年生よりも 2 年生のスコアが低い。これは別々の集団のクロスセクショナルな比較であることに留意したい。

（A 3）2019 年度 1 年生の PP でのスコア（総合）と 2020 年度の OL でのスコア（総合）はほとんど変化がない。これは同一集団の経年的な比較であることに留意したい。

表 6　A 大学の年度（モード）別・学年別平均値の一覧

学年	Mode	年度	N	総合	L	L1	L2	L3	R	R1	R2	R3
1	PP	2019	185	556.6	540.7	551.4	522.4	535.2	582.0	573.5	575.0	573.6
1	OL	2020	195	574.6	581.2	558.2	579.8	577.4	577.4	575.6	554.0	575.2
2	PP	2019	178	543.2	536.5	542.8	534.8	518.4	556.3	559.7	550.7	545.8
2	OL	2020	173	558.2	560.1	542.5	570.9	546.0	563.8	552.5	549.2	566.4

A1 の傾向が見やすくなるように、2020 年度 OL 版の平均値から 2019 年度 PP 版の平均値を引いた数値を、表 7 に示した。

表 7　A 大学の年度（モード）の異なる同一学年（＝別々の集団）の平均値の差

学年	Mode	年度	N	総合	L	L1	L2	L3	R	R1	R2	R3
1	PP	2019	185									
1	OL	2020	195	18.0	40.5	6.8	57.4	42.3	-4.6	2.1	-21.0	1.6
2	PP	2019	178									
2	OL	2020	173	15.0	23.6	-0.3	36.0	27.6	7.5	-7.2	-1.6	20.6

セクション別に見ると 1 年生、2 年生ともに、L2 の得点差が最も大きく、次に L3 が大きい。これは異なる集団が異なるモードのテストを受けた結果なので、直ちに集団の力の差であるともモードによる差であるとも決め

ることはできないことは、再度確認したい。ただし PP と OL のモードの違いによってテストとしての難易度が変わってくると考えられる合理的な理由が存在しないことは上に述べた通りである。よって 1 年生に関しては、2019 年度入学の集団よりも 2020 年度入学の集団のほうの英語力がもともと高かった、2 年生に関しても 2 年時の夏の時点で 2019 年度の集団よりも 2020 年度の集団の英語力が高かった、という可能性が高いと考える

次に A 2 の傾向が視認しやすくなるように、2019 年度に PP を受けた 1 年生と 2 年生の、2020 年度に OL を受けた 1 年生と 2 年生の、それぞれの得点差を表 8 に示す。いずれの年度（＝モード）においても、1 年生よりも 2 年生のほうのスコアが低い傾向がある。最も差が大きいのは 2020 年度に OL で測定した L3（リスニングの内容理解力）である。確認しておくと、これは別々の集団の比較であり、同一の集団の英語力が 1 年生から 2 年生にかけて低下したということではない。それぞれの年度内の 1 年生と 2 年生の比較は、それぞれ PP のみ、OL のみによる結果なので、モード効果は無関係である。つまり 2019 年度にも、2020 年度にも、夏の時点において 1 年生は 2 年生よりも英語力が高かったのである。すなわち A 大学では年々入学者の英語力が高くなっている可能性がある。A1 の傾向と考え合わせるならば、この解釈が概ね当たっているのではないかと思われる。

表 8　A 大学の同一年度（モード）内の、1 年生と 2 年生の平均値の差

学年	Mode	年度	N	総合	L	L1	L2	L3	R	R1	R2	R3
1	PP	2019	185									
1	OL	2020	195									
2	PP	2019	178	-13.4	-4.2	-8.6	12.4	-16.8	-25.7	-13.7	-24.3	-27.8
2	OL	2020	173	-16.3	-21.1	-15.7	-9.0	-31.4	-13.6	-23.1	-4.8	-8.8

最後に A 3 の傾向が視認しやすくなるように、2019 年度入学の集団の 1 年次（PP で測定）と 2 年時（OL で測定）のスコアの差を表 9 に示す。ベルクスコア総合の平均値の差はわずかに 1.7 で事実上差はない。これは同一集団の英語力を約 1 年のインターバルを隔てて測定したものである。つまり当該集団は 1 年時の夏と 2 年時の夏の英語力がほぼ等しいということが示唆された、ということである。これは当該大学の英語授業効果という観点からすると喜べない結果である。しかし同一集団が時を隔てて 2 度受験した結果のスコアがほとんど同じだったということは、測定結果を出したテストの安定性、信頼性を表していると考えられる。つまりこの A 3 の傾向が PP と OL の等価性を間接的に示す「状況証拠」だとすることは可能である。なおセクションスコアを見てみると、特に L 2（リスニングの音声解析）、ついで L 3（リスニングの内容理解）ではスコアが伸びたが、リーディングと L1 などその他のセクションでスコアを落としたことで結果的に相殺されたことが伺われる。大学の授業によって、高校時代などそれまでは十分でなかったリスニング分野のスキルがやや強化され、反面、語彙を中心にリーディング分野の知識が低下したということかも知れない。

表 9　A 大学の同一集団（2019 年度 1 年生→2020 年度 2 年生）の経年変化

学年	Mode	年度	N	総合	L	L1	L2	L3	R	R1	R2	R3
1	PP	2019	185									
2	OL	2020	173	1.7	19.4	-8.9	48.5	10.8	-18.1	-21.0	-25.9	-7.2

4.1.2　B 大学

B 大学についても同様に、まず年度、集団ごとの平均値を表 10 に示す。以下の観察が可能である。

（B 1 ）A 大学と同様、いずれの学年でも 2019 年度の PP 版よりも 2020 年度の OL 版のほうが、平均値が高い。パート別にみるとその内訳は A 大学とは対照的に、リスニングでもリーディングでも同程度に見られる。

（B２）同一年度（＝同一モード）で比較すると、A 大学と異なり、１年生よりも２年生のスコアが高い。これはクロスセクショナルな比較であるが、順当か結果と言える。

（B３）2019 年度１年生の PP でのスコア（総合）と 2020 年度の OL でのスコア（総合）は、466.2→505.6 と向上している。これは同一集団の経験的な変化であり、A 大学と異なり望まれるべき結果が出ている。

表 10　B 大学の年度（モード）別・学年別平均値

学年	Mode	年度	N	総合	L	L1	L2	L3	R	R1	R2	R3
1	PP	2019	77	466.2	489.9	474.5	490.9	502.7	437.7	444.9	456.1	436.8
1	OL	2020	69	484.6	509.3	491.4	524.0	506.8	466.0	477.2	460.6	461.7
2	PP	2019	86	481.1	512.9	489.5	508.6	529.9	447.1	469.3	458.0	428.5
2	OL	2020	69	505.6	538.6	519.9	544.6	535.2	481.4	509.8	463.7	469.7

B1 の傾向を視認しやすくするために、2020 年度 OL 版の平均値から 2019 年度 PP 版の平均値を引いた数値を表 11 に示した。１年生も２年生も 2020 年度の得点のほうが高い。１年生については 2019 年度入学の集団よりも 2020 年度入学の集団のほうが英語力があったと考えられる。２年生に関しても 2020 年度生のほうが英語力が高かったと考えられる。リスニングもリーディングも同様の傾向がある。ただし L3（リスニングの内容理解）と R2（文法解析）については差が小さい。

表 11　B 大学の年度（モード）の異なる同一学年（＝別々の集団）の平均値の差

学年	Mode	年度	N	総合	L	L1	L2	L3	R	R1	R2	R3
1	PP	2019	77									
1	OL	2020	69	18.4	19.4	16.9	33.0	4.1	28.3	32.3	4.5	24.9
2	PP	2019	86									
2	OL	2020	69	24.5	25.7	30.5	36.0	5.3	34.3	40.6	5.7	41.3

次に B2 の傾向が視認しやすくなるように、2019 年度に PP を受けた１年生と２年生の、2020 年度に OL を受けた１年生と２年生の、それぞれの得点差を表 12 に示す。PP においては総合で 14.9 の差が見られ、OL においては 21.0 の差が見られる。その差はどちらかと言えばリスニングの差から来ているが、R1（読む場合の語彙）の差もかなりある。つまり B 大学においては、例年、１年生よりも２年生の英語力が高い、という傾向が見られ、それは測定モードが 2019 年度のように PP であっても、2020 年度のように OL であっても変わらない、ということである。つまり OL が PP と同様の測定をしていることの状況証拠がある、と言える。

表 12　B 大学の同一年度（モード）内の、１年生と２年生の平均値の差

学年	Mode	年度	N	総合	L	L1	L2	L3	R	R1	R2	R3
1	PP	2019	77									
1	OL	2020	69									
2	PP	2019	86	14.9	23.0	15.0	17.7	27.2	9.4	24.3	1.9	-8.3
2	OL	2020	69	21.0	29.3	28.5	20.7	28.4	15.4	32.6	3.1	8.1

そして B３の傾向を視認しやすくするために、2019 年度入学の集団の１年時（PP で測定）と２年時（OL で

測定）のスコアの差を表 13 に示す。B 大学ではこの集団は 1 年時から 2 年時にかけて明らかに英語力が向上している。総合スコアで 40 点近い向上をしている。もっとも向上の幅が大きいのは R 1（読む語彙力）で、実に 64.9 である。ついで伸び幅が大きかったのは L 2（音声解析）である。B 大学では例年、 1 年時から 2 年時にかけてバランスよく英語力が向上する傾向がある、という仮説を設定する。

表 13　B 大学の同一集団（2019 年度 1 年生→2020 年度 2 年生）の経年変化

学年	Mode	年度	N	総合	L	L1	L2	L3	R	R1	R2	R3
1	PP	2019	77									
2	OL	2020	69	39.3	48.7	45.4	53.7	32.5	43.7	64.9	7.6	33.0

そこでこの仮説を検証すべく、B 大学における 2017 年度～2020 年度入学者のベルクスコア経時変化のデータを調べて一覧にしたのが表 14 である。

表 14　B 大学 2017 年度～2020 年度入学者のベルクスコア経時変化

テストモード	PP	PP	PP	PP	PP	PP	PP	OL
時期	2017年4月	2017年8月	2018年4月	2018年8月	2019年4月	2019年8月	2020年4月	2020年9月
2017年入学者	451.8	465.3	488.1	509.6				
2018年入学者			440.5	465.5	478.6	481		
2019年入学者					444.1	466.2	492.9	505.2
2020年入学者							455.3	484.6

表 14 に示した平均値の中で、網掛けのある 2020 年 9 月の 505.2（2019 年度入学者）と 484.6（2020 年度入学者）のみが OL による数値であり、ほかはすべて PP による数値である。我々の興味は、このふたつの OL による数値が、他の PP による数値と異質でないとみなせるか、という点である。そこで視認しやすくするために、1 年生 4 月を Time 1、1 年生 8 月もしくは 9 月を Time 2、2 年生 4 月を Time 3、2 年生 8 もしくは 9 月を Time 4 と呼び、Time 1 から Time 2 、Time 1 から Time 3、Time 2 から Time 3、Time 1 から Time 4 までのスコアの向上幅を、異なる年度の入学者グループ毎にまとめて表 15 として示す。

表 15　B 大学 2017 年度～2020 年度入学者のベルクスコア向上幅

	Time 1→Time 2	Time 1→Time 3	Time 2 →Time 4	Time 1→Time 4
	1 年春→ 1 年夏	1 年春→ 2 年春	1 年夏→ 2 年夏	1 年春→ 2 年夏
2017 年入学者	13.5	36.3	44.3	57.8
2018 年入学者	25.0	38.1	15.5	40.5
2019 年入学者	22.1	48.8	39.0	61.1
2020 年入学者	29.3			

注：網掛けしてある数値は、PP および OL による測定に基づいている

網掛けしてある数値は、OL による測定スコアから PP による測定スコアを減じたものである。Time 1 から Time 4 までの向上幅を見ると 2019 年度入学者の 61.1 は 2017 年度入学者の 57.8 とほぼ等しい。Time 2 から Time 4 までの向上幅をみると 2019 年度入学者の 39.0 は、2017 年度入学者の 44.3 よりも小さい。また Time 1

から Time 2 の向上幅を比較すると、2020 年度入学者の 29.3 は 4 つの値で最大ではあるが、2018 年度入学者の 25.0 よりも格段に大きいとまではいえない。以上を総合的に判断すると、B 大学においては例年（その幅については年度によりムラがあるものの)、入学時から 1 年間、 1 年と半年間の英語学習を経て着実にスコアが向上しているといえ、2020 年 9 月に実施した OL の数値もその傾向を捉えていると考えるのが妥当だと言える。すなわち OL は従来の PP と同じように機能しているものと判断される。

4.1.3　C 大学 S 学部

S 学部についてのベルクスコアを、2019 年度に PP を受けた 1 年生、2020 年度に OL を受けた 1 年生、に分けて一覧にした (表 16)。数値が高いほど赤色が濃くなるように色分けしてある。2019 年度の PP 版よりも 2020 年度の OL 版のほうが明らかに、かつ格段にスコアが高い。確認しておくと、これは別々の年度に入学してきた別々の集団の比較である。

表 16　S 学部の年度（モード）別平均値の一覧

学年	Mode	年度	N	総合	L	L1	L2	L3	R	R1	R2	R3
1	PP	2019	217	407.5	407.5	407.0	416.3	420.8	394.1	417.3	409.8	393.6
1	OL	2020	342	463.0	470.0	481.3	470.1	462.8	457.0	511.1	418.7	453.9

2019PP と 2020OL の差が見やすくなるように、2020 年度 OL 版の平均値から 2019 年度 PP 版の平均値を引いた数値を、表 17 に示した。最も大きな差は語彙パートである L1 と R 1 に見られる。L1 で 74.4、R 1 では実に 93.8 の差があり、2020OL のスコアが高い。他のパートも軒並み 2020OL のほうが 40〜60 点ほど高いのだが、唯一 R 2 （構文解析）だけは差が 8.9 と、他のパートに比べると格段に小さい。

表 17　S 学部の年度とモードの異なる同一学年（＝別々の集団）の平均値の差

学年	Mode	年度	N	総合	L	L1	L2	L3	R	R1	R2	R3
1	PP	2019	217									
1	OL	2020	342	55.5	62.5	74.4	53.8	42.0	62.9	93.8	8.9	60.3

これは異なる集団が異なるモードのテストを受けた結果なので、直ちに集団の力の差であるともモードによる差であるとも決めることはできないことは、再度確認したい。ただし PP と OL のモードの違いによってテストとしての難易度が変わってくると考えられる合理的な理由が存在しないことは上に述べた通りである。また万一何らかのメソッド効果が存在したにせよ、これほどのスコア差となって現れることは考えられない。よって C 大学 S 学部の 1 年生に関しては、2019 年度入学の集団よりも 2020 年度入学の集団のほうがもともとかなり英語力が高かった、という可能性が極めて高いと考える。

これを確かめるために、C 大学 S 学部の過年度データを調べてみた。2017 年度と 2018 年度の同学部の 1 年生の平均スコアの推移は表 18 の通りである。いずれの年度も 410 点前後で入学してきており、そのレベルは 1 年時が終わる時点になってもそれほど変化していない。これはまさに表 3 に示された 2019 年度入学生のスコア帯とほぼ同一である。つまり C 大学 S 学部では少なくとも過去数年間はベルクスコア平均値が 410 点前後の学生が入学してきていた。それを考えると 2020 年度に OL で測定した結果の平均 463 は明らかにかつ大幅に例年と異なる。(※このような過年度の推移データが利用可能だったのは S 学部に関してのみである。T 学部は受験していない。)

表 18　S 学部の 2017 年度、2018 年度入学生の 1 年時の平均スコア推移

	4月			12月		
	N	平均値	標準偏差	N	平均値	標準偏差
2017年度 1 年生	169	408	40	159	407	53
2018年度 1 年生	139	413	53	137	417	50

以上を考え合わせる C 大学 S 学部における 2019PP と 2020OL の大きな差は、PP と OL の違いによるメソッドイフェクトではなく、入学年度の異なる 1 年生の英語力レベルの差を反映していると考えられる。

4.1.4　C 大学 T 学部

S 学部と同様に、まず年度、モードごとの平均値を表 19 に示す。以下のことが観察できる。

（1）1 年生同士を比べると OL 版を受けた 2020 年度のスコアのほうが PP 版を受けた 2019 年度のスコアよりも格段に高い。OL 版の平均値は 463.6 で、これは T 学部の 463.0 という数値はほぼ同一である。S 学部同様、特に R1（目で見る語彙）のスコアの高さが際立っている。506.4 という数値は、S 学部の 511.1 にかなり近い。C 大学では S 学部も T 学部も、2020 年度には特に読む場合の語彙力が例年よりも格段に高い学生が入学したと考えられる。

（2）2019 年度に PP 版を受けた 1 年生と 2 年生を比べると、2 年生のほうがわずかに高い。

表 19　T 学部の年度（モード）別・学年別平均値

学年	Mode	年度	N	総合	L	L1	L2	L3	R	R1	R2	R3
1	PP	2019	152	399.4	395.5	403.6	414.3	395.7	399.4	403.3	404.0	403.7
1	OL	2020	239	463.6	471.0	487.1	466.7	464.9	457.1	506.4	420.0	458.1
2	PP	2019	136	408.0	405.1	413.5	419.9	407.9	406.4	419.7	398.8	419.9

（1）の傾向を視認しやすくするために、2020 年度 OL 版の平均値から 2019 年度 PP 版の平均値を引いた数値を表 20 に示した。総合的ベルクスコアの差は 64.3 で、CS 大学の場合の 55.5 よりも更に差が大きい。どのパートでもほぼまんべんなく 2020OL の数値が高い。R1 については実に 103.1 の差である。R 2 についても 15 点近くの差がある。

表 20　T 学部の年度（モード）の異なる同一学年（＝別々の集団）の平均値の差

学年	Mode	年度	N	総合	L	L1	L2	L3	R	R1	R2	R3
1	PP	2019	152									
1	OL	2020	239	64.3	75.5	83.6	52.4	69.2	57.7	103.1	15.9	54.5

次に（2）の傾向が視認しやすくなるように、2019 年度に PP を受けた 1 年生と 2 年生の得点差を表 21 に示す。おおよそ 5〜15 点くらいの幅で、2 年生のほうの英語力が高い。（ただし R 2 については逆転している。）ほぼ 1 年間の英語学習の成果として十分かどうかはともかく、順当な結果と解釈される。

以上の結果を総合すると、C 大学では S 学部でも T 学部でも、2020 年度はそれまでよりも格段に英語力、とくに読む場合の語彙力の高い層の学生が入学して来たと考えられる。

表 21　T 学部の同一年度、同一モードの 1 年生と 2 年生の平均値の差

学年	Mode	年度	N	総合	L	L1	L2	L3	R	R1	R2	R3
1	PP	2019	152									
1	OL	2020	136	8.6	9.6	9.9	5.6	12.2	7.0	16.4	-5.2	16.2

そこで事務局を通じて C 大学の事務担当者に入学者のレベル感について照会したところ、予想通り「教員に聞くところでは明らかに上がっているそうである」「昨年は受験生が浪人を嫌ったため、レベルを落としてでも受かりたいという慎重派が多く見られ、本学も高い競争率をマークした。」「そういった外的要因が良い方向に作用し、入学者のレベルアップにつながったのだと分析している」との回答を得た。

以上の両学部の学年別のベルクスコアの分析からは、S 大学の 2019PP と 2020OL のスコアの違いは、PP と OL のモードの違いからくるものではなく、ふたつの年度の入学者の英語力レベルの違いがきていることの状況証拠が得られたと考える。

以上の 3 大学のスコアデータ分析をまとめると、あくまで状況証拠としてではあるが、従来の PP による測定と OL による測定に質的な違いはないことが示唆されていると考えられる。

4.2 信頼性、受験者分離、受験者階層

つぎに Rasch モデリングのソフトウェアである Winsteps によって信頼性、受験者分離、受験者階層を比較した。ここでいう信頼性とは Winsteps では person reliability と呼ばれるもので、当該の受験者集団が測定誤差を考慮したときにどの程度よく弁別されているかを示す指標であり、素点の場合の Cronbach's alpha に相当する。受験者分離は 0.0〜1.0 の範囲の値しかとれない信頼性係数の限界を補い、0.0 から理論的には無限大までの値をとる指標である。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 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。

この分析には生データ（各受験者が各項目で正解したか誤答だったかがわかるデータ）が必要だが、生データは管理の都合上、個々の受験者の学年情報が不明であるフォーマットで保存されている。そこで以下の分析は大学ごとに、2019 年度の PP 版データセット、2020 年度の OL 番データセットを単位として行った。すなわちひとつのデータセットには 1 年生、2 年生、および少数のそれ以外の学年に所属する学生が含まれている。つまり受験者の学年に関わりなく、当該大学でその時期に PP 版を、あるいは OL 版を受験した学生集団全体がどのように英語力によって弁別されたか、という観点で分析する。

4.2.1　A 大学

A 大学で 2019 年度に PP を受験したのは 370 名、2020 年度に OL 版を受験したのは 377 名である（いずれもそれぞれの年度の 1 年生、2 年生、その他を含む)。それぞれのデータセットに対して Winsteps を走らせて得られたアウトプットの中から、素点情報および信頼性、受験者分離、受験者階層の数値を表 22 に示す。

VELC Test®は 120 項目からなるテストであり、素点の取りうる範囲は 0〜120 である。表 22 を見ると、PP と OL の素点が分布している範囲は非常に似通っていることがわかる。最低素点が 35 と 36、最高素点が 113 と 117 であり、どちらの集団でも、最も力のない受験者も 0 点は取らず、最も力のある受験者も満点は取れていない。信頼性係数は.87 と.89 であり、どちらも十分に高いが OL のほうが僅かに値が高い。これを受験者階層に換算してみると、3.85 と 4.09 である。すなわち PP は当該集団を統計的に有意に区別されうるほとんど 4 つの能力レ

ベルに、OL は 4 つわずかに超える能力レベルに弁別したということがわかる。

表 22　A 大学の 2 つのデータセットの素点の記述統計、信頼性および受験者分離

	N	最低素点	最高素点	平均値	標準偏差	信頼性	受験者分離	受験者階層
2019 年度 PP 版	370	35	113	82.7	13.0	.87	2.64	3.85
2020 年度 OL 版	377	36	117	86.3	13.0	.89	2.82	4.09

注: 素点の満点は 120 点

4.2.2　B 大学

同様に B 大学で 2019 年度に PP を受験したのは 164 名、2020 年度に OL を受験したのは 140 名である。やはりそれぞれに異なる学年の学生を含むデータセットである。素点情報および信頼性、受験者分離、受験者階層の数値を表 23 に示す。

表 23　A 大学の 2 つのデータセットの素点の記述統計、信頼性および受験者分離

	N	最低素点	最高素点	平均値	標準偏差	信頼性	受験者分離	受験者階層
2019 年度 PP 版	164	32	104	63.1	14.1	.88	2.72	3.96
2020 年度 OL 版	140	31	110	68.3	15.4	.90	3.07	4.43

注: 素点の満点は 120 点

A 大学と同様に、PP でも OL でも、最も力の低い受験者も 0 点は取らず、最も力のある受験者も満点は取れていない。信頼性係数は.88 および.90 と十分に高く、受験者階層に換算した場合には、PP はおおよそ 4 レベルに、OL はおおよそ 4 レベル半に当該の受験者グループを弁別したことがわかる。なお受験者分離や受験者階層の数値が OL のほうが高いことは、測定ツールとして OL のほうが PP よりも弁別力があるという意味ではない。あくまで別々の集団を測定しているのであって、OL 版を受けた 2020 年度の受験者集団のほうがわずかに英語力の多様性があったという解釈のほうが妥当である。

4.2.3 C 大学

C 大学に関しては、2019 年度 12 月に PP を受けた全学生 $n = 505$ （S 学部と T 学部をあわせたもの）と、2020 年度 7 月に OL を受けた全学生 $n = 1{,}679$（S 学部 T 学部に加えて他学部もあわせたもの）の生データを分析対象とした。ベルクスコアを算出する方式通り項目難度をアンカー（固定）した状態で、Winsteps を走らせた結果を表 24 に示す。

表 24　C 大学の 2 つのデータセットの素点の記述統計、信頼性および受験者分離

	N	最低素点	最高素点	平均値	標準偏差	信頼性	受験者分離	受験者階層
2019 年度 PP 版	505	10	102	45.6	13.5	.85	2.37	3.49
2020 年度 OL 版	1679	22	99	61.7	13.4	.86	2.45	3.60

注: 素点の満点は 120 点

いずれのモードでも、最も力の低い受験者も 0 点はとらず、最も力の高い受験者も満点（120）はとらない、というのは従来の VELC Test®の傾向と共通である。平均素点は PP 版より OL 版のほうが約 15 点高いが、これは PP 版を受けたのが 2 学部、OL 版を受けたのが全学部であり、学部間の英語力の違いの反映だと考えられる。いずれの場合でも、受験者信頼性は PP で.85、OL で.86 と十分に高い。受験者階層を見ると、いずれのテスト

も受験者グループ全体を、おおよそ 3 ～ 4 レベルの統計的に分離可能な階層に分離している。

4.2.4 従来の PP データとの比較

上の 4.2.1～4.2.3 をあわせて見ると、同一大学のおおよそ人数の近い集団が受験した時の、PP と OL の信頼性および受験者分離、受験者階層には違いが見られないと言ってよい。PP セット、OL セットともに、同一大学の集団を統計的に識別可能な 3 ないし 4 のレベルに分けている。なおこれは、靜(2020）で確認した 5 つの大学データセットから得られた結果（表 25）とほぼ共通のものである。このことからしても、OL 版は従来の PP 版と同様の信頼性と識別性能があることが示唆されている。

表 25　レベルの異なる 5 集団の得点信頼性および受験者分離（靜, 2020 より）

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

Rasch モデリングの特長として受験者とテスト項目を同一次元上に配置して、それぞれの相対的な位置を視覚的に捉えることができることがある。それを表すマップは Wright map と呼ばれるが、B 大学の 2019 年度 PP と 2020 年度 OL の Wright map を図 1 に示す。

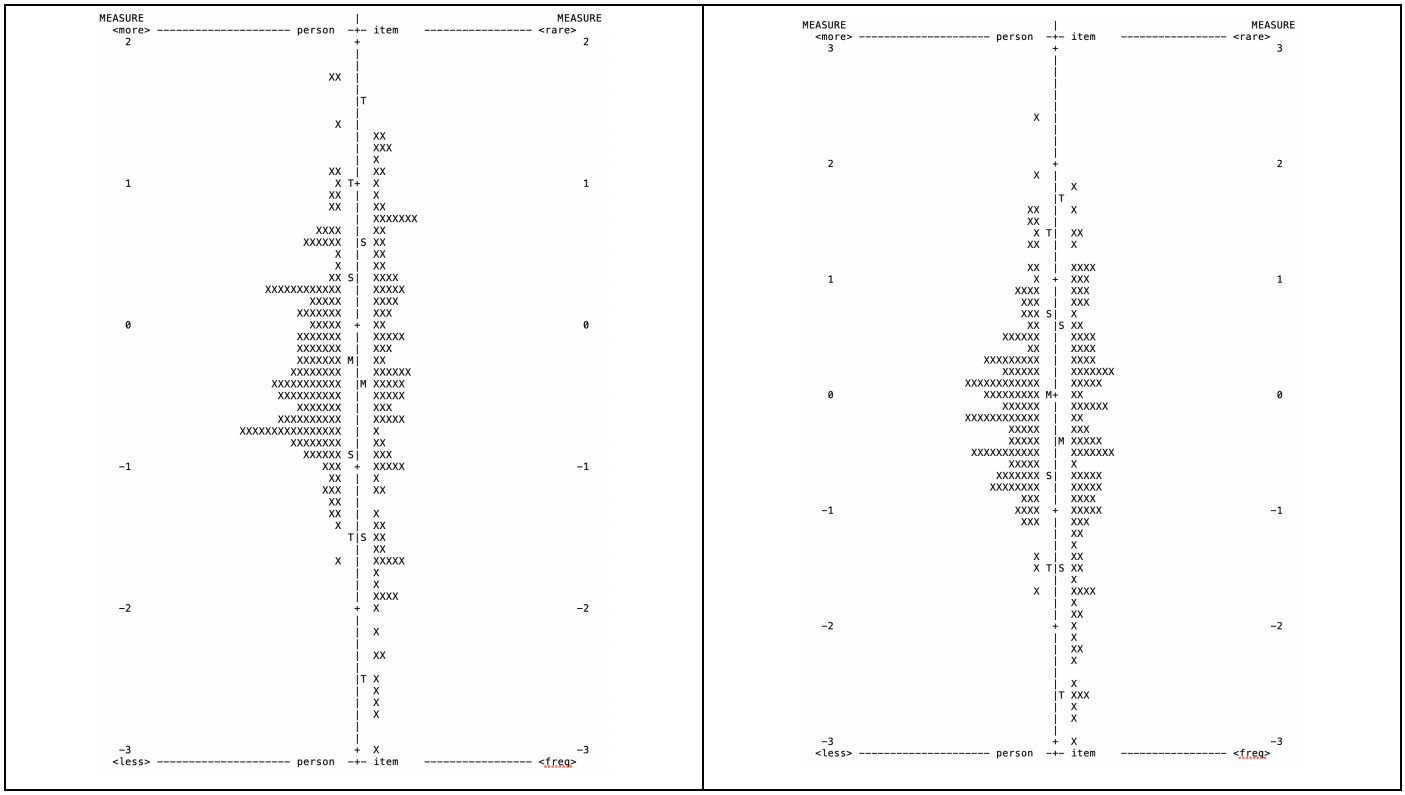


図 1　B 大学の 2019 年度 PP（左）と 2020 年度 OL（右）の Wright Map

左が PP、右が OL である。それぞれの真ん中に縦に貫いているのが VELC Test®が測定している英語力の次元軸であり、軸の左の X が個々の受験者を、右の X が個々項目を表す。軸は上に行くほど英語力／項目難度が高く、下にいくほど英語力／項目難度が低い。PP と OL はフォームが異なるので、個々の項目の位置は全く同一

ではないが、すべての項目の平均難度は左右のフォームで等化されている。いずれのフォームでも軸の下方に位置する難度の低い（易しい）項目から上方に位置する難度の高い（難しい）項目まであることが見て取れる。軸の左の受験者の分布を見ると、下方の英語力の低い層から上方の英語力の高い層まで幅広く分布しており、厳密ではないもののおおよそ中間層が厚い正規分布に近いかたちになっていることがわかる。

4.3 選択肢の選択頻度の分析

VELC Test®の項目はすべて 4 肢選択である。同じ項目であれば、PP であっても OL であっても A, B, C, D の 4 つの選択肢が選ばれるパターンは基本的には違いがないはずである。たとえばPP において選択肢 A が 40%、選択肢 B、C、D がそれぞれ 30%、20%、10%の受験者に選ばれたならば、OL であっても同じような割合でそれぞれの選択肢が選ばれるはずである。

今回 A 大学と B 大学が受けたすべてのフォームの中から PP と OL で共通だった項目の中から 30 を選び、そのすべてについて、4 つの選択肢が選ばれた頻度についてカイ二乗検定を行ってみた。例えば、ある R1 の項目の PP 版と OL 版の頻度カウントは表 26 のようであった。

表 26　ある項目の PP 版と OL 版における選択肢の選択頻度

	選択肢 A	選択肢 B	選択肢 C	選択肢 D
PP 版	114	10	21	18
OL 版	79	14	26	21

この頻度の分布が PP 版と OL 版の間で有意に異なるかを検定したということである。この項目については、Pearson's chi-squared = 6.0656, $df = 3$, $p = .108$ で、有意差はない。このような検定を 30 項目のすべてについて行った結果、30 項目中、3 分の 2 にあたる 20 項目では有意差がなかった。残りの 10 項目では有意差があったが、Cramer's Vによる効果量を確認してみると、0.16〜0.23 の範囲であり、すべて「効果量小」である。

たとえば、$p = 0.027$ で有意であった R1 項目は、$V = 0.17$ であり、実際の選択状況は表 27 のようである。誤答選択肢 B、C、D の頻度パターンこそ多少異なるものの、全体としては概ね似通っていると考えて問題ないことがわかる。

表 27　有意差が見られたある項目の PP 版と OL 版における選択肢の選択頻度

	選択肢 A	選択肢 B	選択肢 C	選択肢 D
PP 版	133	11	9	11
OL 版	129	5	1	5

4.4 同一フォームでの、PP と OL の項目正答率の比較

C 大学の 2020OL に使用したフォーム（n =1,679)は、A 大学の 2019.7 月 PP に使用したフォーム（n = 188）と同一である。そこでそれぞれのデータにおける項目難易度を比較することとした。VELC Test®では複数回のトライアルによって定められた値に各項目の難易度を固定（アンカー）した上で、それとの比較において受験者のベルクスコアを算出している。そこでそのアンカー値にかかわらない項目難度をみるため、正答率（正解人数/総人数）を 2019PP と 2020OL で比べることとしたものである。

まず 120 項目全体の正答率を比較するため、プロットしたのが、図 2 である。$r = .825$ であり、強い正の相関が確認された。

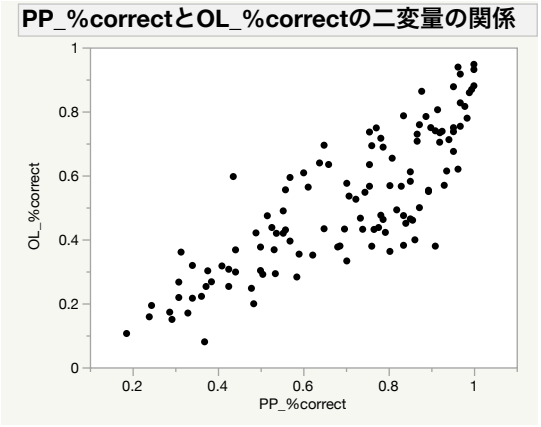
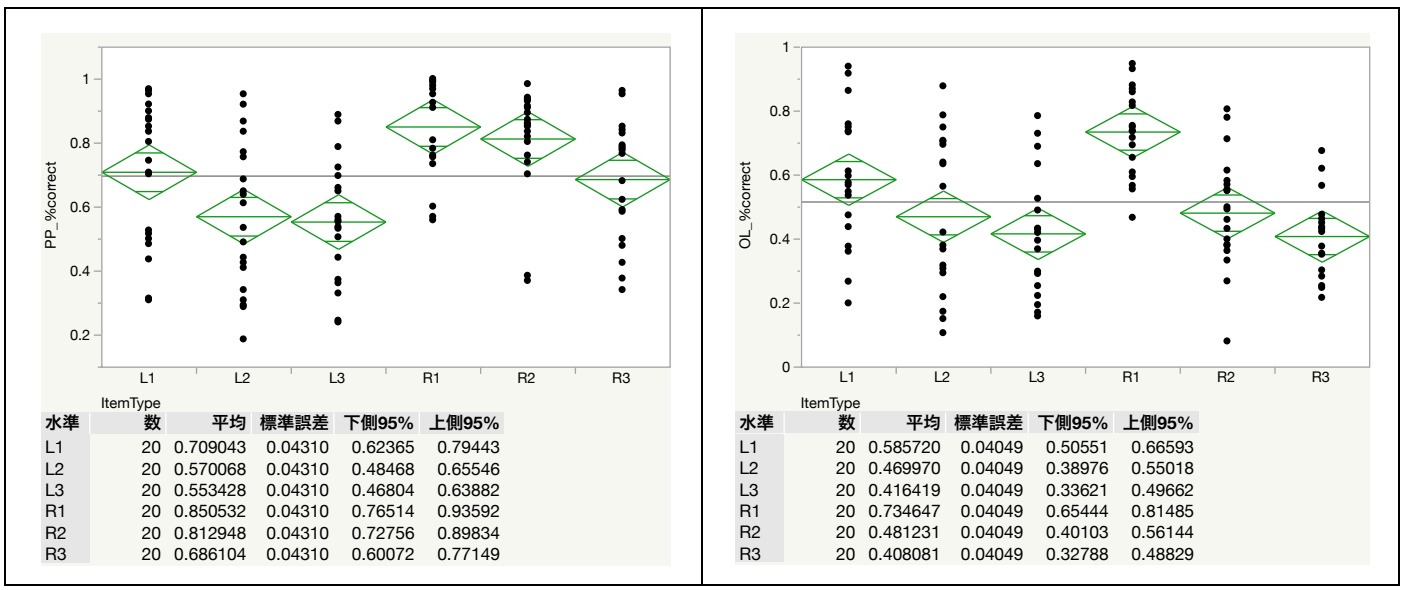


図 2　2019PP での正答率と 2020OL での正答率の散布図

次に、リスニングセクションの、L1, L2, L3,　リーディングセクションの R1, R2, R3 べつに正答率値をグルーピングしてグラフにしたのが図 3 である。左が PP で右が OL である。左右のグラフはまったく同一ではないものの、共通して見られるパターンがある。（1）全体としてリスニング項目の正答率のほうが、リーディング項目の正答率よりも低い。（2）リスニングの 3 セクションの間には、L1＞L2＞L3 という傾向が見られる。（3）リーディングの 3 セクションの間には、R1 ＞ R2 ＞ R3 という傾向が見られる。



水準	数	平均	標準誤差	下側95%	上側95%
L1	20	0.709043	0.04310	0.62365	0.79443
L2	20	0.570068	0.04310	0.48468	0.65546
L3	20	0.553428	0.04310	0.46804	0.63882
R1	20	0.850532	0.04310	0.76514	0.93592
R2	20	0.812948	0.04310	0.72756	0.89834
R3	20	0.686104	0.04310	0.60072	0.77149

水準	数	平均	標準誤差	下側95%	上側95%
L1	20	0.585720	0.04049	0.50551	0.66593
L2	20	0.469970	0.04049	0.38976	0.55018
L3	20	0.416419	0.04049	0.33621	0.49662
R1	20	0.734647	0.04049	0.65444	0.81485
R2	20	0.481231	0.04049	0.40103	0.56144
R3	20	0.408081	0.04049	0.32788	0.48829

図 3　PP（左）と OL（右）での項目タイプ別の相対的難易度のパターン

なお、正答率の絶対値をみるととくに R2, R3 について PP 版と OL 版にかなりの差が見られる。しかしこれは異なる集団（A 大学と C 大学）の英語力特にリーディング力の差の反映であると解釈される。絶対値は異なるものの、PP 版と OL 版で、項目タイプ別の相対的難易度のパターンが似通っているという点が重要であると考える。

4.5 項目のモデル適合度の確認

モデル適合度に関しては、C 大学で 2020 年度 OL に使用したフォームの 120 項目について確認した。項目の Infit Mean Square の分布状況を図 4 に示す。Infit Mean Square は Rasch 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)。ただし Linacre (2005)はより広く範囲をとり、0.5～1.5 の項目が productive of measurement であり、1.5～2.0 の項目は Neither constructs nor degrades

measurement (p. 197)としている。図 4 を見ると最大値が 1.85 なので、1.3 を超える項目を数えてみると、1.3～1.4 の項目が 5 項目、1.5～1.85 の項目が 6 項目であった。Linacre の基準によると測定に役だってもいないが害にもなっていないとされる項目が 120 項目中 6 項目あったことになるが、2.0 を超える項目は皆無であったので、この程度であればテスト全体としての結果には影響がないと判断する。

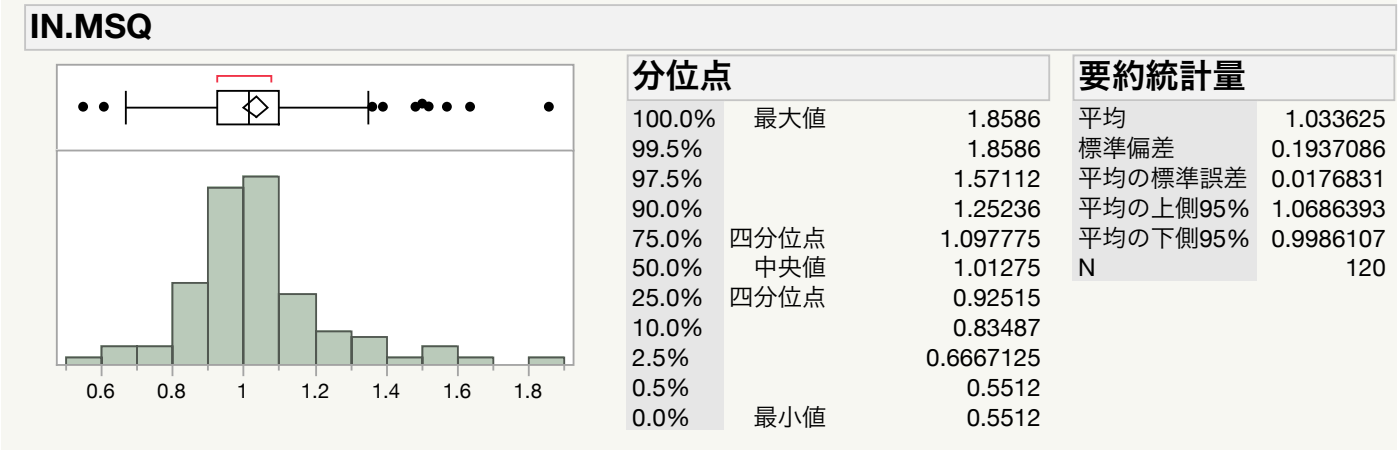


図 4　2020OL での項目適合度 Infit Mean Square の分布

5. 考察とまとめ

本研究は、A 大学、B 大学、C 大学で、2019 年度に PP 版を、2020 年度に OL 版を受験した、のべ 3,236 名のデータの全部あるいは一部を利用して、（1）学年別のスコアの分布および経年変化、（2）PP 版と OL 版全体としての受験者信頼性と受験者分離度、（3）PP 版と OL 版の同一項目間の選択肢の選択頻度、（4）同一フォームの PP 版と OL 版の項目正答率、（5）OL 版データからみる項目のモデル適合度、の分析を行った。

（1）の分析からは、PP 版と OL 版によるスコアには違いがあることを示す傾向は見られなかった。むしろ大学ごとに、これまで PP 版によって明らかになってきた傾向と類似の傾向が、今回 OL 版を用いた場合にも検知された。また C 大学に関しては当該大学担当者からの情報なども総合すると、2019 年度までの学生層とは明らかに異なる英語力をもった学生層が 2020 年度に入学してきたことによって大きくスコアが伸びたのだろうという解釈ができた。またその英語力の異なりは語彙セクションにおいて最も大きいという示唆を、OL の結果が示したと言える。これは語彙力がすべての基礎となることを考えると、納得できる結果である。

（2）の分析からはやはり PP 版と OL 版による違いは発見されなかった。2020 年度の OL 版は 2019 年度の PP 版と同等かそれ以上に、個別大学の受験者集団を統計的に区別可能な異なるレベルに弁別していることが示された。そしてその弁別の程度は、靜（2020）で報告した従来の PP 版が同程度の人数の個別大学の学生集団を弁別してきた程度とよく似たものであった。受験者能力と項目難度の相対的な分布を視覚的に表示する Wright Map からは VELC Test®は項目難度の幅が広いテストであり、受験者集団の能力が正規分布に近く幅広く分布していることが確認された。

（3）の分析からもやはり PP 版と OL 版の間の、意味のある違いは発見されなかった。選択肢の頻度に統計的な有意差が見られた場合にも、効果量の点ではいずれも「小」であり、実際上の意味のある違いではないとみなすのが適当だと解釈される。

（4）の分析からは PP 版と OL 版では、正答率で表した場合の項目の難度には強い正の相関があり、かつ項目のタイプ別の正答率のパターンが共通であることが確認された。

（5）の分析からは、OL 版のデータは Rasch モデルに概ね適合していることが確認できた。

以上に総合すると、PP 版と OL 版の等化性が担保されていない、と懸念する材料は、実施データからもみつ

からなかったといえる。そもそも PP 版をそのまま PC での実施が可能なように移植したのが OL 版である。紙がスクリーンに変わっただけでテスト特性、項目特性が変化することは考えにくい。その想定の妥当性を裏付けることができたと言えるだろう

ただテスト自体として OL 版は PP 版と等価であるとしても、一般論として、その場に試験監督がいない OL テストという受験環境が PP 版と比べた場合にセキュリティ面において潜在的な課題があることは否めない。OL 環境では密かに手元の辞書を使用するなどの不正な受験方法を完全には排除できない。スマートフォンなどを使用して他者と通信することも不可能ではない。またそもそもいわゆる「替え玉受験」さえも可能ではある。これらの不正行為の防止のために、大学によっては Zoom などを用いて教員がリアルタイムで「監視」している状態で受験させるところもあるようだ。それもひとつの方法であろう。

しかしながら学年内での「プレイスメント」や「授業効果の測定」といった、極めてハイステークスであるとまでは言えない受験用途に限って考えるならば、果たして受験者の中に懲戒処分を受けるリスクを冒してまでスコアを上げようと試みる者が現実にどれだけいるかは疑問である。また VELC Test® Online に関しては辞書使用などの不正行為が可能なのはリーディングセクションのみであってリスニングセクションでは不可能である。またリーディングセクションであっても、問題テキストのコピーペーストが不可能な仕様になっている。このため単語の意味を調べようとすればスクリーンから視認したスペリングを自分で改めて別の辞書などに正確に打ち込む必要があり、このことも不正な辞書使用を抑止する効果があると思われる。実際、A 大学では PP 版と OL 版の間でリスニングセクションのスコアがかなり異なるがリーディングセクションにはほとんど差がないといったケースも観察された。これは今回のデータに関しては、異なる集団が受験した PP 版と OL 版のスコア上の違いが不正行為などの要因ではなく、実際の英語力の違いを反映したものだという解釈の正しさを強く示唆すると言える。

結論として今回の調査では、正当に受験がなされた場合には VELC Test の OL 版は従来の PP 版と等価であることが強く示唆された。今後は、性善説に立たずともテストセキュリティ面での懸念をさらに小さくするための方策を検討してゆきたい。

引用文献

靜哲人(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.

靜哲人(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.

靜哲人(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.

靜哲人(2014)「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」大学英語教育学会第 53 回国際大会(2014.8.28). 横浜市立大学.

靜哲人(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.

靜哲人(2015b)「VELC Test®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.

靜哲人（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.

靜哲人(2020)「VELC Test® 2012-19 年度実施データの分析および総括」『語学教育研究論叢』第 37 号, 75-89.

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.

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」 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.

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/

Shizuka, T. (2016) Modification of VELC Test® listening section part 2 type multiple-choice 1-blank partial "dictation" items: Effects on distractor discriminations and TOEIC®-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.