VELC Test® 短縮版の信頼性および基準関連妥当性の検証 [1)]

― 項目数の漸減はテスト特性にどの程度影響を与えるか？―

靜　哲人

Reliability and Criterion-related Validity of the 90-item Version of the VELC Test®:

How much does gradual reduction of the number of items affect the test characteristics ?

SHIZUKA Tetsuhito

Abstract

The 90-item versions of the VELC® Test were recently developed based on their original 120-item versions by dropping 5 items from each of the 6 parts in such a way that the mean item difficulty of the whole test would remain unchanged. This study explored the reliability and the criterion-related validity of the newly launched 90-item versions, as well as hypothetical versions with still fewer items. Simulated 90-item data sets and those with gradually smaller number of items (from 84 to 6) were created based on actual 120-item VELC Test® response data sets produced by 130 students. Their TOEIC® L&R scores served as the criterion. Regarding the 90-item versions, it was found that they (1) correlated very strongly with the corresponding 120-item versions (r = .98 or higher), (2) exhibited very high reliability (alpha = .91 or higher), and (3) correlated with TOEIC® L&R practically as strongly as the full 120-item versions did. Regarding versions with smaller numbers of items, 48-item versions exhibited barely satisfactory measurement precision.

キーワード：VELC Test、短縮版、項目数、信頼性、妥当性

[1)] 第 6 回 VELC 研究会（2021.9.3）にて行った同タイトルの口頭発表に基づいている。

1　VELC Test とは

VELC Test® とは英語能力測定・評価研究会［VELC 研究会］によって開発され、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（靜・吉成, 2012）。リスニングとリーディングの 120 項目によって熟達度を測定するテストであり、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た（靜, 2012a; 2012b; 2013; 2014; 2015a; 2015b; 2017; Shizuka, 2016; 靜・ 望 月, 2014; Kumazawa et al. 2016）。実施開始から 8 年間のデータについては靜（2020a）が総括的に分析し、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改めて確認している。2020 年には新型コロナウイルス感染症拡大によってテストの対面実施が困難となったため、オンライン版（以下 OL 版）の運用がスタートした。OL 版と従来のペーパー版（以下 PP 版）との等価性については確認されている（靜, 2020b; 靜・望月, 2020）。

2　90 問版

OL 版と同時に運用開始されたのが 90 問の短縮版である。すなわち 2021 年現在、

Listening Part 1			Listening Part 2			Listening Part 3			Reading Part 1			Reading Part 2			Readig Part 3		
Item L1-1	+	+	Item L2-1	+	+	Item L3-1	+	+	Item R1-1	+	+	Item R2-1	+	+	Item R3-1	+	+
Item L1-2	+	+	Item L2-2	+		Item L3-2	+	+	Item R1-2	+		Item R2-2	+	+	Item R3-2	+	
Item L1-3	+		Item L2-3	+	+	Item L3-3	+		Item R1-3	+	+	Item R2-3	+		Item R3-3	+	+
Item L1-4	+	+	Item L2-4	+	+	Item L3-4	+	+	Item R1-4	+	+	Item R2-4	+	+	Item R3-4	+	+
Item L1-5	+	+	Item L2-5	+	+	Item L3-5	+	+	Item R1-5	+	+	Item R2-5	+	+	Item R3-5	+	+
Item L1-6	+	+	Item L2-6	+		Item L3-6	+		Item R1-6	+		Item R2-6	+	+	Item R3-6	+	
Item L1-7	+		Item L2-7	+	+	Item L3-7	+	+	Item R1-7	+	+	Item R2-7	+		Item R3-7	+	+
Item L1-8	+	+	Item L2-8	+	+	Item L3-8	+	+	Item R1-8	+	+	Item R2-8	+	+	Item R3-8	+	+
Item L1-9	+	+	Item L2-9	+	+	Item L3-9	+	+	Item R1-9	+		Item R2-9	+	+	Item R3-9	+	+
Item L1-10	+	+	Item L2-10	+		Item L3-10	+		Item R1-10	+	+	Item R2-10	+	+	Item R3-10	+	
Item L1-11	+		Item L2-11	+	+	Item L3-11	+	+	Item R1-11	+	+	Item R2-11	+		Item R3-11	+	+
Item L1-12	+	+	Item L2-12	+	+	Item L3-12	+	+	Item R1-12	+	+	Item R2-12	+	+	Item R3-12	+	+
Item L1-13	+	+	Item L2-13	+	+	Item L3-13	+		Item R1-13	+		Item R2-13	+	+	Item R3-13	+	
Item L1-14	+		Item L2-14	+	+	Item L3-14	+	+	Item R1-14	+	+	Item R2-14	+	+	Item R3-14	+	+
Item L1-15	+	+	Item L2-15	+		Item L3-15	+	+	Item R1-15	+	+	Item R2-15	+		Item R3-15	+	+
Item L1-16	+	+	Item L2-16	+	+	Item L3-16	+	+	Item R1-16	+	+	Item R2-16	+	+	Item R3-16	+	+
Item L1-17	+		Item L2-17	+	+	Item L3-17	+		Item R1-17	+	+	Item R2-17	+	+	Item R3-17	+	
Item L1-18	+	+	Item L2-18	+	+	Item L3-18	+	+	Item R1-18	+		Item R2-18	+	+	Item R3-18	+	+
Item L1-19	+	+	Item L2-19	+		Item L3-19	+	+	Item R1-19	+	+	Item R2-19	+		Item R3-19	+	+
Item L1-20	+	+	Item L2-20	+	+	Item L3-20	+	+	Item R1-20	+	+	Item R2-20	+	+	Item R3-20	+	+

図 1　120 問版と 90 問版の使用項目の比較イメージ図

各パート内のアイテムは 1 ～ 20 まで項目難度で昇順にソートされている。120 問版は “+” を付した全項目を使用し、90 問版は “+” のない網掛け部分の項目を使用しない。

VELC Test® には、120 問 PP 版、120 問 OL 版、90 問 OL 版が存在する。

90 問版は 120 問版をもとにしながら、6 つのパート各 20 問から難易度が異なる各５問を削除することにより作成した。図１にイメージを示す。この方法により、90 問版と 120 問版のパート毎および全体としてのテスト難易度はほぼ等しくなるよう調整されている。

理論上、90 問版は 120 問版よりも項目数が 30 問少ないぶんだけ測定誤差が大きくなる。具体的に VELC スコアにどの程度の測定誤差があるかはそのスコアレベルによって異なるが、受験者が多い VELC スコア 400 ～ 600 程度の能力帯では、120 問版で 18 ～ 21 程度、90 問版では 21 ～ 24 程度が測定標準誤差であると想定されている（表１）。

表１　能力帯別の 120 問版および 90 問版の VELC スコアの測定標準誤差

VELC スコア	200	300	400	500	600	700	800
120 問版の誤差	30	23	18	18	21	30	46
90 問版の誤差	37	26	22	21	24	32	53

このような根拠により、90 問版には 120 問版とまったく同じ精度はないものの、それでも主たる用途である学内プレイスメントや授業効果測定などの目的にとっては十分な精度を保持するはずであると考えられる。ただし、これはあくまで理論的な想定であり、実際の受験者データによって得点の信頼性などを検証したものではない。

そこで本研究ではこの想定が実際のテストデータによってどの程度裏付けられるかを確認することとした。また今後のための参考として、90 問からさらに項目数をどの程度減らすと信頼性や妥当性がどの程度影響を受けるかも調査することとした。

２　目的

VELC Test®90 問短縮版が、フルバージョンである 120 問版に比べて相対的にどの程度の信頼性と基準関連妥当性を有しているかを検証する。また今後の参考として、90 問よりさらに項目数を減らした場合に、信頼性と基準関連妥当性がどの程度影響を受けるかも併せて検証する。

３　方法

３.１　分析方法

基準関連妥当性を調べるためには、同一学習者が VELC Test® 受験と同時（あるいは比較的近い時期に）に基準となるもうひとつのテスト（TOEIC® など）を受験している必要がある。しか本研究の実施時点において、90 問版 VELC Test® の受験者でそのような基準テストデータが利用可能なケースは存在しなかった。そこで次善の策として、

TOEIC® L&R スコアが判明している受験者が 120 問版 VELC Test® を受験した解答データに基づき、90 問版をうけた場合の解答状況を模擬的に作り出し、この 120 問版データと模擬 90 問版模擬データを比較することとした。

比較の指標は、（1）信頼性（クロンバック α）と（2）基準関連妥当性（TOEIC® L&R との相関）である。また、90 問版よりもさらに項目数が少ないテストデータを模擬的に作り出し、（2）と（3）の変化を調べる。

VELC Test® の結果としてフィードバックされる「VELC スコア」は単純な素点ではなく、Rasch モデリングによる logits 値を一般ユーザーによりわかりやすい整数に変換したものである。しかし本研究では Rasch モデリングによる信頼性係数や受験者分離などの指標を用いず、あえて素点によるクロンバック α を用いた。これは今回模擬的に作り出す 120 通り（4.2 参照）のデータセットに対応する 120 の control files を作成して、120 回 Winsteps（Rasch モデリングのプログラム）を走らせるのは、コストパフォーマンスの面から賢明ではないと考えたからである。一般に VELC Test® の素点と VELC スコアの間には、$r = .97$ 程度の非常に強い相関関係があるため、今回の調査を素点ベースで行うことには問題がないと判断した。

3.2　データセット

データ提供者は、本研究の目的について説明を受けた上でスコアの提供に同意した首都圏の大学生（$n = 135$）である。そのうち 112 名は、2019 年末から 2020 年春にかけて VELC Test® 120 問 PP 版を受験してかつ TOEIC® L&R のスコアを提供してくれた学生、23 名は 2021 年 1 月～ 2 月に VELC Test® 120 問 OL 版を受験してかつ TOEIC® L & R スコアを開示してくれた学生である。詳細を表 2 に示す。

前述したように OL 版については PP 版との等価性が確認されている（靜, 2020b; 靜・望月, 2020）ため、今回の分析では PP 版と区別せず扱った。そこで、フォーム別に整理し、フォーム 2051（$n = 14$）の解答データを「セット 2051」、フォーム 2052（$n = 76$）の解答データを「セット 2052」、フォーム 2055（$n = 45$）の解答データを「セット 2055」と呼ぶこととする。

表2　データセット別 VELC Test 実施時期、フォーム、モード、受験者属性と人数

	実施時期	フォーム	モード	受験者所属大学	*n*
セット 2051	2019 年 12 月	2051	PP	A 大学	14
セット 2052	2020 年 1 月	2052	PP	B 大学	53
	2021 年 1 ～ 2 月		OL	B 大学	23
セット 2055	2020 年 3 月	2055	PP	B 大学（14）、C 大学（26）、D 大学（5）	45

4　分析および結果

4.1　90 問版と 120 問版の比較

まず模擬 90 問版と 120 問版の比較について報告する。この模擬 90 問版はそれぞれの当該フォームに関して実際に短縮版として使用されているのと同一の項目を含むものであるので、模擬とはいうもののデータとしては実際の 90 問版と変わらないと考えられる。よって以下では単に 90 問版とする。

4.1.1　記述統計

まずセット毎に 90 問版と 120 問版の記述統計を表 3 に示す。平均値としてはセット 2052 と 2055 がおおよそ同じで、セット 2051 がそれよりも 90 問版で 15 点、120 問版で 20 点ほど低い。一方、標準偏差に関してはセット 2051 と 2052 には大きな差がなく、セット 2055 のみ明らかに大きい。標準偏差の傾向についてはセット 2051 とセット 2052 はそれぞれ単一の大学に所属する受験者によるデータであるのに対し、セット 2055 は 3 つの異なる大学に所属する学生のデータを合わせたものであることの反映だと考えられる。

表3　90 問版と 120 問版の記述統計

		90 問版		120 問版	
	n	平均	標準偏差	平均	標準偏差
セット 2051	14	50.14	11.33	66.64	14.40
セット 2052	76	65.82	10.70	85.87	14.27
セット 2055	45	64.51	14.39	87.04	18.47

4.1.2　相関

3 つのセットそれぞれに関しての 90 問版と 120 問版の相関を表 4 に、ヒストグラムおよびプロットを図 2 ～図 4 に示す。すべてのセットにおいて相関係数は $r = .98$ を上回っており、極めて高い。すなわち、プレイスメント目的であれば 90 問版は 120 問版の代替として十分に機能すると言えよう。

表4　90 問版と 120 問版の相関係数および上下 95% 区間

	n	相関係数	下側 95%	上側 95%
セット 2051	14	.987903	.961089	.996274
セット 2052	76	.983780	.974458	.989718
セット 2055	45	.989232	.980372	.994105

4.1.3　信頼性

次に信頼性の指標として、それぞれのセットにおける 90 問版と 120 問版のクロンバック α を表 5 に示す。まず 120 問版の数値を見ると、セット 2051 は、サンプルが $n = 14$ と非常に少ないにもかかわらずほとんど .90 に近い値である。セット 2055 は .9492 と最も高い。誤解を避けるため述べておくと、これはテストフォームとして 2055 が他のフォームよりも「信頼性」の高い得点を得やすい、ということではない。クロンバック α はその算出式からも明らかなように、テスト項目の性能とともに受験者能力の散

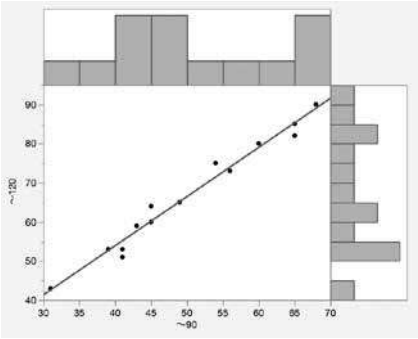

図2　90 問版と 120 問版のヒストグラムおよび散布図（セット 2051）

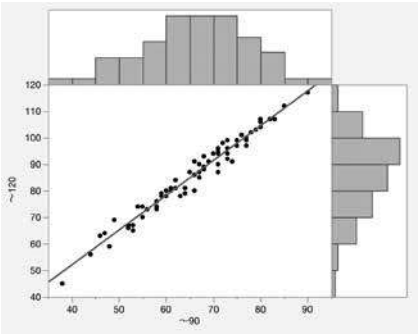

図3　90 問版と 120 問版のヒストグラムおよび散布図（セット 2052）

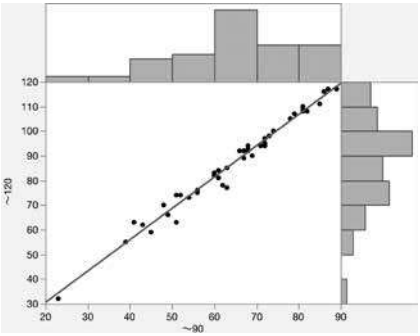

図4　90 問版と 120 問版のヒストグラムおよび散布図（セット 2055）

らばりの関数でもある。セット 2055 の α の値の高さは、異なる 3 つの大学に所属する受験者の標準偏差が他のセットよりも明らかに大きかったことが原因であると考えられる。90 問版の値を見ると最も低いセット 2051 であっても .8798 と十二分に高い。最も高いセット 2055 では .9359 である。セット毎に 90 問版と 120 問版の値の差を比較してみると、0.01 ～ 0.02 程度と小さい。

表５　990 問版と 120 問版の信頼性係数（クロンバック α）の比較

セット	*n*	90 問での α	120 問での α
セット 2051	14	.8798	.8958
セット 2052	76	.8870	.9104
セット 2055	45	.9359	.9492
3 セット合体	135	.9181	.9333

4.1.4　基準関連妥当性

次に TOEIC® L&R のトータルスコアを基準としたときの、基準関連妥当性を吟味するが、数値を示す前に指摘しておきたいことがある。それは、VELC Test® は設計思想からして、TOEIC® L&R を模そうという意図がそもそもないということである。その意味で TOEIC® L&R との相関の高低は、VELC Test® の日本語母語大学生の英語熟達度テストとしての妥当性の高低を必ずしも意味しない。VELC Test® はそれ自体として個別受験者が日本の大学生全体を基準としてスキル別にどのような位置にいるのかという集団基準準拠的情報と、スキル別にどの程度のタスクがどの程度の正確さでできるのかという目標基準準拠的情報の両方をもたらすことのできる測定ツールである、と開発者として考えている。

そのことを確認した上で、TOEIC® L&R との相関を表 6 に示す。3 つのセットは素点ベースでは厳密には等化されているとは言えないが、事実上はおおむね等しいとみなせると思われるため、3 セットを合わせた場合の相関係数も求めてみた。

セット別に比較すると、やはり α の高低と同じように、セット 2051 が一番低く、セット 2055 が一番高い。セット 2051 に関しては 90 問版のほうが 120 問版より値がわずか

表６　90 問版と 120 問版の TOEIC L&R との相関係数

セット	*n*	90 問版	120 問版
セット 2051	14	.7168	.7016
セット 2052	76	.7361	.7592
セット 2055	45	.8283	.8394
3 セット合体	135	.7962	.8080

ながら高いが、これはサンプルサイズの小ささに起因するノイズであると解釈するのが妥当だと思われる。セット 2055 の値の高さは、やはりこのセットを受けた受験者グループの能力幅（= 標準偏差）の広さの表れであると考えられる。３セット合体した場合の相関係数は 90 問版で r = .7962、120 問版で r = .8080 である。

なお、再度指摘しておくと、この相関係数は素点との数値である。Rasch モデリングを経た VELC スコア、さらに６つのパート別 VELC スコアのすべてを使用した回帰式で求める予測点（TOEIC 相当点）を用いたときの数値は、より高い。具体的には３セット合体した場合、120 問版の VELC スコアトータルと TOEIC® L&R の相関は r = .8373、予測点（TOEIC 相当点）との相関は r = .8437 である（表７）。

表７　３セット合体での素点、VELC スコア、予測点の TOEIC L&R との相関係数

	素点	VELC スコア Total	予測点（相当点）
TOEIC L&R Total	.8080	.8373	.8437

すなわち、今回は簡易的に素点ベースで算出した r = .8080（120 問版）という数値に関しては、絶対値に着目するのではなく、同じく素点ベースで算出した r = .7962（90 問版）とそれほど値に差がない、という点にのみ着目するのが適当である。

以上、120 問版３フォームと、そこから模擬的に作り出した 90 問版３フォームを、信頼性係数と TOEIC® L&R を基準にした場合の基準関連妥当性を調査した。その結果、想定どおり、いずれのフォームにおいても 90 問版は 120 問版よりも信頼性・基準関連妥当性ともにやや減ずるものの、テスト目的に照らして十分に高い信頼性と基準関連妥当性を保持していることを確認した。

4.2　項目数をさらに減じた場合のシミュレーション

次に、今後さらなる短縮版を開発する際の基礎データを得るため、項目数がどの程度減ると信頼性と基準関連妥当性がどの程度影響を受けるのかをシミュレーション的に調査した。このために、以下の手順で項目数の異なる解答データを作り出した。

⑴　90 問版の６つのパート（各 15 問）から各１問ずつ計６問ずつを削除していった。どの項目を削除するかはエクセルの RAND 関数を利用してランダムに選んだ。こうして 84 問版（各パート 14 問）、78 問版（各パート 13 問）、72 問版（各パート 12 問）、…18 問版（各パート 3 問）、12 問版（各パート 2 問）、6 問版（各パート 1 問）の 14 種類の短縮版データを作成した。

⑵　⑴で RAND 関数を２回ずつ適用することで、14 × 2 = 28 種類の短縮版データを作成した。

⑶　2051、2052、2055 の３種類のフォームに（１）（２）の手法を適用し、3 × 28

＝ 84 種類の短縮版データを作成した。

⑷ 2051、2052、2055 をもとにした 2 種類ずつの短縮版データ群を合体した、2 × 14 = 28 種類の短縮版 3 セット合体データを作成し、（3）に加えた。こうしてすでに分析した 4 種類の 90 問版データセット 4 つに加えて、新たに 112 種類のデータセットの合計 116 通りの模擬的な短縮版データを作り出した（表 8）。

表 8　3 つのフォームをもとにした 116 種類の短縮版データセットの項目数

フォーム	項目数															
2051	6	12	18	24	30	36	42	48	54	60	66	72	78	84	90	120
	6	12	18	24	30	36	42	48	54	60	66	72	78	84		
2052	6	12	18	24	30	36	42	48	54	60	66	72	78	84	90	120
	6	12	18	24	30	36	42	48	54	60	66	72	78	84		
2055	6	12	18	24	30	36	42	48	54	60	66	72	78	84	90	120
	6	12	18	24	30	36	42	48	54	60	66	72	78	84		
3 セット合体	6	12	18	24	30	36	42	48	54	60	66	72	78	84	90	120
	6	12	18	24	30	36	42	48	54	60	66	72	78	84		

注：2051、2052、2053、3 セット合体の各上段と下段は、RAND 関数をそれぞれ適用して別の項目を削除したもの

4.2.1　120 問版との相関

まず、116 種類のデータセット素点の、当該 120 問版セット素点との相関係数を表 9 に示す。この数値はフルバージョンである 120 問版を基準テストとしたときの基準関連妥当性であるとも言える。見やすくするために .90 以上の数値に薄い網掛けを、.95 以上の数値に濃い網掛けを施している。3 つのフォームのうちどれかひとつを元にした 6 バージョンのうち、いずれの場合でも 42 問あれば数値は .90 以上となり、60 問あれば .95 以上となっている。3 セット合体した場合には、30 問あれば数値が .90 に達し、48 問

表 9　項目数の異なるデータセットの 120 問版との相関係数

フォーム	項目数															
	6	12	18	24	30	36	42	48	54	60	66	72	78	84	90	120
2051	.71	.69	.78	.89	.92	.92	.93	.92	.92	.95	.96	.97	.98	.98	.99	1.0
	.67	.87	.81	.83	.89	.87	.90	.92	.92	.95	.94	.95	.97	.98		
2052	.73	.84	.86	.91	.91	.92	.93	.95	.95	.96	.97	.97	.97	.98	.98	1.0
	.57	.70	.73	.83	.88	.91	.94	.95	.96	.96	.97	.97	.98	.98		
2055	.72	.78	.82	.88	.88	.92	.95	.97	.97	.98	.97	.98	.98	.99	.99	1.0
	.79	.85	.90	.92	.94	.94	.96	.96	.96	.97	.97	.98	.99	.99		
3 セット合体	.74	.81	.83	.90	.91	.92	.94	.95	.96	.97	.97	.98	.98	.98	.99	1.0
	.71	.77	.78	.85	.90	.91	.93	.95	.96	.97	.97	.97	.98	.98		

注：.90 以上に薄い網掛け、.95 以上に濃い網掛けを付している。

あれば .95 以上となっている。

相関の強さの変化を視覚的に確認するため、3 セット合体の上段のセットについて、項目数が異なる場合の散布図の変化を図５に示す。120 問版を基準とした時、48 問であってもかなりの程度類似した結果が得られると解釈できるだろう。

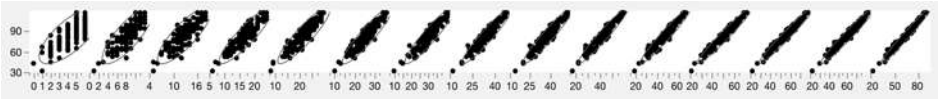

図５　項目数６～ 90 の短縮版の 120 問版との素点の散布図

縦軸が 120 問版の素点、横軸が短縮版の素点。一番左のプロットが項目数 6、左右から 8 番目（中央）のプロットが項目数 48、一番右のプロットが項目数 90。

次に 3 フォーム合体の上段と下段のセットに関して、相関係数の強さの変化を折れ線グラフで視覚化してみた（図６）。項目数が小さいうちはおそらくランダムな項目選定の影響のため２つの線に隔たりがあるが、42 問あるいは 48 問あたりから 2 つの線は事実上重なり、1.0 にかなり近くなっていることが視認できる。

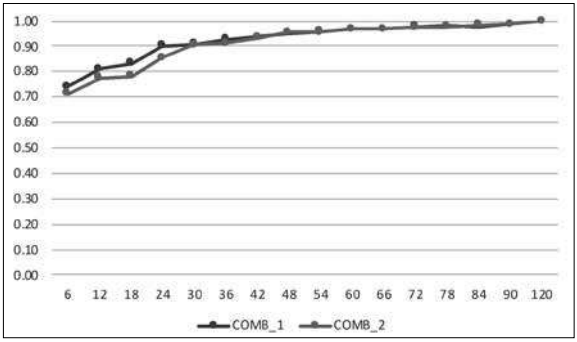

図６　項目数の漸増による 120 問版との相関の変化

COMB_1 は 3 セット合体の上段、COMB_2 は下段。

4.2.2　信頼性

116 種類の短縮版および４種類の 120 問版のデータセットのクロンバック α を表１０にまとめた。

視覚的に確認するため、3 セット合体の場合の値の変化の折れ線グラフを作成してみた（図７）。項目数が 60 に達したあたりから事実上２本の線が重なっていることが見て取れる。

一般に信頼性のひとつの基準とされる .70 以上に薄い網掛けを、.80 以上に濃い網掛

表 10　項目数の異なるデータセットの信頼性係数：クロンバック α

フォーム	項目数															
	6	12	18	24	30	36	42	48	54	60	66	72	78	84	90	120
2051	.38	.54	.58	.64	.69	.76	.80	.83	.84	.84	.86	.88	.88	.88	.88	.90
	.35	.54	.65	.58	.64	.65	.65	.73	.75	.79	.81	.83	.86	.87		
2052	.56	.59	.61	.69	.73	.75	.78	.81	.82	.84	.87	.87	.88	.88	.89	.91
	.22	.53	.53	.62	.71	.75	.77	.81	.84	.84	.85	.86	.87	.88		
2055	.34	.59	.64	.74	.78	.81	.85	.87	.88	.90	.91	.92	.92	.93	.94	.95
	.56	.66	.73	.81	.85	.88	.89	.89	.90	.91	.92	.93	.93	.93		
3 セット合体	.53	.55	.56	.69	.75	.76	.80	.82	.84	.86	.88	.89	.89	.90	.91	.93
	.53	.68	.74	.78	.79	.83	.85	.86	.87	.87	.88	.89	.90	.90		

注：.70 以上に薄い網掛け、.80 以上に濃い網掛け、.90 以上に非常に濃い網掛けを付している。

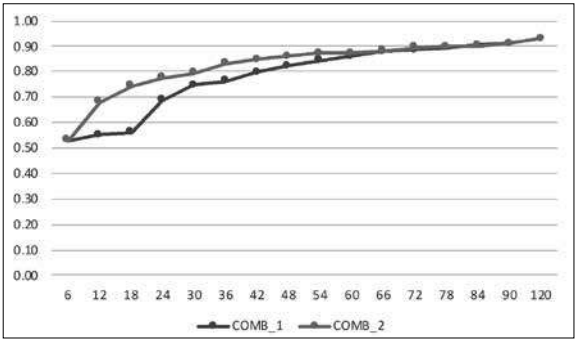

図 7　項目数の漸増によるクロンバック α の変化

COMB_1 は 3 セット合体の上段、COMB_2 は下段。

けを、.90 以上にさらに濃い網掛けを施している。48 項目あれば、n = 14 と受験者人数の少なかったフォーム 2051 でも .70 以上の数値が得られている。3 セット合体した場合には、30 項目あれば .70 以上、48 項目あれば .80 以上が得られている。やはり項目数 48 が、ひとつの最低基準と言えるかもしれない。

4.2.3　基準関連妥当性

最後に、TOEIC® L&R を基準とした場合の相関係数を求めてみた（表１１）。サンプル数が n = 14 であったフォーム 2051 は、上段では 42 問でいったん .70 に達した数値が 66 問でふたたび .70 を下回るなどノイズと解釈できるパターンが見られるので、とりあえず除外して考えることとする。他の 2 つのフォームでは 54 問あれば最低でも .70 が得られている。

相関の強さの変化を視覚的に確認するために、3 セット合体の上段について散布図を

表 11　項目数の異なるデータセットの基準関連妥当性：TOEIC L&R との相関係数

フォーム	項目数															
	6	12	18	24	30	36	42	48	54	60	66	72	78	84	90	120
2051	.58	.57	.57	.66	.63	.65	.70	.73	.70	.70	.67	.66	.68	.70	.72	.70
	.50	.64	.63	.67	.69	.59	.64	.64	.64	.65	.66	.63	.67	.70		
2052	.49	.61	.64	.67	.63	.68	.68	.71	.70	.73	.72	.75	.74	.73	.74	.76
	.34	.44	.42	.53	.62	.64	.67	.69	.72	.72	.72	.71	.72	.74		
2055	.54	.59	.68	.74	.75	.77	.79	.81	.81	.82	.81	.81	.81	.82	.83	.84
	.64	.68	.71	.74	.76	.78	.80	.81	.81	.82	.84	.84	.83	.83		
3セット合体	.54	.61	.65	.71	.71	.73	.74	.76	.76	.77	.76	.78	.78	.79	.80	.81
	.56	.60	.60	.67	.72	.74	.76	.77	.78	.79	.79	.79	.79	.80		

注：.70 以上に薄い網掛け、.80 以上に濃い網掛けを付している。

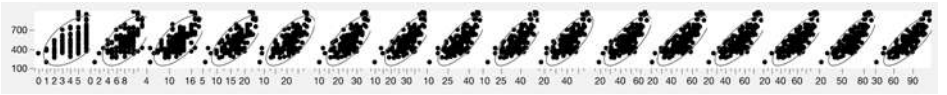

図8　TOEIC L&R スコアと 6 問版～ 120 問版 VELC Test 素点のプロット

縦軸が TOEIC スコア、横軸が VELC 素点。一番左のプロットが項目数 6、一番右が項目数 120。

作成したものを図 8 に示す。3 セット合体では 48 項目の時点で .75 に達している。項目数 48 のプロットは左から 8 つ目のものである。かなり強い相関があることが視認できる。

相関係数の変化を折れ線グラフで可視化してみた（図 9）。やはり項目数が 48 を超えるあたりから、90 問版が到達する上限にかなり近い域に達していることが見て取れる。

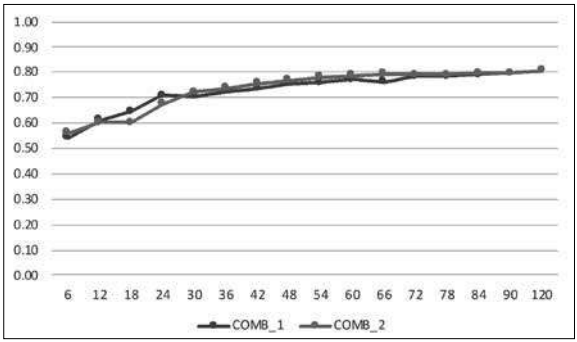

図9　項目数の漸増による TOEIC L&R との相関係数の変化

COMB_1 は 3 セット合体の上段、COMB_2 は下段。

5　考察および結論

本研究は VELC Test® の３種類のフォームの実際の受験データをもとにして模擬的に作り出した短縮版データについて、90 問版の信頼性と妥当性を検証し、かつ 90 問よりさらに項目数を減らした場合の信頼性と妥当性への影響を調べた。その結果、運用を開始している 90 問版は（１）120 問版との素点の相関係数が３つのフォームとも r = .98 以上と極めて高く、（２）信頼性係数も α = .91 以上であり、（３）TOEIC® L&R との相関の強さも 120 問版とほとんど変わらなかった。TOEIC® L&R との相関を今回は素点ベースで計算しているが、実際の短縮版の場合はパート別の VELC スコアをもとにした回帰式によって算出されるため、本研究での数値よりも高いものになることは再度確認しておきたい。すなわち VELC Test® の 90 問版は 120 問版と事実上ほとんど変わらない精度を持って、日本語を母語とする大学生に関するプレイスメントや授業効果測定が行えるツールであることが確認できたと言えよう。

さらに項目数を漸減した場合のシミュレーションの結果としては、テスト全体で 48 項目（６つのパートに各 8 項目）あれば、ある程度の精度（120 問版との相関は .95 以上、クロンバック α はおおよそ .80 以上、TOEIC L&R との素点による相関 .75 以上）を持って測定が行えそうである、という見通しが得られた。極めてハイステイクスであるとまでは言えない一般的な英語熟達度テストの目的に照らした時、信頼性係数 .80 というのは満足できる水準であると考えられる。本研究により、今後さらにコストパフォーマンスのよい英語力測定を実施してゆくための貴重な示唆が得られたと言えるだろう。

引用文献

静哲人（2012a）「VELC Test® 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）「VELC Test® 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 Test® 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® 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）「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.

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.

靜哲人(2020a)「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（2020b）「VELC Test® Online と VELC Test® P&P の等価性を検証する（その 1)」言語教育 EXPO2021（2020.10.25）Zoom 上にて開催.

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・望月正道（2020）「VELC Test® Online と VELC Test® P&P の等価性を検証する（その２)」日本言語テスト学会（2020.12.12）Zoom 上にて開催.

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.