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

第 2 号

2013 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

第 2 号

2013 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 (あいうえお順)

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
研究代表(筑波大学名誉教授)

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
研究副代表(上智大学外国語学部教授)

目次

はじめに

本報告書は、規準の設定(Standard Setting)をテーマとした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の2年目の研究成果をまとめたものである。共同研究はさまざまな形態をとりうる。研究発表や論文を目指して一つのトピックについて文献を研究しデータを収集分析し結果を解釈した後何らかの結論を出すというのはそのうちの一つである。しかしながら、このようなアプローチをとるためにはかなり特定化された共通の研究課題について全員が持っていなければならない。一方、ある程度幅の広い融通のきくテーマを選び、それについて個々の研究者の立場からある程度独立して調査を行うという進め方もある。今回私たちがとったのは後者の方である。とはいえ、規準の設定をテーマとし、数回の会合を開催し、発表をし、質疑応答を行いながら進めてきた。互いの発表から刺激を受け学びあったその成果の一端がこの報告書である。前年度は、研究史、文献、残された課題をまとめたが、その中から最も重要だと思われるテーマを各自が選び、データを分析しながら考察を深めた。

前年度から引き続き本年度も、英語教育研究センター長小笠原剛士氏には、詳細な点にいたるまでご教示を頂いた、改めて感謝申し上げる次第である。

2013年3月31日

研究副代表　渡部　良典

予備調査：*CITO* Variation on the Bookmark Method
A Pilot Survey of the *CITO* Variation on the Bookmark Method

大友賢二
Kenji OHTOMO

ABSTRACT

Standard setting in educational measurement can be defined as a process by which a standard or cut score is established. Unless cut scores are set appropriately, the results of any given assessment could be questioned. The bookmark method, one of the important standard-setting methods, has been developed to be used with tests that are scored using Item Response Theory. CITO in the Netherlands is one of the most prestigious institutes for educational measurement in the world. We have recently found some interesting points for further investigation concerning the CITO Variation on the Bookmark Method in their manual for relating language examinations to the CEFR. We are therefore planning to do further research on the revised Bookmark Method developed by CITO so that we may implement it in our country.

In order to gain a deeper understanding of the CITO Bookmark Method, extensive experimentation and research on the present method is needed. The present method often follows the procedure as found in (Zieky, Perie & Livingstone. 2008, p. 113) "・・・ to place a bookmark at the point between the last question that borderline test takers would probably answer correctly and the first question that borderline test takers would probably not be able to answer correctly." We suspect, however, that the placing of the bookmark may often be influenced by subjective judgment as an artifact of the procedure. This is a report of the results of our study on how to place the bookmark systematically and effectively without subjective judgment by the participants. We are sure that, based on our data, this new study will help refine the bookmark method to open a new road for the further development of language testing practice.

1. 規準設定の意味と必要性

1. 1. 教育における 3 つの目標

われわれが行っている教育の中の「教育目標」は、教授や学習の結果として期待される学習者の状態を表現したものであり、概して言えば、内容的要素や能力的要素などから構成されている。このことに関連する議論は、たとえば、文部科学省や日本教育心理学会などで、これまで数多く行われてきている。さらに、これと関連する「学力」とはどんなものを指しているのであろうか？それは、ごく簡単に言えば、教育目標の達成度や達成状況を指していると考えることができる。内外における多くの教育目標の分類を検討してみると、認知的領域、情意的領域，運動的領域の 3 つに分けていることが一般的である。教育目標を体系的に分類したものには、梶田・渋谷・藤田訳(1973) による教育目標のタキソノミー(分類学)が有名であるが、ここでは、主に、わが国における教育目標の分類に関連してその実態を探してみることとする。

梶田(1983, pp. 80-83)は、到達目標が行動目標として表現されなければならないという考えに対して、到達目標を、達成目標、向上目標、体験目標に分けることを提唱している。達成目標とは、「特定の具体的な知識や能力を完全に身につけることが要求されるといった目標」を指している。また、向上目標を「ある方向へ向かっての向上や深まりが要求されるといった目標」としている。基本的には、個人内での比較や他の人との比較という形でしか進歩あるいは向上・深化などが把握しがたいという性格のものであり、論理的思考力とか鑑賞力、指導性とか社会性といったような包括的で、総合的な高次の目標が、これに属すると考えられている。第 3 番目の体験目標に関しては、「学習者側における何らかの変容を直接的な狙いとするものではなく、特定の体験の生起自体をねらいとするような目標」としている。

最近では、新しい高等学校学習指導要領(平成 21 年告示)のねらいを実現するために文部科学省・国立教育研究所(2012)を発行しているが、目標と教育は常に大きく重要な課題となっている。

1. 2. 規準設定の意味

教育評価における「規準設定」に関連して、英語の「standard」という語の持つ意味は、さまざまなひとによって、さまざまに用いられている。そこで、まず、ここで用いる「standard」の意味を明確にしておくことが必要である。Standard という語の持つ意味は、外国の評価関係の文献において、たとえば、Fulcher, G. (2010, p. 323) では、その意味を 6 つにわけて示している。

(1) A code of practice, or guidelines, designed to guide test development and use.
(2) A set of hierarchical descriptors of levels of achievement.

(3) A level of performance required to pass a test, be classed as a ‘master’, or receive certification.
(4) A comprehensive list of content standards for what it is expected learners will master at specific educational levels.
(5) ‘standard-setting’ or ‘aligning tests to standards’----establishing cut scores against performance standards, or aligning test content to content standards.
(6) A non-technical expression indicating the role of tests in improving educational progress, as in the phrase ‘raise standards.’

ここでは、「規準設定」ということに論を進める前に、まず、この「規準」という言葉の意味を、明確にしておかなければならない。これは、英語での standard という語に当てることとする。教育の中の standard は、上に述べたように、これまた多くの意味を持っている。これをさらにしぼると、たとえば、standard-setting とか aligning tests to standards での意味を思い浮かべていただくこととする。つまり、具体的には、“ Standard setting can be defined as the process by which a standard or cut score is established ”(Cizek G.J., 2006, p. 226)というコンテクストでの standard を指すこととする。すなわち、「規準設定」というのは、規準または分割点を設定する過程をさしているものである。教育においてはある目標を設定して、学習した学習者がその目標に到達したかどうかを考えなければならない。ごく、簡単に言えば、その目標があるテストで 75 点以上であるとした場合、その 75 点は、本当に目標到達と判断するのに適切かどうかを検討することと考えればよい。つまり、到達と、未到達を決定する「分割点」(cut-score)を、どのようにしたら、最も適切に設定できるかを考えることである。

1．3．必要性：観点別評価、CEFR、CAN-DO statements

「規準設定」ということの意味と関連する作業は、教育の場においては、きわめて重要な役割を果たしている。わが国における観点別評価の現実は、まさにこの規準設定の分野に関連している。外国語における「コミュニケーションへの関心・意欲・態度」、「外国語表現の能力」、「外国語理解の能力」、「言語や文化についての知識・理解」などの観点別評価においては、その A,B,C 評価、さらに、その総合評価としての「評定」における 1,2,3,4,5 評価は、まさに、この規準設定の分野に関連している重要な要素である。さらに、わが国にも影響を与えている CEFR(Common European Framework of Reference for Languages)の動向、ごく最近では、文部科学省の「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」に関連する動向は、まさに、この「規準設定」に関連する課題である。

観点別評価は先に取り上げた「目標」との関係で、多くの課題を投げかけている。

どのような状態になったら、目標が達成されたと判断するのか、という大きな課題がある。それを明確にしないまま、目標が設定されていたということはなかっただろうか？つまり、目標として設定されている方向への向上が見られたかどうかという「向上目標」が、評価規準を設定する大きな視点ではなかったかという反省である。たとえば、「十分満足できる」、「おおむね満足できる」「努力を要す」という決定にも係わらず、結局は、従来の相対評価にとどまったという現実も、この向上目標に関連するものと思われる。どのような状態になった場合に、そのように判断するのか、それが十分検討されずに行われていたのではなかろうか？つまり、向上目標と達成目標との混同があったのではないだろうか？観点別評価の中の総合的な評価「評定」は、絶対評価へ移行するといわれていながら進展できなかった原因は、この向上目標と達成目標との混同にあったのではなかろうか？

さらに、「評価規準」と「判定基準」という日本語から生まれる混乱も見逃すことはできない。これは、ごく最近、眼にとまった用語である。北尾倫彦監修(2012、p. 15)には、つぎの一節が見られる。ここでの「規準」と「基準」の意味の混同は、ないのだろうか？

> これらの点を考慮すると、まず第 1 に、基本となる観点別評価をたしかなものとするために､1 つには教科の観点を単元ごとに具体化した評価規準を正しく設定し、2 つめには、「おおむね達成された」か「十分達成された」かの違いを的確に示す判定基準を設ける必要がある。それらの具体的手順については次節で詳しく説明しているが、それを参考にしてよい評価規準と判定基準を作成することが極めて大切である。

「基準」と「規準」の違いは、ここで、明確にしておくことが必要である。この議論は、1983 年という 30 年前の話題に戻るが、橋本(1983: 28)では、criterion には「規準」を、standard には「基準」をと述べている。その後、皆見(2008)など、様々な議論があったが、これに関する筆者の立場をここで明確にしておかなければならない。筆者は、池田(監訳)(2008, p. 12)に準じて、criterion を「基準」、standard を「規準」とする。

2. 規準設定のための方法

規準設定の方法に関しては、多くの研究がなされてきている。年代順にその跡を見ると、Livingston and Zieky(1982)、Cizek(2006)、Hambleton & Pitoniak(2006)、Cizek & Bunch(2007)、Zieky, Pirie, and Livingston(2008)などがある。この中で、規準設定の方法

を 4 つに分類している Hambleton & Pitoniak (2006:, p. 440)によれば、それは、つぎのようになる。

(1) Methods that involve review of test items and scoring rubrics
(2) Methods that involve review of candidates
(3) Methods that involve looking at candidate work
(4) Methods that involve panelist review of score profiles

2. 1. テスト項目中心の方法：

Methods that involve review of test items and scoring rubrics

この分類で(1)に該当するものとしては、Angoff Method, Extended Angoff and Related Methods, Nedelsky Method, Jaeger Method, Bookmark and Other Item Mapping Methods, Direct Consensus Method をあげることができる。

このすべてを紹介する余白がないので、ここでは、こうした規準設定の方法を開発した方を知るための基本的な参考文献を以下紹介しておくこととする。

Angoff Method は、Angoff , W.H.(1971,pp. 508-597)に示されたのが初めである。その後、Hambleton & Plake(1995,pp.41-55)などによって、extended Angoff procedure が利用されるようになった。また、Ebel Method は、Ebel, R.L.(1972)で知られている。Nedelsky Method は、Nedelsky, I.(1954)が参考になる。Jaeger Method を知るための参考文献としては、Jaeger, R.M.(1989)がある。これは、わが国で行われた翻訳でも見ることができる。池田，藤田、柳井、繁桝(編訳)(1992)のなかの第 14 章「学生のコンピテンスの証明」(井上俊哉訳)がそれである。

Bookmark Method and Other Item Mapping Methods に関しては、このあとで、詳しく述べるので、ここでは、ごく簡単に、述べることとする。

この method が初めて紹介されたのは、1996 年である。これは、アリゾナの Phoenix で開催されたシンポジウム Symposium conducted at the Council of Chief State School Officers National Conference on Large-Scale Assessment で発表されたのが初めてであると言われている。発表者と、そのタイトルは、Lewis, D.M., Mitzel, H.C., and Green, D.R.(1996)の‘IRT-based standard setting procedures utilizing behavioral anchoring,’である。この method の特徴は、いくつかあげるとすれば、Mitzel, Lewis, Patz, and Green, D.R.(2001, p. 250)で述べているつぎのようなことである。

(a) integrates selected-response (SR) and constructed –response (CR)item formats.

(b) simplifies the judgmental task by reducing and or refocusing the cognitive load on the judges,
(c) connects the judgment task of setting cutscores to the measurement model, and
(d) connects test content with performance level descriptors.

Direct Consensus Method を理解するための参考文献の一つは、Sireci, S.G., Hambleton, R.K., & Pitoniak, M.J. (2004)である。また、(2)に該当するものとしては、Borderline -group Method と Contrasting -groups Method などが考えられている。

2．2．受験者中心の方法: Methods that involve review of candidates

この受験者中心の方法としては、2 つの方法が考えられてきている。しかし、いずれも、いくつかの問題が指摘されており、多くは使われてはいないのが現状である。

受験者中心の方法の一つ、Boderline-group method は、Zieky & Livingstone(1977)によって最初に提案された規準設定法である。これは、受験者を 3 つの group に分けることから開始される。まず、審査員を決定する。つぎは、「合格」グループ, 「不合格」グループ、さらには「境界線グループ」という 3 つのグループの技能とはいったい何かを検討する。そして「境界線グループ」を決定する。そのグループにテストを実施する。その結果を見て、テスト得点の中央値(median test score)を求め、それを分割点とするというものである。この場合では、述べた 3 つのグループをどうして決定するかが大きな問題であろう。

もう一つの方法 Contrasting-groups method の手順は、非常に簡単に述べれば、つぎのようになる。まず、最小限容認可能な能力とは何かを審査員に検討してもらう。そして、たしかな目標到達者と目標未到達者を審査員に決定してもらう。つぎは、2 つのグループにテストを実施する。テスト後、2 つのグループの成績分布を描く。そして、この 2 つの曲線の接点に分割点を設定する。

この 2 つの method に関しては、Zieky, Perie, & Livingston(2008, p. 79)でも述べているように、その disadvantages 問題点として、つぎのような状態が取り上げられている。たとえば、Contrasting groups method では、comparable evaluations of test takers for jurisdictions such as a state と、basic, proficient, advanced に関して、米国の州での受験者の統一した評価を設定することの難しさを述べている。また、Borderline group method においては、borderline test takers は、全学習者のうちの極めて少ない割合を占めるために、多くの関係者からの情報を求めなければならないなど、そこにも問題が多い。

以上のような受験者を中心にして分割点を設定しようと言う場合、多くの問題は、審査員の訓練にもあると考えられる。規準設定に関する大きな課題がこの審査員の訓練にあることは、テスト項目を中心にして行われた基準設定においても同じことであ

る。つまり、測定や判断の誤差による問題であろう。何を基準として到達者と判断するのか、あるいは、何を基準として未到達者と判断するのか、という点であろう。これが、主観的な判断であれば、多くの規準設定は問題を抱えたままの状態から抜け出すことは、極めて困難である。

2．3．その他の方法：

規準設定の方法として、テストを中心としたもの、受験者を中心としたものの 2 つをあげて、その概略を検討した。

テストを中心としたものと、受験者を中心としたものの分類に関しては、Hambleton & Pitonia (2006)、Cizek & Bunch(2007)、Zieky, Perie, & Livingston(2008)では、ほぼ同じような見方をしている。しかし、その他の分類では、それぞれが各自の方法で行っている。

Hambleton & Pitoniak(2006)における分類としては、前の分類のほかにあと 2 つがあげられている。3 番目の分類としては、methods that involve looking at candidate work、そして 4 番目の分類としては、methods that involve panelist review of score profiles である。ここでは、この 3 番目の分類と 4 番目について取り上げることとする。第 3 番目の分類に属するものとして、Item-by-item approaches, Holistic approaches, Hybrid approaches をあげることができる。第 4 番目の分類に属するものとしては、Judgemental policy capturing method, Dominant profile method, Item Cluster method をあげている。さらに、Compromise Methods である。これに関しては、Hofstee Method, Beuk Method, de Gruijter Method をあげている。

Cizek & Bunch (2007)における分類としては、Section 2．Standard-setting Methods として、12 に分類しているが、その中にある関連事項としては、The Hofstee and Beuk Methods がある。この方法は、「折衷的な方法」(compromise method)とも呼ばれるものである。Hofstee, W.K.B.(1983)、Beuk, C.H.(1984)が参考になる。

Zieky, Perie & Livingston(2008,pp.85-86)では、この分類は、Methods Based on Compromises between Absolute and Normative Judgments として取り上げられている。その中での、The Beuk Method では、In the Beuk method each participant specifies both a passing score and a pass rate. が見られる。また、The Hofstee Method においては、In Hofstee’s method, each participant specifies the highest and lowest acceptable passing score and the highest and lowest acceptable pass rate. などがある。

3．規準設定法に関するこれまでの評価：

これから究明しようとしている「規準設定法」に関して、わが国ではあまりその検討の跡が見えないけれども、外国における研究の流れはきわめて明白である。以下、

否定的評価、中立的評価、肯定的評価、妥当性検討などの流れを概観することとする。

3．1．否定的見方

Kaftandjieva, F.(2004, p. 31)は否定的評価の一つと考えることができる。

> “To summarize---there is no ‘ gold standard ’, there is no ‘true ’ cut-off score, there is no best standard setting method, there is no perfect training, there is no flawless implementation of any standard setting method on any occasion and there is never sufficiently strong validity evidence.”

つまり、規準設定に関しては、no gold standard, no true cut-off score, no best standard setting method, no perfect training など、極めて、厳しい評価をしている。これと、類似した否定的な評価としては、AERA, APA & NCME (1999, p. 53) では、つぎのような発言が見られる。

> “There can be no single method for determining cut scores for all tests or for all purposes, nor can there be any single set of procedures for establishing their defensibility.”

さらに、Jaeger and Mills (2001, p. 314) でのつぎの発言も、軽視することはできない。

> Standard setting has been called the ‘Achilles heel’ of educational testing(Hambleton & Plake(1998)largely because there is no clear consensus on the best choices among numerous methods and because the results of applying any method cannot easily be validated(Kane、1994)

3．2．中立的見方

こうした否定的な評価に対して、完全に否定はしないという、いわば中立的な立場を取っているのは、つぎの Cizek, G.J. and Bunch, M.B. (2007, p. 320)に見られる発言である。

> According to Segal, ‘A man with a watch knows what time it is. A man with two watches is never sure.’ Because there is no equivalent of atomic clock in the field of standard setting, our recommendation is simply for practitioners to invest in a single watch of greatest quality given available resources. ”

こうしたCizek らの中立的、しかも、建設的な意見は、今後の規準設定法の明るい方向を示していると考えることができる。さらに、努力して見事な時計を一つ見つけることこそ重要であるという、今後の研究に大いに期待する姿勢は、決して見逃すことはできない。

おなじように、データ収集には主観的な要素が入り込むけれども、決定した規準は極めて客観的で、重要であるという建設的な意見も見られる。Zieky, Perie & Livingston(2008, p. 197)でのつぎの結論も、決して見逃すことはできない。

> In this sense, all cutscores are subjective. Yet, once a cutscore has been set, the decisions based on it can be made objectively. Instead of a separate set of judgments for each test taker, you will have the same set of judgments applied to all test takers. Cutscores cannot be objectively determined, but they can be objectively applied.

3．3．肯定的見方

Nicholes, Twinge, Mueller, and O'Malley(2010, pp. 14 - 24)は、規準設定に関するきわめて新しい意見である。ここでは、これまでの規準設定法が「明らかに恣意的」(blatantly arbitrary)であったという意見に対し、「精神物理学的尺度」(psychophysical scaling)として知られている stimulus-centered scaling methods と、これまでの規準設定法を比較するなどして、まったく角度を変えて検討し直さなければならないとしている。

> Some writers in the measurement literature have been skeptical of the meaningfulness of achievement standards and described the standard-setting process as blatantly arbitrary. We argue that standard setting is more appropriately conceived of as a measurement process similar to student assessment. The construct being measured is the panelists' representation of student performance at the threshold of an achievement level.

時を同じにして、2011 年には、Bookmark-Based Methods の適切性を訴える論文も見られ、いわば、規準設定法の肯定的な方向が見られる。つぎの Peterson, Schulz & Engelhard (2011, pp. 3 - 14)は、その一例である。

> This research is used to evaluate Bookmark-based methods on key criteria originally considered by the Governing Board. Findings suggest that Bookmark-based methods have comparable reliability, resulting cut scores, and panelist evaluations to Angoff. Given that Bookmark-methods are shorter in duration and less costly, Bookmark-based methods may be preferable to Angoff for NAEP standard setting.

3．4．規準設定を評価する視点

Kane, M.T.(1994)や Fulcher, G.(Editor, *Language Testing*),(2010, pp. 225-252)では、規準設定を評価する視点を述べている。

おおくの規準設定法が開発されているが、その設定法がほんとうに適切か否かは何を基準として評価するのがよいか、それは、きわめて重要な事項である。これに関するこれまでの検討は、多く見られるが、その中で、まず、Kane(1994)の視点を探ることとする。Kane は、つぎの 3 つの視点をあげている。 第 1 は procedural evidence、第 2 は internal evidence、そして第 3 は external evidence といわれている事項である。この内容に関しては、Cizek(2006, p.235)に詳しく説明がある。そのなかで、出典 Pitoniak(2003)として示されている「規準設定評価の要素」(Standard-Setting Evaluation Elements)がその鍵であろう。

第 1 の観点としては、手続き上のことであるが、明示性(explicitness)、実用性(practicability)、手続きの実行(implementation)、審査員のフィードバック(feedback)、文書化(documentation)をあげている。第 2 の観点である内的課題であるが、方法の一貫性、審査員内の一貫性、審査員間の一貫性、決定の一貫性、そして他の測定値との関連を課題にしている。第 3 の観点としては外的な要素との検討である。その中では、他の規準設定法との比較、他の情報源との比較、そして、分割点の合理性などを課題にしている。

規準設定の方法は、きわめて重要である。しかも、その方法は、はたして妥当であったかどうかを検討しなければならない。では、何を基準としてその妥当性を検討することがよいのかは、今後の重要な課題として残されていると考えられる。

4．Bookmark Method の開発と課題

4．1．Bookmark Method の誕生と特徴

Bookmark Method が初めて言語テスト界に紹介されたのは、1996 年と言われている。筆者が生まれて初めて外国にでて、Georgetown University で Robert Lado 教授の指導をうけた年が 1965 年で、この Bookmark Method が誕生する 31 年前であった。この bookmark method も Item Response Theory もまったく耳には入ってこなかった時期である。この誕生に関しては、「2．1．テスト項目中心の方法」のところで、ごく簡単に述べているとおりである。つまり、Lewis, D.M. ,Mitzel, H.C. and Green, D.R.(1996)の発表がその誕生ということができる。

Bookmark Method の特徴としては、5 つのことをあげることができる。

まず、第 1 は、「項目応答理論」の活用である。古典的テスト理論の「正答数に基づく得点」(number right score)では、たとえば、38 点の意味は適切に捉えることはできない。それが、受験者の能力の低さを示すのか、それとも、テスト項目の困難度を示

すのかが、説明できないからである。周知の通り、項目応答理論では、そうした課題を解決でき、より正確なテスト項目困難度、より適切な受験者能力のもとで、基準判定が可能になるからである。

第 2 は、複数の分割点を設定できるからである。テストを 1 回実施すれば、複数の分割点を設定することができるということは、いままで開発された多くの規準設定法では、不可能であった。それが、可能であるのは、たとえば、ある一定の困難度を持つテスト項目を、正答確率 0.67 という程度で処理できるには、どのぐらいの能力を持った受験者が必要であるかなどを算出できるからである。

第 3 は、テスト項目が多肢選択形式でも、記述形式でも、いずれの場合にも使うことができるということである。これまでの方法では、例えば、多肢選択形式のテストの場合のみ使用可能ということがあった。しかし、この方法では、単に多肢選択形式のテスト、または、単に記述式のテスト、あるいは、多肢選択形式と記述式テストの混合の場合でも、データの処理は可能である。

第 4 は、審査員の作業を極度に簡素化することができるということである。たとえば、Nedelsky 法でも Ebel 法でも、審査員に課せられた作業と責任は大きいものがあった。「誤答と思われる選択肢はいくつあると判断されるか」とか「この選択肢は誤答であると最小限度達成者は判断できるか」とか、その作業と責任は大きい。Bookmark method では、その大半をコンピュータにまかすことができ、審査員の作業は極度に簡素化することができる。

第 5 は、テスト項目の内容も反映したといえる。評価が可能であるということである。この方法では、テスト項目判断の資料として、「順番付き項目冊子」(Ordered item booklet: OIB)が審査員に配布される。その冊子には、テスト項目それ自体、項目困難度、その正答確率を算出するための受験者の能力水準などが含まれていて、審査員の判断・評価の正確さを高めることが可能である。

以上の特徴は、これまでの規準設定法の効率化を高めるのに大いに役立ったと言える。

4．2．Response Probability の課題

Bookmark Method のなかで、頻繁に用いられている用語に response probability というものがある。この用語は、Bookmark Method を理解するのに、きわめて重要な概念である。この用語を理解するために、以下、2 つの説明を取り上げてみることとする。まず、Cizek & Bunch(2007, p. 162)での説明である。

> In the Bookmark procedure, the basic question participants must answer is “Is it likely that the minimally qualified or borderline examinee will answer this SR item correctly(or earn this CR item score point)?” Obviously it is important to define “likely” or to operationalize this decision rule. In practice, the Bookmark procedure employs a 67% likelihood(or sometimes a 2/3 chance)of desired response(i.e. of getting the SR item correct or of achieving a certain CR score point or higher).

ここで述べているように、審査員が答えなければならない、最も基本的な質問は、「このテスト項目に対して、最低の能力保持者、あるいは、境界線にある受験者は、正解を出す可能性があるか？」ということである。しかし、これをここで言う「可能性」とは、どんなことを意味するのであろうか？これを明確にしておかなければならない。実際問題として、bookmark procedure では、その正解を出す可能性を 67%としている、ということである。つまり、3 回の回答で、2 回の正解を出す可能性を指している。

同じように、関連事項に対する Zieky, Perie & Livingston(2008, p. 113)の説明は、以下の通りである。

> Ask participants to read through the Ordered Item Booklet from the easiest question to the hardest question and to place a “bookmark” at the point between the last question that border-line test takers would probably answer correctly and the first question that borderline test takers would not be able to answer correctly. The word “probably” is typically defined for this purpose as a probability of at least two-thirds, 2 out of 3, or .67. This probability is able called a *Response Probability* of .67 or RP67.

つまり、境界線上の受験者が正解することが、多分、できると思われる最後のテスト項目とその受験者が、多分、正解できないと思われる最初の項目との間に「しおり」を置くように，審査員は依頼される。しかし、この場合の“probably”(多分)の意味を明確にしておかなければならないとしている。その「多分」が、67%の確率を意味するのであれば、つまり、3 回の試行において 2 回の正解を出せるような確率であれば、

その確率を「正答確率.67」(response probability of .67)あるいは、「正答確率 67」(RP67)と呼ぶこととしている。

正答確率は、すべて.67 でなければならないというのではない。多くの実験研究では、2PLM の場合は．67 の正答確率が情報関数を最大にするという機能を最も高めるということが証明されている。しかし、1PLM、つまり、Rasch Model が用いられた場合には.50 の正答確率が良いであろうという意見もある。たとえば、Wang, N.(2003)がその 1 例である。.50 の正答確率がよいとされるその理由は、たとえば、1PLM では、P = 1/(1+exp (- (θ- b)))のなかの θ と b を同じ 2.0 とした場合には、P=1/(1+exp(-(2-2)))となり、P=1/(1+exp(-0)))となるので、P＝1/(1+1)となり P=0.5 となるからである。つまり.50 の方が.67 よりもよいという数理的な利点は、受験者の能力と項目の困難度が同じになったときは、正答確率は、ちょうど、0.50 になるからだと説明している。

正答確率を、0.67 や 0.50 だけではなく、RP＝0.80 とすることを勧めている場合も見られる。たとえば、Bock, R.D., Mislevey, R., &Woodson, C.(1982)がその一例である。そのことに関しては、Mitzel, Lewis, Patz, and Green(2001, p. 262)では、”Alternate RP levels have been used or proposed by others. Bock, Mislevey, & Woodson(1982)made an early suggestion of RP = .80 for mastery. In its item anchoring procedures, response probabilities of .80 (Educational Testing Service, 1987) and .65 have been used by NAEP.”と述べている。

このように、正答確率は、0.67, 0.50, 0.80 ということも考えられるが、つぎの Zieky, Peri & Livingston(2008, p. 113)で示されているように、0.67 が Bookmark method では、最も多く用いられている。Response probabilities other than .67, such as .50 and .80 have been used, but .67 is the most commonly used response probability for Bookmark studies.

4．3．Bookmark を置く場所

Bookmark Method のなかで審査員が行わなければならない最も重要なことのひとつは、与えられた OIB (ordered item booklet) を見て、どこに bookmark を置くべきかを決定しなければならないことである。これが十分検討されないままこの方法が実施されると、やはり、Bookmark Method は、主観に頼るしかない方法であるということになってしまう危険がきわめて大きい。これまで、この置き場所に関して与えられている指示は、どんなものであったかを、少し，整理しておくことが必要であろう。

Cizek, G.J., Bunch, M., & Koons, H.(2004, p. 37)における指示では、以下のようになっている。

> Standard-setting participants are instructed to place a marker in their OIB on the page(i.e., item)*immediately after* the page at which, in their opinion, the likelihood criterion applies, that is, to place their bookmarks at the first point in the booklet at

which they believe examinees' probability of marking the desired response drops below .67.

これを要約すると、つぎのようになる。規準設定の審査員たちは、OIB のなかで、彼らの判断する見込みの基準があてはまる頁のすぐ後の頁(テスト項目)に「しおり」(bookmark)を置くように指示される。つまり、受験者が正解する確率が 67%以下になるだろうと審査員が信じる最初の頁にしおりを置くように指示されるというものである。Hambleton, R.M., & Pitoniak, M.J.(2006, p. 443)では、つぎのように説明している。

The task for the panelist is to place a bookmark between the two items in the ordered item booklet such that from his or her perspective, those items before the bookmark represent content that borderline examinees at a given performance standard should be likely to know and be able to do. IT should be noted that although Lewis and colleagues (Lewis, et al, 1996, 1998; Mitzel, et al, 2001) described the placement as being between two items, others have operationalized the task for panelists as putting the bookmark on the last item the borderline examinees would be likely to answer correctly. As Cizek, Bunch, and koons (2005) pointed out, however, both approaches lead to the same result.

また、Cizek, G.J.(2006, p. 247)での説明は、つぎの通りである。

Mitzel, Lewis, Patz and Green (2001) recommend that the probability judgement be referenced to a 67 percent likelihood, which they refer for as the *response probability* (RP). According to Mitzel, et al. (2001), and RP of .67 can be interpreted in the following way: "For a given cut score, a student with a test score at that point will have a .67 probability of answering an item also at that cut score correctly" (p. 260). Thus, participants are instructed to place a marker on the first page in their OIB at which in their opinion, the RP drops below .67.

ごく簡単に、この bookmark の置き方を説明すれば、Cizek, G.J. & Bunch, M.B.(2007, p. 184)"Participants place their bookmarks on the last item in the OIB for which they believe a minimally qualified examinee has a 2/3 chance of answering correctly.ということになる。

まとめとして、Zieky, M.J., Perie, M., & Livingston, S.A.(2008, p. 113)をあげておくと、つぎのようになる。

> Ask participants to read through the Ordered Item Booklet from the easiest question to the hardest question and to place a “bookmark” at the point between the last question that border-line test takers would probably answer correctly and the first question that borderline test takers would not be able to answer correctly. The word “probably” is typically defined for this purpose as a probability of at least two-thirds, 2 out of 3, or .67. This probability is able called a *Response Probability* of .67 or RP67.

4．4．精神物理学の課題

応答確率が 67%である時点を判断するという作業は、審査員にとっては、きわめて困難な作業であり、妥当な説明を継続するに十分ではないのではないかというふうに考えられる。この OIB での準備は、発想としては、かなり単純なものであったようである。しかし、この考え方は、つぎの英文で説明しているように、「古典的精神物理学」(classical psychophysics)という領域に変換してしまったようである。そのことは、審査員の正答確率の取り扱いを説明する方法としては、より適切であると考えられたからであろう。Cizek, Bunch, & Koons(2004, p. 36)では、こう伝えている。

> The idea, however, instantly transformed standard setting into a classical psychophysics experiment in which a stimulus of gradually changing strength or form is presented to subjects who are given the task of noting the point at which a just-noticeable difference (JND) occurs.

つまり、徐々にかわってゆく力や形の刺激に触れれば、「著しい相違」(just-noticeable difference: JND)におこる時点に気づく力が与えられるであろうということである。審査員は、それぞれの次にくる項目は、その前の項目より困難であるということを知って審査を開始する。それをおこなっているうちに、OIB のなかのいくつかの項目のなかに、1 つ、または 2 つの「著しい相違」に気づくであろうということである。それが、bookmark の置くべき場所と関連するであろうということである。

精神物理学(psychophysics)とは、どんなものであるかということの追求は、ここでは控えておく。しかし、この学問は、非常に簡単に言えば、外的な刺激と内的な感覚の対応関係を測定し、また、定量的な計測をしようとする学問であるといえる。ちなみに、「精神物理学的測定法」(psychophysical method)は、東・梅本・芝・梶田(編)(1988, p. 364)には、つぎのような説明が見られる。

精神物理学とは、フェヒナー(Fechner,C.T.)に由来し、精神と身体(物体)とを結ぶ法則を扱う学とされた。フェヒナーは、精神的感覚量と身体的刺激量との間の量的関係を記述するに当たって、精神的感覚の増加には、身体的刺激量に比例した大きさの変化が対応することに気づき、また、物的エネルギーの測定は容易なので、これによって感覚の量を示そうとした。

これだけでは、bookmark の置き場所を見つけ出す手がかりは見えだせない。さらに、究明しなければならない内容である。先に述べた“noticeable difference”を発見した場所は、どんな所なのかを、項目困難度、弁別力、あるいは、その時の受験者の能力などの要素から物理的に発見できるデータをどうしたら求めることが可能であるかが究明できれば、よいのではないかと考える。審査員が決定した時点を詳しく検討して、その時点を、物理的なデータを用いて究明できないか考えることは、きわめて重要な課題であろう。

4．5．応答確率と受験者の能力

大友賢二(監修)、中村洋一、小泉利恵(編集)(2009, p. 107)においては、正答確率の時点での受験者の能力を推定できる計算式を開発し提示している。この計算式は、合否判定の資料として、きわめて重要である。たとえば、67%の正答確率で、bookmark を置く場所が発見された場合、その場合の受験者の能力を推定し、それをもって合否判定のための分割点の糸口を見いだすことができるからである。1PLM, 2PLM, 3PLM において、正答確率、項目困難度、弁別力指数、当て推量などがわかっていれば、つぎのようにして、その受験者の能力は、推定できる。

1PLM：$P=1/ (1+\exp (- (\theta - b)))$ $\theta= \ln (P/ (1-P)) + b$
2PLM: $P = 1/ (1+\exp (-Da (\theta- b)))$ $\theta=\ln (P/ (1-P))/ (Da) + b$
3PLM: $P = c + (1- c)* (1/ (1+\exp (-Da (\theta-b))))$
$\Theta = \ln ((P/ (1-P))* (1-c)-c)/ (Da)+b$

たとえば、1PLM を用いて、データの分析を行った場合、0.67 の正答確率で、困難度が-3.242 のテスト項目に答えられる受験者の能力は、$\theta = \ln (0.67/ (1-0.67))+ (-3.241)$ から、-2.533 であることが推定できる。たとえば、2PLM を使ってデータの分析をした場合は、0.67 の正答確率で、困難度 2.41、弁別力 0.94 のテスト項目に答えられる受験者の能力は、$\theta = \ln (0.67/ (1-0.67))/ (1.7*0.94) + 2.410$ から 2.853 であることが推定できる。また、3PLM を使ってデータの分析をした場合は、0.67 の正答確率で、困難度-0.260、弁別力 1.16、当て推量 0.18 のテスト項目に答えられる受験者の能力は、$\theta = \ln ((0.67/$

(1-0.67))* (1-0.18)-0.18)/ (1.7*1.16)-0.26 から-0.060 であることが推定できる。

5. データによる分割点の推定

5．1. Schagen and Bradshaw(2003)をめぐって

Bookmark Method におけるデータ分析の例は、多く見られるが、ここでは、Cizek, Bunch, and Koons(2004, pp. 39-40)で取り上げている Schagen, I. and Bradshaw, J.(2003, September)のデータを検討してみることとする。

Table 5.1.1. Ordered Item Booklet Parameters and Associated Theta Values

PNO	TIN	Difficulty	(b)Discrim	(a)Theta@RP=.67
1	19	-3.395	0.493	-2.550
2	13	-2.770	0.997	-2.352
3	01	-2.757	1.441	-2.468
4	22	-2.409	0.461	-1.505
5	04	-2.282	0.527	-1.492
6	02	-2.203	0.607	-1.517
7	12	-2.141	0.503	-1.313
8	03	-1.781	0.520	-0.980
9	14	-1.737	0.931	-1.290
10	31.1	-1.710	0.817	-1.240

Table 5.1.1.は、Ordered Item Booklet(OIB)に関する表である。この項目の配列は、Difficulty の易しい項目から難しい項目へという順序となっている。この表は、PNO, TIN, Difficulty, Discrimination, Theta の列で構成されている。最初の PNO は、OIB を構成している page の番号を指している。これを Page Number in OIB(PNO)と呼ぶ。ここでは、この booklet を構成しているのが 10 頁ある。グラフを作成したりする場合は、この PNO 順にデータを使う。つぎの TIN というのは、このデータを作成する元になったテスト項目の番号である。これを Test Item Number(TIN)と呼ぶこととする。Difficulty と Discrimination は、それぞれの TIN の項目を 2PLM の IRT で分析した場合に算出された parameter である。Theta は、そのテスト項目を正答確率.67 で回答できる受験者の能力を示すものである。

実際の表では、テスト項目が 50 である。紙面の都合上、以上の 10 項目に関するデータをここでは示すこととする。Theta@RP=.67 は、先に説明してあるように、正答確率が 0.67 の場合の受験者の能力を示す。たとえば、最初に示してある項目 19 の場合を取り上げると 2PLM で分析した結果は、この項目の困難度は、-3.395、弁別力は 0.493 である。この項目を正答確率 0.67 で答えることができると推定される能力はいくらかを算出した結果が、-2.550 ということである。この算出方法は、先に示した数式を使

えば、θ ＝ln(0.67/(1-0.67))/(1.7*0.493)+(-3.395)=-2.550 となることは、明らかである。

このデータを用いて、12 名の審査員が示した Bookmark の置き場所がこの論文に示してある。それを要約すると、TIN=2(PNO=6)としたものが 7 名、TIN=04(PNO=5)としたものが 3 名、TIN=13(PNO=2)としたものが 2 名であると述べている。もっとも多いのは、Difficulty-2.203、Discrim 0.607 である TIN=2 で、この項目を正答確率 0.67 で回答できる受験者の能力は-1.517 となっている。ここで解明したい最も大きな問題は、なぜ、どんな理由で、TIN=2 に多くの審査員が bookmark を置いたのかということである。

さきに示したように、審査員が行うことは、「正答確率が.67 以下に下がると思われる OIB の最初の頁に bookmark を置くこと」(to place a marker on the first page in their OIB at which, in their opinion, the RP drops below .67)(Cizek(2006, p. 247))である。正答確率が.67 以下に下がる「と思われる」(in their opinion)とした場合、その「思われ方」は、審査員によってまちまちであろう。審査員に対して、そう思わせる要素、あるいは、データは何であろうか？それは、困難度であるのか、弁別力であるのか、あるいは、正答確率であろうか？あるいは、他の要素であるのか？それを調べるのが、解決策になるのだろうか？様々なことが考えられる。一つの試みとして、求めているデータの変化を、直感ではなくて、グラフのデータで捉えることは可能なのであろうか？それを解決するために、困難度、弁別力、受験者能力をそれぞれ低い方から高い方に順に並べて、その状況を判断してみることにする。

Table 5.1.2. 低から高へ配列した DIF, DISC, THETA

PNO	Difficulty (PNO)	Discrimination(PNO)	Theta (PNO)
1, 4, 1	-3.395 (1)	0.461(4)	-2.550 (1)
2, 1, 3	-2.770 (2)	0.493(1)	-2.468 (3)
3, 7, 2	-2.757 (3)	0.503(7)	-2.352 (2)
4, 8, 6	-2.409 (4)	0.520(8)	-1.517 (6)
5, 5, 4	-2.282 (5)	0.527(5)	-1.505 (4)
6, 6, 5	-2.203 (6)	0.607(6)	-1.492 (5)
7, 2, 7	-2.141 (7)	0.997(2)	-1.313 (7)
8, 3, 8	-1.781 (8)	1.441(3)	-0.980 (8)

「正答確率が 0.67 以下に下がると思われる OIB の最初の頁に bookmark を置く」という判断は、何を基にして判断するのかを確かめるために、DIFF, DISC, THETA の変動を示す以下のようにグラフを書いてみた。数値という視的な感覚が、その判断を

決定する要因となっているかを知りたいからである。Table 5.1.2.のデータをグラフで見ると以下のようになる。

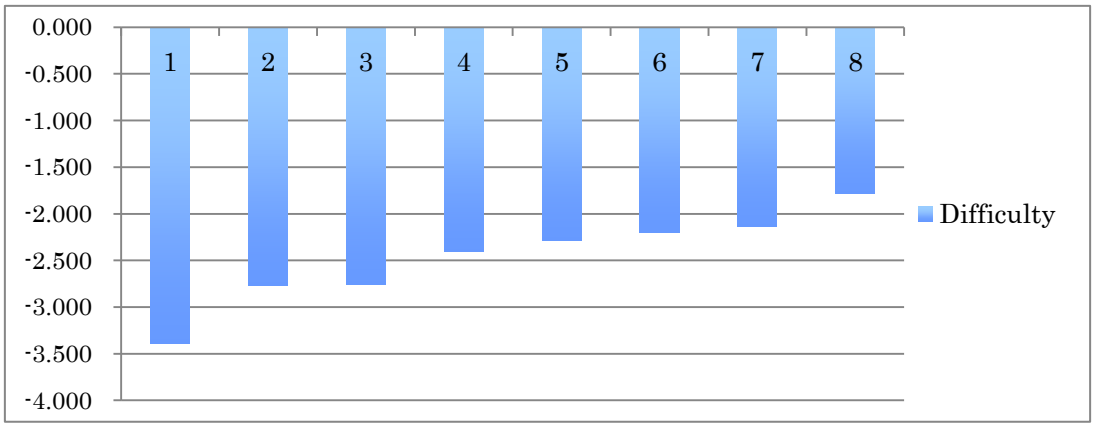


Figure 5.1.1. DIFFICULTY

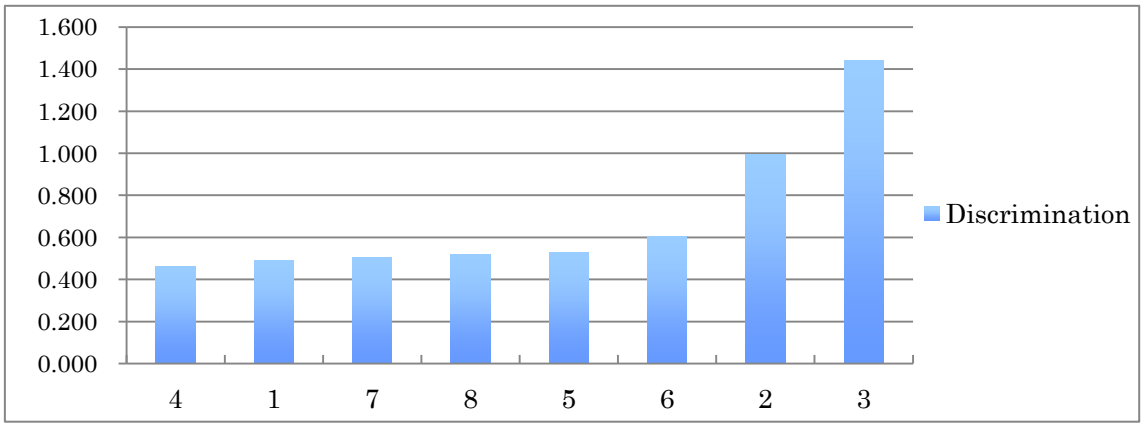


Figure 5.1.2. DISCRIMINATION

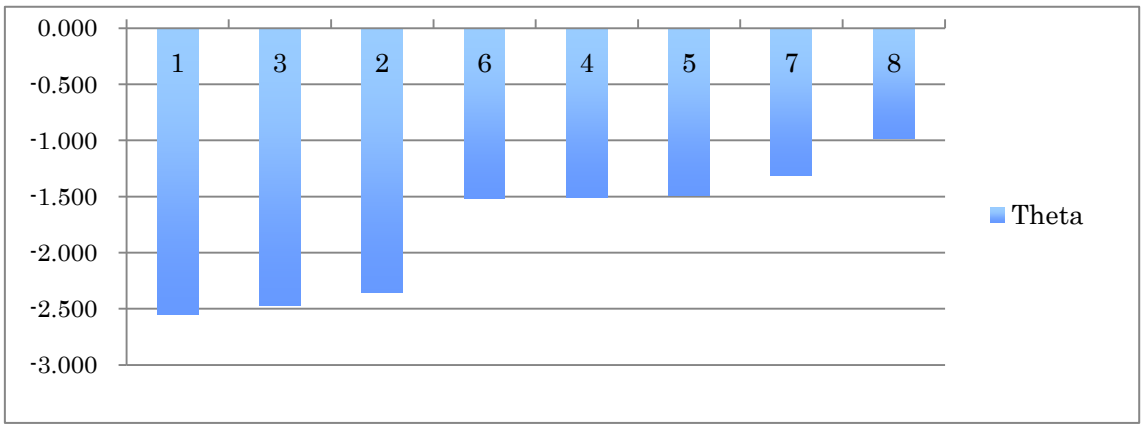


Figure 5.1.3 THETA

このグラフを見る限り、項目の間では、「著しい相違」を見いだすことは、やや困難である。

最初は、DIFFICULTY を資料にしたものである。このグラフ(Figure 5.1.1.)を見てわかることは、最初は-3.395 が困難度であり、最後は-1.781 の困難度を示しているものである。この最初から最後まで見渡して、感じることは、「著しい相違」はさほど感じない。あえて言えば、PNO(1)から(2)への移動における少ない相違が見られる。また、PNO(3)から(4)への移動におけるこれまた少ない相違である。それ以外の「著しい相違」はとくにあるとは言いがたい。

第 2 番目のグラフ:Figure 5.1.2.は、DISCRIMINATION を資料にしたものである。この資料を構成しているものに注目してみる。最初は、0.461 であり、最後は、1.441 の弁別力を示しているものである。最初から最後まで見渡して、感じることは、「著しい相違」と言えば、PNO6 から PNO7 への移動に関するものである。同じように、PNO7 から PNO8 までの相違も、他と比べれば、大きいと言えよう

第 3 番目のグラフ(Figure5.1.3)は、Theta@RP=0.67 を資料にしたものである。この能力値の配列は、最初は、-2.550 で、最後は-0.980 のものである。このグラフを最初から最後まで見渡して感じる「著しい相違」は、あえて言えば、PNO3 から PNO4 への移動に関するものである。それ以外は、分割点を意味するような大きな相違は、見当たらない。

5. 2. 「PNO 間の数値差」を利用した推定

さきに、Table 5.1.2.を用いて、分割点の設定を求めるためのグラフ作成を試みた。しかし、その結果は、審査員の求めた bookmark の置き場所に近い位置を見いだすための適切にして十分なデータを求めることはできなかった。そのため、さらなる推定法を見いだすために、グラフデータの修正を試みた。Table 5.1.2 で、それぞれのデータ間にあまり「著しい相違」を見つけることはできなかったが、「PNO 間の数値差」を明確に示すことによって、「著しい相違」を見いだせるのではないかと考えた。Page number in OIB 間の数値差が大きければ、「著しい相違」がより明確になるのではないかという発想である。

DIFFICULTY

GDN(PNO-PNO)	Difference		
1(1—2)	-0.625	-3.395	-2.770
2(2—3)	-0.013	-2.770	-2.757
3(3—4)	-0.348	-2.757	-2.409
4(4—5)	-0.127	-2.409	-2.282
5(5—6)	-0.079	-2.282	-2.203
6(6—7)	-0.062	-2.203	-2.141
7(7—8)	-0.360	-2.141	-1.781

DISCRIMINATION

GDN(PNO-PNO)	Difference		
1(4—1)	-0.032	0.461	0.493
2(1—7)	-0.010	0.493	0.503
3(7—8)	-0.017	0.503	0.520
4(8—5)	-0.007	0.520	0.527
5(5—6)	-0.080	0.527	0.607
6(6—2)	-0.390	0.607	0.997
7(2—3)	-0.444	0.997	1.441

THETA

GDN(PNO-PNO)	Difference		
1(1—3)	-0.082	-2.550	-2.468
2(3—2)	-0.116	-2.468	-2.352
3(2—6)	-0.835	-2.352	-1.517
4(6—4)	-0.012	-1.517	-1.505
5(4—5)	-0.013	-1.505	-1.492
6(5—7)	-0.179	-1.492	-1.313
7(7—8)	-0.333	-1.313	-0.980

ここでは、page number 間の数値差を求めるので、その差をグラフで示す順序をGDN(graph data number)という記号で示すこととする。そのために、たとえば、difficultyにおいては、PNO 1(-3.395)と 2(-2.770)との差(-0.625)を求め、その差を示すデータをGDN<1>として定めておくという方式をとった。つぎのデータは、PNO2 と 3 との差を示すデータを GDN<2>とすることであった。同様にして、最後の GDN<7>は、PNO7(-2.141)と 8(-1.781)との差(-0.360)を示すものである。

この「PNO 間の差」のデータを用いて作ったグラフは、以下のようなものである。

以下の 3 つのグラフの最初のグラフ(Figure 5.2.1 DIFFICULTY)は Difficulty に関するものである。上の表でわかるように、GDN(PNO-PNO)の下に示されている数値、たとえば、1-2 というのは、PNO1(-3.395)から 2(-2.770)までの差(-0.625)をデータとして示しているものである。グラフで見ると、横線のデータ<1>からデータ<7>までの番号のうち、このグラフで直感的に気づくのは、GDN<1>から<2>への変動の大きさである。GDN<2>は、difference の最大(-0.625)を示す GDN<1>の直後にあるからである。さらに、この両者に共通に含まれている PNO は 2 である。また、GDN<6>から<7>への変動の大きさである。この GDN に含まれている PNO は 6 である。先にあげた 12 名の審査員のうち 7 名が選んだ bookmark の置き場所は、PNO6 であった。この GDN<6>は、PNO7 と 8 の差を示す GDN<7>という大きい変動を持つ時点の直前の位置にあり、これが「著しい相違」を示す重要な地点であると考えることができる。したがって、PNO2 と PNO6 に bookmark を置くのが最も適切と判断できる。

こうしてみると、この「PNO 間の差を利用した推定」を探れば、審査員が bookmark の置き場所とした、PNO6 と 2 に最も近い場所を求めることは可能であると考えられる。そして、もう一つの大きな前進は、多くの審査員の示した bookmark の置き場所は、精神物理学的な直感や主観だけではなく、データに基づく「PNO 間の差を利用した推定」という方法でも取得可能ではないかと言うことである。

つぎの Figure 5.2.2 の<discrimination>でも，「PNO 間の差を利用した推定」は有効であるかどうかを検討してみることにする。グラフ上の「著しい相違」は、どこに見いだすことができるであろうか？GDN<5>と<6>の周辺にそれを見いだすことができる。GDN<5>というのは、PNO6(0.607)と 2(0.997)との大きい差(-0.390)を示す GDN<6>の直前の位置にあり、DIFF の場合と同じように、「著しい相違」を示す重要な地点であると考えることができる。GDN<5>を構成しているのは、PNO5 と 6 である。また、GDN<6>を構成しているのは、PNO6 と 2 であり、両者を共通に構成しているのは、PNO6 である。また、GDN<6>は、difference が最大(-0.444)を示す GDN<7>の直前にある。さらに、この両者に共通に含まれる PNO は 2 である。したがって、PNO6 と PNO2 に bookmark を置くのが、最も適切と判断される。

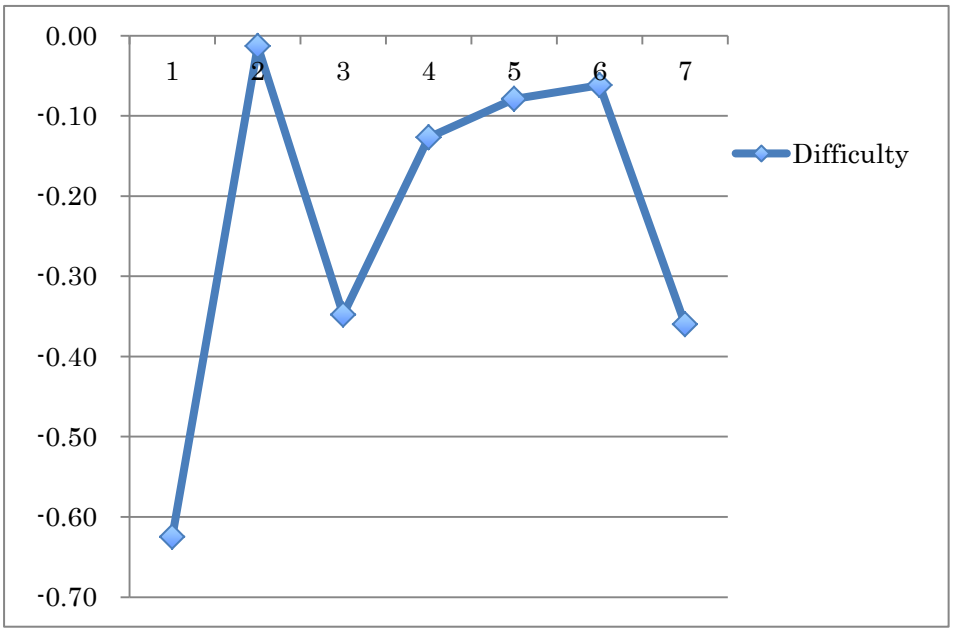


Figure 5. 2. 1.<DIFFICULTY> : PNO 間の数値差

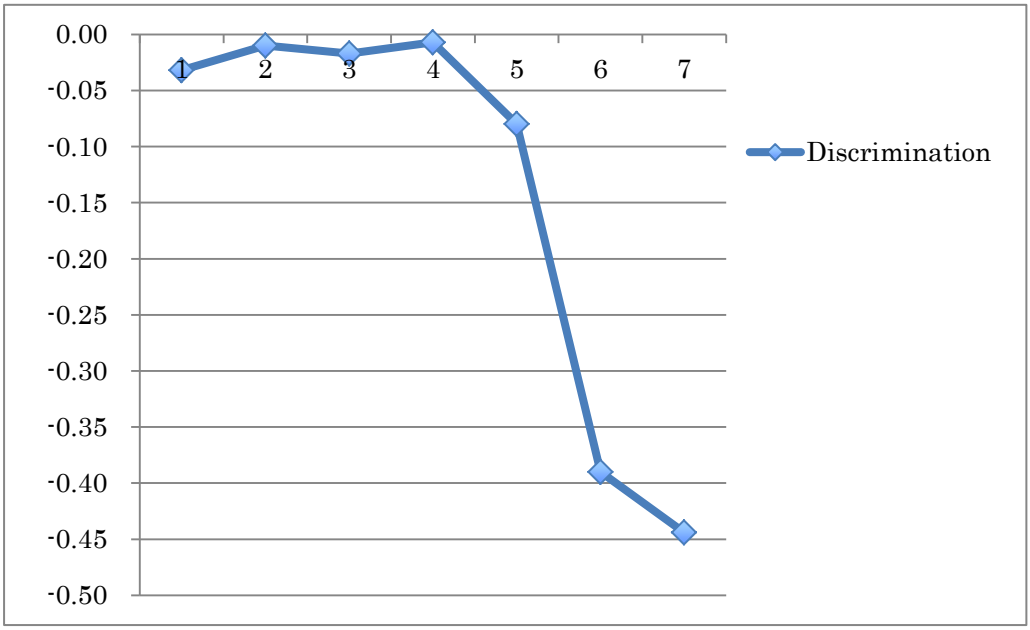


Figure 5. 2. 2. <DISCRIMINATION>: PNO 間の数値差

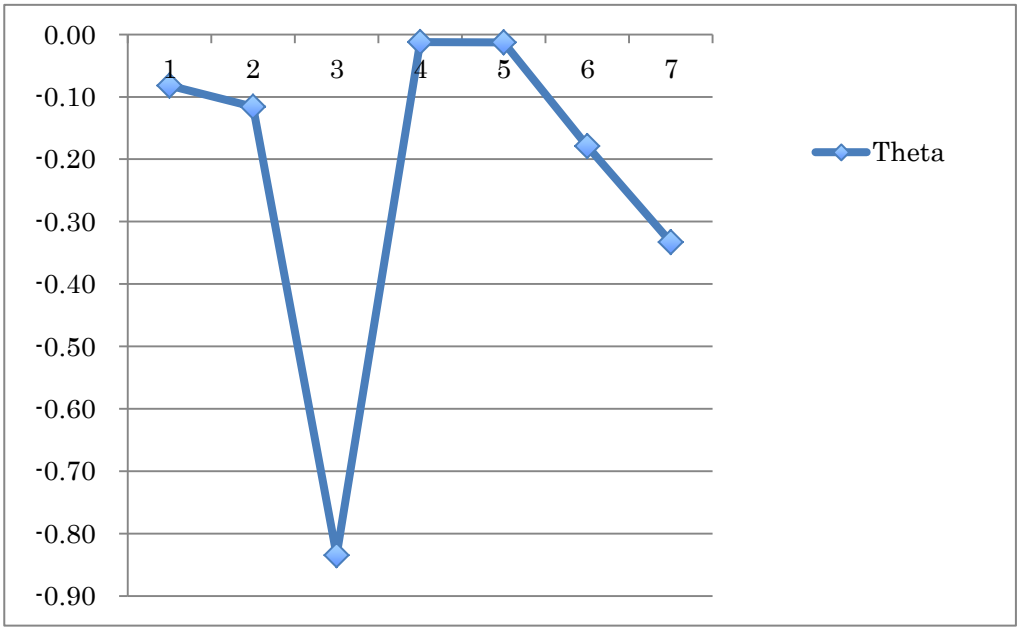


Figure 5. 2. 3. <THETA>: PNO 間の数値差

最後の Figure 5.2.3.<THETA>でも「PNO 間の数値差を利用した推定」は有効であるかどうかを検討してみることにする。グラフ上の「著しい相違」はどこに見いだすことができるであろうか？ここでは、GDN<2>から<3>への移動、また、GDN<5>から<6>、<7>への移動にその傾向が見られる。この GDN<2>というのは PNO2(-2.352)と 6(-1.517)との大きい差(-0.835)を示す GDN<3>の直前の位置にある「著しい相違」をしめす重要

な地点である。この GDN<2>を構成しているのは、PNO3 と 2 である。また GDN<3>を構成しているのは、PNO2 と 6 であり、両者を構成しているのは PNO2 である。したがって、PNO2 に booklet を置くのが、最も適切と判断される。

また、GDN<5>、<6>、<7>の周辺にそれを見いだすことができる。この GDN<5>というものは、PNO5 と 7 との大きな差(-0.179)を示す GDN<6>の直前の位置にあり、また、GDN<6>というのも、PNO7 と 8 との大きい差(-0.333)を示す GDN<7>の直前にあり、やはり、「著しい相違」を示す重要な地点であると考えることができる。この GDN<5>を構成しているのは、PNO4 と 5 である。また、GDN<6>を構成しているのは、PNO5 と 7 であり、両者を共通して構成しているのは、PNO5 である。したがって、PNO5 に bookmark を置くのが、最も適切と判断される。

以上の 3 つのグラフから判断して言えることは、つぎのとおりである。

DIFFICULTY のグラフから言えることは、GDN<6>を構成している PNO6、および、GDN<1>と GDN<2>を共通に構成している PNO2 は、審査員の判断と密接に関係し、bookmark の置き場所の候補として適切である。

DISCRIMINATION のグラフから言えることは、GDN<5>と GDN<6>を共通に構成している PNO6、および、GDN<6>と GDN<7>を共通に構成している PNO2 は、審査員の判断と密接に関係し、bookmark の置き場所の候補として適切である。

THETA のグラフから言えることは、GDN<2>と GDN<3>を共通に構成している PNO2、また、GDN<5>、<6>を共通に構成している PNO5 は，審査員の判断と密接に関係し、bookmark の置き場所の候補として適切である。

以上、この PNO の間の数値差を利用すれば、審査員の判断した bookmark の置き場所である PNO6, 5, 2 を明確に推定することができることが判明した。

6. Wright & Stone データの検証

PNO 間の数値差を利用すれば、bookmark の適切な置き場所を探すことがでできるという糸口が Schagen and Bradshaw (2003) のデータを使って探し求めることができた。しかし、この糸口発見は、ほかのデータでもたしかに可能なのであろうか？それを検証するために、ここでは、Wright, B. D. and Stone, M. H. (1979, p. 31).Table 2.3.1. Original Response of 35 Persons 18 Items on the KNOX CUBE TEST. (In *Best Test Design,* Chicago, MESA Press)の 10 データを ABC データに変換して使ってみることとする。

Table 6.1.
KNOX CUBE TEST: ABC data

		ABCDA	BCDAB	CDABC	DAB	
10001	‘10001	ABCDA	BCABC	DABCD	ABC	07
10002	‘10002	ABCDA	BCDAB	ABCDA	BCD	10
10003	‘10003	ABCDA	BCDAA	BDDAB	CDA	10
10004	‘10004	ABCDD	ACCAD	ABCDA	BCD	06
10005	‘10005	ABCDA	BCDAB	ABCDA	BCD	10
10006	‘10006	ABCDA	BCDAB	DABCD	ABC	10
10007	‘10007	ABCDA	BCDAB	CDAAC	ABC	14
10008	‘10008	ABCDA	BCDAB	BCDAB	CDA	10
10009	‘10009	ABCDA	BCDAB	BCDAB	CDA	10
10010	‘10010	ABCDA	BCDAB	CABCD	ABC	11
10011	‘10011	ABCAA	BCDAA	BCDAB	CDA	08
10012	‘10012	ABCDA	ACAAB	ABCDA	BCD	08
10013	‘10013	ABCDA	DDDAB	CDDAB	CDA	10
10014	‘10014	ABCDA	BCDAB	CABCD	ABC	11
10015	‘10015	ABCDA	BCDAB	CDAAB	CDA	13
10016	‘10016	ABCDA	BCDAD	CABCD	ABC	10
10017	‘10017	ABCDD	BCDAB	ABCDA	BCD	09
10018	‘10018	ABCDA	BCDAB	ABAAB	CDA	11
10019	‘10019	ABCDA	BCDAA	BCDAB	CDA	09
10020	‘10020	ABCDA	BCDAB	ABAAB	CDA	11
10021	‘10021	ABCDA	BCDAB	CAAAB	CDA	12
10022	‘10022	ABCDA	BCDAB	CDDAB	CDA	12
10023	‘10023	ABCDA	BCDAB	BBABA	BCD	12
10024	‘10024	ABCDA	BCDAB	CAACD	DAA	14
10025	‘10025	ABCCA	BABCD	ABCDA	BCD	05
10026	‘10026	ABCDA	BCDAB	DABCD	ABC	10
10027	‘10027	ABCDA	BCABC	DABCD	ABC	07
10028	‘10028	ABCDA	BCDAA	CBCDA	BCD	10
10029	‘10029	ABCDA	BAAAB	CBBBA	BCD	10
10030	‘10030	ABCDA	BCDAA	BCDAB	CDA	09
10031	‘10031	ABCDA	BCDAB	XXXXX	XZZ	10
10032	‘10032	ABCDA	BCDAB	CBCDA	BCD	11
10033	‘10033	ABCDX	XCCCB	ABCDA	BCD	06
10034	‘10034	ABCDA	BCDAB	ADXBA	BCD	12
10035	‘10035	ABCAB	CDABC	DABCX	XXZ	03

Table 6. 2.

KNOX CUBE TEST: Person Score

01-10: 07, 10,10,06,10,10,14,10, 10.11

11-20: 08,08,10, 11, 13, 10, 09, 11, 09, 11

21-30: 12, 12, 12, 14, 05, 10, 07, 10, 10, 09

31-35: 10, 11, 06, 12, 03

Table 6. 3.

KNOX CUBE TEST: Item Score

01-09: 35,35,35,32,31,30,31,27,30

10-18: 24,12,06,07,03,01,01,01,00

Table 6. 4.

KNOX CUBE TEST: 2PLM by XCALIBRE (tm)for Windows 95/NT Version 1.10, (1995) Assessment System Corporation

ITEM	**a-parameter**	**b-parameter**
1	deleted	deleted
2	deleted	deleted
3	deleted	deleted
4	0.93	-1.93
5	0.92	-1.76
6	0.90	-1.57
7	0.86	-1.78
8	0.95	-1.10
9	0.95	-1.55
10	0.87	-0.73
11	0.85	-0.69
12	0.85	1.47
13	0.93	1.31
14	0.85	1.97
15	0.94	2.41

16	0.95	2.40
17	0.95	2.39
18	deleted	deleted

Table 6.5 Wright&Stone,

PNO

	TIN		DIFFICULTY	DISCRIMI	THETA
1		4	-1.93	0.93	-1.482
2		7	-1.78	0.86	-1.296
3		5	-1.76	0.92	-1.307
4		6	-1.57	0.9	-1.107
5		9	-1.55	0.95	-1.112
6		8	-1.10	0.95	-0.662
7		10	-0.73	0.87	-0.251
8		11	0.69	0.85	1.18
9		13	1.31	0.93	1.758
10		12	1.47	0.85	1.96
11		14	1.97	0.85	2.46
12		17	2.39	0.95	2.828
13		16	2.40	0.95	2.838
14		15	2.41	0.94	2.853

データ

DIFFICULTY
GDN (TIN-TIN)

		C	D	C-D
1	(4—7)	-1.93	-1.78	-0.15
2	(7—5)	-1.78	-1.76	-0.02
3	(5—6)	-1.76	-1.57	-0.19
4	(6—9)	-1.57	-1.55	-0.02
5	(9—8)	-1.55	-1.1	-0.45
6	(8—10)	-1.1	-0.73	-0.37
7	(10—11)	-0.73	0.69	-1.42
8	(11—13)	0.69	1.31	-0.62
9	(13—12)	1.31	1.47	-0.16
10	(12—14)	1.47	1.97	-0.5
11	(14—17)	1.97	2.39	-0.42
12	(17—16)	2.39	2.4	-0.01
13	(16—15)	2.4	2.41	-0.01
14	(15--	2.41		

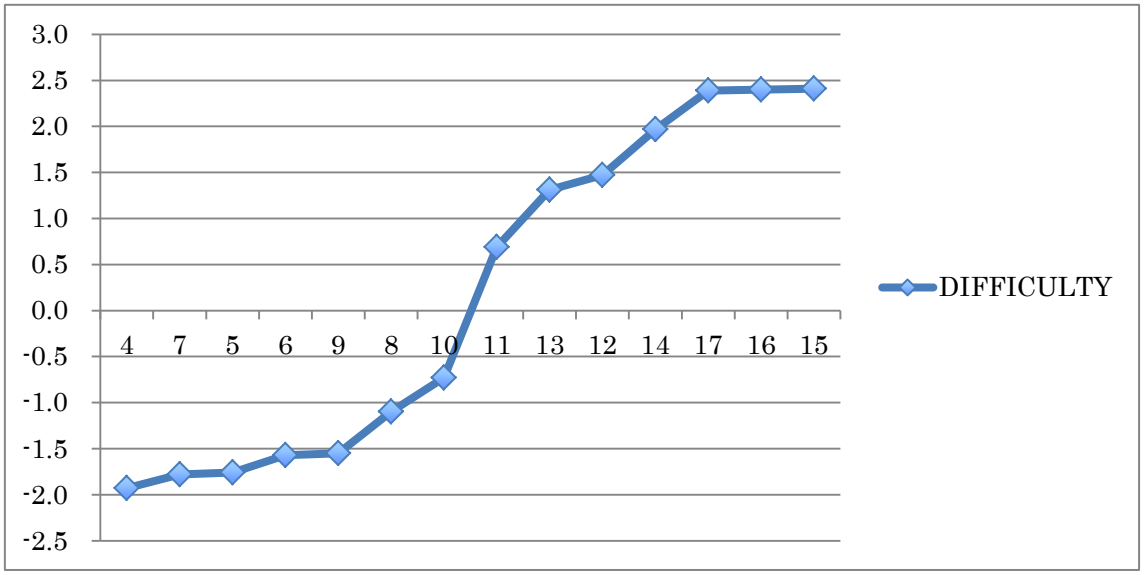


Figure 6.1. Wright & Stone, 2PLM, RP=0.67, DIFFICULTY

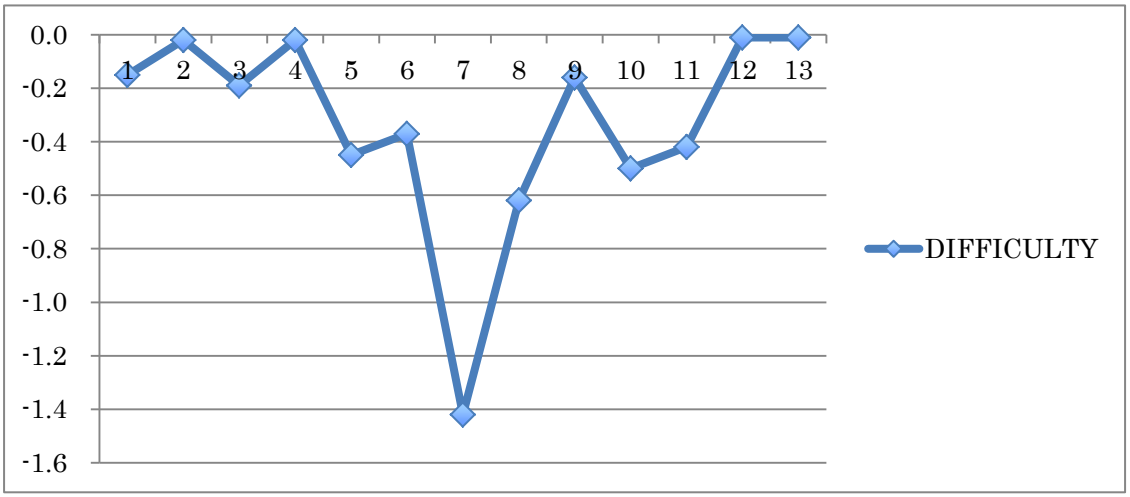


Figure 6.2. Wright & Stone. 2PLM. RP=0.67. DIFFICULTY (C-D)

DISCRIMINATION
GDN (TIN-TIN)

データ		Item	C	D	C-D
	1	(11—12)	0.85	0.85	0
	2	(12—14)	0.85	0.85	0
	3	(14—7)	0.85	0.86	-0.01
	4	(7—10)	0.86	0.87	-0.01
	5	(10—6)	0.87	0.9	-0.03
	6	(6—5)	0.9	0.92	-0.02
	7	(5—4)	0.92	0.93	-0.01
	8	(4—13)	0.93	0.93	0

9	(13—15)	0.93	0.94	-0.01
10	(15—9)	0.94	0.95	-0.01
11	(9—8)	0.95	0.95	0
12	(8—17)	0.95	0.95	0
13	(17—16)	0.95	0.95	0
14	(16--	0.95		

THETA
GDN(TIN-TIN)

データ

	Item	C	D	C-D
1	(4—5)	-1.482	-1.307	-0.175
2	(5—7)	-1.307	-1.296	-0.011
3	(7—9)	-1.296	-1.112	-0.184
4	(9—6)	-1.112	-1.107	-0.005
5	(6—8)	-1.107	-0.662	-0.445
6	(8—10)	-0.662	-0.251	-0.411
7	(10—11)	-0.251	1.18	-1.431
8	(11—13)	1.18	1.758	-0.578
9	(13—12)	1.758	1.96	-0.202
10	(12—14)	1.96	2.46	-0.5
11	(14—17)	2.46	2.828	-0.368
12	(17—16)	2.828	2.838	-0.01
13	(16—15)	2.838	2.853	-0.015
14	(15--	2.853		

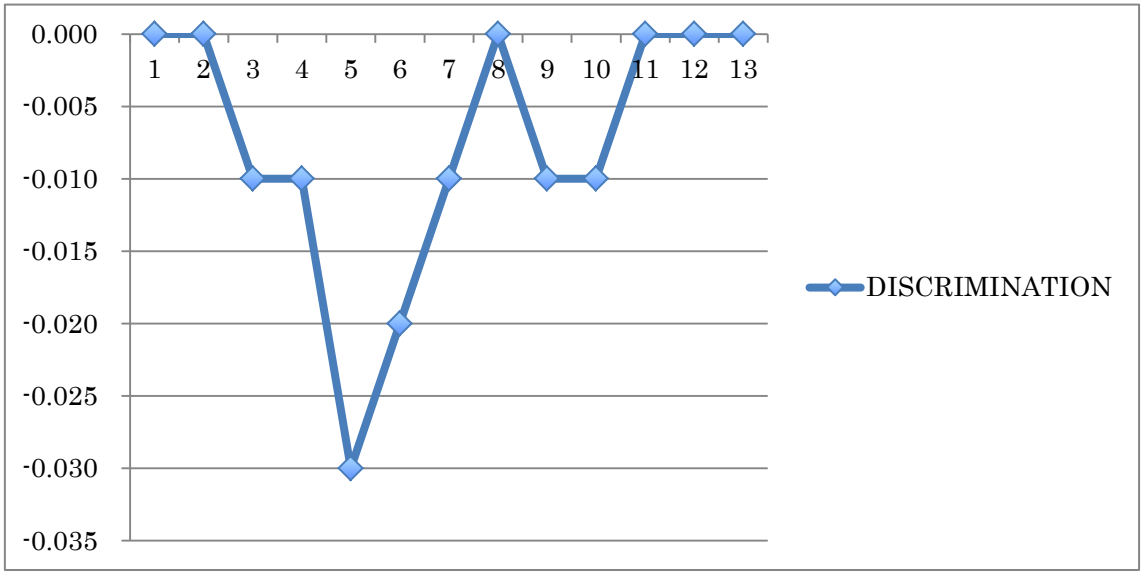


Figure 6.3. Wright & Stone (1979). 2PLM. RP=0.67. DISCRIMINATION (C-D)

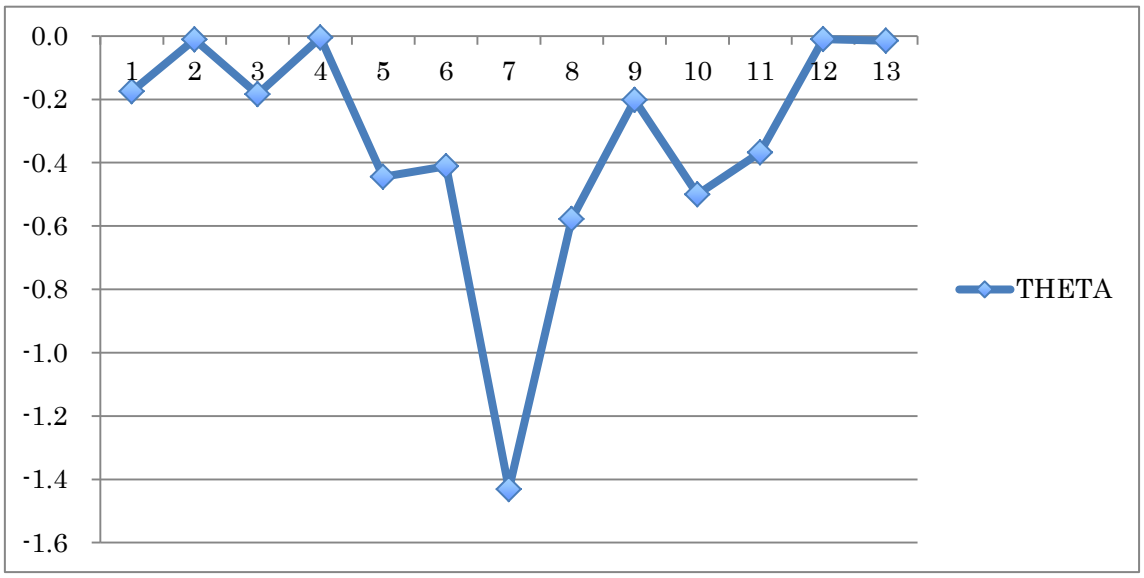


Figure 6.4. Wright & Stone (1979). 2PLM. THETA (C-D)

Figure 6.1.から Figure 6.4.までは、Wright and Stone(1979, p. 31)の Table 2.3.1. の表を用いて求めたデータを元にして作られたものである。元のデータは、person number 35, item number 18 であるが、計算上、delete しなければならないデータがあるために、ここで示されているパラメータの算出結果では、item number は 14 項目になっている。

大きく分けて、関連する資料は 2 種類ある。その 1 つは、グラフ作成のための数値による 5 つのテーブルである。もう一つは、それを使って作成したグラフである。グラフの Figure 6.1. 及び、Figure 6.2. を検討する前に、Table 6. 5. Wright & Stone (1979). 2PLM. RP=0.67, DIFFERENCE に注目する必要がある。

最初のテーブル(Table6.5.)は、page number や item number の数値を利用した bookmark の置き場所を探すための基礎データである。つまり、Ordered Item Booklet 作成に必要なもので、テスト項目困難度順にデータは並べて作成されている。PNO, TIN, DIFFICULTY, DISCRIMINATION, THETA の値が示されている。

つぎのテーブルは、DIFFICULTY の差を利用した分割点の推定に必要なデータである。最初の GDN<1>から<14>は、グラフに示される順番をさしている。つぎの TIN は、重要なデータを求めるための item number が示されている。たとえば、(4—7)というのは、TIN4 の difficulty から 7 の difficulty を引く、という意味である。その具体的な数値は、つぎの C と D に示されていて、その計算結果は、C-D のところに示されている。具体的に示すと、1 番目のデータ：GDN<1>は、TIN4(-1.93)から 7(-1.78)を引いた(-0.15)を<C-D>のところに記入し、その数値をグラフの GDN<1>としていることを意味している。

つぎの DISCRIMINATION, THETA に関するテーブルも、同じ手順で示されている。

つぎは、グラフの見方である。Figure 6.1.においては、グラフの上では、審査員が「著

しい相違」の場所を発見して、booklet を直ちに置くことは、いささか困難である。「著しい相違」をグラフの上には、明確に示されていないからである。

Figure 6.2.では、それに比べると、「著しい相違」の場所を発見するのは、比較的容易である。Figure 6.2.のなかに見える横線の 1-13 までの番号は、Figure 6.1.のなかに見える横線の 1-13 までの番号とは、その内容が異なるので、これを GDN という語の後に< >でかこった番号で示すこととする。ここでは、GDN<6>から、GDN<7>までの変動は、大きいので、それを直感的に発見することは可能である。-0.37 から-1.42 という大きな差があるからである。GDN<6>というのは、TIN10 から 11 という大きな差(-1.42)を示す GDN<7>の直前にあり、「著しい相違」を示す重要な地点であると考えることができる。こうした地点を発見した方法は、前に述べた Cizek, Bunch, and Koon(2004)での差を利用した推定方法とおなじである。ここで見られる GDN<6>を構成している TIN は、8 と 10 である。また、GDN<7>を構成している TIN は、10 と 11 であり、両者を共通に構成しているのは、TIN10 である。したがって、ここに booklet を置くのが最も適切と判断される。依頼する審査員が定める booklet の置き場所も、この TDN10 と推定することができる。

Figure 6.3. DISCRIMINATION(C-D)を検討すると、つぎのようなことが理解できる。ここでは、GDN<4>から<5>までの変動は、大きいので、それを直感的に発見することは可能である。-0.01 から-0.03 というこの表では大きい差があるからである。GDN<4>は、TIN10 から 6 という大きな差(-0.03)を示す GDN<5>の直前にあり、「著しい相違」を示す重要な地点であると考えることができる。こうした地点を発見した方法は、前に述べた Cizek, Bunch, and Koon(2004)での差を利用した推定法と同じである。ここで見られる GDN<4>を構成している TIN は、7 と 10 である。また、GDN<5>を構成している TIN は、10 と 6 であり、両者を共通に構成しているのは TIN10 である。したがって、ここに booklet を置くのが最も適切と判断される。依頼する審査員が定める booklet の置き場所も、この item number 10 と推定することができる。

Figure 6.4.THETA(C-D)を検討すると、つぎのようなことが理解できる。ここでは、GDN<6>から<7>までの変動は、大きいので、それを直感的に発見することは可能である。-0.411 から-1.431 という大きな差があるからである。GDN<6>は、TIN10 から 11 という大きな差(-1.431)を示す GDN<7>の直前にあり、「著しい相違」を示す重要な地点であると考えることができる。こうした地点を発見した方法は、前に述べた Cizek, Bunch, and Koon(2004)での差を利用した推定法と同じである。ここで見られる GDN<6>を構成している TIN は、8 と 10 である。また、GDN<7>を構成している TIN は、10 と 11 であり、両者を共通に構成しているのは、TIN10 である。したがって、ここに booklet を置くのが最も適切と判断される。依頼する審査員が定める booklet も、この TIN10 と推定することができる。

以上、Wright and Stone(1979, p. 31)の Table2.3.1.のデータを用いて、分割点の設定場所を探し求めた結果である。この結果は、DIFFUICULTY, DISCRIMINATION, THETA すべてのデータにおいて、TIN10 が最も適切であろうという判断を下す道を切り開くに至った。

7. まとめ　CITO　Variation を検討するための予備調査結果

この研究課題は、CITO Variation on Bookmark Method を検討するための予備調査に関するものである。この CITO Variation の是非を問う前に、行わなければなければならない事項は、数多く存在する。そうした課題を検討して初めて、Bookmark Method に関する CITO (Centraal Instituut Voor Toetsontwikkeling: National Institute for Educational Measurement, The Netherlands)による提案を検討することができると考えられる。したがって、今回の研究は、その予備調査を行うことであった。

7. 1. 分割点の設定法

これまで、規準設定に関して、どのような意味を持つのであろうか、どのような研究結果が見られるのか、また、その研究に対する様々な議論、さまざまな評価はどんなものがあったのかを検討した。そして，多くの規準設定法の中で、わが国でも使用できる可能性の高い Bookmark Method を取り上げて、その詳細な手順と方法を検討してきたのがこの研究である。

その大枠は、1. 規準設定の意味と必要性、2. 規準設定のための方法、3. 規準設定法にかかわるこれまでの評価、4. Bookmark Method の開発と課題、5. データにおける分割点の推定、6. Wright and Stone(1979)データの検証、である。この方法の中で、もっとも主観的ではないかと批判されてきている分割点の設定方法に関して、とくに、重点を置いて考察してきた。テストデータを用いて、受験者の分割点を設定するために使われた方法の一つは、審査員の判断に基づくものであった。しかし、この方法は主観的と批判された重要な点である。そのために行われた審査員への指示では、「受験者が正解する可能性が、67%以下になるであろうと審査員が信じる最初の頁に booklet を置くのがよい (to place a marker on the first page in their OIB at which, in their opinion, the RP drops below .67)」というものであった。しかし、これには、第 1 に、審査員の主観が介入する可能性が高いことである。それを取り除く方法としては、第 2 に、審査員による審査の最終決定までの審議回数を増やすことであった。審議回数は、3 回程度必要であろうとされてきた。しかし、第 3 には、長すぎる審査時間という問題があった。したがって、複数の審査員の決定にもとづく判断にとって代るべく、第 4 として、データ分析の客観的方法の開発が望まれていたわけである。本研究は、従来の審査員のこうした判断をより客観的にするための方法の一つを開発することであ

った。そのために、具体的に、その方法を提案し、分析したものである。開発を試みたのは、以上述べたように、「PNO/TIN 間の数値差を利用した推定法」である。

7．2．予備調査・実験の結果:　PNO/TIN 間の数値差を利用した推定法

その予備調査・実験は、一つには、Schagen and Bradshaw (2003)のデータに関して行われた。審査員を用いた Bookmark の置き場所である PNO6,5,2 は、この「PNO/TIN 間の数値差を利用した推定法」を利用すれば、明確に推定することができた。さらに、Rasch Measurement の開発のための最も重要な文献のひとつ：Benjamin D. Wright & Mark H. Stone (1979). *BEST TEST DESIGN*. Chicago, MESA Press で使用されたデータを使って、「PNO/TIN 間の数値差を利用した推定法」を、再度、分析し検討してみた。その結果は、調査項目 DIFFICULTY, DISCRIMINATION, THETA いずれの分野においても、きわめて明確な bookmark の置き場所として TIN10 を求めることが可能であった。

「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、次のようになる。

(1)使用したテストの結果を IRT(Item Response Theory)を用いて分析する。(2)RP(response probability)を設定し、Theta@RP を算出し、OIB (ordered item booklet) 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/TIN 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。

7．3．CITO Variation on the Bookmark Method

Bookmark Method は、ヨーロッパにおいても、その注目を浴びていることは、つぎの文からも理解することができる。これは、Frank van der Schoot (2009, p. 2) Cito variation on bookmark method. In Language Policy Division, Strasbourg, *Reference Supplement to the Manual for Relating Language examinations to the CEFR (Common European Framework of Reference for Languages: learning, teaching, assessment)* Council of Europe. からの一節である。

> Section 6.9 of the Manual for Relating Examinations to the Common European Framework of Reference for Languages (CEFR)describes the Cito variation of the bookmark method. This method 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 panelists are fully informed about the level of mastery for all items in the item pool or test at every point of the ability scale. This informs panelists about the relative difficulty of the item in the test or item pool. Furthermore it prevents panelists making inconsistent decisions. Usually, however, panelists are not familiar with the psychometric concepts involved. Therefore, the standard setting method should be introduced carefully.

次年度においては、こうした CITO の提案等も加味して、わが国の英語学習者に対する、もっとも適切な規準設定法は何かをさらに究明することとする。

規準設定の関するわが国の議論は、まさに、始まったばかりである。そこでは、多くの問題を抱えて進まなければならないことが予想される。しかし、Cizek, G.J.(President, National Council on Measurement in Education)が、Cizek and Bunch(2007, p. 320). *Standard Setting: A guide to Establishing and Evaluating Performance Standards on Tests*. Sage Publications で述べているつぎのような発言は、きわめて真剣な研究者の声として決して見逃すわけにはいかない。"According to Segal, 'A man with a watch knows what time it is. A man with two watches is never sure.' Because there is no equivalent of an atomic clock in the field of standard setting, our recommendation is simply for practitioners to invest in a single watch of greatest quality given available resources."

参考文献

AERA, APA & NCME (1999). *Standards for Educational and Psychological Testing* , (p.53). AERA.

Angoff, W.H. (1971). Scales, norms, and equivalent scores. In Thorndike (Ed.). *Educational Measurement (second Ed.)* (pp.508-600), ACE.

Beuk, C.H. (1984). A method for reaching a compromise between absolute and relative standards in examinations. *Journal of Educational Measurement,* 21, 147-152.

Bock, R.D., Mislevey, R., & Woodson, C. (1982). The next stage in educational assessment, *Educational Researcher*, 4-16.

Cizek, G.J (1996). Standard- Setting Guideline, *Educational Measurement: Issues and Practice, Spring, 1996. 14.*

Cizek, G.J. (2006). Standard Setting, In S.M. Downing & T.M. Haladyna (Eds.) *Handbook of Test Development:* (pp.225-258). Lawrence Erlbaum Associates.

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

Cizek, G.J. and Bunch, M.B. (2007). *Standard Setting, A Guide to Establishing and*

Evaluating Performance Standards on Tests, (pp.65-217). Sage Publications.

Cizek, G.J., & Bunch, M.B. (2007). The Bookmark Method, *Standard Setting, A Guide to Establishing and Evaluating Performance Standards on Tests*, (pp.155-192). Sage.

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

Council of Europe (January, 2009). 6.9. *Cito* Variation on the Bookmark Method: *Relating Language Examinations to the CEFR: A Manual*. (pp.82-83). LPD, Strasbourg.

Council of Europe (October, 2009). *Section 1: Cito variation on the bookmark method*, *Reference Supplement to the Manual for Relating Language Examinations to CEFR*, (pp.1-17). LPD, Strasbourg.

Downing, S.M. & Haladyna, T.M. (Eds.) (2006). *Handbook of Test Development,* Lawrence Erlbaum Associates, Publishers.

Ebel, R.L. (1972). *Essentials of Educational Measurement*. Printice-Hall.

Fulcher, G. (2010). *Practical Language Testing.* Hodder Education.

Frank van der Schoot (2009:2).Cito variation on bookmark method. In Council of Europe, *Reference Supplement of the Manual for Relating Language examinations to the CEFR, Language Policy Division.*

Hambleton, R.M. and Pitoniak, M.J. (2006). Setting Performance Standards. In Brennan, R.L. (ed.) (2006). *Educational Measurement (Fourth Edition),* (pp.442-470). ACE.

Hambleton, R.M. & Plake, B.S. (1995).Using an extended Angoff procedure to set standards on complex performance assessments. *Applied Measurement in Education,* 8, 41-56.

Hofstee, W.K.B. (1983). The case for compromise in educational selection and grading. In S.B.Anderson & J.S. Helmick (Eds.) *On Educational Testing.* (pp.109-127). Jossey-Bass

Jaeger , R.M. and Mills, C.N. (2001). An Integrated Judgment Procedure for Setting Standards on Complex, Large-Scale Assessments. In Cizek, G.T. (ed.) *Setting Performance Standards,* (pp.313-338). Lawrence Erlbaum Associates, Publishers.

Jaeger, R.M. (1989). Certification of Student Competence. In Linn,R.L.(Ed.)(1989). *Educational Measurement (Third Edition),* ACE.

Kaftandjieva, F. (2004:31). Section B: Standard Setting, Reference Supplement to the Preliminary Pilot Version of the Manual for Relating Language Examinations to the CEFR, Language Policy Division, Strasbourg. Council of Europe.

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

Lewis, D.M., Mitzel, H.C. & Green, D.R. (1996, June). 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 Assessment, Phonix, AZ.

Livingston, S.A., and Zieky, M.J. (1982). *Passing Scores: A manual for setting standards of Performance on educational and occupational tests,* ETS.

Mitzel, H.C., Lewis, D.M., Patz, R.J. & Green, D.R. (2001). The Bookmark Procedure: Psychological Perspectives. In Cizek, G.J. (ed.)(2001). *Setting Performance Standards: Concepts, Methods, and Perspectives.* (pp. 249-282), Lawrence Erlbaum Associates, Publishers.

Nedelsky, I. (1954). Absolute grading standards for objective tests. *Educational and Psychological Measurement.* 14. 3-19.

Nichols, P., Twing, J., Mueller, C.D., and O'Malley, K. (2010). Standard-Setting Methods as Measurement Processes, *Educational Measurement: Issues and Practice,* 29(1), 14-24.

Peterson, C.H., Schulz, E.M., Engelhard Jr., G. (2011). Reliability and Validity of Bookmark-Based Methods for Standard Setting: Comparisons to Angoff-Based Methods in the National Assessment of Educational Progress, *Educational Measurement: Issues and Practice,* 30(2), 3-14.

Pitoniak, M.J. (2003). Standard setting methods for complex licensure examinations. Unpublished doctorial dissertation, University of Massachusetts, Amherst.

Schagen, I. and Bradshaw, J. (2003 September). *Modeling item difficulty for Bookmark standard setting.* Paper presented at the annual meeting of the British Educational Research Association, Edinburgh.

Sireci, S.G., Hambleton, R.K., & Pitoniak, M.J. (2004). Setting passing scores on Licensure exams using direct consensus. *CLEAR Exam Review,* 15 (1), 21-25.

Wang, N. (2003). Use of the Rasch IRT Model in Standard Setting: An Item-Mapping Method. *Journal of Educational Measurement* 40(3), 231-253.

Wright, B.D. & Stone, M.H. (1979). *BEST TEST DESIGN,* MESA Press.

Zieky, M.J. (2001). So Much Has Changed: How the Setting of Cutscores has Evolved Since the 1980s. In Cizek, G.J.(ed.)(2001). *Setting Performance Standards: Concepts, Methods, and Perspectives* (pp. 19-52). Lawrence Erlbaum Associates, Publishers.

Zieky, M.J. & Livingston, S.A. (1977). *Manual for setting standards on the basic skills assessment tests*, Educational Testing Service.

Zieky, M.J., Perie, M. and Livingston, S.A. (2008). *Cutscore: A Manual for Setting Standards of Performance on Educational and Occupational Tests,* Educational Testing Service.

東 洋・梅本堯夫・芝 祐順・梶田叡一(編)(1988).『現代教育評価事典』,金子書房.

池田 央・藤田恵爾・柳井晴夫・繁桝算男(編訳)(1992).『教育測定 第 3 版』,みくに出版.

池田 央(監訳)(2008).『テスト作成ハンドブック』,(p.12)(Downing and Haladyna (Eds.)(2006). *Handbook of Test Development,* Lawence Erlbaum Associates, Publishers)教育測定研究所.

大友賢二(監修).中村洋一・小泉利恵(編)(2009).『言語テスト:目標の到達と未到達』ELPA.

北尾倫彦(監修)(2012)『平成 24 年度版観点別学習状況の評価規準と判定基準：中学校外国語』,図書文化.

文部科学省・国立教育研究所(2012).『評価規準の作成・評価方法等の工夫改善のための参考資料』 (高等学校外国語),教育出版.

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』第一法規出 (Bloom, Hastings, and Madaus (Eds.)(1971). *Handbook of Formative and Summative Evaluation of Student Learning,* McGraw-Hill, Inc.)

橋本重治(1983).『続・到達度評価の研究―到達基準設定の方法』,日本図書文化協会.

梶田叡一(2005).『教育評価(第 2 版補訂版)』,有斐閣.

皆見英代(2008).「規準」と「基準」、'criterion' と'standard' の区別と英和照合―教育評価の専門用語和訳に戸惑う、『国立教育政策研究所紀要 137』.

井上俊哉(訳)(1992). 「学生のコンピテンスの証明」,(原文 Richard M. Jaeger, Certification of Student Competence, In Linn, R.L. (Ed.), (1989). *Educational Measurement: Third Edition,* NCME, ACE)原著第 3 版下巻,日本語版編集委員,池田央他 『教育測定学(下巻).S.L.学習研究所会.

"Can-do statements" の比較・研究
Comparative studies on practices of Can-do statements

伊東祐郎
Sukero Ito

Abstract

To help language learners understand the dimensions of each level of proficiency, there are "can-do" statements (CDS). CDS are generally positive: they describe what a learner is able to do each level. Therefore, CDS help learners understand the types of tasks they must accomplish to be proficient at the various levels. However, some CDSs describe what a learner cannot do or does wrong at the lower levels. This does not help learners, even those at the lowest levels, see that learning has value and that they can attain language goals.

This paper reviews CDS of ALTE (the Association of Language Testers in Europe) , CEFR (The Common European Framework of Reference for Languages), ACTFL (American Council on the Teaching of Foreign Languages), and CLB (The Canadian Language Benchmarks). as well as Can-do lists of EIKEN and Can-do guide of TOEIC.

CDS endorse language use in all phases from beginning level through to advanced. They reflect performance descriptors for all levels and are mapped against the scales. Within each stage or level the descriptors are progressive but may address different aspects of each skill.

This paper examines the descriptions of each level of proficiency provided by those CDSs above, and tries to analyze the structures and functions taken into the CDS. Differences in performance of tasks in each level were investigated. Central to the study was the use of a taxonomy based on Bloom's Taxonomy for characterizing performance tasks which were described in CDS.

1. 問題と目的

“Can-do statements”(以下「CDS」と称す)は、コミュニケーション活動にかかわる能力が言語化されたものである。言語能力の構成概念を外的な社会的機能に焦点を当てて、現実的でより観察可能なものとして捉えようとしたものである。最近、外国語教育の分野で、「スタンダード」「ガイドライン」「フレームワーク」「ベンチマーク」(以下「標準」と称す)という言葉を頻繁に耳にするが、それらには共通して、CDS が盛り込まれている。社会学的な観点から新たなコミュニケーション能力のモデルを提示し、教育の方法や評価のあり方への枠組みに新たな解釈の基礎を提供しようとしていることがうかがわれる。

一方、外国語教育における大規模テストでは、テスト結果から得られる得点を具体的な能力の解釈として活用できるよう、得点に対する意義付けをこれまで以上に重要視するようになってきた。正答数を合計して算出した得点を提示するだけでは、学習の成果としての弱点と優れている点がわかりにくい。また、点数という数字による情報やその管理のみで終始してしまい目標や目標基準に対する達成度が具体的な形でフィードバックされにくい。教師ならびに学習者双方にとって、数値以上の有益な情報は得られないことになる。そこで、得点に対して、意味ある解釈ができるよう、尺度を設け、それぞれの尺度に対応した知識や能力の特徴を記述した言語能力記述文(「CDS」の邦訳)が提示されるようになってきた。

しかしながら、CDS には様々な記述の仕方が存在し、構造自体が明確に把握されているわけではない。また、CDS の活用方法についても、開発の意図や趣旨とは無関係に第三者に導入されたり活用されたりすることも少なくなく、妥当性の検証や活用の望ましい方法についても検討する必要がある。そこで本稿では、2011年度の報告でまとめた残された課題、以下の(1)から(3)をテーマに、既存の CDS を取り上げて構造分析を試みる。あわせて、一部の CDS の比較検討を行い、CDS の活用の可能性を探ることを目的とする。

(1) 規準設定(スケール化)の目的と規準の活用実態の検討

(2) 規準設定にかかわる背景理論の研究

(3) 第2言語としての言語習得と外国語としての言語習得における相違点の検討

なお、分析の手順としては、最初に標準における CDS を、続いて大規模テストにおける CDS の分析を試みる。最後に、CDS の構造等について総括的分析と考察を行う。

2. 1. ALTE(The Association of Language Testers in Europe)

ヨーロッパでは、複言語主義の名の下に教育の統合化が行われている。外国語教育においても、欧州各国の言語テストの能力レベルを相互に比較可能にすることが求め

られ、テストが測定した能力を同一尺度で判定する必要性があった。そのため、ALTE では CDS を作成し、目標言語を使って具体的に何ができるかを明文化した。結果として、他言語の能力と比較ができるようになっている。文法項目の異なる諸言語の能力比較には、パフォーマンスを基準にした CDS が開発され、評価や問題作成の際の基準枠として活用されている。

2．2．CEFR(Common European Framework for References)

ヨーロッパでは各国の統合と相まって、教育の標準化や言語政策の推進をかかげて、10年以上もかけて外国語教育の理念が議論され、その結果、CEFR が誕生した。CEFR は欧州評議会がいわゆる言語教育の統一化を実現するために作成したものである。この点において ALTE の CDS 開発の趣旨とは異なる。CEFR では、言語教育という視点から包括的な開発を試み、シラバスやカリキュラムの策定、評価の方法の参考になるよう広範囲に及ぶ内容が盛り込まれていることが、特徴として挙げられる。

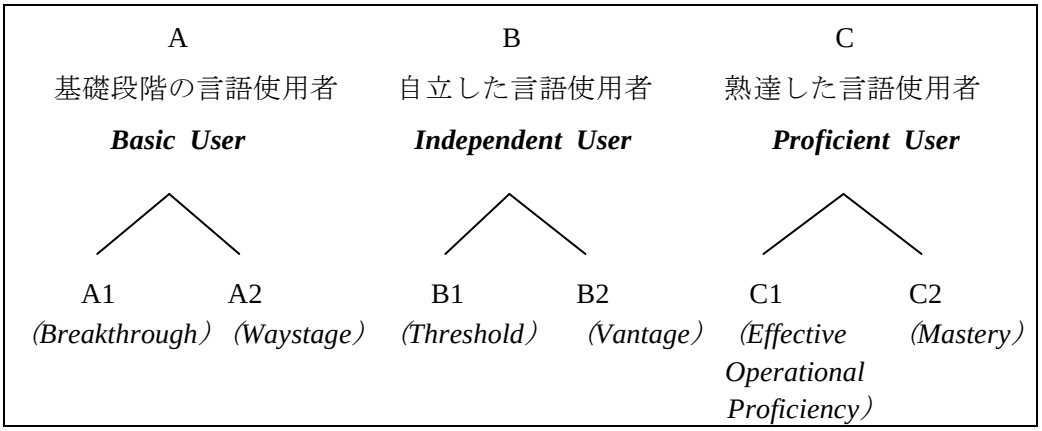


図 1　CEFR と ALTE の能力レベル

ALTE の能力レベルは全部で6段階で構成されている。この6段階というのは欧州評議会の CEFR も6レベルから構成されているので、欧州の能力記述というのは6レベルで統一されていると言えよう。そして、レベルは大きく ABC という3段階に分けらている。さらに2レベルに分けられていて、A1、A2と番号を附して段階差を明示している。番号が大きいと能力が高くなる。したがって、A2の次は B1、B2、そして C1、C2 と能力レベルが明示される。ALTE の Breakthrough Level というのは A1に相当するので、一番下のレベルとなる。

2．3．ACTFL(American Council on the Teaching of Foreign Languages)

ACTFL の Proficiency Guidelines の特徴は、学校におけるアカデミックな外国語科目

の中での外国語能力を明確化したものとなっている。Proficiency Guidelines は、1950年代に開発された Interagency Language Roundtable(ILR)が基になっている。開発の主な目的は、米国における外国語教育の理念の明確化と教育内容の充実であった。外国語学習の意義、教育内容の構造化、言語能力観の明示化、外国語教育を通しての人間育成等が盛り込まれている。

ACTFL は、主レベルとして次の5つを設定している。 Distinguished、Superior、Advanced、Intermediate、Novice で、後者3レベルについては、さらに High、Mid、Low に階層化している。Distinguished レベルは、最近設定されたレベルで、『ACTFL-OPI 入門』には記述がなかったので、空欄にしてある。なお、CDS に相当するところは、「判定尺度」としてタスクのレベルと達成度を基にした記述になっている。口頭表現力のレベルを評価するための目安として活用している。

表 1　ACTFL の判定尺度

Distinguished		
Superior		意見の裏付けができる。仮説が立てられる。 具体的な話題も抽象的な話題も論議できる。 言語的にも不慣れな状況に対応できる。
Advanced	- High	主な時制／アスペクトを使って叙述、描写できる。 複雑な状況に対応できる。
	- Mid	
	- Low	
Intermediate	- High	自分なりに言語が使える。よく知っている話題について簡単な質問をしたり答えたりできる。 単純な状況や、やりとりに対処できる。
	- Mid	
	- Low	
Novice	- High	コミュニケーションができるのは、決まり文句、暗記した語句、単語の羅列、簡単な熟語でのみ。
	- Mid	
	- Low	

(記述は『ACTFL-OPI 入門』から転載)

2．4．CLB(Canadian language benchmark)

CLB は、移民に対する言語教育を背景に CDS が開発されている。他国とは異なる視点から能力基準を作っている。カナダでは、現在不足している労働力は看護士であると言われている。多くの看護士を南米やフィリピンから受け入れて、カナダ国内の医療や福祉の現場の充実を図ろうとしている。このような事情から、カナダはどちらか

と言えば雇用者が移民を雇用する際に、言語能力を判定・評価する際の評価基準として CDS が活用されている。CLB は、以下のように3ステージ(StageⅠ、StageⅡ、StageⅢ)で、各ステージがさらに4レベルに分けられ、全体としては12レベルで構成されている。

表 2　CLB の能力レベル

	StageⅠ				StageⅡ				StageⅢ			
	1	2	3	4	5	6	7	8	9	10	11	12
Speaking												
Listening												
Writing												
Reading												

3. 各標準における CDS の共通点

3．1．基盤はコミュニケーション能力

先行事例を概観したが、開発の経緯にはそれぞれ固有の背景や理由が存在し、目指すべき目標や掲げられる理念は異なっている。共通点をあげるならば、外国語あるいは第二言語教育における CDS は、全てコミュニケーション能力の記述とスケール化が目指されている点である。これまでの日本国内の外国語教育は、どちらかと言えば、文法事項や語彙の配列や分類など、言語的要素にかかわる内容の記述が中心になる傾向があった。この点において、各標準で明示されている CDS は、日本の外国語教育に一石を投じる形になっている。

3．2．指導項目の配列ではない

2つ目の共通点として、コミュニケーション能力の漸増性に注目し、段階的に記述されている点である。コミュニケーション能力というのは連続性を持ったものである。それゆえに、それらを明示するためには、段階を示しながらも包括的な記述になっている。特定のカリキュラムやプログラムのための CDS ではないために、ACTFL をはじめ CEFR でも文法や文型、語彙などの教授細目の配列ではないことを断っている。またそれぞれの外国語教育においては、様々なアプローチ、すなわち教授法がとられているところから、ある特定の教授法や、指導法の例示ではないということも明示されている。

3．3．焦点は outcome

3つ目として、コミュニケーション能力に関わるパフォーマンスの課題に注目している点である。あくまでも言語能力の outcome、要するに言語を使って何ができるか、学習のプロセスよりも、最終的な outcome に焦点を当てた領域を記述している点である。初級レベルから徐々に中級、上級と記述内容が拡大するが、漸増性、連続性を反映させながら、レベルごとに、パフォーマンスの課題に注目した記述がなされている。記述されている課題については広範囲に及ぶ。したがって、課題の設定、言語使用領域の特定、課題の難易度の調整など、CDS を記述する作業を行うに当たっては、検討すべき事項は少なくない。

3．4．学習者中心

共通点の4つ目として、実際のコミュニケーション能力に焦点を当てた、学習者中心の記述、別の言葉で表現するならば、現実の言語運用場面中心ということが挙げられる。CLB が、教師が事前に頭の中で考えたコミュニケーション能力やそれを背景として場面ではなく、現実の生活場面で実際に用いられる言葉の機能と概念に基づいているという点である。実際にどのような場面で、どのような目的のために言葉が話され、学ばれているかという学習者中心の視点は見逃せない。

4．CDS の構造

4．1．言語運用場面

ALTE や CEFR 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、social、work、study の3つの領域で言語運用場面を規定している。言語使用領域は無限大であり、このような規定を設けないと能力基準の枠作りで苦心することになる。筆者は以前に大規模試験としての日本語能力試験の CDS を策定した経験があるが、一般的な日本語力を測定する試験を目指すと言う議論の中で、対象範囲が広がりすぎてしまい、CDS の作成を困難に感じた経験を持つ。その点、ビジネス日本語能力試験は場面も機能もビジネス場面に限られているので、CDS は作りやすくなる。また、第3者にとってもわかりやすく、内容などに関する助言などもしやすくなる。

4．2．言語運用領域

言語運用領域と言えば、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。CEFR では、「話す」を「Spoken Interaction(会話／対話)」と「Spoken Production(独話)」と分けて5領域で構成している。コミュニケーション能力を技能別に

記述することになると、言語能力の連続性を反映しつつ、低いレベルから上位レベルに記述内容を高度化していかなければならない。能力の各レベルにおけるコミュニケーション能力の特定が複雑な作業になる。能力レベルの根拠や拠り所をどのように確定するかという最も重要な課題に直面することになる。

日本語で「読む」「書く」の CDS や CEFR を参照しながら作成する場合、直面する課題は、漢字の知識や運用にかかわる記述をどのようにするかである。アルファベットを書き言葉として共有している言語では、CEFR を活用することに問題は生じないかもしれないが、日本語や中国語など漢字や他の文字を使う言語では、漢字の認識力や再生力について異なる参照枠を設けて CDS を記述する必要が出てくる。

5. CDS の記述

能力発達段階の観点の分類に関しては、和田(2004)が、CEFR の CDS がどのような観点から表記されているかを詳細に調査・分析している。分類の観点として、ポイントを2つ挙げている。一つは言語形式から記述している点であり、もう一つは、内容から記述されている点である。

5．1．言語の形式面からの記述

言語の形式に焦点を当てると、正確さ、流暢さ、繰り返しとかポーズ、即興性、長さ、速さが、報告書にまとめられている。その他としては、複雑さ、多様さ、明確さ、このような観点から CEFR の能力の記述がなされていることが分析されている。

5．2．言語の内容面からの記述

内容面とは、一つには場面、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で、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。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。

5. 3. CEFR における CDS の記述内容の分析

5. 3. 1. 聞く

聞くListening (A1)	
はっきりと話してもらう	→明確さ
ゆっくり話してもらう	→流暢さ
自分、家族や身の回りのことについて	→話題
聞き慣れた語句や基本的な表現語句ならば	→複雑さ
理解できる	

この報告書で示されている具体的な分析方法を紹介する。リスニングの A1レベルは「はっきりと話してもらえれば理解できる」とか、「ゆっくり話してもらえれば理解できる」「自分、家族の身の回りのことについてであれば理解できる」「聞きなれた語句や、基本的な表現語句ならば理解できる」このような形で能力記述がなされている。ここでの分析は、「はっきりと話してもらう」は、「明確さ」に関することとして、「ゆっくり話してもらう」はスピードという点で「流暢さ」の視点からの記述であると分析している。そして話題については、初級レベルの学習者にとっては自分自身のことや家族のこなど具体的なテーマで明示されている。語彙については話題に関連して、聞き慣れた基本的という表現で初期段階の語彙レベルを記述している。

5. 3. 2. 話す

会話／対話Spoken Interaction (B2)	
流暢に自然に会話ができ	→流暢さ
母語話者と普通のやりとりができ	→複雑さ
身近なコンテクストの議論に	→話題
積極的に参加し、	→知識
自分の意見を説明し、弁明できる	→機能

次の spoken interaction、話す能力であるが、B2レベルと言うこともあり、A レベルに比べ、やや包括的な書き方がなされていると述べている。例えば、「流暢に自然に会話ができ、母語話者と普通のやり取りができ、身近なコンテクストの議論に積極的に参加し、自分の意見を説明し、弁明できる」の記述には、実場面のイメージが想定しにくく、具体性に欠ける書き方であると分析している。視点としては、流暢さ、複

雑さ、話題、知識、機能が明示されていた。

5．3．3．書く

ライティングWriting (C1)	
適当な長さで	→流暢さ
いくつかの視点を示して	→明確さ
明瞭な構成で	→正確さ
自己表現ができる	→機能
自分が重要だと思う点を	→話題
強調しながら	→機能
手紙やエッセイ、レポートで	→媒体
複雑な主題を扱うことができる	→複雑さ

ライティングについては、C1レベルが取り上げられていた。具体的には、「適当な長さでいくつかの視点を示して、明瞭な構成で自己表現ができる、自分が重要だと思う点を強調しながら手紙やエッセイ、レポートで、複雑な主題を扱うことができる」という記述である。C1は上位レベルであるためか抽象性が増していることがわかる。分析の観点は、流暢さ、明確さ、正確さ、機能、話題、媒体、複雑さ、が挙げられていた。

5．4．CDS の技能別特徴の考察

5．4．1．聞く

リスニングに関しては、やはり聴解内容の話題のなじみ度が主な記述の中心になっていることが報告されている。そしてテキストの速さや長さ、流暢さが記述の端々に出ていることも述べられている。

5．4．2．読む・書く

リーディングとライティング、これは、読むと書くにかんすることである。前者の場合は読解素材、媒体、すなわち何を読むかということが難易度と非常に関係があることがわかっている。初級レベルだと「メモが書ける」とか、「メモが読める」というようなレベルである。ところが、C2レベルだと、「自分の感情を相手に感動的に与えるエッセイが書ける」となっていて、書き手の内在化された感情や思いに踏み込んだ記述になっている。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。作文媒体が言語形式や内容と密接に関係していることがわかる。

5. 4. 3. 話す

次に、spoken interaction と spoken production であるが、前者は会話及び対話として、後者は独話に相当するもの、スピーチや講演など一方的に話すものと捉えることができる。

spoken interaction の場合には話題のなじみ度から記述がなされている。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。話をしていて話がうまくいかないと、自分の母語で言い換えたり、易しい言葉を使ったりして、会話を円滑に進めるためにストラテジーをとることになるが、CEFR の CDS ではこのようなことまで言及していることが特徴として挙げられる。

spoken production、これについては話題が決め手となるようである。何を話すかによって影響を受ける。一方的に話す場合、話題の易しさ、難しさというのが能力の高低にも連動することになる。そして発話の長さに加え、語彙の複雑さ、特に日本語の場合は、和語を使うのか漢語を使うのか、例えば「あそこに新しいビルが建てられています／建設されています」。「建てられています」は初級で勉強するが、「建設」となると漢字熟語のために高度になる。「新しい店が開かれました」という場合と、「新しい店が開店しました」という場合とでは「開店」を使うと、同様にレベルが高いと評価される傾向がある。やはり語彙の複雑さや、漢字系抽象語彙が使えるかどうかなど和語使用との関係からも能力記述をしていく上で考慮すべき視点である。その他、言語の機能、例えば依頼する行為と、断るという行為は心情的にも複雑な状況がある。留学生に「今日これからコーヒーでも飲みに行きませんか？」「行きません」なんて断られるとショックであるが、丁寧に「いやあ、ちょっと」とか社会言語学的なことも含めた返答が返ってくると、その後の人間関係もうまくいくことがある。そういう意味で言語の機能というのは、人間関係維持機能も併せ持っており、このような部分が spoken の場合には大切になる。

6. CDS と大規模テスト

冒頭でも述べたが、国内外で実施されている大規模テストは、能力レベルとそれに対応したテストの結果について、「具体的に何ができるのか」「どのようなレベルであるのか」という得点以外の情報が得られるように CDS が活用されている。本稿では、英検と TOEIC の例を取り上げて、CDS の記述と活用の実態についてまとめてみる。

6. 1. 英検 Can-do リスト

英検では、このリストを作成するために、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いる。「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、リスト化したものを「Can-do リスト」としてまとめている。

表3　英検「話す」Can-Do リスト

	話す
1級	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。
準1級	社会性の高い話題について、説明したり、自分の意見を述べたりすることができる。
2級	日常生活での出来事について説明したり、用件を伝えたりすることができる。
準2級	日常生活で簡単な用を足したり、興味・関心のあることについて自分の考えを述べることができる。
3級	身近なことについて簡単なやりとりをしたり、自分のことについて述べることができる。
4級	簡単な文を使って話したり、質問をすることができる。
5級	初歩的な語句や定型表現を使うことができる。

表4　英検「話す」2級 Can-Do リスト

	日常生活での出来事について説明したり、用件を伝えることができる。
話す	・日常生活の身近な状況を説明することができる。(遅刻や欠席の理由など) ・印象に残った出来事について、話すことができる。(旅行、イベントなど) ・自分の学校(会社)について、簡単な説明をすることができる。(場所、人数、特徴など) ・簡単な道案内をすることができる。(例：Go straight and turn left at the next corner.) ・買い物で店員に欲しいものや好みを伝えたり、簡単な質問をすることができる。(色、サイズ、値段など) ・簡単な伝言をすることができる。 (例：Tell Jane to call me back./Tell John I can't go to the meeting today.)

英検の Can-Do リストは WEB 上で公開されていて、その一部を以下に紹介する。1級から5級までの各段階における言語運用力が具体的に記述されている。アンケート結果の分析において自信の高いものを精選したとしており、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。英検では、このリストに「英検合格者の実際の英語使用に対する自信の度合い」というサブタイトルをつけ

て公開している。

6. 2. TOEIC Can-do Guide

TOEIC の能力記述は、Can-do Guide として能力レベルの記述を公開している。以下の表は、スコアとそれに対応した能力レベルを示している。この表を見る限り、受験者の結果は相対的な位置づけしかわからないが、後続する「レベル別評価」表によって具体的な能力がわかるようになっている。

表5　TOEIC スコアと能力レベルの対応

スピーキングスコア	Proficiency Level(能力レベル)
190－200	8
160－180	7
130－150	6
110－120	5
80－100	4
60－70	3
40－50	2
0－30	1

表6　TOEIC 能力レベル別評価

能力レベル	スピーキング
［8］ スコア 190～200	一般的に、レベル8に該当する受験者は、一般の職場にふさわしい継続的な会話ができる。意見を述べたり、複雑な要求に応えたりする際の話の内容は大変わかりやすい。基本的な文法も複雑な文法もうまく使いこなし、正確で的確な語彙・語句を使用している。また、質問に回答し、基本的な情報を提供することができる。発音、イントネーション、強調すべき部分がいつも大変わかりやすい。

能力レベル	スピーキング
［2］ スコア 40～50	一般的に、レベル2に該当する受験者は、意見を述べることも、意見の裏付けを述べることもできない。複雑な要求に応えることもできない、また、まったく的外れな応答をする。 質問に答える、基本的な情報を提供するなど、社会生活や職業上の日常的な会話も理解しにくい。 書かれたものを読み上げる際の英語は理解しにくいことがある。

7. 英検と TOEIC における CDS 比較

英検の CDS は、実際の言語場面を提示して、どのような課題を達成できるかを単刀直入に記述している。すべての記述文は「～できる」で統一されている。一方、TOEIC の CDS は、レベル8であれば上位級であるために、記述文の多くが「～できる」となっているが、下位級であるレベル2では、「～できない」「～(できない)ことがある」「～にくい」など、受験者の能力の限界についての記述が多くなっている。受験者がコミュニケーションにおいて挫折を起こしている場面や状況、また言語的限界を具体

的に明記することによって、レベルを特定しようとしている。ただ一方で、レベル2でできることが何であるのかの記述がほとんどなく、必ずしもCDSが全レベルにおいて能力記述として統一されているわけではないことがわかる。

このように既存のCDSの分析から、その記述の仕方と活用の方法に2つの側面のあることに気づかされる。そのひとつは、目標規準(criterion)としてのCDSで、もうひとつは、評価指標(descriptor)としてのそれである。前者の場合は、教育や測定対象の目標として明示するものである。到達目標として期待されるべき事項として記述するものである。したがって、条件付き記述、例えば、「ルビが振ってあれば、新聞が読める」「繰り返し話してもらえれば、わかる」のような記述は存在しない。「新聞が読める」「会社説明会での説明がわかる」が目標となるのである。英検「話す」1級から5級のCan-Doリストは、各級の目標、ゴールであると解釈できよう。

一方、評価指標としてのCDSの例としては、TOEIC能力レベル別評価の記述などに見られるもので、試験結果の能力レベルの記述としてまとめたものである。評価尺度の各レベルに対応した言語能力を具体的な内容で記述したものである。TOEICの評価尺度では、レベルとスコアとともに、評価のガイドラインとして言語能力が発達段階を踏まえて記述してある。例えば、先述のレベル2のような「限界」を記述したものや、条件付き記述「ゆっくり話してもらうか、繰り返しや言い換えをしてもらえば、○○できる」「限定された範囲内では、○○できる」「Non-Nativeとして十分なコミュニケーションができる」などの記述である。テスト課題、ここでいう目標に対する習熟度の度合いを、条件や状況を附して記述したものになっている。

8. CDS作成の際の課題

8. 1. 認知的負担度と言語能力の難易度

次に、言語技能をどのような段階で記述していくかについて述べたい。言語活動における認知的負担度と言語形式や語彙は関係があると言われている。とすれば、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るが、タスクそのものが思考力を求めるもので、認知的負担が高まると考えられよう。

8. 2. 認知的負担度とブルームのタキソノミー(Bloom's Taxonomy)

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。右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、CDS を記述する上で、参考になるのではないかと思われる。

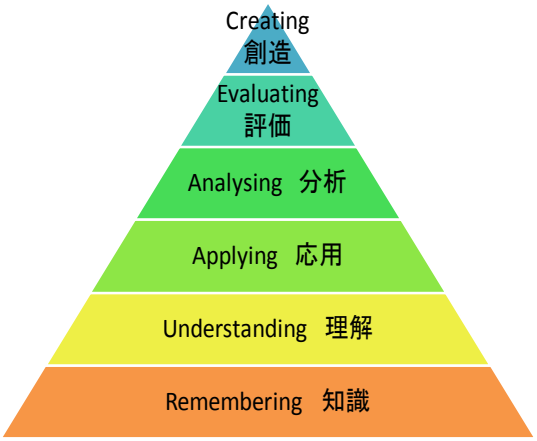


図2　Bloom's Taxonomy
(改訂版)(筆者作成図)

そこで、ここでは、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)を仮説的に対応させ、CDS と認知的負担度の関係を概観してみたい。もちろん6つの認知活動が ALTE の6レベルと対応するかどうかについては別途検証する必要があることを断っておく。なお、認知的活動に下線を引いてあるが筆者によるものである。

【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
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。

（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）

「ありきたり」「基本的」「事実に基づく」「時間」「日時」などの評点によって、認知的負担度の少なさが読み取れる。

【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
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。

「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。

【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
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。

「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。

【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
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。

「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。

【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
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。

「自分の仕事の範囲内」「抽象的」「かなりの流暢さ」「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。

【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
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。

「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。

上記、ブルームのタキソノミーの概略を示したが、CDS の難易度が認知力と関係するならば、CDS を分析したり記述したりする場合は、どのような要素、例えば、言語活動のどのような行為から記述すべきかを検討する上で重要な視点になる。あわせて、CDS の難易度やレベル判定をするための根拠を示すことによって妥当性の検証につなげたい。

9. CDS における包括性と個別性

能力記述をしていく段階で、課題は無視できない。CEFR の記述を見ればわかることであるが、包括的に書こうとすれば書くほど、抽象的記述になってしまうことである。例えば、「高度な論文が読める」となると、高度な論文とはどんな論文なのか具体性を欠くことになる。一方、詳細に記述しようとすると、個別性、多様性が求められることになる。それは具体的な記述で非常にわかりやすくなるのであるが、個別性

が強すぎて、包括的に解釈することがむずかしくなる。結果的に、能力を記述する段階では包括的に書くのか個別的に書くのかという点で、ジレンマに陥ることになる。

能力の発達を段階的に記述するには個別性も必要になる。初級レベルだと「メモが書ける」とするか、「平仮名が書ける」とするかである。平仮名が書けると言う行為に対象物を加え、具体的に「メモが書ける」とか、「板書のコピーができる」というような書き方が「現実的」「リアル世界」ということになる。人の言語活動という点からするとわかりやすい。しかし、上級レベルで、「板書が書ける」でよいのかということがある。「板書が要約できる」とか、あるいは「内容をまとめてレポートが書ける」のような記述になってくる。したがって、個別性というのは、場面、あるいは機能と要素が必然的に加わることになると言えよう。

10. 開始レベルと最終目標の設定

能力の記述のスタートと最終の目標をどのように設定するかも大切になる。CEFRの場合、一応はゼロスタートの部分もあるが、そうではない部分もある。一番低いA1とA2に該当する能力がない領域もある。能力記述の始点、スタートをどこにするかということも、考えなければならない。

例えば日本の大学に入学してくる学生の場合、英語については中学校、高等学校で基本的なことは学んでいることを前提としている。したがって、「アルファベットが読める・書ける」や「挨拶が言える」は不要になる。では、大学の英語教育の始点をどのレベルが始点になるのだろうか。と同時に、最終の目標をどこにおくかということも重要になる。CEFR の「自分の感情を読者に的確に伝えられるようなエッセイが書ける」という部分であるが、日本の学生が全員エッセイを書くことを目的にカリキュラムが組まれているだろうか。したがって、CEFR の CDS を注意深く見ると、日本の大学生には必要ではない記述も含まれていることになる。注意して検討する必要がある。始点を何もできないゼロ設定とすると、終点は理想的なネイティブスピーカーになるだろうか。日本の英語教育はネイティブスピーカーに近い人を理想としているのだろうか。また日本で学ぶ留学生は日本人の成人をモデルにすることになるだろうか。アカデミックな分野で言語行動として最終目標を規定するのであれば、能力記述のありかたも変わってくる。

11. 目標言語の位置づけ

能力記述の課題としてもう一点注意しておきたいのは、目標言語の学習環境の違いからくる記述の違いという点である。日本国内での日本語教育の場合には、第二言語としての日本語を習得することになる。生活に密着したものになる。学生の場合は、

勉強するのに必要なアカデミックな日本語ということで規定しやすい。環境が規定されることによって学習者にも認識されやすくなる。しかし、アメリカやアジア等の外国で日本語を学び習得する場合には、必ずしも日本での日本語教育に基づく能力記述がそのまま当てはまるとは限らない。日本で英語を学ぶ場合と現地で英語を学ぶ場合を考えればその違いは容易に想像できる。外国語科目としての日本語なのか、教養科目としての日本語なのか、職業のための日本語なのかなど、どのような学習環境で目標言語を学ぶかによって留意すべき点は少なくない。

12. CDS の記述内容

言語能力の記述を考える場合、やはり can-do 調査が必要になる。しかしこの can-do 調査は何のために実施するのか十分に検討しておくことが大切である。各能力のレベルを把握するために実施するのか、最終目標の熟達度に達している人の能力を見極めるためにやるのかによって、調査の内容や方法も変わってくる。最終目標を明確にするということであれば、最終到達目標として期待される can-do だけでもよいが、テストの得点の意味づけや、各段階の能力の定義や違いを明示化するということであれば、「ゆっくり話せばわかる」レベルから、「多少速く話してもわかる」レベルまで、CDS の書き方も変わってくる。記述の際に段階的な違いをどのように表記、表現するか検討しておかなければならない。また認知的負担度を言語行動から明示することも必要になる。今回の分析等から、CDS の記述には、以下の視点から多面的に言語能力が記述されていることがわかった。必ずしもパフォーマンスに限定しているわけではない。特に、評価指標としての CDS には、言語的側面について言及することが多いことがわかった。

(1)文の質・文法的正確度
(2)語彙の豊富さ
(3)話し言葉(発音)／書き言葉(表記)
(4)社会言語学的側面
(5)会話運用的側面
(6)流暢さ
(7)談話の質
(8)機能・タスク(課題)
(9)話題・内容

13. 言語運用場面の特定

言語運用場面の特定には、2つの方法がある。一つは質問紙によるアンケート調査で、もう一つは職務分析である。前者については、まずは学習者に日々どんな言語行動を

しているかを調査する。学習者向けに実施すると同時に、言語教育のプロである教師にもアンケート調査を行うことも選択肢の一つとしてある。

職務分析については、例えば、フライトアテンダントの言語運用力を調査する場合、フライトアテンダントが機内で使う日本語を調査して、どういう機能や表現があるかを分析する。必要に応じて職場、学校、あるいはその他の場面や環境での言語運用とその実態を分析する。これが能力を特定するための職務分析になる。観察という手法を用いておこなうものである。そして観察の結果を、学習者と教師双方が内省して、意見交換を行い、言語運用力の具体的な内容を記述していくことになる。

質問紙調査によるアンケート調査は、ALTE の報告書によると、数万人に対して実施したと報告されている。結果的に能力記述は何百にも及んだという報告がなされている。能力記述を代表的で、頻度の高いものに絞り込んでいくにはかなりの労力が必要になることがうかがわれる。

14. まとめ

最後に、CDS の教育的機能を改めてまとめておきたい。

①学習者自らが自分自身の該当する能力レベルと目標言語を使って何ができるか具体的な中身についても把握できるチェックリストとしての機能。

②診断的試験の開発とともに、言語活動を基本にしたカリキュラム、教材の開発にかかわる基盤としての機能。

③教育内容の透明化の基盤整備に寄与する機能。これにより、異なる外国語間での能力の枠組みを比較検討したり、同じ状況下に存在する、教育や教材の目的や内容を比較したりする手段としての機能。

④③と関連して、学習者が異なる教育機関で継続して学習する場合、学習の接続を有機的なものにし、効率のよい継続学習が実現できる機能。

⑤外国語学習者への指導や試験にかかわる者に対して、実用的な情報や資料を提供する機能。試験結果を活用しようとする者が、あるレベルでの試験の認定証の意味をよりわかりやすく解釈できる機能。

⑥研修や人事管理にかかわる人にとって、職務内容にかかわる職能を策定する際に、また、新しい職務について外国語能力の必要条件を特定する際の参考情報として活用できる機能。

⑦外国語の訓練および企業の人材採用にかかわる人々に役立つ、活動ベースの言語学的調査を実施する手段としての機能。

本稿では、上記の項目のうち、いくつかを対象に考察はできたが、Standard setting の観点からは、引き続き研究を継続する必要があることを述べておきたい。

参考文献

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.

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.

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.

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

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.

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.

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.

参考・引用ウェブサイト

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

Council of Europe：http://www.coe.int/

TOEIC Can-Do Guide：http://www.ets.org/o.pdf#search='TOEIC+can+do+guide'

http://www.coun.uvic.ca/learning/exams/blooms-taxonomy.html　(2012年2月14日)

http://www.nwlink.com/~donclark/hrd/bloom.html　(2012年2月14日)

公益財団法人 日本英語検定協会：http://www.eiken.or.jp/eiken/

Can-do statements (CDS) の規準設定
Standard setting for can-do statements

藤田智子
Tomoko Fujita

Abstract

It has become a current trend for English language programs at Japanese universities to introduce can-do statements (CDS) to their curriculum for many reasons. The Common European Framework (CEFR) is probably the most established CDS for language learners in European countries, but CDS should ideally be tailor-made for the target learners who study at the specific language program. Therefore, case studies investigating efficient ways to create valid CDS for specific language programs are quite important. This research focuses on CDS for a listening course at an English language program in a Japanese university. Firstly, a panel of teachers selected 28 CDSs for three different proficiency levels of students. Then, before the listening course began students answered the CDS as a form of questionnaire referred to as can-do self-checklist (SCL). After students completed the listening course they answered SCL again. The results were analyzed with the item response theory (IRT) one parameter model, and students' ability levels (θ) were estimated. The average θ of the students' SCL increased toward the end of the semester, and the students in the basic level increased θ the most. The results of SCL were compared with θ of students listening tests, but the correlation coefficient among these tests were mid-range, although relationships were strong at the basic level. These results might indicate that the basic level students' self-evaluation may be more reliable than teachers' expectations.

1. 問題と目的

Can-do statement (CDS) の規準設定に関して、「英語教育における習熟度レベルとCan-do statements」、「Can-do statements が英語の授業において果たす役割」、「IRT を活用した規準設定」、「プレースメントテストの研究」、これらを下位テーマにして英検委託研究2012 年度報告書にまとめた。その中でいくつかの今後の課題が浮かび上がってきた。

CDS には、Common European Framework (CEFR) (Council of Europe, 2001)や、英検、GTEC for STUDENTS, TOEFL, TOEIC などがそれぞれの規準で、「どのようなテスト結果を得た学習者は何ができる」という Can-do statements (CDS)を設定している(テストスコアの解釈規準としての CDS)。そしてまた、European Language Portfolio (ELP)の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 CDS として、Can-do チェックリストがある。

しかし、これらの CDS の妥当性の検証は充分に実施されていると言って良いのであろうか。Weir (2005)は、もっと慎重に CEFR の妥当性検証を行い、多言語共通参照枠として完成度をより高いものにするべきだと述べている。また、彼は、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)する必要があると主張している。

例えば、文化や言語環境が異なるヨーロッパの言語学習者のために作られた枠組みである CEFR を日本の言語学習者にそのまま適用させるには無理があり、変更や工夫をする必要がある。CEFR の枠組みを参照してもらい、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 を作った人々の考えでもある(Trim, 2001)。そこで、これを日本人に適用した CDS にする必要性が強調されている(境、2009;根岸、2006a)。しかし、これも日本人に適用することだけでは十分ではなく、日本人学習者の中にも子供、大人、学生、社会人など、より細かく識別して、本来は、そこで学ぶ学習者に対応した CDS を作成するべきである。

このように妥当性が高く、その英語教育プログラムの履修者の英語能力に可能な限り適応した CDS を作成する試みが行われているが、その典型的なものが、項目応答理論(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である。North and Schneider (1998)、Sato (2010)、筒井、近藤、&中野 (2007) は、CDS を自己評価のツールとして、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実施し、IRT を用いた分析を行ってその妥当性を確認した。これらの研究は主に Can-do チェックリストの結果を、IRT1 パラメタまたは 2 パラメタモデルを利用して、各 CDS の困難度パラメタを推定して難易度の規準設定の目安にする方式を採用している。

今後、妥当性の高い CDS を日本の言語教育の現場に普及させるにあたって、その重要なカギとなるのは、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できるかぎり適応した CDS の設定を追求することである。Green (2010) も、研究者、教員と学習者が実際

に使っている教材や言語運用の実践的な例を持ち寄って、より妥当な CDS のレベルの設定のために意見交換し、積極的に協力し合うことの重要性を強調している。本論が、ある日本の大学英語教育プログラムにおいて、妥当性の高い CDS を作成し規準設定するための事例研究の一つとなれば幸いである。

2. 先行研究

2. 1. テストスコアの解釈規準としての CDS

テストスコアの解釈規準としての CDS の利用は、Cambridge ESOL が CEFR と合体してテストを開発したことから、さまざまなテストがその解釈規準として独自の CDS を公表している。International English Language Testing System (IELTS) をはじめとする Cambridge ESOL(English for Speakers of Other Languages)の英語能力テストは、Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001) の 6 段階レベルと表裏一体のように合体したものであると言う(Taylor, 2003)。これは、テストの受験者たちに自分たちの得たスコアの本来の意味を、詳細な記述によって理解することを可能にする意味で非常に有用である。例えば、TOEIC Can-do Guide、TOEFL iBT as competency descriptors、などもこの動きに追随している。また、国内で代表的かつ日本語で平易に書かれているのは英検 Can-do リストである。

国内でのテストスコア解釈規準として利用される CDS に関する研究として、根岸(2005, 2006a)が、GTEC for STUDENTS という英語テストにおいて、そのテストで測った言語能力を示すガイドラインとして CDS を作成する過程について述べている。これは GTEC for STUDENTS Can-do statements としてウェブ上でも公開されていて、高校生の初級、中級、上級を中心に、リーディング、リスニング、ライティング、スピーキングの英語の 4 技能ごとに 7 つのレベルに分けている。そして、7 つのレベルに対応する GTEC for STUDENTS の 4 技能ごとのテストスコアと 4 技能それぞれの、日常、または教室内での学習タスクに基づく能力記述文(CDS)が表示されている。これは受験者たちが正解したテスト問題の特徴を、レベルごとによく調査して、その問題がどのような実際の場面に関連しているのかを記述したものである。

次に DIALANG は、CEFR をもとにした言語能力診断をオンラインで実行できるように開発された言語能力テストである(Alderson & Huhta, 2005)。ヨーロッパの 14 言語に対応でき、受験者がどの言語のテストを受けるか選択できるようになっている。 はじめに、どの言語でテストを受けるか決め、そのあと語彙テストを受けるかどうか、自己評価をするかどうかなどは、受験者が決めることができる。次いでリーディング、リスニング、ライティングの能力テストを受ける。DIALANG は、言語能力を測定することだけを目的にしているわけでなく、言語能力診断をして今後の学習に役立てるために開発されたものである。またテストの結果が、素点ではなく受験者が CEFR の A1～C2 のどのレベルに該当するかで判定されるのも特徴である。そして受

験者が自己評価をすることにより、自己のテスト得点と自己評価の相関を知ることもできる。結果レポートには、自分のそれぞれの技能が CEFR のどのレベルであるか判定されたものと、そのレベルの学習者は典型的にどういうことができるかを通知してもらえる欄がある。これは CEFR のポリシーである、「学習者が自律的に自己修正しながら学習を進めることのサポート」を提供することに対応している。

DIALANG で判定する CEFR A1～C2 のレベルの規準設定は、14 言語のそれぞれの専門家たちを集め、各技能に対して大がかりに実行された。専門家たちは、CEFR を熟知するためのトレーニングを受け、「CEFR で記述されている、あるレベルの能力を持つ受験者が、そのテスト問題に正解できるかどうか。」を判断基準にし、一つずつのテスト問題にレベル判定を下していった。 さらに、評価者間信頼性や予備テストの結果との相関係数など、量的分析の結果も踏まえ最終的に CEFR 判定レベルの分割点(カットポイント)を決めている。

斉田(2008)は、DIALANG を使って日本の大学 1 年生 130 人の CEFR でのレベルを調査した。参加者の約 8 割が、「日本で 6 年間の英語教育を受け、海外滞在経験はない。」いわゆる標準的な日本人大学 1 年生である。このテスト結果の平均は、Listening は A1, Reading は A2, Writing は A2、Structure は B1, Vocabulary は A2～B1 であった。ここで、Structure と Vocabulary を「言語知識」とし、Listening, Reading, Writing を「言語運用能力」とするならば、この被験者たちの「言語運用能力」は「言語知識」よりも 1～2 レベル低い傾向にあると言える。そして、この学生たちのテスト結果と自己評価による CEFR レベルを比較すると、一致した割合は Listening, Reading, Writing のセクションであり、いずれも一致度は 62～65%であった。

Naganuma(2008, 2010)や Naganuma & Miyajima (2006) によると、テストスコア解釈尺度として開発された CDS の中には、(1)日常、職場、学校などの場面での行動を、コミュニカティブ／アカデミックベースのタスクとして「…が、できるであろう」というように段階的に描写したタイプと、(2)テスト項目を分析してテストタスク上どのようなことができるか(例:そのリーディングテストで何点とった人はどのようなリーディングのテストタスクができる等)の指標を表したタイプに分けられると述べている。彼はまた、CDS として能力記述文で表現することによって、テストスコアという数量的な指標では具体的に分かりにくいものを、そのスコアの学習者が、実際にどのようなことができるのかを質的能力指標として示すことができるようになったと指摘している。

Weir (2005)は、妥当性の観点からテストスコアと CDS を安易に対応させることは、危険であると述べている。CEFR の各レベルの難易度に適応するようにテストを作成するには、能力記述文の内容的パラメタの難易度を決める規準の構成概念妥当性が不十分であるため、現状の CEFR には難しいと言うのである。また妥当性を満足できるようにするには、それぞれのテストが根拠とする仕様や規格を包括した独自の CDS でなければ不適格であるとも述べている。しかし、Weir(2005)は、CEFR で英語能力レベルの規準を表現することを全否定したのではない。彼は、これからの方向性として、テスト開発者たちは CEFR の 6 レベルで、何が、どのようにできる(Can-do)についての研究をさらに深め、どのような状況下でアクティビティーが実行さ

れ、そのパフォーマンスが特定の規準についてのどのような質的レベルと対応するのかについて、詳細に至るまで追及する必要があると指摘しているのである。

2. 2. 自己評価としての CDS ：Can-do 自己チェックリスト

CDS にもとづいて、学習者が自己の能力を診断したり、教員が学習者のレベルを判断する手段として利用するための自己評価チェックリストを Can-do チェックリストと言う。この代表的なものが、CEFR に基づく Can-do チェックリストとして開発された European Language Portfolio (ELP)である。ELP は、技能ごとに 6 段階の CEFR それぞれのレベルにおいて、目標とする学習行動のなかでできること(Can-do)をリストにしたもので、このリストを、学習者が自己評価としてチェックすることによって、自分の能力レベルを診断することができる。このようにして ELP は、能力と目標の 2 つの面から学習のプランを立て、学習者が自ら目的をはっきりと持って学習できるようにし、最終的には、学習者の自律的学習を促進することをめざしている。そしてまた、学習の記録を残すことができるようにするために、ポートフォリオのスタイルをとっている。ELP は、CEFR の 6 段階(A1, A2, B1, B2, C1, C2)のレベルごとに、領域、場面、状況に合わせた能力記述文が設定されている。

North (1995, 2000), North & Schneider (1998) は、難易度の論理的な段階的尺度を作成するために、テスト項目と同じように多くの能力記述文を、IRT(Rasch モデル)を利用して分析検証した。彼らは、言語能力を communicative language activities, strategies, qualitative aspects of language proficiency のようにカテゴリーに分ける大枠を作り、さらにその中で細分化してからそれぞれにあてはまる能力記述文を作成した。次に、その能力記述文を利用して教師が学習者を評価し、その結果を同一尺度化するためにラッシュ(Rasch)モデルによる項目バンク作成手法を用いて分析した。その後、Lenz & Schneider (2004)は、作成した英語の能力記述文の項目困難度を、能力記述文項目バンク (Bank of Descriptors) としてウェブ上で公開している。

Sato (2010)は、英検 CDS を自己評価ツールとしてその妥当性の確認をした研究を実施した。彼は、英検 CDS のうち、5 級～準 2 級までの 16 項目の CDS を、2571 人の日本の中学 1～3 年生に自己評価として回答してもらい、そのデータを、Rasch モデルを使って分析した。その結果、16 項目が被験者の中学生たちにとって、比較的困難度が低めで、また、16 項目に対する自己評価による項目困難度と 5 級～準 2 級までの設定されたレベルは、ほぼ一致した。さらにまた、この受験者の自己評価結果と彼らの英語能力のレベル、さらに英語学習に費やした時間とも比例した。しかし、研究対象とした 16 項目は英検 5 級～準 2 級までの CDS の限られた一部であるため一般化することは難しいが、これら 16 項目の英検 CDS については妥当性が高いと言うことができる。

最後に、CDS と規準設定に関する研究で、日本人学習者のスピーキング能力の CDS と規準設定に関するものとしては、筒井、近藤、＆中野(2007)が挙げられる。これは、後に述べる

North & Schneider (1998)の研究で開発された能力記述文をもとにして、ある日本の大学で、スピーキング能力の自己評価と教師評価を CEFR の 6 レベルに分かれた習熟度別レベルに分けて実施・比較したものである。英会話力育成コースで学ぶ約 2600 人の学生は、能力記述文の中からスピーキングの項目を 99 選んで作った自己評価チェックリストに回答した。これと同時に、担当教員たちには同じ 99 項目のチェックリストで学生を評価し、これら学生自己評価と教師評価を比較した。BILOG-MG3.0 を使用して、2 パラメタ IRT モデルでこの結果を分析したところ、学生自己評価と教師評価の項目困難度の相関はかなり高いが、学生自己評価と教師評価そのものの相関は低いということが分かった。また、このコースの 6 段階に分かれた習熟度別レベルごとの学生自己評価と教師評価の両方を、項目特性曲線を描いて比べてみたところ、両方の曲線ともに CEFR と同じように 6 段階になった。

2. 3. 日本の大学英語教育に CDS を取り入れる動き

この約10年、高等教育の英語の授業にCDSを導入する動きは既に緩やかに広まりつつあったが、2012 年には、文部科学省に「外国語教育における「CAN-DO リスト」の形での学習到達目標設定に関する検討会議」が設置され、その動きはますます本格的に拡大しつつある。

ここで、CEFR に基づいた CDS を日本の大学英語教育に導入する動きに焦点を当てる。まず、茨城大学、大阪大学、慶應義塾大学では、その英語プログラムに CEFR や CEFR をベースとしたCDSを導入しようとした。茨城大学では、CEFRのレベルを基準にして習熟度別クラスを編成し、また総合英語プログラムを開発し、自律的に英語学習ができる人材養成をめざしている(Ano et al., 2007; Fukuda, 2009; Nagai & Fukuda, 2004)。大阪大学では、25 の専攻語すべてにおいて、到達目標を CDS で表して公開するという 「透明」「共通」「強制しない」姿勢で CDS を中心としたカリキュラム改革を行ってきた(真嶋、2010)。さらに、Majima (2010)では、日本で CEFR を取り入れた言語教育を行っている事例を 7 つの活用分野に分けて紹介し、そのうちの一つが「CEFR のレベルと教育機関の言語プログラムの到達目標を関連づけたもの。」である。 さらに慶應義塾大学では、小中高大一貫教育の中に、CEFR を基にした英語教育を実現しようとしている。この中心的な取り組みの一つとして、English Language Portfolio (ELP)の日本版と言える慶應 ELP を開発、試行している (Horiguchi, et al., 2010)。最後に、大学生が英語の授業で必要とされる能力に関して、清泉アカデミック Can-do Scale として 4 技能ごとに 20 の CDS が作成された(Naganuma & Miyajima, 2006)。

2. 4. 日本人学習者に適応する CDS へ

ヨーロッパの言語学習者のために作られた CEFR は、日本人学習者にそのまま適用するには無理があり、修正や工夫をしてより日本人学習者に適応させる必要があると言われている(境, 2009; 根岸, 2006b)。

例えば、中島・永田(2006)は、CEFR 準拠の自己評価アンケートである DIALANG

self-assessment (SAS) を使用して、CEFR がどのくらい日本人学習者に適用可能かを検証した。彼ら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。さらに根岸(2006b)は、この研究の中で、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、CEFR の Reading の A1 レベルの項目にある「葉書などに書かれた、短く簡単なメッセージを理解することができる。」に対して、日本人学習者はより困難である A2 レベルと判定した。これは、おそらく CEFR の基準では、カードに Happy Birthday! や Congratulations! にプラスして、とても簡単な短いメッセージを付け加える程度を設定していたと思われる。ところが日本人学習者たちが、「post card = 葉書」に書かれたメッセージとして連想する内容が、もっと長い情報量であったからだと思われる。そしてまた、「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英語でこれらの経験をしたことがほとんど無いために、困難度が高いと思ったからだと推測できる。このように、学習者が自己評価するとき、彼らが体験したことがない内容を自己評価のための質問にしても、その回答はあまり正確ではないと言われている(伊東・川口・太田、2008)。

Negishi (2005)や根岸(2006)では、このように CEFR レベルと日本人学習者の判定が異なった項目に、学習者が具体的に内容を理解するための工夫として参考資料を付けることで成果をあげたと報告している。例えば、前述した Reading A1 レベルの項目には、参考資料として具体的なカードの見本を示し、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した。両方とも改良後の項目の困難度は、ほぼ CEFR 設定どおりの順序となった。

さらに、CEFR をもっと日本人学習者に適用させる動きのなかで、日本版 CEFR(CEFR-J)のフレームワークを構築しようとする取り組みも行われている。ここではまず、一般的な日本人学習者のレベルは、CEFR の下位レベルをさらに細かく分ける必要があると認識し、ヨーロッパで 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、A1 を 3 つに、A2, B1, B2 はそれぞれ 2 つに分ける日本人学習者向きレベルの設定を提唱している(岡、2008)。

そしてこの動きと符合する研究として、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)を使って調査した斉田(2008)によると、日本人大学 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 レベルという結果になった。これは、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、やはり A1、A2、B1 の 3 レベルのなかに、より詳細なレベルを設定したほうが現実的である

という方向性をサポートしている。

2. 5. 項目応答理論(IRT)の利用

テストスコアの解釈規準としての CDS についても、またここで挙げた自己評価としての CDS に関する先行研究のほとんど全てにおいて研究結果の分析に使用されている IRT についてここで説明を加えたい。

IRT の代表的なモデルとして、1 パラメタ(1PL)、2 パラメタ(2PL)、3 パラメタ(3PL)の 3 モデルがある。1～3PL モデルはそれぞれに違う式で表され(式 1～3)、それぞれの特徴や、項目パラメタを安定して推定するために必要な被験者数もモデルによって異なる(Bond & Fox, 2001; Brown & Hudson, 2002; Hambleton, Swaminathan, and Rogers, 1991; McNamara, 1996; 大友, 1996; 芝, 1991)。

$$P_j(\theta) = \frac{1}{1+\exp(-(\theta - b_j))} \quad \cdots\cdots (1)$$

$$P_j(\theta) = \frac{1}{1+\exp(-Da_j(\theta - b_j))} \quad \cdots\cdots (2)$$

$$P_j(\theta) = c_j + (1-c_j)\frac{1}{1+\exp(-Da_j(\theta - b_j))} \quad \cdots\cdots (3)$$

Rasch モデルとも呼ばれる 1PL モデル((1)式)は、この式に含まれているように b パラメタ(項目困難度)の推定をするもので、2PL モデル((2)式)は、b パラメタに加えて a パラメタ(項目弁別力)の推定もできる。また、3PL モデル((3)式)は、b、a パラメタに加えて c パラメタ(当て推量)も推定できる。これらの 3 モデルそれぞれに対して、安定した推定に必要とされる受験者数は異なり、1PL モデルでは、500 人以下、2PL モデルでは 500 人から 1000 人(eg, Ayala, 2009)、3PL モデルでは 1000 人以上の受験者が必要だと言われている(Lord, 1968)。これに加え最近の研究では、野上(2009) や野上、小林、& 林 (2010)が、3PL モデルの下方漸近線パラメタを推定せずに、選択肢数の逆数に固定する方法(3PLcFix)を利用すると、3PL モデルに比べて少ない人数の被験者数であっても比較的安定した項目パラメタ推定を行える可能性があると提案している。

ここで IRT モデルと被験者数の関係について、3PL モデルのほうが 2PL モデルより多い受験者数を必要とする理由を例にとってみる。3PL モデルは 2PL モデルより項目ごとのパラメタ数が多いということに加えて、当て推量パラメタの影響を受けて、能力推定値が高い受験者の情報量を過大評価し、能力推定値が低い受験者の情報量を過少評価する傾向がある。 従って、2PL モデルより 3PL モデルの有効サンプル数は少なくなるので、2PL モデルより受験者数が多くなければ 3PL モデルの項目パラメタ推定値はより不安定になる傾向がある(張,

2009)。

このように、どの項目応答モデルを採用するかによって、分析結果の精度が変わることがあるので、データや目的を良く考慮して、どの項目応答モデルが最も適応しているか慎重に吟味する必要がある(e.g., Choi & Bachman, 1992；Kolen & Brennan, 2004)。

2. 5. 本研究で解明しようとすること

日本の大学英語教育において、CDS を学習到達目標に設定する動きが加速している。しかし、実際に日本人大学生を対象として、どのような方法で、どのような CDS を導入すべきかに関する事例研究は、まだ充分実施されているとは言い難い。設定される CDS は、その英語教育プログラム独自の CDS であることが望ましく、その作成にあたって、まずは学習者による自己評価(Can-do チェックリスト)をもとに習熟度レベルの規準設定をする方法を検討する。本論では、英語リスニング能力についての Can-do チェックリストに注目し、その英語プログラムを履修する大学生の自己評価による困難度に適応した Can-do チェックリストを作成するための調査を実施した。学生の Can-do チェックリストの反応と、学生の習熟度レベルやリスニングテストのスコアとして測定された能力との相関性を調査して、CDS の規準設定にあたって、どのような点に留意することが必要なのか調査することにした。

まず、(1)事前事後(学期初めと終わり)で実施した日本人大学 1 年生の自己評価としての Can-do チェックリストの結果は、事前事後どのように変化するのか、さらにその変化の度合いは 3 つの習熟度別レベルごとに違いがあるかを調査する。そして、(2)Can-do チェックリストとテストスコアの比較をし、この相関係数が 3 つの習熟度レベルごとにどのように異なるのか比較する。最後に(3)Can-do チェックリストを作成した時、教員たちが想定した項目の困難度と、IRT を用いて分析した項目困難度推定値がどのくらい一致しているか検討する。

3. 研究方法

3. 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 445 人(全体の約 8%)が本研究に参加した。彼らはプレースメントテストのスコアによって 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)に分けられて 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(SCL)を実施してもらった。2 回の自己評価回答者の習熟度レベル別の内訳は、表 1 のようになっている。445 人が学期が始まってすぐ(4 月)に、本研究における一回目の Can-do チェックリスト(今後 SCL1 と呼ぶ)に回答した。そのうちの、331 人が約 3 か月半後の学期末(7 月末)に 2 回目の Can-do チェックリスト(SCL2)に回答した。

表 1. SCL1 と SCL2 の回答者レベル別人数

習熟度レベル	SCL 1	SCL 2
Basic	106	48
Intermediate	250	203
Advanced	89	80
total	445	331

3. 2. Can-do チェックリスト(SCL)

本研究で使用する Can-do チェックリスト(SCL)は、CEFR、European Language Portfolio, City & Guilds (International English Qualifications), TOEIC, 英検の CDS, CEFR の日本語版(吉島・大橋, 2004)を参考にした。また、本論での SCL は、学生に分かりやすくするために、日本語で書くことにしたので、これらの中でも、日本人学習者のために日本語で書かれた英検 CDS は、最も参考にした部分が多い。また、根岸(2006b)が示していたように、より的確に日本人学習者に内容を理解してもらうための手掛かりとして( )に例を入れているところも英検 CDS を参照した。従って本論の SCL にも( )に短く具体例を入れている。例えば、SCL20:「買い物に行った場合、商品について店員からの情報(サイズ、機能、割引、在庫など)を聞いて理解することができる。」のようである(Appendix A 参照)。「商品についての情報」だけに止めるよりも、( )内のような具体的な例があると、被験者は容易に SCL に書かれている内容を理解することができると思われる。

本論の SCL は、3 種類(初級、中級、上級、それぞれ 3 レベルに到達目標として 14 ずつの能力記述文がある。以下に示すように、合計 28 の能力記述文が 7 文ずつ別のレベルと重なる構成になっている(図 1 参照)。

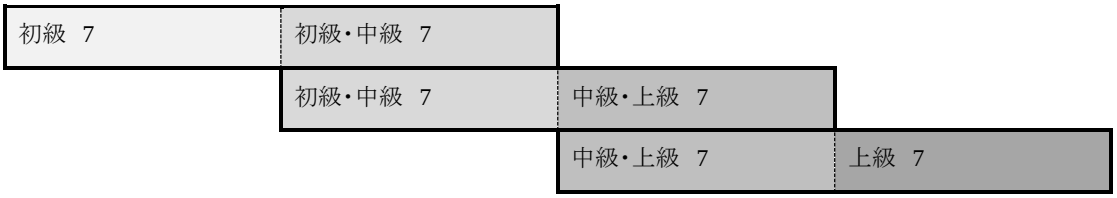


図 1. 3 種類(初級、中級、上級)の Can-do 自己評価チェックリスト(SCL)

このような形の SCL にした理由は二つある。一つは、1 人の学生が多くの質問に答える必要がなくなるようにすることで、もう一つは、初級の学生が上級の SCL に答える必要がなく、回答者の習熟度に適応した質問をすることができるからである。

これら 28 項目をより妥当なものにするために、この英語プログラムに所属するリスニングコース担当教員のうち、10 人にご協力いただき、アドバイスやフィードバックをいただいた。各担

当教員には、(1)それぞれの SCL の難易度レベルが想定したレベルと合っているか、また、(2)SCL の内容が学習者に問題なく理解できるような表現になっているか、(3)能力記述文の表現に誤りがないかなど、これら 3 点を中心に修正・変更したほうが良いと思われる点に赤入れしたり、書き出したりしてもらった。これらを回収して、修正、変更、削除を行い最終版の 28 項目からなる SCL を作成した。

3. 3. リスニングテスト 1 とリスニングテスト 2

リスニングテスト 1 は、学期初めに実施する学生の英語能力を判定するための実力テストで、所要時間は 90 分間、そのうちリスニング 30 問、文法 30 問、リーディング 40 問の合計 100 問に多肢選択(4 択)で解答するテストである。表 3. 4. の文法とリーディングテストは、これらのサブテストのことである。また、リスニングテスト 2 は、学期末に実施されたリスニング能力の到達度を測る 70 問の多肢選択(4 択)問題のテストである。

3. 4. IRT による分析

被験者の能力値 θ が変化し、項目パラメタは同じという前提で BILOG-MG3.0 を使用して、1 パラメタ IRT モデルで分析した。2 パラメタモデルを使用しなかった理由は、被験者数が 500 人以下なので、パラメタの推定が不安定になることを避けるためと、アンケートタイプの項目であるため、*a* パラメタの必然性がそれほど高くないと判断したためである。

ここで SCL1 に応えた被験者を、それぞれ初級 B、中級 I、上級 A という習熟度レベルごとに初級から順に G1-B、G2-I、G3-A とし、SCL2 に応えた被験者も同じように G4-B、G5-I、G6-A とラベリングした(図 2 参照)。この時、SCL 1 の分散が 1、平均が 0 として、SCL 1 では G1-B を、SCL 2 では G4-B を基軸にして、SCL 1 の 6 グループを比較した。さらに、SCL 2 と 6 グループを比較するため SCL 2 は Ref = 1 で 6 グループは Ref = 4 にして BILOG-MG3.0 にかけた。

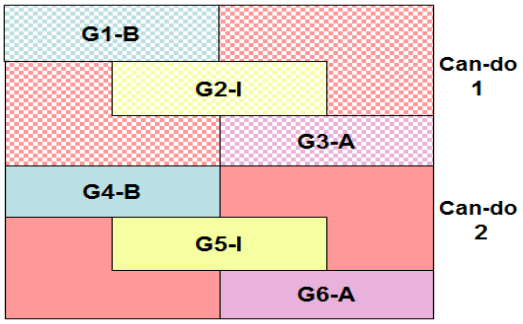


図 2. 分析方法:SCL 1 と SCL 2 に応えた 3 習熟度別レベル

4. 結果

4. 1. SCL1 と SCL2 への反応の変化

前述した分析方法で、6 つのグループの平均能力値(θ)を比較したのが表 2 である。これによると、SCL1 から SCL2 への変化は、全体的に上昇しており、学期初めより学期末の方が平均で θ = 0.322 上昇していた。また、習熟度別の θ の平均値は、学期末のほうが初級レベルは 0.310、中級レベルは 0.315 上昇したが、上級レベルだけ 0.172 にとどまり、他のレベルの約半分の上昇となった。

表 2. SCL 1 と SCL 2 の習熟度別能力値(θ)の平均変化

	SCL 1	SCL 2	SCL 2 - SCL 1
初級	-0.410	-0.100	0.310
中級	-0.254	0.061	0.315
上級	0.417	0.589	0.172
全体	-0.157	0.165	0.322

4. 2. Can-do チェックリスト(SCL)とテストスコアの比較

表 3 は、SCL1 と英語テストスコアの相関関係を表している。SCL 1 とすべての英語テストは、$p<0.01$ で全ての相関係数が有意だと認められたが、その相関係数は $0.275 < r < 0.329$ と全体的にあまり高い相関関係だとは言えない。SCL 1 と各英語テストとの相関関係に比べ、英語筆記テストどうしの相関関係(例えば、リスニング 1 とリスニング 2 は $r = 0.752$)は高く、リスニング能力を測るテスト(リスニング 1、リスニング 2)とリーディング能力を測るテストであっても英語筆記テスト間では、高い相関係数を示している(リスニング 1 x リーディング、$0.746 < r < 0.753$)。これは自己評価である SCL と筆記テストとの形式の違いからくるものと推測される。

表 3. SCL1 と英語テストスコアの相関関係(Peason)

	SCL1	文法	リスニング 1	リーディング	Total
文法	.275**				
リスニング 1	.313**	.666**			
リーディング	.325**	.716**	.746**		
Total	.322**	.860**	.852**	.920**	
リスニング 2	.329**	.702**	.752**	.719**	.790**

Note. すべての相関係数は有意差 $p<0.01$. $N = 426$

さらに、表 4 は SCL 2 と英語テストスコアの相関関係を示している。SCL 2 と英語テストのスコアも、$p<0.01$ で全ての相関係数が有意義であることを示しているが、その相関係数は $0.210 < r < 0.327$ で、SCL 1 とほとんど大きく変わらずあまり強い関係があると認められなかった。

表 4. SCL2 と英語テストスコアの相関関係(Peason)

	SCL2	文法	リスニング 1	リーディング	Total
文法	.210**				
リスニング 1	.325**	.660**			
リーディング	.273**	.719**	.753**		
Total	.286**	.858**	.849**	.929**	
リスニング 2	.327**	.683**	.780**	.738**	.797**

Note. すべての相関関係は有意差 $p < 0.01$. $N = 324$

表 5 は、習熟度レベル別に SCL 1、SCL 2 と、リスニングテスト 1、リスニングテスト 2 との相関関係を集計して比較したものである。SCL 1 は、習熟度レベルによって大きな変化はなく、全体的に良く似た相関係数を示している。しかし、SCL 2 には、特徴があり、リスニングテスト 1、リスニングテスト 2 ともに初級レベルが最も相関係数が高く($r = 0.414$、$r = 0.395$)、次に中級レベル($r = 0.175$、 $r = 0.230$)、上級レベル($r = 0.211$、 $r = 0.115$)の順になっている。これは習熟度が低いほど、リスニングテストとの相関係数が高いという結果を示している。

表 5. 習熟度別 SCL1＆ SCL2 と英語テストスコアの相関関係

	SCL1			SCL2		
	初級 B	中級 I	上級 A	初級 B	中級 I	上級 A
リスニング 1	.367	.256	.397	.414	.175	.211
リスニング 2	.326	.332	.289	.395	.230	.115

Note. すべての相関関係は有意差 $p < 0.01$.

4. 3. IRT による項目困難度推定値と想定した困難度

SCL の 28 項目を IRT による分析し、項目困難度順に並べたものが Appendix A である。項目困難度パラメタの値が小さいものを上から順にならべたのが「パラメタ順」で、その次の列に

は項目困難度(*b* パラメタ)の値が示されている。その次の列には「想定順」として、SCL を作成したときに教員グループで想定した困難度の順が表示してある。この「パラメタ順」と「想定順」、2 つの順位が 6 位以上違っている SCL を探った。その中で、想定順よりパラメタ順が上位となったのは、28 位⇒16 位、20 位⇒13 位、12 位⇒6 位、そして想定順よりパラメタ順が下位となったのは、15 位⇒21 位、8 位⇒14 位 で、合計 5SCL あった。

まず、パラメタ順が想定順の困難度より低くなった SCL から調べてみると、パラメタ順 16 位が想定順 28 位より 12 位も上になった SCL28:「いろいろな種類のドラマ、ドキュメンタリーや映画などを楽しみながら理解することができる。」については、対象となるドラマ、ドキュメンタリーや映画が、そのテーマや内容によって難易度は大きく異なること、また「楽しみながら理解する」のは、どの程度の深い理解であるのか、など限定しづらいところが、順位を大きく変えた理由ではないかと考えられる。他に、パラメタ順 13 位で、想定順 20 位であった SCL20:「買い物に行った場合、商品について店員からの情報(サイズ、機能、割引、在庫など)を聞いて理解することができる。」 は、店員とのやりとりが買う商品によって、また会話の内容の奥の深さによって大きく変化することが考えられる。例えば、その商品がパソコンで、機能についての詳細な内容のやりとりになる場合と、商品が T シャツでサイズや色についての単純なやりとりになる場合とでは、難易度が大きく変わるはずである。最後に、想定順 12 位であったがパラメタ順は 6 位であった SCL12:「自分の良く知っている話題(趣味や好きなこと)で、簡単な内容であれば、話の要点を理解することができる。」 これも、先に述べた SCL28 や SCL20 と同じく、話題の種類によって難易度は異なるうえ、「簡単な内容」の簡単さが受け取り方に個人差がある。

反対に、パラメタ順が想定した順位より低くなった SCL を見てみると、パラメタ順が 21 位となり、想定順 15 位がより困難度が高いと判断される傾向にあった SCL15 「テレビで政治、社会、経済などに関するニュースを見て、映像を見ながらその要点を理解することができる。」 これは、ニュースの話題である、政治、社会、経済が、大学 1 年生にとってはあまり身近でなく、関心がない場合は母国語でも難しいと感じたかもしれない。また、パラメタ順が 14 位で想定順 8 位の SCL8 「テレビのニュースのトピックや天気予報、商品の宣伝などの要点を理解することができる。」については、英語放送のテレビを見た経験がないということが、想定した順位よりも学生たちが難しいと考えた原因だと思われる。

5. 考察

英語リスニング能力に関する Can-do 自己チェックリスト(SCL)で、実際に学生たちがリスニングコースを履修したあと、履修前とどのような変化が履修したことで、あるのか確認するために、学期初めと終わりに同じ Can-do 自己チェックリスト(SCL1 と SCL2)による自己評価を行い、その平均能力値(θ)の変化を比較した。その結果、被験者たちは学期初めに実施した SCL1 より、学期末に行った SCL2 のほうが、平均で θ = 0.322 上回る傾向にあった。そして習熟度別による調査では、中級と初級レベルの学生の平均 θ が SCL1 と SCL2 を比較して、ほぼ同じく

らいの伸び(θ = 0.310、0.315)があったが、上級レベルのみが 0.172 の伸びにとどまった。 コースを履修した事前事後で学習者たちが自己評価を行うと、ふつう自分の能力が「上昇した。」と答える人が「下降した。」と答える人を上回る。したがって、事前事後で θ が 0.322 伸びたことは特筆するべきことではないが、習熟度別の θ の変化では、上級レベルの θ の伸びだけが、他のレベルの半分であったことは、「上級者ほど自分の能力の伸びを慎重に評価する。」、また 「初級者ほど能力が伸びる余地を多く残している。」などの理由が推測できる。

次に、SCL1、SCL2 と英語筆記テストスコアとの相関関係であるが、本論で用いた英語筆記テストは、CEFR のレベルに合わせて作成されたものではなく、また何らかの規準設定を目指して作成されたものでもない。SCL のように、リスニングに関する能力記述文を読んで自己評価をアンケート形式の 4 択で答えるものと、正解不正解のある筆記テストの相関関係は、リーディングとリスニングテストのように違う技能のテストであっても、筆記テストどうしの相関関係のほうが高くなることが多い。SCL1 も SCL2 も英語筆記テストとの相関係数は $0.210 < r < 0.329$ で決して高いとは言えない。それでも、リスニング能力に関する SCL2 は、リーディングに比べ、リスニング筆記試験と高い相関を示している。

また、この SCL と英語筆記テストの相関係数について、習熟度レベル別に調査したところ、初級レベルが筆記テストとの相関関係が一番高くなった。この結果は、言語能力レベルが低い人ほど SCL と筆記テストのスコアとの相関がはっきりし、能力レベルが高い人の SCL は筆記テストのスコアとあまり相関が高くないと言える。本研究に参加した教員の何人かは、「初級レベルの学習者は、あまり真剣でなく、いい加減に自己評価をしたと思う。」と報告したが、実は初級レベルのほうが上級レベルより、自己評価の結果が自らの英語筆記テストのスコアに近い可能性がある。

SCL 作成時に教員たちが想定した項目の困難度による順序(想定順)と、それに学生が応えた結果を、IRT を用いて分析して得られた項目困難度推定値パラメタによる順序(パラメタ順)のくい違いを調査した結果、くい違いが 6 位以上のものが 5 項目あった。この 5 項目の問題点を総合的に考察すると、第 1 に、聞く対象となる英語、つまり会話、テレビ、映画、店でのやりとりなどの「話題」をシステマティックに限定して、難易度を設定する必要がある。話題の難易度に関する規準設定は、CEFR に関する本に詳細に示されているが、その設定の規準はヨーロッパの言語学習者である。日本人学習者を対象にした難易度設定の規準には、日本人学習者の環境や条件を加味した設定にしなければならない。そして、「話題」だけでなく、会話、映画、テレビ、ラジオなど、英語をどのような媒体で聞くかについても、日本人向けの規準設定に配慮をする必要がある。SCL28 のように「いろいろな種類のドラマ、ドキュメンタリーや映画」は、日本では字幕放送や吹き替えで見る機会も多い。日本人学習者が、これらを見たことがあり、ある程度想像することが可能かもしれない。それに反して、SCL15 や SCL8 のように、英語圏で放送されているような、ニュース、天気予報、TVCM は、 実際のテレビ番組を見る機会があまりない日本人学習者にとって、難しいと感じるのは自然な反応である。このように、

「経験したことがない話題」に関して学習者が自己評価するとき、 あまり正確な判断ができないことは、先行研究結果と一致する。また、難しさを表す方法の工夫も必要で、SCL28 の「楽しみながら理解する」のは、どの程度の深い理解であるのか限定しづらい。また、SCL12 の「簡単な内容」も、その簡単さに個人差がある。このようにあいまいな表現方法はあまり使わないほうが、より正確な自己評価が可能な SCL にする方策だと思う。

6. 結論

今後も日本の大学英語教育において、CDS を授業に取り入れる動きは進んでいくと思われる。最も普及しているのは、CEFR を規準にした CDS や SCL であるが、これらはヨーロッパの言語学習者のためのもので、日本人学習者にそのまま適応することは難しい。本来 CDS は、その英語教育プログラムのニーズに適応するように作成されるべきで、例えば日本人学習者で、大学 1, 2 年生というように、対象履修者を絞り込んで、その英語教育プログラムに可能な限り合った独自の CDS を作成するべきだと考える。

本論では、このように、ある大学英語教育プログラム独自の CDS を作成するにあたって、どのような留意点があるのか探るため、その CDS に基づいた SCL を学習者に回答してもらい、習熟度レベルの規準設定をする方法を検討した。そして、英語リスニング能力についての SCL に注目し、その英語プログラムを履修する大学生の自己評価を基にした困難度に適応した SCL を作成することを目指した。学生の SCL への反応と、学生の習熟度レベルやリスニングテストのスコアとして測定された能力との相関性を調査したところ、上級レベルの学生よりも、初級レベルの学生ほど、自分たちの能力が伸びたと実感しやすいと言う結果が得られた。また、教員たちが心配するほどは、初級レベルの学生たちはいい加減に自己評価をしているわけではないという傾向も示された。これらの結論によって、その英語教育プログラムの履修者に適合した SCL を設定しようとするときに、履修者たちに実際に作成した SCL を自己評価してもらい、その結果を、IRT を用いた分析によって困難度パラメタを推定して規準設定に役立てる可能性を示した。

また、本論では教員たちが想定した SCL の困難度による順位(想定順)と、IRTによる SCL 学生自己評価による項目困難度の順位(パラメタ順)のくい違いから、今後の能力記述文(CDS)作成時に参照すべき、さまざまな推論を導けた。リスニング能力に関する CDS の作成に関して、日本人大学生たちが聞く英語の「話題」による困難度と、「媒体」(例:会話、映画、テレビ、公共のアナウンス)について、もっとシステマティックに困難度を規準設定する必要性がある。そしてまた、学習者によって受け入れ方の違う表現(例:「簡単な」、「楽しみながら理解できる」)を減らし、学習者が的確に理解しやすい表現を工夫する努力をすべきであることも痛感した。

最後に、今回の事例研究を通して感じたことは、日本の英語教育において、その英語教育プログラムに適合した CDS を設定するための研究を、もっとさかんに実施するべきである。そ

して、CDS の規準設定に対する、さまざまな角度からのデータを収集し、研究者どうしの連絡や情報交換をもっとさかんに行う必要性を感じた。長い道のりになるかもしれないが、このようなプロセスを経てこそ、英語を学ぶ多くの学習者たちが、CDS をより有効に利用できるようになると考える。

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.

Ano, K., Betts, R. Fukuda, H. Nagai, N. Okayama, Y. Sasaki, M., & Ueda, A. (2007). Can-do statements based on CEFR: A case study of IEP at Ibaraki University. *Studies in Humanities and Communication Ibaraki University, 2,* 1-18.

Ayala, R. J. (2009). *The theory and practice of item response theory.* New York: Guilford.

Bond, T. G., & Fox, C. M. (2001). *Applying the research model: Fundamental measurement in the human sciences.* London: Lawrence Erlbaum.

Brown, J. D., & Hudson, T. (2002). *Criterion-referenced language testing.* Cambridge: Cambridge University Press.

Choi, I. C., & Bachman, L. F. (1992). An investigation into the adequacy of three IRT models for data from two EFL reading tests. *Language Testing, 9,* 51-78.

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

Fukuda, H. (2009). The possibility of applying CEFR to English education in Japan. *Studies in Humanities and Communication Ibaraki University, 6,* 25-41.

Green, A. (2010). Conflicting purposes in the use of Can-do statements in language education. In M. Schmidt, N. Naganuma, F. Dwyer, A. Imig & K. Sakai (Eds.), *Can-do statements in language education in Japan and beyond - Applications of the CEFR* (pp. 35-48). Tokyo: Asahi Press.

Hambleton, R., Swaminathan, H., & Rogers, H. (1991). *Fundamentals of item response theory.* London: Sage Publications.

Horigtuchi, S., Harada, Y. Imoto, Y., & Atobe, S. (2010). The implementation of a Japanese version of the “European Language Portfolio-Junior version-” at Keio: Implications from the perspective of organizational and educational anthropology. In M. Schmidt, N. Naganuma, F. Dwyer, A. Imig & K. Sakai (Eds.), *Can-do statements in language education in Japan and beyond - Applications of the CEFR* (pp. 138-154). Tokyo: Asahi Press.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking methods and practices.* New York: Springer.

Lenz, P., & Schneider, G. (2004). *A bank of descriptors for self-assessment in European*

language portfolios. Strasbourg: Council of Europe.

Lord, F. M. (1968). An analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three-parameter logistic model. *Educational and Psychological Measurement, 28,* 989-1020.

MacNamara, T. (1996). *Measuring second language performance.* New York: Longman.

Majima, J. (2010). Impact of Can-do statements / CEFR on language education in Japan: On its applicability. In M. Schmidt, N. Naganuma, F. Dwyer, A. Imig & K. Sakai (Eds.), *Can-do statements in language education in Japan and beyond - Applications of the CEFR* (pp. 57-65). Tokyo: Asahi Press.

Nagai, N., & Fukuda, H. (2004). Goal setting of general English language program at Ibaraki University based on CEFR. *Studies in Humanities and Communication Ibaraki University, 16,* 75-105.

Naganuma, N. (2008). The potential of Can-do scale to provide better English education. *ARCLE Review, 2,* 50-77.

Naganuma, N. (2010). The range and triangulation of Can-do statements in Japan. In M. Schmidt, N. Naganuma, F. Dwyer, A. Imig & K. Sakai (Eds.), *Can-do statements in language education in Japan and beyond - Applications of the CEFR (pp. 19-34).* Tokyo: Asahi 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.

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

Taylor, L. (2003). 'The Cambridge approach to speaking assessment'. *Research Notes 13*: 1–4. Cambridge: Cambridge ESOL 13. Available online www.CambridgeESOL.org.

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.

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

張一平(2009). 2 パラメータと 3 パラメータ項目反応帆出るにおける比較。『行動計量学』、36(1)15-24.

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.

野上康子(2009). 多肢選択形式のテストの分析に使用する 2 値型IRTモデルの選択に関する検討.『日本テスト学会第 7 回大会発表論文抄録集』、128-131.

野上康子・小林夏子・林則生(2010). 多肢選択形式のテストにおける 2 値型IRTモデルの項目パラメタ推定と受験者に関する検討　Paper presented at the Tokyo 8 th (JART), Tokyo.

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.

大友賢二(1996).『項目応答理論』、東京: 大修館.

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.

境一三(2009). 日本における CEFR 受容の実態と応用可能性について-言語教育政策立案に向けて -『英語展望』、117, 20-25.

芝祐順(編) (1991). 項目応答理論－基礎と応用－東京大学出版会.

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.

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.

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.

Appendix A			
難易度順	Bパラメタ	想定順	CDS
1	-1.23	2	ゆっくりペースで繰り返して話されれば大切な情報(例えば、メールアドレス、電話番号など)を正確に理解することが。。。
2	-1.03	6	ゆっくり話されていれば、声の調子を参考にしてその話者の感情や態度を理解することが。。。
3	-0.99	3	ゆっくりなら日常生活に関する簡単で短い話(家族、趣味、大学、週末など)の大筋やキーワードを理解することが。。。
4	-0.92	1	初めて会った人との挨拶や普段の挨拶などを理解することが。。。
5	-0.85	9	簡単で短かければ日常生活に関する話(家族、趣味、大学、週末、部活など)の内容を理解することが。。。
6	-0.75	12	自分の良く知っている話題(趣味や好きなこと)で、簡単な内容であれば、話の要点を理解することが。。。
7	-0.71	4	ゆっくり繰り返して話されれば簡単な指示(道案内、集合場所、発着時間など)を聞きその内容の大筋を理解することが。。。
8	-0.55	7	ゆっくりペースではっきり話されれば、短く簡単な(駅や館内放送等の)アナウンスを理解することが。。。
9	-0.49	5	ゆっくり話されている会話のテーマが何か理解することが。。。
10	-0.44	11	簡単な内容で短かければ、電話で相手の話(伝言、日時や場所など)を理解することが。。。
11	-0.36	14	短く簡単な内容であれば、話者の主張(賛成か反対か？など)、感情や態度を理解することが。。。
12	-0.05	13	十分な資料、図表や絵などのビジュアルな助けがあれば、英語で行われる簡単な授業、研修、交渉の内容を理解することが、、、
13	0.06	20	買い物に行った場合、商品について店員からの情報(サイズ、機能、割引、在庫など)を聞いて理解することが。。。
14	0.17	8	テレビのニュースのトピックや天気予報、商品の宣伝などの要点を理解することが。。。
15	0.18	10	細かい指示やアナウンス(道案内、集合場所、発着時間など)を聞きその内容を理解することが。。。
16	0.27	28	いろいろな種類のドラマ、ドキュメンタリーや映画などを楽しみながら理解することが。。。
17	0.34	16	興味・関心のある話題に関するまとまりのある話(授業、研修、講演など)の内容を理解することが。。。
18	0.41	17	観光地のガイド、博物館のツアーや施設の説明、使用方法などを聞いてその内容を理解することが。。。
19	0.43	21	テレビドラマや映画などでまとまった長いセリフを聞き、話者の気持ちや感情を理解することが。。。
20	0.54	18	グループワークやディスカッションで話し手の意見の論点を理解することが。。。
21	0.77	15	テレビで政治、社会、経済などに関するニュースを見て、映像を見ながらその要点を理解することが。。。
22	0.88	19	自分の良く知っている内容であれば、電話で問い合わせ、クレーム、交渉などを行い、その相手の話の要点を理解することが。。。
23	0.96	25	ミーティング(イベントの打ち合わせ、社内の会議など)に参加してその内容や他の人たちの意見を理解することが、。。。
24	0.97	24	幅広い、成句(例, give up/ hold out)イディオム(例,be in the same boat / break somebody's heart)、口語表現(話し言葉にしか使われない表現)を理解することが。。。
25	1.69	22	ラジオの政治、社会、経済などに関するニュースを理解することが。。。
26	1.84	26	多様な内容であっても電話で問い合わせ、クレーム、交渉などを行い、相手の話を理解することが。。。
27	1.96	23	専門性の高い様々な話題に関するまとまりのある話(一般教養、社会問題についての講演など)を理解することが。。。
28	2.01	27	仕事や研究に関する専門用語や手順を聞いて理解することが。。。

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

Rasch-LRT Approaches to Setting Standards for a Receptive Vocabulary Size Placement Test

法月　健

Ken Norizuki

Abstract

Setting standards for placement decisions is a time-consuming and complicated task. Classical standard setting methods typically require a large body of well-trained panelists to make hard decisions after a series of technical discussions. Another important condition is that one reliable and valid placement instrument (or even more than one) should be fitted into a tight institutional schedule prior to or at the start of a given educational program. These ideal conditions are difficult to realize in most educational institutions in Japan. The present study thus explores a relatively easy and yet effective standard setting procedure whereby approaches based on Rasch model and LRT (Latent Rank Theory) analyses are applied to a thirty-minute receptive vocabulary size test called SCELP (Survey of Core English Language Proficiency), which was developed in-house and used for placement decisions at a tertiary institution in Japan. The findings suggest that a test like SCELP can be a good measure of placement and that Rasch model and LRT-based approaches to setting standards are a promising area which merits continued research. It is also worth exploring the development of standards- and criterion-based receptive vocabulary tests which may be designed to reflect different vocabulary levels and demands of Eiken tests. (200 words)

1. 問題と目的

あるテストの分割点を事前に設定するには、まず、その尺度上で対象となる受験者の能力値を客観的に位置付けることが求められる。しかしながら、実際には、そのような規準設定を行っていない high-stakes テストが多く、Bramley (2010) は、全体的な試験結果の変化によって成績の境界線を変更する際に、その変化の主要因が問題の難易度にあるのか、受験者の能力によるものかを見極めることが常に問題になることを指摘している。

一方、テストの分割点を仮に(素点もしくは標準化された得点の)70 点に設定して、70 点以上を合格、69 点以下(70 点未満)を不合格にした場合、多くの場合、1 点の差は便宜的な境界線であると言わざるを得ない。

このような規準設定の問題解決法を探るため、2011 年度は、以下の 2 つの課題について文献調査を中心とした研究を行った。

(1)規準設定におけるラッシュモデルの有用性
(2)規準設定における潜在ランク理論の有用性―項目応答理論と古典的テスト理論との比較

研究の結果、段階評価に基づく潜在ランク理論は、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。

2011 年度の文献研究から提示された課題は、大掛かりな評価システムの確立や高度な分析モデル・手法の開発を追求し、日本の教育機関において個別に取り組むことが極めて難しい状況を前提とするものが多かった。また、純粋な順序尺度に基づく潜在ランク理論と間隔尺度の精度向上を追求するラッシュモデルでは、基本的理念において相容れない部分があることも否めない事実である。

しかしながら、その一方で、小泉・飯村(2010)の研究のように、それぞれの理論の特性を活かして、現実的に規準設定の問題に対処することも可能ではないかと考えた。

2012 年度は、理想的な規準設定の条件の充足よりは、実際に規準設定の目的で使用した言語テストのデータに対して、人的・技術的・時間的な制約下での規準設定を想定して、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。

2. 先行研究

本研究で扱うデータは、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、数年間実施された語彙テストの結果のうちの一部である。語彙テストを開発することになった経緯については後述するが、まず、語彙力テストの有用性について先行研究を検証する。次に、ラッシュモデルと潜在ランク理論等によるデータ分析

の結果をどのように規準設定の手続きに結びつけるか、先行研究から大まかな方向性を探ることとする。

2. 1. 語彙力テストの有用性

語彙力を測定するテストには様々な様式がある。語彙力を定義する重要な概念に、語彙知識の広さ (breadth) と深さ (depth)がある。前者は、ある程度知っている語彙の数を表しているが、一般に単語の綴り(form)とその主な意味(meaning)を結びつけることができるのに十分な知識と解釈される。一方、後者は、ある単語に対する知識の度合いを指す場合と前者の知識を含めた発音、形態素、文法、連語関係や使用域等の語彙の総合的知識を指す場合がある (Schmitt , 2011)。

関連する別の区分の仕方として、リスニングやリーディングの活動の中で語彙を理解する受容(receptive)語彙力とスピーキングやライティングの活動の中で語彙を使用する発表(productive)語彙力があるが、語彙力の異なる側面を表しているとも言えるだろう。

Milton (2009)によると、発表語彙力を測定するテストの中にも簡易かつ迅速にプレイスメントを行う目的で利用することが可能なものもあるが、Laufer and Goldstein (2004)は、受容語彙力テストの方が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、さらには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。

Read (2000)は、Meara 等が開発した受容語彙力テスト(Eurocentres Vocabulary Size Test: EVST)がプレイスメントの目的で優れた結果を残した Meara and Jones (1988)の研究等に言及している。このテストは、単語のリストを受験者に提示し、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るが、過剰申告を避けるために存在しない単語をリストに含めて、その単語を選んだ場合は、減点される。Alderson(2005)も DIALANG テストの Version 1 に含まれている類似形式の語彙力テスト(Vocabulary Size Placement Test: VSPT)が読解、聴解、ライティング、文法テストとの間にかなり高い相関が確認できたとしており、プレイスメントテストとして有効に機能していることを示している。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 (Mochida & Harrington 2006; Eyckmans, Van de Velde, van Hout, & Boers, 2007; Stubbe 2012; Pellicer-Sanchez & Schmitt 2012)。

その他に、Nation (1990)の Vocabulary Levels Test(VLT)が日本人高校生や大学生のプレイスメントに活用できることを示した Beglar and Hunt(1999)、望月語彙テスト(MVST)を日本の大学のプレイスメントに活用して、ラッシュモデルや潜在ランク理論を使って分析を行った小泉・飯村(2010)、日本人大学生及び大学院生と英語を母語とする大学院生に Nation (2006)に基づく Vocabulary Size Test (VST)を実施して、ラッシュモデル等の分析を使って、その信頼性と妥当性を検証している Beglar (2010)の研究等が注目される。VLT は、英単語と英語定義の組み合わせ、MVST は英単語と日本語訳や日本語定義の組み合わせ、VST は、英

単語とその単語を含む文を見て単語の定義を選択肢から選ぶ方式と、それぞれ問題形式は異なるが、日本の EFL 環境で Yes/No 形式を含めた受容語彙力テストが十分に機能して、ラッシュモデルや潜在ランク理論の分析の有用性も示唆されていると言えよう。

2. 2. 規準設定法のモデル

規準設定の方法は数多く存在するが、大まかにテスト中心のモデルと受験者中心のモデルのいずれかに分類されることが多い (大友,2008)。受験者中心のモデルについては、個々の受験者の能力についてよく知っている規準設定の評定者がいない場合は不適当であり、テスト中心のモデルについては、境界水準の受験者の個別のテスト項目の正答確率や項目群への正答数等を予測して評定しなければならない(Pitoniak and Morgan, 2012)。本研究については、入学したばかりの 1 年生のクラスのプレイスメントの目的で行ったテストであるため、前者のモデルの条件を満たすことはできないと言って良いだろう。後者のモデルについては、可能性はゼロではないが、特別な評定者訓練を受けない限り、テストが実施される前に意味のある予測を行うことは極めて難しい。

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、年度や学期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。

いずれにしても従前の規準設定法は、綿密な計画な下に実施されても、人間の判定に基盤を置く恣意的なものだと Lissitz (2013)は述べ、将来の規準設定法として、潜在クラス分析(latent class analysis: LCA)を応用した混合ラッシュモデル(mixture Rasch model: MRM)等に代表されるに統計的解決法を提唱している。MRM は、「複数の潜在的な母集団を含んだテストデータを分析するため、ラッシュモデルと潜在クラス分析 (latent class analysis: LCA) モデルを統合した」モデルであり、テストデータと主観的な審査員の判定を融合した結果を導くことができる分類法に基づく規準設定手続きのモデルとして、近年、様々な研究が行われている (Rost & Langeheine, 1994; Cohen, Wollack, Bolt & Mroch, 2002; Jiao, Lissitz, Macready, Wang & Liang, 2011; Lee & Chen, 2011; Templin & Jiao, 2012 等)。

MRM は数千人以上の大規模試験の分析には活用が期待できるが、あいにく、単独の学科規模で実施するプレイスメントテストのような数百名以下のデータ分析には適応しない。しかしながら、ラッシュモデルと潜在ランク理論を融合して、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考慮に入れることによって、分割点のより合理的な設定を探ることが可能になるかもしれない。

3. 研究方法

ある大学で開発され、数年間実施された受容語彙力テストのデータの一部について、ラッシュモデルや潜在ランク理論等の統計手法を使って分析を行った。

3. 1. 被験者

ある日本の大学の 1 年生の英語必修プログラム受講クラスを決定するために実施された受容語彙力テストの受験者 151 名を主な分析対象とする。受験者の中には相当数の留学生が含まれていたが、どの受験者が留学生であったか、何人の受験者が留学生であったかは完全には特定できない。受験者の能力層は、その後の授業等を通じての観察やコミュニケーションから、実用英語検定準1級合格以上から4、5級程度まで分布していたことが予測される。特に留学生間の習熟度の差は顕著だった。

さらに受験者の心理を探るため、上記の受験生とは別の、日常的に英検や TOEIC を学習している大学生 12 名(日本人 10 名、留学生 2 名)に、テストを受験し、アンケートと面接の質問に回答してもらった。

3. 2. テスト

先行研究で示した議論と同様、プレイスメント計画を練る中で、受容語彙知識の広さ(サイズ)を測定するテストが適切だと考えたが、当該大学のニーズを十分に満たす外部テストがないとの判断から、下記の観点から、大学(学科)独自で、目的のプレイスメントテストを開発することになった。

1)対象受験者の能力水準 – 既存のプレイスメントテストは、対象とするすべての学生の水準に対応しているとは言えない。特に当該学科においてほとんど英語を学習したことのない留学生を含む初級学習者と上級学習者を単一テストで測定することは困難である。

2)テストの所要時間 – ペーパー実施する外部テストの多くが相当数の項目を有し、長時間の解答時間を要する。初級学習者にとってはまったくわからない問題に対峙し、その後の学習意欲の喪失につながることも少なくない

3)指示文・選択肢の使用言語 – 日本語の場合は、日本語訳能力に直接的、間接的に依存する問題に対して、留学生の多くが、問題文自体を理解できても効果的に対応できないことが少なくない。逆に英語の場合は、初級学習者が問題の趣旨を理解できなかったり、問題を理解できても選択肢の意味解釈がスムーズにできず、限られた所要時間の負荷も大きくなってくることもある。

4)正確かつ簡便な実施 – 語彙力テストは、長文読解形式や他の統合的テストに比べて、短時間で解答することが可能であり、作文テスト分析の多くで問題となる採点の主観性・恣意

性の可能性が低く、面接テストのような多くのパフォーマンステストと異なり、一斉実施が可能である。初級学習者の負荷も低く、他のプレイスメントテストに比べて彼らの心理的負担が軽減されることが期待される。

5)意味と形の関係 – 単語の意味と形(綴り)の関係についての理解度を測定するのが最も基本的な受容語彙サイズのテストであるが、Yes/No 形式のテストでは、「理解している」と受験者が誤って解釈している可能性もあり、理解の度合いもはっきりとはわからない。選択肢が英語定義のみの VLT、VST や日本語訳のみの MVST では、1)～4)で述べたような問題が生じる。

以上の理由から、対象プログラム受講者集団により広範に対応することが期待できる受容語彙サイズテストを開発することとなった。当時インターネット上に公開されている日本人英語学習者向けの語彙リストはそれほど多くなく、利用可能なものも作成者の許可が必要なものが多かった。教育・研究対象で自由に活用することを認めていた北海道大学英語語彙表(以降、HEV)の便宜上の利点も大きかったが日本人英語学習者向けのリーダビリティの開発においても、英語圏で開発された語彙リストに基づくリーダビリティ指標よりも HEV に基づくリーダビリティ指標の方が日本人英語学習者に適応していることを Norizuki (2004)が報告しており、適切な水準を選ぶことで、目的に照準化されたプレイスメントテストが開発できると考えた。

HEV は、第 1 水準(中学校必修レベル 786 語)、第 2 水準(高校必修レベル 1778 語)、第 3 水準(大学受験レベル 2096 語)、第 4 水準(大学基本レベル 1520 語)、第 5 水準(大学上級レベル 1274 語)の 7454 語で構成されるが、プレイスメントの目的から、第 1～3 水準の単語を用いた測定が妥当と考え、第 1 水準 16 問×2、第 2 水準 16 問×2、第 3 水準 16 問の 80 問でテストを開発し、2005～2008 年度にかけて実施した。

当時、受験者の中には留学生も相当数含まれていたため、指示文は日本語と英語を併用し、選択肢は同義や類義の日本語及び同義・類義や関連する意味を有し、できる限りテスト項目の単語と同一かそれよりも低い HEV 水準の英単語を併記した。テスト実施前には見本用紙を使い、解答手順を説明し、よく理解できていない受験者には、監督補助の教員が手順を説明し、全員が解答手順を理解していることを確認してから、所要時間「30 分」のテストを開始した。テストに対しての受験者の心理的不安を軽減するため、「テスト」と呼ばず、「英語基礎能力調査 (Survey of Core English Language Proficiency：SCELP)」の名称の下に実施した。

3. 3. 分析

SCELP のデータを使って、規準設定におけるラッシュモデルと潜在ランク理論の有用性を探るために、4 つの研究課題を掲げることとした。

1. SCELP 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)
2. SCELP の項目や受験者の解答様式は難易度や能力水準から予測される結果とどの程度適合していたか。(項目・受験者の解答適合度)
3. ラッシュモデルと潜在ランク理論の分析手法を用いることで、SCELP のデータからいかにして説得力のある分割点設定を行うことができるか。(ラッシュモデルと潜在ランク理論を使った分割点設定)
4. SCELP のような受容語彙力テストは、学習者の総合的英語習熟度水準や問題の把握や、それを基にした診断的フィードバックにどのように活用することが可能か。(学習者の総合的英語習熟度との比較及び診断的フィードバックの可能性)

分析は Excel 2010 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 3.75.0 (Linacre, 2012)、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 を用いた。

4. 結果

分析の結果と解釈について、3.3.節で述べた 4 つの研究課題に分けて、以下、4.1.～4.4.節で論じていくこととする。

4. 1. 難易度と語彙レベルの関係

テストは、平均が 80 問中 54.2 点(67.8%)とかなり高く、中心的傾向の他の指標も類似の値を示しているが、得点幅は、様々な学習背景の相違の影響もあって、最高点 79 点(98.8%)から最低点 14 点 (17.5 点)までかなり広く分布している。信頼性係数は .949 と高い数値を示している。

表 1
SCELP の基本統計量

受験者数	項目数	素点平均	素点最頻値	素点中央値	標準偏差	最高点	最低点	KR20
151	80	54.2	56	56	14.6	79	14	.949

表 2 は各項目の正答率と異なる語彙表の語彙水準や HEV 元来の第 1～3 水準の 3 段階区切りと SCELP 作成のために区分した第 1 水準(各 16 項目 2 段階)、第 2 水準(各 16 項目 2 段階)、第 3 水準(16 項目 1 段階)の 5 段階水準別の場合の相関を示している。比較した指標の中で HEV5 が最も高く、しかも.7 を超えるかなり高い相関を示していることから、SCELP 作成時の語彙レベル区分が妥当であったと言える。

表 2
正答率(項目容易度)と語彙レベルの相関

	HEV3	HEV5	J8000	GS-AW
HEV5	.954			
J8000	.644	.586		
GS-AW	.641	.597	.523	
正答率	-.706	-.737	-.617	-.528

*HEV3: HEV の第 1&2(各 32 項目)第 3 水準(16 項目)の語彙別に 3 分割
*HEV5: テストを第 1&2 水準(各 16 項目×2 段階)、第 3 水準(16 項目)の語彙別に 5 分割
*J8000: JACET8000 の語彙レベルで 8 分割
*GS-AW: Healey, Nation and Coxhead (2005)の Range プログラムにより、A General Service List of Words と The Academic Word List の語彙を 3 段階＋リスト外で 4 分割

表 3 は HEV5 の各 16 項目の問題群の正答率を比較しているが、語彙レベルが上がるにつれて、正答率が下がっていることがわかる。レベル 1 の後半項目からレベル 2 の前半項目にかけては、緩やかな減少であるが、レベル 2 の後半からレベル 3 にかけて急激に難易度が増していることが確認できる。

表 3
各問題群 (HEV5) の難易度

問題群	L1: 問 1-16	L1: 問 17-32	L2: 問 33-48	L2: 問 49-64	L3: 問 65-80
正答率	.912	.763	.718	.609	.387

各問題群の難易度分布の差異をより明確に比較するため、ラッシュ難易度をテストの中心(この場合は、受験者能力平均)が 100 になる標準得点 WITs 値に変換した分布を図 1 に示した。WITS 値が高い項目ほど難しく、WITs 値が高い受験者ほど習熟水準が高いことを意味する。5 本の線の箱で囲まれている部分は、データの数を 1:3 に分ける値である第 1 四分位から 3:1 に分ける第 3 四分位 (Quartile) までの幅を示している。レベル 1 の H5-1 とレベル 5 の H5-5 の最高値(Max)と最低値(Min)の値からこの 2 水準間で項目難易度が交わる部分はないが、その他の各レベル間ではかなりの項目が難易度において重複していると言える。

特に表 4 からわかるように、H5-2 の第 3 四分位は、一つ上のレベルの H5-3 のものを上回っていて、第 1 四分位の値も近接していることから、レベル間で難易度に大きな差がないことを明示する結果となっている。数値を詳しく調べてみると、H5-2 の後半 25～31 番の項目は連続で WITs 値が 90 を超えており、このレベルとしてはかなり難しくなっていると言える。その一方で、H5-3 の始まりの 33～37 番は全体の中でも最低値の 58.8 から 80 台の数値にとどまっている。

このため、特定の語彙レベルがどのレベルの学習者の能力水準に合致しているかは、平均値等から述べることができる程度にしか明確にできず、分割点設定に語彙項目の内容を関連付けるためには、難易度順に並べ替える必要がある。

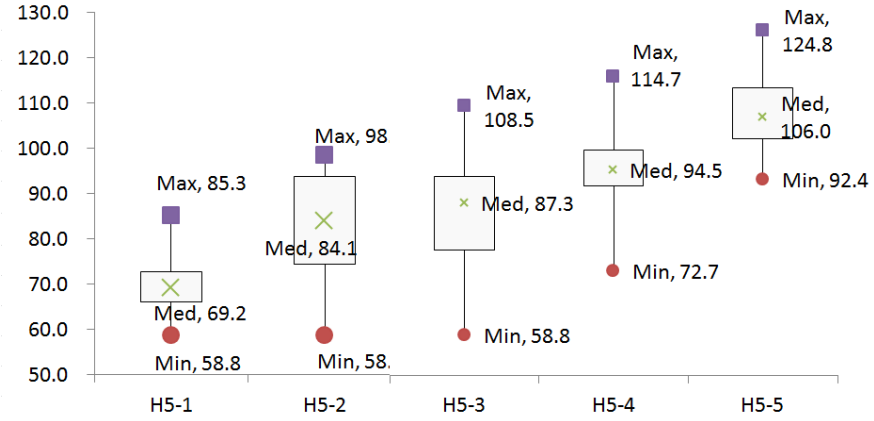


図 1　各問題群（HEV5）の難易度分布(WITs 値)

表 4

各問題群 (HEV5) の第 1 四分位と第 3 四分位(WITs 値)

問題群	H5-1	H5-2	H5-3	H5-4	H5-5
Q1	66.1	74.5	77.2	90.9	101.3
Q3	72.8	93.8	93.1	98.7	112.3

4. 2. 項目・受験者の解答適合度

HEV 水準が上がるにつれて、全体的にテスト項目の難易度も高くなることが確認されたが、以下、個別の項目や受験者の解答が難易度や能力水準に適合しているか検証する。

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)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。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。

Beglar が使用したテストでは、1 項目を除き、「大きな残差 (residual)」が見られたのは 4 名未満の受験者に限られたとされているが、SCELP では、12 名から 18 名の受験者がプラス 2 以上もしくはマイナス 2 以下の残差を示した。SCELP のアンダーフィット項目の 5 項目中 4 項目は中程度の難易度であったが、そのうちの 3 項目(ant, executive, raisin)は Heatley, et al. (2002)の Range のプログラムが基準とする GS-AW 語彙のリスト外であった。HEV 水準からすれば特別に難易度が高い単語とは言えないが、他の国では重要な学習語彙に含まれていない可能性もある。このような状況も反映してか、能力推定値の高い留学生が間違えているケースが多く、逆に「レーズン」のように音声に当てはめることができるとカタカナ語として認識できるのか、能力推定値の低い日本人学生が正解しているケースも目立った。中程度の難易度の 4 項目については、受験者の能力水準(ランク)が上がるにつれて正解率が上がり、正解

の選択肢が最も多くの受験者によって選択されている。アンダーフィット項目の中で、唯一図 2 で示される項目(violate)のみ、正答率が低く(21.2%)、受験者の潜在ランクが上がっても正答率はほとんど変わらず、特定の誤答選択肢の選択率が上がる現象が見られた。

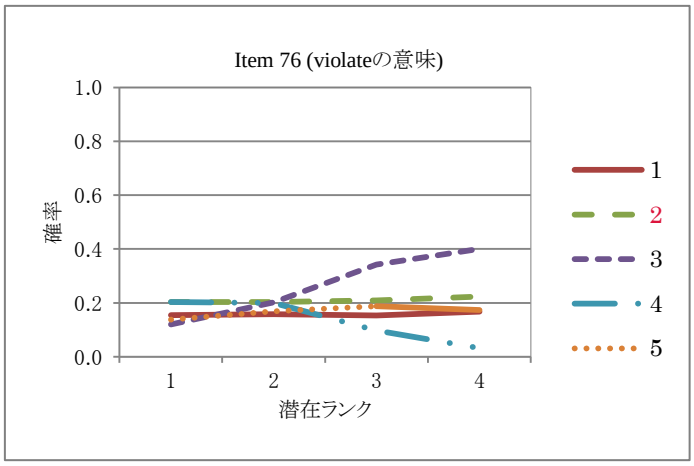


図 2　誤答選択肢が正答選択肢よりも多く選択されたミスフィット項目

潜在ランクが上がるにつれて選択率が高くなっている選択肢は、3 番の「無理に…させる(force)」で、正解選択肢 2 番の「規則を破る(break)」とは明らかに意味的に隔たった内容である。このことから項目内容に問題があったと言うよりは、対象受験者の多くにとって、この単語(violate)が未知の単語であったり、十分に正確な意味が認識されないまま、何らかの理由で特定の誤答選択肢が選ばれたことを示唆していると言えるだろう。しかし、アンダーフィットの直接の原因はいずれも正答確率が低いにもかかわらず正解している 18 名の受験者によるため、予測に反した複雑な解答様式を呈したこのような項目については、修正を検討する価値がある。

Beglar (2010)は、さらにインフィットとアウトフィットの t 値が-2.00 を下回るオーバーフィット項目について点検を行っているが、本テストにおいてオーバーフィット項目は 3 項目(3.8%)のみであった。Beglar は、「オーバーフィット項目が 5%未満の場合は、項目難易度と能力推定値に大きな影響は及ばない」と解釈しており、オーバーフィット項目の数については問題なさそうである。

一方、小泉・飯村(2010)は、項目と受験者のインフィット平均平方値が 0.70～1.30 の範囲を超える場合をミスフィットと呼び、自らのテストデータにおいて点検を行っている。本研究のテストも同じ基準で見てみると、特に問題となるアンダーフィット(>1.3)は、2 項目(2.5%)、11 名(7.3%)、オーバーフィット(<0.70)は、0 項目、6 名(4.0%)であった。受験者のミスフィット数はやや多いが、小泉・飯村のデータと比較して、そん色なく、簡易テストであることを考えれば、まずまずの結果であったと言える。

4. 3. ラッシュモデルと潜在ランク理論を使った分割点設定

ラッシュモデルを活用することで、SCELP の項目難易度と受験者能力を共通の間隔尺度上で比較することが可能だが、SCELP の項目は必ずしも難易度順に配列されておらず、HEV の上位水準の項目が下位水準の項目よりも易しくなる場合もある。そこで、潜在ランク理論のランク区分を参考にしながら、ラッシュ能力推定値を軸に分割点を設定する方法を模索することとした。

表 5 は、潜在ランク理論で、テストを一様分布の潜在ランク数 5 で分析した際に、各潜在ランクに位置する項目と受験者の数を示している。項目については、各潜在ランクに所属する受験者が正答する確率を求め、その値が最も基準値(Exmetrika では.5)に近づく地点のランクを示しているが、受験者は、所属する確率(事後所属確率)が最も高くなる地点のランクが付与されている(植野・荘島 2010)。項目においては、最も低いランクのランク 1 に半数近くが位置し、ランク 3、4 は少なくなっているが、受験者はランク 3、4 が多く、ランク 5、1 が少なくなっている。

表 5
項目及び受験者の潜在ランク数　　(潜在ランク数 5、一様分布の分析の場合)

	R1(初級)	R2(初中級)	R3(中級)	R4(中上級)	R5(上級)
項目	38	15	12	6	9
受験者	22	31	35	35	28

潜在ランク理論では、同じ正答率の受験者や項目が必ずしも同じ潜在ランクに推定されるとは限らない。表 6 は、同じ正答率を示した 6 人の受験者の解答様式、付与された異なるランク、各潜在ランクに所属する確率を示しているが、ランク間の境界水準においては、素点や正答率、ラッシュモデルの分析と異なり、同じ正答数であっても、識別力の高い項目により多く答えた受験者のランクが相対的に高くなる傾向があるとされている。ラッシュ推定値や素点でも機械的な境界区分はできるが、このようなランクの確率が標示されることで、異なる視点を加味したクラス編成を行い、編成後も受験者の習熟度過程を観察していくことが期待される。

表 6
境界ランクの受験者例

	正答数	正答率	LR	R1	R2	R3	R4	R5
受験者 A	59	.738	4	.000	.000	.285	.707	.008
受験者 B	59	.738	4	.000	.000	.377	.621	.002
受験者 C	59	.738	4	.000	.000	.476	.523	.000
受験者 D	59	.738	4	.000	.000	.481	.518	.001
受験者 E	59	.738	3	.000	.001	.840	.158	.000
受験者 F	59	.738	3	.000	.005	.903	.092	.000

ラッシュモデルと潜在ランク理論の手法の併用による分割点設定方法を探るため、次のような手順を取ることとした。

①受験者能力(WITs 値)を数値の高いほうがリストの上に来るように並べ替える。以後の比較参照のため、項目難易度(WITs 値)も同様に並べ替える。

②潜在ランク(1～5)を各受験者能力に付与し、①の並べ替えの際に WITs 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、WITs 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、ランク 5 と 4 の受験者グループ境界付近は、表 8 のような状況であることが確認された。なお、全受験者リストで受験者 1 よりも上に位置する受験者は、全員 WITs 値が 115.5 以上でランク 5 に所属し、受験者 14 よりも下に位置する受験者は、全員 WITs 値が 110.3 以下でランク 4 以下に所属している。

このことから、WITs 値と潜在ランクを基準に規準設定を行う場合は、最上位クラス(以後、便宜的に「上位クラス」と呼ぶ)の受講生の数を最も絞り込む場合は、受験者 1 よりも WITs 値が高い受験者がその対象となるが、潜在ランク 5 の受験生が存在する最も WITs 値の低い地点の受験者を含めるならば、受験者 14 までが上級クラスに選ばれることになる。

植野・荘島(2010)によると、識別力の高い項目に正答する数が増えると潜在ランクが高く推定され、逆にそのような項目に誤答する数が増えると潜在ランクが低く推定される傾向があるとされるが、この境界水準については、正答項目の識別度平均値や誤答項目の識別度平均値の比較を通じて顕著な特徴を確認することができなかった。

表 8
潜在ランク 5/4 境界域受験者の能力値指標と正誤別項目平均識別度の比較

	正答率	WITs	ランク	R_4 確率	R_5 確率	正答項目識別度平均	誤答項目識別度平均
受験者 1(5/4)	87.5%	114.0	5	.248	.752	.459	.370
受験者 2(5/4)	87.5%	114.0	5	.268	.737	.456	.393
受験者 3(5/4)	87.5%	114.0	5	.285	.715	.453	.415
受験者 4(5/4)	87.5%	114.0	5	.289	.710	.453	.416
受験者 5(5/4)	87.5%	114.0	4	.558	.441	.454	.407
受験者 6(5/4)	86.3%	112.6	5	.200	.800	.461	.367
受験者 7(5/4)	86.3%	112.6	5	.333	.667	.459	.380
受験者 8(5/4)	86.3%	112.6	5	.351	.648	.457	.392
受験者 9(5/4)	86.3%	112.6	4	.664	.332	.452	.426
受験者 10(5/4)	85.0%	111.5	5	.421	.578	.462	.368
受験者 11(5/4)	85.0%	111.5	5	.455	.544	.458	.389
受験者 12(5/4)	85.0%	111.5	4	.635	.362	.455	.412
受験者 13(5/4)	85.0%	111.5	4	.643	.356	.463	.364
受験者 14(5/4)	85.0%	111.5	4	.693	.306	.462	.364

一方、表 9 は、表 8 と同じ基準で、潜在ランク 4 と 3、3 と 2、2 と 1 の境界域における能力値指標と正誤別の項目平均識別度を比較したものである。WITs 値と潜在ランクの対応関係は、上級クラスの編成と同様に一律ではないが、下位ランク境界域になるほど、潜在ランクの変動域は小さくなっている。

さらに、正答項目と誤答項目の識別度を比較すると、ランク 4 と 3 の境界域では、ランク 3 が付与された受験者は誤答項目の識別度が正答項目の識別度よりも高い値を示したり、より近似した値となり、ランク 3 と 2、2 と 1 の境界域ではすべての受験者の誤答項目識別度が正答項目識別度よりも高くなっているが、その差は境界域のより低いランクが付与された受験者(例、3 と 2 の境界域では 2 の受験生)のほうが大きくなる傾向が確認できる。

この結果から、変動域が大きいランク 4/3 では、上級クラス編成と同様に WITs 値の 102.5 の受験者までを「中上級」クラスに含め、変動域が小さい 3/2、2/1 では、潜在ランクが変わる地点で「中級」、「初中級」、「初級」クラスを分ける弾力的な分割設定案も考えられる。しかし、同じ得点で異なるクラスに配置された受験者、プレイスメントの最終決定者、結果に関係する当事者等がその決定に異議を唱えた場合、説得力のある決定理由を説明することは難しい。

表 9

潜在ランク 4/3、3/2、2/1 境界域受験者の能力値指標と正誤別項目平均識別度の比較

	正答率	WITs	ランク	R_{n-1} 確率	R_n 確率	正答項目識別度平均	誤答項目識別度平均
受験者 1(4/3)	75.0%	103.4	4	.158	.835	.457	.421
受験者 2(4/3)	75.0%	103.4	4	.334	.660	.457	.422
受験者 3(4/3)	75.0%	103.4	3	.721	.277	.447	.453
受験者 4(4/3)	73.8%	102.5	4	.285	.707	.465	.399
受験者 5(4/3)	73.8%	102.5	4	.377	.621	.455	.430
受験者 6(4/3)	73.8%	102.5	4	.476	.523	.456	.426
受験者 7(4/3)	73.8%	102.5	4	.481	.518	.457	.425
受験者 8(4/3)	73.8%	102.5	3	.840	.158	.449	.448
受験者 9(4/3)	73.8%	102.5	3	.903	.092	.451	.442
受験者 1(3/2)	62.5%	95.4	3	.208	.791	.441	.461
受験者 2(3/2)	62.5%	95.4	2	.572	.427	.432	.474
受験者 3(3/2)	62.5%	95.4	2	.787	.211	.433	.474
受験者 4(3/2)	62.5%	95.4	2	.910	.086	.437	.467
受験者 1(2/1)	50.0%	88.0	2	.159	.841	.428	.468
受験者 2(2/1)	50.0%	88.0	1	.895	.105	.421	.475

本研究で扱った受験者に関しては、テストデータ以外に判定する資料がないため、上記の①～③の規準設定の手続きに続いて、以下の④、⑤の手順を規準設定の最終手続き案として提示し、さらに⑥を事後点検として実践することで、「テストが作成者の計画の通りに所定

の決定を行う目的で活用されている度合い」を意味する決定妥当性(decision validity)(Brown, 1996; Brown & Hudson, 2002)について検証することとした。

④WITs 値とランクが両方変わるところに分割点を設定する。ランクが変動する区域(例、…2、1、2、1、1、2、1、1…)がある場合は、現実的に対応できるクラスサイズを考慮に入れて、(1)変動が完全に終息する手前の地点 (この場合 6 番目) か、(2)ランクが下がる手前の地点のいずれか(この場合 1、3、6 番目)の WITs 値を(暫定的な)分割点とする。

⑤各能力編成クラスに対応する WITs 項目難易度の項目群をまとめる。特定レベルの項目数が少ない場合は、将来の補充や新しいテストの開発の際に、検討する。

⑥アンケートや他の評価データがある場合は参考にして、特にランクの変動区域で、入れ替えが必要な受験者がいないか確認する。補助データがない場合は、④の分割方法のいずれかを採用し、クラス編成後に、診断的指導が必要な学習者に適宜対応したり、同様の学習集団にテスト実施と他の評価手段を併用実施して、決定妥当性を検証する。

①～⑤の手続きに従って規準設定を行った決定案を実際に観測された値を基にまとめると、表 10 の結果となった。5 つの能力編成クラスは、人数的にややばらつきがあるが、①～④の論理的な手順に則って指導可能なクラスサイズに無理なく分割されていると考えれば、手続き的には大きな支障はないと言えるだろう。

人数の若干の不均衡についても、学習困難者も含まれる初級クラスでは少人数のほうが指導しやすいと考えれば、初級クラスの受講該当者が少なくなっていることはさほど問題でないだろう。上級クラスも少人数で早い進度で進めることが望ましいならば、33 名を同一習熟度水準の 2 グループに分けて指導する方法も検討できるだろう。

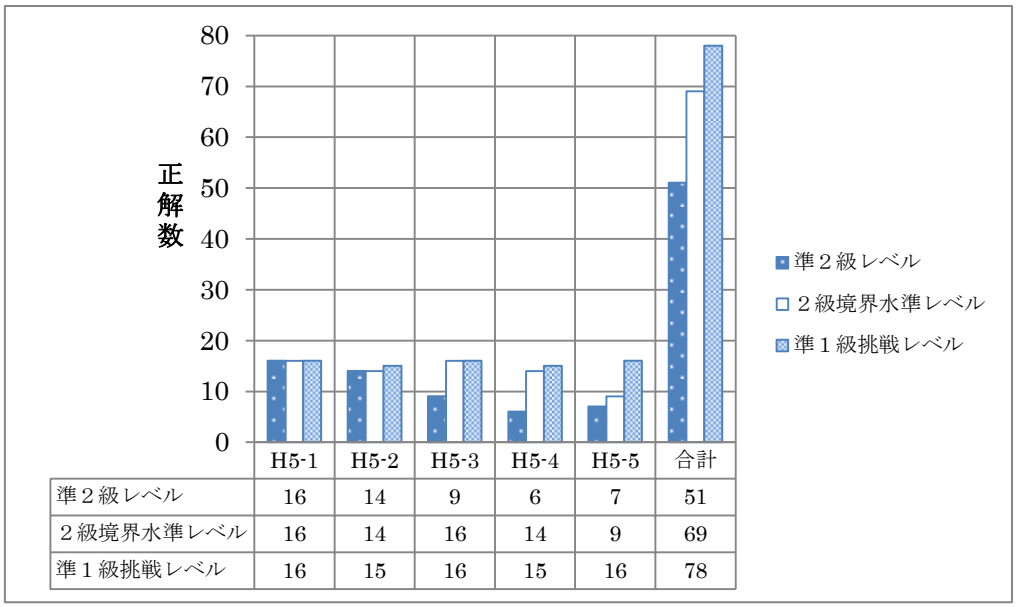
項目難易度は初級が半数弱を占め、上級、中上級は特に少ないことがわかった。受験者がテストに対して不安感を感じることなく、アンケートに回答するような感覚で解くことができるような易しい項目作成に焦点を置いたこともあり、総体的にプレイスメントの機能に大きな問題はなかったと考えられる。しかし、上級と中上級、中上級と中級の分割点設定の精度を高めたり、習熟度の高い学習者についても問題点を把握し、効果的な診断的フィードバックの提供を考慮に入れると、もう少し WITs 値の高い項目も含めることが望まれる。

表 10
観測値に基づく規準設定案(すべての水準で④の(1)の分割方法を採用した場合)

	正答率 (能力)	正答率 (項目)	WITs (能力)	WITs (項目)	ランク (能力)	ランク (項目)	該当 受験者数	該当 項目数
上級	99-85%	11-28%	140-111	125-111	5	5-4	33	6
中上級	84-74%	33-44%	110-103	109-103	4	5-4	33	8
中級	73-63%	46-66%	102-95	102-95	3	4-2	35	14
初中級	60-50%	62-52%	94-88	94-89	2	2-1	29	16
初級	49-18%	74-97%	87-66	87-59	1	1	21	36

4. 4. 学習者の総合的英語習熟度との比較及び診断的フィードバックの可能性

SCELP を受験した 151 名はすでに大学を卒業しており、他の客観的な評価資料は残っていないため、本研究の分析を行った当時、英検や TOEIC を学習していた日本人学習者 10 名と留学生 2 名の協力を得て、SCELP の受験、アンケートへの回答、一部の学習者にはアンケート結果の説明も依頼した。その結果、それぞれの学習者の理解している語彙のレベルと特徴が明らかになった。すべての学習者が表 10 の上級から中級の正答率を示し、総合的習熟度が高い学習者ほど正答率は高く、HEV 上位水準の正答率も高いことが全体的な傾向として確認できた。その一方で、中上級以下の学習者は、正解した項目の中にも(消去法選択によるため)ほとんど意味を理解できていなかったり、習熟度を問わず、下位水準の単語でも意外に難しいと感じられたものが多く存在することが確認できた。また、留学生の誤答の中には、学習語彙の相違を如実に示すものや、意味は理解しながらも、日本語の正答選択肢の意味を誤って解釈したために正答できなかったケースも確認できた。



	H5-1	H5-2	H5-3	H5-4	H5-5	合計
準２級レベル	16	14	9	6	7	51
２級境界水準レベル	16	14	16	14	9	69
準１級挑戦レベル	16	15	16	15	16	78

図 3　学習者の習熟度と SCELP の解答様式の比較

図 3 は、10 人の日本人学習者中、①準 2 級合格習熟水準の学習者、②2 級合格境界水準の学習者、③準 1 級合格に取り組む学習者の解答様式を比較したものである。

①学習者は、分析した年度の SCELP 平均点よりも若干低い合計得点であったが、学習経歴から見て、準 2 級には合格できる水準にあったと考えられる。①、②学習者と異なり、H5-3(HEV 第 2 水準の前半)水準から正答率が大幅に下がり、H5-4、H-5-5 水準では正解している項目もほとんど理解できていなかったことが、アンケート及び面接の結果から確認できた。

②学習者は、2 級合格境界水準にあって、H5-4 までは、H5-2 の allow-permit と excuse-pardon の組み合わせを間違えたり、下位レベルの語彙にも解答にやや自信がないものもあるが、H5-4 水準の後半から徐々に未知の単語や理解が不十分な語彙が増えるようである。H5-5 水準の語彙に対しては、ほとんど十分な理解ができていなかったようである。

③学習者はほとんどすべての項目に対して何らかの受容的知識を有しているようであるが、日本語訳の「許可」の「許」のイメージから H5-2 の excuse を permit と誤って結びつけたり(allow の選択肢としても permit を選択して正解)、どちらかというと文章よりも会話でよく使う indeed に最もなじみがなく、間違った解答をしていることが確認できた。

5. 考察

4 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. 難易度と語彙レベルの関係

SCELP は HEV の第 1、2 水準から各 32 項目、第 3 水準から 16 項目の計 80 項目で構成されるが、第 1、2 水準は作成過程でやや易しめの前半 16 項目、やや難しめの後半 16 項目になるように意識された配列だったため、16 項目ずつ 5 段階の難易度に分かれるように意図された受容語彙力テストであると言える。SCELP の各項目への正答率と項目群の 5 段階の区分との相関は、項目群区分の段階が上がるにつれて項目の正答率が下がることが顕著な -.738 を示した。この値は他の語彙リスト区分と比較しても高い。テスト全体としては、意図された難易度構成になっていると言える。

しかしながら、最も難度が低い HEV1 水準と最も難度が高い HEV5 水準との関係を除いて、WITs 値の分布が重なっている部分があり、重複の度合いが大きく、十分に意図された相対的な難易度構成になっていない箇所も見受けられた。

将来的に、ラッシュモデルの分析をより具体的なテスト内容の分析に結び付けていくためには、語彙水準と項目難易度の対応関係がより明確なテストを開発し、受験者能力との関係を追究していくことが望まれる。

5. 2. 項目・受験者の解答適合度

SCELP は項目や受験者のミスフィットの数や比率から見て、Beglar (2010)が使用した VSP や小泉・飯村(2010)の MVST と比較してもそん色なく機能していることがわかった。しかし、項目 76 のように予期しない応答様式を示したり、ミスフィットにつながる予測に反する誤答の中には、受験当時日本語能力がそれほど高くなく、英語の習熟度が高い留学生によるものがかなり多かった。語彙の学習優先順位は国によってかなり異なるため、その影響も考えられるが、彼(女)らが単語の意味が理解できていながら、間違ってしまった可能性も否定できない。

SCELP は、様々な英語習熟度水準や日本語力の高くない留学生にも対応するため、英語と日本語を選択肢に併用したが、英語の定義解釈に慣れていない大多数の受験生を考えて、英語は関連する単語を提示したものの、関連性がとらえにくかったり、文脈がないため、単語の別の語義をイメージして誤答選択肢と強引に結び付けてしまった可能性もある。

当該テストは現在では使用されておらず、データの一部しか受験生が特定できない状況であるが、解答適合度の低い受験者や項目の問題点を早期の診断的分析で明らかにして、受験者への適切なフィードバックやテスト項目の修正、さらには新たなテスト開発に改善点を反映していくことの教育的意義が示唆される結果と言えるだろう。

5. 3. ラッシュモデルと潜在ランク理論を使った分割点設定

項目難易度と受験者能力を同一の間隔尺度上で直接比較することができるラッシュモデルと、項目と受験者が所属する潜在ランクを示すことで分割点設定につながる段階別評価を導く潜在ランク理論を併用することで、合理的な手順で分割点設定を行うことができることが確認できた。従来の多くの規準設定法と異なり、教科担当の評定者が多数いなくても、何回も協議を重ねるだけの時間的余裕がなくても、実施することが可能である。教科担当者の役割は統計データをどのように評定につなげるかを検討し、可能な場合は、アンケートや面接を行って、受験者の技能や知識の状態をより明確に把握して、最終決定につなげることが望まれる。非テスト情報の収集やプレイスメントの手順としては、Brown (1996)によるハワイ大学の英語教育プログラムの詳細な記述が参考になる。

統計的には、本研究の最終的な一連の手続きの中では活用しなかったラッシュモデルの受験者や項目の分離指標、測定誤差、潜在ランク理論の目標潜在ランク分布と付与されるランクと応答様式の関係等、規準設定の視点から、ラッシュモデルと潜在ランク理論の応用について研究を続けていくことが望まれる。

しかしながら、ラッシュモデルと潜在ランク理論を使った分割点設定を実践化できるかどうかの最大の鍵は、「統計学や心理測定学の専門スタッフが利用できる度合いが規準設定法の選択を考慮する際に重要だ」とした Pitoniak & Morgan (2012, p.356)の指摘に帰結するように思える。統計ソフトを利用することで、両モデルを併用した分割点設定が比較的簡便かつ適切にできることを確証して、実践化に向けての次へのステップへと結びつけていくことが大きな課題となる。

5. 4. 学習者の総合的英語習熟度との比較及び診断的フィードバックの可能性

SCELP は単に受容単語力を測定するテストとしてではなく、総合的英語習熟度を予測できるプレイスメントテストとして利用できることが、少人数の学習者に SCELP を実施した結果、大まかに確認することができた。また、習熟度が高くなるにつれて、より難易度が高い単語への正解率や理解度が高まることもわかった。その一方で、テスト直後に実施したアンケートと面

接の結果から、SCELP の問題点も探ることができた。

SCELP の問題点の一つは、すべての選択式問題に共通する問題であるが、正解している問題が必ずしも理解して正解できているとは限らず、間違った問題よりも理解度が低かったとは必ずしも言えない点である。アンケートで「見たことのない単語」、「よく意味のわからない単語」を特定させたところ、正解した項目番号や選択肢も含まれていることがわかった。面接でそのことを尋ねると消去法で残った 2 つからもっともらしいものを選んでいたり、まったく根拠もなく偶然正解を選んでいるケースもあった。

SCELP の 2 番目の問題点は 1 番目のものと強く関連するものであるが、短時間でより多くの単語の意味と形の関係の理解度をチェックするために考案した各ブロック 4 つの単語の意味を 5 つの選択肢の内容と組み合わせる方式にあったと言える。この問題は作成過程である程度想像できたが、想像していた以上に大きな影響だった可能性が、面接を行って判明した。比較的英語習熟度が高く、学習意欲が高い受験者でも、正答の見極めが難しい場合は、語彙知識からの類推ではなく、単純な当て推量で多くの問題に解答している。SCELP では各ブロック 3 つの問題の選択肢を選ぶと、残りの問題の選択肢はまだ選んでいない 2 つに限定される。問題作成過程では、このような状況にあっても、受験者はすでに選んだ選択肢も含めて、再度問題と選択肢の組み合わせを全体的にチェックする手順を踏むだろうと考えていた。しかし実際に面接した大半の受験者は、その余裕がなかったのか、合理的な判断ができずに、ブロックの大半の組み合わせがずれてしまう状況も見受けられた。

ある問題の不正解・正解が次の問題の応答に影響を与えることは、局所的独立性の観点からも決して望ましいことではない。VST や MVST も含めて組み合わせ方式の問題形式では項目間の完全な独立は望めないが、選択肢の数を増やしたり、問題の数を減らすことで、受容語彙力以外の要因がテスト結果に影響を与える度合いを軽減化していくことが、将来の類似のテスト開発においては求められるであろう。

6. 結論

本研究の結果から、4 つの研究課題が検証されたが、その成果を一文でまとめると、受容語彙力テストの SCELP は、意図とした構成概念を正確かつ適切に測定し、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適切かつ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。

一方、本研究の問題点及び課題として、SCELP の基準となった HEV の語彙水準が個別項目レベルでは必ずしも能力水準と一定の関係にないこと、項目形式の制約、ラッシュモデルや潜在ランク論に基づく規準設定法の確立と実践的普及の問題点等を指摘した。

上記の問題点の克服に加えて、より規準設定の理念を反映した新たな受容語彙力テストの開発を、今後の研究課題に掲げていきたいと考えている。

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ができれば、SCELP 以上に明確な理念に基づいた規準設定を行うことが可能になるだろう。

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.

Bramley, T. (2010). *Locating objects on a latent trait using Rasch analysis of experts' judgments.* A paper presented at the conference "Probabilistic Models for Measurement in Education, Psychology, Social Science and Health," Copenhagen, Denmark (June, 2010). Retrieved from http://www.cambridgeassessment.org.uk/ca/digitalAssets/186233_TB_locating_ objects_Rasch2010.pdf

Brown, J.D. (1996). *Testing in language programs*. Upper Saddle River, NJ: Prentice Hall Regents.

Brown, J.D., & Hudson, T. (2002). *Criterion-referenced language testing.* Cambridge: Cambridge University Press.

Cohen, A.S., Wollack, J.A., Bolt, D.M., Mroch, A.A. (2002). *A mixture Rasch model analysis of test speededness.* A paper presented at the annual meeting of the American Education Research Association, New Orleans, LA. Retrieved from http://www.psyc.jmu.edu/assessment/research/pdfs/JPM%20NCME%20Paper%20SP%2008.pdf

Eyckmans, J., Van de Velde, H., van Hout, R., & Boers, F. (2007). Learners' response behaviour in Yes/No vocabulary tests. In H. Daller, J. Milton & J. Treffers-Daller (Eds.) *Modelling and assessing vocabulary knowledge*. (pp.59-76). Cambridge: Cambridge University Press.

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.

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.

Linacre, M. (2012). *WINSTEPS Rasch measurement computer program* (Version 3.75.0). Chicago: Winsteps.com.

Meara, P. & Jones, G. (1988). Vocabulary size as a placement indicator. In P. Grunwell (ed.) *Applied Linguistics in Society* (pp.80-87). London: Centre for Information on Language Teaching and Research.

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

Mochida, A., & Harrington, M. (2006). Yes/No test as a measure of receptive vocabulary. *Language Testing, 23,* 73-98.

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

Nation, I. S. P., (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review, 63,* 59-82.

Norizuki, K. (2004). In search of new dimensions for readability for Japanese learners of English. *Bulletin of Shizuoka Sangyo University,6,* 167-180.

Pellicer-Sanchez, A., & Schmidt, N. (2012). Scoring Yes–No vocabulary tests: Reaction time vs. nonword approaches. *Language Testing, 29,1-21.*

Pitoniak, M.J., & Morgan, D.L. (2012). Setting and validating cut scores for tests. In C. Secolsky & D.B. Denison (Eds.) *Handbook on measurement, assessment, and evaluation in higher education.* (pp. 343-366). New York: Routledge.

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

Rost, J., & Langeheine, R. (1997). A Guide through latent structure models for categorical data. In J. Rost & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences* (pp.13-37). Munster, Germany: Waxmann. Retrieved from http://www.ipn.uni-kiel.de/aktuell/buecher/rostbuch/inhalt.htm

Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual.* Basingstoke: Palgrave Macmillan.

Stubbe, R. (2012). Do pseudoword false alarm rates and overestimation rates in Yes/No voczabulary tests change with Japanese university students' English ability levels?

Language Testing, 29, 471-488.

Templin, J., & Jiao, H. (2012). Applying model-based approaches to identify performance categories. In G.J. Cizek (Ed.), *Setting performance standards. (Second Edition)* (pp.379-397). New York, NY: Routledge.

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

大友賢二(監修)(2008).『言語テスト:目標の到達と未到達 vol. 2』, 英語運用能力評価協会.

植野真臣・荘島宏二郎(2010).『学習評価の新潮流』, 東京:朝倉書店.

資料 1

英語基礎能力調査　Survey of Core English Language Proficiency　(指示文一部省略、レイアウト修正)

1～80 の単語とその右側のボックス内の単語を比べてください。各ボックスの 1～ 80 の中から左側にある単語と意味が似ているか関連している単語を一つずつ見つけて、マークカードの番号にマークしてください。それぞれのボックスには左側の単語とほとんど関係のない単語が一つあります。

Look at each of the groups of words numbered 1 to 80. Then look at the words in the box to the right of each group of words. Find one word that has nearly the same meaning or has …..

A. 1～4
1. clock
2. girl
3. hat
4. student

A. 1～4
(1) 帽子 (cap)
(2) 足 (leg)
(3) 学生 (school)
(4) 時計 (time)
(5) 少女 (woman)

B. 5～8
5. change
6. ear
7. ship
8. yellow

B. 5～8
(1) 船 (boat)
(2) 黄色 (color / colour)
(3) 耳 (face)
(4) 一覧 (list)
(5) 変化 (turn)

C. 9～12
9. bring
10. build
11. help
12. learn

C. 9～12
(1) 買う (buy)
(2) 作る (make)
(3) 学ぶ (study)
(4) 助ける (support)
(5) 持って行く (take)

D. 13～16
13. example
14. game
15. place
16. wind

D. 13～16
(1) 風 (air)
(2) 場所 (area)
(3) 例 (case)
(4) 草 (grass)
(5) 試合 (match)

E. 17～20
17. glad
18. real
19. short
20. young

E. 17～20
(1) 忙しい (busy)
(2) うれしい (happy)
(3) 足りない (little)
(4) 本当の (true)
(5) 若い (new)

F. 21～24
21. maybe
22. only
23. quickly
24. usually

F. 21～24
(1) 早く (fast)
(2) 次に (next)
(3) 普通は (often)
(4) おそらく (perhaps)
(5) 唯一の (single)

G. 25～28
25. below
26. during
27. else
28. several

G. 25～28
(1) 他の (other)
(2) ～以来 (since)
(3) いくつかの (some)
(4) ～の間の (while)
(5) ～の下の (under)

H. 29～32
29. allow
30. ant
31. excuse
32. pattern

H. 29～32
(1) 様式 (design)
(2) アリ (insect)
(3) 言い訳 (pardon)
(4) 許可する (permit)
(5) おじさん (uncle)

I. 33～36
33. lip
34. origin
35. photograph
36. stranger

I. 33～36
(1) 起源 (beginning)
(2) くちびる (mouth)
(3) 写真 (picture)
(4) 道 (route)
(5) 知らない人 (visitor)

J. 37～40
37. community
38. doubt
39. law
40. seed

J. 37～40
(1) 種 (plant)
(2) 法律 (rule)
(3) 疑い (question)
(4) 地域共同体 (society)
(5) 視覚 (vision)

K. 41～44
41. elect
42. guide
43. prefer
44. win

K. 41～44
(1) 選ぶ (choose)
(2) 直す (fix)
(3) 指導する (lead)
(4) ～をより好む (like)
(5) 勝つ (victory)

L. 45～48
45. firm
46. harmful
47. international
48. severe

L. 45～48
(1) 害のある (damaging)
(2) 効果的な (effective)
(3) 国際的な (global)
(4) 固い (secure)
(5) 厳しい (serious)

M. 49～52
49. grace
50. indeed
51. nearly
52. preparation

M. 49～52
(1) ほとんど (almost)
(2) 準備 (arrangement)
(3) 優雅 (elegance)
(4) 期待 (expectation)
(5) 真に (truly)

N. 53～56
53. argue
54. cross
55. loan
56. realize

N. 53～56
(1) 議論する (discuss)
(2) 貸す (lend)
(3) 気づく (notice)
(4) わたる (pass)
(5) 変化する (vary)

O. 57～60
57. delight
58. inch
59. speed
60. stuff

O. 57～60
(1) 材料 (material)
(2) 速さ (movement)
(3) よろこび (pleasure)
(4) 解決 (solution)
(5) 2.54 cm (unit)

P. 61～64
61. block
62. publish
63. remind
64. replace

P. 61～64
(1) 入れ替わる (change)
(2) 発明する (invent)
(3) 出版する (print)
(4) ふせぐ (stop)
(5) 気づかせる (tell)

Q. 65～68
65. appetite
66. athlete
67. executive
68. legend

Q. 65～68
(1) 経営幹部 (business)
(2) 食欲 (food)
(3) 運動選手 (sport)
(4) 伝記 (story)
(5) 望遠鏡 (universe)

R. 69～72
69. inclination
70. necessity
71. priest
72. raisin

R. 69～72
(1) 教会の司祭 (church)
(2) 気持ち (feeling)
(3) 放送 (media)
(4) 必要性 (need)
(5) 干しぶどう (grape)

S. 73～76
73. compel
74. suspend
75. transmit
76. violate

S. 73～76
(1) 驚かせる (amaze)
(2) 規則を破る (break)
(3) 無理やり～させる (force)
(4) 伝える (send)
(5) 中止する (stop)

T. 77～80
77. incredible
78. intentional
79. significant
80. sympathetic

T. 77～80
(1) 重要な (important)
(2) 意図的な (planned)
(3) 同情的な (sorry)
(4) 信じられない (unbelievable)
(5) 活気のある (vigorous)

Setting Lexical Standard for CLIL Courses
CLIL における語彙による規準設定[1]

渡部良典
Yoshinori Watanabe

Abstract

Setting a standard for assessing CLIL courses is challenging in two ways. First, CLIL is intended to teach more than two elements at a time in an integrative manner, so it is extremely difficult, if not possible, to identify them separately to assess performance in each of these elements. Second, in order to assess the effectiveness of the course and/or the achievement of the students in the course it is necessary for the assessor to examine first whether the teachers teach what to teach and then if students learn what teachers teach, and only after going through this process it becomes possible to establish a connection between the two. In order to establish the process as a routine component of the assessment procedure of CLIL courses, it behooves assessors to establish a specific set of observable constituents for evaluating teaching, learning and ultimately the entire programme. In order to do so, the best way would involve identifying the vocabulary that uniquely characterises the course. The present paper illustrates a sample of procedure and the product of such an attempt. In so doing, the project is placed in a larger framework derived from the taxonomy of educational objectives, and then proceeds to lexical analyses of classroom observation data.

[1]本稿をさらに詳しく報告した研究の成果は、Profiling lexical features of teacher talk in CLIL courses – The case of an EAP programme at higher education in Japan. *International CLIL Research Journal: Special issue - CLIL in Japan: beyond the European context.*として出版される予定である。同じデータを称したものであるが、読者対象が異なるため、精度等がやや異なるところがある。

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、さまざまな実践力を伸ばすことを目的とした教育原理である。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ひとつである。CLIL の最も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。

この 3 要素を同時に扱うとはいえ、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)。ところが、教科内容も同時に扱うという特色が強調されるあまり言語そのものの位置づけがあいまいになる傾向があることが指摘されている(e.g. Dalton-Puffer, 2007)。たしかに、従来の言語教育では読解の際でも文法構造の分析等言語そのものを意識しすぎるあまり、言語については知識があるが、運用能力が身につかないということが批判されることがあった。そこで、理解できる言語に触れる機会を多くして自然に習得(acquisition)を促す指導方法が提唱された(e.g. Krashen,1983)。しかしながら、その後イマージョン教育の調査結果等から、外国語の習得においては言語そのものを意識に乗せることも必要でありしたがって有効であることが理解されるようになってきた(e.g. Ellis, 2005；van Patten、2003)。

確かに、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。しかしながら、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 も例外ではありえない。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。

しかし、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する必要がある。そのうえで、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、今回は言語のもっとも基本を成す語彙を扱った。

2. CLIL の評価システムとその基盤となるモデル

本報告書では評価システム全体を考察の対象とはしていない。しかしながら、本プロジェクトの全体の枠組みを示す必要があるので、本節で簡単にまとめることとする。

CLIL の評価には Bloom(1949)およびその改訂版である Anderson 他(2001)が用いられることが多い。図 1 は Bloom のオリジナル版を示したものである。

knowledge(知識)→ comprehension(理解) → application (応用)→ analysis (分析)→ synthesis(統合) → evaluation(評価)

図 1 Bloom のオリジナル版 (Bloom、et al, 1956 を参考に現筆者が単純化したもの)

改訂版教育目標の分類(以下、改訂版)(Anderson, et la., 2001)は、Tyler (1949)の Content aspect と Behavioral aspect との 2 次元で教育目標を立てることを試みたものである。簡単に図式化したのが図 2 である。

remember(記憶する)→understand(理解する)→apply(応用する)→analyze(分析する)→evaluate(評価する)→create(創造する)

knowledge(知識)＝ factual(事実)、conceptual(概念)、procedural(手続き)、metacognitive(メタ認知)

図 2 改訂版(Anderson, et al, 2001)(Anderson et al, 2001 を参考に現筆者が単純化したもの)

改訂版では、1)知識(knowledge)を独立させ、認知プロセス(cognitive processes)とは異なる次元に設定した。その結果知識の次元とそれを運用する認知プロセスの次元の、2 次元の構成となった。2)Bloom 版では構成要素がすべて名詞で記載されていたが、改訂版では動詞となりプロセスを強調している。3)Bloom 版の知識は動詞化されまた認知プロセスをあらわすために、remember(記憶する)となった。4)知識や事実に関する知識(factual knowledge)、概念に関する知識(conceptual knowledge)、手続きに関する知識(procedural knowledge)、メタ認知に関する知識(metacognitive knowledge)の 4 種類から成るとした。階層性については、その妥当性をパス分析(Estrand, 1982 等)、因子分析(Hill, 1984 等)、共分散構造分析(Hill, 1984)等さまざまな実証研究の成果を援用して行ったとしている。また認知心理学の影響にあ

ることは明らかである。最も基本にあるのは、Gagnë(1977)である。

このように教育の目標を 2 次元で分類することにより、1 次元における知識を他の次元にある認知プロセスで処理するというふうにより応用力が高まった。例えば、付録 A に掲載したように、同じ事実に関する知識(factual knowledge)に対しても、記憶する(remember)場合、その知識を応用する(apply)場合、などのように目標設定がきめ細かく行えるようになり、ひいては評価も行えるようになった。このシステムは何より単純で教育目標を整理する際には便利である。

しかしながら、図から容易に見て取れるが、やはり 6 つの認知プロセスが記憶する(1 のレベル)から創造する(6 のレベル)に移るにつれて複雑になるという階層をなすという前提はかわっていない。したがって、改訂版の問題点として、本当に 1 から 6 に移るにつれて困難な認知プロセスを経ているのかどうかについては、必ずしも実証的に証明されているわけではなく、かなり恣意的であるといわなければならない。これはすなわち、あくまで分類であり習得の理論ではないことを示している。また、改訂版で対象となっているのは認知領域(cognitive domain)だけであり、情意領域(affective domain)は全く考慮されていないが、これは片手おちである。さらに、運動神経系統(psychomotor)(Simpson、1965)について全く触れられていないので、外国語学習では発音などの目標設定をする余地がない。

これらの問題点を解決すべくさらに改訂を行ったのが、Marzano & Kendall(2007)である。Marzano &Kendall 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、Anderson 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。

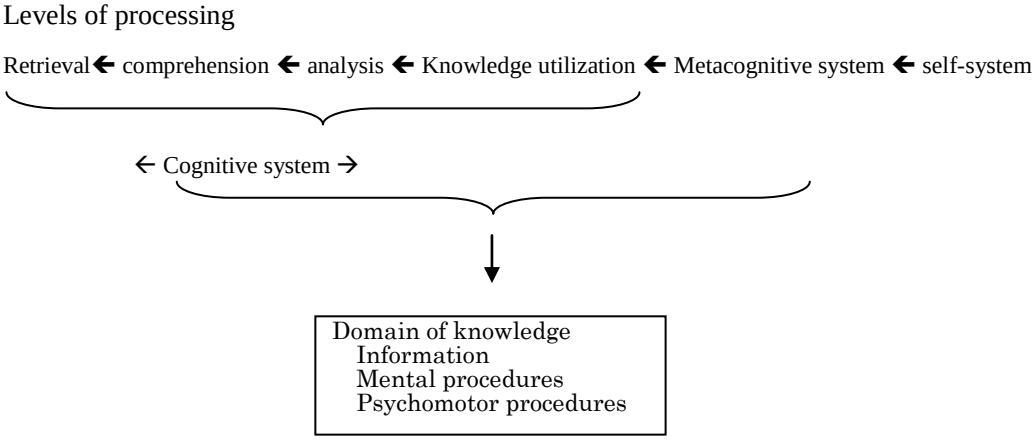


図 3 Marzano & Kendall(2007)のシステム (Marzano & Kendall, 2007 を参考に現筆者が単純化したもの)

学習対象が重要であると認識したり、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化し、下位 3 層を Lower-order thinking 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。

	THE COGNITIVE PROCESS DIMENSION					
THE KNOWLEDGE DIMENSION	1. REMEMBER	2. UNDERSTAND	3. APPLY	4. ANALYZE	5. EVALUATE	6. CREATE
A. FACTUAL KNOLWEDGE						
B. CONCEPUTAL KNOWLEDGE						
C. PROCEDURAL KNOWLEDGE						
D. METACOGNITIV E KNOWLEDGE						

図 4 改訂版の分類表　　Anderson, et al (2001)、改変

確かに、構成要素を動詞化して動きを表すようにしてあるし、またこの図式には表れていないが、別に知識の次元を設け、例えば「事実に関する情報(factual knowledge)」を「記憶する」、あるいは同情報を「理解する」というふうにいくつかの組み合わせでとらえることができるようになっている。そのために、教育目標を立てる際には大変臨機応変で使いやすい。

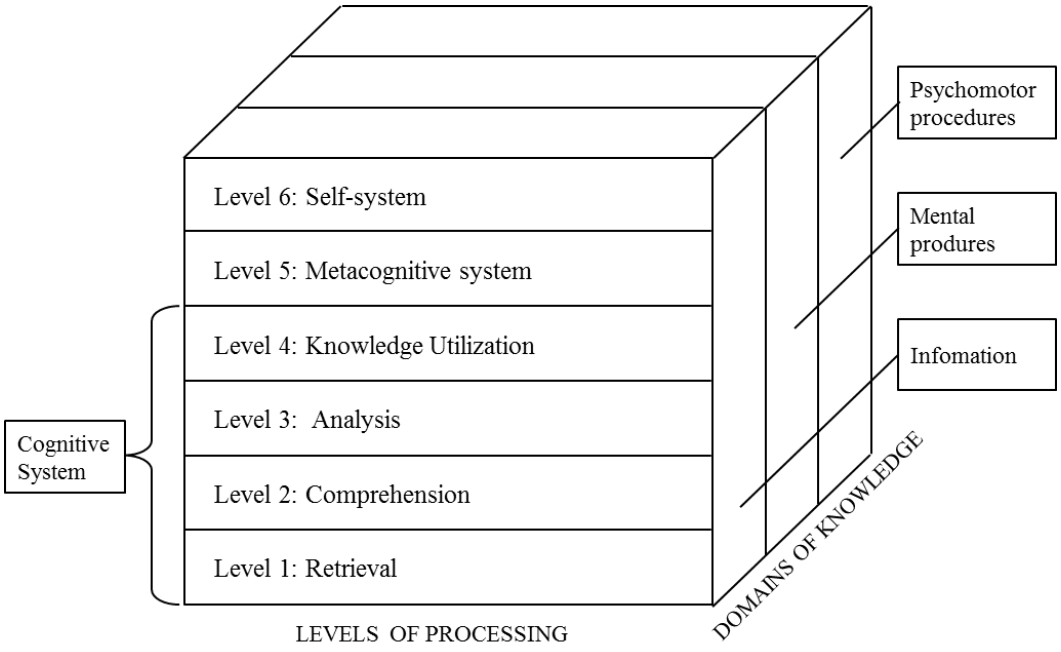


図 5 Marzano & Kendall (2007)の The new taxonomy of educational objectives

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。

表 1 新版の分類表

		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.

改訂版にしても、Marzano & Kendall(2007)も指摘する通り、これら階層を形成する構成要素について、操作が複雑だからという理由で実際の運用に高度な思考力が要求されるという証拠があるわけではない。また Bloom のオリジナルに存在し、またおそらく我が国の学校教育も直接の影響を受けている情意領域(affective domain)が含まれていない。情意領域は、困難だったということは、その提唱者たち、すなわち Bloom らですら認めている。("The success of *Taxonomy of Educational Objectives, Handbook 1: Cognitive Domain*, has spurred our work on the *Affective Domain*. As is indicated in the text, we found the affective domain much more difficult to structure, andwe are much less satisfied with the result. Bloom, et al, 1964, p. v).

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。

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)。

3. 先行研究

CLIL に関する実証研究はその多くが教育効果の検証である。その結果の示すところ教科に関する知識が増しかつ総合的な言語運用能力が高まる(e.g. Sylvën, 2004; Väzques, 2007; Zydatiß, 2007)という報告がある一方で、方法論に問題があることも指摘されている

(Përez-Canado, 2012)。また CLIL の授業における談話の型を検証しようとする試みもなされている(e.g. Dalton-Puffer & Smit, 2007)。Initiate、Response、Feedback からなるいわゆる IRF パターンは時に、教員が中心の授業形態を示すものとして批判されることもある(Lyster, 2007)。しかしながら、一方では教科学習には有効でもあり、したがって CLIL の特徴とみなすこともできるという報告もある(Katjia, 2007; Tarja, 2007)。CLIL の授業における語彙に関しては、特に教科の内容を習得することが焦点となることから特定の分野に関する語彙の習得に寄与するところが大きいという報告もある(e.g. Sylvën, 2004; Wode, 1999; Seregëly, 2008; Jexenflicker & Dalton-Puffer, 2010)。

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。

4. 研究方法

上智大学における CLIL の授業を観察の対象とした。上智大学では 2010 年度より実験的に CLIL を取り入れた外国語教育プログラムを行っている。春学期(4 月から 7 月まで)15 週間各 90 分の授業を並行して 4 クラス、さらに秋学期(10 月から翌 1 月まで)15 週間各 90 分の授業を並行して 4 クラスを開講しており、いずれも選択科目、定員は 25 名から 30 名である。春学期は学術研究に必要とされる一般的な学習技能(study skills)および 4 技能の統合の習得を目的としている。秋学期は、自然科学、人文学、社会科学など異なる分野から特定のテーマに絞り、春学期に習得した基礎能力を土台に学術能力を鍛えることが目標である。なお詳しくは、渡部、池田、和泉(2010)および和泉、池田、渡部(2012)を参照のこと。

4. 1. 対象授業科目および教員

今回観察の対象としたのは、春学期に 2 名の教員 A、B それぞれが担当している合計 2 つの授業 C、および秋学期、教員 A が担当している授業 D である。表 1 に図式化した。教員 A は

40 歳代、B は 30 歳代ともにイギリス人の母語話者である。教員 A の授業 D は文学であり、2011 年 11 月に授業観察を行った際には文学、特に詩の読み方を指導していた。教員 A と B の担当していた授業はどちらも同時期、すなわち 2012 年 6 月に行った。春学期はどの授業も共通のカリキュラムおよびシラバスと授業計画によって進められるためどちらの教員とも扱っているのは、学術論文の書き方を中心にしてプレゼンテーションの方法などについても指導を行っていた。

表 1

	春学期　授業 C　学術基礎技能	秋学期　授業 D　特定の分野
教員 A		
教員 B		

4. 2. データ収集および分析

授業はすべて DVD に録画し、教員の発話に焦点をおいて文字起しをした。授業観察中 DVD カメラは教室の後方部に設置し、教員の動きに合わせて随時方向を変えた。聞き取れない部分については XXX とし、分析対象からは除外した。データは語彙分析ソフト AntConc(Antony, 2011)を用いて分析した(www.antlab.sci.waseda.ac.jp/software.html)。

5. 結果と考察

5. 1. 基本的な量的傾向

基本統計は表 2 に示した通りである。

表 2　各教員の使用語彙基礎統計

		延べ語数 Token	異なり語数 Type	異なり語数／延べ語数 Type-token ratio	語彙密度 Lexical density
C クラス					
	教員 A	5,212	689	0.13	0.46
	教員 B	3,871	550	0.14	0.51
D クラス					
	教員 A	4,650	694	0.15	0.46

表 1 が示すように、延べ語数と異なり語数による特徴を見る限りにおいては、教員どうしの差の方が授業の目的による差よりも大きい。教員 A は C クラスにおいても D クラスにおいても教員 B よりも延べ語数、異なり語数ともに値が高い(教員 A、C クラスの延べ語数＝5,212、教員 B、C クラスの延べ語数＝3,871;教員 A、C クラスの異なり語数＝689、教員 B、C クラスの延べ語数＝550)。また語彙密度も教員 A は教員 B に比べて低い値を示している。このことから、

教員 A は全体として比較的多くの発話を行い、さまざまな語彙を使っているが、より会話に近い特徴を示している。今回の調査の目的に直接関係があるわけではないが、指導の効果測定評価を目的とした調査を将来行う際には考慮すべきデータとなる。

5. 2. 使用語彙の特徴

次に、使用されている語彙の特徴を検証した。前置詞や接続詞などの機能語については、今回は関心の対象外だったので、名詞および動詞に焦点を置いた。結果は表 3 から表 6 に示した通りである。

表 3
教員 A が C(学術技能クラス)に比べて D(内容中心クラス)で使用頻度の高かった語彙

	順位	頻度	特徴係数[2]	語彙
1	6	75	21.006	poem
2	20	32	8.963	Sylvia
3	22	30	8.403	Plath
4	26	37	5.185	know
5	27	17	4.761	background
6	30	32	4.070	think
7	31	14	3.921	cultural
8	32	14	3.921	move
9	33	13	3.641	life
10	35	12	3.361	group
11	36	12	3.361	poetry
12	38	11	3.081	holocaust
13	39	11	3.081	metaphors
14	40	27	3.004	read
15	41	25	2.595	right
16	42	9	2.521	analysis
17	43	9	2.521	father
18	44	9	2.521	information
19	45	9	2.521	shoe
20	48	8	2.241	enjoy
21	49	8	2.241	vampires
22	52	22	2.005	good
23	55	7	1.961	confusing
24	56	7	1.961	Daddy
25	57	7	1.961	German
26	58	7	1.961	historical
27	59	7	1.961	literature
28	60	7	1.961	metaphor
29	62	6	1.681	criticism
30	63	6	1.681	died

表 4
教員 A が D(内容中心クラス)に比べて C(学術技能クラス)で使用頻度の高かった語彙

順位	頻度	特徴係数	語彙
14	48	12.163	draft
17	47	11.910	essay
18	41	10.389	presentation
20	59	9.036	good
23	50	7.083	think
25	44	5.816	important
27	41	5.195	first
30	16	4.054	presentations
31	34	3.791	give
34	32	3.403	partner
35	13	3.294	peer
36	12	3.041	presenting
42	10	2.534	review
44	9	2.281	research
45	9	2.281	same
46	9	2.281	workshop
48	8	2.027	easy
49	8	2.027	mistakes
50	8	2.027	paragraph
51	8	2.027	pictures
54	24	1.942	people
55	7	1.774	email
56	7	1.774	fine
57	7	1.774	list
58	7	1.774	number
59	7	1.774	statement
60	7	1.774	thesis
61	7	1.774	visual
62	7	1.774	walrus
67	22	1.605	need

[2] Keyness の試訳

表 3 と表 4 には便宜上上位 30 位までもっとも頻度の高い語彙をリストした。表 3 は教員 A が学術技能中心のクラスに比べて内容中心のクラスでどのような特徴のある語彙を使ったかを示している。表 4 は同じく教員 A が内容中心のクラスに比べて学術技能中心のクラスでどのような特徴のある語彙を使ったかを示したものである。

表 5
教員 A を教員 B と比較した場合の学術技能クラスにおける語彙使用の傾向

	順位	頻度	特徴係数	語彙
1	14	48	9.797	draft
2	17	41	8.369	presentation
3	19	59	6.536	good
4	22	50	5.027	think
5	24	47	4.537	essay
6	25	44	4.056	important
7	27	41	3.583	first
8	28	16	3.266	presentations
9	30	13	2.653	peer
10	32	34	2.524	give
11	33	12	2.449	feel
12	34	12	2.449	giving
13	36	12	2.449	presenting
14	39	32	2.235	partner
15	40	10	2.041	interesting
16	41	10	2.041	review
17	43	9	1.837	guys
18	44	9	1.837	workshop
19	49	8	1.633	mistakes
20	50	8	1.633	pictures
21	51	8	1.633	things
22	53	7	1.429	fine
23	54	7	1.429	hand
24	55	7	1.429	visual
25	56	7	1.429	Walrus
26	57	7	1.429	way
27	59	6	1.225	book
28	60	6	1.225	circle
29	61	6	1.225	confident
30	62	6	1.225	kind
31	63	6	1.225	language
32	65	6	1.225	score
33	66	6	1.225	style
34	69	24	1.167	people

表 6
教員 B を教員 A と比較した場合の学術技能クラスにおける語彙使用の傾向

順位	頻度	特徴係数	語彙
7	75	18.007	topic
9	67	15.575	paragraph
12	56	12.279	sentence
22	22	7.304	transitional
24	36	6.513	essay
26	15	4.98	aspect
27	15	4.98	phrases
28	27	4.091	structure
29	12	3.984	signals
30	26	3.833	body
31	26	3.833	one
32	26	3.833	read
33	25	3.578	statement
36	24	3.326	thesis
38	9	2.988	introductory
40	22	2.832	look
42	8	2.656	information
45	20	2.355	main
46	20	2.355	sentences
47	7	2.324	best
48	7	2.324	building
49	7	2.324	second
50	19	2.123	ideas
52	6	1.992	indicate
53	6	1.992	Jefferson
54	6	1.992	material
55	6	1.992	Thomas

表中 keyness(特徴係数)は対数尤度比(logliklihood ratio)[3]を示しており、通常モデルとの適合度を表すが、ここでは値が高ければ高いほど比較対照のグループと差別化するための特徴の度合いを表している。

2 つの表から明らかなように、基礎統計の示した量的傾向は類似していたものの、使用されていた語彙は大きく異なることが見て取れる。内容中心のクラスでは詩の読み方をテーマとして指導が行われており、対象となっていた詩人は *Sylvia Plath* であった。したがって、*poem*、*background*、*metaphor* などの専門用語が多く使われている。一方、学術技能のコースでは、*presentation*、*paragraph*、*research* などが多く使われている。これらは授業の目的から十分予測される語彙であるが、その一方で、後者の授業では *easy*、*fine* などの情意的な語彙も含まれている。これは学生を励ましたり、評価したりする際に使われている語彙であり、この教員の指導法の特徴を表していると見ることができる。

さらに同じ学術技能コースを指導している二人の異なる教員の語彙使用の傾向を見てみよう。2 名の教員は使用語彙においても異なっていることがわかる。教員 A は draft、presentation、essay などのほかに、good、easy、guys など情意的な語彙を使う傾向がある。しかしながら、授業の指導目的は同じなので教員 B も同様に paragraph、sentence、phrase など学術論文を書く際に必要となるメタ言語が多く使用されていることがわかる。この点において目的が同じであれば使用語彙も同じジャンルに属する語彙を使う傾向があることもわかる。

5. 3. AWL(Academic Word List)との比較分析

学術基礎語彙についてはすでに、Academic Word List(AWL)が広く使用されている。今回使用されている語彙の中で AWL には記載されていない語彙はあるのか、あるとすればどのような語彙なのかを調べるために、*VcabProfile* (http://www.lextutor.ca/vp/eng/) (Laufer & Nation, 1995)でさらに分析を進めた。結果は表 7 に示した通りである。

表 7
AWL に記載されていないが今回使用されていた語彙

Teacher B AE1
African_[1] ambassador_[2] America/n_[2] atomic_[1] banned_[2] Hillary Clinton_[1] colon_[5] desserts_[1] diagrams_[1] divorce_[3] efficiently_[4] email_[6] emotions_[1] essay_[36] European_[1] feedback_[1] France_[2] Germany_[1] handout_[3] hint_[1] homework_[6] importing_[1] Japan_[1] Japanese_[1] Thomas Jefferson_[6] laundry_[1] number_[1] paraphrase_[1] phrase/s_[20] practiced_[1] punctuation_[1] restate_[1] restatement_[1] shy_[1] standup_[1] television_[3] video_[1] Virginia_[1] vocabulary_[2]

[3] 対数尤度比は次の式で算出される。$\sum_{i=1}^{N}[Yi \ln(P(Yi)) + \ln(1 - P(Yi))]$ (Field, 2009, p. 267).

表 7
AWL に記載されていないが今回使用されていた語彙(続き)

Teacher B AE2
beach_[1] bibliography_[1] bored/boring_[3] classroom_[1] crazy_[1] damn_[1] deadlines_[1] email_[7] embarrassing_[3] England_[1] essay/s_[50] facebook_[1] feedback_[4] guys_[9] handouts_[2] handsome_[1] homework_[2] hopefully_[1] Japan_[1] Japanese_[4] Jurassic_[2] karaoke_[1] microphone_[4] mirror_[1] moodle_[1] movie_[1] nervous_[1] peer/s_[14] portfolio_[1] presentations_[16] presenter_[2] professors_[1] rebuild_[1] references_[5] rehearse_[1] sandals_[1] scared_[1] score_[6] shy_[5] sigh_[1] silly_[1] spider/s_[6] Steve_[5] tick_[2] underline_[5] update_[1] usage_[1] video_[1] visuals_[2] vocabulary_[3] walrus_[7] workshop_[9]

Teacher A AE2
adjective_[2] American/s_[4] Ariel_[2] auto-_[1] autobio_[1] autobiographical_[3] bio-_[1] Boston_[1] boyfriend_[1] British_[1] brutal/ly_[6] bull_[1] career_[1] considerate_[1] controverse_[1] criticisms_[7] daddy_[7] divorce_[2] electronic_[1] England_[2] feminism/ist_[5] Sigmund Freud_[5] genres_[2] German_[7] girlfriend_[1] googled_[2] graph_[1] handsome_[1] holocaust_[11] homework_[1] horrible_[2] Hughes_[5] -ical_[1] impressions_[1] innocence_[2] James_[1] jar_[2] Jew_[1] Jewish_[4] linguistics_[1] metaphor/s_[18] movies_[1] nationality_[1] novel_[2] Oedipus_[1] Paltrow_[1] Sylvia Plath_[32] posthumous/ly_[6] professor_[2] psychoanalysis_[2] quiz_[1] reference/s_[10] rhyme/s_[5] rhythm_[6] Russian_[1] skip_[1] stake_[1] stance_[3] stanza_[3] suicide_[3] summarizer_[1] tragic_[2] trailer_[1] twilight_[2] underline_[2] unpacked_[2] vampire/s_[13] video_[4] vocabulary_[3] Wikipedia_[1]

表に見る限り、特に顕著な特徴は認められないが、African, America、Virginia、などの固有名詞が少なからず使用されていたことが見て取れる。学術基礎語彙を特定するには一般化されて、普遍的な語彙を探ることが重要だが、CLIL のような教科に特化した語彙を特定する場合には、このような固有名詞の習得も重要であることを示している。将来 CLIL 用の語彙リストを作成する場合の参考のために、最も基本的な 1,000 語レベルにおいて、AWL にリストされた語彙に記載されていないが、教員Aが専門(文学)を指導するために使用した語彙のリストを付録に記載する。

6. 結論

本報告書では大学における CLIL の授業を記録し、その語彙的な特徴を分析した。その結果 CLIL の学術技能(study skills)、教科内容(topical knowledge)、言語(language)を目的とした授業では、それぞれに特徴的な語彙が使われていることが明らかになった。さらに、教科内容によっては特定の固有名詞も基礎語彙として記載する必要があることが示唆された。今回は語彙のみを分析の対象としたが、それぞれの目的および異なる分野によって必要とされる言語機能、言語構造を特定することにより、CLIL における言語習得を促進する基盤となることが期待される。

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.

Admiraal, W. G., Westhoff, and de Bot, K. (2006). Evaluation of bilingual secondary education in The Netherlands: Students' language proficiency. *English educational research and evaluation*, 12, 2, 75 – 93.

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.

Bloom, B. S. (1949). *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 Zarobe, 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 language integrated learning: Evidence from research in Europe*. (pp. 81 – 92). Bristol: Multilingual Matters.

Coxhead, A. (2000). A New Academic Word List Author(s): Averil Coxhead Source: TESOL Quarterly, Vol. 34, No. 2, (Summer, 2000), pp. 213-238, Downloaded March 31 from http://edc448uri.wikispaces.com/file/view/Coxhead+2000+Acad +Word+List.pdf

[4] 本稿は学術論文というよりも、報告書という性質上、本稿では直接参照しなくても、調査の過程で参考となった文献は記載することとした。

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

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

Dalton-Puffer, C. (2007). *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.

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.

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

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., & Smit, 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 Smit, U. (Eds.). *Empirical perspectives on CLIL classroom discourse.* (pp. 119 – 138). 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). Vocabulary size and use: Lexical 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 language in CLIL*. 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 language integrated learning: Evidence from research in Europe*. (pp. 112 – 129). Bristol: Multilingual Matters.

Lynch, B., & Davidson, F. (1994). Criterion-referenced language test development: Linking curricular, teachers and tests. *TESOL Quarterly*, 28, 4, 727 – 744.

Lyster, R. (2007). *Learning and teaching languages through content: A counterbalanced 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 Wolff, D. (eds.) (2007). *Diverse contexts – converging goals*. Frankfurt am Main: Peter Lang.

Marzano, R. J. & Kendall, J. S. (2007). *The new taxonomy of educational objectives*. Oaks, Cal.: Corwin Press.

Marzano, R. J. & Kendall, J. S. (2008). *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.

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 bilingualen sachfachunterrichts an gymnasien. Kontext, kompetenzen, konsequezen. Frankfurt-am-Main: Peter Lang.

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

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

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.

付録

教員 A が文学(詩の読み方)を指導するために Academic Word List (AWL)以外から使用していた語彙のリスト(1,000 語レベル)(VocabProfile を使用)

0-1000 [families 312 : types 456 : tokens 3931]

OFF types: [?:106:367] adjective_[2] American_[3] Americans_[1] Ariel_[2] auto_[1] autobio_[1] autobiographical_[3] bio_[1] Boston_[1] boyfriend_[1] British_[1] brutal_[5] brutally_[1] bull_[1] career_[1] considerate_[1] controverse_[1] criticism_[6] criticisms_[1] daddy_[7] divorce_[2] doesn't_[1] don't_[2] electronic_[1] England_[2] feminism_[1] feminist_[4] Freud_[3] genres_[2] German_[7] girlfriend_[1] gonna_[5] google_[1] googled_[1] graph_[1] guy_[1] guys_[1] handsome_[1] holocaust_[11] homework_[1] horrible_[2] Hughes_[5] ical_[1] ii_[3] impressions_[1] innocence_[2] it's_[1] James_[1] jar_[2] Jew_[1] Jewish_[4] linguist_[11] movies_[1] nationality_[1] Nazi_[4] Nazis_[1] novel_[2] Oedipus_[1] Paltrow_[1] Plath_[30] posthumous_[1] posthumously_[5] professor_[2] psychoanalysis_[2] quiz_[1] reference_[6] references_[4] rhyme_[2] rhymes_[3] rhythm_[6] russian_[1] sigmund_[2] skip_[1] stake_[1] stance_[3] stanza_[3] suicide_[3] summarizer_[1] Sylvia_[32] ted_[5] that's_[2] tragic_[2] trailer_[1] twilight_[2] uh_[18] underline_[2] unpack_[1] unpacked_[1] vampire_[5] vampires_[8] video_[4] vocabulary_[3] Wikipedia_[1] yeah_[11]

おわりに

最近は、英語教育に対する風当たりは、きわめて強い。たとえば、「大学入試にTOEFL」(朝日新聞5月1日)「官僚もTOEFL必須」「TOEFLの入試導入, 慎重に」(朝日新聞5月5日)「英語能力試験の義務付けとグローバル化」(『英語教育』6月号)などは、その一例であろう。そこでは、やはり、教育の中の評価・テストの意味と機能をよく検討しなければならない時期に来ていることを感じるのは、筆者だけではあるまい。

テストや評価の意味と機能は、その時代の流れを深く検討して、求めなければならない。この報告書は、まさにこの時代のテストと評価の課題を背負って書かれたものである。テストと評価が、英語教育の中で持っている役割はどういうものであるか？そうした多くの課題の中から、「言語テストの規準設定」というテーマでとらえたものが、この報告書である。2011 年度においては、1. 規準設定の意味と歴史、2. 内容言語総合型学習(CLIL)における規準設定、3. Can-do statements における規準設定、4. テスト理論と規準設定、5. ヨーロッパ共通参照枠(CEFR)と規準設定という角度から、規準設定に関する先行研究を探し求めたものである。本年度においては、その先行研究の中で、最も重視しなければならないと思われる課題を、それぞれ選択し、その課題をさらに深めるための研究・実験を行った。その成果を、ここに示したものである。

2012という本年度では、2011年度の先行研究をふまえ、さらに、2013という来年度の準備体制を築いているものである。「予備調査:CITO Variation on the Bookmark Method」「Can-do statements の比較・研究」「Can-do statementsの規準設定」「受容語彙力を測定するプレイスメン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」、「CLIL における語彙による規準設定」と様々な角度からの規準設定を論じたものである。研究構成員、渡部良典、伊東祐郎、法月　健、藤田智子各氏には、教育の現場において、多忙を極めている中、きわめて貴重な成果を収めていただいた。とくに、研究副代表の渡部良典氏には、献身的な協力をいただき、この場を借りて、御礼を申し上げたい。これを基盤に、2013 年度における、さらに実りある前進を心から期待するものである。

2013 年 3 月 31 日

研究代表　大友賢二

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(上智大学外国語学部教授)：研究副代表

(あいうえお順)

言語テストの規準設定　報告書(2)
2013 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
