公益財団法人　日本英語検定協会

英語教育研究センター　委託研究

言語テストの規準設定

報告書（３）

２０１４年3月３１日

研究代表　大友賢二

研究副代表　渡部良典

言語テストの規準設定

報告書（3）

２０１４年3月３１日

公益財団法人　日本英語検定協会

英語教育研究センター　委託研究

研究構成員

伊東祐郎（東京外国語大学留学生日本語教育センター教授）
大友賢二（筑波大学名誉教授）：研究代表
法月 健（静岡産業大学情報学部教授）
藤田智子（東海大学外国語教育センター教授）
渡部良典（上智大学大学院言語学専攻教授）：研究副代表

（あいうえお順）

目次

Measurement of Change
3年間を振り返って

研究代表　大友賢二　・　研究副代表　渡部良典

「言語テストの規準設定」を主題としたこの研究成果は、2012年3月の第1号（179頁）、2013年3月の第2号（122頁）、そして、2014年3月の第3号（122 頁）の中にまとめられている。その内容の詳細に関しては各報告書に示しているが、「言語テストの規準設定」を、きわめて多角的な視点から捉えようとしたものである。その視点は、規準設定の意味と手順、CLIL（Content and Language Integrated Learning）、英語教育・日本語教育における Can-do statements、項目応答理論、潜在ランク理論などである。

海外に於ける規準設定の研究に比べて、この分野に関するわが国の研究と開発は、残念ながら歩調を合わせるまでには至っていない。Kane (1994); Hambleton & Pitoniak (2006); Cizek & Bunch (2007) 等に加えて、いま話題の CEFR の Can-do statements や CLIL とも関連してくる Frank van der Schoot (2009) などの規準設定法に関する提案などにもあまり興味が示されていない。

わが国の外国語教育の評価で最も欠けているものの一つは、設定した目標にほんとうに到達したかどうかという学習の“change”をどうしたら確かめることができるか、という分野である。それは、最近出版された McCoach, Rambo & Welsh（2012, p.216）のなかでの“Many of the most interesting research questions in education and social science involve the measurement of change”とも関連する。 そのさらなる開発のためには、言語テストの尺度化（scaling），規準化（norming），等化（equating）の方法，さらに、それを推進させるテスト理論等へともう一度戻って考察しなおすことであろう。そうした再出発のスタートラインとしての役割をこの報告書が果たせれば、この上ない喜びである。

この研究は、公益財団法人日本英語検定協会の委託研究助成、研究構成委員である伊東祐郎、藤田智子、法月　健という諸氏の献身的な努力の賜物であり、ここに心から感謝の意を表するものである。

CITO Variation on the Bookmark Method の一考察
Investigating the effects of the *CITO* Variation on the Bookmark Method

大友賢二
Kenji Ohtomo

Abstract

In the field of the standard setting methods in educational measurement, many agree that the bookmark should be placed at the point between the last question that borderline test takers would probably answer correctly and the first question that borderline test takers would not be able to answer correctly. But there is as yet no consensus on the method of placing the bookmark in that way. The main reason is that the method may often be influenced by subjective judgment. It may also be time-consuming for processing and interpreting scores by test users.

This paper explores the effect of the *CITO* variation on the Bookmark Method. This variation uses a rather simple display on which difficulty and discrimination values of all items are presented graphically in relation to the ability scale. An important feature of this display is that the panelists are fully informed about the level of mastery of all items in the item pool at every point of the ability scale.

The data developed by KNOX CUBE TEST, Wright and Stone was used in the calculation for the response probabilities of 50, 67 and 90. The important information we got from *CITO* Variation is the valuable procedure for the transformation of the latent scale: from the original values including negative numbers to transformed values on a scale from 100 to 400 and from transformed values to original values. When comparisons are made among the data produced by 1PLM, 2PLM and 3PLM, there has been no drastic difference among them in the results of the data for finding the cut points. This article suggests that if you use the *CITO* variation, there is no significant difference in the use of the different parameter logistic models even when the sample size is small.

CITO Variation on the Bookmark Method の一考察

『言語テストの規準設定』報告書（第3号）

大友賢二

1． 2012年度までの研究内容：

この委託研究は、2011年の4月にさかのぼる。2012の3月には，1年にわたって議論した成果をそれぞれの分野でまとめ，『言語テストの規準設定（第1号）』として発表している。筆者は，その中の第1章：規準設定の意味と歴史においては，1.1. 海外に於ける規準設定法の研究とその動向，1.4. 規準設定法：Bookmark Method の開発と発展，さらに、第5章：ヨーロッパ共通参照枠 ELP と規準設定では，5.2. CEFR 誕生とその背景：複言語・複文化主義など，また、5.4. CEFR と担当テストとの比較：その手段と方法を執筆している。その内容は，一言で言えば，規準設定ということのこれまでの研究の跡を探ってみた、いわゆる，先行研究である。

次の第2年度；2012年4月から2013年3月までは，予備調査：CITO Variation on the Bookmark Method と題して，焦点を Bookmark Method にあてた先行研究といくつかの筆者独自の視点に関する発表を行った。その柱は，1．規準設定の意味と必要性，2．規準設定のための方法，3．規準設定法に関するこれまでの評価などがある。特に，これまでの評価は，重要な視点であり，3．1．否定的見方，3．2．中立的立場、3．3．肯定的見方を明らかにした。さらに、焦点を Bookmark Method にしぼり、独自の考察を行った。 4．1．Bookmark Method の誕生と特徴，4．2．Response Probability の課題、4．3．Bookmark をおく場所，4．4．精神物理学の課題，4．5．応答確率と受験者の能力等これまでに行われた Bookmark Method の基本的な考えに関する整理をおこなった。

この第2号の報告書においては，こうした先行研究を顧みた結果，これまでの Bookmark Method よりすぐれた方法を見出すことができないかと考えるようになった。それが、5.「データによる分割点の設定」である。ここでは，まず，5．1．Schagen and Bradshaw(2003)をめぐって，5．2.「PNO 間の数値差」を利用した推定を提案した。つまり、Cizek, Bunch, and Koons (2004)で用いたデータ分析法は，視点を変えた独自の推定法でも分析は可能であろうと考えた。その独自の推定法をさらに推進する意味で，Wright and Stone (1979) で示されたデータを用いてその独自の推定法の検証を試みたのである。

その結果，PNO/TIN 間の数値差を利用した推定法は，より正確に，より短時間で，bookmark の置き場所を推定することができることを明らかにした。

現在，執筆中の日本英語検定協会:英語教育研究センター委託研究『言語テストの規準設定』報告書（第3号）「*CITO* Variation on the Bookmark Method の一考察」の構成は，すでに述べたように，（1）2012年度までの研究内容．それに加えて，（2）*CITO* variation on the Bookmark Method の特徴，（3）潜在ランク理論からの視点、（4）Sample Size と PLM の種類、（5）規準設定の手順、（6）むすび: という構成で進めることとする。

2. *CITO* Variation on the Bookmark Method の特徴

ここでは，おもに（2）*CITO* Variation on the Bookmark Method に関する内容のあらましを検討することとする。

Council of Europe (2009). *Reference Supplement to the Manual for Relating Language Examinations to the Common European Framework of Reference for Languages: learning, teaching, assessment.* Language Policy Division, Strasbourge. で記されている Frank van der Schoot による論文: Section 1 *CITO* Variation on the Bookmark Method の内容は，1.1. The construction of the item display, 1 .2. Introduction of the display to the panel members, 1. 3.The standard setting procedure, 1. 4. Practical considerations, APPENDIX: A.1. Finding the points RP50 and RP80, A.2. Transforming the latent scale, A.3. Decision making である。

規準設定法に関しては，Cizek, Bunch & Koon (2004) で示された DIFFICULTY, DISCRIMINATION, THETA@RP=.67 （その項目に67%の正答率が可能な能力水準）の方法，さらには Cizek (2006:247)で示されている「正答率が．67以下に下がるであろうと思われる OIB (ordered Item Booklet)の最初の頁に bookmark を置くこと」などという考えに加えて、大友（2013）『英語教育とテスト:第二言語習得における規準設定をめぐって』の中の，大友（2013）「PNO/TIN 間の数値差を利用した推定法」、法月(2013)「Rasch Model と LRT を併用した分割点設定法」などの考察が注目されている。

van der Schoot (2009)で提案されている CITO variation の主な特徴は，次のようなものがある。

特徴1:ability scale は、3つの領域に分けた方が便利で，意味がある。

(1) a segment that corresponds to insufficient or poor mastery where RP<0.50

(2) a segment that corresponds to moderate mastery where RP falls between 0.50 and 0.80.

(3) a segment that corresponds to full mastery, where RP is greater than 0.80.

しかし、80%の正答確率をもって，完全習得の範囲と決定するかどうかは，恣意的なものであり，そのように決定しなければならないという心理測定の理由は存在しないという見方をしている。

表2．1．2PLM で算出した The KNOX CUBE TEST, Wright and Stone (1979:31)のデータを用いた RP50, RP67, RP80 の設定

TIN	DIF(b)	DIS (a)	RP50	RP67	RP80
4	-1.93	0.93	-1.93	-1.48	-1.05
7	-1.78	0.86	-1.78	-1.30	-0.83
5	-1.76	0.92	-1.76	-1.31	-0.87
6	-1.57	0.90	-1.57	-1.11	-0.66
9	-1.55	0.95	-1.55	-1.11	-0.69
8	-1.10	0.95	-1.10	-0.66	-0.24
10	-0.73	0.87	-0.73	-0.25	0.21
11	0.69	0.85	0.69	1.18	1.65
13	1.31	0.93	1.31	1.76	2.19
12	1.47	0.85	1.47	1.96	2.43
14	1.97	0.85	1.97	2.46	2.93
17	2.39	0.95	2.39	2.83	3.25
16	2.40	0.95	2.40	2.84	3.26
15	2.41	0.94	2.41	2.86	3.28

上の表は，RP（Response P robability）=0.67 で示してあるものに、RP50と RP80のデータを付け加えたものである。たとえば、2PLM における RP50 は、ln (0.5/(1-0.5))/(1.7＊a)+b で求めることができる。

例：TIN7 (Test Item Number 7)における RP50 は、In(0.5/(1-0.5))/(1.7＊0.86) -1.78=-1.78
例：TIN15(Test Item Number15)における RP80 は、In(0.8/(1-0.8))/(1.7＊0.94)+2.41=3.28

特徴2：latent scale は、負の領域を含めない方が理解しやすく、便利である。A transformation of the latent scale to an ability that ranges from 100-400.

Original scale から Transformed scale（100−400）を求める方法、また、Transformed scale から Original scale を求める方法は以下の通りである。

表2．2．The KNOX CUBE TEST, Wright and Stone (1979:31)のデータを用いた Transformed Scale

	ORIGINAL		TRANSFORMED	
TIN	**RP50**	**RP80**	**RP50**	**RP80**
4	-1.93	-1.05	143.68	179.48
7	-1.78	-0.83	149.80	188.50
5	-1.76	-0.87	150.62	186.78
6	-1.57	-0.66	158.37	195.36
9	-1.55	-0.69	159.19	194.21
8	-1.10	-0.24	177.56	212.58
10	-0.73	0.21	192.66	230.91
11	0.69	1.65	250.63	289.77
13	1.31	2.19	275.93	311.73
12	1.47	2.43	282.47	321.61
14	1.97	2.93	302.88	342.02
17	2.39	3.25	320.02	355.04
16	2.40	3.26	320.43	355.45
15	2.41	3.28	320.84	356.27

２．１．Original Scale から Transformed Scale への変換

まず，PR50 の最小の値から少し小さめの値を求める。ここでは，最小値は，-1.93 であるので，それより少し小さめの値として，-3.00 を設定する。この-3.00 は最小値より，1.07 少ない数値である。さらに，RP80 の最大値より少し大きめの値を求める。ここでは，最大値は、3.28 であるので，それより少し大きめの値として、4.35 を設定する。この 4.35 は最大値より、1.07 大きい数値である。

つぎの式、(1), (2), (3)を設定する。

B＊(-3.00)+A=100 (1)

B＊4.35+A=400 (2)

A=100-B＊(-3.00) (3)

B を求めるために，(2)に(3)を代入すると、B＊4.35+100-B＊(-3.00)=400 となる。これを整理すると、B＊4.35-B＊(-3.00)=400-100。両辺に 1/B を掛けると，1/B＊B(4.35+3.00)=1/B(300)。つまり、7.35=1/B(300)。したがって、B＝300/7.35=40.82 。この B を(3)に代入すると，A=100-40.82＊(-3.00)=222.46 となる。

この結果を、V＝B＊ θ ＋A にあてはめて、original scale から transformed scale を求めることができる。

（例）Transformed RP50(TIN10)=40.82＊(-0.73)+222.46=192.66

（例）Transformed RP80(TIN10)=40.82＊(0.21)+222.46 =231.03

Transformed R P80(TIN10) の表2．2．の数値は 230.91 となっているが，それは小数点以下の数値をすべて用いた計算結果である。

2．2．Transformed scale から original scale への変換

まず，表2．2．の original scale において設定した最小の値(l)は、-3.00 であった。この-3.00 は表にある最小値 -1.93 より 1.07 少ない数値である。さらに、original scale において設定した最大の値(h)は、4.35 であった。この 4.35 は、表にある最大値 3.28 よりも 1.07 大きい数値である。また、transformed scale の最小値(L)は 100, 最大値(H)は 400 と設定している。

つぎの式、(4), (5)を設定する。

B=(H-L)/(h-l) (4)

A=L-B＊l (5)

b =(h-l)/(H-L) (6)

a =l-b＊L (7)

b を求めるために，データを(6) に代入すると、b＝(4.35-(-3.00))/(400-100)=0.025 となる。a を求めるために，データを(7)に代入すると、a=-3.00-0.025＊100=-5.5 となる。いま、transformed scale の値を Vc、original scale の値を θ c とする。そうすると，つぎの関係が成立する．つまり、θ c =b＊Vc＋a ということになる。

この関係にデータを当てはめて、transformed scale から original scale を求めることができる。

（例）Vc=230.91 であれば、θ c = 0.025＊230.91+(-5.5) =0.27: 表 2.2.の数値は，0.21 であるが，それは、少数点以下の数値をすべて用いた計算結果である。

（例）Vc=192.66 であれば、θ c =0.025＊192.66+(-5.5)=-0.68: 表 2.2.の数値は，-0.73 であるが,それは、小数点以下の数値をすべて用いた計算結果である。

特徴3：Interquartile r ange (四分位範囲)の設定などに関する理解があれば、その活用に役に立つことが多い。

この四分範囲（interquartile range）というのは，次の例で理解することが可能である。

例：1、5、7、10、13、16、18、20、24＜奇数のデータ：9個＞

中央値 (median) ＝ 左から5番目の＜13＞, 右から5番目の＜13＞で13が中央値＝ この中央値が第2四分位数(second quartile) と呼ばれるもの。第1・第3四分位数＝中央値を除いた8個のデータを下半分と上半分の2つに分ける。下半分の中央値（5+7）/2=6 が第1四分位数(first quartile)、上半分の中央値（18+20）/2=19 が第３四分位数(third quartile)となる。四分位範囲(interquartile range)＝第３四分位数ー第１四分位数で、19-6=13。 四分位偏差(interquartile deviation)＝四分位範囲/2 で、13/2=6.5。

例:1、5、7、10、14、18、20、24＜偶数のデータ:8個＞

中央値＝(10+14)/2=12。 第2四分位数＝中央値＝12。第1・第3四分位数:8個のデータを下半分と上半分に分ける。下半分の中央値（5+7）/2=6 が第1四分位数、上半分の中央値（18+20）/2=19 が第3四分位数となる。四分位範囲＝第3分位数ー第1分位数で、19-6=13。四分位偏差＝四分位範囲/2 で、13/2=6.5。

van der Schoot (2009:9-11)では、standard setting procedure の第3段階で、この四分位範囲（interquartile range）に触れて, つぎのような考えを示している。

このグラフ（Figure 8: Item map with interquartile range of judgments and with five percentile points of the ability distribution for a population of reference） では、審査員の2回目の検討結果の最終決定においては, 茶色の縦の線（つまり, 235から255の能力尺度に入る受験者）が、interquartile range であることを示すものである。つまり, 審査員の50％の方が, この2つの能力尺度の間に分割点は入ると判断している。そして, 分割点は235より低いと判断した審査員は25％、残りの25％は255より高いとろに分割点は設定するのがよいとしている。この範囲が最終段階まで審査員の間では一致できない場合と考えられる。しかし、この一致しない部分は、規定設定過程では報告されなければならない。

> The two thick vertical lines (with horizontal values of about 235 and 255, respectively) display the interquartile range of the final decisions after the second round. This means that 50% of the panel members arrived at a standard between these two values, 25% had a standard lower than 235 and 25% came up with a standard higher than 255. This range gives a picture of the remaining disagreement between panel members after a thorough discussions round, and this disagreement certainly must appear in the report on the standard setting procedure.(p.11)

このグラフの中のもっとも重要な点は, これは, impact information (衝撃的情報)を与えるということである。つまり、個々の決定値の中央値が最終決定とされた場合は, それは, 250に非常に近いものになり, それが, 全審査員の求めた値の中央値となるであろう。したがって, もし, この規準

が受験者の正否を決定するものとなれば，50％はこのテストに不合格となるであろう。このことは、審査委員にとっても，また情報を修正しようとする者にとって，また，最終決定に責任ある権威者にとっても，そして、審査員の能力以外の審査員からの最終忠告を変えようとしている者にとっても，きわめて重要な情報であるとしている。

> The most important feature of Figure 8, however, is that it provides impact information (see Section 6.2.1 of the Manual). If the median of the individual decisions is taken as the final group decision, it is seen to be very close to 250, which is also the median of the ability distribution in the population. If this standard were to be used for deciding on success or failure in an examination, it follows that about 50% of the population would fail the examination. This is important information for the panel members, who might wish to revise their decisions, but also for the authority that is responsible for the final decision, and who might change the final advice from the panel because of reasons outside the competency of the panel members. (p.11)

特徴4：困難度や弁別力等の状況は，「能力尺度」に関連したグラフで示されている。

Bookmark Method の改善を述べている van der Schoot (2002:2)の考察のいくつかを特徴1，2，3で述べてきたが，その最も基本的考察は、以下の通りである。

この方法では，すべてのテスト項目の困難度と弁別力の状況は，「能力尺度」に関連したグラフで示されるということである。この示し方の重要な特徴は，審査員に対しては、すべての項目の達成度は，「能力尺度」のあらゆる時点で，テストあるいは，項目群の中の項目における達成度に関連しているということである。このことで，審査員は，テスト項目の相対的困難度がわかり，さらに，一貫性のない規準設定を避けることができるということである。

この特徴を明確に示すためのグラフ作りを行った結果を以下に示すとしよう。

（1） ABILITY SCALE: Wright and Stone: Original RP(2PLM)

このグラフは，さきに作成した RP（response pr obability）50 と RP80とを表に表したものである。たとえば、前に述べた RP67というのは、その該当する項目に対して67％の確率を持って正解できるであろうと思われる能力を指すことである。67％の確率で正解できるというのは、3回の試行において，2回の正解が出せるであろうと思われる能力である。2/3=0.666=0.67 で計算できる。また、RP80というのは，したがって、10回の試行において8回の正解が出せるであろうと思われる能力である。8/10=0.8 で計算できる。同様に、RP50というのは，10回の試行において5回の正解が出せるであろうと思われる能力をさすことになる。そうした能力は、しかし、どうすれば求めることがで

きるか。その求め方は，すでに述べたように，項目応答理論による公式を用いて算出することができる。ここでは、２PLM を用いているので，P=1/(1+exp(-Da(θ -b)))を変換した θ ＝ ln(P/(1-P))/(Da)+b で求めることができる。したがって，該当する項目の困難度パラメータ(b)，弁別力パラメータ(a)，そして、正解するであろう 確率(p)が解っていれば，それを算出することができる。たとえば、b＝-0.73、a=0.87 の項目（10）に対し，10回の試行において8回の正解を出せるであろうと思われる能力である RP80 を求めたい場合は， θ ＝ln (0 .8/(1-0.8))/(1.7＊0.87)+(-0.73) =0.207 で求めることができる。この項目に対して、10回の試行において5回の正解を出せるであろうと思われる能力である RP50を求めたい場合は，θ ＝ln (0 .5/(1-0.5))/(1.7＊0.87)+(-0.73) = -0.73 で求めることができる。

さらに、RP50 と RP80 から推測される項目（14）の状況について考えてみる。項目14に対しては，この項目に完全に解答可能な能力は，どのぐらいかを求めれば，それは，2.93 以上と推定できる。したがって能力が2.93 以上であれば，full mastery of item 14 という解釈することができる。さらに、能力が1.970から2.93 の間であれば，moderate mastery of item14 と解釈することができる。そして、能力が，1.970以下であれば、poor or insufficient mastery of item 14 と解釈することができるということである。

さらに、項目（13）と項目（17）とを比較検討してみることとする。RP50に関しては，項目（13）は1.31、項目（17）は 2.390，RP80に関しては，項目（13）は 2.19、項目（17）は 3.25 である。
したがって、この2つの項目に解答する場合，つぎの5つの分野に分けて考えることが可能である。

（1）1.31 より少ない能力を持っている受験者は，項目（13）と（17）の両方の項目に正解することは困難である。
（2）1.31 と2.19 の間の能力を持っている受験者は，項目（13） には適切な正解ができるが，項目（17）に正解することは困難である。
（3）2.19 から 2.39 の能力を持っている受験生には，項目（13）には正解することができるが，項目（17）に正解することは困難である。
（4）2.39 から 3.25 の能力を持っている受験生には，項目（13）は完全に正解できるし、項目（17）に関しては、適切な正解が可能である。
（5）3.25 以上の能力を持っている受験生にとっては，項目（13）(17)両方の項目に対して完全に正解することができる。

（2）ABILITY SCALE: Wright and Stone: Transformed RP (2PLM)

ここでは，Transformed RP のデータを用いて，項目（5）と項目（10）とを比較検討してみること

とする。RP50に関しては，項目(5)は150.62、項目(10)は192.66，RP80に関しては，項目(5)は186.78、項目(10)は230.91である。

したがって、この2つの項目に解答する場合，つぎの5つの分野に分けて考えることが可能である。

(1) 150.62 より少ない能力を持っている受験者は，項目(5)と(10)の両方の項目に正解することは困難である。

(2) 150.62 と186.78 の間の能力を持っている受験者は，項目(5)には適切な正解ができるが，項目(10)に正解することは困難である。

(3) 186.78 から192.66 の能力を持っている受験生には，項目(5)には正解することができるが，項目(10)に正解することは困難である。

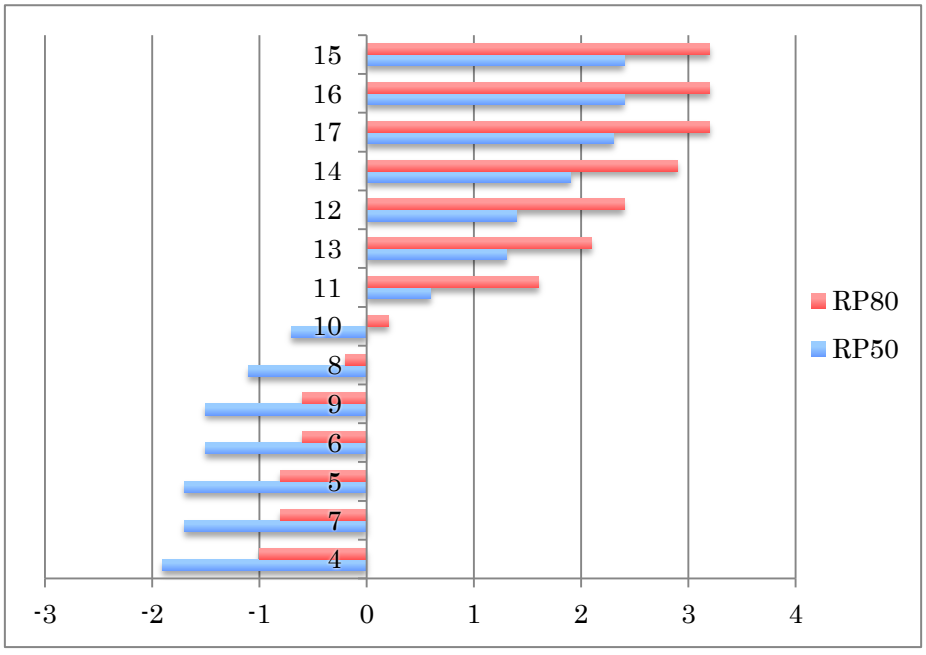
(4) 192.66 から 230.91の能力を持っている受験生には，項目(5)は完全に正解できるし、項目(10)に関しては、適切な正解が可能である。

(5) 230.91以上の能力を持っている受験生にとっては，項目(5)(10)両方の項目に対して完全に正解することができる。

The KNOX CUBE TEST, Wright and Stone (1979):

Original Scale: RP50, RP80

	4	7	5	6	9	8	10	11	13	12	14	17	16	15
RP50	-1.9	-1.7	-1.7	-1.5	-1.5	-1.1	-0.7	0.6	1.3	1.4	1.9	2.3	2.4	2.4
RP80	-1.0	-0.8	-0.8	-0.6	-0.6	-0.2	0.2	1.6	2.1	2.4	2.9	3.2	3.2	3.2



上の表で示されている横棒は、赤が RP80を示し、青が RP50を示すものである。例えば、項目10

（下から7番目）を例にとってみると、RP80＝0.2、RP50＝-0.7を示すものである。この場合は、項目10に関しては、0.2の能力を持っている者は10回の受験のうち、8回は正解を得る可能性を持っているということができる。また、-0.7の能力を持っている者は10回の受験のうち、5回は正解を得る可能性を持っているということができる。

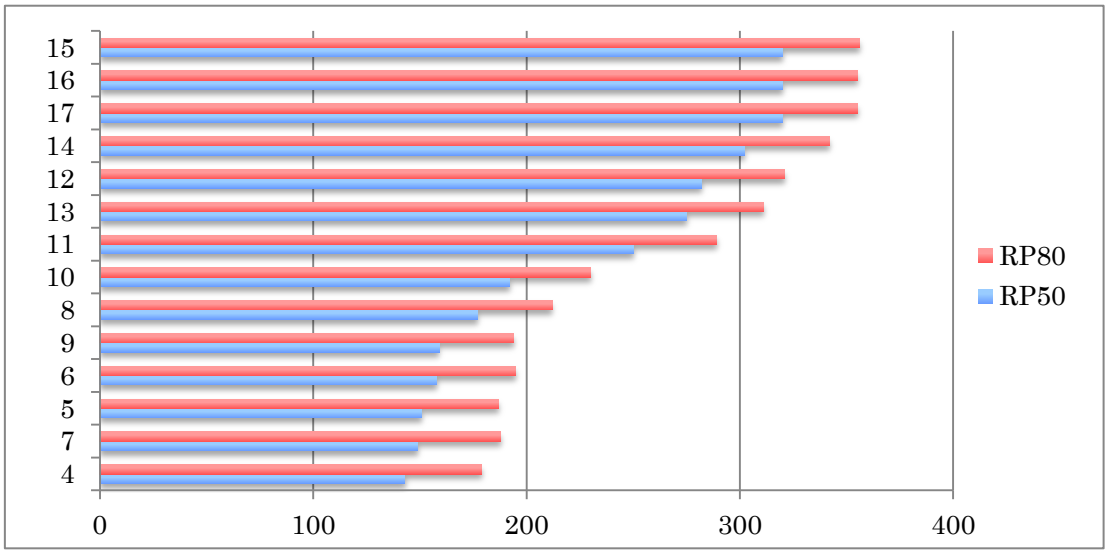
この項目だけに限って言えば、能力-0.7以下では正解を得る可能性は低い. -0.7から0.2 の能力を持つ者は、適切な正解ができる。さらに、0.2以上の能力を持つ者は、完全に正解できると解釈することができる。

このグラフは、操作の都合で、赤棒も青棒も能力ゼロから表示しているが、van der Schoot (2009)では、-0.7から0.2までを結ぶ横棒で示している。

The KNOX CUBE TEST, Wright and Stone (1979):

Transformed Scale: RP50, RP80

ITEM	4	7	5	6	9	8	10	11	13	12	14	17	16	15
RP50	143	149	151	158	159	177	192	250	275	282	302	320	320	320
RP80	179	188	187	195	194	212	230	289	311	321	342	355	355	356



上の表で示されている横棒は、赤がRP80を、青がRP50を示すものである。例えば、項目10を例にとってみると、RP80＝230、RP50＝192を示すものである。この場合は、項目10は、230の能力を持っている受験者は、10回のテストで8回は正解を得る可能性を持っているということである。また、192の能力を持っている受験者は、10回のテストで5回は正解を得る可能性を持っているということになる。

この項目だけに限れば、したがって、能力192以下では正解を得る可能性は低い。そして192から230の能力を持つ受験者は、適切な正解を得ることができる。そして、230以上の能力を持っている者は、完全に正解できると解釈することができる。

このグラフは、操作の都合で、赤棒も青棒も能力ゼロから表示しているが、van der Schoot (2009)では、192から230までの範囲を横棒で示している。

3. 潜在ランク理論からの視点

言語テストに限らず、テストと呼ばれる手段で求められた情報の解釈はさまざまである。例えば、2013年度の入学試験の英語の平均は、68点であった。しかし、2014年度の平均は、60点となってしまった。こうした現象に対しては、きわめて注目すべき解釈を必要とする。8点の低下をどう捉えるかという問題である。平均点が8点も下がったので、受験者の英語能力は下降をたどっていると見る。その結果、わが国の英語教育の質の低下ではないかという声もでてくる。そこで、最も大きい原因は，やはり、英語教員の質の問題だとかの声も出てくる。

しかし、ごく単純に考えただけでも、これを英語教員の質が原因という解釈は正しいと考えることには相当の矛盾を含んでいることがわかる。そこで、まず、テスト問題が違っているではないかという指摘である。2013年度のテスト問題よりも、2014年度のテスト問題が難しかったのが原因であって、英語教員の質や、受験者の能力が低下しているということにはならないという反論である。しかし、テスト問題の難しさを一定に保つにはどうするかという課題が出てくる。

そこで顔を出してくるのが、古典的テスト理論（Classical Test Theory: CTT ）に対する項目応答理論（Item Response Theory: IRT）である。 この IRT の利点は、ごく簡単に言えば、（1）test-free person measurement であり、（2）sample-free item calibration であり、（3） multiple reliability estimation である。（1）は、どんな異なったテストを用いても、共通の尺度で能力測定が可能であるということである。つまり、被験者の能力推定は、その被験者に実施された特定のテスト項目とは切り離して独立に求めることができるということである。（2）は、どんな受験者集団にたいしても、共通の項目特性に関する値を求めることが可能であるということである。つまり、項目困難度パラメータ、あるいは、項目弁別力パラメータの値等の項目特性は、受験者集団とは独立して求めることができるということである。また、（3）は、能力毎にわかる測定の精度を持っているということである。たとえば、適応型テストでは、情報関数の値が最大になるように、個人の能力に応じたテスト項目をコンピュータに集めさせて、被験者の能力にあった無理のない項目で能力を効率的に測定することができる，ということである。

しかし、もう一つ、考えなければならない課題がある。たとえば、ある受験者の英語テストの結果が65点であるとき、その得点が、64点でもなく66点でもなく、まさに65点であると信じることができるであろうか？それは、測定誤差の課題でもあろう。しかし、最近の研究結果では、学力を段階評価するための「潜在ランク理論」（Latent Rank Theory: LRT）が顔を出していることに注目しなければならない。

この考え方を押し進めている研究者の一人、荘島（2010:84）での次の発言は注目に値する。

第2章で紹介した古典的テスト理論（Classical Test Theory: CTT）や、第3章で紹介した

項目反応理論（item response theory: IRT）は、学力を連続尺度上で評価している。例えば、CTT は、T という連続尺度、IRT は、θ という潜在的な連続変数を仮定し、受験者の学力を評価する統計モデルである。一方で、テストは、5-20レベルぐらいに学力を段階評価するくらいの解像度しかないと考え、そのような興味から出発して作られたテスト理論がある。本節では、学力を段階評価するためのテスト理論であるニューラルテスト理論（neural test theory: NTT)(Shojima, 2008a, 2008b, 2009; 荘島、2009)について紹介する。

また、最近，文部科学省「各中・高等学校の外国語教育における「CAN-DO リスト」の形での学習到達目標設定のための手引」や「グローバル化に対応した英語教育改革実施計画」等と関連して、CAN-DO statements などの議論が盛んに行われているが、その点に関しても、このテスト理論は、深く立ち入るものである。そのことは、荘島（2010:107-108）に見られるつぎの一節でも理解できる。

IRP (item reference profile: 項目参照プロファイル)は、各潜在ランクに所属する受検者の各項目に対する正答率であるので、各潜在ランクに所属する受検者たちがいったいどのような項目群（個別能力）にパスし、どのような項目群が未到達であるかについて考察することができる。たとえば、R2 に所属する受検者は、リスニングによる地図の読み取りをはじめとする基礎的な単語能力は獲得しているが、文法力と読解力を獲得していない、などと、各潜在ランクに所属する受検者たちの能力のプロフィールが浮き上がってくる. それをもとに、Can-Do Statement (学習進度記述文)に要約し、その記述文を参考に各潜在ランクにタイトルがつけられる。

こうした理論の開発に伴って、言語テスとの規準設定には、この潜在ランク理論は大きな貢献をなすことが考えられる。大友（2013b）「英語教育とテスト：第二言語習得における基準設定をめぐって」（第7回日本テスト学会賞記念講演）では、PNO/TIN の数値差を利用した推定法を発表しているが、それをより理論的に証明する手段の一つとして、法月（2013）「Rasch Model と LRT を併用した分割点設定法」が研究協力者のひとりとして発表していることは、注目に値する。

Lewis, Mitzel, and Green (1996), Standard Setting, A Bookmark Approach の発表以来注目を浴びている規準設定法は、（1）IRT の活用、（2）複数の分割点の設定、（3）多肢選択形式テストでも、記述式テストでも活用、（4）審査員の作業は極度に簡素化、（5）テスト項目の内容を反映した評価、等の特徴をもっている、いわゆる「Bookmark Method」である。その後の Cizek, Bunch, and Koons (2004) や、Cizek and Bunch (2007) を背景に、その改善の第一歩として示したのが、大友（2013a, 2013b）における規準設定の方法「PNO/TIN の数値差を利用した推定法」である。

「PNO/TIN 間の数値差を利用した推定法」の手順は、それを要約すると、大友（2013a：33）

で述べたつぎのようなものになる。(1)使用したテスト結果を IRT で分析する。(2)RP(response probability) を設定し、Theta@RP を算出して、OIB (ordered item book) を作成する。 (3)低から高へ配列した difficulty, discrimination, theta を作成する。(4)PNO(page number in OIB)/TIN (test item number) の間の数値差を求め、GDN (graph data number) にそって表とグラフを作成する。(5)PNO/TIN の間の数値差が最大の GDN とその前後の GDN を選定する。(6)以上の2つの GDN に共通に含まれる, あるいは、単独で含まれる PNO/TIN を選定する。(7)以上の PNO/TIN を bookmark の置き場所とする。ここで設定した bookmark の置き場所は、TIN=10 と推定された。

この大友(2013a)で設定した TIN=10 が分割点として適切であるかどうかを、Rasch Model と LRT を併用した分割点設定法で検証してみることが、法月(2013)の課題であった。使用した分析のためのプログラムは、Rasch Model に関しては、Winsteps Ver.3.80.1 (Linacre, 2013), 潜在ランク理論では、Exametrika Ver.5.3 (荘島, 2011) であった。

その手順に関する詳細は、法月(2013a:81-103)「受容語彙力を測定するプレイスメントテストにおけるラッシュモデルと潜在ランク理論に基づく規準設定の試行」、法月(2013b)「 Rasch Model と LRT を併用した分割点設定法」に委ねるが、その手順と結論のあらましは、以下の通りである。

分析データは、KNOX CUBE TEST, Wright and Stone (1979:31)の受検者35名、テスト項目18である。

分析手順(1)

S1: LRT(Exametrika) 分析ファイル(Excel)の＜Examinee＞のシートに、RM(Winsteps)分析で得られた受験者能力と項目難易度の値を挿入する。

S2: (1)受験者能力降順、(2)潜在ランク降順、(3)RMP の Rank2 降順に並べ替える。

S3: (1)RMP ランク2＝＞1、(2)受験者能力の数値の変化、(3)項目難易度の数値の変化を検討して、TIN の分割点を検討する。

＊ – 1.37 の受験者が終わる地点、つまり、テスト項目＜10＞が項目難易度の観点から分割点として妥当と思われる。

分析手順(2)

LRT 分析ファイルの＜item＞シート

(1)項目参照プロファイル(IRP)の Rank 1 昇順 (2) IRP 指標の Beta 降順 ＝＞並べ替え。

＊ LRT の項目指標の観点からも、テスト項目＜10＞が分割点領域に位置する。

以上のように、Rasch Model とLRTを併用した分割点設定法においても、TIN間の数値差を利用した推定法を用いた推定法を支持する結果となっている。

ここでは、これまでのデータ処理のまとめとして、正答確率、受験者能力、項目困難度、弁別力、などとの関係を、例をあげて示してみることとする。これに関しては、最も多く使われている1PLM や2PLM に限定して、その例を取り上げてみることにする。

たとえば、法月(2013)の＜分析結果(1−2)＞の中にある受験者17の例を挙げるとしよう。ここの1PLM で求められた困難度パラメータが − 1.57 の項目に対し、正答確率が0．55である場合の受験者能力を求めたければ、$\theta = \ln(P/(1-P))+b$ に、そのデータを使って，$\theta = \ln(0.55/(1-0.55))-1.57 = -1.37$ で求めることができる。また、これと関連して、受験者能力が -1.37 である場合、項目困難度パラメータが -1.57 の項目を受験して得られる正答確率を求めたければ、$P = 1/(1+\exp(-(\theta-b)))$ に、そのデータを使って、$P = 1/(1+\exp(-(-1.37+1.57))) = 0.55$ で求めることは可能である。

また、Cizek, Bunch and Koons (2004: 39) でのデータを例に取るとつぎのようになる。2PLM で求められた困難度パラメータが、-3.395、弁別力パラメータが0.493である項目に対し、正答確率が0.67である場合の受験者能力を求めたければ、$\theta = \ln(P/(1-p))/(Da)+b$ にそのデータを使って、$\theta = \ln(0.67/(1-0.67))/(1.7*0.493) - 3.395 = -2.55$ で求めることができる. これと関連して、受験者能力が − 2.55 である場合、困難度パラメータが − 3.395, 弁別力パラメータが 0.493 の項目を受験して得られる正答確率を求めたければ、$P = 1/(1+\exp(-Da(\theta - b)))$そのデータを使って、$P = 1/(1+\exp(-(1.7*0.493)*(-2.55+3.395))) = 0.67$ で求めることが可能である。

4. Sample Size と PLM の種類

4．1．Sample size の課題:

項目応答理論の利用に関しては、それに用いられる標本の大きさが課題になることがある。用いられる標本数は、どの程度のものが適切と考えられるのであろうか。最低、必要である標本数はどのくらいと考えたらよいかという課題である。この予備調査では、その標本数は大きな問題としないで、その算出の手順に重点をおいて考察してきている。手順は可能であるが、その標本数に課題がありとした場合は、解決すべき課題になるので、この場で、それを確認し、結果の解釈にはそのことを含めておくのが穏当な方向であろうと思われる。したがって、これまでの検討結果を、ここで押さえ、今後の検討課題題の一つとしておくこととする。

Robert Linn Ed (1989) *Educational M easurement (Third Edition),* National Council on Measurement in Education, American Council on Education のなかの Hambleton, R. K. 著

Principles and Selected Applications of Item Response Theory は、野口(1992:211-282)「項目応答理論の基礎と応用」として日本語の訳されているが、そのなかに、4-3: 適切なテストの長さと標本数が示されている。

> 多数の研究者が満足な最尤推定値を得るために必要なテストの長さおよび標本数のガイドラインを示唆してきた。Wright & Stone (1979) は1パラメターモデルに対して、少なくとも20項目の長さと200名の標本数を用意することを勧めている。Hulin, Lissak, & Drasgow (1982)は少なくとも、次に示すテストの長さと標本数を用意することを勧めている。すなわち、2パラメータ・ロジステック・モデルに対して30と500、3パラメタ・ロジステック・モデルに対して60と1000。Swaminathan & Gifford (1983) は、20項目という短いテストそして1,000名の受験者という状況で LOGIST を用いた場合満足なモデルパラメータ推定値が得られたことを報告している。さらに、彼らは80項目テストではすべてのモデルパラメータについてよい推定値が得られたこと、a および θ パラメータは特にテストの長さの増加が良い結果をもたらすこと、そしてテストが短いとき (n<15 項目)、a パラメータで質の悪い推定値が得られたことを報告している。

さらに、Henning, G. (1987: 116-117), *A Guide to Language Testing,* Newbury House での以下の指摘も、IRT の標本数に関する言及として、注目に値するものである。

> As Table 8.1. indicates, the Rasch One-Parameter Model is probably to be preferred by teachers and language testers over the other models for the majority of testing situations. Sample size constraints alone may dictate this choice since the Rasch Model is fully operative with a sample of from 100 to 200 persons, while the Two- Parameter Model requires 200-400, and the Three-Parameter Model depends on the availability of 1,000 to 2,000 persons for parameter estimation to proceed meaningfully.

4．1．1．1PLM での分析

さきに、2PLM に関する規準設定の考察を行ったが、それに用いた手順は、1PLM や3PLM でも同様に適応できるかを確かめなければならない。つまり、標本数が十分ではないと思われる1PLM でも、3PLM でも、その手順は適応可能であるかという課題である。ここでは、前に用いたものと同じデータ:Wright & Stone (1979:31) KNOX CUBE TEST を用いることとする。

1PLM: RASCAL Ver 3.50 (Assessment System Cooperation)
Wright & Stone (1979: 31) KNOX CUBE TEST

1PLM: P =1/(1+exp(-(RP θ - b))),

RP θ = ln (P/(1-P)) +b

TRP (Transformed Response Probability) = 28.93＊RP θ ＋224.40

From ORIGINAL VALUE to TRANSFORMED VALUE

A bit smaller than the smallest RP50 -4.21 + (-0.09) => - 4.30

A bit larger than the largest RP80 5 .98 + (0.09) => 6.07

つぎの式を設定する。

B＊(-4.30) + A = 100

B＊(6.07) + A = 400

A = 100 – B＊(-4.30)

B＊6.07 + 100 – B＊(-4.30) = 400

B = (400-100)/(6.07- (-4.30)) = 300/10.37 = 28.93

A = 100 – 28.93＊(-4.30) = 224.40

V = 28.93＊RP θ + 224.40

したがって、例えば RP50＝- 4.21 の場合は、その TRP50を求める場合は、
V= 28 .93＊(-4.21)+224.40 =102.60 となる。下の表では、102.58 となっているが、これは小数点以下の数値の使い方の違いによる値である。

また、difficulty b ＝-4.21 がわかっている時、その RP80 を求めるには、RP θ ＝ ln (P/(1-P)) + b を用いて、 RP80 = ln (0.8/(1- 0.8)) + (- 4.21) = - 2.82 となる。

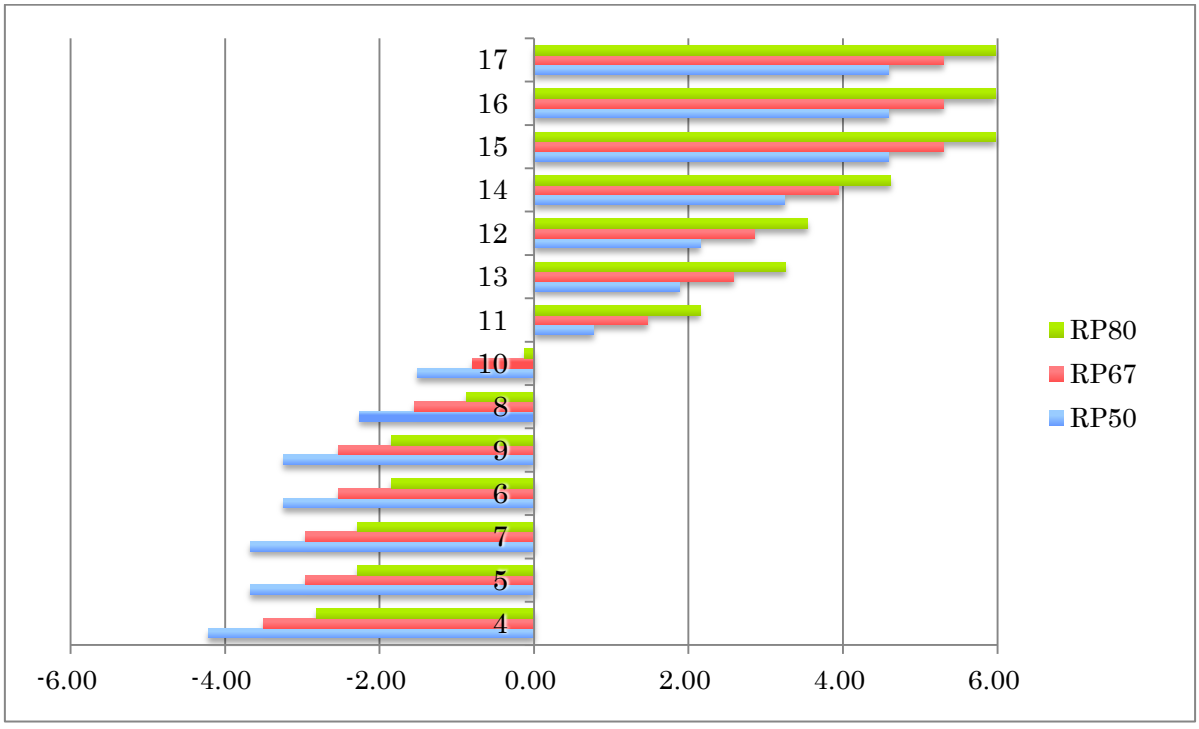
TIN	DIF	RP50	RP67	RP80	TRP50	TRP80
4	-4.21	-4.21	-3.50	-2.82	102.58	142.68
5	-3.67	-3.67	-2.96	-2.28	118.20	158.30
7	-3.67	-3.67	-2.96	-2.28	118.20	158.30
6	-3.24	-3.24	-2.53	-1.85	130.64	170.74
9	-3.24	-3.24	-2.53	-1.85	130.64	170.74
8	-2.26	-2.26	-1.55	-0.87	159.13	199.24
10	-1.51	-1.51	-0.80	-0.12	180.80	220.91
11	0.77	0.77	1.47	2.15	246.53	286.64
13	1.88	1.88	2.58	3.26	278.64	318.75

12	2.15	2.15	2.86	3.54	286.63	326.73
14	3.24	3.24	3.94	4.62	318.02	358.12
15	4.59	4.59	5.30	5.98	357.22	397.32
16	4.59	4.59	5.30	5.98	357.22	397.32
17	4.59	4.59	5.30	5.98	357.22	397.32

試みに、上のデータを用いて、1PLM で求めたRP50、RP67、RP80のグラフを以下に示す。その課題は、さきに示した2PLM の図と同じ傾向にあり、その結果は、きわめて類似していると理解することが可能か、そして、データ処理に関しては、さきに求めた2PLM の手順と類似していると言えるかである。

以下のグラフを見ると、例えば、TIN10 (7) のRP50、RP67、RP80におけるグラフの形は、2PLMと1PLM とでは，きわめて類似していることが解る。2PLM においては、それぞれ、- 0.73,-0.25, 0.21 という数値であるが、1PLM においては，それぞれ、-1.51, -0.80, -0.12 となっている。
それぞれ、RP50、RP67、RP80の間の数値は異なる。しかし、グラフ全体から見ると、2PLM と1PLM で共通するところがある。それは、7番目のデータ（TIN=10）と8番目のデータ（TIN=11）では、これを境にして、グラフが大きく分かれていることが解る。つまり、TIN=10 までは、能力パラメータは負の領域になっているが、TIN=11 では、能力パラメータは正の領域に転換しているということである。

この現象は、2PLM の場合も、1PLM の場合も同じ現象を示していることに注目しなければならない。このことは、分割点・規準の設定に関しては、2PLM でも、1PLM でも，同じように利用できるということであろう。



4．1．2．　3PLM での分析

3PLM：　XCALIBRE Ver. 1.10 (Assessment System Cooperation)

Wright & Stone (1979: 31) KNOX CUBE TEST

3PLM:　P　= c + (1 - c)＊(1 / (1 + exp (-Da (θ - b)))

RP θ = ln ((P/(1-P))＊（1 - c）- c)/(Da) + b

TRP (Transformed Response Probability) = 60.98＊RP θ + 212.81

From ORIGINAL VALUE to TRANSFORMED VALUE

A bit smaller than the smallest RP50　-1.74+ (-0.11)==> - 1.85

A bit larger than the largest RP80　2.　96+ (0.11)==>　3.07

つぎの式を設定する

B＊（- 1.85）+ A = 100

B＊(3.07) + A =400

A = 100 – B＊（- 1.85）

B ＊ 3.07 + 100 – B＊（- 1.85）= 400

B = (400 – 100)/(3.07- (-1.85)) = 300/4.92 = 60.98

A = 100 – 60.98＊(-1.85) = 212.81

V = 60.98＊RP θ + 212.81

したがって、例えば RP50＝- 1.74 の場合は、その TRP50 を求める場合は、V = 60.98＊(-1.74) + 212.81 = 106.70 となる。下の表では、106.89 となっているが、これは小数点以下の数値の使い方の違いによる値である。

また、difficulty b = -1.52, discrimination a = 1.21, guessing c = 0.18 がわっている時、その RP80 を求めるには、RP θ = l n ((P /(1-P))＊(1- c) – c) /(Da) + b を用いて、RP80 = ln ((0.8/(1-0.8))＊(1 - 0.18) – 0.18)/(1.7＊1.21) + (-1.52) = - 0.97 となる。

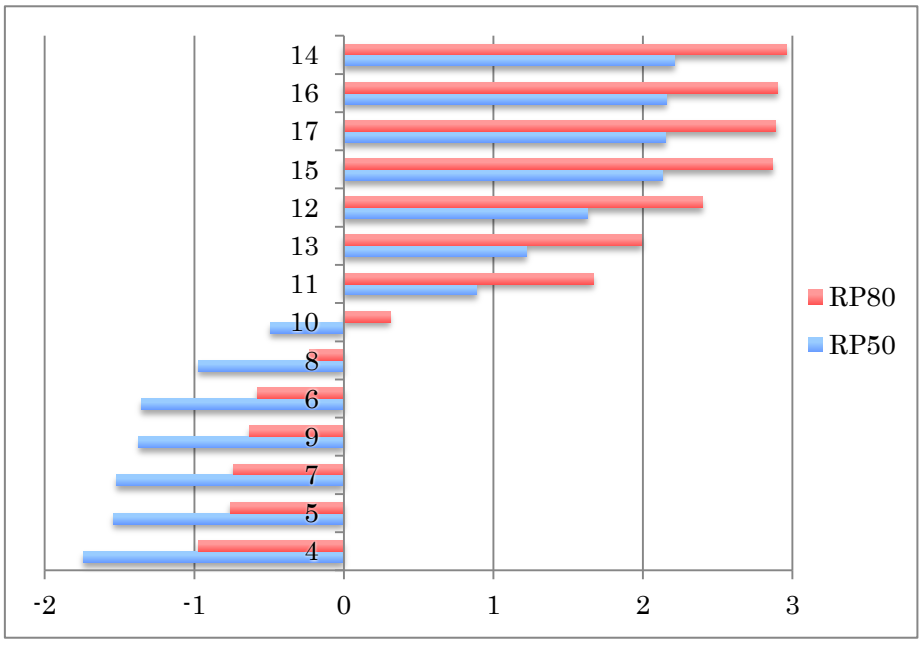
TIN	3b	3a	3c	RP50	RP80	TRP50	TRP80
4	-1.52	1.21	0.18	-1.74	-0.97	106.89	153.66
5	-1.32	1.19	0.18	-1.54	-0.76	118.86	166.42
7	-1.30	1.19	0.18	-1.52	-0.74	120.08	167.64
9	-1.16	1.25	0.18	-1.37	-0.63	129.27	174.54
6	-1.13	1.21	0.18	-1.35	-0.58	130.67	177.44
8	-0.76	1.26	0.18	-0.97	-0.23	153.76	198.67

10	-0.26	1.16	0.18	-0.49	0.31	183.15	231.94
11	1.11	1.18	0.18	0.89	1.67	266.93	314.89
13	1.44	1.22	0.18	1.22	1.99	287.50	333.89
12	1.85	1.21	0.18	1.63	2.40	312.39	359.16
15	2.33	1.24	0.17	2.13	2.87	342.87	388.09
17	2.35	1.24	0.17	2.15	2.89	344.09	389.30
16	2.36	1.24	0.17	2.16	2.90	344.70	389.91
14	2.41	1.22	0.17	2.21	2.96	347.55	393.51

試みに、上のデータを用いて、3PLM で求めたRP50、RP80のグラフをいかに示す。その課題は、さきに示した2PLM の図と同じ傾向にあり、その結果は、きわめて類似していると理解することが可能か、そして、データ処理に関しては、さきに求めた2PLMの手順と類似していると言えるかである。

以下のグラフをみると、例えば、TIN 10 の RP50、RP80におけるグラフの形は、2PLM と 3PLM ではきわめて類似していることが解る。2PLM においては、それぞれ 0.73, 0.21 という数値であるが、3PLM では、それぞれ –0.49, 0.31 となっている。それぞれ, RP50、RP80の間の数値の違いは異なる。しかし、グラフ全体から見ると、それは、7番目のデータ（TIN=10）と8番目のデータ（TIN=11）では、これを境にして、グラフが大きく分かれていることが解る。つまり、TIN=10 までは、能力パラメータは、負の領域になっているが、TIN=11 では、能力パラメータは正の領域に転換しているということである。

この現象は、2PLM の場合も、3PLM の場合も、同じ現象を示していることに注目しなければならない。このことは、分割点・規準の設定に関しては、2PLM でも3PLM でも、同じように利用出来るということであろう。



5. 規準設定の手順:

Bookmark Method による規準設定の手順は、多くの研究者によって示されているが、そのうちの明確な、そして、簡潔なものの一つは、Lissitz (2013:165) で以下のように示されている。

1. Define PLDs (Performance Level Descriptors) and focus on minimal performance levels.
2. Create an ordered item booklet.
3. Present the ordered item booklet and elicit a bookmark for each cut-off.
4. Collect the judgments of each standard setter.
5. Calculate the median judgment for each PLD cut-off.

また、Zieky, Perie, & Livingston (2008:105-118) に於いては, 6．6 Procedures f or the Bookmark Method の中で、Steps 1 -13 まで示されている。さらに、Cizek & Bunch (2007: 180-189)では、ROUND ONE of a Bookmark Procedure として, obtaining preliminary bookmark cut scores, a caveat and caution concerning bookmark cut scores, round one feedback to participants, をあげている。さらに、ROUND TWO of a Bookmark Procedure, ROUND THREE of a Bookmark Procedure を示している。 Bookmark Method には限定しないが、規準設定のため一般的手順として、Hambleton & Pitoniak (2006: 436-439) で示されている TYPICAL STEPS IN SETTING PERFORMANCE STANDARDS は、興味のある記述を以下のように示している。

Step 1: 規準設定法選択する。
Step 2: 審査員と実施計画を決定する
Step 3: 設定する各段階の記述を行う。
Step 4: 規準設定法の使い方について審査員を訓練する。
Step 5: テスト項目の評価を収集する。
Step 6: 収集した評価に対するフィードバックをして検討のための会議を行う。
Step 7: 審査員の評価を集め、規準設定を行う。
Step 8: 審査員の評価を行う。
Step 9: 妥当性を検討、最終記録を 用意する。

こうした手順を踏んで規準設定の最終段階に近づくのであるが、規準設定の最終段階で必要なことは、ある測定を行った後で、その結果は、信頼性や妥当性の高い、適切なものであったかどうかということの検討である。

規準設定の評価として必要な根拠としては、Fulcher (2010:241-243) では、 Kane (1994: 425-461) を取り上げて、次の3つの根拠を論じている。

第1は手順に関する根拠である。これを、procedural evidence としている。そこでは、基準設定は組織的に行われたかという視点である。つまり、審査員は、この方法に関して適切に訓練されていて、自分の考えを自由に述べることができたかということである。ここでは、いわゆる「適正手続」(due process)が重要であることを強調している。

第2は、測定内部の根拠のことである。これを、internal evidence としている. そこでは、手順から到達した結果の一貫性である。 ここでは、例えば、審査員は、自分の評価に関してはどのぐらい自信があると言えるかといった課題である。この点に関しては、審査員の間の評価の一致度は、きわめて重要である。この一致度の測定に関しては、さまざまな方法があるが、ここでは、Cohen's Kappa をとりあげている。

ここで用いられている kappa coefficient というのは、分割表による解析で用いられる測定の一致度の指数である。たたえば、ある測定を2回繰り返して、その再現性、つまり、信頼性があるかどうかを調べたり、審査員 A と審査員 B の評価結果が一致するかどうかを調べる場合等に利用されるものである。信頼性係数とも呼ばれているものである。たとえば、次のようなデータあり、その一致度を検討するとしよう。以下、Bachman（2004: 201 – 202）での例を示すこととする。

		R (RATER) A				
R		Master	N	on-master	M	arginals
	Master	15		2		15+ 2=17
B	Non-master	1	2			2+ 1=3
	Marginals	15+1=16		2+2=4		16+ 4=20

上の表は、Rater A と Rater B の二人の審査員が20名の受験者の評価を目標達成者と目標未達成者に評価した結果を示すものである。

ここでの coefficient kappa は、 k = (Po – Pc) / (1 – Pc) で求めることができるが、Po, Pc はそれぞれ、Po = agreement coefficient , Pc = the proportion of agreement that is due to chance を示すものである。実際のデータを求めてみると、Po = (15+ 2)/20 = 0. 85, Pc = (((15+2) * (15+1))+((2+1) * (2+2)))/20^2 = 284/400 =0.710, K = (0.85 – 0.710)/(1 – 0.710) = 0.483 となる。

一般的に言えることは、Kが0.80以上の場合は、高い一致度（high rates of agreement）, 0.8-0.7の場合は、穏当な一致度（reasonable rates of agreement）, 0.7-0.6の場合は、普通の一致度（moderate rates of agreement）, 0.6以下では要注意（attention）と解釈されるのが普通である。

第3は、測定外部の根拠である。これは、到達と未到達の境界線にある受験者と、他のテスト結果との相関を検討すること等で調べることができる。例えば、2つの規準設定手順での結果を比較するという方法も考えられる. もしも、その2つの手順で結果が一致出来れば、正しい時点（defensible p oint）でその分割点は設定されたと考えることが可能であろう。これは、Livingstone, S.A. & Zieky, M.J. (1982). *Passing Scores: A manual for Setting Standards of Performance on Educational and Occupational Tests.* Educational Testing Service. でも言及されている点である。

また、Hambleton, R .K. (2001:89-116).Setting Performance Standards on Educational Assessments and Criteria for Evaluating the Process. In Gregory J. Cizek (Ed.) *Setting Performance Standards*. Lawrence Erlbaum Associates, Publishers. では、Criteria For Evaluating A Performance Standard-Setting Study として、20の質問に答えるような準備が必要であるとしている点は、注目に値する。

規準設定の最終手順としてこれまで、検討データの信頼性、一貫性を求めてきたが、テストの信頼性と並んで必要なことは、テストの妥当性, である。テストの妥当性に関しては、Kane, M .T. (2006: 17 -64). Validation. In R .L. B rennan (Ed.). *Educational Measurement: Fourth Edition,* American Council on Education and Praeger Publishers でその詳細が述べられているが、Can-Do statements など CEFR の開発等に貢献している O'Sullivan, B . & Weir, C .J. (2011: 13 -32). Test Development and Validation. In B. O'Sullivan (Ed.) *Language Testing: Theories and Practice.* Palgrave Macmillan. の述べている 背景に関する妥当性（context validity），認知的妥当性（cognitive validity），得点に関する妥当性（scoring validity），結果妥当性（consequential validity），規準関連妥当性（criterion-related validity） という5つの視点からの検討も今後の課題として注目しなければならない。

6. むすび

「言語テストの規準設定」を主題として行った委託研究に関して、筆者は、Bookmark Method を取り上げ、その先行研究を整理して、いくつかの課題を投げかけ、その解決のための手段と方法を提供してきた。Bookmark Method は、米国を中心とした開発に引き続き、オランダのテスト研究所 CITO でも修正案が提示され、CEFR の研究課題の一つとなって今日に及んでいる。この研究報告書第3号では、そのうちの CITO Variation を調査・検討してきた。

これまでの Bookmark Method に加えられた視点の最大の特徴は、テスト項目の困難度、弁別力、当て推量等のパラメータの状況は、「受験者の能力」に関連した考察を可能にした点である。困難度等が求められてその項目に対し、50、67、80パーセントの確率で解答出来る

能力は、RP（response probability）として求められる．これに加えて、能力の original scale から transformed scale に変換し、さらに transformed scale から original scale への変換も可能にしているのは大きな特徴である。これが、実際、どのような手段と方法で可能であるかを、Wright and Stone (1979) の KNOX CUBE TEST のデータを用いて、実際に検討しているのがこの研究の大きな特徴でもある。さらに、項目応答理論の活用では、そのデータのサンプル・サイズと PLM の種類の違いが課題にされることが多い。 そこで、少ないサンプル、1PLM、2PLM、3PLM いずれの場合も意味のある差異は示すことなく、普通の規準設定が可能であることを確認している。もちろん、その信頼性と妥当性は、今後の課題であるが、その手順と方法の基本的な検討は一応な段階まで到達していると考えられる。今後の課題は、この基本的テな手順と方法を用いて、実際の膨大なデータに適応するにはどうするかを考えればよい。

わが国を取り巻く英語教育に関する複雑な環境では、そのあるべき姿を捉えることは、かなり困難になってきている．大学入試改革、中教審「生煮え」報告案、わずかに合格ラインに届かなくとも「暫定入試」を認めるという発想は、一点刻みを改め段階評価と関連するのであろうか？古典的テスト理論、項目応答理論、潜在ランク理論等の多角的視点が今日ほど望まれている時期はない。混乱であるが故に、適切なそして妥当なテストの原理に関する理解：language assessment literacy, つまり、an understanding of the principle of sound assessment は、いま、きわめて必要な時なのである。

参考文献

Bachman, L. & Palmer, A. (2010). Procedures for setting cut-scores. In *Language Assessment in Practice* (pp.373-375). Oxford University Press.

Bachman, L. F. (2004). Threshhold loss agreement indices. In *Statistical Analyses for Language Assessment* (pp.199-202).Cambridge University Press.

Brennan, R.L. (Ed.) (2006). *Educational Measurement (*Fourth Edition). American Council on Education and Praeger Publishers.

Brown, J.D. & Hudson, T. (2002). Threshhold-loss agreement methods. In *Criterion-referenced Language Testing*. (pp.169-175). Cambridge University Press.

Cizek, G.J. (Ed.) (2001). *Setting Performance Standards: Concepts, Methods, and Perspectives*. Lawrence Erlbaum Associates, Publishers.

Cizek, G.J., Bunch, M.B., & Koons, H. (2004). Setting Performance Standards: Contemporary Methods, *Educational Measurement: Issues and Practice*, 23, (4). 31-50.

Cizek, G. J. (2006). Standard Setting. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of Test Development* (pp.225-260). Lawrence Erlbaum Associates, Publishers.

Cizek, G.J. & Bunch, M.B. (2007). *Standard Setting: A Guide to Establishing and Evaluating Performance Standards on Tests.* Sage Publications.

Fulcher, G. (2010). 10. Evaluating standard-setting. In *Practical Language Testing* (pp.241-243). Hodder Education.

Hambleton, R.K. (2001) Setting Performance Standards on Educational Assessments and Criteria for Evaluating the Process. In G.J. Cizek (Ed.) *Setting Performance Standards,* Lawrence Erlbaum Associates, Publishers

Hambleton, R.K. & Pitoniak, M.J. (2006). Setting Performance Standards. In R. L. Brennan (ED.) *Educational Measurement: Fourth Edition* (pp.433-470). American Council on Education, Praeger.

Henning, G. (1987). *A Guide to Language Testing. Development, Evaluation, Research,*. Newbury House Publishers.

Kane, M.T. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research* 64, 3. 425-461.

Kane, M.T. (2006). Validation. In Brennan, R.L.(Ed.) *Educational Measurement.* 4th edition . American Council on Education and Praeger. 17-64

Lewis, D.M., Mitzel, H.C., & Green, D.R. (1996). Standard Setting: A Bookmark Approach. In Green, D.R. (Chair), IRT-based standard-setting procedures utilizing behavioral anchoring. Symposium conducted at the Council of Chief State School Officers National Conference on Large-Scale Asssessment, Phonix, AZ.

Lissitz, R. W. (2013). 10 Standard Setting: Past, Present and Perhaps Future. In M. Simon, K. Ercikan, & M. Rousseau (Eds.) *Improving Large-Scale Assessment in Education* (pp.154-174). Routledge Taylor & Francis Group

Luechat, R. M. (2006). Designing Tests for Pass-Fail Decisions Using Item Response Theory. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of Test Development* (pp. 575-596). Lawrence Erlbaum Associates, Publishers.

McCoach, D.B., Rambo, K.E. & Welsh, M. (2012). Issues in the Analysis of Change. In Secolsky, C. & Denison, D.B. (Eds.) *Handbook on Measurement, Assessment, and Evaluation in Higher Education.* Routledge.

Nitko, A. J. (1983). Percent Agreement and Kappa Coefficient. In *Educational Tests and Mesasurement* (pp.406-407). Harcourt Brace Jovanovich, Inc.

O'Sullivan, B. & Weir, C. (2011). Test Development and Validation. In B. O'Sullivan (Ed.) *Language Testing: Theories and Practice* (pp.13-32). Palgrave MaCmillan.

Pitoniak, M.J. & Morgan, D.L. (2012). Setting and Validating Cut Score for Tests. In C. Secolsky & D.B. Denison (eds.) *Handbook on Measurement, Assessment, and Evaluation in Higher*

Education (pp.343-366). Routledge Taylor & Francis Group.

van der Schoot, F. (2009). CITO Variation on the Bookmark Method. In Council of Europe. *Reference Supplement to the Manual for Relating Language Examinations to the Common European Framework of Reference for Languages: learning, teaching, assessment.* Language Policy Division, Strasbourg.

Wright, B.D. & Stone, M.H. (1979). *Best Test Design.* MESA Press.

Zieky, M.J., Perie, M. & Livingstone, S. A. (2008). 7.4. Evaluate the Cutscores. In *Cutscores: A Manual for Setting Standards of Performance on Educational and Occupational Tests* (pp.163-168). Educational Testing Service.

法月　健(2013)「Rasch Model とLRT を併用した分割点設定法」. 大友賢二「英語教育とテスト:第二言語習得に於ける基準設定をめぐって」、成蹊大学.

野口裕之訳(1992). 項目応答理論の基礎と応用. 池田、藤田、柳井、繁桝編『教育測定学原著第3版上巻』(ロバート・L・リン編). みくに出版

大友賢二(2013a)「予備調査:CITO Variation on the Bookmark Method」.『言語テストの規準設定　報告書(第2号)』、日本英語検定協会　英語教育研究センター

大友賢二(2013b)「英語教育とテスト:第二言語習得に於ける基準設定をめぐって」『第7回日本テスト学会賞記念講演会』 成蹊大学.

荘島宏次郎(2010)「ニューラルテスト理論:学力を段階評価するための潜在ランク理論」. 植野真臣・荘島宏二郎『学習評価の新潮流』. 朝倉書店.

"Can-do statements" の比較・研究−Ⅱ

Comparative studies on practices of Can-do statements Ⅱ

伊東祐郎

Sukero Ito

Abstract

This paper reviews Can-do lists of the EIKEN tests in comparison with the Can-Do Statements (CDS) of ALTE (the Association of Language Testers in Europe) .The EIKEN tests are well known as a high-stake test of English in Japan, which are a seven-level set of tests. The seven levels of EIKEN are designated as “grades,” and range from Grade 5 (beginner) to Grade 1 (advanced), with two bridging levels (Grades Pre-1 and Pre-2).

The Can-do list was published and provides CDS describing the ability to use English in each of the four skill areas (reading, writing, listening, speaking) for each of the seven EIKEN grades. It should be noted that CDS was designed to elicit information from test takers about what they believe they can do in English outside the testing situation. Therefore being able to perform language activities included on the list for a particular grade does not guarantee that a person would be able to use English properly or to pass the grade of the EIKEN tests. The primary aim of the EIKEN Can-do list to help test users gain a better understanding of the grade or levels of language ability targeted by the EIKEN tests, and also aims to contribute to a better understanding of typical language learners in Japan.

This paper mainly examines the descriptions of Grade 1 and Grade 4 of proficiency provided by the Can-do list above, and tries to analyze the structures and functions taken into the CDS. Differences in descriptors of tasks in each level were investigated. Central to the study is the use of a taxonomy based on Bloom’s Taxonomy of characterizing performance tasks which were described in CDS.

１．問題と目的

2012 年度の研究において、Can-Do Statements（CDS）作成の際の課題として認知的負担度と言語能力の難易度に言及した。

言語運用能力は、認知的負担度と言語形式や語彙と深い関係があると言われている。下位級の認知的負担度が低いレベルは、「馴染みのあること」「よくわかっていること」が対象になり、必然的に頭を使う必要が低くなる。一方、上位級になるにしたがって、「不慣れな状況」「社会性の高い話題」「抽象的なテーマ」などが対象となっていて、物事の分析力、知識の統合力、判断力など高度な思考力が必要となる。これらは言語活動におけるタスクと密接にかかわるものであり、タスクを分析の対象にすることによって、認知的負担度、言語運用力の発達段階が推察できると考えられる。

本稿では、英検が公開している出題の基準や範囲となっている試験問題の内容を参考にしながら、CDS の記述文の分析を試みる。その際に、認知的負担度と言語能力の難易度を CDS の表記から考察し、記述文の構造について考察する。

本稿の構成は以下のようになっている。

１）英検が測定しようとしている知識や能力の明確化（級別の能力規準の明確化）
２）現在のテストの言語運用場面と言語運用領域の明示化
３）英検が提示している言語能力記述文（「Can-do リスト」）の構造分析
４）英検の「Can-do リスト」と他の CDS との比較・分析
５）Standard setting における英検の役割・機能の考察

を目的として本論を進める。

２.「英検」が測定しようとしている知識や能力：「各級の審査基準」から

英検のウェブサイトを見ると、「英検」が測定しようとしている知識や言語能力は、7 レベルに分けられて明示されている。このことから、英検では、言語能力の総体を段階的に 7 分割し、それぞれに対応するテストを実施している。ウェブで公開されている「各級の審査基準」を見ると、上位級から「１級」「準１級」「２級」「準２級」「３級」「４級」「５級」と７段階から構成される。

ではこれらのレベルの違い、すなわち言語能力の発達段階はどのような視点から記述されているのだろうか。また、「言語行動」という観点からどのような構成でまとめられているのだろうか。最初に「各級の審査基準」を参考に、各級の言語能力の違いを明示していると思われる、鍵となる言葉を選び出してみる。すると、１級、準１級、２級では、「社会生活」、準２級では「日常生活」、３級では「身近なこと」、４級では「簡単な」、５級では「初歩的」という表現が使われていることがわかる。しかしながら、この段階にお

いては、上位級間の能力差の違いはわかりにくい。

各級の審査基準

レベル	基　　準
１級	広く社会生活で求められる英語を十分に理解し、また使用することができる。
準１級	社会生活で求められる英語を十分理解し、また使用することができる。
２級	社会生活に必要な英語を理解し、また使用することができる。
準２級	日常生活に必要な英語を理解し、また使用することができる。
３級	身近な英語を理解し、また使用することができる。
４級	簡単な英語を理解することができ、またそれを使って表現することができる。
５級	初歩的な英語を理解することができ、またそれを使って表現することができる。

さらに、審査領域として「読む」「聞く」「話す」「書く」の４技能を構成し、それぞれの能力基準が明示されている。それらから、タスクの難易度に影響を与える分野・内容・話題の記述を取り出してみると、以下のような特徴のあることがわかる。

１級：社会性の高い幅広い（分野・内容・話題）
準１級：社会性の高い（分野・内容・話題）
２級：社会性のある（内容・話題）
準２級：日常生活の話題
３級：身近なことに関する内容
４級：簡単な内容
５級：初歩的な語句や文

以上のことから、英検では取り上げる内容や話題については、「社会性」「日常生活」「身近なこと」という観点からレベルの大枠をとらえられていることがわかる。この段階での上位級間の違いは、「社会性」について「高い」「幅広い」「ある」によって解釈することになる。では、「社会性」「日常生活」「身近なこと」とは具体的にどのような事象を指すのだろうか。さらに探求してみたい。ここで、今後の表現を「社会性」はそのままの表現で、「日常生活」を「日常性」、「身近なこと」を「自分性」として扱うことにする。

３．英検の構造：「英検で求められる能力と検定形式」から

英検のウェブのホームページでは、各級別に「求められる能力と検定形式」欄にて、出題内容や出題形式が公開されている。ここでは、１級と４級について紹介し、その後、具体的にどのような問題が出題されているか考察してみる。

【１級】一次試験:「筆記（100 分）」「リスニング（約 30 分）」

求められる主な能力	形式･課題	形式･課題詳細	問題数	問題文の種類	解答形式
語彙力	短文の語句空所補充	文脈に合う適切な語句を補う。	25	短文 会話文	4肢選択（選択肢印刷）
読解力	長文の語句空所補充	パッセージの空所に文脈に合う適切な語句を補う。	6	説明文 評論文など	
	長文の内容一致選択	パッセージの内容に関する質問に答える。	10		
作文力	英作文	指定されたトピックについての英作文を書く。	1	（英作文なので問題文はない）	記述式
聴解力	会話の内容一致選択	会話の内容に関する質問に答える。（放送回数1回）	10	会話文	4肢選択（選択肢印刷）
	文の内容一致選択	パッセージの内容に関する質問に答える。（放送回数1回）	10	説明文など	
	Real-Life形式の内容一致選択	Real-Life 形式の放送内容に関する質問に答える。（放送回数1回）	5	アナウンスなど	
	インタビューの内容一致選択	インタビューの内容に関する質問に答える。（放送回数1回）	2	インタビュー	

【１級】 主な場面・題材	
場面・状況	家庭、学校、職場、地域（各種店舗・公共施設を含む）、電話、アナウンス、講義など
話　　題	社会生活一般、芸術、文化、歴史、教育、科学、自然・環境、医療、テクノロジー、ビジネス、政治など

【４級】一次試験：「筆記（35 分）」「リスニング（約 30 分）」

求められる主な能力	形式･課題	形式･課題詳細	問題数	問題文の種類	解答形式
語彙力	短文の語句空所補充	文脈に合う適切な語句を補う。	15	短文 会話文	4肢選択（選択肢

読解力	会話文の文空所補充	会話文の空所に適切な文や語句を補う。	5	会話文	印刷）
	長文の内容一致選択	パッセージの内容に関する質問に答える。	10	掲示・案内 Ｅメール（手紙文）説明文	
作文力	日本文付き短文の語句整序	日本文を読み、その意味に合うように与えられた語句を並び替える	5	短文	
聴解力	会話の応答文選択	会話の最後の発話に対する応答として最も適切なものを補う。（放送回数2回、補助イラスト付き）	10	会話文	3肢選択（選択肢読み上げ）
	会話の内容一致選択	会話の内容に関する質問に答える。（放送回数2回）	10		4肢選択（選択肢印刷）
	文の内容一致選択	短いパッセージの内容に関する質問に答える。（放送回数2回）	10	物語文 説明文	

【４級】主な場面・題材	
場面・状況	家庭、学校、地域（各種店舗・公共施設を含む）、電話、アナウンスなど
話　　題	家族、友達、学校、趣味、旅行、買い物、スポーツ、映画、音楽、食事、天気、道案内、自己紹介、休日の予定、近況報告、海外の文化、人物紹介、歴史など

第一次試験は紙筆試験と　　試験から構成される。紙筆試験によって、「語彙力」「読解力」「作文力」が測定される。１級と４級の「語彙力」「読解力」は４肢選択形式で、「作文力」については、１級では記述式であるが、４級では語句の並べ替えで、記述式ではない。

４．言語運用場面と言語運用領域

ALTE（The Association of Language Testers in Europe）や CEFR（Common European Framework of Reference for Languages）で明示されているコミュニケーション能力の枠を概観すると、ALTE では、広範囲の言語運用場面を職業や勉学、生活といったそれぞれの場面に応じて、社会一般（social）、仕事（work）、勉学（study）の 3 つの領域で言語運用場面を規定している。テスト理論でいうところの目標言語使用領域（TLU＝target language use）である（Bachman & Palmer,1996）。言語使用領域は無限大であり、ALTE のように領域を限定しなければ能力基準の枠作りで苦心することになる。英検の場合は、「社会性」という言葉でひとくくりになっているので、この段階での言語運用場面の特定はむ

ずかしい。「社会一般」という幅広い目標言語使用領域をあらかじめ意図した結果の表れであると推察される。また、言語能力の７段階化については、連続性を反映しつつ、低いレベルから上位レベルに能力の発達段階を示す必要がある。その高度化の記述が、先に述べたように、下位級の「初歩的」「簡単」なレベルから「日常生活」を経て、「社会性の高い」という表現で記述されているが、特に４級と５級という入門期のレベルが、「初歩的」「簡単」が多用され、言語行動の視点からの記述が少ないことがわかる。

一方、言語運用領域については、一般的には、「聞く」「話す」「読む」「書く」の４技能（４領域）が挙げられる。英検は一次試験と二次試験において、伝統的な４技能を測定の目的としている。伝統的な４技能と述べたのは、CEFR では、「話す」をその能力の特徴から「Spoken Interaction（会話／対話）」と「Spoken Production（独話）」とに分け、５領域で構成しているからである。ただし、下位級である「４級」と「５級」においては、口頭試験は設定されていない。また、二次試験の口頭試験は、「面接形式のスピーキング」テストが実施される。

５．英検の出題にかかわる場面・状況・話題

次に、ウェブの各級の「求められる能力と検定形式」のところでは、各級が取り扱う[8]ような場面・題材」が明示されている。７レベルの出題内容や出題対象領域にかかわる類似性や相違性などの特徴を知るために、一覧にまとめてみた。次の表は「英検全級の場面・状況、話題一覧」はその結果である。これによって、「社会性」「日常性」「自分性」を構成する項目が具体的に示されることになる。あわせて、級別の違いや特徴がわかる。

下記の一覧表を分析してみると以下のことが読み取れる。

１）「１級」と「準１級」の類似性が高い
２）「２級」と「準２級」の類似性が高い
３）「３級」と「４級」と「５級」の類似性が高い。
４）「講義」「社会生活一般」「芸術」「文化」「政治」は１級と準１級に特化されている。
５）「医療」「テクノロジー」「ビジネス」は、上位級（「１級」～「２級」）に限られる。
６）「教育」「科学」「自然・環境」は、「１級」～「準２級」に限られる。
７）「歴史」については、「４級」と「５級」では対象となっていない。
８）「仕事」については、「２級」のみで、他の級では対象となっていない。
９）「人物紹介」については、「準２級」と「３級」に限られる。
10）「２級」と「準2級」は、比較的幅広い分野・領域を扱っている。
11）「３級」～「５級」で扱われている分野・領域は、「日常生活」や「自分自身」のことが対象となっている。

英検全級の場面・状況、話題一覧

		1級	準1級	2級	準2級	3級	4級	5級
場面・状況	家庭	○	○	○	○	○	○	○
	学校	○	○	○	○	○	○	○
	職場	○	○	○	○	－	－	－
	地域	○	○	○	○	○	○	○
	電話	○	○	○	○	○	○	○
	アナウンス	○	○	○	○	○	○	－
	講義	○	○	－	－	－	－	－
話　題	社会生活一般	○	○	－	－	－	－	－
	芸術	○	○	－	－	－	－	－
	文化	○	○	－	－	－	－	－
	歴史	○	○	○	○	○	－	－
	教育	○	○	○	○	－	－	－
	科学	○	○	○	○	－	－	－
	自然・環境	○	○	○	○	－	－	－
	医療	○	○	○	－	－	－	－
	テクノロジー	○	○	○	－	－	－	－
	ビジネス	○	○	○	－	－	－	－
	政治	○	○	－	－	－	－	－
	学校	－	－	○	○	○	○	○
	仕事	－	－	○	－	－	－	－
	趣味	－	－	○	○	○	○	○
	旅行	－	－	○	○	○	○	○
	買い物	－	－	○	○	○	○	○
	スポーツ	－	－	○	○	○	○	○
	映画	－	－	○	○	○	○	○
	音楽	－	－	○	○	○	○	○
	食事	－	－	○	○	○	○	○
	天気	－	－	○	○	○	○	○
	道案内	－	－	○	○	○	○	○
	海外の文化	－	－	○	○	○	○	－
	人物紹介	－	－	－	○	○	－	－
	家族	－	－	－	－	○	○	○
	友達	－	－	－	－	○	○	○
	自己紹介	－	－	－	－	○	○	○
	休日の予定	－	－	－	－	○	○	○
	近況報告	－	－	－	－	○	○	○

以上のことから、上位級における「社会性の高い」話題というものが、具体的にどのような範疇から出題されているか理解できよう。そして、中位級における「社会生活」「日常生活」の範疇が、そして下位級の「身近な話題」が具体的に何を対象としているか理解できる。

あわせて、ホームページで紹介されている「各級の目安」を見ると、各級の出題内容は概ね、日本の学校制度で英語を学ぶ年齢や学年を基準とした目安となっていることがわかる。ある意味では、日本で使用されている教科書の語彙、文法、表現、場面、状況を基にテストの出題内容が決定される、目標基準準拠テストとして特徴付けられる。

各級の目安

習得目標	級	推奨目安	出題目安
リーダーの英語、4技能総合力	1級	大学上級程度	相手に伝える発信力と対応力 世界で活躍できる人材の英語力
	準1級	大学中級程度	実際に使える英語力
使える英語、海外留学、履歴書	2級	高校卒業程度	海外留学・国内での入試優遇、単位認定など コミュニケーション力
	準2級	高校中級程度	教育や科学などを題材、入試対策にも最適
使える英語の登竜門基礎力定着高校入試レベル	3級	中学卒業程度	海外の文化などが加わる
	4級	中学中級程度	身近なトピック
	5級	中学初級程度	家族のこと、趣味やスポーツなど身近な話題

６．英検－CEFRの研究プロジェクト

実は、英検は2007年度に英検-CEFRに関する研究プロジェクトを発足させ、２年間の調査を行っている（ホームページ）。このプロジェクトの目的は、英検とCEFRとの関連性を探り、CEFRに対する英検のレベルの解釈を容易にすることにあった。ここでは紙面の関係で詳細を省略するが、下表は、英検各級の内容とCEFRの各レベルを比較した結果の英検とCEFRとの対応表である。この報告書では、後述する英検のCan-doリストも参考資料として活用したとある。英検の３級～５級の３段階がA１レベルに相当することについては、学校教育や初期段階の言語教育では、身近な目標が必要となるところから、CEFRの６段階はレベルが広く、教育的な動機付けなどに混乱をきたすという説明があり、英検が中学校の指導要領や教室活動を考慮して目標設定されてあることが述べられている。英検の出題内容について検討するには、このような背景を理解することは重要である。なぜなら、英検の理念にもかかわることであるからである。

英検とCEFRの対応表

CEFR	英検
C２	－
C１	１級
B２	準１級
B１	２級

A２	準２級
A１	３級 ４級 ５級

７． CDS の記述：能力発達段階の観点の分類から

言語能力にかかわる発達段階の観点の分類に関しては、和田（2004）が、CEFR の CDS を詳細に調査・分析しどのような観点から表記されているかを報告している。分類の観点として、ポイントとなる２つの視点を挙げている。一つは「言語（形式）」から記述されている点であり、もう一つは、「内容」から記述されている点である。

和田（2004）の分析では、CEFR の能力の記述においては、各レベルで言語が使用される状況において想定される言語活動が記述され、その段階付けを下記の分類に基づいて行っているとしている。

能力発達段階の観点の分類（和田（2004））

言語面に焦点を当てると、「正確さ」「流暢さ」「繰り返し」とか「ポーズ」「即興性」「長さ」「速さ」などが、報告書にまとめられている。それらに加えて、「複雑さ」「多様さ」「明確さ」などが上げられている。このような点は、言語能力の質的観点からの記述とみることができる。発話や作文の内容面というより、言語能力のなかの、文法能力にかかわる点であると考えてもよいだろう。詳細については、後述部分を参照されたい。

内容面とは、まず「場面」「話題」にかかわることである。話題というのは本人にとって「なじみ」が有るか無いか、そして具体性の高い事項であるのか抽象性の高い内容であるのか、そして「日常的なこと」なのか否か、また、興味関心にもかかわることである。

そして、言語の「機能」についてもかかわっている。「機能」とは、何のために言葉を使うかに関係するものである。そして「媒体」が関与するとしている。何かが読めると言った場合、読む対象が新聞なのか、また新聞に入ってくる折り込みチラシなのか、あるいは学術書なのかという、何を通してその読解という行為をしているかという具体物を指すことになる。その他として既有の「知識」が挙げられる。

８． 英検における CDS の記述内容の分析（CDS の技能別特徴の考察）

英検の CDS は、Can-do リストとして公開されている（英検ホームページ）。このリストは、2003 年 5 月から約 3 年の歳月をかけ、延べ 20,000 人を超える 1 級から 5 級の合格者（合格直後）に対し、数回に渡る大規模アンケート調査を行って、「具体的にどのようなことができる可能性があるか」ということを各試験の実施団体が調査し、リスト化したものである（英検ホームページ）。回答者である合格者が自信の高いものを精選したもので、言語運用力の発達段階について、１級から５級までの７段階で具体的に把握することができる。もちろん、該当級合格者全員が「必ずできる」ということを保証するものではないと断っている点には留意する必要はある。

この Can-do リストは、英検における言語能力観を知る上で、また、試験内容を予測する際に参考になると思われる。 以下に、上記の「能力発達段階の観点の分類」に基づいて、英検１級と４級の Can-do リストを分析してみる。

９． 英検１級「Can-do リスト」の分析

１級

＜読む＞社会性の高い幅広い分野の文章を理解することができる。

・雑誌の	→媒体
社会的、経済的、文化的な記事を	→話題
理解することができる。	→機能（Time/Newsweek など）
・文学作品を	→媒体
理解することができる。（小説など）	→機能
・資料や年鑑などを読んで、	→媒体
必要な情報を	→話題
得ることができる。	→機能（報告書、統計的な資料など）
・留学や海外滞在などの手続きに必要な	→話題
書類を	→媒体
理解することができる。	→機能

＜聞く＞社会性の高い幅広い内容を理解することができる。

・幅広い話題に関する　→話題
まとまりのある話を　→複雑さ
理解することができる。　→機能（一般教養的な講演や講義など）
・社会的な話題に関する　→話題
話を　→媒体
理解することができる。　→機能（環境問題に関する講演など）
・会議に参加して、　→場面
その内容を　→話題
理解することができる。　→機能（イベントの打合せ、会社のミーティングなど）
・テレビやラジオの　→媒体
政治・経済的な　→話題
ニュースを　→複雑さ
理解することができる。　→機能
・いろいろな種類のドラマや映画の　→媒体
内容を　→話題
理解することができる。　→機能

＜話す＞社会性の高い幅広い話題についてやりとりをすることができる。

・社会的な話題や時事問題について、　→話題
質問したり　→機能
自分の考えを述べたりすることができる。→機能
・会議に参加して　→場面
やりとりをすることができる。　→機能（イベントの打合せ、会社のミーティングなど）
・幅広い内容について、　→話題
電話で　→媒体
交渉することができる。　→機能（予定の変更、値段の交渉など）
・相手の状況に応じて、　→話題
丁寧な表現やくだけた表現を　→複雑さ
使い分けることができる。　→機能

＜書く＞社会性の高い話題についてまとまりのある文章を書くことができる。

・社会的な話題について　→話題
自分の意見を　→知識？
まとまりのある文章で　→複雑さ
書くことができる。　→機能（環境問題に関してなど）
・自分の仕事や調査について、　→話題
まとまりのある文章を　→複雑さ

書くことができる。	→機能（レポート、報告書、仕事のマニュアルなど）
・商品やサービスについて、	→話題
苦情を申し立てる文章を	→複雑さ
書くことができる。	→機能（商品の故障、サービスの内容など）
・社会的な話題に関する	→話題
雑誌記事や新聞記事の	→媒体
要約を	→複雑さ
書くことができる。	→機能（社説や論文など）
・講義や会議の	→場面／話題
要点の	→複雑さ
メモをとることができる。	→機能

9.1　1級「読む」

「読む」に関しては、やはり読解内容のジャンル、ここでは、媒体が取り上げられている。雑誌、文学、報告書、資料など、幅広い分野をカバーしている。話題については社会的、経済的、文化的な記事、話題、資料としての年鑑など統計的資料も含まれている。海外留学などの手続き書類が 1 級に含まれているのが興味深い。

9.2　1級「聞く」

「聞く」については、媒体は「テレビ」「ラジオ」「映画」などのメディアが上げられている。話題については広範囲を扱っている。政治・経済的なニュースが具体的な記述にとどまる。「会議」が場面として取り上げられている点で、会議は認知的レベルの高い行為として位置づけられていると推察できる。

9.3　1級「話す」

「話す」については、話題として社会的な代表である時事問題、会議での話題、あとは、相手に応じてと広範囲な話題対応力が上げられる。CEFR のように、spoken inter action と spoken production と分けられていないので、スピーチや講演など一方的に話す独話としての能力については触れていない。電話で交渉する場面を取り上げているところが興味深い。

CEFR では、複雑さや流暢さ、そして運用上の方略、これはストラテジーにかかわる事項であるが、これらについての言及があるが、英検については皆無である。

9.4　1級「書く」

「書く」は、長文を書くことが要求される行為となっている。意見書、レポート、報告書、説明文、要望書、要約など「媒体」が課題として挙げられている。長文は、何を書くかによって構成やそこで使用される語彙、表現なども異なるので、使い分けができるかどうかメタ認知的知識も求められるレベルであると言えよう。日本人であっても日本語で書けそうにないような高度な記述となっている。作文の課題が言語形式や内容と密接に関係していることがわかる。

10. 英検 4 級「Can-do リスト」の分析

4級

＜読む＞簡単な文章や表示・掲示を理解することができる。

・短い手紙（Eメール）を　→媒体
理解することができる。　→機能
（家族の紹介、旅行の思い出など）　→話題

・イラストや写真のついた　→複雑さ？媒体？　（子供向けの絵本など）
簡単な物語を　→話題
理解することができる。　→機能

・日常生活の身近なことを　→話題
表す文を　→媒体
理解することができる。　→機能
（例：Ken went to the park and played soccer with his friends.）

・公共の施設などにある　→場面
簡単な　→複雑さ　（例：No Smoking/Closed/No Dogs ）
表示・掲示を　→媒体
理解することができる。　→機能

・簡単な　→複雑さ
英語のメニューを　→媒体
理解することができる。　→機能
（ファーストフード・レストランにあるメニューなど）→場面

・パーティーなどの　→場面
招待状の内容を　→媒体／話題　（日時、場所など）
理解することができる。　→機能

＜聞く＞簡単な文や指示を理解することができる。

・簡単な　→複雑さ
自己紹介を聞いて、　→話題　（名前、住んでいるところ、家族など）

その内容を理解することができる。　→機能

・簡単な　→複雑さ

文を聞いて、　→媒体

その内容を理解することができる。　→機能　（例：I like dogs , bu t she likes cats .）

・簡単な　→複雑さ

指示を聞いて、　→話題

その意味を理解することができる。　→機能（例：Open your textbook./Close the door, please.）

・人や物の位置を聞いて、　→話題

理解することができる。　→機能（例：The book in on the TV .）

＜話す＞簡単な文を使って話したり、質問をしたりすることができる。

・簡単な　→複雑さ

自己紹介をすることができる。　→機能

（名前、住んでいるところ、家族など）→話題

・簡単な　→複雑さ

質問をすることができる。　→機能

（時刻、好きなもの、相手の名前など）→話題

・相手の言うことがわからないときに、　→複雑さ

聞き返すことができる。　→機能（例：Pardon?/ Could you speak more slowly?）

・日付や曜日を　→話題／複雑さ

言うことができる。　→機能

＜書く＞簡単な文やメモを書くことができる。

・短い文であれば、　→複雑さ

英語の語順で書くことができる。　→機能（例：I went to the park yesterday.）

・語句を並べて　→複雑さ

短いメモを　→話題

書くことができる。　→機能（例：birthday party at 6 p.m.）

・文と文を接続詞（and/but/so/when/becauseなど）でつなげて　→複雑さ

書くことができる。　→機能

・日付や曜日を　→話題／複雑さ

書くことができる。　→機能

10.1　４級「読む」

「読む」については、ジャンルは限╟れている。「日常生活の身近なことを表す文」については、英語の例文は出ているが、場面が不明である。それに対してメニューは具体的で

わかりやすい。複雑さの程度は「短い」「簡単」で明示されている。パーティーなどの日時や場所は単語レベルの易しさとして位置づけられている。

10.2　４級「聞く」

「聞く」に関しては、指示、自己紹介、など文脈によって状況がわかるもので複雑さ（容易さ）を明示している。話題についても「今／ここ」と文脈依存によるものに限られている。

10.3　４級「話す」

「話す」については、「聞く」同様に文脈依存による話題である。日付や曜日など単語レベルの発話力、単文レベルの質問など言語的側面で記述されている。自分自身について語る自己紹介などの課題で複雑さを明示している。

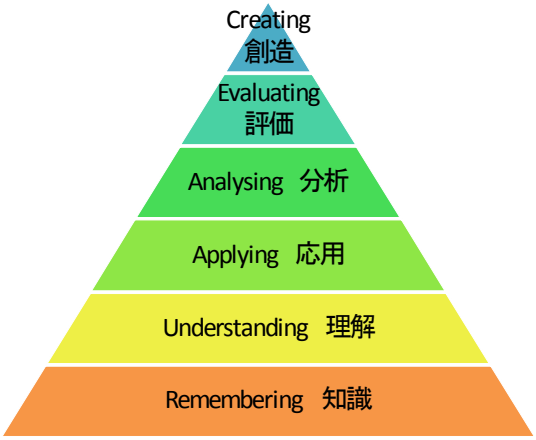
10.4　４級「書く」

「書く」については、話題としてはメモとなっている程度である。複雑さは単語レベル、単文レベルである。

11.　認知的負担度（Bloom's Taxonomy）と CDS との比較

ブルーム（1956）が“Taxonomy of educational objectives”の中で提唱した「教育目標のタキソノミー（分類学）」は、教育目標の能力面を階層的に整理したものである。ブルームは、教育目標（＝授業目標）を３次元、すなわち、①認知的領域（cognitive domain）、②情意的領域（affective domain）、③精神運動的領域（psychomotor domain）の３領域から構成されるとしている。ここでは言語能力の関係から、①に焦点をあてて考察することにする。

認知的領域（cognitive domain）とは、組織的原理は思考力操作の複雑化と捉えることができる。図の上位のカテゴリーは下位のカテゴリーより複雑で、抽象的あるいは内在化された能力となっている。認知活動は、知識→理解→応用→分析→評価→創造というかたちで高次化していくことがわかる。各段階の内容については以下に概説するが、言語能力を段階的に記述する場合、認知的領域がどのように関わっているか考察してみる。



本論では、上記の 6 つの認知活動の特徴を記した後に、ALTE の CDS（聞く／話す／読

む／書く）を提示し、CDS と認知的負担度の関係を概観してみたい。あわせて、英検の「１級」と「４級」の Can-do リストから能力記述文を追記で掲載してみる。その際に、「１級」を上位の「創造」に、「４級」を下位の「知識」に配置し、記述の仕方や表現の分析を試みる。断っておくが、ALTE の 6 レベルのブルームの Taxonomy への配置、ならびに、「１級」＝「創造」、「４級」＝「知識」の対応は、著者の独断によるもので、今回の比較のために配置したものである。なお、以下に続く記述は、下位級から上位級の順になっている。また、能力の難易度を示す用語に下線を引いてあるが筆者によるものであることも記しておきたい。

12. 英検と ALTE における CDS 比較

【1 Remembering 知識】：客観的な知識・情報を暗記したり、記憶したりして、必要に応じて想起できるレベル。単語や文字、文法規則の暗記に相当する言語活動。

	聞くこと／話すこと	読むこと	書くこと
ALTE	基本的な説明・指示を理解し、またはありきたりの話題に関する基本的で事実に基づく会話に参加することができる。	基本的な掲示、説明・指示、または情報を理解することができる。	基本的な用紙に記載し、時間、日付、場所を含むメモを書くことができる。
英検４級CAN・DOリスト	４級＜聞く＞簡単な文や指示を理解することができる。 ・簡単な自己紹介を聞いて、その内容を理解することができる。（名前、住んでいるところ、家族など） ・簡単な文を聞いて、その内容を理解することができる。（例：I like dogs, but she likes cats. ） ・簡単な指示を聞いて、その意味を理解することができる。（例：Open your textbook./Close the door, please.） ・人や物の位置を聞いて、理解することができる。（例：The book in on the TV. ） ４級＜話す＞簡単な文を使って話したり、質問をすることができる。	４級＜読む＞簡単な文章や表示・掲示を理解することができる。 ・短い手紙（Eメール）を理解することができる。（家族の紹介、旅行の思い出など） ・イラストや写真のついた簡単な物語を理解することができる。（子供向けの絵本など） ・日常生活の身近なことを表す文を理解することができる。（例：Ken went to the park and played soccer with his friends.） ・公共の施設などにある簡単な表示・掲示を理解することができる。（例：No Smoking/Closed/No Dogs） ・簡単な英語のメニューを理解することができる。（ファーストフード	４級＜書く＞簡単な文やメモを書くことができる。 ・短い文であれば、英語の語順で書くことができる。（例：I went to the park yesterday.） ・語句を並べて短いメモを書くことができる。（例：birthday party at 6 p.m.） ・文と文を接続詞（and/but/so/when/becauseなど）でつなげて書くことができる。 ・日付や曜日を書くことができる[9]

	・簡単な自己紹介をすることができる。（名前、住んでいるところ、家族など） ・簡単な質問をすることができる。（時刻、好きなもの、相手の名前など） ・相手の言うことがわからないときに、聞き返すことができる。（例：Pardon?/ Could you speak more slowly?） ・日付や曜日を言うことができる。	・レストランにあるメニューなど） ・パーティーなどの招待状の内容を理解することができる。（日時、場所など）	

（出典：*Common European Framework of Reference for Languages: Learning, teaching, assessment.* 国際交流基金による翻訳版、以下出典同じ）

・ALTE の CDS では、最下位のレベルを「ありきたり」「基本的」「事実に基づく」「時間」「日時」などの表現によって、認知的負担度の低さを示している。
・「英検」では、「簡単な」「短い」「身近な」で負担度を示している。ALTE 同様に、「日付」「曜日」が単語レベルで読み書きできるレベルを明示している。習ったばかりの言語知識に依存したレベルを記述化している。

【2 Understanding 理解】：客観的な知識・情報の内容や論理の展開を把握して、必要に応じて知識を活用できるレベル。音声や文字で入手した知識や情報を理解、解釈する言語活動。

	聞くこと／話すこと	読むこと	書くこと
ALTE	慣れた環境の中で、単純な意見や要求を表現することができる。	周知の範囲内で率直に書かれた情報、たとえば製品に関する情報や、標示、簡単なテキストブック、またはよく知っている事柄に関するレポートを理解することができる。	用紙に記載し、個人情報に関係する短い簡単な手紙やハガキを書くことができる。

・ALTE では、「慣れた」「単純な」「周知の範囲内」「よく知っている」「個人情報」「短い」「簡単な」から認知的負担度を示していることがわかる。この段階では、意見や要望を表現するという自発性を示す表現がある。また、「標示」「レポート」「手紙」「ハガキ」など具体的な媒体や場面において、言語活動が可能なレベルとして明示している。

【3 Applying 応用】：学習した基本的な知識・理論・情報を活用して、与えられた新たな

応用問題を解決できるレベル。既習の言語知識や情報を他の場面や状況で応用することができる言語活動。

	聞くこと／話すこと	読むこと	書くこと
A L T E	限られた方法で抽象的・文化的な事柄について意見を述べ、あるいは周知の範囲内で助言をし、説明・指示や公示を理解することができる。	日常的な情報や記事を理解し、精通している分野内の非日常的な情報について全般的な意味を理解することができる。	よく知っている事柄またはありきたりの事柄について、手紙を書きメモを取ることができる。

・ALTE では、「限られた」「周知の範囲内」「日常的」「精通している」「よく知っている」「ありきたり」から認知的負担度を示している。「抽象的」「文化的」「記事」「手紙」から、言語行動としては、ある程度現実社会で対応できるレベルを明示している。

【4 Analyzing 分析】：問題の状況や観察した事象を『複数の構成要素』に分けて、その傾向・特徴・確率などを分析できるレベル。未習語彙があっても語形成の知識や文脈から内容を推察したり分析したりして、より深く理解する言語活動。また、比較したり分類したり、また因果関係を探ったりする活動。

	聞くこと／話すこと	読むこと	書くこと
A L T E	よく知っているトピックを題材に会話ができ、話についていくこともでき、またはかなり幅広い話題について会話を維持することができる。	関連する情報を得るために文章を検索して、細かい指示や助言を理解することができる。	人が話している間にメモを取り、あるいは非標準的な依頼を含む手紙を書くことができる。

・ALTE では、「よく知っている」「かなり幅広い」「関連する情報」「細かい」「非標準的」から認知的負担度を規定している。「トピック」「文章を検索」「メモをとる」から、言語活動が維持できる、詳細について理解できる、自主的な行動がとれるなど、複数の状況に対応できるレベルを示している。

【5 Evaluating 評価】：自分の学習経験や分析力・統合力を生かして、現実世界で直面する問題・課題・危機に対して効果的な判断を下せるレベル。意見や批評など自己の思いや考えを表現する行為。

	聞くこと／話すこと	読むこと	書くこと
A L T E	自分の仕事の範囲内で会議やセミナーに効果的に貢献し、抽象的な表現に対処しながらかなりの流暢さでうち解けた会話を維持することができる。	学習コースに十分対応できるほどに早く読み、情報を得るために媒体を読み、非標準的な通信文を理解することができる。	職業上の通信文を下書きしたり作成したりし、会議で適度に正確なメモを取り、コミュニケーションできる能力を示すエッセイを書くことができる。

・ALTE では、「自分の仕事の範囲内」「抽象的」「かなり流暢さ」「学習コース」「早く」「情報」「非標準的」「職業上」「会議」「正確な」などから認知的負担度が示されている。言語行動としては、それぞれの状況に主体的かつ積極的に関わることができるレベルであると言えよう。

【6 Creating 創造】：複数の構成要素を適切に分析した結果として、新たな理論・独自の価値観などを論理整合的に統合できるレベル。自己の主張や新たな考えを発信する行為。

	聞くこと／話すこと	読むこと	書くこと
ALTE	**口語的発言を理解し、敵意のある質問に対して自信を持って対応し、複雑な問題や微妙な問題について助言し話すことができる。**	**複雑な文章の細かい点を含め、文書、通信文、報告書を理解することができる。**	**優れた表現と正確さで、どのような題材についても手紙を書くことができ、また会議やセミナーについて完全にメモを取ることができる。**
英検1級CAN・DOリスト	**1級＜聞く＞**社会性の高い幅広い内容を理解することができる。 ・幅広い話題に関するまとまりのある話を理解することができる。(一般教養的な講演や講義など) ・社会的な話題に関する話を理解することができる。（環境問題に関する講演など） ・会議に参加して、その内容を理解することができる。（イベントの打合せ、会社のミーティングなど） ・テレビやラジオの政治・経済的な話題ニュースを理解することができる。 ・いろいろな種類のドラマや映画の内容を理解することができる。 **1級＜話す＞**社会性の高い幅広い話題についてやりとりをすることができる。 ・社会的な話題や時事問題について	**1級＜読む＞**社会性の高い幅広い分野の文章を理解することができる。 ・雑誌の社会的、経済的、文化的な記事を理解することができる。（Time/Newsweekなど） ・文学作品を理解することができる。（小説など） ・資料や年鑑などを読んで、必要な情報を得ることができる。（報告書、統計的な資料など） ・留学や海外滞在などの手続きに必要な書類を理解することができる。	**1級＜書く＞**社会性の高い話題についてまとまりのある文章を書くことができる。 ・社会的な話題について自分の意見をまとまりのある文章で書くことができる。（環境問題に関してなど） ・自分の仕事や調査について、まとまりのある文章を書くことができる。（レポート、報告書、仕事のマニュアルなど） ・商品やサービスについて、苦情を申し立てる文章を書くことができる。（商品の故障、サービスの内容など） ・社会的な話題に関する雑誌記事や新聞記事の要約を書くことができる。（社説や論文など） ・講義や会議の要点のメモをとることができる。

	、質問したり自分の考えを述べたりすることができる。 ・会議に参加してやりとりをすることができる。（イベントの打合せ、会社のミーティングなど） ・幅広い内容について、電話で交渉することができる。（予定の変更、値段の交渉など） ・相手の状況に応じて、丁寧な表現やくだけた表現を使い分けることができる。		

・ALTE では、「敵意のある」「自信を持って」「複雑な」「微妙な」「通信文」「報告書」「優れた」「どのような（話題）」「完全に」から認知負担度がわかる。

・「英検」では、「社会性の高い」「幅広い」「いろいろな種類の」「丁寧な」「くだけた」で負担度を示している。また、場面や媒体としての「会議」「講義」「文学作品」「レポート」「報告書」「書類」などで、認知度の高さを明示している。言語行動としては、高度なコミュニケーション活動に支障なく参加できるレベルと言えよう。ここでいう高度とは、「やりとり」ができ「まとまりのある」表現構成ができるレベルと解釈できる。

上記の 2 つの CDS を比較してみると、ALTE では、言語能力の発達段階の上位級の複雑さの記述において、「抽象的」「流暢さ」「非標準的」「敵意のある」「複雑な」「微妙な」などの「複雑さ」の視点から記述する傾向があるのに対して、英検では、「社会性の高い」「幅広い話題」といったジャンルの広さを明記する傾向のあることがわかる。

話題に関しては、「ありきたり」「慣れた」「周知の範囲内」「よく知っている」「個人情報」「限られた」「日常的」「精通している」「関連する情報」「非標準的」など「なじみ度」で表現されている。一方、英検では、「簡単な」「短い」という言語形式に焦点を当てた表現から上位級に進むにしたがって、「一般教養的な」「いろいろな種類の」「幅広い」などの表現になっている。

13. 考察:英検における CDS の認知的負担度と言語能力の記述から

言語能力の発達段階を CDS として記述する場合、どのような観点から記述するかは重要である。英検の審査基準の場合、「社会性」「日常性」「自分性」という括りで難易度が区別されていた。また、「簡単な」「短い」などで各レベルの段階の特徴を記述している。

以下に述べる考察では、和田（2004）の「能力発達段階の観点の分類」を援用し、考察を試みる。

英検の級別の CDS では、「場面」「話題」「機能」「媒体」「知識」にかかわる用語を多様な形で組み合わせることによって、表現しようと試みていることがわかる。

「場面」についていえば、「会議」「講義」を使うことによって、これらの言葉が内在している言語活動における認知度の高さを示すことになる。下位級の 4 級レベルでは、これらの場面は出てこない。むしろ「公共の場」「レストラン」「パーティー」などが例示されている。

「話題」にかんしては、「なじみ度」が認知度の高低を示すことになる。例えば、「社会的」「経済的」は日常性からみるとなじみ度は低くなり、認知度を高めることになる。一方、「招待状」「自己紹介」などは、「自分性」に密接にかかわることになるため、認知的負担度は低くなるのである。「話題」は、背景知識とも深く関わることになるため、「公共の施設の表示・掲示」「レストランのメニュー」「旅行の思い出」などは、既有知識の支えによって、易しいレベルとなる。

次に「機能」について考察してみたい。実は「機能」については、はっきりと明示したものが多くない。社会的な言語活動にはさまざまな目的があって、言語を使用することになる。感謝する・お礼する・依頼する・要求する・謝罪する・お詫びする・断る・主張する・指示する、など多様な言語行為を行っている。このような具体的な行為を記述したCDSが少ない。英検 1 級を見ると、「会議でやりとりする」「ニュースを理解する」とあるが、抽象的で包括的な記述になっている。一部に見られるような「（電話で）交渉する」「（自分の）意見を述べることができる」「苦情を申し立てる」など、認知・思考力をともなう言語活動を記述したものが少ないことがわかった。

次に「媒体」を見てみる。何を読むかによって求められる認知度は異なってくる。「読む」の場合、文学作品である小説（1 級）と通信文の一形態である E メール（4 級）とでは、文体も異なり、使用されている語句も違う。書き言葉と話し言葉の違いも現れ、認知度の違いを暗示することになる。媒体自体が、求められる言語能力や認知能力の難易を規定しまうことが読み取れる。どのような媒体がどの言語運用レベルで記載されているかによって、あるレベルが定義づけられてしまう可能性のあることがうかがわれる。

CDS の記述を能力の発達段階として活用する場合、言語活動がどこで行われているかという「場面」ごとの分類、そしてテーマが何であるかという「話題」別の分類、そして、言語活動がどのような目的のためになされようとしているのかという「機能」の明示化、そして、どのような手段で、あるいはどのような手段を用いて言語活動を行っているかという「媒体」を明示し、話題についての「知識」の有無やなじみ度を明示することが必要となってくる。結果として、CDS が「社会性」を帯びたものか、あるいは「日常性」に関連したものであるのか、あるいは「自分性」に深く関わるものであるかによって、発達段

階に応じた適切なレベルへの分類や配置が可能になるものと思われる。

14. 結語

本稿では、英検が測定しようとする言語能力とその発達段階を示す Can-Do リストの構造分析を試みた。考察からわかったことは、１）CDS の記述には、多様な視点が存在し構造自体も異なること、２）英検の Can-Do リストは、受験者集団の一般的な言語運用力のイメージ化のために作成されていること、３）Can-Do リストで記述されていることが必ずしも英検が測定しようとしている能力を明示化したものではないこと、である。そして、４）英検が日本で英語を学ぶ学習者にとって、学習奨励や学習の目安になっていて、CEFR などのような社会生活における能力設定ではなく、英語を外国語として学ぶ日本人学習者を想定している、ということである。その結果、Can-Do リストの活用については、CEFR などと異なることに留意すべきかもしれない。そして、５）英検の Standard setting は学年ごとによる構造化されたものである可能性が高いことが考えられる。しかし、この点については、今後の検証を待ちたい。

参考文献

石井英真（2004）「『改訂版タキソノミー』における教育目標・評価論に関する一考察」『京都大学大学院教育学研究科紀要』50: p.p.172-185.

牧野成一他（2001）『ACTFL－OPI 入門』アルク

和田朋子（2004）「TUFS 言語能力記述モデル開発のための試み : Common European Framework（of Reference for Languages）の考察」『言語情報学研究報告５』p.p.89-102.　21 世紀 COE プログラム　東京外国語大学大学院地域文化研究科編

ブルーム、B.S.他（梶田叡一、渋谷憲一、藤田恵璽訳）（1973）『教育評価法ハンドブック－教科学習の形成的評価と総括的評価』第一法規出版

ACTFL (1986): ACTFL Proficiency Guidelines. In: Byrnes, H. and Canale, M (eds.) 1987: *Defining and Developing Proficiency: Guidelines, Implementations, and Concepts.* Lincolnwood (Ill.): National Textbook Company.

Alderson, J. C (1991) 'Bands and scores' In: Alderson, J.C. and North, B. (eds.) *Language testing in the 1990s*. London: British Council / Macmillan, Developments in ELT, 71.86.

ALTE (1994) European Language Examinations: Descriptions of examinations offered by members of the Association of Language Testers in Europe(ALTE) *ALTE Document 1*, Cambridge, EFL Division, University of Cambridge Local Examinations Syndicate, Version 2 January 1994

Bachman, L. F. (1990) *Fundamental Con siderations in Lang uage Testing* . Oxfo rd University Pr ess.（池田央・大友賢二監修 (1 997)『言語テスト法の基礎』C.S.L. 学習評価研究所）

Bachman, L. F. & Pal mer, A. S. (199 6) *Language Testing in Practice : Designing and D eveloping Useful Lang uage Tests* . O xford University Pres s.（大友賢二他監訳（2000）『<実践>言語テスト作成法』大修館書店）

Brown, J. D . (1996) *Testing in Language Pro grams*. Pre ntice-Hall. （和田稔（1999）『言語テストの基礎知識』大修館書店）

Brown, H . D. (2004) *Language Assessmen t: Prin ciples and Classroom Practice s* .Longman.

Canale, M. & Swain, M. (1 980) ‘Theoretical bas es of communicative approaches t o second langu age teaching and testin g' in *Applied Ling uistics I/I* .

Bloom, Benjamin S. (1956). *Taxonomy of educational objectives: Handbook I: Cognitive domain*. New York: David McKay.

Bloom, Benjamin S., Hastings, Thomas J & Madaus, George F. (1971). *Handbook on formative and summative evaluation of student learning*. New York: McGraw-Hill.

Center fo r Canadian Lang uage Ben chmarks（2000）*Canadian La nguage Ben chmarks 200 0.* Minister of Public Wo rks and G overnment Serv ices Canada.

Council of Europe (2001) *Common European Framework of Reference for Languages : Learning,teaching, assessment*. Cambridge University Press.（吉島茂、大橋理枝訳編 (2004)『外国語の学習、教授、評価のためのヨーロッパ共通参照枠』朝日出版社）

Council of Europe (2001) *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge University Press.

Davies, A. et a l. (1999) *Dictionary o f Lang uage Testing*. C ambridge Univers ity Press Hennin g, G . (1 987) *A Guide to L anguage Te sting: De velopment, Evaluation, Research* . Newbu ry House.

Dunlea, Jamie (2009)「英検と CEFR の関連性について Pa rt1」『STEP 英語情報 11・12 月号』

Dunlea, Jamie (2010)「英検と CEFR との関連性について Part2 」『STEP 英語情報 1・2 月号』

TOEIC Service International and The Chauncey Group International (1998) *TOEIC Can-Do Guide*, Chauncey Group.

Hymes, D.H .（1972）On Co mmunicative co mpetence. In J.B . Pri de & J. Hol mes (Ed s.) *Sociolinguistics: Se lected reading*. Har mondsworth: Pengu in.

参考・引用ウェブサイト

公益財団法人 日本英語検定協会：http://www.eiken.or.jp/eiken/（2014 年 3 月 14 日）

The Centre for Canadian Language Benchmarks (CCLB): http://www.language.ca/（2014 年 2 月

15 日）

Council of Europe：http://www.coe.int/ （2014 年 2 月 2 日）

TOEIC Can-D o Guide：http://www.ets.org/o.pdf#search='TOEIC+can+do+guide'（2014 年 3 月 25 日）

http://www.coun.uvic.ca/learning/exams/blooms-taxonomy.html （2014 年 3 月 14 日）

http://www.nwlink.com/~donclark/hrd/bloom.html （2014 年 3 月 14 日）

Can-do self-checklist の規準設定と妥当性

Standard setting and validity for can-do self-checklist

藤田智子

Tomoko Fujita

Abstract

A can-do self-checklist (checklist) is a self-evaluation tool which provides great opportunities for students to discover their strengths and weaknesses to become better self-regulated learners. This is one of the purposes of the Common European Framework (CEFR). This study for a can-do statement (CDS) - based English language program at a Japanese university focuses on "question order effects" of tailor-made can-do checklists for a listening course. Questions in checklists are usually presented as a series of similar items, or as items organized by difficulty level. It is possible, however, that the question order may influence answers (question order effect). In order to investigate the influence of any order effect on the checklist, and any differences among three different proficiency levels, about 600 students were asked to answer three different forms of the checklist. Therefore, 200 students each answered either Form R (randomly situated questions), Form C (a series of the questions in the similar content), or Form L (situated by its difficulty level). The results showed that Form R is different from others, and Form C and L had somewhat higher reliabilities, and they were also considered the easier forms to use. Considering the results by proficiency level, Form L showed the fewest differences among the three proficiency levels. In conclusion, considering the well-balance between order effects and convenience is the important for creating more valid and reliable can-do checklists.

1. はじめに

先の報告書第 1 号と 2 号にまとめたことや、研究結果を踏襲しながら、本年度は、Can-do チェックリストと英検Can－doリストの両方に共通する問題点として、question order effects（質問順序効果）に注目した。CDS には Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001)に準拠した European Language Portfolio (ELP) があるように、学習者が自己評価として自分の英語能力を診断し、また教員も学習者のレベルを判断する手段として利用可能なものがある。この ELP のように CDS を基に編集したり、そのまま自己評価としてのツールにしたものが、Can-do チェックリストである。これは、CEFR の目的の 1 つ、「自律学習者を養成すること」に由来する。つまり学習者が、学んだことがどのくらい身についたか自己評価し、それが十分でなければ何が良くないのか自分で考え、自己修正する「振り返り」の機会を持つことができるようにすることを意図している。いわば、Can-do チェックリストは、自律学習者のための重要なツールの 1 つなのである。

Can-do チェックリストの規準設定や妥当性に関する研究には、英検 CDS を自己評価として利用し、その妥当性を研究したり（Sato, 2010）、あるいは学習者のレベルを判断する教師評価の手段として自己評価と比較する研究（筒井、近藤、＆中野, 2007）もある。また、IRTを用いて困難度パラメタを推定し、CDSの規準設定をする方法がよく用いられ、困難度の分かった能力記述文を、Bank of descriptors としてウェブで公開もしている（North,1995, 2000; North and Schneider, 1998; Lenz & Schneider 2004)。さらに、Can-do チェックリストを、日本人学習者に適応させるための試み（中島・永田 2006；根岸 2006b）、その英語プログラムに適したものとして、どのように作成すべきかについて研究したもの（藤田・前川、2013）もあり、注目度は高い。しかし、Can-do チェックリストの question order effects についての研究は筆者の知る限りでは、ほとんどない。

さて、Can-do チェックリストは、その質問項目にスムーズに回答してもらうため、一般的に項目を難易度の低いものから高いものへ並べる配置になっていることが多い。そのため、学習者が、その項目に対して、できる/できない などと回答しているとき、真剣に質問の内容を読まず、チェックリストのその項目の位置で難易度を推定していることがあるかもしれない。

「英検 Can-do リスト」を自己評価のツールとして活用する場合も、問題となるのは学習者が上の級の能力記述文を、下の級の能力記述文と比べ、それらの内容ではなく、その項目の書かれた場所から、できる/できない の判断をするということである。例えば、準 2 級と 2 級の能力記述文を比べ、2 級の能力記述文の内容を、良く読んで理解することなく、準 2 級の能力記述文より高度な内容だと判断することが問題となる。

今後、妥当性・信頼性の高い Can-do チェックリストを日本の高等言語教育の現場に普及させるにあたって、その原動力となるのは、充分に多くの事例研究を実施して、その英語教育プログラムに適応し、妥当性・信頼性の高い Can-do チェックリストを作成するための知見を集めることだと思う。Can-do チェックリストの question order effects の影響を探求することが、日本の高等英語教育プログラムにおいて、Can-do チェックリストの作成、規準設定、妥当性検証に関わる希少な実証研究の一つとなれば幸いである。

2. CDS、Can-do チェックリストに関する先行研究

2.1 日本人学習者に適応するCDSと Can-do チェックリスト

ヨーロッパだけでなく他の地域にも影響を与えるようになった CEFR を、その国や地域に適合させたものを、国・地域言語参照レベル記述 (Reference Level Descriptions for National and Regional Languages: RLD)という。これは、ヨーロッパの言語学習者のための CEFR を、世界中の言語学習者にそのまま適用させるには無理があり、大きな変更や工夫をする必要性から出てきた動きである。そして、CEFR の枠組みは参照のためであり、その言語学習の原場に適用する形に修正して使ってほしいというのが、CEFR を作った側の考えでもある（Trim, 2001）。Weir（2005）は、CDSはそれを使用する国ごと、さらに教育機関ごと、言語カリキュラムごと、テストごとに、その学習者や受験者に適したCDSとして誂える（Tailor made）必要があると言っている。

この CEFR を日本人学習者に適合させた RLD にする試みが次に述べる CEFR-J やジャパン・スタンダード（Japan Standards for Foreign Language Proficiency based on CEFR:JS）である。このとき CEFR を、レベルを示す尺度としてのみ日本人学習者に合わせるのではなく、

CEFR の理念もともに盛り込む必要があるが、日本版は CEFR は、これらをうまく RLD 化させている「フィンランド版 CEFR」を最も踏襲している (笹島、2013)。

まず、CEFR を日本人学習者に適応させる RLD の動きのなかで、CEFR-J のフレームワークを構築しようとする取り組みが行われ、2012 年に公開された。日本人の平均的な英語能力を CEFR の 6 段階にすると、中学 3 年間はすべて A1, 高校の 3 年間から大学生は、すべて A2 になる可能性がある。日本人の英語学習者の 8 割が A レベルであり（投野、2013）、日本人学習者全体のほぼ全てが 6 段階の下 4 レベル（A1, A2, B1, B2）を占めているので、これらの下位レベルをさらに細かく分ける必要があると認識された。そこで、CEFR の下位レベルをより細かく分けているCEFRフィンランド版を参考にして、まず、A1 を3つに分ける(A1.1, A1.2, A1.3)。さらに、A2, B1, B2 はそれぞれ2つに分ける（A2.1, A2.2, B1.1, B1.2, B2.1, B2.2）方法をとって、日本人学習者に適応したレベルの設定を提唱した（岡、2008）。

また、英語運用能力に関するジャパン・スタンダードが開発された（川成、2013）。このプロジェクトでは、システマティックに構成された非常に緻密な「JS言語能力記述一覧表」が作成されている。また、JS の言語材料参照表（http://kawanarikaken.blogspot.jp）は、CEFR の理念にもとづき、学習者の自律学習を促すために、ディスクリプタ―は学習者が「自己評価」するために利用し、それによって自己の学習について「振り返り」の機会を持つことをめざしている。

2.2. 日本人学習者に適応するCDSへ

日本人学習者の英語能力の特徴をより理解するための研究としては、CEFR のレベルでテスト結果が判定される言語能力テスト DIALANG の英語版（Alderson & Huhta, 2005）を使って調査したものがある（斉田、2008）。DIALANG の結果、ある日本の大学の 1 年生のリスニング能力は CEFR の A1 レベル、リーディング能力は A2 レベル、ライティング能力は A2 レベル、文法能力は B1 レベル、語彙能力は A2 から B1 レベルとなった。これは、文法や語彙能力は高いが、リスニング能力を含む英語コミュニケーション能力は低いという、日本人英語学習者の一般的な特徴と符合している。また、日本人大学生の大多数が A1～A2 という非常に狭いレベル範囲に入るという可能性を示唆していて、CEFR を日本人学習者に適用させるようにレ

ベル設定をするには、A1、A2、B1 の3レベルではなく、より詳細なレベルを設定したほうが適切であるということを方向づけている。

また、この DIALANG の自己評価アンケート、DIALANG self-assessment (SAS) を使用して中島・永田（2006）は、日本人学習者たちが、各 CEFR の能力記述文に対してどのような困難度レベルとして認識しているかを調査した。この研究を踏まえ、根岸（2006b）は、日本人学習者たちが答えた困難度レベルと CEFR の設定している困難度レベルの間にはっきりとした相違があった項目に注目した。例えば、「お店や郵便局、銀行で簡単な用事を済ませることができる。」という CEFR　Listening A2 の項目に対して日本人学習者たちは、CEFR 設定より困難度ランクが 1 つ上の B1 レベルと判定した。これは日本人学習者が郵便局や銀行で、英語で簡単な用事を済ませる経験をしたことがほとんど無いために、困難度が高いと感じたのではないかと推測できる。そして、学習者が自己評価するとき、体験したことがないことについて質問しても、回答はあまり正確ではないことがあることを意味している（伊東・川口・太田、2008）。

このような場合、つまり CEFR レベルと日本人学習者の判定が異なった CDS に、参考資料を付けることで、学習者が具体的に内容を理解するための工夫として成果をあげ、もともと CEFR が想定していた難易度レベルに近づけることができることがある（Negishi、2005； 根岸 2006）。前述した Listening A2 レベルの項目には、銀行や郵便局での簡単なやりとりの例を示すことで改良後の項目の困難度は、ほぼ CEFR 設定どおりの順序になったと報告されている。

これらの実証研究の結果から、英検 Can-do リストの能力記述文には、その記述に説明を加えるための典型例を（　）を用いて説明している。前述の例のように、説明を加えることで、より内容の本質を理解できることもあるからだ。しかし、反対に、能力記述文の内容が特定のことに限定されすぎるという面もある（柳瀬、2013）。

3. 英検 Can-do リストへの提言

3.1. 英検 Can-do リストのなりたち

前述した文部科学省による提言のなかにも、学校は、学習到達目標をCAN-DOリストの形で設定・公表することが望ましいというような表現が用いられ、また、「外国語教育における『CAN-DOリスト』の形での学習到達目標設定のための手引き」が出されたことで、「英検Can-doリスト」を活用しようとする動きも出てきた（柳瀬、2013）。2006年に公開された「英検Can-do リスト」は、1級から5級まで（準1級と準2級を含む）合計7つの級があり、それぞれの級の合格者が、英語で何ができるのかを具体的に表すことと、英語教育関係者への情報提供を目的として作成された。

英検の特徴は、受験者が自分で受験級を決めなければならないところである。そして、受験級の選択が間違っていれば、本来の実力にあわない級を受験するという、あまり意味がないことになってしまう。従って、英検受験者は、自分の受験級を知るために、ウェブで公開されている英検 Can-doリストを試してみることも多いのではないかと想像できる。このことから、TOEICやTOEFLに比べれば、英検にとって「各級の受験者が英語でできること」をCan-doリストとして表すことは、とても重要である。

柳瀬（2013）によると、英検 Can-doリストの作成は2003年から、各級、約2000人の任意抽出した合格者を対象に作成のための調査を開始した。4技能別になった能力記述文を被験者に自己評価として5段階（1. ほとんどできない。2. 少しできる。3.ある程度できる。4.だいたいできる。5. よくできる。）で回答してもらう方法で実施した。この調査に先立ち能力記述文の作成は、中学校・高等学校学習指導要領、各種英語検定教科書、英検のテスト課題、さらにACTFL, ALTE, Canadian Language Benchmarks, CEFR, DIALANG Self-assessment List, TOEIC Can-do Guideなどを参考にして作成された。

4技能別に7つの級に分類した能力記述文を作成し、被験者の回答する項目をよりレベルにあった、限定したものにするために、隣接する級の項目を共通項目とした5フォームの質問紙を作成した。例えば、フォーム1は1級、準1級、2級の項目が含まれ、フォーム2には準1級、2級、準2級の項目が含まれる。フォーム1とフォーム2の共通項目は、準1級と2級の項

目となる。これにより、IRT を利用して 5 フォームすべての項目を同じ 1 つの尺度に載せることができる。

しかし、最終的に採用する各級の能力記述文は、選択肢 3（ある程度できる）以上を選んだ回答者の割合が 80%以上、かつ選択肢 4（だいたいできる）以上を選んだ回答者の割合が 50%以上という条件を設定して選択した。1 つ 1 つの能力記述文が、どの級のものとして採用するかは、この条件を満たしていることが基準となった。

3.2. 英検 Can-do リストのインパクト

2006 年に、英検 Can-do リストが発表されてから、「英検 Can-do リスト」の妥当性をテーマにした研究が行われるようになった。特に STEP BULLETIN には、「英検 Can-do リストの妥当性」に関する研究が掲載されている（eg. 臼田悦之、2009；竹村雅史、2008）また、英検受験者が自分たちの受験級を決めるとき、この「英検 Can-do リスト」を最終判断のよりどころにしている可能性も大きい。さらに、2012 年に文部科学省に「外国語教育における「CAN-DO リスト」の形での学習到達度目標設定に関する検討会議」が設置されて以来、日本の英語教育の現場に CDS を取り入れる動きが加速していて、日本語で書かれた「英検 Can-do リスト」を参考にして、先生方が対象とする学習者に合わせた CDS を作成する機会も増えてきたと思われる。このように、英検 Can-do リストの影響は大きく、今後もその影響は拡大すると予想される。

このように与えるインパクトの大きさを考えると、発表されてから 8 年になることも鑑み、英検 Can-do リストの妥当性を高めるために、修正や改訂をする時期に来ているのではないかと考えられる。もしその機会があるとしたら、以下の点に配慮することが望ましいと思う。

1. Can-do リスト作成と選別の過程に専門家グループが関わり、検証する。
2. Can-do リストの作成過程は、どのように行われたのか公開する。例えば JS ディスクリプターは、4 つの構成要素を決めて 1 つずつの能力記述文をシステマティックに作成し、言語材料参照表で非常に細かく全体の整合性を測った作成過程をウェブで公開している。
3. 調査の質問紙に対する回答方法は 5 件法でなく、4 件法にして真ん中に答えやすい傾

向を排除する。

4. Can-do リストの選択・最終決定のとき、IRT で推定された各項目の困難度 θ を基準にする。

5. 調査用の質問紙を作成するとき、Question order effects に配慮する。

4. Can-do チェックリストの Order Effect

一般的に、Can-do チェックリストは、「できる」「まあできる」「あまりできない」「できない」で答える 4 件法、または 5 件法のものがよくみられる。その質問項目の並べ方は、1)レベルごと、2)隣接の複数レベルが一緒、3)すべてのレベルが一緒、になっている場合がある。本研究で使用した Can-do チェックリスト・フォームL(付録参照)は、2)であり 3)でもある。このようなチェックリストは、ほとんどの場合は習熟度別に簡単から困難な項目の順、内容のジャンル別、に並んでいることが多い。

一般的にアンケート調査などの項目は、単独で質問されることはありえない。一続きの項目のかたまりとして質問されることや、一連の質問のなかのその項目の位置によって、その回答に影響を与えることがある。学習者がチェックリストに回答するときも、一部の学習者は、項目の内容をよく読まず、リストの中の項目の位置が後ろか前かで難易度を推定して答えている可能性がある。しかし、ほとんどのアンケート調査についての研究は、「項目の順序による回答への影響について」言及していない。Schuman & Presser (1996) は、アンケート等の質問の順序による影響についての研究は膨大な数には及ばず、一般的なアンケート調査に対象を限ると、過去 50 年間で多くて 24 件くらいの調査しか報告されていないと述べている。つまり、一般的なアンケート調査に関する研究においても、項目の順序が回答に与える影響について、あまり研究されていないと推測できる。

Schuman & Presser (1996)は、アンケート用紙に書かれる質問項目の順番や位置によって起こる回答への影響を order effects(順序効果)と呼び、アンケートを作成するときの重要事項としている。ある順番で質問したとき、その順番が回答に影響を与えることがあるような場合は、そのアンケート結果を一般化することが難しくなる。そして order effects は、内容が似た質問

の間で発生する可能性が高い。この影響を数値化して報告している研究はほとんどないが、Duverger (1964) が行ったフランスのある世論調査では、order effects による結果への影響は6%であった。彼らによると、order effects は避けるべきではあるが、似た内容の問題をまとめて質問するほうが、潤滑に都合よく質問できることが多い。従って、order effects を警戒しつつスムーズに質問が進むようにバランスを考えて質問を配置することもできると述べている。

また、Knowles et al (1996) は、質問紙の最初のほうに位置する質問項目は、後に続く項目の背景となって影響を与えると報告している。そして内容が関連のある項目をまとめて一連の項目として質問すると、項目どうしの相互に与える影響は増す。項目がひとまとまりにされることで、回答する側に項目どうしの関連性をより強調して受け取る傾向があるからだ。Roberson & Sundstrom (1990)の雇用者に対するアンケートの研究では、この order effects は内容によって結果への影響の与えかたが変化し、その中でも給与収入に関するものに最も影響が大きく表れたと述べている。これはおそらく、他の項目よりも回答者に強いインパクトがあるからだと考えられている。最後に、Couper, Traugott, & Lamias (2001) は、実施したウェブアンケートの実験のなかで、関連項目を1スクリーンに納めた場合と、1 スクリーンに1項目にした場合を比較した。1スクリーンに 5 問を載せた時、1スクリーンに 1 問ずつにした場合と比べ 5 問の平均点は、1 スクリーンのほうが低い場合と高い場合があった。しかし、他にも何回か同様の実験をしたが、はっきりとした違いは見られなかった。

Research Questions

Can-do チェックリストで一続きの項目のかたまりとして質問されることや、チェックリストのなかの質問される順番で、その内容が推測できるような場合がある。質問の順序が回答に影響を与える(order effects)が、実際、Can-do チェックリストで生じる可能性があるのか調査する。

(1) Can-do チェックリストに、アトランダムに項目を並べる(フォーム R)、内容別に項目を並べる(フォーム C)、想定した難易度順に項目を並べる(フォーム L)、これら 3 種類のフォームへの回答のしかたに違いはあるのか調査する。

(2) 3 種類のフォームへの回答結果と、被験者の習熟度レベルの関係を調べる。

（3）　被験者にとっては、どのフォームが回答しやすいのか調査する。

5. 研究方法

5.1. 被験者

ある日本の大学で必須英語教育プログラムを履修する1年生627人（全体の約8%）が本研究に参加した。表1. は被験者の詳細を示している。彼らは入学時に、英語プレースメントテスト(リーディング、リスニング、文法)のスコアによって3つの習熟度別レベル（初級レベル: Basic、中級レベル:Intermediate、上級レベル:Advanced)に分けられ、2 年間で約 168 時間の英語の授業を履修する。レベルによって使用する教科書も異なっていて、その習熟度レベルに適応した授業内容を実施することになっている。リスニングコースを1学期間履修する 1 年生の中で、できるだけ全体の比率と近くなるように、各レベルからアトランダムに選んだ学生たちに Can-do チェックリストを実施した。

表 1. 各フォームのレベル別回答者数

フォーム	Form R	Form C	Form L	Total
Basic	56	56	56	168
Intermediate	118	115	115	348
Advanced	38	36	37	111
total	212	207	208	627

5.2. 3 フォームの Can-do チェックリスト

本研究で使用した CDS は、日本のある大学の英語教育プログラムのリスニングに関する Can-do チェックリストを作成するための予備研究のために作成された。リスニング能力に関する CDS を Can-do チェックリストの形にして、学生が自己評価として回答するのは、到達目標にどれくらい達したかを測り、また、学習者の「振り返り」の機会をつくるためである。

このプログラムのネイティブ教員、日本人教員合計 8 人からなる委員会で、どのような CDS がプログラム履修学生たちに相応しいか話し合った。2013 年に公開された CEFR-J や、ジャパン・スタンダード(JS)などはまだ無かったので、CEFR の日本語訳（吉島・大橋、2004)を基

にして、英検 CDS、清泉アカデミック Can-do framework（Naganuma & Miyajima, 2006）などを参考にしながら、この大学の履修学生に適応するように CDS を作成した。この作成の過程で、CDS を 3 名の日本人の担当教員が読み返して、日本人大学生に理解できるように書かれているか確認した。

次に Can-do チェックリストのレベルについては、委員会において慎重に吟味した。斉田（2008）によれば、日本人大学 1 年生のリスニング能力は CEFR の A1レベルであるため、初級（B レベル）を A1 レベルに設定しようとする意見も出た。しかし、他の研究や高等教育機関では、大学生を A2 レベルとしているところも多いことや、本論で扱う英語教育プログラムにおいては、CDS を到達目標として設定することが目的であることを考慮した。そして、最も低い CEFR-A1 ではなく、せめて A2 レベルを大学 1 年生の到達目標に設定したいという意見や、さらに、対象の学習者たちは英語の習熟度は高くないが、大学生として相応の話題や内容でなければ学習意欲を減退してしまうのではないかという心配もあり、最終的に、本研究の Can-do チェックリストは A2, B1, B2 の3つのレベルで作成することになった。

次に、Can-do チェックリストの記述文については Negishi (2005), 根岸(2006)での報告に従って、できるだけ具体的な例を示すように努力した。英検 CDS を前例として倣い、典型例を(　　)を用いて説明する手法（柳瀬、2013）も導入した。最終的に日本語で書かれたリスニングの能力記述文を 30 作成し、リスニングコース担当の教員 3 名が、言葉の言い回しが学習者にうまく理解されるかどうかチェックして完成版とした。被験者たちには、これら 30 問に対して「できない」、「あまりできない」、「まあできる」、「できる」の 4 件法で答えてもらう形式をとった。4 件法を採用したのは、柳瀬（2013）での英検 CDS 作成時などで生じたように、5 件法にして真ん中の選択肢　3「ふつう」を入れると、3 と答える日本人被験者が多くなる傾向があるためである。

これらの 30 能力記述文を、作成者側で想定した難易度低～高の順序に並べたフォームLを最初に作成し、フォームC、フォームRを次に作成した。各フォームについては、以下に詳細を示す。また、項目番号はすべてフォーム L のものとする。

フォーム R: アトランダムに 30 項目を質問紙に配置した。内容が同じものをまとめることなく、

難易度にも関係なく配置された。

フォーム C: 以下の 10 種類の内容があり、それぞれの種類につき 3 つずつ項目がある。1．話す速度と内容 （speed & content）、2．メインアイディア （main idea）、3．指示と説明（instruction & explanation）、4．音声教材 （audio materials）、5．会話 （conversation）、6．スピーチとレクチャー （speech & lecture）、7．クラスルームの英語 （classroom English）、8．ビジュアル教材 （visual material）、9．語彙 （vocabulary）、10．文の複雑さ （sentence complexity）

フォーム L: CEFR の A2、B1、B2 の 3 レベルの項目を難易度の低い項目から高い項目の順に並べた。

上記の 3 種類のフォームを用意し、これらの 3 フォームを、フォーム R、フォーム C、フォーム L、フォーム R、フォーム C、フォーム L、、、というように重ねて配ってもらった。被験者にはフォームが 3 種類あることを通知していない。30 項目は、Can-do チェックリストとしての質問項目であるが、最後に 1 項目アンケートの使用に関する質問「このアンケートの答えやすさを教えて下さい。」を加えた。回答選択肢は、答えにくい、やや答えにくい、まあ答えやすい、答えやすい、の4件法である。

6. 結果

6.1. フォームの違い

3 種類それぞれのフォームの 30 項目に対する 4 件法による回答を、1～4で点数化して項目1から30までの平均値を計算した。表2は、それぞれのフォームごとの信頼性と記述統計である。信頼性は、どのフォームも非常に近い値を示している（0.916 ≦α≦0.932）しかし、この中で最も高かったのは、フォームL、次いで僅差でフォーム C であった。

表2. 3 フォームの信頼性と記述統計

フォーム	フォーム R	フォーム C	フォーム L	Total
信頼性（α）	0.916	0.929	0.932	0.926
合計点平均値	74.34	75.64	74, 86	74.95
項目平均値	2.46	2.51	2.51	2.49
標準偏差	15.11	15.11	12.77	14.33

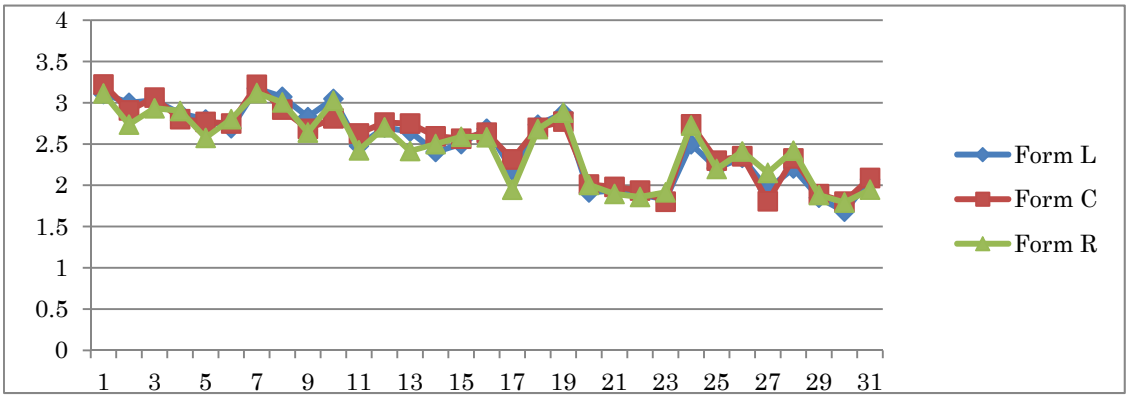


図1. 3 種類のフォーム 30 問に対する回答の平均値

図1は、3 つの異なったフォームごとの各項目平均値を表したものである。どのフォームも似た傾向になっていて重なりが多いように見える。次に、これらのデータを、SPSS を使用して一元配置分散分析にかけ、Bonferroni による多重比較による検定を行って、どの水準間に有意差があるかどうか調査した。表3に示した項目は、5%水準で有意差があるものである。フォーム R と L が最も有意差がある項目が多く 10 項目で、次にフォーム R と C が 6 項目であった。これは、フォーム R だけ他2つのフォームと差があり、フォーム L とは1番違いが大きく、フォーム C とは 2 番目に違いがあるということを示している。

表3.　フォームの違いによる多重比較

Question	フォームRとC	フォームRとL	フォームCとL
2	*	*	
5	*	*	
9		*	
10	*	*	
11		*	
13	*	*	
17	*	*	
24		*	*
27	*	*	
28		*	

Bonferroni , $p<0.05$ で有意差があるもの＊

6.2. フォームごとの習熟度レベルによる違い

6.2.1. 習熟度レベルによる違い

次に、フォームごとの習熟度レベルによる違いを調べる前に、学習者の習熟度レベル　初級（B）、中級（I）、上級（A）の違いによって回答がどのように異なるのか確認した。図2は、そ

れぞれのレベルの回答の平均値をグラフにしたものである。順当に、習熟度レベルが高い A、I、B の順に「できる。」と答えた人が多かったことが示されている。

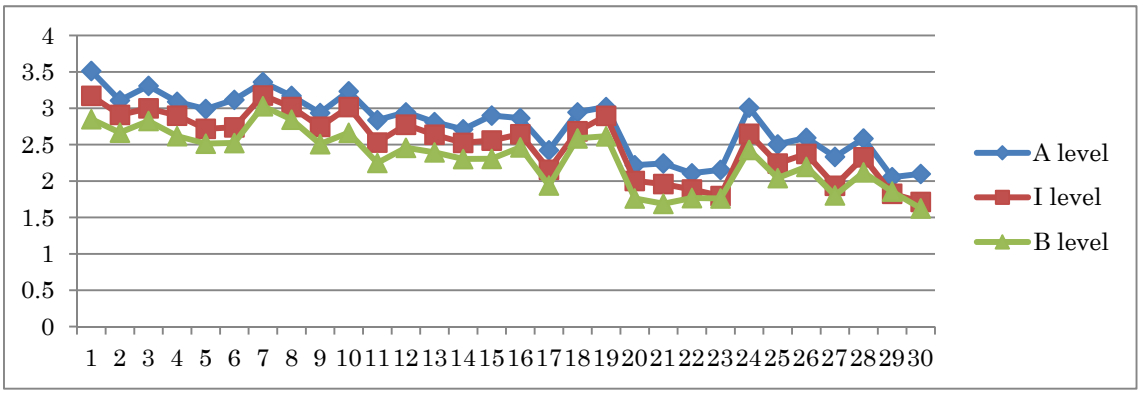


図2. 習熟度レベル毎の回答平均値

習熟度レベルごとのデータを SPSS を使って一元配置分散分析にかけ、Bonferroni による多重比較による検定を行って、どのレベル間（レベル A と I, レベル A と B, レベル I と B）に有意差があるのか調べた。その結果、30 項目中 18 項目において 3 レベル間すべてに有意差が認められた。また、どの水準（A,I,B レベル）間にせよ、有意差がないという判定結果になったのは項目番号8、9、12、13、14、18、19、22、23、27、29、30 の 12 項目だけであった。従って、習熟度レベルの違いによる回答の差はあることが確認できた。

6.2.2. フォーム R

さらに、フォームごとに習熟度別レベルの違う被験者の反応を調査した。まず、フォームRに関しては、図3によると、レベル A と他の 2 レベル（B と I）は違いがあるように見える。これに反し、レベル I と B は非常に似た結果になっている。

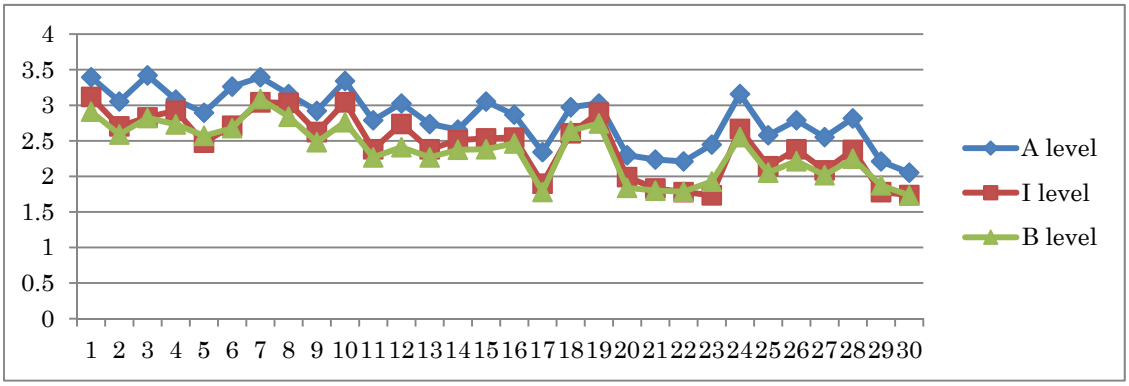


図3. フォーム R に対するレベルごとの回答平均値

これらのデータを SPSS を使用して一元配置分散分析にかけ、Bonferroni による多重比較による検定を行ってどの水準間に有意差があるかを調査した。表4 に示した項目は、5%水準で有意差があるものである。フォーム R の受験者のうち、A と B レベルの受験者間では 20 項目、A と I レベルでは 15 項目、I と B では 4 項目有意差があることを示している。項目がアトランダムに配置されたフォームでは、A レベルの受験者と他の 2 レベルの受験者の回答のしかたに違いがある。

表4. フォームRの習熟度レベル A, I, B による多重比較

Question	レベル A と I	レベル A と B	レベル I と B
1		*	
2		*	*
3		*	*
5	*		
6			
7	*		
9		*	
10		*	
11	*	*	
12		*	*
13	*	*	
14			
15	*	*	
16		*	
17	*	*	
18	*		
20		*	
21	*	*	
22	*	*	
23	*	*	
24		*	*
25	*	*	
26	*	*	
27	*	*	
28	*	*	
29	*		

Bonferroni， $p<0.05$ で有意差があるもの＊

6.2.3. フォームC

図4　では、フォーム R とは反対に、フォーム C では、習熟度レベルが低い B レベルだけ、他の 2 レベルと異なる回答のしかたをしているように見える。

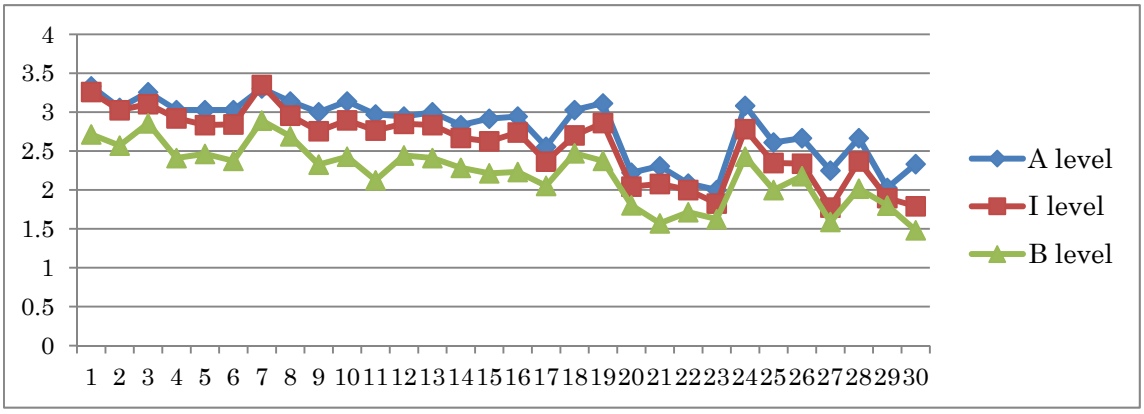


図 4. フォーム C に対するレベルごとの回答平均値

表5.　フォームCの習熟度レベル A, I, B による多重比較

Question	レベル A と I	レベル A と B	レベル I と B
1		*	
2		*	*
3		*	
4		*	*
5		*	*
6		*	*
7		*	*
8		*	
9		*	*
10		*	*
11		*	*
12		*	*
13		*	*
14		*	*
15		*	*
16		*	*
17		*	*
18		*	
19		*	*
20		*	
21		*	*
22		*	*
23		*	
24		*	*
25		*	*
26		*	
27	*	*	
28		*	*
29			
30	*	*	

Bonferroni ,　$p<0.05$ で有意差があるもの＊

これらのデータを一元配置分散分析にかけ多重比較による検定を行った結果、5%水準で有意差があるものは、表5によると、A と B レベルが 29 項目、 B と I レベルが 20 項目もあるのに対し、A と I レベルでは 2 項目だけであった。 フォーム C では、A とIレベルは、どの項目でも非常に近い平均値を示している。項目が内容別に配置されたフォームでは、B レベルの受験者と他の 2 レベルの受験者の回答のしかたに違いがある。

6.2.4. フォームL

図5によると、フォーム L では、どの習熟度レベルも似たような平均値を示しているようである。

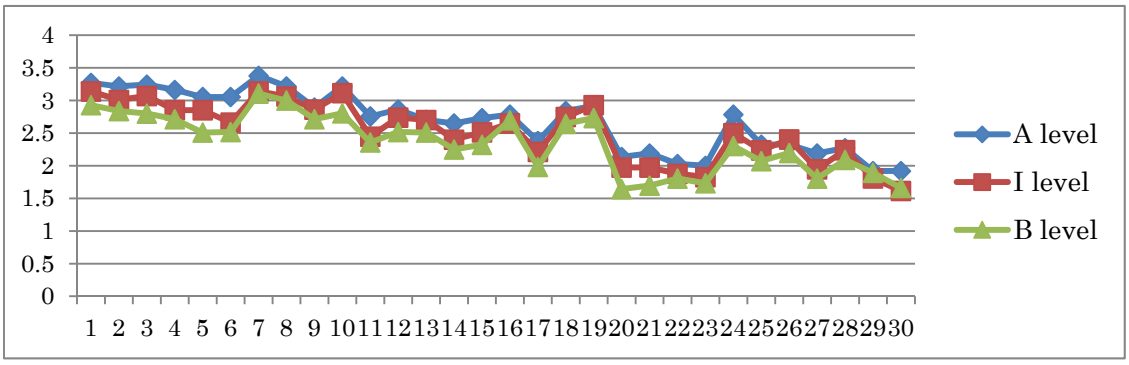


図5.　フォーム L に対するレベルごとの回答平均値

表6.　フォームLの習熟度レベル A, I, B による多重比較

Question	レベル A と I	レベル A と B	レベル I と B
1		＊	
2		＊	
3		＊	＊
4		＊	
5		＊	＊
6		＊	＊
10	＊	＊	
11	＊	＊	
14		＊	
15		＊	
16		＊	
17		＊	
20		＊	＊
21		＊	＊

Bonferroni , $p<0.05$ で有意差があるもの＊

表6に示す、一元配置分散分析の結果では先のフォーム R,C に比べ、全体的に、どのレベル間においても有意差がある項目が少ない。最も有意差がある項目が多いのは、レベル A と

B で 14 項目あるが、レベル I と B では 5 項目、レベル A とIでは 2 項目だけであった。特に困難度が高いと推定される後の方の項目（22～30）ではどのレベル間においても有意差がなかった。これは、フォーム L を使用すると、どの習熟度レベルの被験者も、大きな違いがなく回答する可能性を示唆している。

6.3. 使いやすさ

最後に、表7は3つのフォームごとに、どれが最も使いやすいか尋ねた時の回答の平均値である。大きな差ではないが、フォーム C、フォーム L、フォーム R の順に使いやすいという回答を得た。しかし、これらのデータを一元配置分散分析にかけて多重比較した結果、どのフォーム間においても有意差はなかった。回答を習熟度レベルごとに調査した結果は、初級（Basic）、中級(Intermediate)、上級(Advanced)レベルの学生の順に使いやすいと回答している。

表7. フォームごとの使いやすさ（レベル別）

フォーム	Form R	Form C	Form L	Total
Basic	2.00	2.36	2.17	2.18
Intermediate	1.93	2.07	2.00	2.00
Advanced	1.92	1.75	2.14	1.94
total	1.95	2.09	2.07	2.04

7. 考察

7.1 3 種類のフォームによる違い

記述統計上では、3 種類のフォームには大きな違いがなかったが、信頼性（α）に関して、フォーム C と L がフォーム R より少し高かった。また、3 種類のフォームの 30 問に対する回答の平均値を表すグラフは、ほとんどの項目でとても近い値となっていたが、このデータを一元配置分散分析で多重比較した結果、フォーム R と L が最も有意差がある項目が多く、有意差がある 10 項目の困難度は、ほぼ均等に散らばっていた。次に有意差がある項目が多いのはフォーム R と C（6 項目）で、フォーム R とフォーム C・フォーム L 間には違いがあるが、フォーム

CとL間にはほとんど違いが無いことが分かる。

フォームRと他の2フォームが異なる傾向にあるのは、フォームRだけがアトランダムに項目を並べたものであるのに反して、フォームCやLは、内容が同じであったり、難易度順に並んでいたり、回答する被験者に何らかの手がかりを与えている点で共通している。同じ30項目の質問であっても、フォームが違うことで回答に影響が出るということは、question order effectsであると推定できる。この結果は、order effectsは内容に関連がある項目の間で発生する可能性が高く、最初の方に位置する項目は、後に続く項目の背景となって影響を与えるなどとする 先行研究のorder effectsに関する結論と相反しない。この結果は、order effectsを避けることを優先するべきか、それとも被験者がスムーズに回答できるフォームを避けるべきか、質問紙の作成者は、この2つのバランスを考えて項目を配置する必要があることを示唆している。

7.2 フォームごとの習熟度レベルによる違い

フォームRでは、習熟度が高いAレベルの被験者が他のIとBレベルの被験者たちと異なった反応をしている。それに反してフォームCでは、Bレベルの被験者が他のAとIレベルの被験者たちと異なった反応をしている。最後にフォームLでは、習熟度レベルの違いによって回答にあまり違いがない。これらを考察すると、フォームRではAレベルが、フォームCではBレベルがフォームの違いによる影響を受けやすいと言える。これはおそらく、フォームRは使いにくいフォームであるため、習熟度レベルの高いAレベルの被験者は、フォームによる影響をあまり受けずに回答する傾向にあるが、IやBレベルの被験者はフォームによる影響を受けやすいことを示していると推察できる。しかし今回の結果では、フォームLは、どの習熟度レベルの被験者も、フォームの違いに影響をうけることなく回答することができるため、他の2フォームに比べ万人向きであると言える。

7.3 フォームごとの使いやすさ

非常に少ない差ではあるが、3つのなかではフォームRについて「答えにくい」と回答した被

験者が多いという結果になった。最も「答えやすい」のは、フォームCで、僅差であるが次いでフォームLである。フォームRが「答えにくい」のは、フォームCやLに比べて追加の情報がないからであろう。フォームCのように、同じ内容の項目がまとまっていると、1つずつ項目に答えるより、関連づけることで同じ内容の他の項目に答えやすくなると考えられる。また、Can-doチェックリストの場合は、できる/できないで答える性質上、フォームLのように難易度が予測できると「答えやすい」と感じるのも道理である。

8. 結論

本研究の項目配置順の異なるCan-doチェックリスト3種類のフォームの分析結果から、難易度順や内容別のフォームを使用した被験者は、アトランダム順のフォームを使用した被験者と異なる結果を示し、フォームの項目配置順序が結果に与える影響（question order effects）がある可能性が高いことが分かった。また、習熟度レベル別の被験者の反応を、フォームごとに比べた場合、習熟度の高い被験者ほど低い被験者に比べ、フォームの違いによる影響を受けにくい可能性が示唆された。フォームは、内容別、難易度順フォームがほぼ同じくらい使いやすいが、アトランダム順フォームは他のフォームより使いにくいと被験者たちが感じる傾向があることも判明した。

これらを総合すると、項目を難易度順に並べたり、内容ごとにまとめると、question order effectsの影響は受けやすいが、被験者の習熟度レベルの違いを問わず、スムーズに回答しやすくなると結論づけることができる。つまり本研究の結果は、order effectsに対する考慮と、被験者が円滑に回答しやすいこと、この両面のバランスを考えてCan-doチェックリストの項目を配置するべきであることを示唆していると考えられる。例えば、1）難易度順や内容別に項目を並べる場合、できるだけ難易度順や内容別であることが一目瞭然にならないフォームを作成する。2）order effectsを避けるために項目はアトランダムに並べるが、一つ一つの項目を簡単明瞭にして回答しやすくするなどが考えられる。そして、より妥当性・信頼性が高いCan-doチェックリストフォームを作成するには、対象とする学習者の習熟度レベルに合わせたり、その学習内容や到達目標に合わせたり、千差万別の対応をその都度考える必要がある

が、今後は、order effects と使いやすさのバランスを考慮することを付け加えたいと思う。

これからも CDS を到達目標とする英語教育プログラムにおいて、妥当性・信頼性の高い Can-do チェックリストの重要性は高まるであろう。本研究は、1万人以上の学生が履修するある日本の大学英語統一プログラムに於いて、大学独自の CDS を作成し、それを到達目標として設置するプロジェクトの一環として実施された。このようにして開発された CDS や Can-do チェックリストを根幹とした、大規模な統一英語カリキュラムを維持するのに最も重要なのは、教員と学生の間に「対話」を作ることではないかと思う。このために Can-do チェックリストの役割はとても大きいと思われる。なぜなら、この統一英語カリキュラムでは、学期のはじめ、半ば、おわりの 3 回、学生による Can-do チェックリストを使った自己評価を実施して、教員の評価と比較する機会を作っているからである。これら 3 回の Can-do チェックリストは、学生にとっては、これまでの学習を「振り返る」機会であり、教員にとっては学生の自己評価と教員評価をすり合わせて、学生を自己修正に導く機会となる。そして重要なことは、教員たちが僅かな時間であっても、学生ひとり一人の Can-do チェックリストの回答を見て、教師評価とのすりあわせをしたり、フィードバックを与えるなど、Can-do チェックリストを通じて学生と「対話」するようにしている点である。このように「対話」のきっかけとなることも、Can-do チェックリストの大切な役割の一つだと思う。今後も Can-do チェックリストに関する研究がより多く実施され、妥当性・信頼性の高い Can-do チェックリストが作成されることで、CDS を基盤にした英語教育プログラムがより活性化されることを期待する。

参考文献

Alderson, C., & Huhta, A. (2005). The development of a suite of computer-based diagnostic tests based on the Common European Framework, *Language Testing, 22* (3), 301-320.

Council of Europe (2001). *The Common European Framework of Reference for Languages: Learning, teaching, assessment.* Cambridge: Cambridge University Press.

Couper, M., Traugett, M., & Lamias, J. (2001). Web survey design and administration. *Public Opinion Quarterly. 65*(2), 230-253.

Duverger, M.(1964). *Introduction to the Social Sciences. London*: George Allen and Unwin.

Knowles, Eric S., Byers, & Brenda. (1996). Reliability shifts in measurement reactivity: Driven by content engagement or self-engagement? *Journal of Personality and Social Psychology, 70*(5), 1080-1090.

Sato, T. (2010). Validation of the EIKEN Can-Do statements as a self-assessment measure using Rasch measurement. *JLTA Journal, 13.* 1-20.

Schuman, H., & Presser, S. (1996). *Questions and answers in attitude surveys.* SAGE Publications, Thousand Oaks, CA.

Trim, J. (2001). Chapter 1: Guidance for all users. In Council of Europe (Eds.), *The Common European Framework of Reference for Languages: Learning, teaching, assessment.* (pp.1-7). Cambridge: Cambridge University Press.

Naganuma, N., & Miyajima, M. (2006). The development of Seisen academic Can-Do framework. *Bulletin of Seisen University, 54,* 43-61.

Negishi, M. (2005). The development of an English proficiency scale in Japan. *ARELE. 16,* 191-200.

North, B. (1995). The development of a common framework scale of descriptors of language proficiency based on a theory of measurement. *System, 23*(4), 445-465.

North, B. (2000). *The development of a common framework scale of language proficiency.* New York: Peter Lang Publishing.

North, B., & Schneider, G. (1998). Scaling descriptors for language proficiency scales.

Language Testing, 15(2), 217-263.

Roberson, T., & Sundstrom, E. (1990). Questionnaire design, return rates, and response favorableness in an employee attitude questionnaire. *Journal o f Applied Ps ychology, 75*(3), 354-357.

Weir, C. J. (2005). Limitation of the common European framework for developing comparable examinations and tests. *Language Testing, 22*(3), 281- 300.

伊東田恵・川口恵子・太田理律子（2008） 外国語能力の自己評定における言語タスク経験の影響.『JLTA Journal』、11. 156-169.

臼田悦之（2009）「英検 Can-do リストのスピーキング分野における Can-do 項目の妥当性検証」財団法人日本英語検定協会『STEP BULLETIN』vol. 21.

岡秀夫 （2008） 英語教育の基準を求めて-日本版 CEFR への取り組み.『英語展望』、116, 13-23.

川成美香（2013）CEFR準拠の新たな英語到達基準JS「ジャパン・スタンダード」の策定 『英語展望』、121、8－13.

斉田千里（2008）ヨーロッパ言語共通参照枠（CEFR）による日本人大学生英語力診断の試み-英語教育達成目標への CEFR 適用可能性の-検討- 『JACET Journal』、47, pp. 127-140.

笹島茂 （2013） JSにおける言語材料参照表の概要と利用. 『英語展望』、121、14－19.

竹村雅史（2008）「英検 Can-do リストによるWriting技能に関する妥当性の検証」 財団法人日本英語検定協会『STEP BULLETIN』, vol. 20.

筒井英一郎・近藤悠介・中野美知子 （2007） 日本人英語学習者の実践的発話能力に関する評価規準の検討 -Common European Framework of References を基盤として-. Paper presented at the Nippon Test Gakkai (JART), Tokyo.

投野由紀夫（編）（2013）『英語到達度指標 CEFR-J ガイドブック』東京：大修館書店.

中島正剛・永田真代（2006）CEFRの日本人外国語学習者への適用可能性.『外国語教育研究』、8、 5-23.

根岸雅史 （2005）「日本における英語能力記述の枠組みの開発」『ARELE: annual review

of English language education in Japan』、全国英語教育学会, 16, pp. 191-200.

根岸雅史 (2006) GTEC for STUDENTS Can-Do Statements の妥当性検証研究概観. 『ARCLE REVIEW』、1, pp. 99-103.

根岸雅史(2006b)CEFR の日本人外国語学習者への適用可能性の向上に向けて.『言語情報学研究報告』、14、79-101.

藤田智子(2013)Can-do statements (CDS) の規準設定. 『言語テストの規準設定』公益財団法人日本英語検定協会　英語教育研究センター委託研究報告書　第 2 号, pp. 60-80.

藤田智子・前川眞一(2013)日本の大学英語教育プログラムに於ける Can-do statements の規準設定.『日本言語テスト学会誌』第 16 号, pp. 147-166.

柳瀬和明(2013)CAN－DOへの関心の高まりと「英検Can－doリスト」『英語展望』、121、32－37.

吉島茂・大橋理枝(訳編)(2004)『外国語教育 II- 外国語の学習、教授、評価のためのヨーロッパ共通参照枠』、東京:朝日出版社.

付録.

あなたの英語リスニング能力についての質問です。最も当てはまるところの〇を一つ塗って下さい。なお、このアンケートはこれからの英語プログラムの改善のために学校としてお願いするもので、皆さんの成績評価とは全く関係がありませんので、正直に答えてください。

Form L

英語リスニングについて		できない	あまりできない	あまりできる	できる
1	ゆっくり話されれば基本的で学習者にとってごく身近な話題（例：基本的な個人や家族の情報、買い物、近所）についてその要点を理解することが。。。	○	○	○	○
2	シンプルで短いメッセージのメインアイディア(話者が最も言いたいこと)を理解することが。。。	○	○	○	○
3	短い説明や簡単な指示（例：道案内、集合場所や時間など）の要点を理解することが。。。	○	○	○	○
4	日常の事柄に関する、短い録音（例、教材など）の一部を理解し、必要な情報を理解することが。。。	○	○	○	○
5	身近な内容に関する会話の話題を理解することが。。。	○	○	○	○
6	身近な内容に関する簡単で短いストーリーの要点を理解することが。。。	○	○	○	○
7	教員の英語の指示は、簡単であれば理解することが。。。	○	○	○	○
8	映像がほとんど説明してくれるならば、どのような出来事や事故を伝えるテレビのニュースであるかメインポイントを理解することが。。。	○	○	○	○
9	身近なものに関する基本的なコミュニケーションの要求をみたすことのできる単語（例: 個人や家族の基本情報、買い物、近所のこと）からなる話を理解することが。。。	○	○	○	○
10	シンプルな構造の文が多く使われた話を理解することが。。。	○	○	○	○
11	標準的な速さで話されれば、学校や余暇などの場面で出会う身近な話題を理解することが。。。	○	○	○	○
12	良く知っている話題であれば、メインアイディア(話者が最も言いたいこと)と補助的な詳細を理解することが。。。	○	○	○	○
13	よく知っていることについてのアナウンス、指示や説明（例：毎日使っている設備の取り扱い説明など)を理解することが。。。	○	○	○	○
14	身近な話題に関するラジオの短いニュースや、簡単な内容の録音された音声素材の要点を理解することが。。。	○	○	○	○
15	自分の周りで話されている身近な内容の会話を理解することが。。。	○	○	○	○
16	よく知っている内容の短く簡単なスピーチの要点を理解することが。。。	○	○	○	○
17	教員の英語の指示は、やや複雑でも理解することが。。。	○	○	○	○
18	映像が大筋を説明していれば、身近な話題について事実を伝えるニュースの要点を理解することが。。。	○	○	○	○
19	日常的また自分のよく知っていることに関する単語からなる話を理解することが。。。	○	○	○	○
20	複雑な構造の文が含まれていても話の要点を理解することが。。。	○	○	○	○

21	自然な速さで話されても、毎日や普段の大学で話すような内容について、事実の情報を細部まで理解することが。。。	○	○	○	○
22	社会性や専門性の高い話題でもメインアイディア（話者が最も言いたいこと）と補助的な詳細を理解することが。。。	○	○	○	○
23	社会性や専門性の高い分野のアナウンス、指示や説明を理解することが。。。	○	○	○	○
24	よく知っている話題であれば、録音され、放送された音声素材の内容を理解することが。。。	○	○	○	○
25	普段大学で話すような内容について、自分の周りで話されている会話を理解することが。。。	○	○	○	○
26	よく知っている内容の明確に構成された講義であれば理解することが。。。	○	○	○	○
27	教員の英語の指示や解説は、複雑な内容であっても理解することが。。。	○	○	○	○
28	よく知っている話題についてのインタビュー、短い講演、ニュースレポートなど多くのテレビ番組や映画の内容を理解することが。。。	○	○	○	○
29	本人の専門性や社会性の高い単語が含まれる話を理解することが。。。	○	○	○	○
30	複雑な構造の文を多く含む話を理解することが。。。	○	○	○	○

最後に、このアンケートの答えやすさを教えて下さい

31	答えにくい ○	やや答えにくい ○	まあ答えやすい ○	答えやすい ○

実用英語検定の級別頻出単語に基づく英語受容語彙力テストの開発と規準設定
Setting Standards for Two Versions of a Receptive English Vocabulary Size Test Aligned with Different Grades of Eiken Tests

法月　健
Ken Norizuki

Abstract

This paper explores the practical application of the Rasch model and LRT (Latent Rank Theory) techniques to setting standards for a receptive English vocabulary size test aligned with Eiken tests of Grades 4, 3, Pre-2, 2, Pre-1 and 1. The first step was to develop two versions of the test to measure the receptive knowledge of important words frequently used in these six grades. Each completed version had 50 odd-numbered synonym-matching items and 50 even-numbered items in which examinees were asked to rate the extent of their knowledge about the word in each odd-numbered item on the four-point scale. The two versions of the test named as the Vocabulary Knowledge Survey (VKS1 & VKS2) were administered to 347 university students in Japan. Despite some problems that require further research, the present findings suggest that VKS1 and VKS2, which were successfully equated through 12 common items, placed examinees into appropriate ability levels in relation to difficulty levels of words as well as specific Eiken grades that the words were associated with. It was also found that the test had an additional advantage by allowing testers/teachers to scrutinize the extent of examinees' knowledge about individual items with both correct and incorrect responses.

1. 問題と目的

あるテストにおいてX点以上を合格、X点未満を不合格とした場合、X点を分割点とする理由が便宜的なものになることは少なくない。このような問題点を解決するため、2011 年度の本委託研究において、筆者は「ラッシュモデル」と「潜在ランク理論(LRT)」の規準設定における有用性について、文献調査を中心とした研究を行った (法月、2012a; 2012b)。

研究の結果、段階評価に基づく LRT (植野・荘島、2010)は、規準設定の基盤となる分割点を決定するのに有用なランク関連指標を提供するのに対して、ラッシュモデルの分析は、様々な規準設定法の審査判断における客観性を高め、順序尺度と間隔尺度を融合した統計モデルへと発展させることも可能であることが明らかになった。

2012 年度 (法月、2013) は、人的・技術的・時間的な制約下での規準設定の遂行を想定して、ラッシュモデルや潜在ランク理論の手法等を実践的に応用することを研究目的に掲げた。実際にある大学の 1 年生の英語必修クラスの能力編成 (プレイスメント) を迅速に決定する目的で開発され、数年間実施された受容語彙力テスト (SCELP) の結果のうちの一部を、ラッシュモデルと LRT を活用して、分析を行い、現実的な規準設定の方法を探り、試行した。

分析の結果、SCELP は、信頼性係数の数値も非常に高く、ラッシュモデルと LRT を併用することで、整合性の高い規準設定の方法を導き出すことができた。

SCELP は、北海道大学で開発された第 1 水準から第5水準で構成される英語語彙表を基に開発されたプレイスメントテストであるが、実際には対象受験者の能力域や習熟度の低い受験者への負担に配慮して、第 1～3の水準の単語に限定して、問題は作成されている。そのため、より幅広い能力層の学習者に十分に対応できるかどうかは不明である。また英語資格試験に照準化されたテストではないため、結果から、詳細なフィードバックや規準設定の意味づけを行うことは困難である。

SCELP は、特定の大学内で限定的に利用するプレイスメントテストとしては、効率的な分割機能を有する手法であったと言えるが、仮に他の教育機関にも活用され、様々な利害関係者から質的フィードバックの提供が求められる状況にあったならば、十分に効果を発揮できなかったかもしれない。2013 年度はこのような問題点を念頭に新たな受容語彙力テストを開発し、ラッシュモデルと LRT の規準設定について、さらなる検証を行うこととした。

2. 先行研究

2. 1. 受容語彙力テストとしての SCELP の有用性と課題

近年の研究において、語彙力テストは、発表語彙力と受容語彙力を測定する両面において、様々な目的で活用され、その効果が幅広い角度から議論されているが、Laufer and Goldstein (2004) は、受容語彙力テストのほうが発表語彙力テストよりも、受験者の将来のリーディング、ライティング、総合的言語能力や学術的達成の成否を予測するのに適していて、クラス編成や入学許可の目的で使用するのに優れていると主張している。

受容語彙力を測定する代表的なテストとして、知っている単語に「Yes」、知らない単語には「No」の欄にチェックさせ、存在しない単語に「Yes」を選んだ場合は減点される European Vocabulary Size Test があるが (Read, 2000)、このような「Yes/No」語彙力テストの有用性については、近年も盛んに議論されている（Alderson, 2006; Stubbe, 2012 等）。「Yes／No」テストは簡易で短時間で多くの語彙項目の知識の有無を確認することができるが、「Yes」を選んでも別の単語と勘違いしている場合もあり、実際の理解の度合いは見当がつかない。

これに対して、Beglar and Hunt (1999) や小泉・飯村 (2010) は Nation (1990) が開発した Vocabulary Levels Test (VLT) や望月語彙テスト (MVST) を日本人学習者のプレイスメントに活用した分析を行っている。意味（語義や訳）と形式 (綴られた単語) の理解状況について確認ができる点で、「Yes／No」よりも深い受容語彙力を確認できるが、その分、各項目への解答に時間を要し、時間の制約がある場合は、設問数が制限される。また前者 (VLT) は英語の語彙定義を理解できる習熟度の学習者でないと適用は難しく、後者は訳語がすべての受験者に同等に理解できる言語環境でないと使用できない。

問の単語の日本語訳と英語の類義語の選択肢を組み合わせる SCELP は、ある大学の 1 年次必修の英語クラスの決定の目的で開発された 30 分程度で終了する 80 問のテストである。日本語と英語の2カ国語の選択肢情報を提供して意味と形式の知識を確認することで、日本語力の劣る留学生や習熟度が高くない学習者にも適応するように設計されている。法月 (2013) は、実際に数年間プレイスメントテストとして使用された SCELP のデータの一部（150 名）を分析し、新たに 10 名程度の学習者に、実験参加者としてテストに加えてアンケートと面談を実施し、補佐的な検証を行った。

分析の結果、SCELP は信頼性 (KR20 = . 949) が非常に高く、若干のミスフィット項目や受験者はあるものの、大きな問題があったとは考えられず、30 分で終了する簡易テストであったことを考えると、きわめて効率性がよく、正確なテストであったと言える。また、後述のようにラッシュモデルと LRT を使った規準設定においても有効に機能していることがわかった。

SCELP は、単一の教育機関のプレイスメントのニーズを十分に満たしたテストであったと言えるが、便宜的に分割した5つの水準が何を意味しているのかを明確に規定することはできなかった。Milton (2009) は、近年の研究結果から、受容語彙サイズと IELTS の評定、Cambridge FCE の合否、CEFR の水準との対応関係を明確に提示している。日本では幅広い英語習熟度の学習者が実用英語検定を受験しており、教師は初めて指導する学習者に対しても英検合格級で総合的な英語習熟度を判断することが多い。英検の級別によく出題される単語が頻度順に分類された単語集も出版されているが、このような単語集と実際のテスト問題を参考に、各級の水準や基準を反映した受容語彙力テストを開発することによって、より明確な理念に基づいた規準設定を行うことが可能か検証する価値がある。

2．2． ラッシュモデルと LRT を活用した規準設定の有用性と問題点

規準設定の方法は数多く存在するが、少ない人員で一連の判定手続きを行う状況において、複雑な方法を実施することは容易ではない。Pitoniak and Morgan（2012）はアメリカの大学のプレイスメント実施の際には、様々な専門家の意見を結集するために、評定者グループは最低 10 名、理想的には 15 名必要であるとしているが、日本の大多数の大学において、これだけ多数の評定者を集めるのは非現実的な制約と言える。また、大半の規準設定法において複数回の評定の点検が課せられているが、現場の関係者はみな、学年度の初めの繁忙期にそれほど時間をかけられない状況にある。

綿密な計画を立て、専門家の議論のもとに規準設定が行われても、人間の判定には恣意性が伴う。より客観的で、説明力の高い判定を行うために、Lissitz (2013) はラッシュモデルと潜在クラスモデルを統合した混合ラッシュモデル (Mixture Rasch model: MRM) 等に代表される統計的解決法を提唱している。しかしながら、規準設定の目的で MRM の分析を行うには、大きな受験者とテスト項目のサンプルの使用が望まれる (Jiao, Lissitz, Macready, Wang & Liang, 2011) ため、一般的に小規模の教育プログラムにおけるプレイスメント決定には十分に適応しないと考えられる。

法月 (2013) は、ラッシュモデルとLRT を融合することで、同一の間隔尺度上でテスト項目の難易度と受験者能力を直接比較しつつ、統計的に付与された潜在ランクを参考にして、分割点をより合理的に設定する方法を探った。分析の結果、ラッシュモデルとLRT を併用することで、150 名の SCELP のデータにおいて簡便な規準設定が可能であることが確認できた。

具体的な規準設定の方法は、ラッシュモデルの能力推定値と LRT のランクが変化する地点の中から適切な分割点を選び、分割点を含めた各能力推定値に近接する難易度の項目を対比するものだった。

大友 (2013a) は、項目応答理論を使って、各項目の難易度、識別度、及び正答確率が.67 になる能力推定値を算出し、それぞれの指標がもっとも大きな変化を示す中心的項目に分割点を定めているが、大友 (2013b) において同データをラッシュモデルと LRT を併用して分析した結果、同じ分割点決定に至ったことが報告されている。

ラッシュモデルの能力推定値と LRT のランクが変わる地点は一致しないことが多いため、分割点を決めるためには多様な状況に対応できる判定基準を明確にする必要があるだろう。また、大友 (2013b) のように、受験者能力ではなく項目難易度を軸に分割点を設定する場合、分割点に位置する項目の特徴づけを具体化する必要があるだろう。

3．研究方法

3．1．テスト

SCELP の問題点や 2012 年度の研究の問題点を克服するために、以下のような改善指針を念頭に、2 種類の語彙力テスト(Vocabulary Knowledge Survey: VKS) を開発した。

①より広域な能力層に対応するため、50 項目ずつで構成される 4、3、準2、2、準 1 級の頻出重要語彙を対象にした Version1 (VKS1) と、準2、2、準1、1 級の頻出重要語彙を対象にした Version2 (VKS2) のテストを開発した。級別の頻出重要単語は、「英検でる順パス単」(旺文社)シリーズの4級から1級の単語集の中から、頻度区分や収録語の品詞割合を考慮に入れながら、Excel で発生させた乱数を参照して、できるだけ無作為に抽出した。VKS1 については、各級 10 項目ずつで構成されるが、VKS2 については、準2級から1級までの問題を均等に4分割できないため、準2級と2級の単語集に共通に含まれる7単語(2項目＋5選択肢) によって準2・2級共通2項目を設け、残りは各級12項目で構成した。各 50 項目のうち、準2級6項目、2級4項目、準1級4項目をテスト等化 (equating) のために共通項目とすることを意図したが、選択肢を共有する2級2項目の5つの選択肢がバージョン間で1つずつ異なっている状態でテストを実施したため、この2項目については、非共通項目として扱うこととした。その結果、バージョン間で異なる 38 項目、両バージョン共通の 12 項目 (準2級6項目、2級2項目、準1級4項目) を分析することになった。

②SCELP は、習熟度の低い学習者や日本語能力が高くない留学生のために選択肢に日本語訳と英語の類義語を設ける形式を取ったが、実験目的のために受験し、アンケートや面接に応じた日本人学習者のコメントから、問題項目の単語の意味がわからなくても選択肢の日本語訳から答えを類推することに依存する学習者が多くいることがわかった。実際の英検の語彙問題は短文の文脈が用意される穴埋め形式である点で異なるが、日本語訳を削除し、意味の対応関係から英語の類義語を組み合わせる解答様式を取ることとした。

③2012 年度の研究では、一部の受験者にアンケートと面接を行ったが、すべての受験者の解答心理を追究するため、各項目に解答した直後に、その「項目の単語」に対する理解度を偶数番号の「回答欄」に4段階(5－かなり知っている単語　4－何となく意味がわかる単語　2－見たことはあるが意味は分からない単語　1－見たこともないし、意味もわからない単語)で評価させることとした。たとえば、項目1、3、5の解答が３、４、１で、理解度がそれぞれ、5、4、2の場合、100 項目のマークカードの1、3、5番の解答欄に3、4、1をそれぞれマークし、2、4、6番の回答欄に4、4、2とマークを入れることになる。テスト解答及び理解度回答は合計で100 項目になる。理解度の 4 段階評価に3を含めなかったのは、実際の理解度について学習者がよく考えないままに中間値を選ぶことがないように配慮し、彼らの肯定的回答(5、4)と否定的回答(2、1)の心理的な差をより顕著に反映させることを意図したものである。図1は、奇数番号の1番と3番の項目に解答しながら、偶数番号の2番と4番で理解度をチェックさせるシステムを説明したものであり、実施前に受験者に配布した説明プリントからの抜粋である。

例（Example）

1　clock
2　1 の理解度　(your degree of knowledge about the word 'clock' for question 1)
3　girl
4　3 の理解度　(your degree of knowledge about the word 'girl' for question 3)

(1) time
(2) woman
(3) ship
(4) leg
(5) hat

図1　VKS への解答・回答方法説明で使用した例　(テスト前配布の解説プリントからの抜粋)

3. 2. 被験者

下位級の語彙を多く含む VKS1 は、日本の5大学で学ぶ 213 名の学習者に実施した。これらの学習者の一部は留学生であることが確認できているが、正確な数は把握できなかった。一方、より難度の高い級の単語を多く含む VKS2 は、日本の5大学で学ぶ 134 名の学習者に実施した。そのうちの 1 名は韓国人留学生で、その他は日本人学習者であった。VKS1 を受験した 213 名のうち、2大学 30 名の受験者に対しては VKS1 受験後、数日から 1 か月程度の間隔で VKS2 も実施した。これらの 30 名を加えた 164 名の VKS2 テストデータについても別途分析し、同一受験者の異なるテスト間での能力推定の正確さについて検証を行った。

3. 3. 分析

VKS の規準設定の有用性を検証するため、下記の研究課題を掲げることとした。

1. VKS は SCELP と比較して、どのような利点や問題点があるか。(VKS の利点や問題点)
2. VKS の項目の難易度はどの程度語彙レベルと関連していたか。(難易度と語彙レベルの関係)
3. ラッシュモデルと潜在ランク理論 (LRT) の分析手法を用いることで、いかにして VKS に対して説得力のある規準設定を行うことができるか。(ラッシュモデルと LRT を使った規準設定)
4. VKS のような受容語彙力テストの結果から、学習者へどのような診断的フィードバックを提供することが可能か。(学習者への診断的フィードバックの可能性)
5. VKS の項目難易度や他の項目情報は、規準設定においてどのような意味を持っているか。(規準設定における項目情報の意味)

分析は Excel 2 010 に入力されたデータを基に、ラッシュモデルの分析には WINSTEPS Version 3. 81.0 (Linacre, 2014)、潜在ランク理論の分析には Exametrika Version 5. 3 (荘島 2011)を使用した。基礎統計値や相関等は、Excel の表計算で処理し、信頼性等の一部分析には、IBM SPSS Statistics Version 20 も用いた。英検の習熟度級区分と他の語彙レベル指標と比較するため、AG eneral Service List of Words と The Academic W ord L ist に基づく

Heatley, Nation & Coxhead (2002) の Range プログラムと JACET 8000 LEVEL MARKER (http://www.tcp-ip.or.jp/~shim/J8LevelMarker/j8lm.cgi) を使って語彙レベル分析も行った。さらには、項目難易度と項目や選択肢の単語の特徴との関係を探るため、Graesser, McNamara, Louwerse & Cai (2004) が詳しく解説している単語やと文法構造の特徴を解析するツールである Coh-Metrix Version 3.0 も分析に使用した。

4. 結果

4. 1. VKS と SCELP の比較

VKS は Version 1、 Version 2 ともに平均点は 30 点 (正答率 60%) 前後で、正答率が約 68%に達していた SCELP に比べて、テストの難易度は受験者集団全体の能力により適応した関係にあったと言える。SCELP は一つの教育機関を対象としたクラス編成の目的で作成されたが、VKS は全国の大学生の英語語彙能力を想定して、より広範な規準設定に活用することを目指すものであったため、意図した結果がある程度実現できたと言える。しかしながら信頼性係数においては、問題数の差を考慮しても SCELP (80 問) のデータ (.949) に比べて低く、特に大きな利害を伴う現実の規準設定の場面では、VKS2 の信頼性の数値水準は問題になる可能性が高い。

表1
Version 1 (VKS1) の基本統計量

受験者数	項目数	素点平均	最頻値	中央値	標準偏差	最高点	最低点	KR20
213	50	30.8	35	31	8.2	49	10	0.888

表2
Version 2 (VKS2) の基本統計量

受験者数	項目数	素点平均	最頻値	中央値	標準偏差	最高点	最低点	KR20
134	50	28.7	31	29	5.5	42	10	0.757

VKS にテストとして問題があったとすれば、どこに問題があったのか、ラッシュモデルの適合度指数と潜在ランク理論のグラフを参考に検証することとする。

Beglar (2010) は、ラッシュモデルの応答適合度の指標であるインフィット平均平方値 (mean square、以降、MnSq) と標準化されたインフィット値 (standardized infit、以降、t 値) が +2.00 を上回る項目をアンダーフィットと見なし、その結果、実施した語彙サイズテスト 140 項目中5項目がアンダーフィットであったとしているが、SCELP についても同じ基準でアンダーフィット項目がないか調べたところ、80 項目中5項目が、t 値においてのみ基準値を上回った (法月, 2013)。

これに対して、VKS1 においては 50 項目中6項目が t 値においてのみ基準値を超えたが、VKS2 においては2項目にとどまった。

一方、小泉・飯村 (2010) は項目と受験者のインフィット MnSq が 0.70～1.30 の範囲を超え

る場合をミスフィットと呼んでいるが、この基準で SCELP を点検したところ、1.30 を超えるアンダーフィットは、2項目 (2.5%)、11 名 (7.3%)、0.70 を下回るオーバーフィットは、0項目、7名(4.0%)であった (法月、2013)。これに対して、VKS1 では、特に問題となるアンダーフィットは2項目(4.0%)、オーバーフィットは 0 項目であったが、受験者のアンダーフィットは 19 名(8.9%、うち 2 名は>+2.00)、オーバーフィットも 19 名 (8.9%) に達し、VKS2 では、項目はアンダーフィット、オーバーフィットともに 0 項目であったが、受験者はアンダーフィットが 20 人(15%)、オーバーフィットも 10 人 (7.5%) だった。

VKS1 でインフィット MnSQ が 1.30 を超えた項目 1091 (MnSq=1.39) と項目 1097 (MnSq=1.54) について、潜在ランク理論の名義モデルの分析を通じて、項目参照プロファイルのグラフで点検すると、以下のような解答様式が示された。

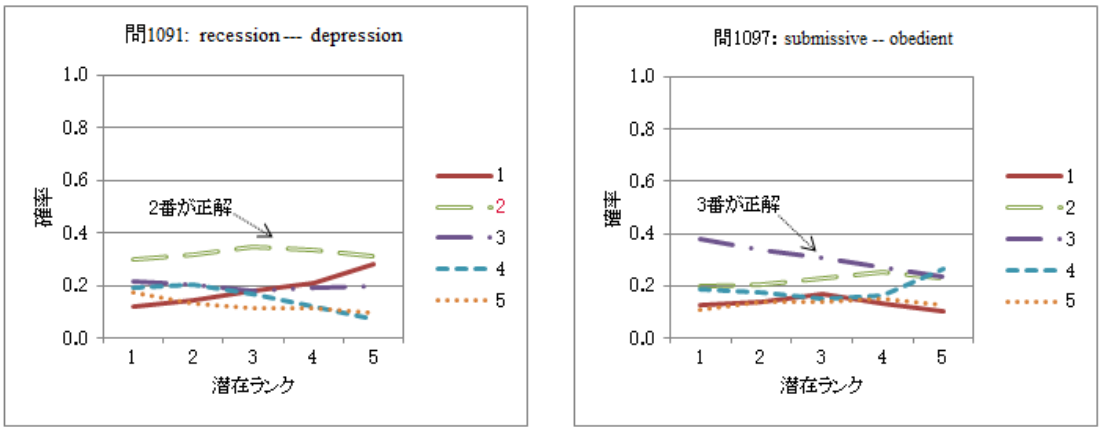


図2 ミスフィット項目の解答様式

問 1091 は VKS1 の 91 番目の解答・回答項目でテストの 46 番目の項目になるが、recession (不況) の同義として選ばれるべき2番の正答選択肢 (depression) を選んだ受験者の割合はランク3でピーク(.345)に達するものの、ランク4で.333、ランク5で.309 と微減している。ランク間でほとんど正答率が変わらない正答選択肢に対して、1 番の誤答選択肢 (circumstance) は、ランクが上がるにつれて選択率が上がっていることがわかる。

一方、VKS1 の 49 番目のテスト項目である問 1097 は、submissive (従順な) と同義である3番選択肢 (obedient) がランク1の受験者では.377 の選択率であるが、ランクが上がるにつれて数値は下がり、ランク5では.234 まで落ち込んでいる。これに対して、誤答選択肢2 (courteous) と4 (partial) は、ランク5で正解選択肢とほぼ同程度の選択率になっている。

いずれの項目も準1級項目で、項目難易度はラッシュ値で問 1091 が 0.91、問 1097 は 1.03 と VKS1 の中ではかなり高くなっている。ラッシュモデル適合度分析の個別の応答様式を詳しく見ると、ミスフィットの要因になった受験者の大きな残差は、問 1091 の場合は 22 人中 21 人が能力推定値の低い受験者による正解を示す+2.00 以上の解答に起因しており、問 1097 番

は 30 人全員が+2.00 以上を示している。このことから、これら2項目には、習熟度の低い受験者を正解に導くような、語彙力以外の何らかの構造的要因が働いていた可能性がある。

ここまで述べてきた数値の検証は、VKS の問題点を示唆する内容が多かったが、SCELP に無い2つの特徴も利点として確認できた。1点目は、解答の正誤や応答様式だけでなく、受験者の理解度や各項目の理解困難度を点検できることであり、2点目は、2つのバージョンのテストによって、SCELP よりも広域の受験者能力を測定できる可能性が高いことである。

理解度のチェックについては、SCELP においても補足的な実験への参加者に対してのみ、テスト解答終了後に総括的な調査を実施したが、80 問すべての項目について問題を解きながら理解度を回答させることは徹底できなかった。VKS は各バージョンのテスト解答を50問に絞り込み、テスト項目、理解度項目を交互に配置することで、すべてのテスト受験者にすべての解答項目についての理解度をテスト解答の直後に、逐次記録させるシステムを設けた。

理解度は、前述の通り、5、4、2、1の4段階の評定を基本としたが、「3」の回答が1つ以上見られた受験者は、VKS1 で 29 人(13.6%)、VKS2 で 18 人（13.4%） あり、VKS1 で最大 10 回答、VKS2 で最大 9 回答の受験者が見られた。大半は 1、2回の回答に限られ、偶発的なミスや無意識下での選択、もしくは4と2のいずれにもあてはまらないと「例外的な判定」をしたことによると考えられるが、「多数回答者」は試験開始前の指示書を使った指示内容が、理解できていなかった可能性が高い。分析の過程で、「3」の回答を削除する方法も検討したが、テスト得点との相関や理解度の信頼性に大きな変化が見られなかったため、「3」の回答をそのまま「4」、「2」の「中間評定値」と見なして、削除せずに5段階評定として扱い、分析を行った。また、VKS1 の理解度アンケートに無回答の受験者が 1 名あったため、相関分析からは除去したが、全問「5」を選択した受験者については、実際に受験者がすべての項目の単語を「かなり知っている」と解釈したことを否定する根拠はないため、そのまま含めて分析した。

受験者得点と受験者理解度平均の相関は、VKS1 において .764、VKS2 では .640 とやや差が出たが、項目正解率と項目理解困難度平均の相関は VKS1 で .946、VKS2 では .932 といずれも高い数値になった。なお、VKS1 の信頼性は .94、VKS2 の信頼性は .92 だった。

表3は、VKS1 と VKS2 の項目難易度と項目理解困難度の平均とその標準化された平均を比較したものである。CHIPs は、尺度の中心が 50 になるようにラッシュ推定値を変換した標準化された得点である (Wright & Stone, 1979)。難易度は正誤の2値、理解度は「中間評定値3」を含めて5段階の評定尺度と数値の性格は異なるが、大まかな比較ができるものと仮定した。なお本分析では、VKS1、VKS2 ともに受験者能力の平均を尺度の中心に設定している。

表3
VKS1 (V1) と VKS2 (V2) の項目難易度と項目理解困難度 (CHIPs 値) の比較

	V1 難易度	V1 理解困難度	V1 標準平均差	V2 難易度	V2 理解困難度	V2 標準平均差
平均	46.0	47.9	-0.30	47.3	50.1	-0.41
分散	61.1	15.5	－	66.5	29.2	－

VKS1、VKS2 ともに、項目難易度よりも項目理解困難度のほうが全般的に高い数値を示していて、VKS2のほうがその差が大きくなっている。個別に項目の難易度と理解困難度を比較すると、VKS1内で相対的に易しい4級項目の差が大きく、いずれも理解困難度が高くなっている。その一方で、VKS1 内では相対的に難しいと考えられる準1 級項目については、それほど大きな差ではないが、10 項目中8項目までは難易度が理解困難度よりも高くなっている。これに対して VKS2 においてもテスト内では易しい準2級等の項目の理解困難度が難易度よりもかなり高くなっているが、VKS1 とは異なり、最も難しいと考えらえる 1 級項目の 12 項目中8項目についても理解困難度がわずかではあるが難易度を上回っている。

このことから受験者は VKS のバージョンの違いに関係なく、易しい項目には高い確率で正解する力を持っていても、十分に理解しているとは必ずしも考えていない傾向が高いようである。また、より習熟度が低い受験者が多いと思われる VKS1 の難関項目については、正解するだけの十分な知識を持っていなくても見たことや聞いたことがあったり、意味を誤って類推している傾向が高かったことが示されている可能性がある。その一方で、より習熟度が高い VKS2 の受験者は、知っている単語と知らない単語の区別がかなり正確にでき、単語に対する十分な理解はなくても、既存の知識や合理的なテスト解答方略を使って、より正確に答えを推測したり、導き出したりしていた可能性がある。

SCELP にない VKS の 2 つめの有益な特徴は、共通 12 項目を介して、等化が可能な難易度が大きく異なる2つのバージョンのテストを兼ね備えていることである。表4のように共通 12 項目の項目難易度の平均には、VKS1、VKS2 の各バージョンの受験者平均 CHIPs 値を 50 に設定すれば、大きな差が生じる。このことから VKS1受験者よりも VKS2 受験者の習熟度がかなり高いことが確認できる。本研究の主要データは VKS1 を受験した 213 名と VKS2 のみを受験した 134 名であるため、両バージョンのテストを受験した 30 名を VKS2 のデータに加えて 164 名のデータとして比較分析すると、難易度が異なるバージョンのテスト間で同じ受験者の能力がどの程度正確に測定できるかを調べることができる。

表4では、共通項目の VKS1 項目難易度の値を係留項目 (anchor items) として固定して、164 名の VKS2 データを VKS1 尺度に等化した場合の共通受験者 30 名の能力推定値を比較することができる。大きく項目難易度が異なる垂直等化 (vertical equating) にもかかわらず、両バージョンから得られた当該受験者集団の能力推定値は驚くほど一致していたと言える。

表4
VKS1、2 の共通項目の難易度差と等化された共通受験者の能力推定値 (CHIPs) の比較＊

	①V1 共通項目	②V2 共通項目	①－②標準平均差	③V1 共通受験者能力	④V2 共通受験者能力	③－④標準平均差
平均	50.2	42.7	1.10	57.1	57.2	-0.05
分散	24.2	70.4	－	13.5	9.4	－

＊②の欄は等化前のデータ、④の欄は VKS1 尺度に等化後の数値を示している

4. 2. 難易度と語彙レベルの関係

規準設定を行う前に項目難易度と項目理解困難度が語彙レベルとどのような関係にあったか検証することとする。表5は VKS1 項目の難易度と理解困難度 (CHIPs 値) が級別にどのように分布しているかを視覚化した図3の基礎データをまとめたものであり、法月 (2013) でも使用した四分位数 (quartile) の計算に基づくものである。表5の 75%は第3四分位、25%は第1四分位の地点を示している。

図3で級別の項目難易度を比較すると、3級問題の分布域が広く、最小値で4級問題よりも低く、最高値で準2級問題の最高値とほぼ同じ水準に達している項目があることがわかる。一方、2級問題の 75%(第3四分位)の地点は準1級問題の最小値の水準であるが、その最大値は、準1級問題の最大値を上回っている。

VKS1 の級別項目難易度を図3右側の級別項目理解困難度と比較すると、後者は前者に比べて、分布域は全体的に小さく、分布の重なりの度合いは大きくなっているように思えるが、最大値と最小値は級が上がるにつれて常に高くなっていることがわかる。大きな差ではないが、級間の語彙レベルの差を、受験者もそれとなく認識していたのではないだろうか。

表5
VKS1 の級別項目難易度・項目理解困難度 (CHIPs 値)

	4 級 難易度	3 級 難易度	準 2 級 難易度	2 級 難易度	準 1 級 難易度	4 級 理解難度	3 級 理解難度	準 2 級 理解難度	2 級 理解難度	準 1 級 理解難度
最大値	41.7	49.0	49.4	59.3	57.7	45.9	46.8	51.0	52.5	56.0
75%	38.9	44.9	48.3	52.4	56.1	44.8	46.2	49.4	51.0	54.7
中央値	36.1	42.1	46.9	50.7	54.7	43.2	45.2	47.2	49.8	53.3
25%	35.2	37.9	45.5	50.3	54.5	42.7	44.4	46.3	49.0	51.6
最小値	32.1	28.0	41.7	47.0	51.0	42.0	40.9	44.7	47.4	50.8

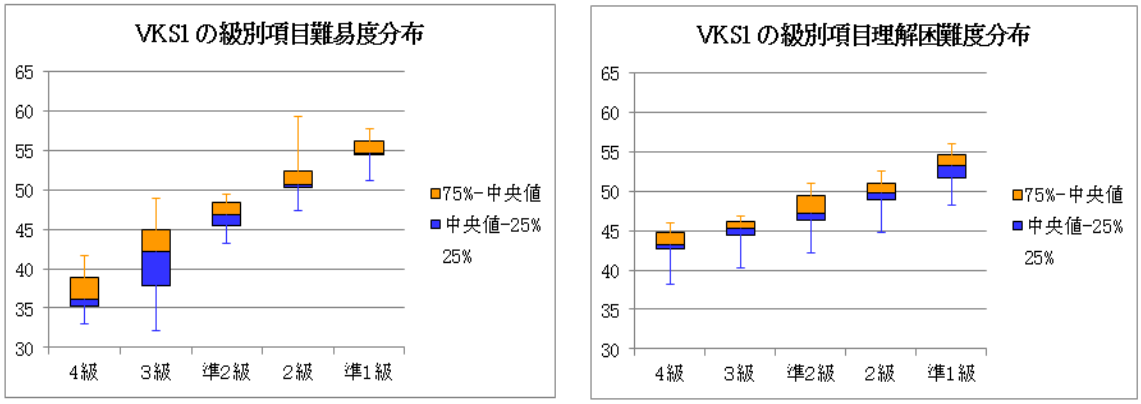


図3 VKS1 の級別項目難易度・項目理解困難度(CHIPs 値)分布比較

表6と図4は VKS2 項目の難易度と理解困難度の関係を示している。準2級・2級項目は、いずれも難易度が低く、準2級に含めて分析することが妥当と考えて、準2級 14 項目、2級以上

は各12項目と数えて、扱うこととした。VKS1と異なり、相対的に上位の2つの級（1級・準1級）と下位の2つの級（2級・準2級）の能力分布がより顕著な差として表れている。詳しくデータを見ると、特に2級の75%（第3四分位）の43.5と準1級の最小値の49.9の間には、実際に観測された項目は一つしかなく、後述の分割点の設定の議論を予測させる結果となっている。

表6
VKS2の級別項目難易度・項目理解困難度（CHIPs値）

	準2級 難易度	2級 難易度	準1級 難易度	1級 難易度	準2級 理解困難度	2級 理解困難度	準1級 理解困難度	1級 理解困難度
最大値	49.7	51.1	55.5	58.5	47.0	48.7	54.7	60.3
75%	41.2	43.5	54.2	58.0	46.2	47.5	54.4	58.5
中央値	37.9	41.8	53.2	56.2	45.2	46.3	52.9	57.9
25%	36.8	40.5	51.5	55.3	43.3	45.4	51.5	56.3
最小値	29.5	36.9	49.9	54.0	41.7	44.0	49.0	54.7

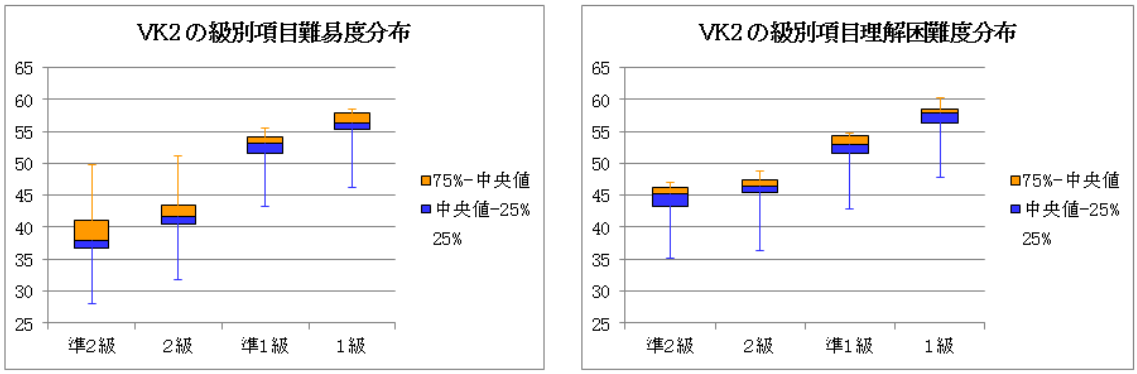


図4　VKS2の級別項目難易度・項目理解困難度(CHIPs値)分布比較

SCELPと同様に、VKS1とVKS2の結果から、ある語彙レベルが特定の学習者の能力水準に合致しているか否かは、大まかにしか判定することはできず、分割点設定に語彙項目の内容を明確に関連付けるためには、項目を難易度順に並べ替える必要がある。

4．3．ラッシュモデルとLRTを使った規準設定

規準設定の分割点を決定するために、法月（2013）及び大友（2013b）の中で述べられた方法に基づいて、ラッシュモデルの能力推定値と項目難易度をLRTのランク・メンバーシップ・プロファイル(RMP)の表に位置づけることとした。ランクの数や目標潜在分布の様式を複数検討したが、結局、各テストの対象級の数に合わせて、VKS1は5つのランク（4つの分割点）、VKS2は4つのランク（3つの分割点）を設けることとし、一様分布を指定した。規準設定の手順は、VKS1、VKS2ともに、以下の通りである。

①Exametrika の RMP の Excel 表内に受験者能力推定値（CHIPs 値）が入った列を挿入し、数値の高いほうがリストの上に来るように並べ替える。この表に「対応する項目」の難易度(CHIPs) や番号等の情報を追加する。対応する項目として、(A) その難易度がある受験者能力値と同じかそれよりも低く(正答確率 .50 以上)、(B) 次に高い受験者能力値よりもその難易度が上回っているものを表の受験者能力値の横の欄に記載した。(A)の条件を満たし、(B)の条件を満たす項目が無い場合は、その受験者能力推定値に最も難易度が近い項目を「対応する項目」として、その情報を記載した。

②①の並べ替えの際に CHIPs 値が同じでランクが異なる場合は、ランクの高いほうがリストの上に来るように設定する。

③各ランクに所属する確率も提示する。①の並べ替えの際に、②の条件に加えて、CHIPs とランクがともに同じ場合は、隣接する境界ランクの「より高い」ランクに所属する確率(例、境界ランクが5と 4 の場合は、5の確率)が高い方がリストの上に来るように設定する。

上記のような手順でデータの並べ替えを行った結果、VKS2 の最上位グループと2番目のグループの分割点候補と考えられるランク4と3の受験者グループ境界領域は、図5のような状況であることが確認された。

	A	B	C	D	E	F	G	H	I	J	K
1			潜在ランク	能力	対応する項目	Item Number	級	ランク・メンバーシップ・プロファイル			
2	正答数	正答率	推定値	CHIPs	CHIPs	δ(≤θ)	δ(≤θ)	Rank 1	Rank 2	Rank 3	Rank 4
14	36	0.720	4	54.55	54.19 & 54.00	問2057. 2069. 2073 & 2087	準1級&1級	0.000	0.001	0.118	0.881
15	36	0.720	4	54.55	54.19 & 54.00	問2057. 2069. 2073 & 2087	準1級&1級	0.000	0.003	0.127	0.870
16	36	0.720	4	54.55	54.19 & 54.00	問2057. 2069. 2073 & 2087	準1級&1級	0.000	0.006	0.187	0.807
17	36	0.720	4	54.55	54.19 & 54.00	問2057. 2069. 2073 & 2087	準1級&1級	0.000	0.005	0.304	0.691
18	36	0.720	3	54.55	54.19 & 54.00	問2057. 2069. 2073 & 2087	準1級&1級	0.001	0.057	0.626	0.316
19	35	0.700	4	53.913	53.5035	問2067	準1級	0.000	0.002	0.145	0.853
20	35	0.700	4	53.913	53.5035	問2067	準1級	0.000	0.003	0.147	0.850
21	35	0.700	4	53.913	53.5035	問2067	準1級	0.000	0.011	0.176	0.813
22	35	0.700	4	53.913	53.5035	問2067	準1級	0.000	0.005	0.248	0.747
23	34	0.680	4	53.276	52.821	問2065	準1級	0.000	0.022	0.476	0.502
24	33	0.660	4	52.639	51.41 & 51.09	問2063 & 2035	準1級&2級	0.000	0.005	0.112	0.883
25	33	0.660	4	52.639	51.41 & 51.09	問2063 & 2035	準1級&2級	0.000	0.012	0.148	0.840
26	33	0.660	4	52.639	51.41 & 51.09	問2063 & 2035	準1級&2級	0.000	0.002	0.204	0.793
27	33	0.660	4	52.639	51.41 & 51.09	問2063 & 2035	準1級&2級	0.000	0.020	0.189	0.790
28	33	0.660	4	52.639	51.41 & 51.09	問2063 & 2035	準1級&2級	0.000	0.006	0.282	0.713
29	33	0.660	4	52.639	51.41 & 51.09	問2063 & 2035	準1級&2級	0.000	0.005	0.292	0.703
30	33	0.660	4	52.639	51.41 & 51.09	問2063 & 2035	準1級&2級	0.000	0.017	0.308	0.675
31	33	0.660	4	52.639	51.41 & 51.09	問2063 & 2035	準1級&2級	0.000	0.013	0.362	0.624
32	33	0.660	4	52.639	51.41 & 51.09	問2063 & 2035	準1級&2級	0.000	0.007	0.396	0.597
33	33	0.660	3	52.639	51.41 & 51.09	問2063 & 2035	準1級&2級	0.004	0.156	0.611	0.229
34	32	0.640	4	52.0475	51.41 & 51.09	問2063 & 2035	準1級&2級	0.000	0.016	0.468	0.516
35	32	0.640	3	52.0475	51.41 & 51.09	問2063 & 2035	準1級&2級	0.000	0.024	0.618	0.358
36	32	0.640	3	52.0475	51.41 & 51.09	問2063 & 2035	準1級&2級	0.001	0.058	0.677	0.264
37	32	0.640	3	52.0475	51.41 & 51.09	問2063 & 2035	準1級&2級	0.003	0.153	0.758	0.086
38	32	0.640	3	52.0475	51.41 & 51.09	問2063 & 2035	準1級&2級	0.007	0.220	0.691	0.082
39	32	0.640	3	52.0475	51.41 & 51.09	問2063 & 2035	準1級&2級	0.036	0.177	0.709	0.078
40	31	0.620	4	51.4105	51.41 & 51.09	問2063 & 2035	準1級&2級	0.000	0.016	0.296	0.687
41	31	0.620	3	51.4105	51.41 & 51.09	問2063 & 2035	準1級&2級	0.001	0.060	0.581	0.358
42	31	0.620	3	51.4105	51.41 & 51.09	問2063 & 2035	準1級&2級	0.002	0.083	0.568	0.347
43	31	0.620	3	51.4105	51.41 & 51.09	問2063 & 2035	準1級&2級	0.001	0.040	0.638	0.321
44	31	0.620	3	51.4105	51.41 & 51.09	問2063 & 2035	準1級&2級	0.004	0.074	0.607	0.316
45	31	0.620	3	51.4105	51.41 & 51.09	問2063 & 2035	準1級&2級	0.000	0.035	0.668	0.297
46	31	0.620	3	51.4105	51.41 & 51.09	問2063 & 2035	準1級&2級	0.003	0.086	0.664	0.247
47	31	0.620	3	51.4105	51.41 & 51.09	問2063 & 2035	準1級&2級	0.001	0.049	0.719	0.231
48	31	0.620	3	51.4105	51.41 & 51.09	問2063 & 2035	準1級&2級	0.004	0.120	0.662	0.213
49	31	0.620	3	51.4105	51.41 & 51.09	問2063 & 2035	準1級&2級	0.003	0.091	0.772	0.134
50	31	0.620	3	51.4105	51.41 & 51.09	問2063 & 2035	準1級&2級	0.002	0.111	0.756	0.131
51	31	0.620	3	51.4105	51.41 & 51.09	問2063 & 2035	準1級&2級	0.005	0.109	0.757	0.129
52	31	0.620	3	51.4105	51.41 & 51.09	問2063 & 2035	準1級&2級	0.022	0.283	0.590	0.106
53	31	0.620	3	51.4105	51.41 & 51.09	問2063 & 2035	準1級&2級	0.010	0.299	0.590	0.102
54	31	0.620	3	51.4105	51.41 & 51.09	問2063 & 2035	準1級&2級	0.005	0.217	0.682	0.096
55	31	0.620	2	51.4105	51.41 & 51.09	問2063 & 2035	準1級&2級	0.249	0.615	0.131	0.004
56	30	0.600	3	50.819	50.77 & 50.46	問2049 & 2053	2級&準1級	0.011	0.082	0.526	0.382
57	30	0.600	3	50.819	50.77 & 50.46	問2049 & 2053	2級&準1級	0.002	0.053	0.683	0.262
58	30	0.600	3	50.819	50.77 & 50.46	問2049 & 2053	2級&準1級	0.006	0.214	0.635	0.145

図5 VKS2 の最上位グループ決定の分割点候補

一番上の赤丸内は正答率が .72 で最初に潜在ランク推定値（一様分布）が3を示した位置をマークしている。この上の4の地点を最上位グループの境界地点にすることも可能だが、正答率が同じ .72 であるため、テストの利害関係者の多くに、不公平感を与える結果になりかねない。一方、2番目の赤丸の .66 の地点にもランク 4／3の変動が見られるが、正答率が最も低い4まで含めると .62 の地点になるため、同率正答者の分割を避けるならば、4番目の黄色の塗りつぶし線で分割することになる。この図に示されているだけで、3つの赤丸と4番目の黄色の塗りつぶし線の4もしくは5地点を、分割点候補として考えることができることがわかる。法月 (2013) の研究では、上位グループの可能性があっても下位グループに位置づけられてしまうことを避けるため、当該ランクの一番低い地点の正答率が途切れるところ（この場合ならば、正答率が .62 からそれ未満に代わる直前 <4番目の黄色の塗りつぶし線>）の地点に分割点を定めたが、今回のデータではそのような分け方をすると、上記の例で最上位グループが、134 名の受験者中 53 名に達するなど、VKS1、VKS2 ともに上位グループが膨らみ、最下位グループがほとんど残らなくなる状況だった。そこで、ランクが替わり、かつ正答率が下がる最も高い地点に分割点を置くこととした。これにより、VKS2 の最上位のレベル4 (V2L4) のグループの最下位能力値は.720 で、ランク3の受験者 1 名までが含まれる結果となった。

VKS1 は 4 級～準 1 級までの5段階、VKS2 は準 2 級～1 級までの 4 段階の級を照準にした語彙テストであるため、そのような観点からラッシュモデルと LRT の分析の結果を進め、最終的に級区分と同じ数の区分で行った各テストの規準設定の結果を、表7、8にまとめてある。VKS1 の問 1081 は上記の規準設定手順①の(B)の条件を満たしていないが、V1L5 分割点能力推定値の受験者の正答確率が .50 を超え、他の項目の難易度よりもこの地点に近接しているため、V1L5 の下限に対応する項目として位置づけた。VKS2の問 2037 も同様の理由で、V2L2 分割点下限に対応する項目として位置づけた。いずれの能力別グループ内にも複数の RMP ランクが混在する結果となったが、VKS1、VKS2 テストともにグループ区分の番号と主要ランクの数字が一致する結果となった。今回は上位グループを少なめに分割したため、下位グループの人数が多くなる傾向が見られるが、現実のクラス編成では、レベル内を任意の正答率で区分したり、無作為に等分割することも理にかなっていると思われる。

表7
VKS1テストによる能力グループの規準設定案と潜在ランク及びラッシュ (CHIPs) 値の関係

VKS1	能力値域	難易度域	能力分割域	分割域項目No.	R5 (人数)	R4	R3	R2	R1
V1L5	68.4-56.2	59.3-55.3	56.2-55.5	1081*	27	1	0	0	0
V1L4	55.5-53.5	54.7-53.1	53.5-52.8	1069	12	21	1	0	0
V1L3	52.8-49.9	52.5-49.4	49.9-49.3	1051	0	23	29	1	0
V1L2	49.3-46.5	49.0-46.2	46.5-45.9	1055	0	0	16	29	1
V1L1	45.9-36.5	45.6-28.0	－	－	0	0	0	16	36

*L5 と L4 の分割点領域に位置する項目 1081 の難易度は L4 の上限受験者の能力推定値よりも低いが、V1L5 分割点に対応する項目として分類

表 8
VKS2 テストによる能力グループ規準設定案と潜在ランク及びラッシュ (CHIPs) 値との関係

VKS2	能力域	難易度域	能力 分割域	分割域 項目 No.	R4 (人数)	R3	R2	R1
V2L4	58.8-54.6	58.5-54.0	54.6-53.9	2067	15	1	0	0
V2L3	53.9-51.4	53.5-51.1	51.4-50.8	2035	16	20	1	0
V2L2	50.8-49.0	50.8-43.7	49.0-48.3	2037*	0	16	18	1
V2L1	48.3-37.4	43.5-29.5	－	－	0	1	17	28

*L2とL1の分割点領域に位置する項目2037はL1の上限受験者の能力推定値よりも低いが、V2L2分割点に対応する項目として分類

表9と表 10 は、VKS1 と VKS2 の規準設定が、級別の項目の分布にどのように対応しているかをまとめたものである。両テストとも図3、4の分布からも確認された通り、級の区分できれいに能力グループが分かれることはなかったが、VKS1 については、習熟度の最も高い V1L5 とその次の V1L4 において、この問題の最も高い級である準 1 級項目が最も多く照準化されており、中間水準の V1L3 では2級、V1L2 では準 2 級が照準の中心となり、最も習熟度の低い V1L1 グループには3級や4級項目が多く含まれていることがわかる。

級別項目の分布状況は、VKS2 においても VKS1 と類似の傾向が見られた。最上位グループの V2L4 は 1 級語彙項目中心、V2L3 は準 1 級中心、V2L2 は2級項目、V2L1 は準2級項目中心へと分布は変わっていく。VKS1、VKS2 とも明確な区分ではないが、習熟度の高いグループほど、より上位級の単語の難易度が受験者の能力に適応している傾向が確認できる。

表 9
VKS1 の級別項目規準設定結果

VKS1	項目数	受験者数	1級項目	準1級項目	2級項目	準2級項目	3級項目	4級項目
V1L5	5	28	0	4	1	0	0	0
V1L4	6	34	0	5	1	0	0	0
V1L3	8	53	0	1	6	1	0	0
V1L2	9	46	0	0	2	5	2	0
V1L1	22	52	0	0	0	4	8	10
合計	50	213	0	10	10	10	10	10

表 10
VKS2 の級別項目規準設定結果

VKS2	項目数	受験者数	1 級	準1級	2級	準2&2級	準2級	3級	4級
V2L4	17	16	12	5	0	0	0	0	0
V2L3	6	37	0	5	1	0	0	0	0
V2L2	5	35	0	2	2	0	1	0	0
V2L1	22	46	0	0	9	2	11	0	0
合計	50	134	12	12	12	2	12	0	0

VKS 区分の妥当性を検証するため、VKS1 による5段階、VKS2 による4段階の難易度区分を英検の級、JACET 8000 の語彙レベル区分(レベル1～8＋リスト外のレベル9)、Healey, Nation & Coxhead (2002) の Range のプログラムの区分（レベル 1～3＋リスト外のレベル4）と

比較した。JACET 8000 と Range については、各ブロック(2項目＋5選択肢)の単語の語彙レベルの平均値を使用した。相関分析の結果、VKS1 は、英検の級区分と .869、JACET 8000 とは .754、Range とは .815 の相関を示し、VKS2 の相関は英検と .897、 JACET 8000 と .883、Range とは .887 だった。いずれもかなり高い数値を示していることから、VKS の区分は妥当であったと考えられる。

表4の VKS1 と VKS2 の共通項目の難易度差から、両テストの受験者の能力に大きな差が存在することは明白であるが、今回の分析で使用したすべての項目に対して大まかな位置づけを行うことを目的に、共通 12 項目を VKS1 の難易度値に係留し、VKS2 の残りの項目を VKS1 の尺度に等化させることにした。これにより、VKS1 と VKS2 の規準設定の結果を統合して、表 11 のような結果を得ることができた。

VKS1 の値に係留した共通 12 項目を除く 76 項目と両テストの受験者 347 名の比較を行うと、VKS2 項目はやや低いランクに分類(*例、L1 とL2 の2級項目はすべてVKS2、VKS2 の準2・2 級項目はすべてL1*)される傾向が見られたが、L1 から L5 までは VKS1 の尺度で分割し、L6 は VKS2 の L4 や VKS1 の L5 の最上位項目・受験者を位置づけることでスムーズに分類することができた。係留項目の難易度と VKS2 の相当項目の元々の難易度の値の差異からも誤差の大きい垂直等化であることは確実であるため、あくまでも参考分析であり、VKS1 で高い能力値を示した受験者が VKS2 で同じ水準に達するとは限らないが、VKS1 と VKS2 の間のレベルの問題を開発すれば、さらに汎用性が広がる可能性も考えられる。

表 11
VKS2 を VKS1 に等化した場合の規準設定結果

VKS1&2	項目数	受験者数	1級(項目)	準 1 級	2級	準2&2級	準2級	3級	4級
L6	16	19	12	4	0	0	0	0	0
L5	9	78	0	5	3	0	1	0	0
L5／L4*	0	8	－	－	－	－	－	－	－
L4	5	61	0	4	1	0	0	0	0
L4／L3*	0	9	－	－	－	－	－	－	－
L3	10	71	0	1	8	0	1	0	0
L2	11	47	0	0	5	0	4	2	0
L1	25	54	0	0	1	2	4	8	10
合計	76	347	12	14	18	2	10	10	10

＊VKS1の尺度で L1 から L5 まで分割した際に、隣接する上位側グループの下限能力値と下位側グループの上限能力値間の数値を示した VKS2の受験者は、便宜的に L5／L4 と L4／L3 に分類

4. 4. 学習者への診断的フィードバックの可能性

一般的に受験者は易しい項目に正解して、難しい項目には正答する確率が低くなるが、表 12 のE1のように、級が高くなるにつれて正答率が低くなる明瞭な特徴を示す受験者はそれほど多くない。一見すると不規則な正答率のパターンが、学習者の学習履歴や背景を反映している可能性もある。

E2は3級問題で 70%、準2級問題で正答率が 50%まで下がるが、2級問題では正答率が

80%に達している。この受験者は VKS1を受験する1カ月ほど前から2級の試験勉強を始めているため、その効果が影響を与えているかもしれない。E3は、VKS1 の中で最も易しいと考えられる4級問題の正答率 (10%) が最も低く、最も難しい後半の2級問題や準1級問題 の正答率(60%、40%)が顕著に高くなっている。受験者が選択した理解度の数値を見ると、正答率に比例した変化は見られないため、当て推量などによる偶然の正解の可能性が高いと言えよう。E4もE3と同様に4級問題の正答率(10%)は低いものの、準2級、2級問題の正答率(いずれも 50%)は他の級に比べて高めになっている。E4はE3とは対照的に、理解度においても正答率が高い準2級、2級問題、準1級問題は、正答率が低い4級、3級問題に比べてもかなり高くなっていることがわかる。

E3とE4のような難項目の正答率が高くなる傾向を示した受験者が全部で3名あったが、いずれもラッシュモデルの分析で受験者ミスフィットを示している。しかしながら、理解度の数値を参照すると、E4とE3の受験者とでは全く異なる扱いが必要かもしれない。中学校、高校時代にあまり英語を勉強していなかった学習者が大学に入ってからの英語の授業や専門の授業の中で、一見すると難解な単語を習得していたならば、E4のような学習者に対してはE3とは異なる語彙指導やアドバイスが必要になるかもしれない。

自己評価の理解度が高い項目に対して不正解が目立つ受験者に対しても注意を喚起させる必要があるかもしれない。E2の学習者は VKS1 と VKS2 を両方受験してくれたため、共通 12 項目の成否の理由について、試験後、確認したところ、接頭・接尾辞の不十分な知識を過剰に一般化してしまい、全く誤った解釈をしていたところもあった。よく知らない単語について類推する力は必要であり、この学習者に大きな問題があったとは言えないが、過剰解釈する傾向の強い学習者には必要に応じて適切なアドバイスが求められるかもしれない。

表 12
受験者の級別単語正答率・理解度(平均)比較の例(VKS1)

受験者	4級 正答率	4級 理解度	3級 正答率	3級 理解度	準2級 正答率	準2級 理解度	2級 正答率	2級 理解度	準1級 正答率	準1級 理解度	テスト 正答率	理解度 全平均
E1	100%	5	80%	4.9	60%	4.1	50%	3.3	30%	1.8	74%	4.22
E2	100%	5	70%	4.6	50%	4.3	80%	4.4	30%	2.8	66%	4.22
E3	10%	2.4	20%	2	20%	2	60%	2	40%	2.2	30%	2.12
E4	10%	1.9	20%	1.7	50%	2.2	50%	3.2	30%	2.4	32%	2.28

4. 5. 規準設定における項目情報の意味

VKS1 は5段階、VKS2 は4段階の規準を設定することが可能で、共通項目を通じて VKS2 を VKS1の尺度に等化させることで、6段階の規準を設定できることも確認した。現実的なクラス編成においては、さらなる分割を講じる必要もあるが、分析を通じて得られた各規準について一定の意味づけを行わない限り、統計手法を使った分割点設定も恣意的と見なされかねない。各規準を具体的かつ客観的にどのように定義するかについては、別の研究の機会に

委ねたいが、本稿では類似の難易度項目群や分割点周辺の基本的な特徴を探ることで、規準設定における項目難易度や他の項目情報の意味について探索することとする。

表 13 は VKS1 項目の難易度順1～5番、16～20 番、26～30 番の項目の難易度や理解困難度等の情報が項目の単語と正解選択肢の単語とともに提示されている。1～5番はいずれもこのテストで最も難しいレベルであるV1L5であるが、抽象度が高く視覚化しにくい動詞が多く含まれている。16～19 番は V1L3 に属し、テスト尺度の中央付近の値を示す項目であるが、1～5番に比べてより日常的なレベルで使われる概念の単語が多い。V1L2 レベルの 20 番になるとさらに日常性を増している。29、30 番はこのテストで最も易しいレベルの V1L1 になるが、日本語に訳すと、イメージし易い単語と言えるかもしれない。

表 13
VKS1 項目の規準設定例

難易度順	級	項目番号	項目通過率	理解困難度(1-5)	難易度(CHIPs)	理解困難度(CHIPs)	レベル	問単語	正解選択肢	品詞
1	2	1067	16.0%	2.3	59.3	52.5	VlL5	urge	recommend	動
2	準1	1083	20.2%	1.6	57.7	56.1	VlL5	preclude	hinder	動
2	準1	1095	20.2%	1.9	57.7	54.2	VlL5	compound	hybrid	名／形
4	準1	1089	24.4%	2.5	56.4	51.5	VlL5	scholarship	grant	名
5	準1	1081	28.2%	2.3	55.3	52.4	VlL5	exceed	surpass	動
16	2	1061	48.4%	3.0	50.4	49.6	VlL3	divide	separate	動
16	2	1073	48.4%	2.9	50.4	50.0	VlL3	quantity	amount	名
18	2	1077	48.8%	2.6	50.3	51.0	VlL3	aware	conscious	形
19	準2	1051	52.6%	2.7	49.4	51.0	VlL3	faith	belief	名
20	3	1031	54.5%	3.8	49.0	46.8	VlL2	dining	meal	名
26	2	1075	62.4%	3.3	47.0	48.6	VlL2	merit	benefit	名
27	3	1023	63.8%	3.8	46.7	46.6	VlL2	reach	arrive	動
28	準2	1055	65.7%	3.9	46.2	46.3	VlL2	average	medium	名／形
29	準2	1041	68.1%	3.6	45.6	47.5	VlL1	fix	repair	動
30	準2	1049	69.0%	3.8	45.4	46.9	VlL1	effect	influence	名

表 14 は VKS2 の項目例であるが、やはり最難関の V2L4 は抽象度の高い動詞が多くなっている。最も難しい項目 ‘plague’ については、理解困難度は他の難関項目に比べて低いため、名詞（伝染病）の意味は比較的よく知られているものの、動詞（苦しめる）として使われる意味については、なじみがなかったのではないかと思われる。V2L4 の分割点に対応している項目 2087 は全項目の中で最も理解困難度が高くなり、V2L3 に対応する最も難しい2項目が、日本語でよく使われる「ハイブリッド車」や大学生の多くに身近なテーマである「奨学金」のようになじみがありそうな語彙項目であるにもかかわらず、VKS1 でも最難関レベルの V1L5 に位置していることが興味深い。28 番目の項目の難易度は分割点に位置する受験者の能力推定値と大きく離れているが、テスト内で他に対応する項目がなく、便宜的に V2L2 の下限項目に位置づけられている。27 番目の項目の通過率とは大きな隔たりがあり、V2L1 レベルの 29、30 番と同様に、他の項目に比べて、日常的な語彙事項になっている。

表 14
VKS2 項目の規準設定例

難易度順	級	項目番号	項目通過率	理解困難度(1-5)	難易度(CHIPs)	理解困難度(CHIPs)	レベル	問単語	正解選択肢	品詞
1	1	2081	15.7%	1.7	58.5	54.7	V2L4	plague	torment	動
2	1	2077	16.4%	1.3	58.2	58.1	V2L4	quell	stifle	動
3	1	2083	17.2%	1.5	58.0	56.1	V2L4	censure	rebuke	動
3	1	2095	17.2%	1.5	58.0	56.6	V2L4	arbitrary	capricious	形
5	1	2097	18.7%	1.3	57.5	57.8	V2L4	genial	amicable	形
16	準1	2073	31.3%	1.7	54.0	54.7	V2L4	submissive	obedient	形
16	1	2087	31.3%	1.3	54.0	59.0	V2L4	infringement	transgression	名
18	準1	2067	33.6%	2.3	53.5	52.1	V2L3	compound	hybrid	名／形
19	準1	2065	36.6%	3.3	52.8	49.1	V2L3	parallel	equivalent	形
20	準1	2061	41.0%	3.3	51.9	49.0	V2L3	scholarship	grant	名
26	準1	2071	50.7%	1.8	49.9	54.3	V2L2	yardstick	measure	名
27	準2	2015	51.5%	3.9	49.7	47.0	V2L2	policy	government	名
28	2	2037	77.6%	3.8	43.7	47.3	V2L2	gain	increase	動
29	2	2033	78.4%	3.5	43.5	48.5	V2L1	spoil	ruin	動
30	2	2051	79.1%	3.7	43.3	47.7	V2L1	enormous	huge	形

単語の難易度を決定する要因は恐らく無限にあると言えるが、近年、Coh-Metrix と呼ばれる文章の中で使用される語彙や文法構造の特徴を解析するオンラインツールを使って、テストの得点等との関係を分析する取り組みや議論が盛んに行われている (Graesser, et al., 2004; Graesser, McNamara & Kulikowich, 2011; Nation & Webb, 2011; Crossley, Salsbury & McNamara, 2012 等)。Coh-Metrix は語彙の場合でも文脈の中で分析することが基本であるが、文脈がなくても分析できる特徴もあるのではないかと考えて、VKS1 の項目及び選択肢の単語をブロック別に、VKS2 は項目と正解選択肢の単語のペアについて一部、分析を試みた。

VKS1 については、項目の難易度 (CHIPs) との相関を調べたところ、CELEX と呼ばれるコーパスに基づく単語の使用頻度との相関が -.778、単語の文字数との相関が .691、MRC 心理言語学データベースの英語母語使用者による評定データに基づく単語のイメージのしやすさとの相関が -.558、同評定データに基づく単語の親密度との相関が -.550、同評定データに基づく単語の具体性との相関が -.460、WordNet というオンライン語彙データベースに基づく多義性との相関が -.418 であった。

注意しなければならないのは、数はそれほど多くないが、いくつかの語彙指標が0になっていたことである。これは項目や選択肢の単語が語彙リストや評定のデータベースに含まれていないことを意味しているようである。また0になっていない場合でも項目もしくは選択肢の単語が一つでもリストにない場合、正確な計算はできていないものと考えられる。

VKS2 の一部の項目と正解選択肢のペアの単語を Coh-Metrix で分析したところ、リストに基づく指標の中には0が含まれるものが多く、その中で唯一0がなかった指標は WordNet に基づく多義性の判定であった。多義性については VKS1 の相関結果と同様に、難しい単語は特殊な単語が多いためかリスト化されている語義が少なく、易しめの単語には多くの語義が

含まれる傾向が確認できた。

Coh-Metrix の結果から VKS1については多くの語彙関連指標が語彙難易度に影響を与えている可能性が確認できたが、VKS2のようにやや特殊な語が多い問題についてはリスト外の単語が増えてしまい、VKS のように文脈がない問題については顕著な影響が出てしまうようである。

5. 考察

5つの研究課題に沿って分析を行ったが、分析結果を総括し、結果から示唆される問題点や今後の研究指針について議論する。

5. 1. VKS と SCELP の比較

VKS は、SCELP に比べて、信頼性係数が低めで、ミスフィット傾向もやや強かった。特定の教育機関を対象に開発した SCELP は実際の受講クラスを決定するプレイスメントテストとして実施したのに対して、今回の VKS を受験した学習者にはそのような現実的な目的意識がなかったことも結果に影響しているのかもしれない。しかしながら、VKS には SCELP にない2つの大きな特徴の効果が確認できた。

1番目の特徴は、テスト項目、理解度項目を交互に配置することで、すべてのテスト受験者にすべての解答項目についての理解度をテスト解答の直後に、逐次記録させるシステムを設けたことである。特に項目通過率と項目理解困難度の相関は両テストとも .9 を超える高さを示したが、両者の関係をラッシュモデルの CHIPs 値で比較すると、易しい項目には正答していても完全に理解して答えている自信の度合いは項目難易度に比べて低めで、VKS1 の難関項目においては正解するだけの十分な知識を持っていない受験者が誤った類推や過剰な一般化のためか理解していると思う度合いが高く、VKS2 については、上位学習者が十分な理解はなくても関連知識や解答方略を運用して正答を導いている度合いが高い可能性を示唆する結果が得られた。

受験者理解度と項目理解困難度が効果的に機能している結果から、客観的な比較はできないが、4項目の答えを5つの選択肢の中から組み合わせる形式の SCELP に比べて、当て推量だけで正解している解答の度合いは低くなっている可能性が高い。

2点目の特徴は、2つの難易度の異なるテストながら、共通項目を通じて等化を行うことが可能なことである。VKS1 と VKS2 を両方受験した受験者 30 名の能力推定値は、テスト間の難易度を係留項目によって調整することで、非常に高い一致度を示した。難易度差が非常に大きい2つのテストを等化することで、SCELP よりも広い能力層の受験者に対して規準設定を行うことができるようになったと言える。

VKS は両バージョンとも修正の余地を残しているが、同じ対象級の同義や類義語の項目と選択肢の単語のペアが、文脈が無い制約の中で各ブロック内の唯一の正解組み合わせにな

るような問題にすることは決して容易とは言えず、単語の多義性や曖昧性を考慮すれば、特に習熟度の低い受験者にとっては認知的負担が大きかった可能性がある。Elgort (2013) は、Nation and Beglar (2007) の Vocabulary Size Test (VST) における使用言語の効果について分析を行い、選択肢が学習者の第1言語（ロシア語）になる2カ国語（英語／ロシア語）使用版が英語のみの単一言語使用版に比べて、正答率が非常に高くなり、習熟度の低い学習者に対してより正確な測定を行うことができると主張しているが、VKS についても将来的には、若干の文脈を設けたり、2カ国語使用版を開発して効果を比較する価値もあるかもしれない。

5．2. 難易度と語彙レベルの関係

各テストの級別の項目難易度と項目理解困難度を比較すると、後者のほうが級の変化に緩やかに対応し、前者は級やテストにより、分布がかなり異なっていることが確認できた。

VKS1は特に3級の項目難易度が最も低い（易しい）項目において4級の最低値よりも低くなっており、逆に項目難易度が最も高い（難しい）項目においては準2級の最高値に近い水準になっている。同様に2級の項目難易度が最も高い項目がこのテストで最も習熟度の高い級である準1級の最高値を上回っている。

VKS1 受験者の中には、中学校、高校と必ずしも段階的に英語学習の習熟度を高めてきていない学習者も見られるため、全体としては級別の区分が機能しているものの、個別項目において級区分とはやや異なる結果を示しても、必ずしも特異な現象とは言えないだろう。

一方、VKS2 においては、下位級の準2級・2級と上位級の1級・準1級との難易度の差が非常に大きくなっている。VKS2 の受験者は習熟度の高い学習者が多いため、高校時代までに習得している比重の高い下位級の単語とまだ十分に習得しきれていない上位級の単語とでは大きな差が生じたのかもしれない。細かくデータを見ると、2級の第3四分位 (75%) と準1級の最低値との間に1項目しか含まれていないことがわかる。この付近は V2L2／L1 の境界領域であり、分割点設定の精度に問題があった可能性は否めない。

5．3. ラッシュモデルと潜在ランク理論を使った規準設定

本研究の規準設定の手続きは法月(2013) に基盤を置くものであったが、最終的に採択した方法は、それとはかなり性格の異なるものであった。法月（2013）では、当該ランクの最も低い領域の正答率が切り替わる地点に分割点を置いたが、本研究では当該ランクの最も高い地点の正答率が切り替わる地点が選ばれた。規準設定の目的にもよるが、受け入れ人数に制限のあるクラス編成や教育プロラムへの入学・参加許可においては、どの方法を選ぶかは実際の受験者のスコア分布に依存する傾向が高いのではないだろうか。Zeiky, Perie & Livingston (2008) が主張するように、「分割点は客観的に決めることはできないが、客観的に適用することができるもの」であるならば、2つの統計手法を使って、一定の条件下で分割点設定の方法を客観的に決定した本研究の手法は、相応に評価できるだろう。

本研究ではVKS1は5つの習熟度水準、VKS2は4つの習熟度水準に分割したが、上位グループの人数を絞っていることから下位グループの人数が大きくなっている。たとえば、VKS1では、上位グループから28 人、34 人、53 人、46 人、52 人に分かれているが、授業実施のためにクラスを細分する場合、人数が多めの下位3グループは正答率の変わる地点で、27／26人、22／24 人、28／23 人のように分割することができる。異なるキャンパスや学部・学科の学生で構成されるため、このような分割ができない場合や、授業運営・理念上、さらなる習熟度分割が望ましくない場合は、無作為的あるいは能力ができるだけ偏らないように分割することも考えられるだろう。

大友 (2013b) においては、項目難易度を軸に分割を行ったが、本研究では、受験者能力推定値を軸に分割を行い、正答確率が .50 以下で、受験者能力に近接する難易度の項目を最低一つ以上、各受験者能力推定値に割り当てる方法を採択した。項目難易度が普遍的な習熟度水準を決める意味を持つならば、項目難易度を軸に規準を設定することが望ましいと考えられるだろう。しかしながら、特にVKSのような文脈のない受容語彙力テストにおいては、習得する単語の順番が必ずしも各受験者に対して一定でないため、今後も受験者能力推定値に沿って規準設定をその都度行い、共通項目を通じて等化作業を行いがら、異なる受験者集団を比較していくことが現実的な方法となるだろう。

本研究においては、2つの異なる難易度層のテストが、共通 12 項目を係留する等化手順を経て、347 名の受験者を6つの能力水準に分割することが可能であることが確認できた。各テストの規準は英検の級別項目の区分ともおおむね一致していて、他の語彙水準リストとの相関も高い。今後は、より合理的な規準設定法を探るとともに、各規準や等化の精度を高めるため、問題の補充や改善、中間層のテストバージョンの開発等を検討する価値がある。

5．4．学習者への診断的フィードバックの可能性

SCELPと異なり、VKSは各受験者に対して級別の習熟度状況の情報を提供することができる。また、受験者理解度と正答率を比較することで、受験者の解答心理やテスト解答方略、難易レベルに対応しない特殊な単語習得状況の可能性についても探ることが可能であることが確認できた。各項目についても項目通過率と項目理解困難度を比較することで、受験者全体の理解困難度の感覚と実際の難易度とのずれが大きい項目について探索することが考えられる。

規準設定の研究を進める中で、教育プログラムの施策者が正確に分割点を決定できるようになるだけでなく、受験者・学習者にも規準到達の方向性を示せるような、診断的な情報提供の促進につなげるような研究を発展させていくことも、望まれるだろう。

5．5．規準設定における項目情報の意味

個別単語の習得が各受験者において必ずしも一定でない状況にあって、項目難易度や他

の項目情報から規準を定義することは容易ではない。しかしながら、各習熟度レベルに割り当てた対応する項目の特徴や分割点周辺の項目の情報を詳しく見ていくことで、大まかな傾向は把握できることが確認できた。Coh-Metrix を使って単語の特徴を解析したところ、項目の難易度と単語の使用頻度、単語のイメージのしやすさ、親密さ、具体性等の指標との相関が高いことがわかった。しかしながら、データベースにない単語が分析データに含まれている場合は正確に計算できていないこと等、解釈には注意が必要である。今後 Coh-Metrix のような単語や文法構造の自動解析ツールを効果的に活用しながら、規準設定に影響を与える項目情報の特徴について研究を進めていくことが望まれる。

6. 結論

本研究の結果から、5つの研究課題が検証された。受容語彙力テストの VKS はいくつかの解決すべき問題点は抱えながらも、受験者理解度の測定機能や等化可能な難易度の大きく異なるテストを兼ね備え、SCELP よりも広域な能力層の受験者に対して、より多角的な視点から、意図とした構成概念を測定し、ラッシュモデルと潜在ランク理論の手法を併用することで、プレイスメントの観点から合理的な規準設定を行うことが可能であることが確認できた。

今後検討すべき課題として4点、ここに記したい。①使用言語や文脈の追加等、テスト形式の変更・修正は必要か、②テスト形式を変更しない場合でも問題の補充や改正、新たなバージョンのテストの開発に価値や意義があるか、③より合理的かつ効率的な規準設定の方法はあるか、 そして、④規準設定の意味づけを明確にする項目や受験者情報を効果的に分析して、必要に応じて、テストの利害関係者にわかりやすく提供する方法はあるか、である。

これらの課題を検証することで、教育現場のニーズに応える、より実践的な規準設定のあり方が見えてくるのではないだろうか。

参考文献

Alderson, J.C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment*. London: Continuum.

Beglar, D. (2010). A Rasch-based validation of the Vocabulary Size Test. *Language Testing, 27*, 101-118.

Beglar, D., & Hunt, A. (1999). Revising and validating the 2000 word level and university word level vocabulary tests. *Language Testing, 16*, 131-162.

Crossley, S.A., Salsbury, T., & McNamara, D.S. (2012). Predicting lexical proficiency in language learner texts using computational indices. *Language Testing,* 29, 243-263.

Elgort, I. (2013). Effects of L1 definitions and cognate status of test items on the Vocabulary Size Test. *Language Testing, 30,* 253-272.

Graesser, A. C., McNamara, D. S., & Kulikowich, J. (2011). Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational Researcher, 40,* 223-234.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods*, *Instruments*, *and Computers*, *36*, 193–202.

Heatley, A., Nation, I.S.P. and Coxhead, A. (2002). RANGE and FREQUENCY programs. [Software] Available from http://www.vuw.ac.nz/lals/staff/Paul_Nation

Jiao, H., Lissitz, B., Macready, G., Wang, S., & Liang, S. (2011). Exploring using the mixture Rasch model for standard setting. *Psychological Test and Assessment Modeling, 53*, 499-522. Retrieved from http://www.psychologie-aktuell.com/fileadmin/download/ptam/4-2011_20111217/06_Jiao.pdf

Laufer, B., & Goldstein, Z. (2004). Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language Learning, 54,* 469-523.

Linacre, M. (2014). *WINSTEPS Rasch measurement computer program* (Version 3.81.0). Chicago: Winsteps.com.

Lissitz, R.W. (2013). Standard setting: past, present, and perhaps future. In M. Simon, K. Ercikan & M. Rousseau (Eds.) Improving large-scale assessment in education: Theory, issues, and practice. (pp.154-174). New York: Routledge.

Milton, J. (2009). *Measuring second language vocabulary acquisition.* Bristol, UK: Multilingual Matters.

Nation, I. S. P. (1990). *Teaching and learning vocabulary.* New York: Newbury House.

Nation, I. S. P., & Webb, S. (2011). *Researching and analyzing vocabulary.* Boston, MA: Heinle, Cengage Learning.

Nation, I.S.P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher*, *31,* 9-13

Pitoniak, M .J., & Morgan, D .L. (2012). S etting a nd validating cut s cores for t ests. In C . Secolsky & D .B. D enison (Eds.) *Handbook on measurement, as sessment, an d evaluation in higher education.* (pp. 343-366). New York: Routledge.

Read, J. (2000). *Assessing vocabulary.* Cambridge: Cambridge University Press.

Stubbe, R . (2012). D o ps eudoword false a larm r ates a nd overestimation r ates i n Yes/No vocabulary tests change with Japanese university students' English ability levels? *Language Testing, 29,* 471-488.

Wright, B., & Stone, M. (1979). *Best test design: Rasch measurement.* Chicago: Mesa Press.

Zieky, M.J. , P erie, M., & Livingston, S.A. (2008). *Cutscore: A Manual for Setting Standards of P erformance o n E ducational and O ccupational T ests*, Princeton, N J: Educational Testing Service.

小泉利恵・飯村英樹 (2010).「ニューラルテスト理論の特徴:古典的テスト理論・ラッシュモデリングとの比較から」『日本言語テスト学会紀要』、 13, 91-109.

荘島宏二郎 (2011). Exametrika (Version 5.3) [Software] Available from http://antlers.rd.dnc.ac.jp/~shojima/exmk/jindex.htm

法月 健 (2012a).「規準設定におけるラッシュモデルの有用性」『言語テストの規準設定報告書』、 財団法人英語検定協会英語教育センター委託研究. (pp.117-126).

法月 健 (2012b).「規準設定におけるニューラルテスト理論の有用性:項目応答理論と古典的テスト理論との比較」『言語テストの規準設定 報告書』、 財団法人英語検定協会英語教育センター委託研究. (pp.127-136).

法月 健 (2013).「受容語彙力を測定するプレイスメントテストにおけるラッシュモデルと潜在ランク理論に基づく規準設定の試行」『言語テストの規準設定 報告書第2号』、 公益財団法人英語検定協会英語教育センター委託研究. (pp.81-103).

大友賢二 (2013a).「予備調査:CITO v ariation on the bookmark m ethod」『言語テストの規準設定 報告書第2号』、 公益財団法人英語検定協会英語教育センター委託研究. (pp.1-38).

大友賢二 (2013b、12月).「英語教育とテスト:第二言語習得における規準設定をめぐって」、『第7回日本テスト学会賞記念講演会』、 東京:成蹊大学.

植野真臣・荘島宏二郎 (2010).『学習評価の新潮流』、東京:朝倉書店.

英検は教科の知識測定の道具として使えるか
―CLIL の評価基準設定の準備としての固有名詞使用検証
Does EIKEN help measure topical knowledge?
Setting standard for CLIL by identifying the use of proper nouns

渡部良典
Yoshinori Watanabe

Abstract

Setting standared for assessing CLIL courses invol ves an extremely com plex task, beginni ng with establishing its construct consisting of la nguage, cogni vie skills and c ontent or topical knowledge, a nd t hen operationalizing t hese elements in observable terms. W atanabe (2012) conducted an observation study in an attempt to identify the vocabulary that would constitute a unique feature to the CLIL. The result showed that the words used in the CLIL course and regular mainstream EAP course dif fered between t eachers as well as betw een course ty pes, which implied that teaching sty le would be equally important to the principles in the course in causing differences observed in the token, ty pe and type-token ration in th e use of vocabulary. Aand y et the result also indicated that dif ferences did exist even between the courses of different pur poses taught by the sam e teacher. Besides these major fidings, the use of proper nouns, inclu ding personal nam es, place nam es, th e name of the book, and so forth, made a differential characteristi c features to the cont ent-oriented CLIL course. The present paper capitalized on this finding and explored the use of proper nouns in the past examination papers of EIKEN. The result of the analy sis o f the reading co mponent of the past exam papers of EIKEN Grade 1 and Grade Pre-1 revea led they differend not only in the total num ber of words but in terms of the frequency and type of proper nouns. It was shown that Eiken Grade 1 co uld be a useful source of assessing CLIL with its focus on the content component.

1. CLIL（内容言語統合型学習）における評価と規準設定

CLIL (Content and Language Integrated Learning) とは、ある特定の教科を語学教育の方法を通して学ぶことにより、効率的にかつ深いレベルで修得し、習得対象言語を学習手段として使うことで、実践力を伸ばすことを目的とした言語指導の原理である。外国語習得のみならず学習上の技能を向上することも大きな目的の一つである。CLIL の中心をなす考え方は、言語が扱う教科 内容（content）、学習技能（study skills）、言語（language）（Coyle et al. 2010 、48-85 頁）、これら3つの要素を同時に扱うことである。この3つの要素は CLIL を構成する3つの観点（基準）ということができる。すなわち、CLIL における課題は、これら3つの互いに独立しているが、同時に関係づけられている要素それぞれについて、どのような規準を設けるのが適切なのかということである。

CLIL はあくまでも言語教育の指導原理である（e.g. Mehisto, Marsh & Frigols, 2008; Coyle, Hood & Marsh, 2010 ; Dale & Tanner, 2012; Harmer, 2012）。外国語の指導にあたって言語環境を整えることが重要であることは言うまでもないが、限られた時間の中で行われ、また教室を離れれば対象言語を使う必要がない環境にある場合、当然のことながら意識的に語彙を増やしたり、文構造を理解したり、といった作業はどうしても必要となるはずである。そして、これは言語教育である限り、 CLIL も例外ではない。その一方、CLIL では、ある特定の教科内容や研究分野、ジャンル等を限定してその中で言語習得を目指すので、そのような限定的な枠組みのない一般的な内容を扱う言語指導よりも効率よく習得できるということが期待されるのである。

そのためには、指導対象とする特定の分野においてどのような言語機能、文法構造を扱うのか、特に当該分野に特有の語彙を特定し、指導の際に教員は積極的に機能、構造、語彙を使い、そして学習者にも使いながら習得するようにする必要がある。必要な言語要素を特定するためには実際に言語が使われている状況を観察記録し、そこから特有の言語を記述するという作業が必要となる。しかも、CLIL は特定の教科を対象とするので、自然環境で行われている言語使用状況ではなく、あくまで教室で行われている言語を記述の対象とする必要がある。また、CLIL は非母語話者の教員であることがイマージョン教育などとは異なる特色の一つであるが（Llinares, et al., 2012 ）、当目的のためにはあえて母語話者の教員をモデルとして彼らがどのような言語を使うのかを記録する。しかしながら、対象となる学習者は対象言語の非母語話者である。すなわち、母語話者の教員が非母語話者の学習者を対象に教室で指導している場面を記録分析するという作業である。

上述のような作業を通してはじめて CLIL における規準の設定が可能になる。言語機能、構造、語彙のうち、本稿では前回に引き続き言語のもっとも基本を成す語彙を扱った。

2. 2012 年報告書の要約

2012 年の報告書では、CLILの授業観察に基づきどのような語彙が最も典型的にCLILの

授業を成り立たせているのかを考察することを目的とした。総計 270 分の授業を分析した結果、通常のEAP（English for Academic Purposes）においては*paragraph*、*draft*、*presentation*, *essay*, *review*などの語彙が多用されるが、同じ教員が担当した詩の鑑賞をテーマとした文学の授業ではそのような学術関係の基礎用語はほとんど使われることがなく、それに替わって *poem*, *background, poetry* などの文学用語が多様さていることがわかった。さらに、当該の授業の特徴となっていたのは作家の名前や、文化的な背景を表す固有名詞[1]の多用であった。例えば、Nationの規定する 1,000 語レベルでかつCoxheadのAcademic Word Listに含まれていない固有名詞には、American_[3] Americans_[1] Ariel_[2] Boston_[1] British _[1] England_[2] Sigmund Freud_[3] German_[7] Hughes_[5] James_[1] Jewish _[4] Nazi_[4] Nazis_[1] Oedipus_[1] Paltrow_[1] Sylvia Plath_[30] （[ ]内は頻度）があった。これら固有名詞の中には、Sylvia Plath のように当該授業のテーマとなる固有名詞もあり、この特定の人命の使用頻度が多いのは当然でもあり、また学生にとって知識がなくても教員の詳細な解説があるので問題はない。しかしながら、前提知識がなければ授業内容の理解に支障を来す語彙がほとんどだ、と言ってよいであろう。すなわちCLILにおいては、固有名詞（Proper Nouns）が知識内容（content）の重要な構成要素となるのである。

3. CLIL のテストにおける固有名詞の重要性

固有名詞に関する知識をHirsh（1987）に倣ってCultural Literacyといってもよいかもしれない。通常のテストでは、人名、地名などはあえて中立にすることが求められる。その結果、誰でも知っている名称を使う、解答するのに特別な知識は必要ない文章を使う、問題文に固有名詞が入っていてもその固有名詞に前提知識のある受験者とそのような知識のない受験者に差がないような問いを作成する、このうちいずれかを行うのである。一方、CLIL は普通のテストではバイアスとして排除されるような要素―すなわち話題に関する知識（topical knowledge）を積極的にテストに採り入れようとする。したがって、英語を使って知識を試すテストが必要となるのである。

前節で述べたように、授業観察の結果一般の EAP の授業と特定の教科内容に重点をおいた授業では教員の使う語彙に明らかな違いが認められた。そしてその違いを生み出す要素のひとつが固有名詞の使用であった。そうすると、英語を使って固有名詞を試すテストというのは CLIL における評価測定に必要となるということができる。本稿では、本来英語能力を測定するために開発されたテストーすなわち話題に関する知識はバイアスとして排除されるよう配慮されているはずのテストーである英検を使って、このテストがCLILの内容（content）を

[1] 固有名詞、すなわちProper nounsは本来Proper names と区別されるべき文法上の概念である（例えば、Jespersen, 1909 – 1949；Huddleston & Pullum, 2002 等を参照のこと）。本稿では文法上の用法を考察することが目的ではないので、単に固有名詞とProper nounsを同義で用いている。

測定するための可能性があるのかどうかを検証することを目的とした。

本調査は CLIL の評価測定システムを構築するための一部である。したがって、本論に移る前に、本研究の枠組みを概観することにしよう。

4. CLIL の評価システムとその基盤

2012 年度の報告書に続いて、本調査でも Marzano & Kendall（2007）の枠組みを用いる。図2にはオリジナル版を、図1には簡素化した図を掲載した。Marzano & Kendall のモデルは、人間の思考のモデルあるいは理論であり、単なる枠組み（framework）ではないのだということを強調している（p. 16）。このモデル（図3）もやはりプロセスと知識の2次元からなるとしている。しかし、Andesron et al（2001）とは異なり、情意領域が自己システム思考（self-system thinking）として組み込まれ、大変重要な役割を果たすとしている。また知識についても、情報（information）、心的手続き（mental procedures）、運動神経上の手続き（psychomotor procedures）から構成されるとする。それぞれの、要素の関係は単なる層（hierarchy）や分類（taxonomy）の代わりに使われているのが、それぞれの要素の支配関係（control）という概念である。

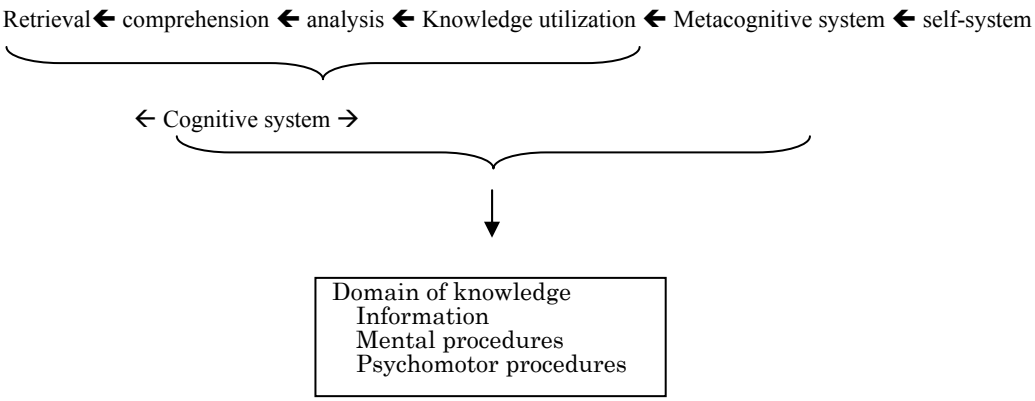


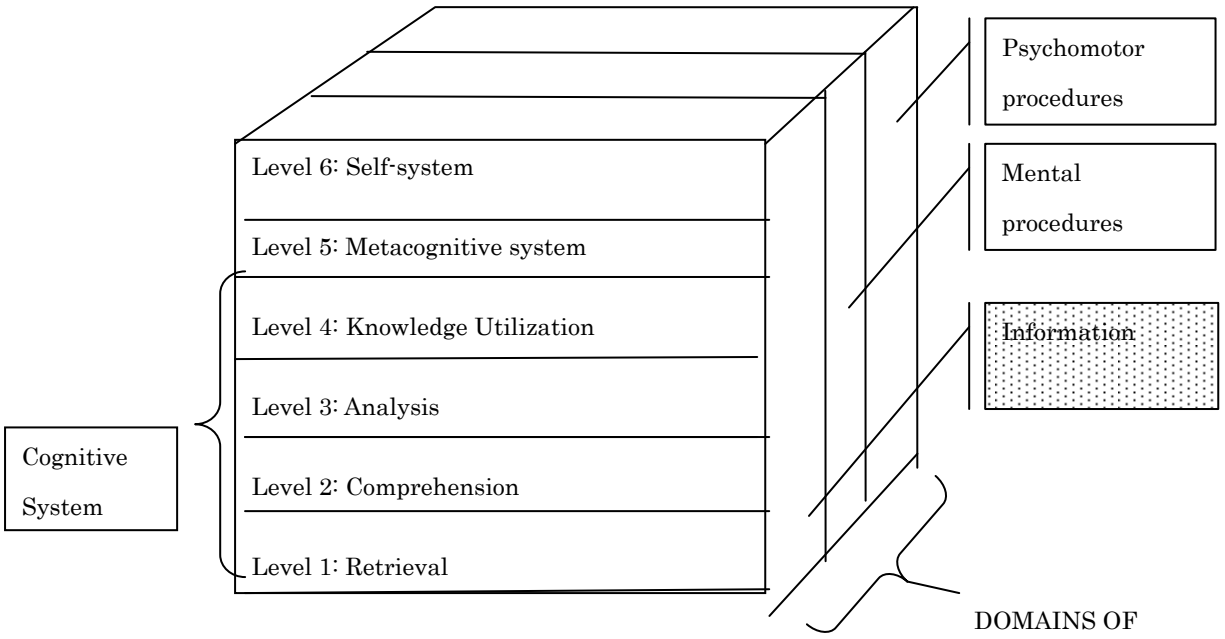
図1　Marzano & Kendall（2007）のシステム（Marzano & Kendall, 2007 を参考に現筆者が単純化したもの）

学習対象が重要であると認識し、興味関心があると、メタ認知が働き、学習や知識の運用が始まるというシステムである。

CLIL では、認知心理学の知見を援用しながら、知識の理解や暗記を中心とする、浅い、表面的な学習（shallow/surface learning）、および学んだ内容を既存の知識や経験と結びつけたり、批判的に考察を行ったりする深い学習（deep learning）の2種類の学びがあるとする。両者を学習活動にバランスよく取り込むために援用しているのが Anderson, et al（2001）である。現在のところ、CLIL の研究や指導で行われているのは、Benjamin Bloom の教育目標の分類で行われている思考の6段階モデルである。このモデルでは、Remembering（記憶する）→ Understanding（理解する）→ Applying（応用する）→ Analyzing（分析する）→

Evaluating （評価する）→ Creating（創造する）という認知技能を階層かし、下位３層を Lower-order thinki ng skills（低次思考力）とし、上位3層を Higher-order thinking skills（高次思考力）とするのである。

Marzano & Kendall（2007）は New Taxonomy of Educational Objectives として、図2に示したような3次元の枠組みを提唱している。ここでは、認知領域が retrieval、comprehension、analysis の3次元でとらえられており、さらに別のレベルに knowledge utilization を置き、さらに metacognitive system（学習方略等はこのシステムに含まれる）、さらに動機等を含む self-system（情意領域はここに含まれる）をもって構成されている。また Anderson、et al と同様に、知識を別の次元においているが、そこには外国語の学習でいえば発音などの運動神経系の知識も含まれる。Marzano & Kendall はこれは枠組（framework）ではなく理論（theory）なのだとしている。



Marozano & Kendall (2007), p. 13

図2　Marzano & Kendall (2007)の The new taxonomy of educational objectives

この枠組みを2次元化し教育目標の点検表にしたのが表1である。このモデルでは、何らかの課題を遂行する必要が出てきた際に、最初に Self-system が作動しその課題に価値や必要性を認めた場合、次の下位にある metacognitive system がさらに下位の cognitive system に作用し、その課題を行うために必要な認知活動を行わしめるのである。したがって、Bloom やその改訂版の Anderson がそれぞれのレベルの要素を想定された操作の複雑さを基盤にして

階層化しているのに対し、Marzano & Kendall はそれぞれのレベルに相互作用と有機的な関連性を想定しているのである。したがって、CLIL のような、言語に加え、集団内の相互作用、認知技能、知識を教育の重要な目標としている原理にとっては、理論化に適した理論的基盤となることが期待されるのである。それは、すなわち CLIL の評価のための理論的基盤を提供することにもなろうし、引いては規準の設定およびその妥当性の検証の枠組みとして機能することにもなりうるのである。この枠組みでは、語彙はもっとも下位次元の知識・情報（informational knowledge）に属することとなる（Marzano & Kendall, 2008, pp. 9 – 11）。

表1　Marzano & Kendall (2007)に基づいた細目・点検表

		Information	Mental procedures	Psychomotor procedures
Level 6: Self-system thinking				
Examining importance				
Examining efficacy				
Examining emotional response				
Examining motivation				
Level 5: Metacognition				
Specifying goals				
Process monitoring				
Monitoring clarity				
Monitoring accuracy				
Cognitive system	*Level 4: Knowledge utilization*			
	Decision making			
	Problem solving			
	Experimenting			
	Investigating			
	Level 3: Analysis			
	Matching			
	Classifying			
	Analyzing errors			
	Generalizing			
	Specifying			
	Level 2: Comprehension			
	Integrating			
	Symbolizing			
	Level 1: Retrieval			
	Recognizing			
	Recalling			
	Executing			

Manzano & Kendall (2007), p. 128.

今回の調査は、知識の領域（Domain of Knowledge）の次元の情報（information）を確定するための試みである。

5. 先行研究

CLIL に関する実証研究は語彙に関しては、特に効果を測定することを目的とせず CLIL の授業を特徴づけようとする記述研究も行われている（Espinosa、2007；Llinares, et al 、2012）。語彙の習得に関しては偶発的に（incidental）な習得方法は効果が低く、意識して練習をする必要があることが CLIL の授業に関しても（Admiraal, Westhoff & Bot 、2006）また、ClIL には直接関係しない分野でも同様の結果が報告されている（e.g. Horst 、2010；Tang、2011）。CLIL の授業ではむしろ学習者がそこで興味関心をもって教室外の家庭学習等で対象言語

に触れる機会を作ろうとし、そのような個人学習で語彙が習得される可能性があることも指摘されている（Ackerl、2007）。

上述の研究の示すところを特に語彙に関してまとめると、CLIL では学習対象となっている特定の分野や科目に関する語彙習得を促進する可能性が高いこと、しかしながら偶発的な習得を待つのではなく、意識的に語彙学習を促す必要があること、ということになる。ここから、特定の分野に関する基礎語彙を特定し、それを効果的に指導する必要があるという結論が導き出せる。そして、そのためには序論で述べたように、授業を観察記録し、そこに特徴的な語彙を特定することの意義が認められるのである。

6. 研究方法

6. 1. 分析対象

分析にはウェブ上に公開されている2013年度実施分英検1級[2]および準1級[3]の中からそれぞれ 10 の英文を選んだ。それぞれの表題については表1および2を参照のこと。分析したのは本編だけであり、選択肢を含む問題の部分は含めなかった。分析には、compleat lexical tutor for data-driven language learning on the Web（http://www.lextutor.ca/）を使った。

6. 2 結果と考察

延べ語数（token）、異なり語数（type）、延べ語数に占める異なり語数の割合（type-token ratio）は表2（準1級）および表3（1級）にまとめた。

表2 2013年度英検準1級読解問題の延べ語数、異なり語数、比率

Title of the passage	延べ語数	異なり語数	Type/Token
The Sa'och language	324	195	.602
The gray invasion	412	238	.571
Hospital uniforms in the United Kingdom	253	146	.577
Responding to cell phones	252	159	.631
The miracle bean	249	171	.687
Food deserts	255	158	.620
Octopus intelligence	245	158	.645
Computer junk?	407	246	.604
Harvesting silk	256	159	.621
Braille vs. speech	326	203	.623
Mean	297.9	183.3	.618

準1級と1級の比較を行うことが本調査の目的ではないが、全体の傾向を概観する。延べ語数、異なり語数双方において1級は準1級と大きく異なる。平均延べ語数は、準1級が 297.9

[2] http://www.eiken.or.jp/eiken/exam/grade_1/

[3] http://www.eiken.or.jp/eiken/exam/grade_p1/

語、1級が504.5語である。異なり語数は準1級が183.3、1級が271.2である。1級では総語数が多いことは当然としても、使われている語彙の数も多いことが見て取れる。一方、総語数に占める異なり語数の割合（type/token ratio）は準1級が.618、1級が.550であり、準1級の方が高い値を示している。

表3　2013年度英検　1級読解問題の延べ語数、異なり語数、比率

Title of the passage	延べ語数	異なり語数	Type/Token
Meigs Field	347	200	.576
The end of Maya Civilization	340	205	.603
Ethics and heart transplants	525	282	.537
Titan and life as we don't know it	509	248	.487
A safer world?	770	399	.518
The descent of man	349	210	.601
U.S. Immunization programs	334	195	.584
Golden rice	525	302	.575
Feathered dinosaurs	522	288	.552
Britain and the American civil war	824	383	.465
Mean	504.5	271.2	.550

次に、本稿の目的である固有名詞の使用頻度を見てみよう。結果は表4および表5に示した通りである。なお、固有名詞の特定にあたっては以下の方針にしたがった。

1．文法的に語形変化を得たものもひとつの語とした。すなわち word family を単位とした。（例：Cambodia、Cambodian、Cambodia's は1単位）。
2．文章中初出はフルネームの人名が2度目からは名字だけの場合も1単位とした。（例：Jean-Michel Filippi、Filippi は同一人物なので1単位）。
3．リスト中、冠詞は省略した。（例：the United Nations）

ここでもやはり準1級と1級の違いは歴然としている。平均固有名詞数が準1級 8.00、1級で22.700 なのは当然としても、全語数に対する固有名詞の使用頻度が準1級で.027 なのに対し、1級では.042 とおおよそ2倍である。さらに固有名詞の異なり語数が準1級では 4.300 であるのに対し、1級では 9.100 で、こちらもおおよそ2倍である。それぞれの級で使われている固有名詞を瞥見するだけでも1級でははるかに多くの多様な語彙がつかわれていることがわかる。

このことから、準1級、1級、どちらも直接知識を問う課題はないにしても、可能性としてはある特定の知識を持っている受験者にとって少なくとも心理的に受験しやすい英文が1級と比較すると準1級には多く使われている可能性が高いということができる。さらに CLIL のテストとしては準1級よりも1級のような英文を用いて特定の知識を検証する測定の道具とできる可能性が高いということが言える。ついでながら、後者の可能性を念頭に置いて、表4の固有名詞のリストを見直すと、欧米特に米国と英国に関する固有名詞がほとんどであることがわかる。これを偏りと見るかあるいは、外国語としての英語能力の見るために意図されているのかにつ

ては別の機会を設けて考察すべき事柄である。

表4　英検準1級（2013 年度実施）における固有名詞の頻度

文章のタイトル	固有名詞総数	全語数に対する固有名詞の割合	固有名詞異語数	全異語数に対する固有名詞異語数の割合
The Sa'och language *Cambodia_[3], Educational, Scientific and Cultural Organization, Jean-Mechel Filippi_[4], Khmer_[4], Sa'och_[7], Khmer Rouge, United Nations*	21	.065	8	.041
The gray invasion *Britain_[5], England _ [2], North America , Scotland, Forestry Commission of Great Britain, Scottish Wildlife Trust*	11	.027	6	.025
Hospital uniforms *English, National Health Service, Nottingham_[2], Scottish_[2], Welsh_[2], United Kingdom*	9	.036	6	.041
Responding to cell phones *Danish, Martin Lindstrom_[4]*	4	.016	1	.001
The miracle bean *American, Paraguay, South America, United States_[3]*	6	.024	4	.023
Food deserts *Helen Lee, Philadelphia_[4], U.S*	6	.024	3	.019
Octopus intelligence *Jennifer Mather, Roland Anderson*	2	.001	2	.013
Computer junk? *U.S. Environmental Protection Agency, United Nations Environment Programme*	2	.001	2	.001
Harvesting silk *Randy Lewis_[3], Utah State University*	4	.016	2	.013
Braille vs. speech *American, Braille_[7], Canada, Diana, Doug Brent_[2], Laura J. S loate_[2], Wall Street, United States, University of Calgary*	15	.046	9	.044
Mean	8.000	.027	4.300	.023

表5 英検 1級(2013 年度実施)における固有名詞の頻度

文章のタイトル	固有名詞総数	全語数に対する固有名詞の割合	固有名詞異なり語数	全異語数に対する固有名詞異なり語数の割合
Meigs Field (airport in Chicago) *Chicago_[4], Meigs Field_ [7], Richard M. Daley_[5], Soldier Field_[2], City of Chicago, United States*	20	.058	6	.030
The end of Maya Civilization *Belize, Benjamin Cook, Boston University, Central American, Douglas Kennett_ [3], Endfield, Georgina, Maya_ [8], Norman Hammond, Pennsylvania State University, University of Nottingham*	20	.059	12	.059
Ethics and heart transplants *Barnard_[3], Christian Barnard, Raymond Hoffenberg, South Africa_ [4], U.S. presidential council, United Kingdom_[2], United States_[8]*	20	.038	7	.025
Titan and life as we don't know it *Cassini_[3], Chris McKay_[3], Darell Strobel_[4], Earth_[6], Johns Hopkins University, Mercury, Methane, NASA Allies Research Center, NASA, Saturn, Titan_[7], Cassini project*	30	.059	12	.048
A safer world? *Abel, Aztec Empire, British, Cain, Darwinian, Europe_[2], Harvard University, John Gray, Michael Nagler, Steven Pinker_[12], Better Angels of Our Nature, Western civilization, World War II, University of California, West_[3]*	29	.038	15	.038
The descent of man *African, Gerald Crabtree_[6], Stanford University*	8	.023	3	.014
U.S. Immunization programs *Americans_[2], Dr. Robert Chen_[3], Centers for Disease Control and Prevention, United States*	7	.021	4	.021
Golden rice *British, Golden Rice, Indian, Vandana Shiva_[2], Institute of Science in Society*	6	.011	5	.017
Feathered dinosaurs *Beijing, British, China, Liaoning Province_[2], Thomas Henry Huxley_[2], Tyramunosaursrex, Xing Xu, Yixian, American Museum of Natural History, Chinese Academy of Sciences, Paleontological Museum of Liaoning*	13	.025	11	.038
Britain and the American civil war *Abraham Lincoln_ [11], American_[2], Great Britain_[24], Charles Hubbard, England, London, North_[3], North America, Secretary of State, South_[20], American Civil War, Battle of Gettysburg, Union_[9], United States_[5], William Seward, Civil War, Emancipation Proclamation*	74	.090	16	.042
Mean	22.700	.042	9.100	.033

6．3．使用語彙の特徴

最期に、これは本調査の主たる目的ではないが、準1級と1級とで使われている語彙の全般的な特徴を見るためにコーパスソフトウェアの AntConc を使い Academic Word List（AWL）(Coxhead, 2000)に掲載されている語彙がどれくらい使われていて、このリストにどのような語が使われていないかを確認した。結果の一部を表6、表7、表8、表9に示した。表5と表6は、準1級、1級それぞれで使われている語彙のうち AWL にリストされている語彙、すなわち一般的な学術英語基礎語彙である。表7と表8に示したのは AWL にリストされていない語彙である。AWL に含まれていない語彙は、レベルが低い、あるいはレベルが高いもしくは特定化された語彙のどちらかの理由が考えられる。固有名詞は太字で示した。コーパスのプログラムでは大文字と小文字を区別しなかったので、固有名詞も小文字で示してある。

表6　準1級の語彙のうち Academic Word List に含まれている語彙

1	abandon	35	demonstrate	67	insert	101	region
2	access	36	design	68	intelligence	102	regulate
3	acquire	37	despite	69	intense	103	release
4	adapt	38	device	70	invest	104	relevant
5	administrate	39	display	71	involve	105	rely
6	advocate	40	document	72	issue	106	remove
7	affect	41	drama	73	item	107	require
8	aid	42	enable	74	job	108	research
9	area	43	energy	75	labour	109	reside
10	aspect	44	enormous	76	link	110	respond
11	assist	45	ensure	77	locate	111	revolution
12	attach	46	environment	78	maintain	112	role
13	author	47	equip	79	major	113	sequence
14	available	48	establish	80	medical	114	significant
15	benefit	49	estimate	81	method	115	similar
16	bulk	50	evident	82	minor	116	site
17	capable	51	evolve	83	modify	117	source
18	challenge	52	exceed	84	negate	118	strategy
19	comment	53	expand	85	network	119	style
20	commission	54	expert	86	option	120	sufficient
21	communicate	55	export	87	output	121	survive
22	community	56	expose	88	percent	122	target
23	complex	57	extract	89	period	123	technology
24	compute	58	facilitate	90	perspective	124	text
25	concentrate	59	focus	91	predominant	125	theory
26	conclude	60	function	92	process	126	tradition
27	consequent	61	fund	93	professional	127	uniform
28	consist	62	furthermore	94	promote	128	unique
29	consume	63	generate	95	publish	129	vary
30	convert	64	guarantee	96	random	130	virtual
31	culture	65	identify	97	range	131	vision
32	debate	66	impact	98	react	132	visual
33	decade	67	income	99	recover	133	voluntary
34	decline	68	indicate	100	refine		

表7　1級の語彙のうち Academic Word List に含まれている語彙

1	abandon	53	convince	105	incline	157	predict
2	academy	54	correspond	106	indicate	158	predominant
3	accompany	55	create	107	individual	159	previous
4	accumulate	56	criteria	108	inevitable	160	principal
5	accurate	57	crucial	109	inherent	161	process
6	achieve	58	culture	110	initial	162	project
7	adequate	59	cycle	111	innovate	163	proportion
8	administrate	60	data	112	institute	164	psychology
9	adult	61	debate	113	intelligence	165	publish
10	affect	62	decline	114	interpret	166	pursue
11	alter	63	define	115	intervene	167	react
12	alternative	64	demonstrate	116	investigate	168	region
13	analyse	65	despite	117	involve	169	reinforce
14	annual	66	detect	118	isolate	170	release
15	apparent	67	dimension	119	issue	171	remove
16	approach	68	displace	120	journal	172	require
17	aspect	69	distinct	121	justify	173	research
18	assist	70	drama	122	layer	174	reside
19	assume	71	economy	123	legal	175	respond
20	assure	72	emphasis	124	licence	176	retain
21	attribute	73	enable	125	link	177	reverse
22	author	74	energy	126	maintain	178	role
23	authority	75	enormous	127	major	179	select
24	available	76	ensure	128	manipulate	180	sequence
25	benefit	77	environment	129	mature	181	series
26	capacity	78	establish	130	media	182	shift
27	challenge	79	ethic	131	mediate	183	significant
28	chapter	80	eventual	132	medical	184	similar
29	chemical	81	evident	133	method	185	source
30	civil	82	evolve	134	modify	186	specific
31	coincide	83	expand	135	monitor	187	statistic
32	collapse	84	explicit	136	negate	188	strategy
33	commit	85	export	137	neutral	189	structure
34	complex	86	factor	138	nonetheless	190	subsequent
35	compound	87	feature	139	obtain	191	sufficient
36	compute	88	federal	140	occur	192	supplement
37	conclude	89	fee	141	odd	193	survey
38	conduct	90	final	142	ongoing	194	survive
39	confirm	91	fluctuate	143	option	195	sustain
40	conflict	92	focus	144	outcome	196	team
41	consent	93	function	145	overlap	197	technology
42	consequent	94	furthermore	146	panel	198	tense
43	consist	95	generation	147	parameter	199	theory
44	constant	96	goal	148	perceive	200	trace
45	constitute	97	grant	149	percent	201	tradition
46	consume	98	guideline	150	period	202	trend
47	contact	99	hypothesis	151	phenomenon	203	trigger
48	contrary	100	identify	152	philosophy	204	undergo
49	contrast	101	impact	153	potential	205	underlie
50	contribute	102	implicate	154	practitioner	206	unique
51	controversy	103	imply	155	precede	207	utilise
52	convert	104	impose	156	precise	208	violate
						209	widespread

当然のことながらAWLには固有名詞は含まれていない。しかし、AWLに含まれない語彙をリストした表7と表8を瞥見しただけでも、準1級と1級の間で固有名詞の種類に違いがあるか

がわかるであろう。

表8　準1級の語彙のうち Academic Word List に含められていない語彙

1	**america**	17	**england**	33	**nottingham**	49	soybean
2	**american**	18	extinction	34	nutrition	50	soybeans
3	**braille**	19	**filippi**	35	nutritional	51	spider
4	**brent**	20	genes	36	octopus	52	spiders
5	**britain**	21	gray	37	octopuses	53	squirrel
6	**british**	22	grays	38	patients	54	squirrels
7	**cambodia**	23	impaired	39	**philadelphia**	55	stimuli
8	**cambodian**	24	insular	40	professor	56	territory
9	cell	25	invasion	41	protein	57	tiny
10	cephalopods	26	junk	42	recycling	58	uniforms
11	cortex	27	khmer	43	scanned	59	vanish
12	dings	28	kluner	44	scottish	60	vertebrates
13	dna	29	**lewis**	45	shopkeepers	61	vocabulary
14	electronic	30	lifespan	46	shortages	62	**welsh**
15	electronics	31	**lindstrom**	47	sloate	63	wholesalers
16	endangered	32	**nhs**	48	soy	64	worldwide

表9　1級の語彙のうち Academic Word List に含められていない語彙

1	**abel**	123	cynical	243	immunization	364	province
2	abhorrence	124	**daley**	244	immunized	365	provoked
3	abolished	125	**darrell**	245	imperialism	366	proxy
4	abolition	126	**darwinian**	246	impoverished	367	pumpkin
5	**abraham**	127	dawn	247	impression	368	qualms
6	abrupt	128	decipher	248	impressions	369	rainfall
7	abundant	129	decisively	249	incited	370	rarity
8	abuse	130	deficiency	250	inconsequential	371	ravaged
9	accomplishes	131	deforestation	251	**india**	372	raymond
10	acetylene	132	demise	252	**indian**	373	reassert
11	activist	133	demolished	253	industrialized	374	recedes
12	ad	134	demolition	254	infection	375	recipient
13	adverse	135	denser	255	infectious	376	recurring
14	afflicted	136	dependence	256	insisted	377	rediscovering
15	**africa**	137	deposit	257	insisting	378	remote
16	**african**	138	descent	258	insists	379	reopen
17	aggressively	139	devastating	259	insulation	380	replenishing
18	airport	140	dietary	260	intact	381	reptilian
19	albersdoerferi	141	diets	261	intellect	382	residual
20	alignment	142	diminutive	262	intellectual	383	respiration
21	allergic	143	**dinosaur**	263	intellectually	384	respirators
22	allies	144	**dinosaurs**	264	intercepted	385	resume
23	altruism	145	diplomats	265	intriguing	386	resurgence
24	**amaranth**	146	disasters	266	ironically	387	**rex**
25	amassed	147	disavowed	267	irrefutable	388	rhetoric
26	amazing	148	discern	268	**john**	389	**richard**
27	**america**	149	discount	269	**johns**	390	**robert**
28	**american**	150	dispatching	270	juvenile	391	runway
29	**americans**	151	dispute	271	**kennett**	392	satellite
30	ample	152	dna	272	kinship	393	satellites
31	ancestor	153	donor	273	landmark	394	**saturn**
32	ancestors	154	**douglas**	274	legged	395	sciurumimus
33	anew	155	downfall	275	legitimacy	396	secede
34	angels	156	downtown	276	lethal	397	sediment
35	announced	157	drought	277	**liaoning**	398	seizure
36	anthropological	158	droughts	278	**lincoln**	399	seward

37 anthropologist
38 anthropology
39 anti
40 appeals
41 archaeologists
42 archaeology
43 archaeopteryx
44 aristocratic
45 astrobiologists
46 atmosphere
47 audacious
48 avian
49 aviation
50 **aztec**
51 bacteria
52 baffled
53 banned
54 barnard
55 **beijing**
56 belize
57 **benjamin**
58 beta
59 bets
60 biblical
61 biodiverse
62 biodiversity
63 biological
64 biologist
65 biotechnology
66 **boston**
67 brainstem
68 breastbone
69 **britain**
70 **british**
71 brutality
72 bulldozers
73 bulldozing
74 bustling
75 **cain**
76 calendar
77 **california**
78 campaign
79 campaigned
80 campaigns
81 cardiac
82 cardiopulmonary
83 cardiovascular
84 career
85 carnivores
86 carotene
87 carrier
88 **cassini**
89 cataclysmic
90 cessation
91 **charles**
92 **chen**
93 **chicago**
94 **china**
95 **chinese**
96 **chris**
97 **christian**
98 civilization
99 civilizations
100 climate
101 clinical
159 dysfunction
160 echoed
161 emancipation
162 embrace
163 emotive
164 endfield
165 engineered
166 **england**
167 entitled
168 era
169 eradicated
170 eruptions
171 escalated
172 ethane
173 **europe**
174 **european**
175 euthanasia
176 evaporated
177 evaporates
178 execution
179 executive
180 exemplifies
181 expansionist
182 fatalities
183 favored
184 feat
185 filament
186 flightless
187 flourished
188 foraged
189 forebears
190 forecast
191 fossil
192 fossils
193 furcula
194 gamble
195 gatherer
196 gatherers
197 gene
198 genes
199 genetic
200 genetically
201 geneticist
202 genetics
203 genome
204 geological
205 **georgina**
206 **gerald**
207 **gettysburg**
208 giant
209 gravity
210 gray
211 hailed
212 **hammond**
213 harbored
214 harmonious
215 harmony
216 harness
217 harsh
218 **harvard**
219 hazards
220 headlines
221 heartbeat
222 hedged
223 heels
279 lineage
280 lingering
281 linguistics
282 litany
283 lizards
284 **london**
285 longstanding
286 **mae**
287 magazine
288 makeup
289 massive
290 **maya**
291 mayor
292 **mckay**
293 measles
294 **meigs**
295 **mercury**
296 methane
297 **michael**
298 microbes
299 midsection
300 miracle
301 mistrust
302 mitigate
303 modeler
304 modem
305 monoculture
306 mumps
307 museum
308 mutated
309 mutation
310 mutations
311 **nagler**
312 **nasa**
313 navy
314 negotiated
315 neural
316 nil
317 nonprofit
318 **norell**
319 **norman**
320 nostalgic
321 **nottingham**
322 obstacle
323 offspring
324 opponents
325 opposition
326 organisms
327 outweigh
328 overreliance
329 paleontological
330 paleontologist
331 paleontologists
332 patents
333 patients
334 peak
335 pending
336 **pennsylvania**
337 perpetuated
338 pervasive
339 physician
340 physicians
341 plagued
342 planet
343 planets
400 shed
401 **shiva**
402 siding
403 skepticism
404 skip
405 skyscrapers
406 slaveholding
407 societal
408 socioeconomic
409 solar
410 spacecraft
411 sparked
412 species
413 speculated
414 speculating
415 spinal
416 sported
417 stadium
418 stalemate
419 stales
420 stance
421 **stanford**
422 startling
423 stature
424 **steven**
425 **stobel**
426 stoked
427 strains
428 stranded
429 strobel
430 stunted
431 **sullivan**
432 surgery
433 surroundings
434 swayed
435 tack
436 tale
437 tantalizing
438 territories
439 terrorism
440 textile
441 theorized
442 **thomas**
443 tilted
444 tissues
445 **titan**
446 toed
447 torture
448 traffic
449 trait
450 traits
451 transplant
452 transplanted
453 transplants
454 trent
455 troops
456 **tyramunosaurs**
457 unearthed
458 unending
459 unexpectedly
460 unquestionably
461 unspoiled
462 unthinkable
463 urgent
464 vacant

102	coelurosaurs	224	heightened	344	plausible	465	vaccinated
103	combat	225	**henry**	345	plight	466	vaccine
104	combating	226	hindrance	346	ploy	467	vandana
105	commonplace	227	**ho**	347	plumage	468	vast
106	comply	228	**hoffenberg**	348	pollination	469	veteran
107	concedes	229	honing	349	polls	470	viability
108	condemned	230	**hopkins**	350	**porous**	471	vicinity
109	conserving	231	hostilities	351	precipitation	472	vindicated
110	conspicuous	232	**huali**	352	predator	473	viruses
111	contemplating	233	**hubbard**	353	predatory	474	vital
112	contend	234	humanitarian	354	prehistoric	475	vitamin
113	cord	235	humanity	355	premodern	476	volcanic
114	correlated	236	humdrum	356	prevalent	477	waived
115	**corwin**	237	**huxley**	357	primitive	478	**wan**
116	counterintuitive	238	hydrogen	358	privileged	479	warlike
117	crabtree	239	hype	359	proclamation	480	**westerners**
118	criminals	240	iconic	360	professor	481	whooping
119	crippled	241	ill	361	proponents	482	**william**
120	crisis	242	immunity	362	prosperity	483	worldwide
121	criticisms			363	prototypical		
122	criticizes						

6. 結論

本報告書では、2013 年度に実施された英検準1級と1級の読解問題をランダムに選び、語彙の特徴を分析した。目的は CLIL のテストとして英検を使うことができるかどうかその可能性を確認することであった。その結果、準1級と比較して1級には多様な固有名詞が高い頻度で使用されており、1級の英検読解問題が CLIL のテストとして使用できる可能性を秘めていることが明らかとなった。

参考文献[4]

Ackerl, C. (2007). Lexico-grammar in the essays of CLIL and non-CLIL students: Error analysis of written production. Vienna English Working Papers 16, 3, pp. 6 – 11.

Alba, J. O. (2009). Themes and vocabulary in CLIL and non-CLIL instruction. In de Zorobe, Y. R. & Catalan, R. M. J. (Eds.) *Content and language integrated learning: Evidence from research in Europe*. (pp. 130 – 156). Bristol: Multilingual Matters.

Anderson, L. W., Krathwohl, D. R., Airasian, P. W., Cruikshank, K. A., Mayer, R. E., Pintrich, P. R., Raths, J. & Wittrock, M. C. (Eds.) (2001). *A taxonomy for learning, teaching and assessing: A revision of Bloom's taxonomy of educational objectives, compete edition*. New York: Addison Wesley Longman, Inc.

Anderson, L. W., Krathwohl, D. R., Airasian, P. W., Cruikshank, K. A., Mayer, R. E., Pintrich, P. R., Raths, J. & Wittrock, M. C. (Eds.) (2001). *A taxonomy for learning, teaching and assessing: A revision of Bloom's taxonomy of educational objectives, compete edition*. New York: Addison Wesley Longman, Inc.

[4] 本稿は報告書という性格上、本編で直接引用した文献だけではなく、報告書をまとめるにあたって参考にした文献はできる限り広く掲載することとした。

Anderson, L. W., Krathwohl, D. R., Airasian, P. W., Cruikshank, K. A., Mayer, R. E., Pintrich, P. R., Raths, J. & Wittrock, M. C. (Eds.) (2001). *A taxonomy for learning, teaching and assessing: A revision of Bloom's taxonomy of educational objectives, compete edition*. New York: Addison Wesley Longman, Inc.

Anderson, L. W., Krathwohl, D. R., Airasian, P. W., Cruikshank, K. A., Mayer, R. E., Pintrich, P. R., Raths, J. & Wittrock, M. C. (Eds.) (2001). *A taxonomy for learning, teaching and assessing: A revision of Bloom's taxonomy of educational objectives, compete edition*. New York: Addison Wesley Longman, Inc.

Anthony, R. AntConc Homepage. Retreived on March 15, 2014 from http://www.antlab.sci.waseda.ac.jp/antconc_index.html

Bloom, B. S. (194 9). *A taxonomy of educational objectives*. Opening remarks of B. S. Bloom for the meeting of examiners at Monticello, Illinois, November 27, 1949. Unpublished Manuscript.

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives: Handbook I: Cognitive domain.* New York: David McKay.

Bloom, B. S., Hastings, J. T., & Madaus, G. F. (1971). *Handbook on formative and summative evaluation of student learning*. New York: McGraw-Hill.

Bloom, B. S., Krathwohl, D. R., & Masia, B. B. (1964). *Taxonomy of educational objectives: Book 2 Affective domain*. London: Longman.

Brinton, D. M., Snow, M. An., & Wesche, M. B. (1989). *Content-based second language instruction*. New York: Newbury House.

Burton, W. H. (1944). *Guidance of learning activities*. New York: Appleton-Century Company.

Catalán, R. M. J. & de Z arobe, Y. R. (2009). The receptive vocabulary of EFL learners in two instructional contextSCLIL versus non-CLIL instruction. In de Zorobe, Y. R. & Catalán, R. M. J. (Eds.) *Content and lan guage integrated learning: Evid ence fr om research in Europe*. (pp. 81 – 92). Bristol: Multilingual Matters.

Coxhead, A. (2000). A New Academ ic Word List Author(s): Averil Coxhead Source: TESOL Quarterly, Vol. 34 , No. 2, (Summer , 2000), p p. 213-2 38, Downloaded March 31 from http://edc448uri.wikispaces.com/file/view/Coxhead+2000+Acad +Word+List.pdf

Coyle, D., Hood, P. & Marsh, D. (2010). *CLIL: Content and language integrated learning*. Cambridge: Cambridge University Press.

Dale, L. & T anner, R. (2012). *CLIL activities: A resource for subject and language teachers*. Cambridge: Cambridge University Press.

Dalton-Puffer, C. (20 07). *Discourse in content and language integrated learning (CLIL) classrooms*. Amsterdam: John Benjamins.

Dalton-Puffer, C. and Smit, U. (Eds.). (2007). *Empirical perspectives on CLIL classroom discourse.* Frankfurt am Main: Peter Lang.

Davidson, F., & Lynch, B. K. (2002). *Tesetcraft: A Teacher's Guide to Writing and Using Language Test Specifications*. New Haven: Yale University Press.

de Zarobe, Y. R. and Catalän, R. M. J. (Eds.) (2009). *Content and language integrated learning: Evidence from research in Europe.* Bristol: Multilingual Matters.

Ekstrand, (1982). *Methods of validating learning hierarchies with applications to mathematics learing.* Paper presented at the annual meeting of the American Educational Research Association, New York City. (ERIC Document Reproduction Service No. ED 216 896).

Ellis, R. (2005). Principles of instructed language learning. *System*, 33, 2, pp. 209 – 224.

Espinosa, S. M. (2009). Young learners' L2 word association responses in two different learning contexts. In de Zorobe, Y. R. & Catalan, R. M. J. (Eds.) *Content and language integrated learning: Evidence from research in Europe*. (pp. 93 – 111). Bristol: Multilingual Matters.

Field, A. (2009). *Discovering statistics with SPSS, 3rd ed.* London: SAGE.

Gagnë, R. M. (1977). *Conditions of learning, third edition*. New York: Holt, Rinehart and Winston.

Gill, B. P., & Schlossman, S. L. (2003). A Nation at Rest: The American Way of Homework. *Educational Evaluation and Policy Analysis, Fall, Vol. 25, 3,* 319–337.

Gottlieb, M. (2006). *Assessing English language learners: Bridges from language proficiency to academic achievement.* Thousand Oaks, Cal.: Corwin Press.

Greene, H. A., Jorgensen, A. N., & Gerberich, J. R. (1916). *Measurement and evaluation in the secondary school*. New York: Longmans, Green and Co.

Harmer, J. (2012). *Essential teacher knowledge: Core concepts in English language teaching*. Essex, UK: Pearson.

Hellekjaer, G. O. (2010). Language matters: Assessing lecture comprehension in Norwegian English-medium higher education. In Dalton-Puffer, C., Nikula, T., & Smit, U. (Eds). *Language use and language learning in CLIL classrooms* (pp. 233 – 258). Amsterdam: John Benjamins.

Hill, (1984). Testign hierarchy in educational taxonomies: A theoretical and empricial investigation. *Education in Education*, *8*, 93 – 101.

Hirsch, E. D. (1987). *Gultural literacy: What every American needs to know*. Boston: Houghton Mifflin.

Horst, M. (2010). How well does teacher talk support incidental vocabulary acquisition? *Reading in a Foreign Language*, Vol. 22, No. 1, pp. 161 – 180.

Huddleston, R. & Pullum, G. K. (200 2). *The Cambridge grammar of the Enligsh la nguage*. Cambridge: Cambridge University Press.

Jespersen, O. (1909 – 1949). *A modern English grammar – Part VII Syntax* . London: George Allen & Unwin, Ltd.

Jexenflicker, S., and Dalton-Puffer, C. (2010). The CLIL differential: Comparing the writing of CLIL and no-CLIL students in higher colleges of technology. In Dalton-Puffer, C., Nikula, T., & S mit, U. (Eds). *Language use and language learning in CLIL classrooms* (pp. 169 – 189). Amsterdam: John Benjamins.

Katja, L. (2007). Die mündliche Fehlerkorrektur in CLIL und im traditionellen Fremdsprachenunterricht: en Vergleich. In Dalton -Puffer, C. and Sm it, U. (Eds.). *Empirical perspectiv es on CLIL classr oom discours e.* (pp. 1 19 – 13 8). Frankfurt am Main: Peter Lang.

Laufer, B., and Hulstijn, J. (2001). Incidental vocabulary acquisition in a second language: The construct of task-induced involvement. *Applied Linguistics, 22, 1*, 1- 26.

Laufer, B., and Nation, P . (1995). V ocabulary si ze and use: Lex ical richness in L2 written production. *Applied Linguistics, 16*, 307 – 322.

Lee, Y. W., & Sawaki, Y. (2009). Cognitive diagnosis and Q-Matrices in language assessment. *Language Assessment Quarterly, 6*, 169 – 171.

Linares, A., Morton, T ., & Whittaker, R. (2012). *The roles of la nguage in C LIL*. Cambridge: Cambridge University Press.

Llach, M. d. P. A. (2009). The role of Spanish L1 in the vocabulary use of CLIL and non-CLIL EFL learners. In de Zorobe, Y. R. & Catalan, R. M. J. (Eds.) *Content and la nguage integrated learning: Evidence fr om r esearch in Eur ope*. (pp. 1 12 – 12 9). Bristol: Multilingual Matters.

Lyster, R. (2007). *Learning and teac hing lan guages thr ough content: A co unterbalanced approach*. Amsterdam: John Benjamins Publishing Company.

Maera, P., Lightbown, P., and Halter, R. (1997). Classrooms as lexical environments. *Language Teaching Research, 1, 1,* pp. 28 – 47.

Mager, R. F. (1962). *Preparing instructional objectives*. Palo Alto, CA: Fearon Press.

Marsh, D. and W olff, D. (eds.) (2007) . *Diverse co ntexts – conver ging goals* . Frankfurt am Main: Peter Lang.

Marzano, R. J. & Kendall, J. S. (2 007). *The new tax onomy of ed ucational o bjectives*. Oaks, Cal.: Corwin Press.

Marzano, R. J. & Kendall, J. S. (2 008). *Designing and Assessing Educational Objectives: Applying the New Taxonomy.* Thousand Oaks, Cal.: Corwin Press.

Mehisto, P., Marsh, D., and Frigols, M. J. (2008). *Uncovering CLIL: Content and language integrated learning in bilingual and multilingual education*. Oxford: Oxford University Press.

Perez-Canado, M, L. (2012). CLIL research in Europe: past, present, and future. *International Journal of Bilingual Education and Bilingualism*, 15, 3, May, pp. 315 – 341.

Puerto, F. G. del, Lacabex, E. G. and Lecumberri, M. L. G. (2009). Testing the effectiveness of content and language integrated learning in foreign language contexts: The assessment of English pronunciation. In de Zarobe, Y. R. and Catalän, R. M. J. (Eds.) *Content and language integrated learning: Evidence from research in Europe* (pp. 63 – 80) Bristol: Multilingual Matters.

Remmers, H. H., & Gage, N. L. (1943). *Educational measurement and evaluation.* New York: Harper & Brothers.

Schmidt, R. (2001). Attention. In Robinson, P. (Ed.) *Cognition and second language acquisition*. (pp. 3 – 32), Cambridge: Cambridge University Press.

Seregély, E. M. (2008). *A comparison of lexical learning in CLIL and traditional EFL classrooms*. Vienna: Universität Wien.

Simpson, E. J. (1965). The classification of educational objectives, psychomotor domain. Vocational and Technical Education Grant, Contract No. OE 5-85-104. http://www.eric.ed.gov/PDFS/ED010368.pdf

Tang, E. (2011). Non-native teacher talk as lexical input in the foreign language classroom. *Journal of language teaching and research, 2, 1,* pp. 45 – 54.

Tarja, N. (2007). The IRF pattern and space for interaction: Comparing CLIL and EFL classrooms. In Dalton-Puffer, C., & Smit, U. (Eds.) *Empirical perspectives on CLIL classroom discourse.* (pp. 170 – 204). Frankfurt am Main: Peter Lang GmbH.

Tyler, R. W. (1949). *Basic principles of curriculum and instruction.* Chicago: the University of Chicago Press.

vanPatten, B. (2003). *From input to output: A teacher's guide to second language acquisition*. New York: McGraw-Hill.

Vázquez, G. (2007). Models of CLIL: An evaluation of its status drawing on the German experience. A critical report on the limits of reality and perspectives. RESLA 1, 95 – 111.

Wode, H. (1999). Language learning in European immersion classes. In Learning through a foreign language. Models, methods and outcomes, ed. J. Masih, 16 _25. London: Centre for Information on Language Teaching and Research.

Zydatiß, W. (2007). Deutsch-Englische Züge in Berlin: Eine evaluation des bilingual en

sachfachunterrichts an gy mnasien. Kontext, kom petenzen, konsequezen. Frankfurt-am-Main: Peter Lang.
渡部良典・和泉伸一・池田真（2011）『CLIL（内容言語統合型学習）第1巻』、上智大学出版.
和泉伸一・池田真・渡部良典（2011）『CLIL（内容言語統合型学習）第2巻』、上智大学出版.

研究構成員

伊東祐郎（東京外国語大学留学生日本語教育センター教授）
大友賢二（筑波大学名誉教授）：研究代表
法月　健（静岡産業大学情報学部教授）
藤田智子（東海大学外国語教育センター教授）
渡部良典（上智大学大学院言語学専攻教授）：研究副代表

（五十音順）

言語テストの規準設定　報告書（３）
２０１４年３月３１日

公益財団法人　日本英語検定協会
英語教育研究センター　委託研究
