論　文

VELC Test® Online は 90 問版と 120 問版のどちらを選ぶべきか
～プラス 15 分／30 項目がプレイスメントに与える影響～

靜　哲人（大東文化大学外国語学部）

Choice between the full and shortened versions of VELC Test® Online: The effects of the 15-min/30-item difference on placement results

Tetsuhito SHIZUKA

要旨

大学生のための英語力熟達度テストである VELC Test® Online の 120 問版と 90 問の短縮版を比較し、実際のプレイスメントにどの程度の違いが生じるか、また実際に運用されている 90 問版からさらに項目数を減じて超短縮版を作った場合、信頼性の点でどの程度の項目数までが許容できるかを検証した。分析対象は、A 大学で 120 問版を受験した 2,241 名の実際の応答データと、そこから模擬的に作り出した短縮版および超短縮版の応答データである。まず 90 問の短縮版は信頼性係数が .854 と高く、120 問版との相関係数が .983 と高かった。またレベル分けでは約 87％の学生が 120 問版と同一レベルに割り振られた。次に 84 問から 6 問ずつ項目数を減じていったところ、60 問版で信頼性係数が .80 ほどとなり、120 問版との同一レベルに割り振られる学生の率が約 76％となったため、60 問版を超短縮版の許容できる項目数の下限と解釈した。以上の結果から、運用開始されている 90 問版は大学内でのレベル分けという主要な用途に対して十分な信頼性があり、その信頼性は項目数を 60 問程度に減じるまでは担保されるであろう、という見通しが得られた。

1. はじめに

1.1 VELC Test とは

VELC Test® とは、筆者も所属する英語能力測定・評価研究会によって開発され、2012 年度から運用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）。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、Raschモデリング（Bond & Fox, 2007）を用いて等化された複数フォームがあり、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。

運用開始から8年間にわたり、その信頼性、妥当性、項目特性などについての検証がなされてきた（靜, 2012a; 2012b; 2013; 2014; 2015a; 2015b; 2017; 2020a; Shizuka, 2016; 靜・望月, 2014; 靜・吉成, 2012; Kumazawa *et al.* 2016）が、その結果、もともとの形、すなわち紙ベースで120問を実施するVELC Test®（以下、PP120とする）が日本の大学生の英語力測定のために有効に機能していることは十分に確認されていたと言ってよい。

1.2 オンライン版

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PP120の実施の多くが取り止めとなった。そこでその状況下でも実施できるVELC Test® Onlineが急遽開発され、2020年7月から運用を開始した。

VELC Test® Online（以下、OL120とする）は、PP120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。すなわちPP120とOL120の項目セットは同一、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。PP120とOL120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PP120では試験教室のスピーカーから一斉に聞こえるがOL120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、テストセキュリティの面に関しては、試験監督がいない状況で受験することが多いOL120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、PP120とOL120は等価でなくなる要因は思い当たらない。ただしその想定がデータで裏付けられるかは検証する必要があった。

そこで靜（2022a）は、PP120とOL120が実際に等価とみなしうるかを実際の受験データで調べた。2019年度にPP120を受験して2020年度にOL120を受験した3つの大学の受験データを対象として、(1) 学年別およびパート別のスコア分布、(2) 信頼性および受験者分離、(3) 同一フォームでの選択肢選択状況、(4) 同一フォームでの項目正答率、の4つの観点から分析し、かつ (5) オンライン版データのRaschモデル適合度を確認したところ、120PPのテスト特性はそのまま120OLに引き継がれていると考えてよさそうだ、という結論を得た。（ただしこの研究はあくまで別々の集団がPP120とOL120を受けたデータを比較したものであるため、強いエビデンスに基づく結論とは言えないことに注意が必要である。）

1.3 短縮版

OL120が運用開始されたのとほぼ同時期に、従来の120問から項目数を30減じた90問のVELC

Test® Online（以下、OL90 とする）の運用も開始された。120 問版は所要時間 70 分であるのに対し、90 問版は 55 分である。120 問版でも大学の標準的授業時間である 90 分の枠には楽に収まるが、さらに時間を短縮したいというニーズに応えたものだ。（ただし 90 問版は自宅等で受験することが多いと思われるオンライン版なので、大学の授業時間の枠自体が無関係となる。）

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されたものである。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っている。しかし等価であると想定される 120PP と 120OL との関係とは異なり、120OL と 90OL のテスト特性は、理論上等価ではありえない。測定誤差、いいかえれば得点の信頼性は、他の条件を一定にしたとき、項目数の関数である。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必ず大きくなる。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り、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。120 問版よりも一定程度信頼性が下がることがわかっている 90 問版について重要なのは、それでもプレイスメントや授業効果測定という主たる目的に照らして十分な程度の信頼性また妥当性が確保されているかどうか、である。

靜（2022b）は 4 つの大学に通う 135 名の大学生の PP120 あるいは OL120 の受験データと、自己申告された TOEIC® L&R のスコアをもとに、模擬的に作り出した PP90 あるいは OL90 のデータの信頼性（クロンバックアルファ）および基準関連妥当性（TOEIC® L&R との相関）を調べた。模擬的に作り出したとは、参加学生の 120 問版に対する解答の中から当該フォームの 90 問版に実際に使用されている項目に対する解答のみを抜き出すことで、その参加学生がその 90 問版を受けたならば生み出したはずである 90 問版の解答データセットを作り出した、という意味である。結果としては、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 $\alpha = .91$ 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（相関係数の差が、ポイント 0.01 ～ 0.02 の範囲だった）。すなわち短縮版でも目的に照らして十分な性能があることが示唆された。

2. 本研究

2.1 目的

本研究は靜（2022b）を補う形で、OL90 のテスト特性をさらに別の観点から、またより大きなサンプルサイズのデータを用いて調べることを目的としたものである。靜（2022b）は、4 つの大学からの合計 135 名の学生のデータをもとに 120 問版と 90 問版を比較した。このときのデータは人数こそ 135 名と多くはなかったが、4 つの大学の受験生を合体していた。このために英語力の幅が比較的大きくなり、そのことがかなり高い信頼性係数につながった（クロンバックアルファは、データの標準偏差の関数である）と考えられる。しかし VELC Test® の主たる用途は、ひとつの大

学内での英語力によるレベル分けである。よって同一の大学内でのプレイスメントに90問版がどの程度うまく機能しているかを確認する必要がある。

そこであるひとつの大学の実際のOL120のデータによるレベル分けの状況を確認し、そこから模擬的に作り出したOL90のデータによるレベル分けの状況を比較してみることとした。また90問からさらに問題数を6問ずつ漸減した場合に、レベル分けにどのような影響があるかも、合わせて探ってみた。

RQ1：単一大学内のレベル分けに短縮版OL90を使用すると、OL120を使用した場合に比べてどの程度の違いが生じるか。

RQ2：単一大学内のレベル分けにおいて、90問よりさらに問題数を減らした超短縮版を使用すると、OL120を使用した場合に比べてどの程度の違いが生じるか。

2.2 分析データ

分析対象としては2022年度にOL120を受験した大学のなかで最も受験者数が多かった（n = 2,737）A大学の解答データを選んだ。この解答データのVELCスコアの平均は443.1である。VELC Test®を受験している全大学の平均が例年おおよそ480である（靜, 2020）ことを考えると、A大学の英語力レベルはVELC Test®を受験する全国の大学の平均とそれほど離れておらず、本研究の分析対象として適当と判断した。全2,737名のなかで1問以上の項目に未解答であった496名を除き、全120問に解答のあった2,241名のデータを分析対象とした。

2.3 手順

分析の手順について簡潔に記しておく。それぞれの詳細については結果とともに改めて説明する。

(1) VELCスコアのおおよその分布を確認した。

(2) 2,241名をOL120の正答数によって4レベルに分けた。

(3) OL90を模擬的に作り出し、その正答数によって4レベルに分けた。

(4) OL120によるレベル分けを基準としてOL90のレベル分けがどの程度変わるかを調べた。

(6) OL84～OL6まで、6問刻みに項目数が異なる14の短縮セットをそれぞれ3種類（a, b, c）模擬的に作り出し、OL120との相関を確認した。

(7) OL120との相関に関してa, b, c間の差がほとんどなくなることが判明したOL84～OL60について、bセット（OL60b、OL66b、OL72b、OL78b、OL84b）の正答数によって4レベルに分け、OL120によるレベル分けとの異同を調べた。

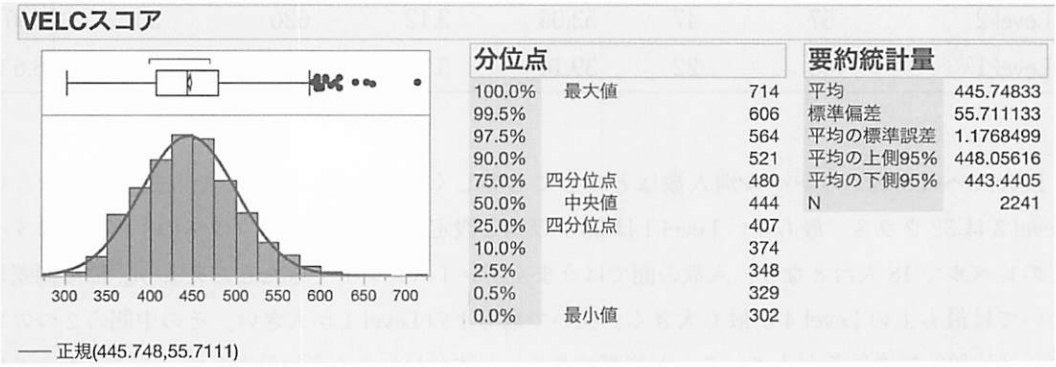
3. 結果

3.1 分析対象学生のVELCスコア

まず全体の分布を確認するため、図1にVELCスコアのヒストグラムおよび記述統計を示す。

視覚的に概ね正規分布に近くなっており、受験生が 302 から 714 まで広い範囲に分布していることがわかる。

なお、VELC スコアは素点ではなく素点を Rasch モデルに当てはめて処理した間隔尺度の数値であるが、本研究ではこれ以下すべての分析に VELC スコアではなく素点を用いる。VELC スコアは VELC Test の異なるフォームを受けた場合の結果を同一尺度上で比較する時には必須であるが、本研究のデータはすべて全員が同一のフォームを受けて得られたものである。その場合、素点と VELC スコアの間には、非リニアではあるが一対一の対応関係が存在する。よって単なるレベル分け（＝順位付け）用途であれば素点で必要かつ十分なのである。項目数を 90 問から模擬的に 84 問、78 問と漸減してゆくとき、42 種類のデータセットを扱うため、仮にそれらの素点を VELC スコアに変換するためには 42 回 Rasch モデリングのソフトウェア Winsteps（Linacre, 2005）を走らせる必要がある。それだけの労力をかけるのは非現実的かつ本研究の目的に照らして不要である。



分位点

100.0%	最大値	714
99.5%		606
97.5%		564
90.0%		521
75.0%	四分位点	480
50.0%	中央値	444
25.0%	四分位点	407
10.0%		374
2.5%		348
0.5%		329
0.0%	最小値	302

要約統計量

平均	445.74833
標準偏差	55.711133
平均の標準誤差	1.1768499
平均の上側95%	448.05616
平均の下側95%	443.4405
N	2241

図 1　A 大学受験者で全問解答した学生の VELC スコア分布状況

3.2　OL120 素点によるレベル分け

レベル分けに先立って、OL120 のデータセット（$n = 2{,}241$）のクロンバックアルファを確認すると、$\alpha = 0.888$ であった。なお、本論文中を通じ、クロンバックアルファの算出には関西大学の水本篤先生によるウェブアプリケーション https://langtest.jp/shiny/rel/（Mizumoto & Plonsky, 2016）を利用している。次に、120 点満点での素点によって 2,241 名を 4 つのレベルに分けた。その際、A 大学での実際のクラス分けの方式に準じて、以下の条件を満たすようにした。

(1) レベルは 4 レベル設定とし、全部で 120 クラスである。
(2) 最も上のレベル（Level 4 とする）は 20 クラスに固定し、その他のレベル（Level 3 ～ Level 1）はそれぞれ 33 クラスを平均とするがそのときの状況により数が増減する。
(3) スコアが同じ学生は、必ず同一のレベルに入る。

120という全体のクラス数の多さを除けば、「4レベル分けとし、最も上のレベルの人数をやや少なくする」というのは多くの大学に当てはまる方式ではないかと思われる。

この条件を満たしながらクラス毎の人数をなるべく等しくすると、20クラス＋(約33クラス×3)の、レベル別の総人数の割合は、Level 4：約16%、Level 3 ～ Level 1：各約28%である。この割合を今回の $n = 2{,}241$ に当てはめると、Level 4：約359名、Level 3 ～ Level 1：各約626名となる。この目安に準じながら「スコアが同じ学生は必ず同一のレベルに入る」というルールを守ってレベル間のボーダーラインを引いた結果、表1のような4つのレベルとなった。

表1　OL120の素点によるレベル分けの結果

	最高	最低	平均	標準偏差	人数	クラス数	平均人数
Level 4	111	71	79.52	7.33	363	20	18.15
Level 3	70	58	63.13	3.58	600	32	18.75
Level 2	57	47	52.05	3.12	626	33	18.97
Level 1	46	22	39.14	5.30	652	35	18.63

Level 3 ～ Level 1のレベル別人数はどうしても等しくはならない。そこで人数の最も少ないLevel 3は32クラス、最も多いLevel 1は35クラスと設定した。その結果クラスの平均人数はすべてのレベルで18人台となり、人数の面ではうまくプレイスメントできたと言えよう。標準偏差については最も上のLevel 4が最も大きく、次いで最も下のLevel 1が大きい。その中間の2つのレベルは比較的標準偏差が小さいことが観察できる。全体がおおよそ正規分布していることから予想された通りのパターンである。4レベルの状態を視覚的に確認するために箱ひげ図にしてみたのが図2である。Level 1とLevel 4で、それぞれ下方と上方にひげが長く伸びており、かつ外れ値もあることが視認できる。

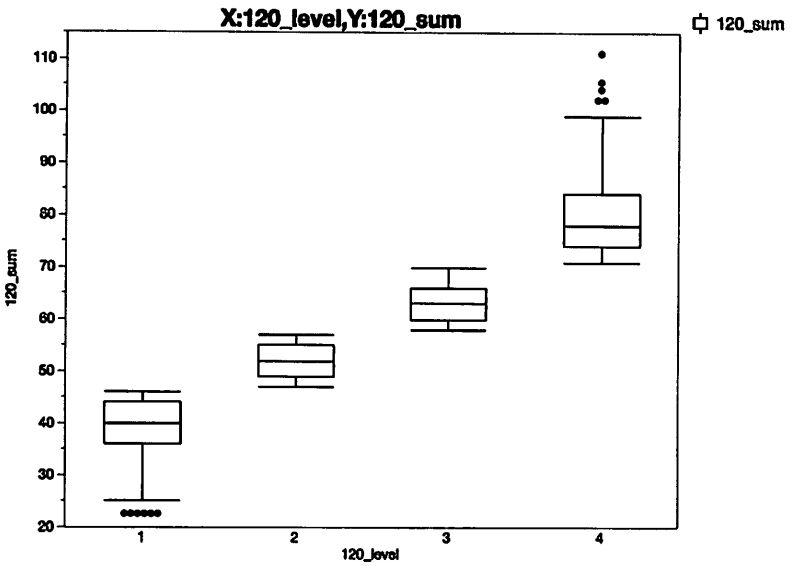


図２　OL120 を用いたレベル分けによる素点分布の箱ひげ図

3.3　OL90 素点によるレベル分け

つぎに 120 項目のなかで実際にこのフォームに対応する 90 問版で使用されている項目のみを抜き出すことにより模擬的な OL90 データセットを作成した。OL120 の時と同様、まずクロンバックアルファを確認してみると、$\alpha = 0.854$ であった。この OL90 データセットによって 90 点満点の素点を出し、OL120 のときと同じ条件によって 4 つのレベルに分けた。素点のレベル間のボーダーラインの位置（＝素点が変化する位置）は、OL120 と同じになるとは限らない。したがって 4 つのレベルに配分する学生の数も表 1 と同じになるとは限らない。今回の OL90 による素点の分布に即しながら、目安となる人数（Level 4：約 359 名、Level 3 ～ Level 1：各約 626 名）にできる限り近づける努力をした結果、最終的に表 2 のようになった。

表２　模擬 OL90 の素点によるレベル分けの結果

	最高	最低	平均	標準偏差	人数	クラス数	平均人数
Level 4	87	54	60.31	5.69	359	20	17.95
Level 3	53	44	48.00	2.79	620	33	18.79
Level 2	43	36	39.46	2.32	601	32	18.78
Level 1	35	16	29.81	4.11	661	35	18.89

OL120 の時と同様、Level 3 ～ 1 に割り振る人数を等しくすることはできず、Level 2 と Level 1 の間には 60 名の人数差が生じた。そこで Level 2 に 32 クラス、Level 1 に 35 クラスを設けること

とした。その結果、クラスの平均人数は Level 4 のみ 17 名台になったほかはすべて 18 名台となり、人数的にはうまく割り振れたと言えるだろう。標準偏差を見ると「Level 4 が最も大きく、次いで Level 1 が続き、間の Level 3 と 2 は比較的小さい」という、OL120 とまったく同じパターンが観察できる。

箱ひげ図を図 3 として示す。図 2 で観察できたのとほとんど同じパターンを視認することができる。

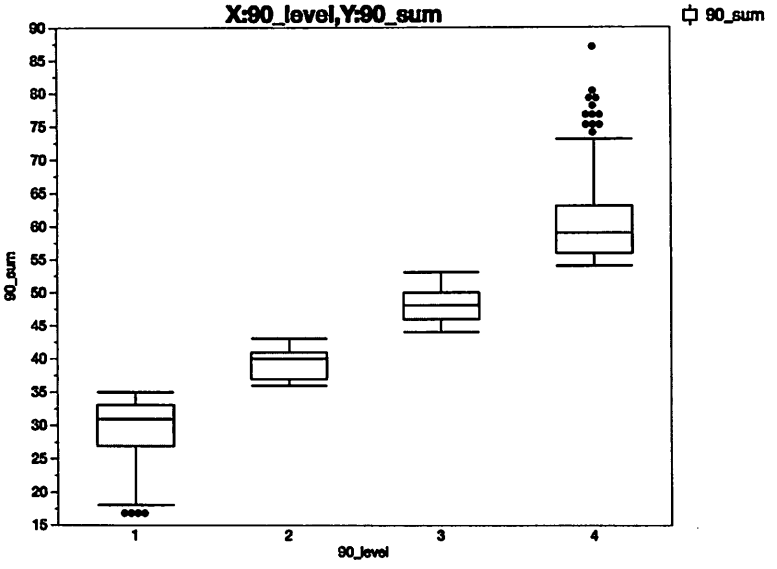


図 3　OL90 を用いたレベル分けによる素点分布の箱ひげ図

3.4　OL120 素点と OL90 素点の関係

OL120 による素点と OL90 による素点の関係を視覚的に確認するため散布図を作成した（図 4）。データポイントは明らかに右肩上がりの直線付近に密集している。次に相関係数を求めてみると、$r = .983$ と非常に高い値が確認できた。

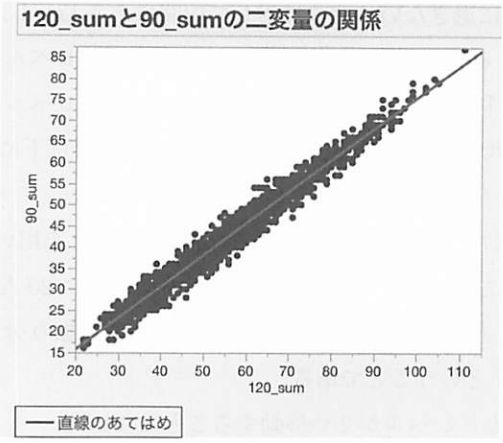


図 4　OL120 素点と OL90 素点の散布図

3.5　OL120 によるレベル分けと OL90 によるレベル分けの関係

では 120 問版によるレベル分けと 90 問版によるレベル分けはどの程度の違いがあるだろうか。120 問版で 4 つのレベルに分けられた学生たちが、90 問版ではそれぞれどのレベルに分けられたかをクロス集計した結果が、表 3 である。

表 3　OL120 と OL90 によるレベル分けのクロス集計

合計	661	601	620	359	2241	同レベル	別レベル
Level 4	0	0	35	328	363	90.36%	9.64%
Level 3	0	56	513	31	600	85.50%	14.50%
Level 2	56	498	72	0	626	79.55%	20.45%
Level 1	605	47	0	0	652	92.79%	7.21%
	Level 1	Level 2	Level 3	Level 4	合計	86.75%	13.25%

注：縦軸が OL120 によるレベル、横軸が OL90 によるレベル。

左端の Level 1 ～ Level 4 が OL120 によるレベルを、下端の Level 1 ～ Level 4 が OL90 によるレベルを表しており、クロスするセルに示しているのは度数（人数）である。度数が大きいほど濃い背景色で表示する設定をしている。圧倒的に背景色が濃く度数の大きいのが、OL120 と OL90 のレベルが同一である 4 つのセルで、左下から右上に対角線上に並んでいる。これら 4 つのセルの度数は全体のなかの 86.75％を占める。つまり 86.75％の学生は、OL120 でも OL90 でも同じレベルに割り振られたということである。それ以外の 13.25％の学生はもとのレベルよりも 1 つ上または下のレベルに割り振られた。もとのレベルよりも 2 つ上または下のレベルに移動した学生はいない。

当然ながら、もともと最も上の Level 4 の学生は、移動する場合はより下のレベルに動くしかなく、逆に一番下の Level 1 の学生は上に移動するしかない。おそらくそのこともあり Level 4 と Level 1 の学生は同レベルにとどまった率が非常に高く、それぞれ 90.36％、92.79％であった。移動したの

はそれぞれ9.64%、7.21%に過ぎない。それに対して中間にあるLevel 3とLevel 2は移動する方向として上下の両方がありうる。そのこともあってか、この中間2レベルに関しては、両端2レベルに比べておおよそ2倍ほど移動が多い。Level 3とLevel 2で同一レベルにとどまったのはそれぞれ85.50%、79.55%であり、残りの14.50%、20.45%はひとつ上または下に移動している。

以上をまとめると次のようになる。

1. OL120によるレベル分けを基準として表現するならば、OL90を用いた場合、全体で86.75%の学生が同じレベルにとどまった。約86.75%というのは、仮に20人のクラスだとすると17.35人にあたる。つまり20人中17人ないし18人が同じレベルに振り分けられ、残りの3人ないし2人が別のレベルに行くということである。
2. 別のレベルに行く場合も、レベルが2つ移動することはない。
3. 最も上のLevel 4および最も下のLevel 1では、同じレベルに振り分けられる率がそれぞれ90.36%、92.79%と特に高かった。20人クラスに換算すると、それぞれ18.1人、18.6人である。つまり20人中18人ないし19人はやはり最も上ないし最も下のレベルに振り分けられることになる。

3.6 項目数の漸減

実際に運用している90問版が120問版を基準としてどの程度違っているかは明らかになったので、次に、仮に90問よりもさらに項目数を減らしたテスト（以下、超短縮版）を使用した場合、レベル分けにどの程度の違いが生じるかを調べた。VELC Test® は6つの種類の項目群からなる6つのパートから構成されている。フルバージョンの120問版は各パート20問あり、そこから各パート5問ずつ減らしたのが90問版である。よって、さらに項目数を減らすにはテスト構成上、各パートで等しい数の項目を削除するのが適切である。そこで90問版を出発点として、そこからさらに各パート1問ずつ（すなわちテスト全体で6問ずつ）減らしてゆくことで、84問版、78問版、72問版、66問版、60問版、54問版、48問版、42問版、36問版、30問版、24問版、18問版、12問版、6問版という14種類の長さの超短縮版を模擬的に作ることとした。各パートで減らす項目を選ぶに際してはエクセルのRAND関数を利用して、項目数ごとに3バージョン（Set A、Set B、Set Cとする）作成した。すなわち14種類×3バージョン＝42のセットができた。

3.7 超短縮版素点と120問版素点の相関

これらの超短縮版（6問〜84問）全42セットおよび90問版（1セット）による素点合計と、120問の素点合計について、まず散布図を作成した（図5〜図7）

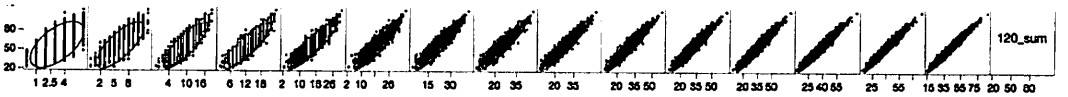


図 5　超短縮版（6 問～ 84 問）Set A および 90 問版の素点の 120 問版素点との散布図

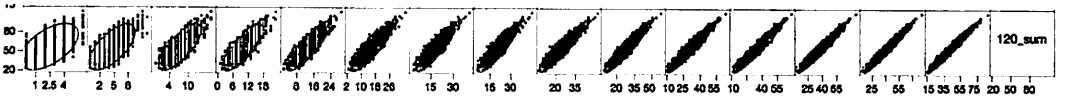


図 6　超短縮版（6 問～ 84 問）Set B および 90 問版の素点の 120 問版素点との散布図

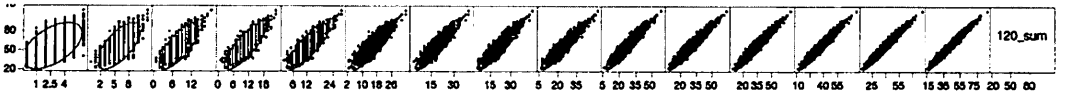


図 7　超短縮版（6 問～ 84 問）Set C および 90 問版素点の 120 問版素点との散布図

各図とも 15 のプロットが表示されている。すべてのプロットに関して Y 軸は 120 問版での素点、X 軸が当該の（超）短縮版の素点である。左端のプロットの X 軸は 6 問版で、その右が 12 問版、と項目数が増えていく。右端のプロットの X 軸は 90 問版であり、すでに図 4 として示したものと同一である。右端から数えて 2 つ目のプロットの X 軸が 84 問版となっている。どのセットでも、6 問版ではデータポイントがかなり広い範囲に散らばっているが項目数が増えるに従って徐々に範囲が狭まり、右上がりの直線に近づいてゆくのが視覚的に確認できる。右端の 5 つのプロット（66 問版～ 90 問版）はいずれのセットでも肉眼ではほとんど見分けがつかないように見える。

それぞれのピアソン相関係数の数値をまとめたのが表 4、それを折れ線グラフにしたのが図 8 である。

表 4　超短縮版（6 問～ 84 問）Set A ～ C および 90 問版素点の 120 問版素点とのピアソン相関係数

項目数	6	12	18	24	30	36	42	48	54	60	66	72	78	84	90	120
Set A	0.596	0.757	0.832	0.869	0.897	0.912	0.928	0.939	0.949	0.954	0.960	0.967	0.973	0.979	0.983	1.000
Set B	0.561	0.670	0.761	0.808	0.864	0.888	0.902	0.917	0.936	0.949	0.959	0.966	0.973	0.978	0.983	1.000
Set C	0.539	0.672	0.761	0.811	0.852	0.879	0.911	0.927	0.942	0.952	0.960	0.968	0.974	0.979	0.983	1.000

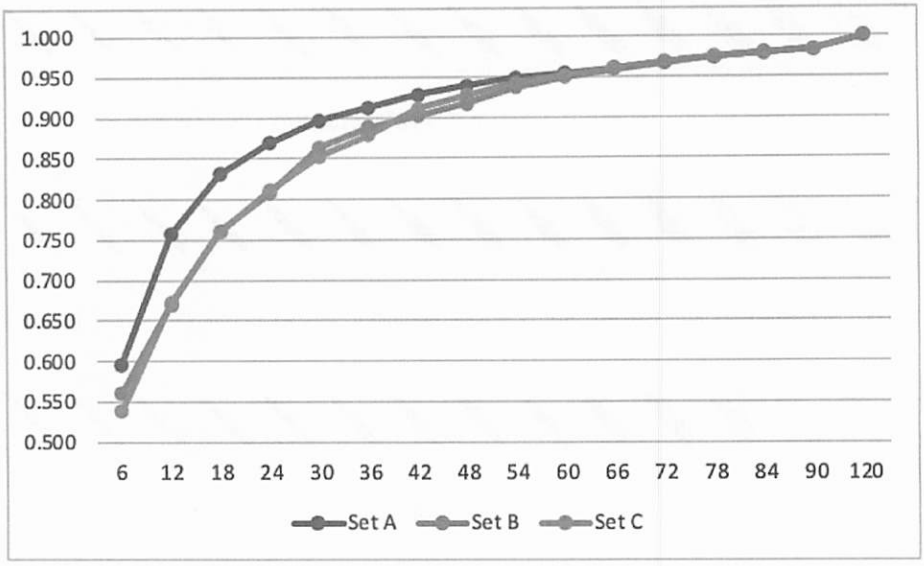


図 8　超短縮版（6 問～ 84 問）Set A ～ C および 90 問版素点の 120 問版素点とのピアソン相関係数

RAND 関数でランダムに項目を選んでいるのだが、項目数が少ない（概ね 50 問以下）うちは「当たり外れ」の影響が大きいのか、3 つのセット間でも相関係数の差が目立つ（なぜか一貫して Set A の値が他の 2 セットの値よりも高かった）。それが 60 問以上になると折れ線グラフ上では 3 つのセットの差は視認できなくなる。表 4 で数値を確認すると、たしかに 60 問版以上では相関係数の数値を小数点以下 2 桁で四捨五入数すると 3 セットで同一になる（60 問、66 問、72 問、78 問、84 問でそれぞれ、r = .95、.96、.97、.97、.98）。すなわち、60 問版以上であれば、Set A ～ C は事実上同一とみなせると言って良いだろう。そこでこれ以降は Set B のみを用いてさらに分析を行う（Set A でも C でもなく B にしたことには意味はない。Set A でも C でも概ね同じ結果が出ると考えられる）。

3.8　超短縮版によるレベル分け

3.7 で、60 問版よりも項目数が多ければ超短縮版 Set A ～ C は事実上同一をみなせることを確認した。そこで Set B に絞って 60 問版、66 問版、72 問版、78 問版、84 問版による 4 レベル分けを行った。

4 レベル分けに当たっては 3.2 に記したルールにより、スコアが同じ学生は必ず同じレベルに割り振ることとしてレベル間のボーダーラインを引いた。項目数が小さくなるほど同点の学生の割合が大きくなる傾向もあり、レベル毎の人数は一定にはならない。Level 4 が 16% で Level 3 ～ 1 がそれぞれ 28% という目安の数値にできる限り近づけようとした結果が、図 9 である。比較のために、基準となる 120 問版、90 問版のパネルも一緒に示している。

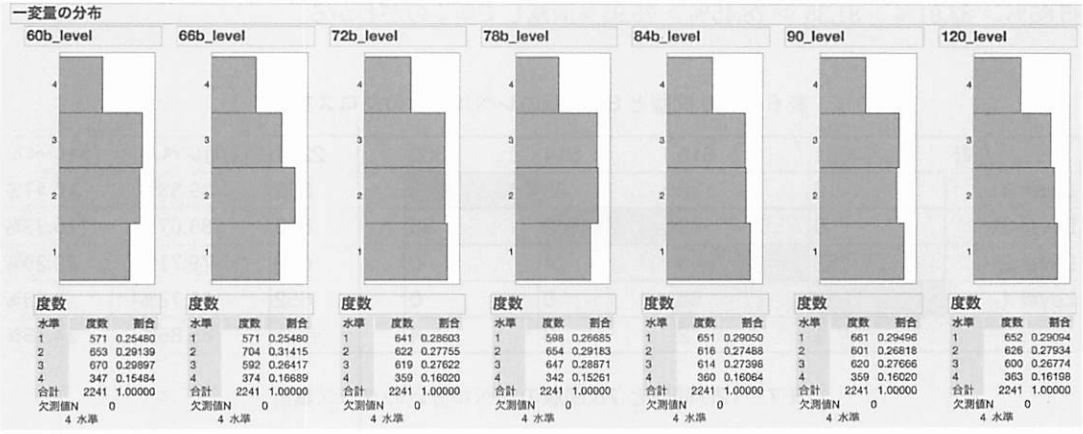


図 9　超短縮版および 90 問版と 120 問版によるレベル分け人数と割合の一覧

このレベル分けに基づいて、レベル毎のクラス数を設定し、ひとクラスあたりの平均人数を算出したのが表 5 である。

表 5　超短縮版によるレベル毎の人数、クラス数、およびひとクラスあたりの平均人数

	60問版(Set B)			66問版(Set B)			72問版(Set B)			78問版(Set B)			84問版(Set B)			90問版			120問版		
	n	c	n/c	n	c	n/c	n	c	n/c	n	c	n/c	n	c	n/c	n	c	n/c	n	c	n/c
Level 4	347	20	17.35	374	20	18.70	359	20	17.95	342	20	17.10	360	20	18.00	359	20	17.95	363	20	18.15
Level 3	670	35	19.14	592	32	18.50	619	33	18.76	647	33	19.61	614	32	19.19	620	33	18.79	600	32	18.75
Level 2	653	34	19.21	704	37	19.03	622	33	18.85	654	35	18.69	616	33	18.67	601	32	18.78	626	33	18.97
Level 1	571	31	18.42	571	31	18.42	641	34	18.85	598	32	18.69	651	35	18.60	661	35	18.89	652	35	18.63
	n: number of students　c: number of classes　n/c: mean number of students per class																				

Level 4 のクラス数は 20 に固定してある。Level 1 ～ Level 3 に関しては、合計クラス数が 100 という縛りを守りながら 31 クラスから 37 クラスに設定した。このレベル分けとクラス数設定ではひとクラスあたりの平均人数が最低 17.10 人、最高 19.61 人となり、我が国の英語クラスのサイズとしては妥当であると思われる。

3.9　超短縮版によるレベル分けと 120 問版によるレベル分けの関係

90 問版と 120 問版のレベル分けをクロス集計したのと同じ手法で、60 問～ 84 問の超短縮版によるレベル分けと 120 問版のレベル分けとクロス集計した結果を表 6 ～表 10 に示す。表 3 と同様に、度数の多いセルの背景色が濃くなるように設定している。どの表も濃い背景色のセル（レベルが同一であったセル）が左下から右上に配置されているのがわかる。「同レベル」の列にはレベル毎および全体で、超短縮版と 120 問版で同一のレベルにとどまった学生の割合が示されている。どの表でも Level 4 と Level 1 では同一レベルにとどまる率が高かった。全体的には同一レベルにとどまる割合が、84 問版＞ 78 問版＞ 72 問版＞ 66 問版＞ 60 問版と、項目数が減少するに従って、

85.85%＞82.91%＞81.35＞78.45%＞75.95%漸減してゆくのがわかる。

表 6　120 問版と 84 問版のレベル分けのクロス集計

合計	651	616	614	360	2241	同レベル	別レベル
Level 4	0	0	38	325	363	89.53%	10.47%
Level 3	0	63	502	35	600	83.67%	16.33%
Level 2	53	499	74	0	626	79.71%	20.29%
Level 1	598	54	0	0	652	91.72%	8.28%
	1	2	3	4	合計	85.85%	14.15%

表 7　120 問版と 78 問版のレベル分けのクロス集計

合計	598	654	647	342	2241	同レベル	別レベル
Level 4	0	0	56	307	363	84.57%	15.43%
Level 3	0	63	502	35	600	83.67%	16.33%
Level 2	43	494	89	0	626	78.91%	21.09%
Level 1	555	97	0	0	652	85.12%	14.88%
	1	2	3	4	合計	82.91%	17.09%

表 8　120 問版と 72 問版のレベル分けのクロス集計

合計	641	622	619	359	2241	同レベル	別レベル
Level 4	0	0	53	310	363	85.40%	14.60%
Level 3	0	70	481	49	600	80.17%	19.83%
Level 2	75	466	85	0	626	74.44%	25.56%
Level 1	566	86	0	0	652	86.81%	13.19%
	1	2	3	4	合計	81.35%	18.65%

表 9　120 問版と 66 問版のレベル分けのクロス集計

合計	571	704	592	374	2241	同レベル	別レベル
Level 4	0	1	49	313	363	86.23%	13.77%
Level 3	1	86	452	61	600	75.33%	24.67%
Level 2	56	479	91	0	626	76.52%	23.48%
Level 1	514	138	0	0	652	78.83%	21.17%
	1	2	3	4	合計	78.45%	21.55%

表 10　120 問版と 60 問版のレベル分けのクロス集計

合計	571	653	670	347	2241	同レベル	別レベル
Level 4	0	0	68	295	363	81.27%	18.73%
Level 3	1	75	472	52	600	78.67%	21.33%
Level 2	66	431	129	0	626	68.85%	31.15%
Level 1	504	147	1	0	652	77.30%	22.70%
	1	2	3	4	合計	75.95%	24.05%

図 10 に、全体としての同一レベル率と別レベル率が、項目数の減少に従って変化する様子を折れ線グラフで示した。120 問版を出発点として表現すると、同一レベル率は徐々に減少するが項目数がちょうど半分の 60 問版になってもおおよそ 4 分の 3 は同一レベルに留まることが視覚的に確認できる。

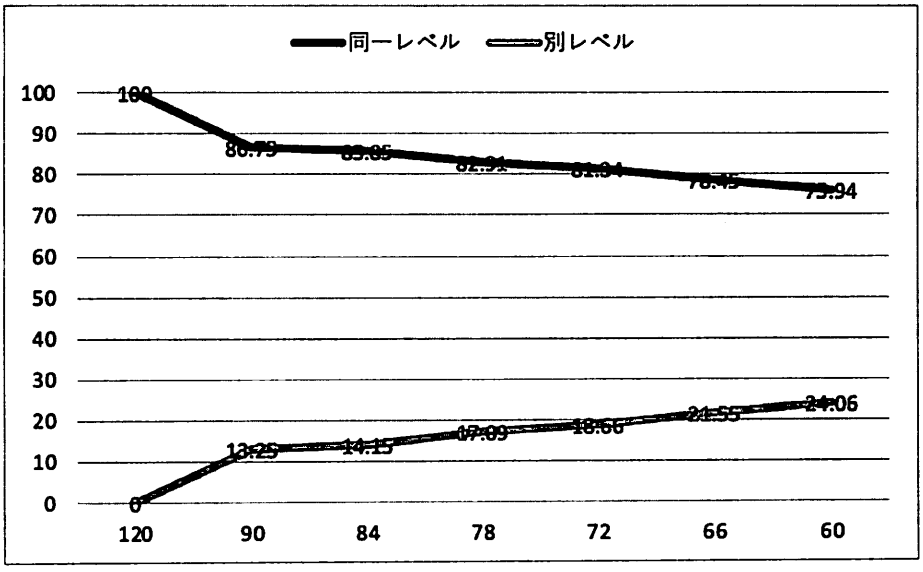


図 10　120 問版、90 問版から 60 問の超短縮版までの同一レベル率および別レベル率の遷移

その率を 20 人クラスに置き換えたときにどの程度の人数が同じレベルに残り、どの程度が別のレベルに移るのかを、表 11 にまとめた。90 問版では 20 人中 17 ～ 18 人が残って 2 ～ 3 人が別レベルに移るのに対して、60 問版になると 15 ～ 16 人が残って 4 ～ 5 人が別レベルに移る、といったイメージであることがわかる。

表 11　短縮版および超短縮版により、20 名クラス中別レベルに振り分けられる人数

	120問版	90問版	84問版	78問版	72問版	66問版	60問版
同レベル%	20.00	17.35	17.17	16.58	16.27	15.69	15.19
別レベル%	0.00	2.65	2.83	3.42	3.73	4.31	4.81
残る人数/20	20人	17～8人	17～8人	16～17人	16～17人	15～6人	15～6人
移る人数/20	0人	2～3人	2～3人	3～4人	3～4人	4～5人	4～5人

3.10　超短縮版の信頼性係数

最後に、84 問版から 60 問版の短縮版セット B を用いて信頼性係数としてのクロンバックアルファ

を求めてみた。すでに示した90問版と120問版の値も再掲し、表12として示す。

表12　120問版、短縮版、超短縮版の信頼性係数

120問版	90問版	84問版	78問版	72問版	66問版	60問版
.888	.854	.850	.845	.837	.821	.803

注：84問版～60問版はいずれもSet Bによる。

4. 考察

本研究では2022年にOL120を受験したA大学の学生のなかで、全項目に回答した2,241名の回答データをもとに、模擬的に90問の短縮版と、6問～84問の超短縮版を作り出し、その信頼性ならびにクラス分け状況を検証した。

＜短縮版＞

まず実際に運用されている短縮版OL90については、信頼性を表すクロンバックアルファは.854と十分に高く、OL120との相関は、r＝.983と非常に高かった。実際にA大学で行われている方式に準じてレベル分けしてみたところ、OL120でもOL90でも同一のレベルに配置される学生が、全体では86.75%いた。とくに最も上のレベルと最も下のレベルに関しては、OL120でもOL90でもレベルが変わらない学生の率が90%を超えた。

このパーセンテージはあくまでOL120の結果を100%としたときの割合である。100%という数値はともすれば「完璧な正確性」を表しているという誤解を招きがちだが、そうではない。測定しようとしている「真の英語力」に相当する「真の得点」とは、当該のテストを無限回繰り返したときの平均値であると定義される。よっていかなる現実のテストの1回限りの得点（観測得点）は100%「正しい」ことはありえず、そこには必ず測定誤差がある。同じOL120を（1度目の受験の後に受験者の記憶を消去して）2回受験したとしても、全員がまったく同じ得点になることはまずありえず、したがってレベル分けにはある程度の違いが生じるはずである。

そのような現象はいかなるテストであっても当てはまるものであり、テスト得点の正確性とは本質的に、絶対的なものではなく相対的なものである。そして問題なのは、その「相対的な」正確性が許容できる範囲にあるか否かである。それを判断するひとつの目安が信頼性係数である。上述したように、クロンバックアルファの値は、OL120がα＝.888、OL90で、α＝.854であり、どちらも十分に高い。靜（2022）が観察したα＝.91以上と比べればやや低いが、靜（2022）のデータは異なる4つの大学からの英語力的に幅の広かった参加者から収集したものであった。単一の大学からのデータである本研究とはそのまま比較できない。いずれにしても今回のOL90で確認されたα＝.854は、実用上十二分に高い信頼性を表しているといえる。よって、同一大学内でレベル分け（ク

ラス分け）をするという用途に照らしては、90 問の短縮版であっても、「十分に」機能していると言えるだろう。

「十分に」の意味をさらに敷衍してみる。もちろん 120 問版のほうが 30 問多い分、合計点に測定誤差が少ない。しかしそれも相対的な話である。180 問版とか 210 問版などを作ることも理論的には可能であって、そのほうが測定誤差はさらに小さくなり、より「真の学力」に近いものに迫ることができる(すべての受験者が疲れずに全力で取り組んでくれるならば、という条件付きであるが)。ただし項目数が多くなると 1 問増で得られる信頼性の増加幅は徐々に小さくなり、カーブの傾きはゼロに近づいてゆく。また、そうなるとテストの所要時間も 120 問版の 70 分よりも格段に長くなるため、受験者の負担および実施上の負荷も大きくなる。そのコスト（所要時間、実施の労力、受験者の疲労など）とパフォーマンス（結果の信頼性）のバランスを考えた上で、「ひとつ」の適切な解として 2012 年に我々がたどりついたのが、70 分で回答する 120 問版であった。仮定の 180 問版や 210 問版に比べればやや信頼性は落ちるが、それでも想定する用途に照らして十二分なパフォーマンスを発揮すると判断したものであり、その判断が妥当であったことはすでに検証されていると考える。

本研究では「もうひとつの」適切な解として 2020 年に運用開始した 90 問の短縮版のパフォーマンスについて、実際にひとつの大学の 120 問版回答データに基づいて、模擬的に検証してみた。その結論が「もちろん 120 問版と全く同じではないが、大きくは違わない、十分に信頼できる結果を出している」というものだということである。

＜超短縮版＞

そしてさらに一歩進めて、「120 問を 90 問にまで短縮しても問題ないのであれば、いったい何問まで短縮しても大丈夫なのか」という疑問に答えようとしたのが、後半の超短縮版についての検証である。90 問の短縮版からさらに 6 問（パートあたり 1 問）ずつ短くした超短縮版を、84 問版から 6 問版まで、それぞれ 3 バージョン作成し、(1) 120 問版との素点の相関、(2) 120 問版とのレベル分けの異同、(3) 信頼性の指標としてのクロンバックアルファを求めてみた。

120 問版との相関に関しては、項目数が概ね 60 問以上あればバージョンによる係数の差が事実上なくなり、$r = .95$ 以上と十分高くなることが確認された。超短縮版の素点に基づいて実際に 4 レベルに分けてみると、120 問版によるレベル分けと同一レベルにとどまる割合が、項目数が減少するにしたがってやはり徐々に減少した。84 問版で約 86%が同一レベルにとどまったが、60 問版ではそれが約 76%となった。本研究ではこの 60 問版を、許容範囲の下限として採用する。

フルバージョンである 120 問版のちょうど半分の 60 問版であっても、4 分の 3 強の学生が同じレベルに割り振られるのであり、かつ 60 問版自体の信頼性係数も .80 を超えるのである。仮に VELC Test® の超短縮版があったならば、60 問版以上であれば、同一大学内のクラス分けという比較的ローステイクスなテスト用途には耐えると言えるのではないだろうか。

もちろんこの後半の検証はあくまで仮定の話であり、現時点においては実際に超短縮版を作成して運用する予定はない。しかし仮定の話ではあっても 60 問の超短縮版でも最低限度に許容できる

機能を持つだろうという見通しが得られたことには一定の価値があると考える。それは現実に運用開始されている短縮版の90問版であれば、許容範囲の下限である60問版よりもはるかに余裕をもって許容範囲内の結果を出すであろう、ということを示しているからである。

謝辞

本論文の草稿に対して、麗澤大学の望月正道先生と東洋大学の熊澤孝昭先生より丁寧なフィードバックをいただきました。

引用文献

靜哲人（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.

靜哲人（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.

靜哲人（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.

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（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（2022a）VELC Test® Online と VELC Test® P & P の等価性を検証する. 大東文化大学紀要 ＜社会科学＞ 60, 173-191 頁

靜哲人（2022b）VELC Test® 短縮版の信頼性および基準関連妥当性の検証：項目数の漸減はテスト特性にどの程度影響を与えるか？ 大東文化大学 語学教育研究所『語学教育研究論叢』39，71-84 頁

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/

Mizumoto, A., & Plonsky, L. (2016). R as a lingua franca: Advantages of using R for quantitative research in applied linguistics. *Applied Linguistics*, 37, 284–291. doi: 10.1093/applin/amv025

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.