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

２０１２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

２０１２年3月３１日

財団法人　日本英語検定協会

英語教育研究センター　委託研究

研究構成員

伊東祐郎（東京外国語大学留学生日本語教育センター教授）

大友賢二　（筑波大学名誉教授）：研究代表

法月　健（静岡産業大学情報学部教授）

藤田智子（東海大学外国語教育センター教授）

渡部良典（上智大学外国語学部教授）：研究副代表

(あいうえお順)

目次

はじめに

規準の設定（Standard Setting）はテスト研究で最も重要なテーマの一つである。しかし、それは言語能力の測定の精度を高める、妥当性を確立する、といった一般的な意味における重要さとは少し違う意味においてである。

きわめて優秀な生徒を指すのに「オール５」といういい方がある。小学校のクラスには必ず１人か２人がいたものである。相対評価の名残かと思っていたが、目標準拠評価の現在でもやはり一般的ないい方なのであろう、Google で検索すると 300,000 件ヒットする。学校教育を受けた人ならばだれでも成績をつけられた経験がある。また、教員にとって最終成績を出すというのは学期や学年を締め括る最も骨の折れる作業である。甲乙丙丁にせよ A、B、C、D にせよ、優良可にせよ、おそらくはだれもが一度は正しく認められていないとか、なぜこのような成績なんだろうとか、考えたことがあるに違いない。普通は、それはそれ、ということで深く考察するというようなことはない。

成績は最もプリミティブな形で表現された規準の設定である。しかしながら、あまりにも当然のこととされているので、その重要さが理解されにくい分野でもある。先に、規準の設定はテスト研究において最も重要な分野の一つであると言ったのは、それが理論や大規模テスト開発に貢献するのみならず、私たちが通常、テストや成績という極めて身近に経験する分野にも関係するという意味においてである。近年 CEFR（Common European Framework of Reference）で行われている A1 レベルや B2 レベルといった言い方も多くの教員や外国語の学習者ですら口にすることが多くなっている。この場合もやはり、規準の設定を意識して敢えて客観化しようとするのは主に研究者にまつわるところが大きい。しかしながら、規準の設定を客観化するための研究と実践の両方が必要である。

本報告書は、規準の設定という言語テストにおける最重要課題を５名の研究者がそれぞれの立場から、１年間に亘り取り組んだ成果をまとめたものである。メンバーは各々が異なる背景を持ち､そして異なる関心を持っている｡しかしそれぞれ言語テストや教育評価、外国語指導の実践に関わっていることでは共通しており、今回はひとつのテーマについて歴史、理論、応用という観点から多元的な考察を行い、今後の研究への橋渡しをするための課題をまとめることを目標とした。固より１年という限られた期間でもあり、また本務校等での職務と国内外における学会活動等と並行して行ったものである。十分な時間と労力を費やすことができたとは言えない。しかしながら、今後への橋渡しとしての作業は果たせたのではないかと思う。本報告書をもとにさらに調査を進める所存である。

平成２４年３月３１日

研究副代表

渡部　良典

1. The Meaning and History of the Standard-setting Methods in Language Testing

1.1. Development of the Standard -setting Methods in Language Testing in Foreign Countries

Kenji Ohtomo

Abstract

Standard-setting is a critical part of educational, licensing, and certification testing. This aspect of test development, however, is not well understood. Standard setting can be defined as a process by which a standard or cut score is established. Unless cut scores are set appropriately, the results of assessment could be questioned.

This report reviews how standard-setting methods in language testing have been developed and used. There have been a number of methods based on judgments about the test questions and judgments about individual test -takers. Some researchers have described the standard setting process as blatantly arbitrary. Some, however, argue that standard setting is more appropriately conceived of as a measurement process to student assessment.

In order to find out new, future directions for standard-setting methods, the report concludes by summarizing three suggestions. First, finding out the best combination of the two methods is needed since there is a view that "there is no best standard setting method ". Second, research on the Bookmark Method is needed. And third, the Ranking Approach developed in Europe has to considered for further development.

1. 規準設定の意味と歴史

1. 1. 海外における規準設定法の研究とその動向

大友賢二

規準設定研究の意味と開発（1）

「海外における規準設定法」を課題として研究を推進する場合、その「規準」という用語の意味を、まず、明確にしておかなければならない。わが国の教育の中では、この「規準」という語は、「測定」「評価」に関連する「指導要録」などに見いだすことができる。わが国における指導要録の告示は、「指導要領」の告示から 3 年ほど遅れてなされるのが普通である。したがって、現行のものは、1998 年の「指導要領」改訂から 3 年あとの 2001 年の告示によるものである。その特徴は、「評定」欄の評価方法も「相対評価」から「目標に準拠した評価」へと完全に移行されたということである。

わが国のこうした状況の中で、測定、評価の概念規定にかかわる課題の一つに、「規準」と「基準」という日本語がある。この日本語「のりじゅん」と「もとじゅん」などの議論は、今回の指導要録の改訂ではじめて生じてきた議論であるかのような印象を与えてしまったことがある。しかし、この議論は、1983 年という 29 年前にその「はしり」があったことを、認識しておかなければならない。橋本（1983)では、クライテリオンには「規準」を、スタンダードには「基準」をと述べている。その後、皆見 (2008)など、さまざまな議論があったが、これに関する筆者の立場としては、池田（監訳）(2008) に準じて、criterion を「基準」、standard を「規準」と呼ぶこととする。

これに関連する用語についての議論は、海外においても見られた。たとえば、「分割点」を意味する用語としては、passing scores は Livingston & Zieky (1982)で、cut-off scores は Nitko (1983)で、cut scores は AERA , APA & NCME (1999)で、cut-scores は Davies, et al. (1994)で、standards は Cizek (Ed.) (2001)で, そして、cutscores は Zieky , Perie, & Livingston (2008)で、それぞれ用いられてきている。

このなかの standard という 1 語をとりあげてみても、その意味は、Fulcher (2010) に示されているように、6 種類という多岐にわたるものがある。そのうちのどの場合の意味を指しているかを明確にして論じなければならない。Standard と関連してよく用いられるのが、cutscores であり、passing scores でもある。つまり、「規準設定（standard setting）」という用語は、「規準、つまり、分割点を設定する手順である」と規定することができる。例えば、この規定は ‘ standard setting can be defined as a process by which a standard or cut

score is established. ’ Cizek (2006) などで示されている。その具体例としては、Cizek & Bunch (2007)*Standard Setting: A Guide to Establishing and Evaluating Performance Standards on Tests* などがあり、最近の傾向としては、この standard という用語が頻繁に用いられてきている。

規準設定研究の意味と開発（２）

海外における規準設定法開発の初期のものとして、第 1 に、Berk (ed.)(1980)をとりあげることができる。これは、1978 年の 10 月に Washington, D.C. で開催された Johns Hopkins University 主催の最初のＮＳＥＲ（National Symposium on Educational Research）の成果をまとめたものである。外国からも参加したテスト専門家260名が、教育測定・評価で当時最も議論の的であった criterion-referenced measurement に焦点を合わせた会議であった。その Introduction では、The term criterion-referenced was first coined in 1962 (cf. Glaser & Klaus (1962)). The distinction between a *criterion-referenced* and a *norm-referenced test* was most clearly defined in Glaser’s seminal essay (1963). と示されているのは、きわめて印象的である。本書には、筆者の UCLA 客員研究員時代(1979-1980)の恩師 James Popham 教授による Domain Specification Strategies を初めとして、Robert A. Berk: Item Analysis, Ronald K. Hambleton: Test Score Validity and Standard-Setting Methods, さらには、後に *Educational Measurement* (The Fourth Edition 2006) の編者として活躍された Robert L. Brennan : Applications of Generalizability Theory などが見られる。

第２の Livingston & Zieky (1982) の内容は、その副題として、*A Manual for Setting Standards of Performance on Educational and Occupational Tests*. と示してあるとおり、まさに、「規準設定法」に関する手引きである。この文献で注目したいことは、その規準設定法を３つに大別している点である。（１）テスト項目 に関する判断を基にしているもの、（２）個人のテスト受験者に関する判断に基づいているもの、そして、（３）は、テスト受験者のグループに関する判断に基づくものとしている。（１）では、Nedelsky’s method, Angoff’s method, そして、 Ebel’s method をとりあげている、また（２）では、the Borderline-Group Method, the Contrasting-Groups Method, the Up-and-Down Method をとりあげている。

第３の文献 Faggen (1994) は、特に、「解答構築式テスト」（constructed-response tests) に関するものである。解答構築式テストは、概して信頼性の低い得点を生み出すので、それに対する慎重な考察が必要である。本研究では、この解答構築式テストの改善を目

指して、以下の４つの方法を比較検討している点は注目に値する。つまり、(１) Benchmark, (2) Item-Level Pass/Fail, (3)Item-Level Passing Score, (4) Test-Level Pass/Fail という方法である。この４つの方法に関し、必用な資料、データ収集、データ分析、参考資料、等の視点からの検討は高い興味をそそるものである。

規準設定研究の意味と開発（３）

規準設定法の研究も、さまざまな角度から検討されるようになってきた。その１例として、ここでは、Cizek, G.J (1996), Standard-Setting Guideline, *Educational Measurement: Issues and Practice*, Spring 1996. および、Zieky, M.J. (2001), So Much Has Changed: How the Setting of Cutscores has Evolved Since the 1980s. In Cizek, G.J.(ed.)(2001) をとりあげて、その検討の跡を探ることとする。

Cizek (1996) では、規準設定の手順として以下のような、Recommended Guidelines for Standard Setting を具体的に示していることに注目しなければならない。

１．Purpose：A.　規準設定の目的を明確にすること、B.　妥当な構成概念を明確に示すこと、２．Method：A.　規準設定の目的と方法を連係させること、B.　測定される特性と方法を連係させること、C.　選択した規準設定法を明示すること、３．Procedures. A. 実施した手順を明記すること、B. 判定委員、判定、分割点などの調整手順を明記すること、４．Technical and procedural analysis.　A. 選択した判定委員と方法を明記すること、B.作業理解の現状を示す証拠を残すこと、C.判定委員による適切な情報利用の記録をとること、D.誤差の大きさの報告をすること。

Cizek, G. J. (ed.)(2001), *Setting Performance Standards*: *Concepts, Methods, and Perspectives,* Lawrence Erlbaum Associates, Publishers は、1980 年代から 2000 年までの規準設定法の経過を知る最も貴重な文献の一つである。本書は、第 1 部は規準設定の基本的課題（Fundamental Issues in Standard Setting）、第 2 部は規準設定法の実際（Standard–Setting Methods in Practice）第 3 部は規準設定における連携する課題（Continuing Issues in Standard Setting）とさまざまな角度からの視点を示している。

第 1 部では、Gregory J. Cizek, Michael J. Zieky, Michael T. Kane, Ronald K. Hambleton など、この時期を代表する著者の顔ぶれである。また、第 2 部では、Body of Work Method や Bookmark Procedure や Setting Standards on Computerized Adaptive Tests など興味深い論文に触れることができる。Zieky, M.J.(2001)のタイトルは、So Much Has Changed であるが、その A Final Word で述べている以下の発言は、きわめて意味深いものである。

In spite of all those changes, innovations, and improvements, some basic characteristics of cutscores remain unchanged. There is no “true” cutscore, and whether or not people find a cutscore to be appropriate depends on their values concerning the relative harm caused by each type of error of classification. (p.47)

規準設定研究の発展（１）

現在まで行われた規準設定に関する研究の跡を探ると、その道は、まさに茨の道であると言える。つまり、規準設定に関するベストな方法は、あるのだろうか。もしあるとすれば、それは何であるか？もし、なければ、なぜないのであろうか？この両者の疑問の間を走り回っているのが現状であろうかと思われる。最善の策は、きわめて求めにくい状況であるが、どんなところに見出すことが可能なのであろうか？そのいくつかを、拾い上げてみることにする。

その１つは、最終的な分割点を決定するときには、どんな研究においても、複数の方法を利用し、統計以外の要因も併せて、すべての結果を考慮に入れるのが賢明であろうとする動向である。例えば、井上訳(1992) 「学生のコンピテンスの証明」、(Richard M. Jaeger, Certification of Student Competence, In Linn, R.L. (Ed.), (1989), *Educational Measurement: Third Edition,* NCME, ACE)が述べていることである。複数の方法とあるが、最も効果的なのは、何と何であるか？すべての結果を考慮とあるが、どんな結果を指しているのだろうか？

第２番目に上げることができるのは、AERA, APA & NCME (1999), *Standards for Educational and Psychological Testing,* AERA における論述：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 (p.53)というものである。

第３としては、Jaeger , R.M. and Mills, C.N. (2001) における論述：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)). (p.314) というものである。

第４の例として挙げることができるのは、Felianka Kaftandjieva (2004),における次のような論述である。

To summarize, there is no ‘gold standard’ , 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. (p.31) という、いわば絶望的なものがある。

規準設定研究の発展（２）

比較的新しい論述としては、Zieky, M.J., Perie, M. and Livingstone, S.M. (2008), *Cutscores: A Manual for Setting Standards of Performance on Educational and Occupational Tests*. Educational Testing Service.がある。その内容としては、次の様なものである。しかし、最後の一節：Cutscores can not be objectively determined, but they can be objectively applied. という論述は、きわめて深い意味を持つものであろう。

> The cutscore will depend on whose judgment are involved in the process and on the method that you use to set cutscores. 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. (p.197)

こうした深い意味を持つ論述は、まことに興味あるものであるが、この論述は、じつは 1982 年におなじく ETS (Educational Testing Service)から出版され、同じ著者の Samuel A. Livingstone と Michael J. Zieky らによって述べられている Conclusion の中の論述ときわめて類似ていることを見出すことができる。それは、Livingston, Ｓ．Ａ． and Zieky, M.J. (1982) , *Passing Score: A Manual for Setting Standards of Performance on Educational and Occupational Tests*. ETS のなかの以下の文である。

> The standard will depend on whose judgments are involved in the process. In this sense, all standards are subjective. Yet once a standard 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. Standards cannot be objectively determined, but they can be objectively applied.(p.67)

Cutscoresに関する１９８２年の論述と２６年が過ぎた２００８年における論述が、さほど変わっていないのは、この課題解決の難しさを物語っているのであろうか？それとも、なにか、ほかの要因と関係があるのであろうか？この規準は客観的に決定できない

ということの意味をさらに究明していかなければならない。

残された課題（１）

これまでは、言語テストの規準設定に関する主な先行研究を検討してきたが、最後に、検討が十分ではなく研究をさらに継続し発展しなければならない課題はいったい何であろうか？現在残されている課題に関して少し考えてみることにする。

その第 1 は、これまで開発された個々の規準設定法ではきわめて客観性を欠き、たとえば、複数の方法を利用することがより賢明であろうとする見方もある。しかし、何と何を利用するのがより効果的なデータを求めることができるのであろうか？それを究明することが必要であろう。残された課題の第 2 は、データの安定性を十分確保できる設定法をさらに求めることである。例えば、ＩＲＴを用いた Bookmark Method などの更なる究明はその一つであろう。残された課題の第 3 は、長い間議論を重ねてきている CEFR に関する研究の跡を探し、その中で、規準設定法研究では何が残されているかを見つけだすことであろう。

規準設定方法の種類を文献で展望してみると、その数は、きわめて大きい。年代順にその流れを検討すると、(1) Livingston and Zieky (1982), (2) Cizek (2006), (3) Cizek and Bunch (2007), (4) Zieky, Perie, and Livingston (2008) などがあり、その中に含まれている規準設定の方法は、次のようなものがある。

まず、文献(1)では「テスト問題に関する判断に基づく方法」（methods based on judgments about test questions）としては、Nedelsky’s Method, Angoff’s method, Ebel’s method, さらに、「テスト受験者に関する判断に基づく方法」（ methods based on judgments about individual test-takers)としては、The Borderline-Group Method, The Contrasting-Groups Method, The Up-and- Down Method などについて検討している。

文献(2) の中で (1) には含まれていない方法としては、The Yes/No Method, The Extended Angoff Method, The Bookmark Method, The Body of Work Method, The Beuk Method, The Hofstee Method などがとりあげられている。文献(3) の中で、(1), (2) には含まれていない方法としては、The Direct Consensus Method, The Item-Descriptor Matching Method などがとりあげられている。文献(4) の中で、(1), (2), (3) には含まれていない方法としては、The Performance Profile Method, The Dominant Profile Method, The Analytic Judgment Method などが取り上げられている。

以上のように開発された規準設定法に対する見方は、肯定的というより否定的な見方

が多い。しかし、角度を変えて、学会誌などでの広い視野に立つ動向は、どうなっているであろうか？この点に関しては、たとえば、次のような興味ある論述を挙げることができる。

残された課題（２）

Nichols, P., Twing, J., Mueller, C. D. , and O'Malley, K．（2010), Standard- Setting Methods as Measurement Processes, *Educational Measurement: Issues and Practice,* 2010, Vol.29, No.1. がそれである。これまでのおおかたの見方は、Standard Setting Methods は、きわめて主観的であるとする声である。この論文は、それに対する反論でもある。次のような文は、明らかに、それを示している。

> 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.(p.14)

残された課題の第 2 は、データの安定性、客観性を十分確保できる規準設定法をさらに求めることである。否定的な見方が継続している規準設定法の中では、たとえば、項目応答理論を活用した Bookmark Method をとりあげてみることができるであろう。

残された課題究明の第 3 の糸口となるデータは、Council of Europe (January, 2009), *A Manual: Relating Language Examinations to the CEFR*, Language Policy Division の中の、Chapter 6. Standard Setting Procedures である。また、この Chapter の 6.11. Conclusion で言及されている　Extra Material provided by Brian North and Neil Jones は、重要な視点を含んでいる。

われわれの近くで活躍されている Neil Jones 氏は、特に、興味ある論文を次々と発表している。最近のものの中では、*Cambridge ESOL: RESEARCH NOTES: Issue 37/ August* (2009) の A Comparative approach to constructing a multilingual proficiency framework: constraining the role of standard setting は、興味深い論文である。この論文の中で示されている、次の発言は、新しい方向の糸口として、適切な要素の一つではなかろうか？" I have proposed that **a ranking approach** offers practical way to align different languages and tests to a common scale, and that is logical to do this as a separate step prior to standard setting." (p.9) ここで言う a ranking approach の理論解明については、Bramley, T.(2005)の論文 A Rank

Ordering Method for Equating Tests by Expert Judgment　や、Council of Europe (April, 2011), の *Manual for Language Test Development and Examining*　をさらに究明することが求められる。

参考文献

AERA, APA & NCME (1999), *Standards for Educational and Psychological Testing ,* (p.53, 59). AERA

Berk, R. A. (ed.) (1980), *Criterion-Referenced Measurement, The State of the Art ,* The Johns Hopkins University Press、

Bramley, T. (2005) A Rank-Ordering Method for Equating Tests by Expert Judgment, *Journal of Applied Measurement 6/2,* 202-223.

Cizek, G. J. (2006), Standard Setting (pp. 225-260, p. 226). In Downing, S.M. & Haladyna, T.M. (Eds.) *Handbook of Test Development*, Lawrence Erlbaum Associates, Publishers.

Cizek, G.J (1996), Standard- Setting Guideline, *Educational Measurement: Issues and Practice*, *Spring, 1*996. 14.

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

Council of Europe (April, 2011), *Manual for Language Test Development and Examining,* Language Policy Division

Council of Europe (January, 2009), *A Manual: Relating Language Examinations to the CEFR*, (pp.57-88) . Language Policy Division

Davies, A. et al. (1994), *Dictionary of Language Testing,* (p.40) Cambridge University Press

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

Faggen, J. (1994), Setting Standards for Constructed –response Tests: An Overview, *Research Memorandum* 94-19.　ETS

Felianka Kaftandjieva (2004), Section B: Standard Setting, *Reference Supplement to the Preliminary Pilot version of the Manual for Relating Language Examinations to the CEFR*: (p.31) Language Policy Dicision, Council of Europe.

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

Glaser, R. & Klaus, D.J. (1962), Proficiency measurement:　Assessing human performance. In

R.M. Gagne, R.M. (ed.) *Psychological principles in systems development,* (pp.419-474). Holt, Rinehart and Winston,

Glaser, R. (1963), Instructional technology and the measurement of learning outcomes: Some questions, *American Psychologist 18*. 519-521.

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*, (p.314). Lawrence Erlbaum Associates, Publishers.

Jones, N. (2009), A Comparative approach to constructing a multilingual proficiency framework: constraining the role of standard setting *Cambridge ESOL: RESEARCH NOTES: Issue 37/ August 2009,* 6-9.

Livingston, S.A., and Zieky, M.J. (1982), *Passing Scores: A manual for setting standards of performance on educational and occupational tests*, (pp.15-43, p.67). ETS

Nichols, P., Twing, J., Mueller, C. D. and O'Malley, K.（2010）, Standard- Setting Methods as Measurement Processes, *Educational Measurement: Issues and Practice, 2010, Vol.29, No.1*. 14-24.

Nitko, A. J.(1983), *Educational Tests and Measurement: An Introduction* (p.454). Harcourt Brace Jovanovich, Inc.

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., Perie, M. and Livingstone, S.M. (2008), *Cutscores: A Manual for Setting Standards of Performance on Educational and Occupational Tests*. (p. 197). Educational Testing Service.

橋本重治（1983）『続・到達度評価の研究―到達基準設定の方法』(p.28) 日本図書文化協会

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 ほか、『教育測定学（下巻）(p.238) C．S．L．学習研究所会

池田 央 日本語版監訳(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

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 』. 273-281

1.2. Exploring the History of Standard Setting in Japan

Yoshinori Watanabe

Abstract

The purpose of the present report is to trace back the history of practice in standard setting in the field of teaching a foreign language with focus on EFL. The history of research into standard setting in Japan is fairly short, whereas its practice has established relatively a long tradition. The report starts by reviewing the practice of setting standard in its simplest form of classroom grading to a highly complex method using the fuzzy set theory. In so doing, it has been found that a huge gap has been present between practice, theory and an official document such as the Ministry of Education guidelines. The report concludes by offering suggestions for the future research, which needs to be conducted to bridge the gap of these three factors.

１．２．　日本における規準設定研究の歴史

渡部良典

本節では、「日本における」、「規準設定」、「研究」の歴史、というふうに3つの条件が課せられている。さらに、分野の規定がないので教育、心理、英語教育、日本語教育、外国語教育等をも含めうることを暗示する。とすれば、本章で扱うべき分野は限りなく広いといわなければならない。限られた紙数と限られた時間内に収めるために、本章では、主に「わが国の外国語としての英語教育における規準設定の実践の近現代史」に範囲を限ることとする。外国語教育における規準設定の実証研究についてまとまった成果が公にされたのは大友（2009）をもって嚆矢してもよいほど最近に属することであり、未だまとめるに足る長い研究の歴史があるとは思われない。本節では過去の成果を肯定的に学びとることを旨とする。現在からみれば素人でもニュートンの古典力学が完全でないことはわかるし、不確　定性原理を少しでも知っていれば相対性原理ですら批判できるし、小澤正直博士の業績を知れば不確定原理ですら完全でないことがわかる。ここでは今後の実証研究ではどのような調査を行うべきなのかを学ぶことにしよう。外国語としての英語教育に関わる範囲において、心理、教育等に触れながら、近現代の実践史から何が読み取れるかを考察する。

規準設定前史

規準設定が最も素朴な形で問われるのは、秀（excellent）、優（very good）、良（good）、可（average）、認（acceptable）、不可（failed）あるいはA、B、C、D、E、Fで表される成績、GPA（Grade Point Average）であろう。90%から、80－89%、70－79％、60－69%、50－59％というようなおおよそ恣意的な分布である。最初に採用したのは、18 世紀のケンブリッジ大学で化学の教授が採用したのが最初である、というようなことを聞いたことがあるが、正確な起源やその科学的根拠については皆目見当がつかない。Durm (1993)は少なくとも米国では1785年エール大学にその起源を求めている。Johnson（2003）、Hue（2005）、Hunt（2009）、Zellner（2011）等は、規準設定が米国の大学ですら、そしてかのバカロレアですら大きな問題となることを示している。

言語テストの研究・実践の歴史を知るには Spolsky（1995）が標準的なテキストであり、特に米国における歴史については Barnwell（1996）がほとんど唯一のまとまった著書である。わが国については Tanaka（2008）の小著および Sasaki（2008）がある。しかし、どれも規準設定に関する詳細はない。日本でいつどのような経緯で使われるようになったのかはもっと定かでないが、教育年表見るといろいろと面白いことに気づく。「1948年（昭和23年）文部省、小学校学籍簿に5段階相対評価法採用を通達」（例えば伊ヶ崎・松島、1990）、ということは、それ以前は相対評価が行われていなか

ったことを示している。またそれに先立ち、次のような記載が目を引く。「昭和 13 年 1938 年文部省が学籍簿を改正し、学習成績は 10 点法、操行は優良可式に表記統一」（下川、2000）（ついでながら、大正 9 年 1920 年には東京帝国大学、大正 10 年度から学年の始期を 9 月から 4 月 1 日に変更することを決定、他の官立大学、高等学校も追随とある。現在の4月年度始まりというのはやはり東大から始まったのである。）

教育・心理の分野では、橋本重治（1959）などが、独自の論を展開しながら、当時の教育心理関係、特に米国の研究を紹介している。橋本の一連の著書がどのくらい現場の教員に普及しているのかは定かではないが、平成 11 年 12 月の中央審議会の答申において、観点別評価、目標準拠評価の重要性が指摘されて後に、1981 年出版の『到達度評価の研究』の新装版が出版されている（橋本、2000）。本書で橋本は、一章を割いて「到達度判定のための基準[ママ]の設定の方法」（123－157 頁）を詳しく紹介している。さらに、信頼性や妥当性も論じており、現在でもこの方面に知識のない教員や研究者の必読となっている。また、個々のテスト・アイテムの分析・判断に基づく方法（今でいう item-centered studies）としてネデルスキー法（Nedelsky, 1954）、アンゴフ法（Angoff, 1971）、エーベル法（Ebel, 1972）を紹介し、また学識経験者へのアンケート調査（いわゆる person-centered studies）による方法も詳細に紹介している。さらに、これらの利点と欠点、また研究なども簡単ではあるが紹介されている。1981 年出版よりほぼ内容はそのままになっていることから、本書がそれだけ先見性に富んでおり現在でも価値を失わないと言うことも言えるが、また同時に教育会にそれだけ進歩がなかったと言うことをも示しているといういい方もできる。

言語教育では、Lado（1961）をもって嚆矢とするが、わが国でこの翻訳が行われたのは丁度 10 年後の 1971 年である。そこではいわゆる相対評価が新しい概念として紹介されている。すなわち第5節 Refining and using foreign language tests がそれで、本節 22 章はそのものずばり Norms である。それに引き続いて正規分布や順位についても詳述されており、全て集団準拠評価が Lado であったとすれば、同じ大修館書店から出版された J. D. Brown（1996）とその翻訳（和田、1999）では、目標準拠評価が詳述されている。これは橋本の再版とほぼ時期を同じくしている。

一方、一般教育学では、Bloom の教育目標の分類が昭和 48（1973）年に翻訳出版されている。本書は大変影響力の大きな理論を紹介しており、従来の規準設定を紹介しているわけではないが、目標を分類しそれを基準として採用することを提案しているわけで、間接的ではあるが現在まで大変影響力が大きい。本書では、規準的標準、基準に基づく測定、基準の達成度、などの用語で規準設定に関係のある概念が解説されている。最初に出版されたのは、Bloom（1949）であり、認知領域（cognitive domain）と情意領域（affective domain）についてまとめられたのが、Bloom, Engelhart, Furst, Hill & Krathwohl（1956）および Krathwohl, Bloom & Masia（1964）である。Bloom の理論を最も体系的に教育実践に生かそうとしたのが Bloom, Hastings & Madaus（1971）である。日本に最も早く紹介されたのはこの本で、梶田・渋谷・藤田（1973）である。訳者の一人梶田は Bloom に直接師事し、その後この分野で多くの著書を著しているが、2002 年度より実施された新指導要領における観点別評価、目標準拠評価においても、引き続き梶田（2004）等でさらに

独自の貢献をし続けた。Bloom の目標分類は教育一般を目指しているので外国語教育に応用するのは大変な作業である。Valette & Disick（1972）は最も早く出版された文献であり、かつ唯一の文献といってもよい。本書でも、“The categories defined by Bloom and his coauthors were designed primarily for the physical and social sciences, history, and literature, rather than for second-language acquisition. For this reason, it has often been difficult in the past to classify foreign-language goals within the Bloom framework”（p. 28）としている。本書の翻訳は大友（監訳）（1980）がある。

規準設定の実践および、理論を応用しようとする試みの歴史

前節に引き続き、本節では実践の歴史を辿る。最初にわが国の文部省（当時）、次に、国内外の規準設定の実践に大きな影響を与えたBloomの教育目標の分類に対する批判、現在大きな影響を及ぼし続けているCEFR、今後応用の可能性のあるファジー理論を概略する。

文部省の実践

我が国における規準設定が特にはっきりと意識され、そして実践に移されたのは 2004 年(平成 16 年)の新学習指導要領の導入においてである。勿論それまでも研究は行われており、特に現役の教育 者向けにわかりやすく書かれた概論書なども出版されていた。教育学における評価研究では橋本重治（1976、2003 他多数）の一連の著作物があるし、外国語教育評価研究でもバレッテ&デシックの翻訳出版も行われている(大友、1980)。これ以前も「相対評価」に代わる評価の枠組みとして「絶対評価」を実践しようとする試みもなされたが、これは成績で人数を限ることなく出来がよければ何人に 5 を与えてもかまわないというような大変ナイーブな解釈に基づいたものだった（渡部、2011）。

2002 年度より実践された文部科学省新指導要領では、コミュニケーションへの関心・意欲・態度、表現の能力、理解の能力、言語・文化についての知識・理解これら4つの観点から学習者の達成度を評価する観点別評価、そしてそれぞれの観点についてA（十分に満足できる状況）、B（おおむね満足できる状況）、C（努力を要する状況）という基準で評価することが求められた。さらに、各学校において、評価規準・判定基準について詳細に記載することが求められた。国立教育研究所より評価規準の作成のための参考資料、評価方法等の工夫改善のための参考資料を公にされている（http://www.nier.go.jp/kaihatsu/shidousiryou.html）。この動きは CEFR の普及とも相俟って can-do statement として達成度を作成することが各学校単位で行われた。この動向は現在でも引き続き行われている。ただ、その効果については、生徒に意図が必ずしも伝わっておらず、十分に機能していないことが報告されている（渡部、2004）。

2004 年以前に実質的な基盤を提供したのは、Benjamin Bloom の著作といわゆる教育目標の分類である。1950 年代にすでに行われていたが、我が国では梶田叡一他（1974）の翻訳出版をも

って嚆矢とする。ここでは目標準拠という考えが行われ、これが 2004 年の以降の動きに直接つながっているということができる。

2004 年以降の実践は規準か基準かといった区別が話題になり、教育研究所なども各教科について基準と規準の実例を記載し公にした。また観点別評価も合わせて注目を浴びることとなった。さらに、ヨーロッパでは Council of Europe が中心となり欧州共通枠(Common European Framework of Reference)、いわゆる CEFR が我が国でも広く知られるようになり、また研究も盛んに行われるようになってきた。初期の研究成果については例えば Morrow（2004）があり、それ以降は Cambridge ESOL Examinations や ALTE などのサイトに報告されている。以前は主に外国語としての英語教育に限定されてきたが、今回は他の外国語教育についても研究や実践が行われている。特に外国語教育としての日本語教育では、日本語版スタンダードなども整備されるようになってきた（国際交流基金、2011）。

このような評価方法等については特段新しいことではなく、すでに昭和 22 年文部省（当時）発行の「学習指導要領・英語編 ［試案］」（小泉、2001）において今から見ても驚くほどの先見性に富んだ内容であることが指摘されている（渡部、2011）。その包括的なことは大村・高梨・出来・佐々木（1980a）の浩瀚な資料集でも、和文で 34 ページ、英文で実に 450 ページ以上にもわたって収録されており、その中には「適当なテストまたは試験の基準」として和文では 16 ページの解説がある。Chapter VII Evaluation of pupil progress in English language、I. Definition, Purposes, and Principles of Evaluation, II. Criteria of a Good Test or Examination、III. Measureable elements、IV. Methods: How to measrue、V. Scoring or keeping record 等各節のタイトルをみるだけでも充実ぶりが覗えるのではないだろうか。

Bloom への先批判

上に述べたように、教育評価における規準設定の実践においては米国の Benjamin Bloom の影響が顕著である。しかしながら、近年 Bloom の分類を見治そうとする動きがある。Anderson 他（2001）は Bloom の延長線上にあるもので、知識、などからなる構成要素はほぼそのまま踏襲し、組織を変更したものである。一方、Marzano & Kendall（2007）は心理学や教育学の実証研究に基づき実際の運用のプロセスを考慮したものとなっている。Bloom 他が単なる分類(taxonomy)でありいわば静的なものであったのに対し、後者は知識や運動能力(psychomotor)まで含めたより包括的で動的な理論である。ただし、Bloom の場合と同様に、外国語教育は想定されておらず、今後の研究や実践を待たなければならない。

CEFR の応用研究

近年我が国における規準設計は CEFR をモデルにしたものが多く行われている。その多くは Can-do statement という形式で各レベルで対象言語を使って何ができるかを設定したものである

（Schmidt 他、2010）。しかしながら、各レベルでどのような言語構造が運用できるのかといった、言語そのものを対象とした研究はほとんどない。投野などのコーパスを援用した研究の成果がまたれるところである（2010）。

ファジー論の規準設定への応用

ユニークなアプローチとしては、ファジー理論を規準設定に適用しようとした Fourali（1994）、Fourali（1997）や Baas & Kwakernaak（1977）などがある。これらは Zadeh（1965）によって形式化されたファジー集合論を適用したものであるが、各レベル間の境界はファジーであるという能力観に基づいており、また各レベルにおける能力記述するのが言語によるものであるのであるから、言語学における同様の傾向、すなわち近年急速な発展を遂げている認知言語学の基礎となった、Lakoff（1973）や Labov（1973）などと同じ流れにあり、ファジー理論を言語能力の規準設定にも適用する意義は十分に認められる。我が国では山下（2001）等の試みがあるが、言語能力の規準設定についてはまだ適用例がない。例外としては外国語学習ストラテジーを検証するために Oxford（1990）の SILL（Strategy Inventory for Language Learning）をファジー理論で分析した例などがある。これらを発展させ、現在行われている方法との比較検討をする意義は十分に認められる。Jin, Mak & Zhou（2012）は最も新しくかつ言語テストにおいて初めて行われた画期的な研究成果である。

残された課題

学習指導要領試案が発行されたのが昭和 22（1947）年、このような試みを見ても外国語教育におけるテストや評価についてはむしろ退歩しているのではないかとすら思われるのである。しかし、それは飽くまで印象にすぎない。限られた資料を見るだけでも、

○指導要領や教科書のような公にされた資料や文書および教材

○理論的基盤

○実践

これら3つの要素が乖離していることが読み取れるのである。すなわち、理論→学習指導要領→教科書→ 指導実践→ 学習の成果 ←到達の評価、これらの間に有機的な関係がみられず、それぞれが時代の空気や、欧米の教育・評価研究の動向、いわゆる有識者の考え等々の不確定な要因によって独立して変更されてゆくように見受けられる。本プロジェクトの母体である日本英語検定協会は、これまでわが国の英語教育に大きく貢献してきたことは疑いようもない。しかし、1963年に発足して以来、具体的に学校教育にどのような形でどの程度資してきたのかがすぐに見てとれる資料がない。延受験者数、各級の受験者数、合格者数等々の情報はわかるにしても、英語検

定試験のために各学校でどのような使われ方がされているのか、受験者はどのような準備学習をしているのか、学校でおこなわれている英語の授業を強化しているのか、あるいは別の受験勉強が必要なのか、他の入試との整合性はどうなのか等々、現実を見てとれるそしてすぐにだれにでも手に入る資料がないのである。

上述したのは、教育すべてにわたって言えることであるが、規準設定というような具体的な特定化されたテーマに焦点を当てるとそれが、さらに具体的に見えてくるのである。理論研究の深化と発展が必要なことは言うまでもないが、同時に理論研究の成果の普及が必要である（Watanabe, 2011）。さらに、英検をはじめとして各種テストで実際にどのような規準の設定が行われているのかを実証的に検証する必要がある。各種テストにはさらに学校教育における成績、入学試験、等々を含むべきだろう。現実に行われていることが分かって初めて、その問題点が明らかになり、解決の手立ても見つかるであろう。そして、教科書、指導要領、参考資料等にどのような内容を記載すべきなのかを明らかにする必要がある。このような研究をするにあたっては、例えば Henrichsen（1989）の教育イノベーションの普及理論などを援用しながら考察する価値のあるテーマであると思われる。これについては、すでに Wall（1996; 2005）等のテストの波及効果を検証した報告があるし、また言語教育一般でも、Markee（1996）等がある。また教育一般の分野では Fullan（2007）等が教育理論の普及にはどのような条件が必要なのかを調査した報告を行っている。

これらの研究は直接規準設定に関するものではないが、これまでの研究の成果が教育現場や、テストの実践に十分に生かされていない歴史を見れば、理論と実践の橋渡しをするための条件を探ることには十分に意味のあることは明らかである。規準設定の研究と同時に進めるべきであることを実践の歴史は示しているのである。

参考文献

Anderson, L. W., Krathwohl, D. R., Airasina, 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.* New York: Longman.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (ed.), *Educational measurement, second edition*. (pp. 508-600). Washington, DC: American Council on Education.

Baas, S. & Kwakernaak, H. (1977). Rating and ranking of multiple-aspect alternatives using fuzzy sets. *Automatica 13,* 47 – 58.

Barnwell, D. P. (1996). *A history of foreign language testing in the United States: From its beginnings to the present*. Arizona:Bilingual Press.

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.

Brown, J. D. (1996). *Testing in language programs.* New Jersey: Prentice Hall Regents.（和田稔訳、(1999)『言語テストの基礎知識―正しい問題作成・評価のために』東京：大修館書店．）

Durm, M. W. (1993). An A is not an A is not an A: A history of grading. *The Educational Forum, 57* (Spring), 294 – 297. http://www.indiana.edu/~educy520/sec6342/week_07/durm93.pdf

Ebel, R. L. (1972). Some limitations of criterion-referenced measurement. In Brach, G. H., Hopkins, K. D., & Stanley, J. C. (Eds.). *Perspectives in educational and psychological measurement*. New York: Prentice-Hall.

Fourali, C. (1994). Fuzzy logic and the quality of assessment of portfolios. *Fuzzy sets and systems*, 68, 123 – 139.

Fourali, C. (1997). Using Fuzzy logic in educational measurement: The case of portfolio assessment. *Evaluation and research in education*, Vo. 11, No 3, 129 – 148.

Fullan, M. (2007). *The new meaning of educational changes, fourth edition*. London: Routledge.

Henrichsen, (1989). *Diffusion of innovations in English language teaching: The ELEC effort in Japan, 1956 – 1968. New York: Greenwood.*

Hunt, L. H. (2009). *Grade inflation: Academic standards in higher education*. New York: the State University of New York Press.

Jin, T., Mak, B., & Zhou, P. (2012). Confidence scoring of speaking performance: How does fuzziness become exact? *Language Testing, 29, 1,* 43 – 65.

Johnson, V. E. (2003). *Grade inflation: A crisis in college education*. New York: Springer.

Krathwohl, D. R., Bloom, B. S., & Masia, B. B. (1964). *Taxonomy of educational objectives: Handbook II: The affective domain.* New York: David McKay.

Labov, W. (1973). The boundaries of words and their meanings. In Bailey, J. N. & Shuy, R. W. (eds). *New ways of analysing variation in English*. (pp. 340 – 373), Washington D.C.: Georgetown University Press.

Lado, R. (1961) *Language testing―The construction and use of foreign language tests*. London: Longmans.（門司勝・本田漠・吉田一衛・松畑熙一（訳）(1971)『言語テスト－外国語テストの作成とその利用』、東京：大修館書店．）

Lakoff, G. (1973). Hedges: A study in meaning criteria and the logic of fuzzy concepts. *Journal of philosophical logic* 2, 458 – 508.

Markee, N. (1996). *Managing curricular innovation*. Cambridge: Cambridge University Press.

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

Morrow, K. (2004). *Insights from the Common European Framework*. Oxford: Oxford University Press.

Nedelsky, L. (1954). Passing score and length for domain-referenced measures. *Review of Educational Research*, *14, 1,* 3 – 19.

Oxford, R. (1990). *Language learning strategies: What every teacher should know.* Rowley: MA.: Newbury House.

Sasaki, M. (2008) The 150-year history of English language assessment in Japanese education. *Language Testing*, 25 (1), 63-83.

Schmidt, M. G., Naganuma, N., O'Dwyer, F., Imig, A., & Sakai, K. (2010) *Can-do statements in language education in Japan and beyond – Applications of the CEFR*. Tokyo: Asahi Press.

Spolsky, B.(1995). *Measured words: The development of objective language teaching*. Oxford: Oxford University Press.

Tanaka, M. (2008) *A history of English language testing in Japan*. Hiroshima: Keisuisha.

Valette, R. M. & Disick, R. S. (1972). *Modern language performance objectives and individualization: A handbook.* New York: Harcourt Brace Javanovich, Inc.（大友賢二監訳）(1980)『英語学習到達目標の設定』玉川大学出版.

Wall, D. (1996). Introducing new tests into traditional systems: insights from general education and from innovation theory. *Language Testing* 13/3: 334-354.

Wall, D. (2005). *The Impact of High-Stakes Testing on Classroom Teaching: A Case Study Using Insights from Testing and Innovation Theory*. Studies in Language Testing, Volume 22. Cambridge ESOL and Cambridge University Press.

Zadeh, L. A. (1965). Fuzzy sets. *Information and control* 8, 338 – 353.

橋本重治(1959)『教育評価法総説』、東京:金子書房.

橋本重治(2000)『到達度評価の研究:その方法と技術』、東京:図書文化.

伊ヶ崎暁生・松島栄一(編)(1990)『日本教育史年表』、東京:三省堂.

下川耿史(編) (2000).『明治・大正家庭史年表』、東京:河出書房新社.

梶田叡一・渋谷憲一・藤田恵璽(訳)(1974)『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』、東京:第一法規.

梶田叡(2004)『絶対評価＜目標準拠評価＞とは何か』、東京:小学館.

梶田叡一・渋谷憲一・藤田 恵璽.(1973)『教育評価法ハンドブックー教科学習の形成的評価と総括的評価』、東京:第一法規.

バレッテ、R. M. & ディシック、R. S (1980)(大友賢二監訳)『英語学習到達目標の設定』、東京:玉川大学出版. Modern language performance objectives and individualization: A

handbook. (1972). Harcourt Brace Javanovich, Inc.
梶田叡一・渋谷憲一・藤田恵璽(訳)(1974)『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』、東京:第一法規.
小泉仁(2001)「第6章学習指導要領における英語教育観の変遷」英語教員研修研究会.『現職英語教員の教育研修の実態と将来像に関する総合的研究』. 平成 12 年度科学究費補助金基盤研究(B)12480055 研究成果報告書.(共著)118-154/ http://www.cuc.ac.jp/~shien/terg/koizumi%5B1%5D.html(2010 年3月 23 日引用).
大村喜吉・高梨健吉・出来成訓(編集)(1980a)『英語教育史資料 第1巻 英語教育課程の変遷』東京;東京法令出版.
大村喜吉・高梨健吉・出来成訓(編集)(1980b)『英語教育史資料 第2巻 英語教育理論・実践・論争史』、東京;東京法令出版.
投野由紀夫(2010)CEFR 準拠の日本版英語到達指標の策定へ.『英語教育』10 月増刊号、60－63 頁.
渡部良典(2004)英語教育における目標基準準拠評価(絶対評価)の動機付け効果の検証. 科学研究助成金、2003 年度～2004 年度、研究課題番号:15520346.
渡部良典(2011)「英語学力評価論」、石川祥一・西田正・斉田智里(編)『英語教育学体系第 13 巻テスティングと評価－4技能の測定から大学入試まで』(pp. 4 －29)、東京:大修館書店.
山下元(2001)『ファジィ理論を応用した教育評価システムの研究 研究報告書』文部省科学研究費基盤研究(C)(10680198).

1.3. Practice in Standard Setting in Japan

Yoshinori Watanabe

Abstract

The present paper reports on the range of attempts to standard-setting practice in context, where a foreign language is being taught. The cases dealt with include amongst many others the pilot version of the MEXT guideline (1947), an attempt by Valette & Disick (1972) to apply Bloom's taxonomy of educational objectives, and CLIL (Content and Language Integrated Learning). A new development will be presented and discussed in relation to the revised version of Bloom's educational taxonomy, by referring to the work of Anderson, et al (2000) and Marzano and Kendall (2007). The report concludes by summarizing a range of issues that would usefully be addressed in the future research in the field exploring the range of issues relating to foreign language teaching and learning in general and language assessment in particular.

１．３．　日本における規準設定の応用と実践

渡部　良典

規準設定についてはすでに多くの実践が行われている。本節の課題は「応用と実践」ということになっているが、過去の実践が何らかの理論の応用と言えるかどうかということには多々疑問が残る。最初に文部科学省の学習指導要領に関して規準設定に焦点を置きながら考察する。次に、今後規準設定の応用が待たれる CLIL（Content Language Integrated Learning）、最後に教育一般において大きな影響を与え続けている Bloom の教育目標の分類（Taxonomy of educational objectives）の改訂版2点について考察する。

文部科学省の学習指導要領

1947 年度試案

正規の学校教育における英語教育においてはカリキュラムや学習指導要領等は存在しなかった。明治 35（1902）年の「中学校教授要目」を持って嚆矢とする。教材の配列等は勿論行われているが、*First Book of English Composition* や *National Reader* 等ある特定の教材に沿って配列してあるだけなので、どのような原理に基づいているのかを知るためには、原点にあたって考察しなければならない。例えば国会図書館の近代デジタルライブラリー（http://kindai.ndl.go.jp/index.html）で確認することができるが、アルファベットから始まり、It is a dog.へと移り、See the boy and the dog. The boy and the dog run. (No 1)。全 55 ページ中 40 ページから始まる short stories では、John and his cat Dick do not like rats. They catch all they can. One time, John set a trap to catch some, and then went away and hid with Dick. … などと続く。言語中心の配列であり、学習者がどれだけ修得したかということについての記載はない。

規準設定について、成績の付け方について明示されるのは昭和 22（1947）年の学習指導要領英語編［試案］であることは別に述べたが、その内容の詳細なることもさることながら、さらに画期的なのはテストや評価についての記載があることであり、さらに画期的なのは pedagogical basis of gradation として、学習者の習得という視点から見た教材の配列の詳細を述べていることである。これは占領下におかれた日本に米国から招聘された米国教育視察団が 1946 年と 1950 年に亘って視察した、その結果に基づいたものとはいえ、まだ画期的なカリキュラム研究の Tyler（1949）が出版される前のことである。

また特に学習者の診断のためにテスト結果を使う必要があることなどを強調し、成績のシステムについても記載がある。パーセンテージによるシステム（すなわち集団準拠評価）と A, B, C による

表点（絶対評価）についてもふれており、後者については “Every effort should be made to develop descriptions to go with the letter marks.”（大村他、1980a、p. 402）とあり、今日の目標準拠評価に近い考え方がすでに記載されている。指導要領試案の典拠となったのは、当時の心理教育評価の分野で当時としては画期的な著書、Remmers & Gage（1943）、Greene, Jorgensen & Gerberich（1916）、および Burton（1944）である。わが国において最初の応用実践例ということができる。特に、Remmers & Gage（1943）は最近も Gill & Schlossman（2003）などで基本文献として引用されており、この分野における影響力を示している。

2011 年現在

現在我が国の中学校、高等学校で行われているのはいわゆる観点別評価、目標準拠評価である。すなわち、英語能力を次の4つの観点を基準にして判定するのである。

- コミュニケーションへの関心・意欲・態度：コミュニケーションに関心を持ち、積極的に言語活動を行い、コミュニケーションを図ろうとする。
- 表現の能力：初歩的な外国語を用いて、自分の考えや気持ちなど伝えたいことを話したり、書いたりして表現する。
- 理解の能力：初歩的な外国語を聞いたり、読んだりして、話し手や書き手の意向や具体的な内容など相手が伝えようとすることを理解する。
- 言語や文化についての知識・理解：初歩的な外国語の学習を通して、言語やその運用についての知識を身に付けるとともにそれの背景にある文化などを理解している。

評価にあたっては、それぞれの観点についてA（充分に満足できる状況），B（おおむね満足できる状況）、C（努力を要する状況）の3段階を規準として評価し、最終成績つまり学年末の通知表に記入する際には5段階に換算するのである。ここでA，B、Cというのはある程度教員の主観的な判断に基づくこともあるし、テスト得点をもとにして、たとえば60点から70点まではCなどとあらかじめ決めておくのである。後者のA，B，Cについては極めて恣意的となりがちであり、この恣意性の客観化する試みが本報告書全体が問題としていることであり、さらに本章の最後にまとめるとおり、今後の課題でもある。

Valette & Disick (1972)の成果について

観点（基準）の方の淵源については、文部科学省発行の文書をはじめとしたその他の関係文書で明示したものはない。また理論的な裏付けについても明示された文書は皆無である。Lado（1961）およびその翻訳（門司他、1971）にはすでに、言語外の自国の文化、少数集団への態度

等がテストすべき対象とされているところにその萌芽をみることができるということもできるかもしれない。しかしながら、よりはっきりとしているのは、おそらく Bloom の教育目標の分類（Bloom、1949; Bloom, et al, 1956; Bloom, et al, 1964）であろう。Bloomの分類システムについては、その改訂案とともに本章で後述するが、ここでは、Valette & Discik（1972）の *Modern Language Performance Objective and Individualization: A Hadbook* が、おそらくBloomを外国語教育に体系的かつ包括的に応用した唯一の試みであり、後に指摘する通りこのようなシステムがないことが今後に残された課題でもあることを指摘したい。本書は、観点別評価には明確な理論的根拠がないことが問題であると前節で指摘したが、現在の評価システムに欠けている理論的根拠を与えてくれる。紙面が限られており全体像は伝えられないが、図1は外国語学習指導分野と教材内容分類の諸段階との関係を表したものであり、Bloomの用語に置き換えると、認知領域を言語技能に関係づけたものである。

図1　外国語学習指導分野と教材内容分類の諸段階との関係

		聞く	話す	読む	書く	身振り	生活様式・文化	文明	文学
第１段階	識別（内面的行動）	X		X		X	X	X	
機械的技能	模倣（外面的行動）		X		X	X	X		X
第２段階	再認（内面的行動）	X		X		X	X	X	X
知識	再生（外面的行動）		X		X	X	X	X	X
第３段階	受容（内面的行動）	X		X		X	X	X	X
転移	応用（外面的行動）		X		X	X	X		
第４段階	理解（内面的行動）	X		X		X	X	X	X
意思伝達	自己表現（外面的行動）		X		X	X	X		
第５段階	分析（内面的行動）	X		X		X	X	X	X
批評	総合（外面的行動）		X		X	X	X		
	評価（外面的行動）	X		X		X	X	X	X

（大友他、1980、57 頁から引用）

図で X で示されているのが、該当部分である。例えば、「聞く」技能は第一段階の識別（内面的行動）から第5段階の分析（聞きとった内容を分析する）および評価（聞き取った内容を評価する）に関係していることを示している。このような概念図が、情意領域についても展開され、また具体的な指導例、評価例なども記載されているのである。

外国語教育における規準設定のための理論新展

CLIL（内容言語統合型学習）における評価と規準設定

CLIL については、別に章をあらためて詳述するので、ここでは、規準設定研究を応用すべき重

要な課題となることのみを指摘するにとどめる。

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)、認知・思考(Cognition)、コミュニティーの創生 (協学 Community)、これらに加えて言語そのものをCommunicationのCと表し、4つのCの構成要素である(Coyle et al. 2010、48-85頁)。この4つの要素はすなわち4つの観点(基準)を成すということができる。すなわち、CLILにおける課題は、これら4つの独立してはいるが、互い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、ということなのである。

CLIL の歴史は浅く、実証研究も漸く端緒についたところであり、実証研究の数も限られている。その成果は、Marsh and Wolff (2007)、Dalton-Puffer (2007)、de Zarobe and Catalän (2009) 等にまとめられているが、テストや評価に関する実証研究は全く報告されていない。本報告書の別章では、特定の目的のための評価(Language for Specific Purposes)や内容重視の言語指導法(Content-based Language Teaching)等の研究で行われた成果を参照しながら、将来の研究への橋渡しをする。

Anderson 他(2001)による Bloom(1949)の改訂版

学習指導要領ではこれまでしばしば観点別評価が話題となってきた。一般に指導目的の分類(taxonomy)を作成する目的には、1)学習者の達成目標を特定すること、2)習得された技能を使うべき条件や場面を特定すること、そして3)習得のレベル、すなわち規準の設定である(Mager、1962)。教育指導の分野では大きな影響を与え続けている教育目標の分類について考察することには、規準設定についても意義のあることだと思う。

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。簡単に図式化すると次のようになる。

図2 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)

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

図3 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)

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)

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）知識に事実に関する知識（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、延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。

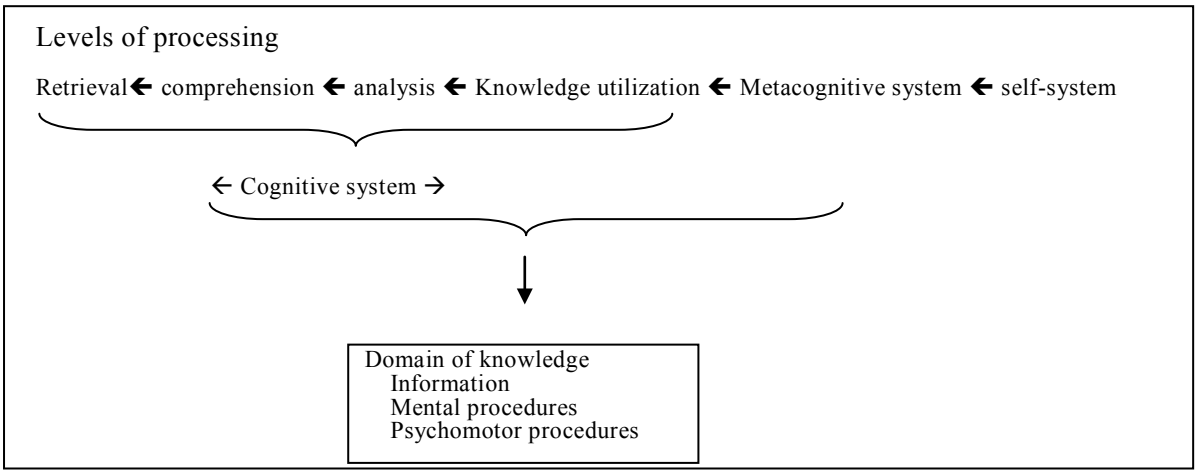
しかしながら、図から容易に見て取れるが、やはり6つの認知プロセスが記憶する（1のレベル）から創造する（6のレベル）に移るにつれて複雑になるという階層をなすという前提はかわっていない。したがって、改訂版の問題点として、本当に1から6に移るにつれて困難な認知プロセスを経ているのかどうかについては、必ずしも実証的に証明されているわけではなく、かなり恣意的であるといわなければならない。これはすなわち、あくまで分類であり習得の理論ではないことを示している。また、改訂版で対象となっているのは認知領域（cognitive domain）だけであり、情意領域（affective domain）は全く考慮されて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

Marzano & Kendall (2007)の新版

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、Andesron et al（2001）とは異なり、情意領域が自己システム思考（self-system thinking）として組み込まれ、大変重要な役割を果たすとしている。また知識についても、情報

（information）、心的手続き（mental procedures）、運動神経上の手続き（psychomotor procedures）から構成されるとする。それぞれの、要素の関係は単なる層（hierarchy）や分類（taxonomy）の代わりに使われているのが、それぞれの要素の支配関係（control）という概念である。

図4　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のである。

残された課題

Anderson 他の改訂版と Marzano & Kendall 版両者ともそれぞれ意義があるが、実践および教育指導への応用ということになると、出版された報告や研究論文の数からしても Marzano & Kendall は Anderson に遠く及ばない。それは、前者にはすでに 50 年に亘る蓄積があるということだけではないように思われる。実践に移すにはなによりも単純でなければならないが、Marzano & Kendall はまだ複雑すぎてすぐに使える理論にはなっていないように思われる。日本においては、教育学関係の文献だけを見ても、石井（2004；2005）、高橋（2001）、尾崎（2009）等の文献があるが、どれも内容の詳細な紹介にとどまっている。中村（2011）などの小学校家庭科への Marzano 理論の応用などの調査があるが、外国語教育については皆無である。Bloom 版については Valette & Disick（1972）がありその翻訳も出版されていることを別章で紹介した。しかしながら、その教育効果の調査や実践例のまとまった報告はない。Anderson らの改訂版の言語教育への応用については、Coyle et al.（2010）等に見ることができるが、しかし役に立つ限りにおいて参照している段階である。

一方では、部分的イマージョンプログラム（partial immersion program）、CBLT（Content-based

language teaching）等、特定の教科を通して外国語の習得を促す方法や、CLIL（Content Language Integrated Learning）などのように、言語が扱う教科内容（Content）、認知・思考（Cognition）、コミュニティーの創生（協学 Community）、これらに加えて言語そのものをCommunicationのCと表し、4つのCの構成要素である(Coyle et al. 2010、pp. 48-85)。これらの分野における、学習評価、教育効果の測定については、Bachman & Palmer（1996）や多くの LSP（Language for Specific Purposes）等で触れられているが、真摯な対象とはならない。まして規準設定の実証研究については皆無である。これも驚くにあたらない。というのは、言語テストの分野では、トピック、個人的な認知能力、他の学習者との協力関係等はバイアスとして排除されてきたからである。しかしながら、CLIL 等で測定したいのは、まさにこれらの誤差を生み出す要因なのである。すなわち、テスト研究では Differential Item Functioning（DIF）（Roever, 2005 等）として好ましくないと判断される項目がまさに必要とされる項目ということになる。

外国語教育においてことさら難しいのは、分類された指導・学習目標（taxonomy）は、外国語そのものが習得の対象でもあり、同時にその言語を使って何らかの作業なり課題なりを遂行できる技能、この2つを同時に測定しているという点である。例えば、Marzano & Kendall 版で、retrieval + psychomotor procedures というのは、少なくとも次の3つの可能性がある。

○ 英語（学習対象の言語）を使って発音の仕方を想起できる
○ 日本語（学習者の母語）を使って英語（学習対象の言語）の発音の仕方を想起できる
○ 例えば水泳などの運動神経を使う動作について英語（学習対象の言語）を使って想起できる

CLIL ではさらに、認知能力、集団形成能力等も測定の対象となるので、更に複雑である。これらを十分に吟味しながら教育目標を設定し、さらに規準を設定するという作業を行わなければならない。しかしながら、改訂版にせよ、Marzano & Kendall にせよ、応用例には外国語に対する例はほとんど皆無といってよい。私たちはBloom版についてValette & Discik（1972）が行った作業を改訂版についても行う必要があるのである。

参考文献

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.

Bachman, L. & Palmer, A. (1996) *Language testing in practice*. Oxford: Oxford University Press.

Bachman, L. & Palmer, A. (1996) *Language testing in practice*. Oxford: Oxford University Press.

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.

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.

Clapham, C. (1996). *The development of IELTS: A study of the effect of background knowledge on reading comprehension*. Cambridge: Cambridge University Press.

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

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

Dalton-Puffer, C. (2007). *Discourse in content and language integrated learning (CLIL) classrooms*. Amsterdam: John Benjamins.

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

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

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.

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

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

Lado, R. (1961) *Language testing—The construction and use of foreign language tests*. London: Longmans.（門司勝・本田漠・吉田一衛・松畑熙一（訳）(1971)『言語テスト－外国語テストの作成とその利用』、東京:大修館書店.）

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*, *second edition*. Thousand Oaks, CA: Corwin Press.

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

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.

Roever, C. (2005). "That's not fair!" Fairness, bias, and differential item functioning in language testing." *SLS Browbag, 9/15/2005*. http://www2.hawaii.edu/~roever/brownbag.pdf

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

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

Valette, R. M. & Disick, R. S. (1972). *Modern language performance objectives and individualization: A handbook*. New York: Harcourt Brace Javanovich, Inc.

Valette, R. M. & Disick, R. S. (1972). *Modern language performance objectives and individualization: A handbook*. New York: Harcourt Brace Javanovich, Inc.（大友賢二監訳）（1980)『英語学習到達目標の設定』玉川大学出版.

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:パフォーマンス評価の位置づけを中心に『京都大学大学院教育学研究科紀要』50、172－185.

石井英真（2005)アメリカの思考教授研究における教育目標論の展開:R. J. マルザーノの「学習次元」の検討を中心に『京都大学大学院教育学研究科紀要』51、302 – 315. http://hdl.handle.net/2433/57526.

中村喜久江（2011）. 子どもの食物選択力を形成する小学校家庭科学習の検討.『広島大学大学院教育学研究科紀要』1, 60、91 – 100.

尾崎博美（2009)教育目的論における「教育目標」概念の分析『東北大学大学院教育研究科研究年報』58、1、13－32. http://hd.handle.net/10097/45517.

大村喜吉・高梨健吉・出来成訓（編集）（1980a)『英語教育史資料　第1巻　英語教育課程の変遷』東京;東京法令出版.

大村喜吉・高梨健吉・出来成訓（編集）（1980b)『英語教育史資料　第2巻　英語教育理論・実践・論争史』、東京;東京法令出版.

1.4. Standard-setting Method: Development of the Bookmark Method

Kenji Ohtomo

Abstract

The Bookmark Method, or the Bookmark Procedure, is one of the methods to establish test cut scores. It was developed in 1966 by CTB/McGraw-Hill researchers Lewis, Mitzel, and Green to address perceived shortcomings in the modified Angoff method, the most commonly used procedure at the time.

The merits of the Bookmark Method are (1) the integration of selected-response and constructed-response item formats, (2) the simplification of judgmental tasks by reducing or refocusing the cognitive load of judges, (3) the connection of the judgment task to setting cut scores to the measurement model, (4) the linking of test content with performance level descriptors, and (5) the ability to retain test data calibrated using IRT (Item Response Theory).

The report concludes with three suggestions for future research on the Bookmark Method. The first is that further research is needed in the field of response probability values. Secondly, comparative research between the Angoff Method and Bookmark Method is needed Finally, further research is needed in the application of the *Cito* Variation of the Bookmark Method.

１．４．規準設定法：Bookmark Method の開発と発展

大友賢二

Bookmark Method の意味と開発（１）

Bookmark Method, Bookmark Approach, または Bookmark Procedure と呼ばれているものは、standard setting, つまり、「分割点の設定」を行う方法の一つである。 その目的に到達するための手段として用いられるのが bookmark、つまり、「(本の）しおり」である。そのしおりを使って、分割点を決定するということである。筆者は、これを「しおり推定法」と呼んでいる。分割点の設定は、きわめて困難である。たとえば、ある学校の入学には、このテストで、何点取らなければならない。ある免許状を取得するには、その試験で、６０点以上をとらないと認められない。など、暗黙の了解があるが、あらためて、「その理由は」と問いただされると、その説明をするのは、容易ではない。では、５５点ではいけないのかという課題がある。採点者の採点の甘さ、辛さによっても６０点の規準は異なるのではないかという課題もある。これを、「いい加減な、でたらめな方法」（'feet on the desk' procedure）と指摘している一人としては、Nitko (1983)を挙げることができる。このように、分割点、つまり、目標の到達と未到達をどう設定することができるかは、きわめて、難問ではあるが、その方法の一つである「しおり推定法」が、ここでの課題である。

Mitzel, Lewis, and Green (2001) によれば、この bookmark method が、テストの世界に初めて紹介されたのは、1996 という年と言われている。1965-1966 は、筆者が Washington, D.C. の Georgetown 大学で Robert Lado 教授のもとで、言語テストの研究をしていたが、この bookmark method も、item response theory も、残念ながら、まったく耳には入ってこなかった。1996 年６月、Council of Chief State School Officers National Conference on Large –Scale Assessment での Symposium “ IRT-based standard setting procedures utilizing behavioral anchoring” が行われたが、その中で、standard setting: a bookmark approach というタイトルで Lewis, D.M., Mitzel, H. C, & Green, D.R.(1996) が紹介したのが、「はじめ」とされている。

この bookmark method の特徴は、大友 (2009)で示されているように、つぎの５つを挙げることができる。第１は、ほかの設定法では見られなかった「項目応答理論（item response theory）」の応用があったこと。第２は、複数の分割点を設定することが可能であること。第３は、解答構築式項目（constructed-response item)でも解答選択式項目（selected-response item)でも、いずれの場合でも利用することが可能であること。第４は、設定作業は、きわめて簡単であること。そして、第５は、テスト項目の内容が、分割点設定の作業に反映で

きるということである。しかし、その最も大きな特徴は、項目応答理論の応用であろう。古典的テスト理論が持っている多くの課題を、項目応答理論の持つ有力な特徴「不変性（invariance）」という武器で克服した点である。

Bookmark Method の意味と開発（２）

分割点設定研究に関する初めの時期については、Linn, R.L.(Ed.)(1989)の Jaeger, R.M.(1989)などでもまとめられている。その種類としては、「テスト中心モデル（test-centered model）」と「受験者中心モデル（examinee-centered model）」がある。テスト中心モデルとしては、Nedelsky 法、Angoff 法、Ebel 法などがあり、受験者中心モデルとしては、境界線グループ法（borderline-group procedure）、グループ対照法 (contrasting-groups procedure) などがある。その内容としては、そののちの Brennan, R.L. (Ed.)(2006) の Hambleton, R.K. & Pitoniak, M.J. (2006) などにもまとめられているので、それを参考に検討することを期待するものである。ここでは、この中で Bookmark Method と比較して検討されることが最も多い Angoff 法も含めて、Bookmark Method の手法の基本的考え方に触れることとする。

William H. Angoff (1919-1993) は、４０年もの間、米国のＥＴＳ（Educational Testing Service）でテスト研究に携わった。彼の研究の中で、最も知られている業績の一つは、Thorndike, R. L. (Ed.) (1971) のなかの”Scale, Norms, and Equivalent Scores” である。彼の分割点設定法は、修正版がいくつかあるが、その基本的手法は、Angoff (1971) で示されている。池田監訳（2006)は日本語で、次のように示している。

> 「合格や優秀レベルをとるのに必要な、最低限の素点を求める為の体系的な手順は、次のとおりである。仮想の「最低限容認可能な人（minimally acceptable person）」を念頭に置きながら、テスト問題を１つ１つ検討し、その人が、検討中のそれぞれの問題に正答できるかどうかを決める。その仮想上の人が正解する問題には、それぞれ１点、その人が正解できない問題にはそれぞれ０点があたえられると、それらの問題の合計が、「最低限容認可能な人」が取得する素点と同じになる。」(p.259)

また、Bookmark Method という方法では、審査員は、特別に用意された「順序付テスト項目冊子」（0rdered item booklet: OIB）を用いて、最も適切な場所に、自分のしおりを置き、そのデータを基に、分割点を設定するというものである。この OIB には、項目応答理論で分析された項目困難度を使って、0.67 という正答確率（response probability : RP）でこの項目に応えられる能力はいくらかを算出し、それが示されている。Mizel, Lewis, Patz, and

Green (2001)によれば、正答確率が 0.67 の場合、「ある分割点においてその点数をとる生徒は、その分割点に位置する問題に 0.67 の確率で正答することができると解釈できる」(p.260) と述べている。したがって、審査員は、RP が 0.67 よりも低くなると考えられる OIB の最初のページにしおりを置くように指示されている。この審査員のしおりの置いてあるページを検討して、分割点を決定するというのが、この分割点の基本的決定手順である。

Bookmark Method の 意味と開発（３）

Bookmark Method がテスト界にはじめて顔を出したのは、1996年であるが、この方法が他の分割点設定と比較検討されてきたのは、2002年から2003年ごろからである。つまり、さきに示した Cizek (Ed.)(2001)が出版された後であるということができる。ここでは、Bookmark Method 開発の初期として、その２つの例をとりあげることとする。ひとつは、Buckendahl, C.W., Smith, R.W., Impara, J.C. and Plake , B.S. (2002), もう一つは、Wang, N. (2003) である。

Buckendahl, C.W. et al (2002). の比較検討は、A Comparison of Angoff and Bookmark Standard Setting Methods というタイトルで述べられている。この実験材料は midwestern school district の Grade 7 Mathematics Assessment である。その詳細に関しては、この論文を参照願いたいが、結論として述べていることは、次のとおりである。

> Although the Angoff method is more widely used, the Bookmark method has some promising features, specifically in educational setting. Teachers are able to focus on expected performance of the “barely proficient” student without the additional challenge of estimating absolute item difficulty. (p.253)

つまり、絶対的な項目困難度を求める為に複数回の測定をするような手間を必要としないで、目標規準の最低限の能力を持った受験者を見つけ出すことができるから Bookmark Method は有望であると記してある。これは、項目応答理論を用いたデータの大きな利点の影響であろう。

Wang, N.(2003) の論文のタイトルは、Use of the Rasch IRT Model in Standard Setting: An Item-Mapping Method である。この中のはじめの部分だけを見ると、「規準設定におけるラッシュ項目応答理論モデルの使用」で、それは、Bookmark Method の事を指しているかのように見る。しかし、次の説明は Bookmark Method ではなく、An Item-Mapping Method と

なっていることにやや当惑を感じるであろう。この２つの方式の似ているところ、異なっているところの説明は本論文に委ねる。異なっていることの１例は、ＲＰに関しては、The bookmark method では 0.67、the item-mapping method では 0.50 などがある。この二つの方法は、bookmark method は教育測定、item mapping method は免許証発行、という別々の分野で開発発展されたものであるが、ＩＲＴを利用している点では、Angoff 法とは異なる。そして、その実験研究の結論では、' Results indicated that the item-mapping method produced higher inter-judge consistency and achieved grater rater agreement than the Angoff method'(p.231)と述べている。

Bookmark Method の発展（１）

この Bookmark Method は、これまで取り上げた特徴をもっているので、多くの研究機関や教育施設で用いられてきていることは、すでに述べたとおりである。ここでは、実際にこの方法を採用しいて分割点を設定する場合、どのような手順でこれを行うのがよいかということについて、これまで行われてきたいくつかの手順を見ることとする。

この実施手順の例は、Mitzel, H.C., Lewis, D.M., Patz, R.J. and Green, D.R. (2001), The Bookmark Procedure: Psychological Perspectives. In Cizek, G.J. (ed.) , *Setting Performance Standards,* Lawrence Erlbaum Associates, Publishers をはじめ、 たとえば、Cizek, G.J. and Bunch, M.B.(2007), *Standard Setting,* Sage Publishers などに見出すことはできる。しかし、ここでは比較的あたらしい Zieky, M.J., Perie, M. and Livingstone, S.A. (2008) *Cutscore: A Manual for Setting Standards of Performance on Educational and Occupational Tests,* Educational Testing Service で示されている手順の要約を紹介することにする。

１）Ordered Item Booklet を作成する。２）項目困難度順、項目番号、項目困難度などを含む Item Map を作成する。３）テーブル主任を指名する。４）審査員を小グループに分ける。５）審査員に対し、Item Map と Ordered Item Booklet を配布する。６）境界線上にある受験者が正解できると考えられる最後の項目と、境界線上にある受験者が正解できないと考えられる最初の項目の間に「しおり」を置くように伝える。７）審査員が自分のしおりを置いた後で、境界線上にある受験者が正解するであろうと考えられるほかの項目についても、同様の作業を継続するよう伝える。８）もし、分割点が一つ以上設定されるようであれば、すべての審査員に対し、Ordered Item Booklet での作業をさらに継続するよう伝える。９）この「しおり推定法」作業は、最終結論に至るまで、3 度ほど繰り返し行うのが一般的である。１０）2 回目の審査が終わった時点で、自分の担当テーブルの審査員に対し、しおりを置いた場所の範囲はどうなっているかを示す。さらに、すべての審査員の

他のテーブルでのしおりを置いた場所の範囲はどうなっているかを伝える。１１）第3回目の判定を行うようすべての審査員に伝える。１２）分割点を設定する。１３）分割点を設定した項目と「行動水準記述文」(performance level descriptors）を比較検討し、最終決定を行う。

以上の説明の内容を深く検討する場合には、この原文にあたることを薦める。

Bookmark Method の発展（２）

ここでは、これまでも言及している　「順序付テスト項目冊子」(Ordered Item Booklet)の実際に関して、その内容等を示すこととする。

Item 22
Ability level required for a .67 chance of answering
Correctly: 1.725

Passage = Yellowstone

Which of these subheadings most accurately reflects the
information in paragraphs 1 and 2?

Ａ．Effects of the Yellowstone File
Ｂ．Tourism Sinc the Yellowstone Fire
＊Ｃ．News Media Dramatically Reports Fire
D.　Biodiversity in Yellowstone Since the Fire

以上は、Cizek, Bunch, and Koons (2004) で示されている　order item booklet の一つである。ここでは、項目番号（item 22)、67%の正答率が求められる能力水準は、項目応答理論に基づいて 1.725 であることが算出され、それが示されている。また、パラグラフ１と２の情報を最も明らかに示す見出しは、次の４つの中でどれかを問い、その正解は、選択肢Ｃであることを示している。

Page,	Item,	Difficulty,	Discrimination,	Theta @ RP=.67
1	19	-3.395	0.493	-2.550
2	13	-2.770	0.997	-2.352
3	1	-2.757	1.441	-2.468
4	22	-2.409	0.461	-1.505
5	4	-2.282	0.527	-1.492

また、こうしたテスト項目に関する情報は、次のように page (booklet の頁), item（項目番号）, difficulty（項目難易度）, discrimination（項目弁別力）, theta @ RP =.67（67%の正答率が求められる能力水準）で纏められている。

以上が、ordered booklet item parameters and associated Theta vales の表である。

残された課題（１）

この Bookmark Method 研究に残された課題の第 1 は応答確率（response probability）に関するさらなる研究であろう。

テストを受験した場合、どの程度の正答率があった場合にある段階の目標に到達したと判断するのか、ということは、きわめて重要な課題である。これまでの実験結果では、０．５０とする考えもあり、０．６７とする考えもある。大方は、０．６７を指示するのが多い。しかし、なお検討の余地があることは、Hambleton, R.K. and Pitoniak, M.J. (2006: 444), Setting Performance Standards, in Brennan, R.L. (ed.) *Educational Measurement (Fourth Edition),* ACE の次の文から判断することができる。

> The results of Williams and Schuls (2005) suggested that an RP value of .67 is easier for panelists to use and yields more reasonable standards. However, as pointed out by the National Research Council (2005), the selection of RP value, as with any other future of a standard-setting method, call for consideration of the specific test and its uses （p.444）

このような「応答確率」(response probability: RP) の課題と深く関連するもう一つ重大な課題がある。それは、項目困難度とそれに対する応答確率 0.67 がわかっても、その確率を保つための受験者の能力値はどの程度必要かを求めなければならない。たとえば、きわめて低い能力では、そのような困難度を持つ項目を 0.67 の確率で正解することはできないはずである。したがって、「ある受験者がある項目に.67 の確率で正解するためには、その項目の困難度と比べてどの程度高い能力が必要か」という課題を解決する必要がある。

これに関する考察は、わが国の文献では、ほとんど見当たらない。Cizek, G.j., Bunch, M., & Koons, H. (2004) 、Cizek, G.J. (2006), Cizek, G.J., & Bunch, M.B. (2007) の文献では、1PLM,2PLMにおいて、正答確率、困難度、弁別力のパラメターが分かっている場合、能力値は、どうして求めることができるかを推定するための数式が示されている。しかし、何故、そういう数式をもとめることができるか、という式の展開手順を詳しく述べている

説明が見当たらない。これに対しては、大友賢二監修、中村洋一・小泉利恵編集（2009)『言語テスト：目標の到達と未到達』ＥＬＰＡでは、その展開解説が見られる。例えば、項目応答理論の１ＰＬＭでは、基本のモデルを変換すると、　θ= ln (P/(1-P)) + b) となる。したがって、Ｐを 0.67 とすると θ＝0.708＋b となり、受験者が 0.67 の確率で応答選択型項目に正解するために必要な能力は、問題の困難度（ｂ）より、0.708 ロジット大きい値であるということが分かる。

残された課題（２）

この研究に残された課題の第 2 は、Bookmark Method とほかの方法：たとえば、Angoff Method との比較検討であろう。

ふたたび、*Educational Measurement: Issues and Practice* (2011), 30(2)　に目を転じると、2011 年においても、Bookmark Method に関する論文は見られる。しかも、かなり、正の方向でのまとめである。論文は、Peterson, C.H., Schulz, E.M. and Engelhard Jr, G., Reliability and Validity of Bookmark-Based Methods for Standard Setting: Comparisons to Angoff-Based Methods in the National Assessment of Educational Progress.,の中に見出すことができる。

この研究は、１９９２年から２０００年までに用いられた NAEP (National Assessment of Educational Progress) における Angoff-Based Method によるデータと、２００５年から用いられてきている Bookmark-Based Method によるデータを基に見出されている結論である。これも、さらなる検討が必要な、残された課題の一つである。

> 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.(p.3)

残された課題の第 3 は、これまでも言及しているＣＥＦＲ関係の文献にもこの Bookmark Method は言及されていることに目を向けなければならないということである。たとえば、Council of Europe (January, 2009), *Relating Language Examinations to the CEFR: A Manual*, LPD, Strasbourg の 6.9. *Cito* Variation on the Bookmark Method や、Council of Europe (October, 2009), *Reference Supplement to the Manual for Relating Language Examinations to CEFR, Section 1: Cito variation on the bookmark method*, LPD, Strasbourg などにも見られる。

この論文のすべてを述べる紙面の余裕はないが、ここで論じていることの概略は、probability correct と受験者の能力の度合い ability scale との関係を検討していることであ

る。たとえば、prob. correct が.80 以上である場合は、ability scale がどの範囲に入ってくるか。Prob. correct が .50--.80 までの場合は、ability scale がどの範囲に入るか、あるいは prob. correct が .50 以下の場合は、どうかといった検討を重ね、最も適切な分割点を見出そうとしているものである。この例を見ても理解できるよう、Bookmark Method は、ヨーロッパ諸国でも、注目を集めているが、とくに、オランダの *Cito* での研究に関しては、残された第 3 の課題として、さらなる究明がぜひ必要である。

参考文献

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

Brennan, R.L. (Ed.)(2006), *Educational Measurement (Fourth Edition),* ACE

Buckendahi, C.W., Smith, R.W., Impara, J.C. and Pake (202), A Comparison of Angoff and Bookmark Standard Setting Method, *Journal of Educational Measurement, Fall 2002, Vol.39, No.3,* 253-263.

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.* (pp.249-282) .Lawrence Erlbaum Associates Publishers.

Cizek, G.J., & Bunch, M.B. (2007), *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

Hambleton, R.K. and Pitoniak, M.J. (2006), Setting Performance Standards. In Brennan, R.L. (ed.)(2006). *Educational Measurement (Fourth Edition),* (p.444). ACE

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

Karantonis, A. and Sireci, S.G.（2006）The Bookmark Standard-Setting Methods: A Literature Review, *Educational Measurement: Issues and Practice,* Vol. 25, No.1, Spring 2006.

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.

Linn, R.L.(Ed.)(1989) *Educational Measurement (Third Edition)*, ACE

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.

Nitko, A.J. (1983) *Educational Tests and Measurement: An Introduction* .(p.460) Harcourt Brace Jovanovich, Inc.

Perie, M. (2005). *Angoff and Bookmark Methods*. Workshop presented at annual meeting of the National Council on Measurement in Education, Montreal, Canada

Peterson, C.H., Schulz, E.M. and 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 , 2011, 30(2).* 3-14.

Thorndike, R.L. (Ed.) (1971). *Educational Measurement* (2nd Ed.),(pp.508-600). ACE

Wang, N. (2003) Use of the Rasch IRT Model in Standard Setting: An Item-Mapping Method, *Journal of Educational Measurement, Fall 2003, Vol. 40, No.3*. 231-253.

Williams, N.J.and Schuls, E.M. (2005). An investigation of response probability (RP) values used in standard setting. Paper presented at the meeting of the NCME, Montreal, Quebec, Canada.

Zieky, M.J., Perie, M. and Livingstone, S.A. (2008) *Cutscore: A Manual for Setting Standards of Performance on Educational and Occupational Tests,* (pp.105-118). Educational Testing Service

池田　央（日本語版監訳）(2006)『テスト作成ハンドブック』(p.259). Downing, S.M. & Haladyna,T.M. (Eds.)(2006). *Handbook of Test Development,* Lawrence Erlbaum Associates, Inc. の日本語訳)（株）教育測定研究所。

大友賢二（2009)「項目応答理論―TOEFL、TOEIC 等の仕組み―」、『電子情報通信学会誌』、*Vol.92, No.12, 2009,* 1008-1012.

大友賢二監修、中村洋一・小泉利恵編集（2009)『言語テスト：目標の到達と未到達』(pp.101-107). ＥＬＰＡ.

2. Standard Setting in CLIL

2.1 Setting Standards for CLIL Courses: With Reference to the Findings Made in the Field of LSP (Language for Specific Purposes) and CBLT (Content Based Language Teaching)

Yoshinori Watanabe

Abstract

Content Language Integrated Learning is the principle of teaching a foreign language by integrating language, topical knowledge, cognitive skills and community building skills into one unit for instruction. In assessing performance and achievement in CLIL, then, these four elements require related but distinct standards to be set, which is a major challenge for researchers in the field of language assessment and evaluation in general, and those in standard setting in particular. The present and the next chapters outline the basic principle of CLIL and subsequently various problems will be presented and discussed to be solved in the future research. In order to provide a basis for the project, an attempt will be made to gather information from various sources of the past research. In the present section (2.1), the findings that have been made to date in the field exploring the issue of Language for Specific Purposes and Content-Based Language Teaching. The next chapter (2.2) will track down information that has been gained in general education with focus on revised new versions of Bloom's Taxonomy of Educational Objectives.

2. CLIL（内容言語統合型学習 Content and Language Integrated Learning）における規準設定

2．1．LSP（Language for Specific Purposes、特定の目的のための言語）及び CBLT（Content-Based Language Teaching、内容重視言語教育）から CLIL への統合及び発展

渡部良典

CLIL の基本原理

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。実際の運営上は、様々な教育原理・技法を取り入れながら、有機的に統合して授業の実践に取り入れるのである。

CLIL に貢献するような実証研究は行われているのだが、どれも間接的である。そこで、以下には今後行うべき実証研究のための整理を行うことにする。

CLIL の中心をなす考え方は、言語が扱う教科（内容 Content）、認知・思考（Cognition）、コミュニティーの創生（協学 Community）、これらに加えて言語そのものを Communication の C と表し、4 つの C の構成要素である（Coyle et al. 2010、pp. 48-85）。CLIL では、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）と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、Anderson, et al（2001）が行った Benjamin Bloom の教育目標の分類で行われている思考の6段階モデルを援用する（Bloom、1949；Bloom, et al、1956）。池田（渡部・和泉・池田、2011）は Community に「協学」という訳語を当て、Coyle 他（2010）では、Culture を使うようになっているが、日本は必ずしも多文化の環境にあるわけではないので、より汎用性のある Community を使うとしている。ここで言う community とは、席の前後左右の生徒、教室等の比較的小規模な集団から、より規模の大きい学校、近隣、市町村、都道府県、国、さらに地域や地球全体など大規模な集まり全てを含む概念である（同書8頁）。すなわち、教室内でペア・ワークやグループ・ワークを行うだけではなく、世界的な規模で起こっている事象などをトピックとして扱うことも含意している。

CLIL における伝統的な4技能の適合性について

さらに、規準設定も含めた学習・教育評価が CLIL で問題となるのは、従来 CBLT や LSP で行われてきたような話題に関する知識と言語の関係だけではない。CLIL では、実際の場面における言語運用を重視するので、伝統的な4技能の分類は適切であるとはいえない。プレゼンテーションを例にとると、準備の段階で資料を読む必要がある。面接調査が必要なこともあるだろう。パワーポイントのスライドの準備をしなければならないし、原稿を準備するためには書く作業も必要である。実際のプレゼンテーションにおいては話すだけではなく、聞き手からの質問に答えることも必要な技能のひとつだろう。読む、書く、聞く、話すという4つの技能が互いに独立して行われるのではなく、それぞれの役割を果たしながら有機的に全体の言語運用を構成する。TOEFL iBT などではすでに、英文を聞いて、それに類似した内容の英文を読み、両方の情報を統合した文章を書かせたり、話させたりといった統合型のライティング・テスト（integrated writing tasks）や、統合型のスピーキング・テスト（integrated speaking tasks）のセクションが設けられている。読んだ内容と聞いた内容を統合して、自分の意見を述べるなどの一連の作業を要求するテスト課題である。CLIL における評価ではこのような技能の見方をむしろ中心とすべきであろう。そこで、従来のような4技能プラス文法、語彙といった静的な言語能力観ではなく、対人関係（interpersonal）、解釈（interpretive）、発表（presentational）の3つのモードから構成された言語運用能力とみなすなどという、Philips（2008）の動的な言語運用能力観もふさわしい。各モードは以下のように定義される。

- 対人関係モード（interpersonal mode）は、話し言葉にせよ書き言葉にせよ、その場に参与している人同士が情報を与えたり受けたり、感情を表現し、意見を交換するのである。学習者は聞き手であると同時に話し手であり、また書き手であると同時に読み手でもある。双方向のコミュニケーションが要求され意味の交渉（negotiation of meaning）も観察される。言語能力の発達に伴い社会文化的な側面も考慮することができるようになったということを示す証拠が現れる。必ずしも音声によるやり取りに限られないことは、電子メールなどに典型的に現れる。
- 解釈モード（interpretive mode）において学習者は、様々なテーマを扱った書き言葉や話し言葉を理解（understand）し解釈（interpret）しようとする。その際学習者は同時に読み手であり聞き手でもある。目の前にある「テキスト」の書き手や話し手は目の前にはいないし、接することもできない。書きことばあるいは話し言葉で表されたオーセンティックなテキスト、すなわち意味のある言語であり内容のある言語で表されたテキストを扱うのである。学習者は言語とその内容に関する理解を深めるために持っている知識、経験、理解の方略を総動員する。これは単に理解（comprehension）なのではない。自らの考えを形成し、そして自分なりの応答を形成するのである。相手が直接目の前にいないので、すぐに意味を尋ねることはできない。（電子ブックなどはこの欠点を補うことができるかもしれない。）
- 発表モード（presentational mode）では、学習者はさまざまなテーマに関する情報、概念、考えなどを読み手や聞き手に提示する。学習者は、ある程度距離を置いた、直接交流することの全

くできない、あるいは非常に限られたやりとりしかできない相手を想定している。コミュニケーションは特定の相手に向けて行われ、はっきりとした目的があり、その分野におけるルールに従って行われるのである。計画的に形式化された発話行為あるいは書きことばによるdocumentであり、正式に提示する前に、草稿を書き、フィードバックを得て、改訂する機会がある。発表モードが対人モードと異なるのは、即効性（spontaneity）が必要とされないという点である。適切な語彙や表現を注意深く意識的に選択する、ということが重要となる。したがって、読者、聴取者を想定し、適切な語彙表現などを使うように指導することなども重要となる。

これらのモードはそれぞれが完全に独立しているわけではない。これは4技能の場合と同様である。例えば、通常の読解コースでは、テキストを解釈して（解釈モード使用）、それを図式化してグループ内で発表させ（発表モード）、内容について質疑応答させる（対人モード）というような一連の流れは普通であろう。CLIL における評価活動においてもやはりこれら一連の流れを反映して、読解問題のあとに、それについて言語を使わない流れ図などに表現させ、それについてサマリーを書かせるなどというタスクも考えられる。テストでは一つの項目の解答が別の解答の前提とならないようにするという、項目独立（item independence）の原則があるが、これとはまた別の問題である。

CLIL におけるテスト・評価研究

言語テストにおいては、トピックなど言語能力以外の要素は誤差を引き起こす変数として排除しようとするのが普通である。しかしながら、そもそも空の容器のような言語などというものは存在しないように、言語能力だけを測定しようとすること自体が不自然な試みだということができる。したがって、言語テストでは、話題の知識、技能等を合わせて測定するのがむしろ自然だということもできる。これは内容ベースの言語教育（content-based language teaching）や特定の目的のための言語（language for specific purposes）におけるテスト研究では以前から考察の対象とされてきた。しかし、認知能力、科目に関する知識、共生、言語、これらを同時に指導の対象と使用としたCLILにおいてもっとも鮮烈な形で問題となる。

CLIL の歴史は浅く、実証研究も漸く端緒についたところであり、実証研究の数も限られている。その成果は、Marsh and Wolff (2007)、Dalton-Puffer (2007)、de Zarobe and Catalän (2009) 等にまとめられているが、テストや評価に関する実証研究はほとんど報告されていない。特に多いのは、教室で行われている教師と生徒の談話分析に関する研究である（例えば、Puerto, Lacabex & Lecumberri, 2009）。まして規準設定などについては全く手つかずの状態であるといってよい。

もちろん、他のあらゆる教育分野と同様CLILにおいても学習評価や教育評価の重要さは認められており、例えば Mehisto, Marsh & Frigols（2008）、Coyle, Hood & Marsh（2010）、渡部・和泉・池

田（2011）などでも、1章を割いてこのテーマを扱っている。しかしながら、それはどう評価したらよいのかといった技術上の紹介にとどまり、テスト研究の本質に迫った記述からは程遠いものである。まして、規準の設定等に関しては触れられることすらまれである。

CLIL と大変関係の深い、内容ベースの言語教授法（CBLT、Content-based language teaching）などでは、すでに Brinton, Snow, and Wesche （1989）などが言語能力と科目に関する知識の関係について詳細に記載しているし、Bachman and Palmer （1996）もさまざまな教育状況および教育目的を設定して、それぞれに応じて言語能力と話題に関する知識（topical knowledge）の関係を論じている。さらに、LSP（Language for Specific Purposes）のテストに関しては、Douglas（2000）に詳しい。また、学校教育における他教科との関係については、Gottlieb（2006）および、O'Malley & Pierce（1996）が具体的な成績のつけ方まで記載されており参考になる。

CLIL における規準設定のために

先に述べたように、CLIL におけるテスト・評価研究は端緒にさえついていないと言える。ただし、過去の研究から重要な示唆が得られることも確かである。ここでは特に重要だと思われる研究の成果を紹介する。1点は先に触れた Clapham（1996）の実証研究である。Clapham（1996）は背景的知識と読解能力および言語能力の関係を探っており、後に見るように大変有益な発見を報告している。

Clapham によると第一に、トピックが特定化されていればいるほどその知識に知悉している受験者が有利となる。たとえば貿易問題について一般の読者が知っている新聞やニュース報道のレベルの知識を要求するようなテスト課題であれば、知識が多くても少なくてもテスト得点を左右する可能性は低い。第二に、背景的知識を生かすためには言語上の境界（threshold）となるレベルがあると想定される。文法のテストで60%以下の得点だった受験者は背景的知識の恩恵に預かることはなかった。一方60%を超える受験者の場合には背景的知識が高い者の方が言語テストの得点も高い傾向が見られた。第三に、文法能力が低い学生は背景知識は生かせず、一方、文法能力が高い学生は知識が無くても文法能力で不足を補うことができる。背景的知識の量が読解テストの得点に影響するのは中級レベルの学習者である。

以上の発見から、第一に、CLIL を通して外国語を習得するためには、内容に関する知識は大変重要だということがわかる。しかしながら、すでに Alderson（1984）が示唆している通り、単に知識が多ければそれだけ外国語の読解も進むということを示しているわけではない。外国語の習得レベルが中級である場合、認知的にも文化的にも（もしレベル設定があるならば）中級レベルである可能性が高い。外国語で中級レベルというのは最も背景的知識が影響するレベルである。したがって、言語とその他の知識・技能を一体化したテスト課題が必要となる。第二に、CLIL では、テストにおいても教材の選択や課題の設定と同様特定の分野の専門家の協力が必要である。特定分野の

専門家は当該の専門に関したテキストを読む場合、非専門家と比べて、結束性、特に語彙の言い換えによる結束性の理解が優れているという報告も Clapham は行っている。第三に、CLIL は言語、内容、認知についてバランスの取れた実力を想定しているので、高度な内容を扱う場合には、言語は単純な語彙や文法構造を使う、タスクは多肢選択式を使うなどテスト方法やテストの形式に対する配慮が必要である。

話題に関する知識の分析的評価基準の例として Bachman & Palmer (1996)は次のような尺度を使うことを薦めている。

話題情報の知識（knowledge of topical information）
構成概念の理論的定義：関連のある話題情報の知識
構成概念の操作的定義：提示された特定の課題において、関連した話題情報の知識があるか。次の尺度で評価される。

能力／習熟の水準		説明
0	ゼロ	関連した話題情報に関する知識が確認できない 使用範囲：ゼロ。受験者が期待された話題についての知識を示していない。 正確さ：不適切
1	限定的	関連した話題情報に関する知識が限られている 使用範囲：狭い。受験者は対象の話題のほんの一部にしか触れていない。 正確さ：劣っている
2	通常程度	関連した話題情報に関する普通程度の知識 使用範囲：普通 正確さ：上の使用範囲内で普通程度か、優れた正確さを備えている
3	広範に亙る	関連した話題情報に関する広範は知識 使用範囲：広範にわたり、ほとんど制限が認められない 正確さ：優れている
4	完全	関連した話題情報に関する完全な知識があることを示している 使用範囲：関連した話題情報について制限が認められない 正確さ：ほぼ完全に正確な知識があることを示している

CLIL においては、このような規準を、話題に関する知識のみならず、さらに思考・認知能力に関する規準、共生能力に関する規準も設定することになる。

また、Brinton, et al.（1989）は、内容ベースの言語指導においては、テストの際、教員はもちろんのこと学習者にとっても、何を測定するのかを明示することは極めて重要である（Brinton, et al、1989、p. 183）。これは、第二言語習得理論の noticing hypothesis（Schmidt、2001）などから見ても大変重要な指摘である。学習者は言語形式（form）と意味（meaning）の同時に焦点をあてることができないとすれば、言語と内容を同時に測定対象にすると、学習者はテストを受ける上で、そして結果においても著しい不利益を被る可能性もあるのである。

上述した Bachman & Palmer（1996）の内容に関する規準は、作文やスピーチ等のすでに得られた言語データを、最初は言語に特化した規準で、そして次に必要に応じて内容に関

する規準で、例えば、正確な事実を伝えているかなど、について判定するのである。しかしながら、実際には、テスト受験者がテストを受けている際にどちらにより重点を置いていたかによって、結果も大きく異なる可能性がある。実証研究が必要な分野である。また、内容が学校教育で行われている科目の場合は、当該科目の担当者との連携をとりながら、注意深く規準を設定することも重要になるであろう。

残された課題

CLIL における評価は、指導法と同様、Content-based Language Teaching や Language for Specific Purposes から学ぶべきことは多い。しかしながら、結局のところはどれも言語と内容について別の規準を立てることとするにとどまっている。言語知識、話題に関する知識、認知技能、言語を使った相互作用、これらの間に有意味な関係付けを行い、言語の習得とそれに関連した能力および技能を統括した理論が必要である。その理論に基づいたうえで４つの技能にそれぞれ独立してはいるが、互いに関連づけられた能力や技能が測定できるような規準の設定を行うことができるようにしなければならない。このような理論がなければ、当然のことながら妥当性の検証は不可能である。

過去および現在においてこのような理論づけに参考になりそうなのは、Widdowson（1984）、Anderson, et al, (2000)、Marzano & Kendall (2007)、Lee & Sawaki (2009)である。Widdowson（1984）は、認知スタイルと LSP と言語運用能力の関係付けを行うべく、認知心理学、言語学、言語習得の関係を理論化しようとしたものである。Anderson, et al（2000）はすでに概略した通り、一般教育学の分野で認知、知識、技能、これらの３分野について関係付けを行っている。さらに、Marzano & Kendall（2007）は Anderson, et al（2000）と同様に Bloom の教育目標の分類を改訂したものであるが、近年の特に認知心理学の知見を有効に参照しながら、３分野に加えて情意領域も包括した、システムを構築している。Lee & Sawaki（2010）は診断テストに関連して、認知言語学を援用することを提唱している。しかしながら、どれも社会的な技能、すなわち言語を使った他者との相互作用には視点が置かれていないし、Anderson, et al（2000）も Marzano & Kendall（2007）も教育一般を対象としているために、言語学習、特に外国語の習得については、ほんの数例が記載されているだけである。しかし、Bloom が創始した当時に比すれば、言語の学習について例示されるだけでも進歩だといわなければならない。

参考文献

Alderson, J.C. （1984）. Reading in a foreign language: a reading problem or a language problem. In J.C. Alderson, & A. H. Urquhart （Eds.）, Reading in a Foreign Language（ *pp*. 1-27）. London: Longman.

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.

Bachman, L.、& Palmer, A. (2006). *Language testing in practice*. Oxford: Oxford University Press.

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.

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

Clapham, C. (1996). *The development of IELTS: A study of the effect of background knowledge on reading comprehension*. Cambridge: Cambridge University Press.

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

Dalton-Puffer, C. (2007). *Discourse in content and language integrated learning (CLIL) classrooms*. Amsterdam: John Benjamins.

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

Douglas, D. (2000). *Assessing languages for specific purposes*. Cambridge: Cambridge University Press.

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

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

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.

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

O'Malley, J. M. & Pierce, L. V. (1996). *Authentic assessment for English language learners:*

Practical approaches for teachers. Reading: Addison-Wesley.

Phillips, J. K. (2008). Foreign language standards and the contexts of communication. *Language Teaching* 41 (1), 93 – 102.

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.

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

Widdowson, H. G. (1984). *Learning purpose and language use*. Oxford: Oxford University Press.

渡部良典・和泉伸一・池田真（2011）『CLIL 内容言語統合型学習』、上智大学出版.

2.2. Setting Standards for CLIL Courses: With Reference to the Findings Made in the Field of Taxonomy of Educational Objectives and Taxonomy of Learning, Teaching and Assessing

Yoshinori Watanabe

Abstract

To be continued from the previous chapter (2.1), the present chapter attempts to seek for the information that has been gathered in the past research this time in the field of Taxonomy of Educational Objectives with special reference to Anderson et al. (2001)'s revised version of Bloom's Taxonomy of Educational Objectives and Marzano & Kendall (2007)'s New Taxonomy of Educational Objectives. It is expected that the outcome will to be referenced to establish a framework for evaluating in CLIL courses by integrating all the sources in the form of test specifications that will serve as a useful blueprint for validating the standard that will be developed for language education in general.

2．2．教育目標の分類（Taxonomy of Educational Objectives）学習、指導、評価の分類（Taxonomy for Learning, Teaching, and Assessing) の CLIL への援用および統合

渡部良典

2．1ではCLILの原理に引き続き、CLILの学習評価において参考にすべきもっとも有効な先行研究の枠組みの一つとして、Anderson, et al（2000）およびMarzano & Kendall（2007）によるBloomの教育目標の改訂版を紹介しながら、本稿では、この二つの枠組みの CLIL への適用可能性について考察を進める。

CLIL（内容言語統合型学習）における評価と規準設定

先に述べたように、CLIL におけるテスト・評価研究は端緒にさえついていないと言えるが、前節で考察した通り、過去の研究から重要な示唆が得られることも確かである。

CLILのテストにおいては、内容に関する知識、学術技能、コミュニティー創生能力などの能力も検証することが必要となる。当然のことながらこれらを全て一回のテストで評価するということは不可能である。そこで毎回の授業の課題をポートフォリオとして収集しておいたり、スピーチを行わせたり、小論文の作成を学外の課題として課したりという多角的な観点から評価しなければならない。その上で、それぞれの観点について規準を設定することになる。CLIL で何をどのようにテストするかは、コースの目的やカリキュラムの構成次第なので一般化はできないが、以下は典型的な例である。

- 英語の不得意な自然科学系専攻の大学生を対象とした CLIL のコースで地学の基本的な内容をトピックとして扱うこととする。テーマは雲の種類と形成過程についてとし、言語領域は比較と対照の英語表現を適切に使った英文の読解能力と小論文作成能力をテストするものとする。評定は言語（文法、語彙、言語機能）、および雲の種類と形成過程の内容が適切に説明できているかどうかが構成概念である。
- 大学の英語教員養成コース履修の学生を対象とする場合などもあろう。英語で授業を行いかつ英語で資料を作成する英語力を確認したい。対象学習者は高校1年生とする。彼らに辞書の使い方について、書き言葉で解説したマニュアルを準備することができるかどうかその能力を推測する。受験者に、辞書で単語を調べるにはどのようにしたらよいかを生徒に説明させる。あるいは英語教員が高校生の授業を行う能力があるかどうかを測定するというような例もあるだ

ろう。測定内容は模擬講義を行ってもらう。構成概念は「読解指導を行う能力」と定義する。

- 心理学の対人関係をテーマにしたトピックで CLIL 授業を進める。その結果を見るためには、専門の文献を読む能力があるかどうかを測定する、文献で学んだことを実際の対人関係に応用できるかどうかを検証する、同テーマについて自分の考えを述べることができるかどうかを検証する、というようなことになるであろう。読解能力を検証するには、基本的な英語で書かれた教科書や雑誌の記事などからいくつかの文章を選び、それらを基にしてテスト課題を設計し、それぞれに多肢選択問題をつけ理解度を測定するというようなことになるだろう。

以上のような多様な能力をテストして検証するためには、先に見た通り分析的基準を用いるのが妥当である。特に、スピーキングやライティングなどのような生産的な能力の検証にあたっては、言語能力と話題の知識それぞれについて別の評価尺度を準備し採点を行う。例えば、始めに正しい文法を使っているか、語彙は正確で生徒に理解しやすいかどうかなど言語能力の観点から採点する。次に同じ英文を内容について採点する。こうすれば書き言葉を使いこなす能力と話題の知識の両方について独立した推測を行うことができる。

言語という複雑な能力を測定するためには、構成概念を定義し、操作化するというプロセスを経なければならないことは常識であるが、このテスト開発の準備段階は、測定対象が複雑であるほど時間と労力が必要とされる。CLIL のようないくつもの観点が必要とされ、そしてそれぞれの観点に異なる規準を設けるために必要なのは、綿密なる青写真、すなわちテスト細目を準備することである。

CLIL のための規準設定における細目の重要性

前章において、話題に関する知識を評価するためには、言語能力を評価するための規準とは別に設定する必要があるということを例示した。そして、単に別個の規準を設定するだけでは十分ではない、というのは、採点者がテストで何を測定するのかということを十分に認識していないからだということを指摘した。規準を設定しても、実際はテスト得点を解釈するにあたって、言語能力だけを考慮して内容に関する知識の方は採点しないなどということが多々起こりうるのでる。そして、それは単に、忙しいからとか、時間がないからとか、そういった単純な理由に帰されることがあるのだが、実際には、設計段階の問題、すなわち、細目を作成する段階で起こっている問題である。すなわち、テストの設計に際して最初に細目に何を指導の最終目的として、何をどのように測定して評価するのかといったことが充分に明示されておらず、採点規準がはっきりしていない場合にこのような問題が起こりがちである。

このような問題を解決するためには、細目を作成する際に構成概念を各テスト課題、及びそれぞれの応答の採点規準について明確に定義して、テストの採点のガイドラインを作成し、得点を運

用する際に内容と言語の両方が区別して解釈できるようにしておくことが何よりも重要である。Lynch & Davidson（1993）は、目標準拠言語テストの細目を作成するためには、チームワークが必要だとし、テスト開発の際の第一の仕事は、適切なメンバーから構成される開発チーム（CRLTD、Criterion-Referenced Language Test Development）を形成することだとしている。Douglas（2000）は、LSP テストの細目作成は、通常のテスト細目を作成のプロセスと変わらないが、特定の分野に関する専門家をメンバーに加えなければならないとしているが、それは CLIL においても同様である。

一般に、テスト細目に記載すべき事項は、テストの目的、対象受験者（年齢、性別、学習レベル、学年、教育内容等）、問題数、長さ、測定対象言語能力の想定使用場面、テキストの種類、言語技能、言語要素、テスト課題（task）、問題数および得点配分、テスト方法、例題、採点基準および規準等である（Alderson, Clapham & Wall, 1995）。さらに、テスト受験者のための細目には、練習問題が必要であろうし、またテストの専門家向けの細目には、妥当性、信頼性の検証方法なども加える必要もあるだろう。

一言で言えば、CLIL における規準設定に必要なのは、第一に細目を作成することであり、そこに記載すべき構成概念を定義することである。

CLIL の規準設定妥当性研究のための枠組

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（想像する）という認知技能を階層かし、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。

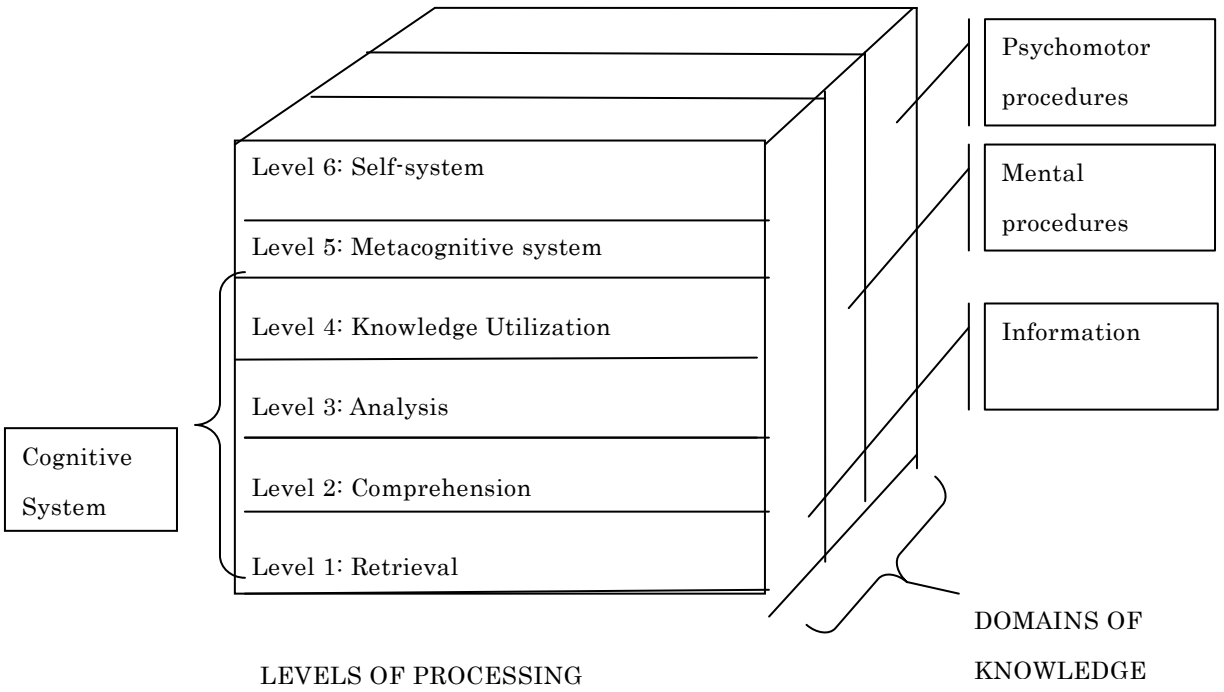
図1 改訂版の分類表

	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. METACOGNITIVE KNOWLEDGE						

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



Marozano & Kendall (2007), p. 13

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。

しかしながら、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 らにすら認めている。（“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）.

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.

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。

残された課題

前節では、Bloom の教育目標の改訂版である Anderson, et al の枠組みや、Marzano & Kendall の理論を援用して CLIL に評価のための理論的基盤を与え、その理論において規準の設定を行い、その妥当性を検証する準備ができるのではないかと、その可能性を示した。しかしながらその際に、最も困難な作業になると思われるのが、私たちの対象が言語であるということである。Bloom にせよ、Anderson, et al にせよ、Marzano & Kendall にせよ、学校教育における科目全般を想定している。その科目群には、言語教育や外国語教育も含まれている。

しかしながら、Anderson, et al（2001）に添付された詳細な適用例を見ても、Marzano & Kendall（2007）の実践版の Marzano & Kendall（2008）を見ても、そこに掲載されている言語教育への適用例は、すでにかなりの言語能力をもった学習者のみが行えるであろう小論文の作成や、大変単純な短音レベルの発音等にすぎない。CLIL は学習者に言語という対象を習得することと、言語を使って何らかの課題をこなす能力と、言語を習得することにより豊かな人間性を育むこと、これらが一体となっているのであり、これらを有機的に評価するということが課せられているのである。

前節では、Widdowson（1984）の認知と LSP の関係、CBLT および LSP におけるテスト・評価に関する成果等の CLIL への適用可能性について示唆した。今後に託された課題は、一般教育学の分野で開発されたシステムを外国語教育に適用することである。すなわち、かつて Valette & Disick（1972）が Bloom の枠組みについて行ったことを、今度は私たちが Bloom の改訂版、新版の Anderson, et al（2001）や Marzano & Kendall（2007）について行うことが必要とされているのである。その成果は第一に、テスト細目と言う形で表す必要がある。その中には、テストの目的、テスト方法等と共に、基準の詳細と規準の設定が記載されるべきである。このような青写真があってはじめて、理論と実践を有機的に関連づけた実証研究が行えるのである。ここで実証研究と言うのは、妥当性検証ということと同義である。

参考文献

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge: Cambridge University Press.

Anderson, L. W., Krathwohl, D. R., Airasian, P. W., Cruikshank, K. A., Mayer, R. E., Pintrich, P. R., Raths, J. & Wittrock, M. C. (Eds.) (2001). *A taxonomy for learning, teaching and*

assessing: A revision of Bloom's taxonomy of educational objectives, compete edition. New York: Addison Wesley Longman, Inc.

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

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

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

Douglas, D. (2000). *Assessing languages for specific purposes*. Cambridge: Cambridge University Press.

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

Marzano, R. J. & Kendall, J. S. (2007). *The new taxonomy of educational objectives*. Thousand 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.

Valette, R. M. & Disick, R. S. (1972). *Modern language performance objectives and individualization: A handbook*. New York: Harcourt Brace.

3. Setting Standards in Can-do Statements

3.1. Proficiency levels and their Can-Do Statements in Japanese Language Proficiency Test (JLPT)

Sukero Ito

Abstract

The JLPT (Japanese Language Proficiency Test) is a standardized test for non-native Japanese language speakers. The test covers reading and listening comprehension as well as language knowledge, such as vocabulary and grammar. The JLPT has five levels: N1, N2, N3, N4 and N5. The easiest level is N5 and the most difficult level is N1. N4 and N5 measure the level of understanding of basic Japanese mainly learned in class. N1and N2 measure the level of understanding of Japanese used in a broad range of scenes in actual everyday life. N3 is a bridging level between N1/N2 and N4/N5.

"JLPT Certificate of Result and Scores"is issued for the purpose of admission to advanced education or finding employment. This certificate contains fail/pass results and scores for each scoring section. It can be used as official proof for schools and companies. However, the certificate does not describe language ability on a scale of levels from N1 to N5. The JLPT office is now trying to develop and validate a set of performance-related scales, describing what learners can actually do in Japanese language or Can-Do Statements. The JLPT needs to define five levels of proficiency in relation to empirically derived difficulty estimates based on test taker's perceptions of what language functions expressed by Can-Do Statements could be successfully performed at each level.

３．Can-Do Statements における規準設定

３．１．日本語能力試験の能力レベルと Can-Do Statements

伊東祐郎

外国語テストと Can-Do Statements

最近の外国語教育では、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、得点に対する意義づけをこれまで以上に重要視するようになってきている。単純に得点という数字を提示するだけでは、満点に対する獲得点数、あるいは達成度しか把握することができない。これでは、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、教師ならびに受験者双方にとって、それ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、いっそうの意味づけをするために、能力基準を設け、それに対する能力の記述文が記されるようになってきた。

大規模試験（high stakes tests）である TOEFL（The Test of English as a Foreign Language）や TOEIC（Test of English for International Communication）などにおいては、言語能力の具体的な内容を記述した"Can-do statements"の研究が行われ、実用化されている。また、ALTE（The Association of Language Testers in Europe＝欧州言語テスト協会、外国語試験の開発・実施機関）では、欧州における外国語能力の共通認定証の促進を図るために、能力評価の基準作りが行われ、外国語能力の認定証の解釈を容易にするために、"ALTE 'Can Do' statements"なるものが開発され活用されている。

本稿では、最近の外国語能力評価の動向を探りながら、具体的な能力記述の作業内容とその結果から得られるいくつかの情報について考察し、今後の言語能力の評価と日本語テスト作成のあり方について考えてみる。

言語知識と言語能力

日本語能力を評価するテストを作成するためには、日本語能力がどのような要素から構成されているか把握しておく必要がある。教室内テストであれば、教育目標や学習内容が基礎資料になる。また、具体的に導入した文型や文法規則、単語や文字、漢字などが出題の対象として考えられる。実際のコミュニケーション力を養成するためのカリキュラムであった場合、具体的なコミュニケーション力を明らかにして、その能力を引き出す手段を

講じなければならない。それによって、テスト問題の内容と方法は決まってくる。

初級レベルであれば､平仮名･片仮名などの文字認識力や発話力が問われることになる。次に、それらから構成される語彙（単語）の理解力が挙げられる。音声面においては、日本語の音素体系、強勢・抑揚などの認識ができるかどうか重要になる。そして、文を理解したり構成したりするために必要な文法力が問われる。これらは、基礎的言語知識を構成する「文字」「音韻（音声)」「文法」にあたるものだ。この他に、実際のコミュニケーションでは､「聴く：聴解力」「話す：会話力」「読む：読解力」「書く：作文力」などの主要言語運用能力が必要になる。「聴解力」と「読解力」は受容技能として理解能力に､「会話力」と「作文力」は産出技能として表現能力として区分されているものである。

言語能力については、次のような Canal & Swain（1980）の言語能力モデルがよく知られている。①文法的能力（Grammatical competence）：目標言語の語彙、文型、構文をはじめ、音韻などを理解して、文を構成できる能力。文法構造、単語の意味に限らず、形態や統語などが含まれる。②社会言語学的能力（Sociolinguistic competence）：目標言語が使用される社会的文脈や状況を理解して、適切に言語を運用できる能力。また、言語の社会的機能を活かして円滑なコミュニケーションを行える能力。③談話能力（Discourse competence）：意味があり、結束性のある発話や文章を構成したり、理解したりする能力。意味と言語形式を結びつけ、発話や文章を機能的に伝達できる能力。④方略的能力（Strategic competence）: 上記３つの能力不足を補うための能力で､コミュニケーションを維持したり、補正したりするための対処能力。言い換え、回避、転移、繰り返しなど。

このような言語知識や言語能力観は多くの研究者によって異なる角度から研究され、新たなモデルが創り出されているわけであるが、テスト項目を作成する段階で必要な具体的な運用力との関係については言及していない。言い換えれば、言語能力を構成要素に分類して、概念的な記述、理論的な枠組みにとどまっていて、言語が使用される場面や状況、また、文脈や人間関係など社会文化的な要因との関連づけは十分であるとは言い難い。一般的に言われている言語能力の枠組みは、教育の現場における指導のあり方やテスト作成への具体的な指針になるまでに至っていない。言語能力の構成概念が観察可能なものとして、わかりやすく定義されていないのである（Bachman, 1990）。

日本語教育における "Can-do" 能力記述文作成の試み

上述のような状況を踏まえ、1997 年に、筆者も委員の一人であった、日本語能力試験の分析委員会のメンバーを中心に、日本語に関する "Can-do statements" の項目作成を始めた（日本語教育学会編 1999, 三枝 2004）。初めての試みということもあり、１級受験者を対象として日本語運用力の記述化を試みた。具体的には、日本語を使ってできる言語活動を

整理・分析し、１級レベルでの実際の運用力を明らかにしようとした。それと共に、どのような言語活動が実際の日本語能力を測定する基準として適切かつ有効であるかを検討している。

調査対象は日本の大学で学んでいる者、及び入学前の学習者に設定し、質問項目を大学生の生活を中心に選定した。ただしアカデミックな分野にのみ項目を絞ったわけではなく、大学外での生活スキルも含むこととした。これは、項目の記述にあたり、その記述が具体的で実際に経験したことがあるタスクのほうが調査対象者の記述の理解に誤差が生じにくく、彼らの言語能力を調査項目に反映しやすいことが過去の統計データから判断されたためである。そこで、被験者が日本語能力試験１、２級受験者であり、その多くが日本国内の大学・専門学校に進学を希望していることに配慮しつつ、項目の記述にあたっては、①大学生が日常生活あるいは大学での勉学に必要とされる行動の記述、②回答者が実際に経験していると想定できる行動の記述、③具体的で現実的な場面の記述などの点に留意した。

項目の作成作業については、先行研究をもとに行動記述の項目の洗い出しを行った。既存の資料として、"TOEIC Can-do Guide"、"ALTE 'Can Do' statements" の行動記述の構成概念を参考にするとともに、国内での先行研究、大学の日本語教育のシラバス、日本語学校などの授業活動内容や使用しているテキストの内容から、行動記述を抽出した。国内の先行研究からは『外国人の日本語評価における Criteria 試案』『運用能力獲得のための基礎日本語能力』『リソース型生活日本語』、また、大学の予備教育、日本語学校で使用されている教科書からは『待遇表現』『An Introduction to Advanced Spoken Japanese』『大学で学ぶためのアカデミック・ジャパニーズ』、『リソース型生活日本語』、『新日本語の中級』を参考にした。この作業で、抽出された行動記述は 173 項目であった。

"Can-do statements" にかかわる調査では、外国語能力を測定して、そのデータを収集し、掌握することが一般的である。筆者らの調査では、日本語能力試験の結果を活用することができた。日本語能力試験の能力記述を試みるものであるから、当然、この試験結果の使用は不可欠なものである。したがって、実施団体からの理解と協力なしでは実現し得なかったことを述べておきたい。また、大規模調査となるため、テスト結果が受験した学生の想定レベルに対応しないのではないかという懸念はあったが、１級受験者に的を絞ることで必要なサンプル数を確保することができた。

調査方法は質問用紙による自己評定方式である。当然のことながら、この方式による利点と不利な点等については検討しておかなければならない。利点としては、①質問用紙を学習者の母語で記述できること、②生活場面を直接反映するものが作成できること、③基準に準拠した、特定の言語使用をベースにできること、などである。一方、不利な点として、①被験者は正直な評定を出さない恐れがあること、②自己評定のスケールとそれから得られるデータが、それとは独立した客観的なソース、つまり外在基準としてのテスト、筆者らの調査に関して言えば、日本語能力試験を構成する「文字・語彙」「聴解」「読解・

文法」と関連させる手続きが必要であることなどである。

"ALTE 'Can Do' statements"の開発にかかわる経緯

先に述べた ALTE では、欧州における外国語能力の共通認定証の促進を図るために、能力評価の基準作りを行った。この試みは、ALTE 枠組み開発プロジェクト（The ALTE Framework Project、以下「枠組みプロジェクト」と称す）と呼ばれているものである。開発の初期においては、以下のような作業が行われている。①学校からのレポートやアンケート調査によって、ALTE 言語テスト使用者を把握。②受験者のニーズを特定。③能力記述文を作成するために、Waystage や Threshold など国際的に認知された言語能力レベル記述を使用。④能力記述文の内容や表現を調整し、受験者能力との妥当性を検証。⑤妥当性と透明性を高めるために、教師および学習者による能力記述文の検討。⑥能力記述文の内容や表現を修正、などである。

枠組みプロジェクトの目的は、学習者が実際に外国で何ができるかを記述して、能力に関連する尺度を開発し検証することにあった。この"ALTE 'Can Do' statements" の中身を、項目作成者向け尺度、評価者向け尺度、テスト利用者向け尺度という３つの視点から分析すると、独自の構想による、テスト利用者向け尺度であるとしている（Alderson, 1991）。この点が、項目作成者向けに編集された日本語能力試験の「出題基準」とは異なっていることがわかる。これらの記述文は、試験プロセスの関係者間におけるコミュニケーション、特に、専門家や関係者以外の者による試験結果の解釈に役立っているとしている。具体的には、能力記述文は次のような機能を有していると考えられている。①外国語学習者への指導や試験に関わる者に対して、実用的な情報や資料を提供する機能。学習者自らが言語を使って何ができるかをチェックするリストとして利用でき、それによって学習者は、自分自身の該当する能力レベルを特定でき、その中身についても定義できる。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外国語の訓練および企業の人材採用に関わる人々にとって役立つ、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。④異なる言語間で、同じ状況下に存在する、講座や教材の目的を比較する手段としての機能。

枠組みプロジェクトは、また二元的な目的を持つものであった。一つには、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ようにすることであった。二つ目として、異なる外国語間での能力の枠組みを統合するための基盤整備に寄与することであった。能力記述文は運用能力を具体的かつ容易に理解できるよう提示されているため、研修や人事管理にかかわる人にとって有用であることが認められている。また、外国語教師に対する資格要件を特定する場合や、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に使用することができるとしている。

"ALTE 'Can Do' statements" は、異なる言語能力レベルでの言語運用力の典型を包括的に記述したものである。具体的には、言語能力と言語使用における脈絡（社会生活、旅行業、職場、学校）に関しての記述である。特徴としては、'Can Do'スケールが、ALTE の各レベルでの典型的な能力を記述したものを基にした言語能力スケールであること。そして、ALTE の試験問題スケールは、ALTE の各レベルにおける試験でのパーフォーマンスによって定義されたものであること。すなわち、実際の言語運用力を記述していることである。したがって、'Can Do'と試験問題スケールが関連づけられていることによって、例えばレベル４の試験に合格した者は、言語を使って何ができるかわかるようになっているのである。

言語運用力の主要レベルにおける枠組を確立することによって、その枠組内で外国語試験を開発することができ、またその内容については客観的な説明が可能となる。その結果、異言語間で中立的な言語熟達度レベルの解説文として、これらの記述文は、さまざまなレベルでのさまざまな言語試験の潜在的な基準枠を構成することになる。これらは、外国語試験に合格した受験者が、現実場面で実際の言語運用力を発揮することで、記述文と試験との間にどのような関連性、共通性、そして妥当性があるのかを実証する機会を与えてくれるものにもなっている。

日本語能力試験（JLPT）「出題基準」にみる言語能力

近年、60 万人以上が受験する日本語能力試験は、今では、大規模試験としては、かなりの影響力を持つ試験として存在している。

日本語能力試験は、非日本語母語話者の日本語能力を測定し、認定することを目的として開発、実施されている試験である。試験は、５つのレベル（N1～N5）に分かれ、受験者の能力に適した級を受験することができる。各級とも「文字・語彙」「聴解」「読解・文法」の３つのセクションから構成されており、レベル別認定基準に基づいて認定される。

本論では、５レベルに移行する前の４レベル（１級～４級）で実施されていた日本語能力試験の問題項目作成者の参考のために作られた『日本語能力試験出題基準〔改訂版〕』（2004）（以下「出題基準」）を基に、出題基準の内容を概観し、日本語能力試験で捉えられている日本語能力観について概観を試みる。

まず、「文字・語彙」についてみてみる。３級ならびに４級の「出題基準」は、日本国内ならびに海外諸国において広く使用されている数種の初級日本語教科書を基礎資料として、日本語教育に関する語彙調査を参考に作成されている。

一方、１級ならびに２級においては、３、４級のように日本語教科書から普遍的な共通

の文字・語彙を選定するのは困難であるとし、日本語における文字（漢字）使用の現状を踏まえ、それに日本語教育の立場からの修正を加えることによって「出題基準」を作成している。特に、「語彙」については、一般社会、日本語教育、学校教育（中学・高校）における語彙の使用を調査したこれまでの語彙調査の資料を基に、日本語教育の立場からの修正を加え、「出題基準」が作成されている。

次に、「文法」をみてみたい。３級ならびに４級で取りあげる「文法・文型事項」を選定する上で､基準が作成された当時の主要初級教科書８種に提出されているものを調査し、『日本語教育』45 号（田中・宮崎 1981）に掲載されている報告「日本語教科書における文法事項とその提示課」を基礎資料としている。この「出題基準」では、４級を初級の前期、３級を初級の後期とするのが妥当であると述べられているが、その根拠についての説明は多様な見解があって困難であるとしている。その結果、各教科書の課数を半分にして、前半と後半に分けたとしている。この中で、主要な文法事項「構文／文型」「活用」「機能語の用法」を挙げ、３級と４級での重要な学習項目として明示している。

これに対して、１級ならびに２級の「文法」については、３級から４級での基本的な助詞、助動詞の学習を前提に、より高度な「機能語」の類を習得することが大切であるとし、文法的な機能語の類が配列されている。教科書としては、『日本語表現文型　中級Ⅰ・Ⅱ』（1983）に言及しているのみで、具体的な教科書を参考に学習事項を選定したことには触れられていない。１、２級の文法・文型についての議論がこれまで十分になされてこなかったことや、ある学習事項が「文法」であるのか「語彙」として扱われるべきかなどの判断のむずかしさ、また、文法事項の守備範囲の広さも手伝って、１級と２級のレベル分けとともに、リスト化のむずかしさにも触れて、「出題基準」のまとめをしている。

次に、技能としての「読解」についてみてみる。「出題基準」の項目として、「テキストの長さ（長文・短文）」「語彙の逸脱率」「漢字含有率」「テキストタイプ」「タスク」「問題形式と選択肢」などが挙げられている。ここでは、テキストの持っている機能に言及し、読解テキストの種類について分類している。例えば、「働きかけ型／指示テキスト」「説明型／記述テキスト」「表現享受型／語りテキスト」「意見・感情等の表出型／論述テキスト」である。そして、出題対象となるテキストタイプの具体例が提示されている。また､出題のタスクの分類もなされていて､出題のねらいとするところも明示されている。級別の出題範囲は、３級と４級においては、形式、事実関係、意味解釈にとどめ、測定対象を広範囲にしないことが述べられている。一方、１級と２級においては、予測、推論、情報総合、テキスト機能も含め、全範囲を対象とすることとしている。ある意味では、１級と２級の読解タスクはすべて含まれることが前提になっていると言えよう。やはり技能としての「読解」については、読む対象としてのテキストとともに、読解能力にも言及し、そのメカニズムを解説している点は、言語知識とは異なった枠組みで対応していることがうかがえる。

最後に、「聴解」についてみてみる。聴解試験の特色として次のような点を挙げている。①学習者が教室を含めた日常生活で出会う課題と同様の課題を解決することが求められる、②情報の正確な聞き取りに留まらず、聞いて何かをすること、つまり課題を解決することが求められる、③（省略）･･･課題の解決にとって最も効果的な聞き方をすることが求められる。･･･（省略）、である。

試験課題一覧では、課題名（「情報完成」「規則・法則・傾向の適用／検証」「情報総合」「意図の汲み取り」「照合」「順序の再構成」「事物・人物の同定」）が提示され、どの級においてその課題が出題されるかを示している。各課題の問題作りにおいては、素材例が複数明示されていて、具体的な問題内容を示している。例えば、「情報総合」では、「ある人がなぜ休んでいるか、数人の友人の情報から推測する」「交通事故や火事などの原因を目撃者の証言などを聞いて推測する」などである。聴解テキストの種類を分類し、どのような聞き方が求められるかに言及している点は、読解力同様に、聴解のメカニズムが解説されていて、テキスト素材を考える上で参考になるものとなっている。

このように、日本語能力試験の「出題基準」は、言語学習の際に当然求められる学習内容、あるいは教授内容からの記述と、実際のテキストがどのような言語能力を機能させているかという視点からの記述でまとめられている。必ずしも、実際の言語運用力が現実の場面と関係づけられて記述されている訳ではなく、どの段階でどのような日本語能力が測定され、現実の生活場面で日本語がどのくらい運用できるかが明らかにされているものではない。

言語能力レベルについて議論するときには、日本語能力試験の１級から４級という級別による区分とともに、「初級」「中級」「上級」そして時に「超級」というレベル区分を示す用語を使用することが多い。日本語教育関係者であれば、だれしもがおおよその能力のイメージを描くことができるが、学習者を含めて、第三者にとっては、具体性を欠くものとなっていることは事実である。「能力基準」は、どちらかと言えば教授する立場からの言語知識を中心とした記述になっていて、学習者が学習成果として身につける能力を明示しているものになっていない。必ずしも項目作成者と受験者双方が共通の認識のもとに共有できる能力記述にはなっていないと言える。しかも、日本語能力試験が測ろうとしているものが、依然、実際に試験で測定されている言語運用力がどの程度級別に分類され、テスト項目に反映されているかは十分に明確にはなっていない。

日本語能力試験における "Can-do" 能力記述文作成の試み

新しい日本語能力試験は、課題遂行能力とそのためのコミュニケーション能力を測定す

る試験をめざしている。あわせて、これらの能力を支える基礎となる言語知識についても測定しているが、特徴的なことは、学習者の実際の言語行動を反映した試験をめざしていることである。

従来の試験（旧試験）は、上の級から１級－２級－３級－４級となっていたが、新試験では、Ｎ１－Ｎ２－Ｎ３－Ｎ４－Ｎ５となる。この背景には、３級に合格しても２級合格が困難であること、１級より一段上のレベルの能力測定の希望があることなどがあった。

Ｎ１：旧試験の１級と合格レベルはほぼ同じ。旧試験より若干高めの範囲まで
　　能力測定ができるように改定。

Ｎ２：旧試験の２級とほぼ同レベル。

Ｎ３：旧試験の２級と３級の間のレベル。

Ｎ４：旧試験の３級とほぼ同レベル。

Ｎ５：旧試験の４級とほぼ同レベル。

現在、2010 年度の改定以来、新日本語能力試験では、問題の改定と平行して、「日本語を使って何ができるか」を記述した Can-Do Statements の開発に取り組んでいる。試験問題の改定に伴い、級のイメージがより明確になり、何を測っている試験なのかがより具体化されることが期待される。

日本語能力試験 ‘Can Do’ statements（試行版）の開発と尺度化

これまでに、日本語能力試験の ‘Can Do’ statements（以下、JLPT-CDS）の開発に関しては、実施団体を中心に研究が進んでいる。その一つが、大隅他（2006）野口他（2006）大隅（2009）の研究である。具体的には、日本語能力試験の「モニター試験」と同日に同じ受験者に対して日本語能力試験 ‘Can Do’ statements（試行版）の調査を行っている。CDS 特性尺度を構成して、各受験者の日本語能力試験共通尺度における特性尺度値と CDS 特性尺度における CDS 尺度値を利用し、JLPT と JLPT-CDS との対応づけを試みている。また、同時に CDS（試行版）と CEFR-DIALANG とを対応づける試みも行っている（大隅・野口・熊谷・石毛・長沼・和田・伊東、2006）。JLPT-CDS（試行版）各項目と日本語能力試験の得点段階（各級認定段階）との対応づけを行った結果の考察もあわせて行っている。

なお CEFR-DIALANG との比較対照については、現時点では順序性の比較に留まる。JLPT-CDS と CEFR-DIALANG 間でほぼ同じ内容を表していると見られる statements が〈聞く〉20 項目中 6 項目、〈読む〉同 9 項目、〈書く〉同 10 項目あり，〈聞く〉〈書く〉については両尺度における順序性が一致したが、〈読む〉に関しては一致しておらず、今後の研究課題となっている。

残された課題

現在、日本語能力試験のウェブ上では、『日本語能力試験 Can-do 自己評価レポート』《中間報告》を公開している。さまざまな学習環境で日本語を学ぶ受験者に対して、日本語を使ってどのようなことができると考えているかについてのアンケート調査のまとめである。

2010 年に、第 1 回試験（2010 年 7 月実施）、第 2 回試験（2010 年 12 月）実施）の受験者を対象に調査を実施している。2010 年 9 月から 2010 年 12 月までの 4 ヶ月間に、日本国内及び海外 5 ヶ国で計 27,165 人の受験者を対象に行ったものである。

このアンケート項目は、「聞く」「話す」「読む」「書く」の 4 つのセクションから構成されている。回答者は、例えば、「駅やデパートでのアナウンスを聞いて、だいたい理解できる」のような日本語を使う行動が記述された文を一つ一つ読んで、まず、1）その行動を実際に日本語で経験したことがあるかどうかを「はい」「いいえ」で答える。その後、2）その行動が日本語でできるかどうかを「4：できる」「3：難しいがなんとかできる」「2：あまりできない」「1：できない」の 4 段階で自己評価する。経験がない場合にも、できそうかどうか想像して答えるよう指示されたものである。

この《中間報告》では、調査の対象となった各レベル合格者による自己評価（4：できる、3：難しいが何とかできる、2：あまりできない、1：できない）の平均値を示している。Can-do の項目は、N1～N3 の表についてはN1 合格者の、N4～N5 の表についてはN4 合格者の判断を基準に、難しいと感じている順に並べ替えている。

ただし、「できる」「できない」の程度の判断は、人によって、また経験したことがあるかどうかによって異なってくるものである。自己評価に基づくものであるところから、調査の結果は実際の合格者の言語行動の実態を正確に表したものではない点で課題が残る。現段階では、合格者が自分の日本語能力についてどう評価しているかを報告しているに留まっている。試験実施団体は、受験者や周りの人々が、「各レベルの合格者が日本語を使ってどんなことができそうか」というイメージ作りの参考情報として活用してほしいと案内している。

最後に、残された課題を以下にまとめておきたい。

(1) 日本語能力試験の各レベルにおける課題と Can-Do Statements との整合性の検討
(2) 日本語能力試験のウェブ上で公開されている、『日本語能力試験 Can-do 自己評価レポート』のさらなる分析
(3) 日本語能力試験の 5 レベルと CEFR の 6 レベルとの比較及び検討

参考文献

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.

Bachman, L. F. (1990) *Fundamental Considerations in Language Testing*. Oxford University Press.（池田央・大友賢二監修 (1997)『言語テスト法の基礎』C.S.L. 学習評価研究所）

Bachman, L. F. & Palmer, A. S. (1996) *Language Testing in Practice : Designing and Developing Useful Language Tests*. Oxford University Press.（大友賢二他監訳（2000）『<実践>言語テスト作成法』大修館書店）

Canale, M. & Swain, M. (1980) ‘Theoretical bases of communicative approaches to second language teaching and testing' in *Applied Linguistics I/I*.

Carroll, J. B. (1961) 'Fundamental considerations in testing for English language proficiency of foreign students' in Center for Applied Linguistics: *Testing the English Proficiency of Foreign Students*. Center for Applied Linguistics.

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

Messick, S. (1989) ‘Validity’in R. L. Linn (ed.): *Educational measurement*. (3rd edn.) Macmillan.（池田央訳（1992）「第 2 章妥当性」『教育測定学―原著　第 3 版―上巻』みくに出版）

Oller, J. (1979) *Language Tests at School*. Longman.（堀口俊一訳者代表（1994)『言語テスト』秀文インターナショナル)

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

伊東祐郎（2005）『日本語教育評価法』<NAFL 日本語教師養成プログラム教本> アルク

伊東祐郎（2005）「これまでの評価／これからの評価」『AJALT』第 28 号、国際日本語普及協会

大隅敦子（2009)「日本語能力試験改定の中間報告」『言語教育評価研究』第 1 号, pp.73-77.

大隅敦子・野口裕之・熊谷龍一・石毛順子・長沼君主・和田晃子・伊東祐郎（2006 年 8 月 26 日）．「日本語能力試験 Can-do-statements（試行版）と CEFR-Dialang との対応付けの試み」第 5 回国際日本語 OPI シンポジウム（ベルリン：ベルリン日独センター）.

国際交流基金・日本国際教育支援協会（2009）『新しい「日本語能力試験」ガイドブック』国際交流基金・日本国際教育支援協会.

国際交流基金・日本国際教育協会（2004）『日本語能力試験出題基準〔改訂版〕』凡人社.

三枝令子他 (2004)『日本語 Can-do-statements 尺度の開発』（平成 13 年度～平成 15 年度

科学研究費補助金・基盤研究 B1　研究成果報告書）
田中和美・宮崎寿子（1981）「日本語教科書における文法事項とその提示課（資料）」『日本語教育』45 号、日本語教育学会．編
寺村秀夫・佐久間まゆみ他（編著）（1983）『日本語表現文型　中級Ⅰ・Ⅱ』筑波大学日本語教育研究室、イセブ出版.
日本語教育学会編 (1999)『Can-do-statements 調査報告』国際交流基金
日本語教育振興協会（2001）『運用能力獲得のための基礎日本語能力』
野口裕之・大隅敦子・熊谷龍一・石毛順子・長沼君主（2006 年 8 月 24 日～26 日）「日本語能力試験 can-do-statements（試行版）の IRT 尺度化と日本語能力試験の得点段階との対応付けの試み」第 5 回国際日本語 OPI シンポジウム（ベルリン：ベルリン日独センター）.
野口裕之・熊谷龍一・大隅敦子（2007）「日本語能力試験における級間共通尺度構成の試み」『日本語教育』135 号、70 - 79.

参考・引用ウェブサイト

国際交流基金「新しい日本語能力試験」http://www.jlpt.jp/ （2011 年 11 月 26 日現在）

3.2. The role and function of Can-Do Statements in Japanese language teaching

Sukero Ito

Abstract

The Japanese language teachers have been interested in teaching what language learners have to know about the language. The most popular areas could be in terms of knowledge about linguistic facts about the syntax and semantics of Japanese language. Teaching what learners should know about the language raises the questions of what we actually mean by language learning outcomes.

The Can-Do Statements in CEFR, National Standards (USA), and Canadian Language Benchmarks have brought the Japanese education system a new way of thought of language proficiency. They assist communication among teachers in the curriculum development process, its implementation and learning outcomes. Reflecting one's language program or describing levels of language learning goals of the program, Can-Do Statements can play a role of:

a) a useful tool for those involved in teaching and testing language learners as a checklist of what language users can do and thus define the stage they are at;
b) a basis for developing diagnostic test tasks, activity-based curricula and teaching materials;
c) a means of carrying out an activity-based linguistic tasks;
d) a means of comparing the objectives of courses and materials in different languages but existing in the same context.

They will be of use to both experienced teachers and less experienced teachers in the program, as they provide easily understandable descriptions of performance, which can be used in specifying requirements to language learning outcomes.

３．２．日本語教育カリキュラムにおける Can-do Statements の役割と機能

伊東祐郎

日本語教育におけるこれまでの言語能力観

日本語教育における教授内容は、言語学習の際に当然求められる学習内容、あるいは教授内容からの記述と、実際のテキストがどのような言語能力を機能させているかという視点からの記述でまとめられている。一般的なシラバスは、構造シラバスと呼ばれ、文型、文法項目、語彙を中心にしたシラバスである。オーディオ・リンガルメソッド全盛期の頃に中核となった構造言語学を基盤としてもので、言語形式を重視している。したがって、初級教科書の構成は、基礎的な文型と文法項目、語彙から配列され積み上げ式という特徴をもつ。中級以降も既習の文型や語彙に基づいていて、コミュニケーションの場面の取り上げ方は二次的である。結果的には日本語の構造をもれなく教育することに主眼がおかれているといってよい。上記の技能別到達目標では、段階的にコミュニケーションの対象となるものが明記されているが、「コミュニケーションのための日本語」「アカデミック・ジャパニーズ」と呼ばれるようになったのはごく最近のことである。

言語能力レベルについて議論するときには、「初級」「中級」「上級」そして時に「超級」というレベル区分を示す用語を使用することが多い。予備教育課程におけるレベルもこのような区分で示されている。日本語教育関係者であれば、だれしもがおおよその能力のイメージを描くことができるが、その場合のイメージとは、文型であったり、語彙や文構造の難易度によるイメージであることが多い。たとえ、運用力をイメージできたとしても大ざっぱなものであろう。学習者を含めて、第三者にとっては、具体性を欠くものとなっていることは事実であろう。「能力レベル」は、どちらかと言えば教授する立場からの構造を中心とした文型や文法の記述が中心となっているといえよう。また、語彙数、漢字数、学習時間数という、言語能力とは直接関係しない「学習量」が能力区分の一項目に加わっていることも特徴的である。日本語能力試験の認定基準を例にとってみても、初級前半のレベルを示す４級は、漢字 100 字程度、語彙 800 程度を習得し、日本語学習時間を 150 時間程度。１級は、高度な文法・漢字 2,000 字程度、語彙 10,000 程度、学習時間を 900 時間程度と明記されていることからもうかがえる。このような数量は、ある意味で、学習者の努力目標として捉えるならば、わかりやすい基準となろう。しかしながら、学習者が学習成果として身につける能力を明示するものには結びつくものではない。このような言語能力観は、教授者と学習者双

方が共通の認識のもとに可視化され共有できる能力記述にはなっていないのである。

最近の外国語教育の動向（スタンダード開発の動き）

欧州では、各国の EU 統合という動きの中で、言語と文化背景の異なる市民の国境を越えた移動にともなう相互理解や協働作業などを推進するために、言語教育にかかわる政策面での整備・統合を実現する必要があった。あわせて、欧州社会の複数言語主義（plurilingualism）に基づく現代語教育システムの統一的実現化を目指してきた。具体的には学習者の学習ニーズ、学習動機、学習内容、学習目標などを包括的に捉え、それらを明示し、言語教育に従事する教師や学習者をはじめ、教育行政関係者、教科書出版社、試験問題作成者などに対して一般性を持たせようとしたのである。Common European Framework of Reference for Languages: Learning, teaching, assessment （以下「CEFR」）はこのような背景から生まれたもので、ここで述べられているコミュニケーション能力は、現在、外国語学習と教授法、そして評価に対して様々な影響を与えている。

米国では、80 年代から 90 年代にかけて初等・中等教育における教育改革が積極的に推し進められた。この中で、American Council of Teaching of Foreign Languages（以下「ACTFL」）が中心になって外国語の全米標準（National Standards）を完成させた。内容は、学校における外国語教育の目標を明確化し、教育内容やアプローチのあり方を示すものとなっている。学習者や学習条件、学習ニーズの多様性に配慮して、柔軟な教育の実現を目指して作成されたものである。この標準の中で、外国語学習者の達成すべき目標を掲げ、外国語によるコミュニケーション力の総体を具体的な形で提示し、小学校から中学校、高校、そして大学につながる外国語教育の一貫性を目指した点は注目に値する。外国語学習の意義や教育指針の認識、また教師訓練のあり方などが具体的な形で提示されていて、米国の外国語教育を全体で捉え直したところが、重要な点であると言えよう。

一方、多くの移民を労働力として受け入れているカナダでは、移民の言語運用力とそれを訓練する言語教育のあり方が重要な課題になっていて、移民に対してどのような言語政策をとり、言語教育を推進していくかが議論されてきた。そこで、政界、財界、教育界が一体となって、言語教育のフレームワーク作りを行ってきた。このような背景の中からできあがったのが、Center for Canadian Language Benchmarks が作成した Canadian Language Benchmarks（以下「CLB」）である。CLB は、労働者としての移民に必要な言語運用力を段階的に示しているので、英語教師のカリキュラムの策定や評価方法の指針となっている。また、雇用者が移民を採用しようとする際には、標準で示されている言語能力基準を参考にして、移民の言語能力の判定や評価を行うよう奨励されているのである。

スタンダードにおける "Can-Do Statements"の役割と機能

欧米各国において実現化されているスタンダードを概観してみると、どのスタンダードにも言語能力を段階的に記述した"Can-Do Statements"（「言語能力記述文」）がある。学習者の言語能力については、文法能力・語彙能力・発音能力などを細分化して捉えるのではなく、コミュニケーションを機能させる要素をグローバルな視点から捉えていると同時に、コミュニケーション能力を成立させている基本的構成能力を明示している。

言語能力記述文は、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ところに特徴がある。最近のスタンダードでは、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。

具体的には、次のような機能や役割を有していると考えられる。

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。

言語能力記述文は、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。Can-Do Statements は、学習者にとっては達成すべき目標が明確になっている。教育者にとっては、言語行動と言語学習に一貫性を持たせられる教育目標の設定が可能で、教育内容と整合させた評価が実現できる。また、テスト開発者にとっては、テスト課題の内容と形式を特定化しやすく、コミュニケーション活動に対する評価の結果の記述が容易になる。このように言語能力記

述文は、これまでの言語運用力を数量や数値ではなく、具体的な運用力の実情と照らし合わせて学習、教授、そして評価することを可能にした点に注目することが大切である。

スタンダードにおける「言語能力記述文」の役割と機能

最近、外国語教育の分野では「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」という言葉を頻繁に耳にする。例えば、米国においては、"ESL Standards for Pre-K-12 Students"、"Standards for Foreign Language Learning"、"The ACTFL（American Council on the Teaching of Foreign Languages：アメリカ外国語教育協議会） Performance Guidelines for K-12 Learners"があり、カナダにおいては、"Canadian Language Benchmarks"を開発している。また、欧州においては、欧州評議会が作成した"Common European Framework of Reference for Languages: Learning, teaching, assessment"（以下「CEFR」）がある。しかしながら日本における外国語教育では、先進事例を参考にするものの、独自の理念や教育目標に基づいて開発された目安や基準はまだ一般的ではない。

欧米各国のスタンダードを概観してみると、言語能力を段階的に記述した Can-Do Statements（「言語能力記述文」）がある。Can-Do Statements は、学習者の言語能力を文法能力・語彙能力・発音能力などを細分化して捉えたものではなく、コミュニケーション活動にかかわる能力を言語化したものである。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ところに特徴がある。最近のスタンダードでは、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。日本語教育において、Can-Do Statements はどのような役割と機能を担っているのか考察を試みたい。

"Can-do Statements" が日本語教育に与えた影響

日本語教育においてもっとも Can-do Statements の影響を受けたのは、欧州内で日本語教育を担っている日本語教師であった。1970 年代から始まった欧州評議会の言語教育への施策を背景として、教育内容を国際的に比較可能で、しかも明瞭な能力評価基準が求められることからカリキュラムの在り方などを検討する状況に直面することになった（国際交流基金 2005）。欧州の複言語主義の促進のために誕生した CEFR は、日本語教育内容の文脈化を考える機会となっている。

山川（2009）は、CEFR が作成された背景にある社会事情や複言語・複文化主義を理解しつつも、日本語が非ヨーロッパ言語である点に着目し、学習者である「個人」を「社会的に行動

する者」と認識することによって「言語でできる」ための環境設定を具体的に明示する必要性を説いている。「日本語」という枠を超え、広く「言語」という観点から Can-do Statements を捉えた場合は、抽象的な記述をどう解釈するかという課題は残るが、言語教育のあり方を歴史的にまた地理的に考察する上でインパクトを与えたことは事実である。

福島（2009）は、CEFR に現れる A1 から C2 までの 6 レベルを機能・行動から考察し、CEFR の特徴は「社会参加」「自己表現」と言う行動を「言語・談話能力」が支えとなっていると分析した。欧州における日本語教育を考える上での文脈の重要性を論じている。また、櫻井・近藤（2009）は、欧州内で教鞭をとっている教師には、CEFR を理解するだけでなく、実践に取り入れていく文脈化の能力が求められていると主張している。CEFR への興味関心の高まりから、知識と活用方法へのニーズに対する教師研修のあり方を考察している。特に Can-do Statements を意識した教案作成や学習評価などを取り上げ、研修の課題をまとめている点は興味深い。

一方、スルツベルゲル三木佐和子（2007）は、スイスの高校での日本語教育を取り上げ、学生達が ELP（European Language Portfolio : ヨーロッパ言語ポートフォリオ）の自己評価の際に、アルファベット文字ではない日本語の「読む」「書く」をどのように評価すべきかを問題点として取り上げている。A1 から C2 に至るまでの漢字語彙をどのように取り扱うかが非ヨーロッパ言語に課せられた課題となっている。

国際交流基金による日本語教育スタンダード

国際交流基金（以下、「基金」）は、1972 年の設立以来、海外における日本語教育の普及に取り組んでいる（嘉数 2009）。昨今日本語教育が多様化する中で、基金は、2005 年から「JF 日本語教育スタンダード」（以下、「JF スタンダード」）の開発に取り組み、2010 年 3 月に、『JF 日本語教育スタンダード 2010』（2009）として発表した。

実はこのスタンダード開発への発端となったのが、CEFR から影響をうけた欧州内の日本語教育関係者からの動きであった。大規模試験である日本語能力試験の能力レベルが、CEFR の A1 から C2 のどのレベルに相当するのかという問い。また、欧州内で使用されている教科書や教材、また開講されている日本語クラスが、A1 から C2 のどれに当たるかという説明を求められることによって、必然的に CEFR との照合や比較に迫られた。このような事情もあり、JF スタンダードは、CEFR を基盤として開発されることなった。

JF スタンダードの理念は「相互理解のための日本語」であり、それは「課題遂行能力」と「異文化理解能力」によって実現されるという考え方を開発担当者が示している。2010 年 3 月に発表された第 1 版で、JF スタンダードが教育現場を支援するために以下の 3 つを提供した。

①能力記述文データ検索ウェブサイト（みんなの Can-do サイト）:「～できる」のかたちで言語能力を記述している、能力記述文（Can-do）を利用した日本語教育実践を支援するウェブサイト。
②ポートフォリオサンプル：ポートフォリオは、異文化理解能力の育成や自律学習支援を目指した評価ツール。JF スタンダードの取り組み事例で実際に使用されたサンプルを提供している。
③事例集：国内外の国際交流基金拠点において、Can-do やポートフォリオを実際に活用した取り組み事例を、日本語教育現場で参照できるよう、わかりやすいかたちで提供したもの。

そして、JF スタンダードは教育現場の様々な実情に応じて利用することができ、教育実践を SEE（現状分析）→ PLAN（目標設定）→ DO（実施）→ SEE（ふり返り）のサイクルで捉えることを提案している。CEFR を基盤としているだけあって、構造もかなり似たものになっている。

概要説明では、基金は、世界中で日本語を通じて相互理解をするためには、日本語を使って何がどのようにできるかという能力（課題遂行の能力）と、さまざまな文化に触れることで視野を広げ他者の文化を理解し尊重する能力（異文化理解の能力）が必要であると考えている。この「相互理解のための日本語」を実現するために、日本語の教え方、学び方、学習成果の評価の仕方を考えるためのツールである「JF スタンダード」の開発を行ってきた。『JF 日本語教育スタンダード 2010』では、CEFR の言語熟達度のレベルにもとづき、日本語の熟達度を「～できる（「Can-do」）」という形式の文で示し、「みんなの「Can-do」サイト」で提供していることが特徴的である。多種多様な日本語教育の現場がいわば同じものさしを使うことで、世界中のどこで日本語を教授または学習していても、どのレベルにおける教育内容であるのかを知ることができるようになることをその利点として挙げている。

日本語予備教育課程におけるスタンダード

筆者が勤務する東京外国語大学・留学生日本語教育センター（以下「センター」）は、日本の大学学部進学を目指す留学生に対して 1 年の集中日本語教育（予備教育課程）のための「日本語スタンダーズ」の構築に取り組んでいる。予備教育課程の目標は、留学生が日本の大学での勉学に必要となる日本語能力を十分に習得すること、また、それぞれの専攻に応じて、人文社会あるいは自然科学の基礎的な学力をつけることになっている。日常生活で必要な日本語能力はもちろんのこと、学部進学後に必須となるアカデミックスキル、いわば「アカデミック・ジャパニーズ」を獲得することを主な教育の目標としてい

る。具体的には教科書や専門書が読める読解力。レポートや論文が書ける文章表現力。また、学生との日常会話のみならず、授業中の講義や学生の発表が理解できる聴解力。そして、授業中に自らが質疑応答でき、研究の成果などを発表できる口頭表現力を伸ばすことに力を注いでいる。

日本語教育は、長年の経験と実績を基盤としている。経験は年月を経るにしたがって教員各自の経験知となり次第に暗黙知となっていく。それが慣行化されることによって、以前は他の教員と共有されていた教育の目標や内容が必ずしも引き継がれていくとは限らなくなる。プロジェクトでは「アカデミック・ジャパニーズ」の概念と予備教育課程の教育目標を再確認した上で、教員各自の教育実践を振り返ることから作業は始まった。実践内容を共有し、教育のゴールと照らし合わせながら、スタンダーズの核となる言語能力記述文＝Can-Do Statements の記述を行った。教員各自の実践を言葉によって明らかにし、相互に共有して自信と確信をもって教育実践していこうとするところに、日本語スタンダーズ開発の意義が見いだされる。

東京外大の取り組みは、国際交流基金と異なり、開発のよりどころを CEFR の 6 段階レベルやそれに対応した Can-do Statements においていない。あくまでも、アカデミック場面における予備教育課程の最終の技能別到達目標を出発点としている点である。

残された課題

Can-Do Statements を作成する作業は、言語教師にとっておそらく馴染みのないものである。構造シラバスに慣れ親しんでいる教師にとって、日々の実践、あるいは教育目標を言語的側面からではなく、言語行動として記述していくことは、これまでの言語能力観や教育観の捉え直しを余儀なくさせるものである。実践の内省と捉え直し、そして新たな枠組み作りは、新しい知を構築していく創造的作業となる。言語的知識の伝授の結果、学生が実際の運用力として身につけるべき能力やスキルを具現化すること、言葉を使って、しかも言語が使用される場面や状況、また社会文化的な要因と関連づけて記述することは、その表現方法においてかなり複雑なものになる。言語習得の過程やパフォーマンスにかかわる概念を把握しておく必要もあるからである。

Can-Do Statements を暫定的にも完成させる過程を通して、これまでの教育内容の妥当性や適切性も浮き彫りになった。アカデミックスキルを身につけさせるという目標と照らし合わせて教育内容の妥当性を点検してみると、使用目的や活用意図が不明な教材や、何のために活用されていたのか不可解なものなどが明らかにされた。また、評価にかかわるテスト問題や試験の実施方法においても、適切なものとそうでないものなどが識別され、教材や試験の見直しをする機会にもなったのである。この意味において、スタンダーズには、

カリキュラムの内容と教育実践、その後の教育成果をスタンダーズに照らし合わせることによって、点検・評価の機能を有していることを確認できたのである。他機関で作られたスタンダードや Can-Do Statements は、やはりその機関や組織の教育理念を反映したものであり、そのための教育実践の見直しや点検にとっては有効であっても、第三者にとっては必ずしも有効になるとは限らない。むしろ、当事者が自らの教育に対する実践知を具現化するという作業にかかわることによって、初めて Can-Do Statements 作成の意義が生まれるものと思われる。その作業プロセスに自らがかかわることが、そしてその主体者となることが最も意味のあると思われる。

最後に、残された課題を以下にまとめておきたい。

(1) 日本語能力と日本語カリキュラムの内容の整合性にかかわる諸問題の共有

(2) 日本語能力と日本語学習及び日本語教授の一貫性、透明性の検討

(3) 日本語教育カリキュラムの可視化、すなわち Can-Do Statements 化の推進

参考文献

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

J.A. van Ek and J.L.M. Trim (1998) *Waystage 1990: Council of Europe Conseil de l'Europe,* Cambridge: Cambridge University Press.

J.A. van Ek and J.L.M. Trim (1998) *Threshold 1990: Council of Europe Conseil de l'Europe,* Cambridge: Cambridge University Press.

J.A. van Ek and J.L.M. Trim (2001) *Vantage: Council of Europe Conseil de l'Europe,* Cambridge: Cambridge University Press.

伊東祐郎（2005）「これまでの評価／これからの評価」『AJALT』第 28 号、国際日本語普及協会

伊東祐郎（2006）「日本語能力"Can-do"記述文作成の試みーテスト得点の妥当化をめざしてー」『高見澤孟先生古希記念論文集』

伊東祐郎 (2006) 「評価の観点から見た日本語教育スタンダード」『日本語学』明治書院、25, pp.18-25.

伊東祐郎 (2009)「日本語教育スタンダードをめぐる議論を終えて」萬美保・村上史展編『グローバル化社会の日本語教育と日本文化』ひつじ書房、pp. 64-71.

嘉数勝美 (2006) 「ヨーロッパの統合と日本語教育-CEF (「ヨーロッパ言語教育共通参照枠」)をめぐって―」)」『日本語学』, 25 (13) , pp. 46-58.

嘉数勝美 (2009) 「国際標準としての「日本語教育スタンダード」の構築-「ヨーロッパ言語共通参照枠組み」(CEFR)の応用と課題」萬美保・村上史展編『グローバル化社会の日本語教育と日本文化』ひつじ書房、pp. 4-27.

金田智子（2010）「日本語教育における CEFR 応用の試み」『英語教育』2010 年 10 月増刊号, pp. 64-67.

金谷憲編著（2003）『英語教育評価論』桐原書店

国際交流基金編（2003）『日本語能力試験企画小委員会口頭能力試験調査部会報告－口頭能力試験科目の創設に向けて－』日本語能力試験企画小委員会口頭能力試験調査部会

国際交流基金編（2009）『JF 日本語教育スタンダード 2010』国際交流基金.

国際交流基金・日本国際教育協会（2004）『日本語能力試験　出題基準[改訂版]』凡人社

三枝令子他 (2004)『日本語 Can-do-statements 尺度の開発』（平成 13 年度～平成 15 年度科学研究費補助金・基盤研究 B1　研究成果報告書）

櫻井直子・近藤裕美子（2009）「CEFR 文脈化のための実践例を取り入れたワークショップ型教師研修」『ヨーロッパ日本語教育』14 号、ヨーロッパ日本語教師会、pp.215-222.

塩澤真季・石司えり・島田徳子（2010）「言語能力の熟達度を表す Can-do 記述の分析－JF Can-do 作成のためのガイドライン策定に向けて－」『国際交流基金日本語教育紀要』、6、pp. 23-39.

靜哲人他 (2002)『外国語教育リサーチとテスティングの基礎概念』関西大学出版部

スルツベルゲル三木佐和子（2007）「CEFR/ELP 能力査定基準の日本語スキル査定への応用を探る－Ａ１の漢字について－」『ヨーロッパ日本語教育』12 号、ヨーロッパ日本語教師会、pp.183-188.

東京外国語大学（2006）『JLC シンポジウム報告書－日本語スタンダードを考える－』東京外国語大学留学生日本語教育センター

東京外国語大学（2011）『世界的基準となる日本語スタンダーズの構築』東京外国語大学留学生日本語教育センター

中村洋一 (2002)『テストで言語能力は測れるか』桐原書店

日本語教育学会編 (1999)『Can-do-statements 調査報告』国際交流基金

日本語教育振興協会（2001）『運用能力獲得のための基礎日本語能力』

福島青史（2009）「CEFR 能力記述文のレベル別特徴とキーワード」『ヨーロッパ日本語教育』14 号、ヨーロッパ日本語教師会、pp.132-139.

山川智子（2009）「日本語教育の文脈化を考える－市民社会における ”plurilingualism/pluriculturalism” 概念の理解と CEFR－」『ヨーロッパ日本語教育』14 号、pp.223-230.

3.3. The relationship of proficiency levels to Can-Do statements in English language education

Tomoko Fujita

Abstract

These days, it is becoming very common to set up Can-Do statements (CDS) based on Common European Framework of Reference for Languages (CEFR) for commercial based high-stake tests such as STEP, TOEFL, and TOEIC. These CDSs for score interpretation are attempts to relate test scores to a wide range of real-life language tasks. However, the validation processes on those CDSs will take a very long time for deliberate processes. According to Weir (2005), who wrote about limitations of the CEFR, tests are required to reflect upon how constructs differ from level to level in terms of the contextual conditions reflecting task performance. He also suggested researchers not only should investigate what learners Can-Do at each level but also under what performance conditions the activity is often performed and in what quality. There are huge demands for conducting validation studies, especially on the context, theory-based, and scoring validity of CDS and the framework. In order to create more accurate and valid CDS, providing opportunities for teachers and learners to engage and contribute their materials and to sample performances from their own contexts for collective moderation is very important.

３．３．英語教育における習熟度レベルと Can-Do statements

藤田智子

Can-Do statements (CDS) は言語学習を進めていくうえで、到達目標であり、自己評価や教師評価の基準でもあり、また自律的学習を促進するために、その有用性は非常に高い。ここでは CDS を尺度として、１．テストスコアの解釈規準を目的とした研究と、２．自己評価を目的とした研究に分けて解説する。これらは英語教育の現場において CDS を学習のサイクルの一翼として組み込んで利用する上で、両方とも欠かせない目的である。本論では、英語教育における CDS の利用について、まず先行研究をまとめたあと今後の課題について言及する。

テストスコアと Can-Do statements

テストスコアの解釈規準

ケンブリッジ大学の英語テスト作成研究機関である Cambridge ESOL（English for Speakers of Other Languages）のテストは、International English Language Testing System (IELTS)をはじめとして Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001) の６段階レベルに対応するように作られている。Taylor (2003)によると、Cambridge ESOL の英語能力テストは CEFR と表裏一体のように合体したものだと言う。このように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が発端となり、今では、さまざまなテスト作成機関がその解釈規準として独自のＣＤＳを公表している。これは、その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している。例えば、TOEIC Can-Do Guide、TOEFL iBT as competency descriptors、などもこの動きに追随し、国内で代表的なのは、英検 Can-Do リストがそれである。

この動きに対応した国内での研究としては、根岸（2005, 2006）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 つのレベルに分けられている。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。

日本の大学や高校などの教育機関で、CDS を作成し、これを基盤とした授業カリキュラムを開発するプロジェクトを数多く手掛けている Naganuma（2006, 2008, 2010）は、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。また、CDS の開発によって、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。

しかし、テストスコアをＣＤＳと対応させる動きに関しては、Weir (2005)が妥当性の観点から警鐘を鳴らしている。彼によると、ＣＥＦＲの各レベルに適応するようにテストを作成する場合、現時点でのＣＤＳでは、能力記述文の内容的パラメタの難易度を上げたり下げたりする規準の構成概念妥当性が不十分であると指摘している。また、テスト理論の妥当性から見ても､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 でなければ対応できないと述べている。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否定したのではなく、将来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）か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も、詳細に至るまで追及する必要があると指摘している。

Can-Do 自己評価チェックリスト

CEFR 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「自己評価チェックリスト」の代表的なものに European Language Portfolio (ELP)がある。これは、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。また、学習の記録を残すことができるように、ポートフォリオのスタイルをとっている。

この 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が、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、能力記述文を同一尺度化して項目バンクに入れて利用する研究を実施した。その後、Lenz & Schneider (2004)は、作成した英語の能力記述文の項目バンク (Bank of Descriptors) をウェブ上で公開している。

筒井、近藤、＆中野（2007）は、前述の North & Schneider (1998)の研究で開発された能力記述文をもとにして、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比較した。CEFR の６レベルと同じ、６つの習熟度別レベルに分かれたクラスで、英語コミュニケーション能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。また、その担当教員には同じチェックリストで学生を教師評価してもらい、これら学生自己評価と教師評価を比較した。この結果を IRT を利用して分析すると、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の能力値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 段階になっていた。

Can-Do statements の今後の課題

妥当性の確認

多くの関係者が Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001)に大きな関心を寄せ、その言語教育プログラムに CEFR の考え方や方式を導入しようとする動きを示すなかで、Weir (2005)は、もっと慎重に CEFR の妥当性検証をし、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主張している。 Weir は、現時点での CEFR の能力記述の内容では、各レベルに適応するようにテストを作成しようとしても、そのテストの難易度を上げたり下げたりする規準となる構成概念妥当性が不十分であると指摘している。言わば、受験者が実際のテスト問題のタスクにどのように答えるかということが、レベルごとに異なるようにテストの内容を設定するべきだということである。テスティングを目的とするフレームワークの構成概念妥当性を提唱し、それぞれ異なったレベルごとに、テスト問題に関する理論ベースのタスクの妥当性、そして、その問題に対する受験者の回答への得点の妥当性まで考慮しなければいけないと主張しているわけである。

また彼は、テスト理論の妥当性の観点から、テストを作成する度に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 を作成しなければ対応できないと述べている。そして、最後に将来の方向性として、テスト開発者たちは、学習者たちが CEFR の 6

レベルのそれぞれにおいて、何がどのようにできる（Can-Do）のか知るだけでは不充分であると主張している。なぜなら、テスト開発者は、受験者たちがどのような状況の下で、普段これらのアクティビティーを行って、そのテストの規準をどのくらい満たすのか予想できなければいけないと強調している。

Green (2010)によると、CDS は言語の指導や学習に関する非常に多様な目的のために用いられる傾向にある。例えば、ニーズ調査、到達目標の設定、学習過程の策定、習熟度の評価や評価結果の比較などのようなさまざまな目的が報告されている。彼は、このように CDS が予想以上に広く応用されるようになったことで、その利用価値が危ぶまれるような問題に直面するのではないかという懸念について論じている。彼の結論としては、CDS を成功させるには、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り、より妥当な CDS のレベル設定のために意見交換したり、積極的に協力しあうことが重要だと述べている。

CDS に基づいたテスト

GTEC for STUDENTS という日本人学習者を対象にしたテストの英語能力習熟度レベルのスケールを開発した Negishi（2005, 2006）は、テスト得点をもとにして、その受験者が実際の生活で、どのくらい英語が使えるかを推定する方法として 2 通りあると述べている。一つは、習熟度別ガイドラインで、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。またもう一つは、CDS としてテスト以外の現実の場面において、受験者が英語でできることに対するテストの得点を導き出したものである。彼は、GTEC for STUDENTS を対象にして、リーディング、リスニング、ライティングの技能について、習熟度別ガイドラインと CDS の両方を開発する過程を報告している。

Alderson & Huhta (2005）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 DIALANG の開発、予備テスト、規準設定について述べている。このテストは、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。初めに語彙テスト、自己評価からはじめ、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。どの言語でテストを受けるか、また、語彙テスト、自己評価を受けるかどうかなども、受験者が決めることができる。DIALANG は言語能力を測定することだけを目的にしたテストというより、初めての大規模な言語能力診断テストである。このテストの結果は、素点ではなく受験者が CEFR の A 1 ～C2 のどのレベルに該当するか判定される。また、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。自分の技能別レベルがそれぞれ CEFR のどのレベルであるかという説明と、そのレベルの学習者は典型的に何ができるかを通知してもらえる。これは CEFR のポリシーであり、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

ポートを提供していることになる。

DIALANG で判定する受験者の CEFR　A1～C2 のレベルの規準設定は、14 言語のそれぞれに専門家たちを集め、各技能に対して実行された。専門家たちは CEFR についてさらに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。」を基準にして一つずつの問題にレベル判定を下した。その後、評価者間の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て、最終的に分割点（カットポイント）を決めている。

この 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のが、斉田（2008）の試みである。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。彼ら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 レベルであった。ここで、Structure と Vocabulary を「言語知識」とし、Listening, Reading, Writing を「言語運用能力」とすると、被験者たちの「言語知識」は「言語運用能力」より 1～２レベル高い傾向にある。また、この学生たちの言語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 のセクションでいずれも 62～65％であった。

自己評価としての CDS

さらに、自己評価としての CDS の妥当性の確認をした研究として、Sato (2010)がある。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、ラッシュモデルを使って分析した。この 16 項目が中学生たちにとって、比較的困難度が低めであるという結果がでたものの、彼らの 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。また、受験者の自己評価と英語能力のレベル、さらに英語学習に費やした時間とも比例した。しかし、今回の研究対象とした 16 項目は英検 5 級～準 2 級までの CDS の一部であるので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、North & Schneider (1998)の研究で開発された能力記述文をもとにして、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し、比較したものである。しかしこのように、日本において CDS とスピーキング能力の規準設定に関する研究はあまり多くなく、前述した DIALANG や GTEC for STUDENTS にしてもライティングテストは含まれているが、スピーキングテストは開発中だと思われる。

残された課題

英検、TOEFL、TOEIC, IELTS, GTEC for STUDENTS などメジャーな英語能力テストは得点とリンクした CDS を設定している。しかし、それらの CDS の妥当性の検証については、あまりにも多様な受験者に対応していることもあり、十分実行されているとは言い難い。また、CDS が予想以上に広範囲に渡る多様な目的に使われ始めていることで、その利用価値が危ぶまれる状況もある。このような懸念を打開するために、研究者、教員、学習者が協力して実践的な例を持ち寄ったり、より正確な CDS のレベル設定のために話し合ったりする機会を多く持つことが妥当性の高い CDS の普及を成功に導く鍵になると思われる。また、比較的研究がなされていないスピーキングテストの CDS に関する妥当性の検証も含め、CDS の規準設定における妥当性の検証研究が急務である。

テストと CDS における規準設定は、テスト得点をもとにして、その受験者が実際の生活でどのくらい英語が使えるかを推定することを意味し、このうち①習熟度別ガイドラインとしての CDS は、受験者たちがテスト問題にどのように回答したのか、そ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状況でどのような英語を使うことを意味するのか記述したものである。もう一つは、②得点後付式の CDS で、これはテストと離れた現実の場面で、受験者が英語でできることを記述し、それに適応するテストの得点を導き出したものである。

方式①の DIALANG は、ヨーロッパの 14 言語に対して作成されたオンラインの言語テストで、規準設定での分割点は CEFR の６つのレベルである。DIALANG の規準設定の過程は、一部に Angoff 法を取り入れて、専門家を５～7 人集めて一つ一つのテスト問題のレベルを吟味するという時間と労力がかかるものであった。またこの膨大な作業を 14 言語について行ったのであるから、さぞかし大がかりな作業であったと思われる。

しかし Weir（2005）の主張に従えば、CDS は厳密に言うと、それを使用する国ごとに、さらに教育機関ごと、テストごと、言語カリキュラムごとに、その学習者、受験者に適した CDS として Tailor made する必要があるわけであるので、ヨーロッパの学習者のために作られた CEFR をそのまま日本人に適用することは考えられず、日本人学習者に適用する CDS を作成する必要がある。また、CEFR を基にして作成された DIALANG によってもたらされた結果を、日本人学習者にそのまま適用することも難しい。

日本人大学 1 年生 130 人が DIALANG を受験し、CEFR のレベル判定を受けた研究結果（斉田, 2008）を踏まえると、日本の大学 1 年生に対応する CDS は、CEFR の A1, A2, B1, B2, C1, C2 レベルのうち下半分の A1, A2, B1 レベルに集中し、あまりに多くの学習者が一つのレベルに入ることになり適応できないため、A1, A2, B1 それぞれをさらに２つずつに分けたりする必要があるだろう。

さらに、日本人学習者の中にももちろん様々な違いがあるので、理想的には、その英語

教育プログラムごとに独自の CDS を作成し、規準設定もするべきであろう。そしてその CDS と、①または②の方式でテスト得点をリンクさせていくことが求められる。

方式①習熟度別ガイドラインを、日本人学習者に適用するように作るには、日本人学習者たちが実際にどんな場面で、どのような英語を使った活動をするのかを長年にわたって注意深く観察し記述する必要がある。DIALANG（Alderson & Huhta, 2005）での規準設定の手順にもあったように、数人の専門家が合議しながら規準設定を進めるべきであろう。非常に時間がかかり労力を要する作業になるはずだ。CEFR は 2001 年に出版されるまで既に 10 年はかかったと聞いているし、現在でも English　Profile のような大きなプロジェクトを実行しながら、変更や改良を加える動きが続いている。最初の段階から 22 年経ってもまだ完璧に完成したとは言えないらしい。

Green (2010)は、CDS が当初の予想をはるかに超えて、非常に多くの国々のさまざまな教育機関において、異なった言語の指導や学習に関する多様な目的のために用いられることになったことに対して、その利用価値が危ぶまれるような問題に直面するのではないかと懸念している。このような事態の解決策の一つであり、今後の大きな課題としては、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り、より妥当な CDS のレベル設定のために意見を交換したり、積極的に協力しあう機会を増やしていく必要があると思う。

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)* pp. 301-320.

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

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.

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

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.

Lenz, P., & Schneider, G. (2004). *A Bank of Descriptors for Self-assessment in European*

Language Portfolios. Strasbourg: Council of Europe.

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

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.

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., & Schneider, G. (1998). Scaling descriptors for language proficiency scales. *Language Testing, 15(2),* 217-263.

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

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

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.

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.

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.

3.4. The role of can-do statements in English language classrooms

Tomoko Fujita

Abstract

In language education in Japan, the popularity of the Common European Framework of Reference for Languages (CEFR) has been increasing rapidly these days. Many English language programs have started introducing the Can-Do statements (CDS) from the CEFR into their curriculum. The cases of four Japanese university language programs, which attempted to introduce CEFR into their curriculum, are illustrated in the first report. In the second report, two problems of introducing CEFR into school curricula in Japan are explained. To begin with, CEFR was written for European language learners with vastly different cultural and linguistic circumstances. Therefore, using CEFR as it is in classrooms in Japan causes problems. Teachers make a better use of CEFR when it is modified for their Japanese students. In addition, CDS for teacher evaluation, self-evaluation, and goal-setting should be customized for each language program in order to form a CDS oriented language framework. Then, tasks driven from the CDS, called can-do tasks will function as learning tasks as well as self- and teacher's evaluation tasks. This will then link together and make efficient relationships for a successful CDS-based curriculum.

３．４．Can-do statements が英語の授業において果たす役割

藤田智子

2001 年に欧州評議会がヨーロッパ言語共通参照枠: Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001） を発表し、その日本語訳（吉島・大橋, 2004）が出版されると、日本における外国語教育や日本語教育に携わる教員たちの間で CEFR と European Language Portfolio (ELP)が脚光を浴び始めた。この影響を受けて、大規模なテスト開発機関ではスコア解釈のために、また教育機関では学生の到達目標の設定や自己評価のための参照枠として Can-Do statements (CDS) が開発されている。そして、学習サイクルのなかに CDS を組み込んで、より効果的なカリキュラムを作り上げようとする試みが行われている。ここでは、日本の大学における、CEFR 導入のための実践的な事例研究について取り上げ、そのあと日本で英語の授業に CDS を取り入れる際の問題点と残された課題について言及する。

日本における Can-Do statements を英語の授業で活用する取組

日本の大学や高校の英語の授業に CDS を取り入れる動きは緩やかに広まりつつあり、これは主に学習者の授業での到達目標、あるいは、言語能力発達段階に関する評価基準として用いられることが多い。Naganuma (2008, 2010)は、日常の英会話ではなく、日本の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 を作成した（Naganuma & Miyajima, 2006）。また、これに先立ち、SELHi 指定高校で、 CEFR と高校の授業内容を対応させて作成された香住丘 Can-Do グレードの開発も行い、日本の教育機関のために開発した CDS を中心にしたカリキュラムを実際の授業に取り入れるプロジェクトに活発に取り組んでいる。

茨城大学での研究と取組

茨城大学では英語の授業を CEFR (Council of Europe, 2001）のＡ１， Ａ２， Ｂ１－１，Ｂ１－２，Ｂ２という５段階のレベルに合わせて習熟度別クラスを編成し、それぞれのレベルの到達目標や学生の自己評価表（Can-Do checklist）を、CEFR を参考にして作成し、これらを中心に据えた総合英語プログラムを開発した。ここでは、自律的学習も推奨され、語学の学習が生涯学習になっていく中で、大学時代に自律的に学習できる人になることが

大切だという考え方をしている（Nagai & Fukuda, 2004; Ano et al., 2007; Fukuda, 2009）。

また、Fukuda (2009)は、CEFR を日本の英語教育に取り入れるには、段階を追って取り組む必要があると述べている。 それは、1. 組織の形成：CEFR を、取り入れる教育機関に適応するようにプログラムデザインし、それをわかりやすく担当教員に説明できる専門家集団を形成する（Muranaka、2010）。2. プログラム開発：CEFR を基にしながらそのプログラム独自の基本コンセプトを作り、レベルを設定し、カリキュラムやシラバスのデザイン、および評価方法を開発する。3. 普及活動：担当教員への説明を徹底するため教員の研修を行い、より多くの教員や関係者に理解を深めてもらうことが必要であるとしている。

Nagai (2010)では、日本の英語高等教育のカリキュラムに CEFR を導入する際のガイドラインとしてその有用性を 3 点挙げている。まず、第 1 点は多様なコミュニケーション言語の活用と学習方法について習熟度別の能力記述を提供していること、第 2 点として言語活動を行う際に必要な一般言語能力とコミュニケーション言語能力を特定すること、さらに第 3 点として、能力記述文を用いて英語カリキュラムや特定の英語科目の目的と学習成果をあらかじめ策定しておくことが可能であることである。そして、CEFR によってカリキュラムを構築する際に、どのように応用可能であるかを具体的な事例を示しながら説明することを可能にし、最後に CEFR に基づいてカリキュラムやコースを構築することにより、日本の英語教育プログラムをより一貫性のあるプログラムにすることができると結論づけている。

大阪大学での研究と取組

大阪大学では、25 の専攻語すべてにおいて、到達目標を CDS で表して公開し、「透明」「共通」「強制しない」姿勢でカリキュラム改革を行ってきた。Majima (2010)は、日本で CEFR を取り入れた言語教育を行っている事例を調査し、活用分野に分けて紹介した。（１）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（到達度目標を CEFR に基づいた CDS で実施したもの)。（２）シラバス・デザインとカリキュラム・デザインに CEFR を利用したもの。（３）ポートフォリオを学習促進の動機づけと代替評価のツールとして使おうとするもの。（４）標準テスト、大規模テスト、検定試験を CEFR に関連づけたもの。（５）アセスメントと評価に CEFR を活用するもの。（６）教材開発に CEFR を活用したもの。（７）教員研修に CEFR を利用したもの。以上、７つの分野である。

しかし、真嶋 (2010) は、CEFR をカリキュラムに導入することに対する批判についても言及している。CEFR は日本で使うことを想定して作られていないので問題が生じるという批判があり、また十分な議論なく安直に 6 段階の共通参照レベルを導入しようとする動きは危険であると指摘している。しかし、彼女は、CEFR を作成した人たちが CEFR を運用するに当たって「透明性」「共通性」を強調していて、CEFR を絶対視せず、教育現場に

合うように変更して使ってほしいという立場であること(Trim, 2002)を忘れてはならないと述べている。

慶應義塾での研究と取組

慶應義塾大学外国語教育センターでの研究プロジェクト（Action Oriented Plurilingual Learning Project (AOP))は、行動中心で自律的な復言語学習環境の整備促進をめざして、慶應義塾の小中高大一貫教育にCEFRやELPを基にした英語教育を実施しようとするものである。それに伴って、慶應義塾内の教師間の協調連携を高めることも目標としている。この中心的な取組の1つとしては、ELPのジュニア版といえる日本版（慶應ELP）を開発し試行したことが挙げられる。この実施後に、生徒と教師を対象にアンケートとインタビューを行った。その結果からの考察として、小中高大一貫教育のなかで、それぞれの学校が自律性を尊重するあまり、一貫したカリキュラム改革に対して動機づけが低くなるが、独自性が高いゆえに緩やかな枠組みこそが重要になることを指摘している（Horiguchi, et al., 2010）。

CEFRと日本の英語教育

CEFR はヨーロッパの言語学習者のために作られたもので、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用したものにする必要性が強調され始めた（境, 2009; 根岸, 2006）。このような動きのなかで、今後の日本の英語教育に CDS が果たす役割について、より探求する必要があるのは、（１）CEFR を日本人学習者向けに変更や工夫をする研究と、（２）日本人学習者向け CEFR を、到達目標、教師評価や自己評価と、実際の授業タスクの内容に有機的に結びつけて一貫したプログラムとして運営するための研究である。

CEFR を日本人学習者向けに変更や工夫をする研究

中島・永田（2006）は、CEFR 準拠の DIALANG self-assessments (SAS)という自己評価アンケートを使って、CEFR がどのくらい日本人学習者に適用可能か検証した。CEFR の CDS 各項目の能力記述に対して、日本人学習者がそれらをどのような困難度レベルと認識しているかを調査した。また、根岸（2006）は、この中島・永田（2006）のなかで、設定されている CEFR の困難度レベルと、日本人学習者が答えた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「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A2 レベルと判定した。おそらく、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のが、Happy Birthday! のような短いメッセージではなく、もっと長い情報量であったからだと思われる。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 の項目に対しては、日本人学習者たちは 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、高い困難度だと思ったからだと推測できる。根岸（2006）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、参考資料を付けることで学習者が具体的に内容を理解するための工夫をして成果をあげた。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取りの例を示した。改良後、それぞれの項目の困難度は、ほぼ設定どおりの順序となった。

また、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実施されている。ここでは、日本人学習者のレベルを考慮して、CEFR の下位レベルをより細かく分けている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つに分けるレベルの設定を行った（岡、2008）。

そして、結果が CEFR のレベルで判定される言語能力テスト DIALANG の英語版（Alderson & Huhta, 2005）を使った斉田（2008）の調査によると日本人大学 1 年生のリスニングは CEFR の A1 レベル、リーディングは A2 レベル、ライティングは A2 レベル、文法は B1 レベル、語彙は A2 から B1 レベルで、ほとんどが A1～A2 という非常に狭いレベル範囲に入るという結果になった。これは、CEFR を日本人学習者に適用させるレベル設定をするには、A1、A2、B1 の中に、より詳細なレベルを設定する必要があることを示唆している。

CEFR による一貫したプログラムとして運営するための研究

このように、CEFR が日本人学習者に適用するように改良されたその次のステップとして、日本人向け CEFR を英語教育の授業に取り込む実践的な研究が必要になる。例えば、到達目標を示す CDS の作成が実現しても、その目標に達するためのコースデザインやカリキュラム、シラバス作成にリンクしなければ、実際の授業に CEFR を取り入れたことにはならない。長沼（2009）は、「Can-Do 評価-学習タスクに基づくモジュール型シラバス構築の試み」として、①Can-Do チェックリスト、②Can-Do 評価タスク、③Can-Do 学習モジュール、これらの開発を３つの過程に分けて説明している。

Can-Do チェックリストは、CEFR における自己評価型の Can-Do リストのことである。

これは、学習段階を示すための指針となり、また学習者にとっては自己の学習段階を確認するための道具として機能することが期待されている。しかし、チェックリストとしてCEFR に対応している European Language Portfolio をそのまま使うことは、外部指標として汎用性が高いというアドバンテージがあるものの、日本人学習者が彼らの教室での学習到達段階に合わせて、具体的に学習段階を把握する材料にすることは難しい。そこで Can-Do チェックリストも日本人学習者に合わせた内容に修正し、実際に授業で学習しているシラバスに基づいた内部指標として開発することが求められる。Can-Do チェックリストは、学習者が自己評価するものであるが、彼らが体験したことがない内容を自己評価として質問しても、その回答はあまり正確ではないことが知られている（伊東・川口・太田、2008）。

これらの問題を解決し、より正確な Can-Do チェックリストにするための試みとして、長沼・永末（2007）は、香住丘高校の Can-Do リストに授業で実際に使っている教科書名をあげ、例えば「『Reading POWER』のテキストを 1 分間に 150 語読むことができる。」というような項目を設けている。しかし、これでは教科書を変えたとき Can-Do チェックリストも変えなければならない。このように Can-Do チェックリストを継続して利用したり、外部に示したりする必要性から、具体性を損ねない範囲で一般的な記述にせざるを得ない場合も多く、そのような Can-Do チェックリストは必ずしも学習者にとって分かりやすい記述とはなっていないのが現実である。

長沼・永末（2009）は CDS で掲げられた到達目標を達成するための「精読スキル」に焦点を当てた学習タスクを開発した。これらは Can-Do 教員評価・学生自己評価と学習が一体化したタスクとなっている。このように、Can-Do 評価タスクは CDS における自己評価の客観的な検証のためのツールとなり（吉池、2006 ; 竹村、2008）、評価タスクが学習タスクとしても機能し、これらが授業に組み込まれることにより、自己効力を育てながら学習を進めることが可能になると考えられている。このように、Can-Do 評価-学習タスクを Can-Do リストと有機的に関連させながら授業内で展開していくことで、初めて自律的な学習が可能になる。

残された課題

日本での英語教育の現場では、早いところでは 2000 年代前半から、ヨーロッパ言語共通参照枠: Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001）や European Language Portfolio (ELP)をカリキュラムに取り入れようとする動きが始まった。Nagai & Fukuda（2004）が CEFR を日本の英語教育の現場（茨城大学）に取り入れることについて書いた初期の論文の一つである。その後、現在に至るまでに、日本の大学や高校の英語教育プログラムで、次々とカリキュラムに CEFR を取り入れる動きが始ま

っている（Naganuma, 2010）。

しかし、大きな組織ほど、到達目標としての CEFR がカリキュラム、シラバス、授業でのタスクや評価まで浸透し、有機的に一貫したものになっているとは言い難いのが現状だと思われる。今、残された課題は、CEFR のめざす理念を日本の英語教育の現場に取り入れるために、CEFR を日本人学習者に適用させ、授業タスクまで一貫した CEFR ベースのカリキュラムを導入することである。

CEFR は、日本人学習者のためではなく、ヨーロッパの言語学習者が外国語教育のシラバス、カリキュラム、テスト、教材などを作成するときに共通の基盤を提供するために作られたヨーロッパ言語共通参照枠組みである(Council of Europe, 2001)。そして、これは多くの研究者たちが長年に渡って達成した大きな成果である。CEFR に基づいた自己評価チェックリストである European Language Portfolio (ELP)や、CEFR のレベルに適応した評価システムである DIALANG も開発されているため､そのまま日本人学習者が利用できれば､非常に便利である。

しかし、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、明らかに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、その言語学習の現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）。日本人学習者に適用させるための研究こそが、今後 CEFR を日本の言語教育の現場に普及させることができるかできないかを決定する鍵になると考えられる。しかしながら、現状では十分に多くの実証研究はされていない。

次の段階としては、この日本人向け CEFR を、到達目標、教師評価や自己評価として実際の授業タスクの内容に至るまで有機的に結びつけ、一貫したプログラムとして運営する取組が求められている。具体的に言うと、ある教育機関で到達目標として CDS を設定し、学習者たちがその目標に達成できるようにするための学習タスクを作成し、それらのタスクを集めた教材を作り、その到達目標がどのくらい達成されているか測定するためのテストを作成し、自己評価のための CDS を用意する。さらにこのプロセスを、学習するスキルごとに、学習者たちの習熟度に合わせていくつか作成する必要がある。CEFR や ELP を授業の一環として取り入れているところはあるが、このように、シラバス、テスト、教材、実際の授業でのタスクまで開発し、一貫して有機的に結びつけた言語教育を現場で実践する段階には、ほとんどのところでは至っていないと考えられる。しかしながら、このように一貫した「Can-Do 評価―学習タスク」の開発まで行うことによってはじめて、CEFR に基づくカリキュラムが実現するわけなので、今後これらの取り組みとその事例研究を実践していかなければならない。

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)* pp. 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.

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.

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.

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.

Muranaka, T. (2010). The significance of constructing program design on national universities : Two cases of reform about English course. *Studies in Social Science,* Ibaraki University, 50, 83-103.

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.

Nagai, N. (2010).Designing English curricula and courses in Japanese higher education: Using CEFR as a guiding tool. 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. 86-104). Tokyo: Asahi Press.

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

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.

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.

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.
真嶋潤子（2010）.『CEFR における評価とアセスメント』佐藤慎司・熊谷由理（編著）アセスメントと日本語教育 - 新しい評価の理論と実践. pp. 38－48. くろしお出版.
長沼君主 (2009). Can-D 評価-学習タスクに基づくモジュール型シラバス構築の試み. 『東京外国語大学論集』第 79 号, 87-106.
長沼君主・永末温子（2007）. 香住丘 Can-Do グレードに基づく Can-Do チェックリストの開発とその運用.『第 33 回全国英語教育学会大分研究大会発表予稿集』323－324.
長沼君主・永末温子（2009）. Post-SELHi 実践における Can-Do タスクによる授業モジュール化の試み.『第 35 回全国英語教育学会鳥取研究大会発表予稿集』230－231.
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（2006）. 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.
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.
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, 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.
竹村雅史 (2008). 英検 Can-Do リストによる Writing 技能に関する妥当性の検証. 『STEP BULIETIN』. 20、251－261.
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』朝日出版社.
吉池陽子 (2006). リーディングの Can-Do statements の妥当性の検証：自己評価と実際のパフォーマンスとの関係について. 『外国語教育研究』9, 25-42.

4. Test Theory and Standard-Setting

4.1. Standard-setting based on IRT

Tomoko Fujita

Abstract

Compared to the classical test theory (CTT), item response theory (IRT) has more advantages. Generally, IRT makes it possible to predict each examinee's performance on a test item based on the level of difficulty of the item and the participant's level of ability. The three main IRT models are the one-, two-, and three-parameter models. Two other models, the three-parameter *c*-fixed model and discrete-item response model were introduced by current research related to IRT. All of the models are based on different formulas and assumptions regarding item properties, and each requires different numbers of examinees for valid estimations. The beauty of IRT analyses is that the test information curve (TIC) indicates how accurately the test measures examinees' ability at each ability level. The TIC should be the highest at the examinees' ability level around the cut-point of a test because critical decisions are made at the cut-point. In this way, TIC provides important information for test developers or teachers whenever they calibrate and revise test items for tailored tests. However, advanced knowledge of psychometrics should not be required for teachers. The studies make it possible to simplify the complicated computing are necessary.

４．テスト理論と規準設定

４．１．IRT を活用した規準設定

藤田智子

規準設定は、教育プログラムでの個人に関する「意思決定」（例えば、入試やクラス分けなど）をするときの大きな要因となることが多い。しかしながら、規準設定は教育的評価の「アキレス腱」と呼ばれていて、多くの方法のなかで、これが最高の方法であると断言できるほどはっきり広く共通認識を持って言えるものではないと言われている。また、どの方法を用いたとしてもその方法の妥当性を検証することは容易ではない(Kane, 1994)。ここでは規準の正確さこそが、これら「意思決定」の妥当性の確立の中核となる（Plake & Hambleton, 2001）。この規準の正確さを追及するための手段のひとつとしてＩＲＴを活用した規準設定が開発されている。本論では IRT を利用した規準設定についての先行研究をまとめ、その中で残された課題について述べる。

項目応答理論（IRT）

IRT のアドバンテージ

古典的テスト理論（ＣＣＴ）は依然としてテストに係る研究の分野で大きな影響力を持ち続けている。教育現場の教師たちはもちろんのこと、テスト開発者たちの中にも、ＣＣＴアプローチを利用してテストの分析を実施している者も少なくはない。（Suen, 1990）。Henning (1987)は、ＣＣＴは項目とテストの分析を中心として、統計学的には相対関係に大きく依存したアプローチが主であると述べている。

しかしながら、このＣＣＴと比べ、項目応答理論（ＩＲＴ）には、主に以下の３つのアドバンテージがある。（１）異なったテストフォームでも受験者の能力が比較可能(Test-free person measurement)。（２）異なった受験者集団でも共通の項目特性を推定できる（Sample-free item calibration）。（３）能力レベルごとに得られる情報量がわかる (Multiple reliability estimation)。 これらは、ＩＲＴモデルがそれぞれの受験者能力を、テスト項目と受験者の能力を切り離して推定することができることに由来する。

IRT の３つのモデル

IRT の代表的なモデルとしては、１パラメタ(1PL)、２パラメタ(2PL)、3 パラメタ(3PL)

の３モデルがあるが、ここでは３パラメタ c fix (3PLcFix)を加えて 4 種類のモデルについて紹介する。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）。

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

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

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

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）の推定もできるが、安定した推定に必要とされる受験者数は 500 人から 1000 人(eg, Ayala, 2009)と言われている。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）も推定できるが、安定した推定には 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、野上(2009) や野上、小林、& 林 (2010)では、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。どの IRT モデルを使うかによって、結果が大きく変わることがあるので、Choi and Bachman (1992) は、分析するデータや目的を良く考慮してどのＩＲＴ項目応答モデルに最も適応しているか、慎重に吟味する必要性を説いている。

IRT を使った分析をするために作られたコンピュータソフトはたくさんあるが、その中でも比較的シンプルなのが、 Xcalibre (1997)や BILOG3 (1989)である。最近は無料でダウンロードできる R を使って、自分で IRT の式を書いて分析し、図や表も簡単に導き出すことができるようになったが、これには統計理論についてのより高い専門知識が要求される。

IRT の前提

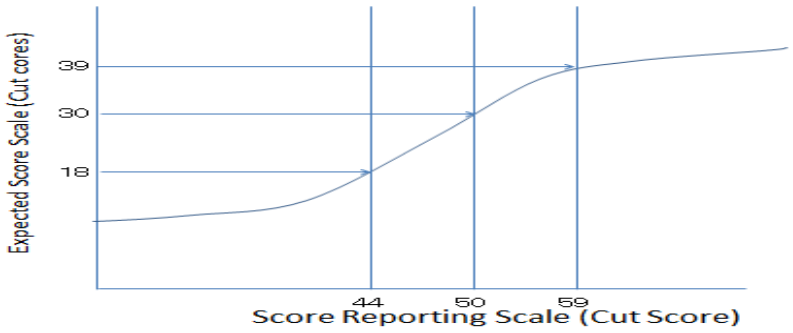
ＩＲＴを利用して規準設定を実行するときは、ＩＲＴの前提条件を満たしておかなければならない。その主なものは、局所独立（local independence）と一次元性（unidimensionality）である。局所独立の前提は、被験者が一つの項目に正解できる確率が他の項目に正解でき

る確率に影響を与えず、テストの中の、ある項目に正解する確率は独立しているという前提である。そして、一次元性の前提とは、一つのテスト項目は唯一の能力を測るものでなければいけないという意味である（Henning, 1987; 大友, 1996）。

テスト特性曲線を利用した規準設定

IRT によって得られた、あるテスト項目一問の受験者の能力、項目困難度と正答確率の関係を表した曲線を項目特性曲線（item characteristic curve: ICC）と呼んでいる。そして項目ごとのＩＣＣを一つのテストとして集合体にしたものがテスト特性曲線（test characteristic curve: TCC）で、能力が θ の受験者に関するテストの平均正答確率の推定値を意味している (大友, 1996)。

図 1.　TCC マッピング法



さて、テスト特性曲線を利用してカットスコアを設定する方法は、ＴＣＣマッピング法（Plake & Hambleton, 2001）と呼ばれている。Figure 1 に描かれたＴＣＣの縦軸は正答確率、横軸が能力値を示している。これに各レベルに設定した正答確率（例：18％初級、30％中級、38％上級）から水平に線を伸ばし、そのＴＣＣとの交点から垂線を下したところの能力値を得点に換算して分割点（カットスコア）とする方法である。

テスト情報関数

IRT による規準設定の優れた点は、項目情報関数が得られ、その項目が最も精度高く測定できる能力値を持つ被験者を特定できることである。それぞれの能力レベルでの項目情報関数を合計し、テスト全体として得られる情報量を表したものを、テスト情報関数とい

う。このテスト情報関数を、受験者の能力値を横軸にとって表した曲線は、そのテストを受験したどの能力レベルの受験者の情報が最も多く得られたのかを表すものである。(第 2 回報告書の図１参照）つまり、この情報曲線によって、そのテストはどの能力値レベルの受験者の能力を最も精度が高く測ることができたかを表している（Embretson, 1999; Hambleton, Swaminathan, & Rogers, 1991; Suen, 1990)。

IRT による規準設定の優位性

IRT による規準設定は、テスト情報関数を利用することで、最も精度を高く測りたい能力水準の項目を多用した「仕立て式テスト」を作ることができる(e.g., Brown, 1997; Henning, 1987; 大友, 1996)。例えば、あるプレースメントテストを実施して、受験者の能力値が θ＝－１を下のカットスコア、θ＝１を上の分割点として初級、中級、上級の３つの習熟度レベルに分けたいとき、θ＝－１と１の項目困難度の問題を多くしてテストを構成することで、分割点周辺の能力水準の被験者能力をより慎重に測ることができる。
これは、個々のテストの目的に合うように設定した能力水準の項目を多く用いてテストを構成することで、最も慎重にどちらのレベルに入れるべきか決定すべき受験者たちの能力をより正確に測り、テストの目的にあった効率の良い意思決定を可能にする。

IRT を使ったテスト情報関数による規準設定

テスト全体として得られる情報量を表したものを、テスト情報関数というが、このテスト情報関数を受験者の能力値を横軸にとって表した曲線は、そのテストを受験したどの能力レベルの受験者の情報が最も多く得られたのかを表すものである。(Embretson, 1999; Hambleton, Swaminathan, & Rogers, 1991; Suen, 1990)。したがって、最も精度を高く測りたい受験者集団の能力値水準、つまり分割点（カットポイント）周辺の困難度を示す項目を多くしてテストを構成することで、「仕立て式テスト」にすることができる（e.g., Brown, 1997; Henning, 1987; 大友, 1996)。

例えば、プレースメントテストを実施して、受験者の能力値が θ＝－１を下のカットスコア、θ＝１を上のカットスコアとして初級、中級、上級の３つの習熟度レベルに分けたいとき、θ＝－１と１の項目困難度の問題を多くしてテストを構成することで、より正確なクラス分けが実施できる。なぜなら、最も慎重にどちらのレベルに入れるべきか決定すべき分割点周辺の能力レベルの受験者たちの能力を、特に正確に測ることができるからである（図２参照)。IRT による規準設定は、一般的にこの特徴を利用して、設定した規準に近い

困難度の項目を多く用いてテストを構成し、より正確で効率の良い意思決定を行うことができる。

図2 テスト情報関数

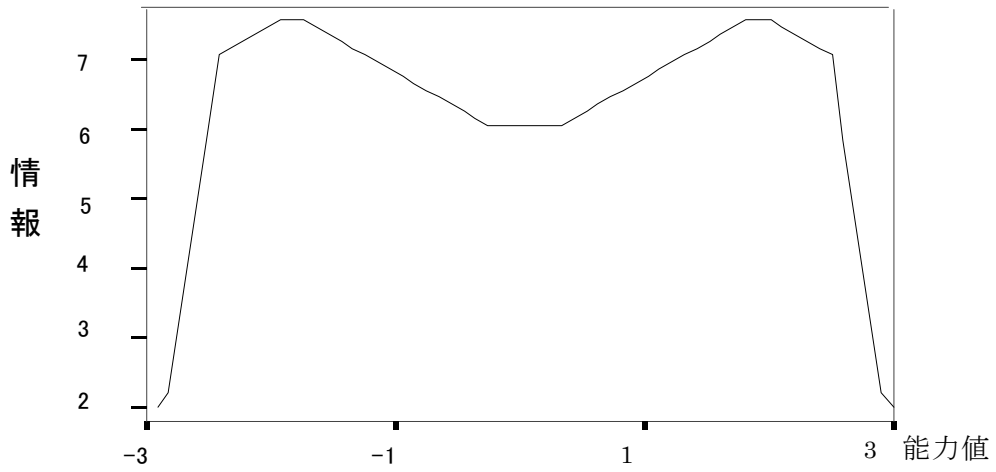


図1. テスト情報関数

離散化 IRT

プレースメントテスト、入試や期末試験などにおいては、受験者一人一人のスコアを連続した値として求める必要はなく、グレードA、B、Cなどのように段階的な情報だけが分かれば十分だという状況のこともある。このような時のために、受験者の能力を順位尺度による段階的評価で示す、潜在ランク理論（Latent Rank Theory: LRT）（Shojima, 2007, 2009: 植野・荘島、2011）が提唱されている。しかしながら、この方法による安定した推定には大量データが必要であるため、データが少数のときでも有効に段階的評価ができる方法として離散化 IRT による方法も可能である（澤谷・前川、2012）。これは項目応答理論の能力特性値 θ を離散化したモデルで離散化 IRT（Discrete-Item Response Theory、D-IRT）と呼ぶことにする（(4) 式）。

$$P(\theta|\xi) = \mathrm{P}(\theta_q|a_j, b_j) = \frac{1}{1 + \exp\left(-\mathrm{D}a_j(\theta_q - b_j)\right)} \quad \cdots\cdots(4)$$

D-IRT は、LRT と比べるとθ_q（ランクごとの能力特性値）と P（項目ごとの正答確率）の間の関数を固定するために、データ数が少なくてもパラメタが安定して推定できると考えられている。また、BILOG などの既存の IRT のためのコンピュータプログラムを使って項目パラメタの推定を行うこともできる。

IRT と古典的テスト理論によるテスト分析結果の比較

Culligan＆Gorsuch (2000)は、日本の大学英語教育プログラムにおいて、同じプレースメントテストのデータを、１．素点をもとにして、分割点を決め受験者を素点で習熟度クラスに分けた場合と、２．IRT（Rasch モデル）によって推定した θ（能力値）をもとにして分割点を推定し、受験者の θ によってクラスに分けた場合の 2 通りの方法で分析して結果を比較した。彼らは、IRT（Rasch モデル）を使って推定した受験者の能力値は、プレースメントテストの一問ずつの項目困難度をもとに推定されているので、古典的テスト理論による一括した計算によるものと比べるとより精度が高く、かつ多くの情報を得られると指摘している。そして、素点と θ、この二つの方法の結果を比較すると、IRT によってクラス分けされた結果と、素点によってクラス分けされた結果の間には相違があり、彼らの研究では、全受験生の 5%にあたる受験者たちが、素点と θ による方法で異なるクラスに入る結果となった。

残された課題

古典的テスト理論に比べ、項目応答理論（IRT）による規準設定は、優位性が大きいことは先に述べた。しかし IRT の問題点としては、教育現場の教員にとって、誰でも容易に理解できるわけではなく、ある程度の時間をかけて専門的な知識を学び、経験を積まなければ日常的に使えるほどにならないという点が挙げられる。この負担をできるかぎり軽減し、優位性の高い IRT による規準設定を広く普及させていくことが、今後の課題だと考える。

まず、3 つの IRT モデルの違いを認識し、分析しようとするデータがどのモデルに最も適応するか決定しなければならない。1PL モデルは、あまり多くの受験者が居なくても、安定した推定ができると言われているが、b パラメタ（項目困難度）だけしか推定することができず、分析によって得られる情報の量はあまり多くない。次に 2PL モデルは、安定した推定に必要とされる受験者数は 500 人から 1000 人であるが、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）も推定できるが、安定した推定には 1000 人以上必要だと言われている。

その上で、忘れてはならないのが IRT の前提である。局所独立の仮定（assumption of local independence）は、ある受験者が一つの項目に正解できる確率が、同じテストの他

の項目の正答確率の影響を受けてはいけないことを意味している。また、一次元性（unidimensionality）とは、一つのテスト項目は、その受験者の一つの能力を測定しなければいけないということである（大友、1996）。

このような前提や、いろいろと複雑な条件はあるが、何と言ってもテストの規準設定をする上で、古典的テスト理論に比べ IRT には多くの優位性がある。IRT によってプレースメントテストを分析して習熟度別クラス分け編成を行った場合は、素点で同じことをするのに比べ、５％もの受験者がより正確なクラスに編成される可能性が高い。また、テスト情報関数によって、分割点の困難度に近い能力値水準の項目を多めに構成した「仕立て式テスト」を作ることもできる。

それでも、実際の教育現場において、専門知識の不足から規準設定に IRT を使用しているところはあまり多くないと思われる。IRT を一般に普及させ、より簡単に教育の現場で正確な規準設定ができるようにしていくことが今後の課題である。

参考文献

Assessment Systems. (1997). *Xcalibre* (version 1.10e). St. Paul, MN: Assessment Systems.

Ayala, R. J. (2009). The Theory and Practice of Item Response Theory. New York: Guilford.

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

Brown, J. D. (1997). Computers in language testing: Present research and some future directions. *Language Learning & Technology, 1,* 44-59.

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.

Culligan, B., & Gorsuch, G. (2000). Using item response theory to refine placement decisions. *JALT Journal, 22*(2), 315-325.

Embretson, S. E. (1999). Issues in the measurement of cognitive abilities. In S. E. Embretson & S. L. Hershberger (Eds.). *The new rules of measurement* (pp.1-15). Mahwah, NJ: Lawrence Erlbaum.

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

Henning, G. (1987). *A guide to language testing.* Boston: Heinle & Heinle.

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

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.

Mislevy, R.J., & Bock, R.D. (1989). *BILOG 3: Item analysis and test scoring with binary logistic models*. Mooresville, IN: Scientific Software.

Plake., & Hambleton. (2001). The analytic judgment method for setting standards on complex performance assessments. In G. J. Cizek (Ed.), *Setting performance standards* (pp. 19-52). Mahwah, NJ: Lawrence Erlbaum.

Shojima, K. (2007). Neural test theory. *DNC Research Note*, 7, 2.

Shojima, K. (2009). Neural test theory model for granded response data. A paper presented at the International Meeting of the Psychometric Society (IMPS), Cambridge.

Suen, H. (1990). *Principles of test theories.* Hillsdale, NJ: Lawrence Erlbaum.

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.

大友賢二 (1996). 『項目応答理論入門』.東京: 大修館.

澤谷秀之・前川眞一（2012）. 離散化 IRT における推定値の傾向およびその精度に関する研究『東京工業大学大学院修士論文』.

芝祐順(編) (1991). 項目応答理論－基礎と応用－東京大学出版会.

植野真臣・荘島宏二郎. (2011). 『学習評価の新潮流』、東京：朝倉書店.

4.2. Study of placement tests

Tomoko Fujita

Abstract

The study of placement tests has been often focused on the difficulty levels for the efficient grouping of students. Test items showing the examinees' ability level around cut-point are more often included in the test because critical decisions are made at the cut-point. However, recent studies focus more on if the tests rank students consistently with respect to the range of skills, knowledge, or abilities taught in the course. For example, a grammar test should discriminate between learners who mastered the past perfect from those who did not. Another factor of placement tests relates to technology because tests are required to be scored in a short amount of time and long complicated procedures should be shortened. A research conducted even 13 years ago still showed a computer-based test (CBT) appears to contribute more to the decision-making process than a pencil-and-paper test. Currently, language programs in Japan place more value on fostering students' speaking skills despite fewer chances to measure learners' speaking skills. Therefore, computer-based placement tests and speaking tests need to be developed soon.

４．２．テストによる規準設定：プレースメントテストの研究

藤田智子

テストで規準設定をし、プレースメントや入学試験の合否などの教育プログラムにおけるレベルの決定をするには、集団規準準拠テスト（norm-referenced test）が適切である。(Brown, 1996; Culligan & Gorsuch, 1999)。Brown (1989, 1990, 1992, 1995a, 1995b)は、プレースメントテストを開発するにあたって、まず、テストは受験者の能力に合っていること、そしてその言語プログラムのカリキュラムやそのプログラムの特徴に適応していることが不可欠であると指摘している。次に、そのプログラムの教育目標にあった内容をテストし、最後に、予備テストを実施して結果を分析し、その能力が分割点に近い受験者たちの能力を最も正確に測定できるように改訂するべきであると述べている。本論では、量的研究や質的研究に重点をおいた、国内外におけるプレースメントテストやその妥当性についての先行研究を解説し、その中で残された課題について論じる。

海外の言語教育プログラムでの IRT を利用したプレースメントテストの研究

Fulcher (1997)はラッシュモデルを利用して Surrey 大学の新入生に対して実施したプレースメントテストを分析した。エッセイ、文法 10 問、読解問題 8 問を 45 分で行うテストは、項目数が少ないこともあり信頼性が低くなった。しかしながら、このテストは上級と初級の２つのレベルに分けるプレースメントテストであったので、分割点はちょうど標準値（θ=０）に近い。そのため、このプレースメントテストのテスト情報関数（test information function: TIF）の値が、受験者の能力水準が θ＝-1.00～0.50 で最も大きくなったということは、分割点 θ＝０に近い能力値をもつ受験者の能力を、より正確に測定できたと報告している。

また、Kondo-Brown and Brown (2000)は３パラメタＩＲＴモデルを使って、ハワイ大学の日本語教育プログラムのプレースメントテスティングシステムをより効率的に、より正確に習熟度別に分けることができるようにするための研究を実施した。日本語テストのなかのリスニング、文法、語彙、作文をそれぞれのサブセクションごとに分けて分析し、項目分析の結果によって弁別力の低い項目を削除したあと、難易度の低い項目から高い項目に並べてさらに吟味して適切ではない項目を削除した。この結果、もともと70問あったテストを50問にし、テストを短くしたにも関わらず、テスト信頼性はほぼ同じ水準を維持していることが確認できた。

また、Green (2004)は、文法の知識を測ることでクラス分けする Global Placement Test (GPT)を使って、受験者がＧＰＴの結果によって配置されたクラスで学ぶ文法知識を、習熟しているかどうか診断できているかを調査した。この研究ではＧＰＴともう一つ別の文法テストを、イギリス、ギリシャ、日本の英語学習者合計 1070 人が受験し、その結果をＩＲＴで分析した。その結果、項目困難度は、テストで測ろうとしている言語の文法的要素よりも、項目のタイプによって影響を受ける可能性が大きいことが分かったと報告している。彼は、この結果によって従来から用いられてきた文法ベースのプレースメントテストが、履修者のパフォーマンスの診断や、ステップアップ教材使用のための充分な情報を提供しているとすることに対する疑いが出てきたと指摘している。

プレースメントテストの結果に対する学生の反応

Bradshaw (1990) はプレースメントテストの結果に対する 96 人からの反応をインタビュー調査した。被験者を母国語、性別、習熟度などによってグループに分け、その反応に違いがあるか分析したが、グループ間では顕著な違いは認められなかった。しかし、低得点グループが高得点グループに比べると、テストそのものを高く評価していないという傾向があることが確認された。著者は、テスト妥当性や信頼性は頻繁に調査されるが、残念なことにテスト結果やそれによるクラス分けに対する受験者の態度、感情、満足度などの質的な部分はめったに調査されることはないと述べている。

日本でのプレースメントテスト

Sugimori (2002)が調査した日本の大学 194 校のうち、64%にあたる大学で統一テストが実施され、そのうちの 56%の大学で、統一テストが習熟度別クラス編成のためのプレースメントテストとして利用された。これらのプレースメントテストは、教員作成による独自のテストが一番多く全体の約 36%を占め、次いで、TOEFL, TOEIC, G-TELP や英検などの市販テストが実施されている。さらに杉森（2003）は、プレースメントテストで測定しようとする技能は、リスニング、リーディング、文法・語法が中心で、英語発信能力であるライティングは約 5%､スピーキングは約 3%以下の大学でしか測定されていないと報告している。

Culligan and Gorsuch (1999)は、市販のプレースメントテストを実施することの長所は、便利さ、経済性、実施と採点の簡単さを挙げているが、対象とする学生やそのカリキュラムに適応しないなどの問題点も指摘している。彼らはＳＬＥＰ(Second Language English

Proficiency Test) という商業ベースのテストを 487 人の日本人大学生に実施し、結果を分析したところ、受験した学生の能力に合わないという結論に至った。

また、前年に発表した研究をベースに、Culligan and Gorsuch (2000)は、Rasch モデルによって推定した分割点と素点による分割点を比較する研究を実施した。その結果、Rasch モデルによる分割点のほうが正確で、素点による分割点では受験者の５％が不適切なクラス分けをされる可能性があると報告している。また、彼らはプレースメントテストを実施するときは、IRTによって分析すると受験者の情報がより多く得られると結論づけている。そして、プレースメントの意思決定のためには、一つだけでなく、複数の方法を根拠にして分割点を決めることで、より妥当な結果を導くことになると述べている。さらにテスト開発者は、学生の授業成績やその伸びなどの長期にわたるデータを集めて、習熟度別クラス決定の検証をすることが重要であると指摘している。

斉田（2008）は、DIALANG を使って日本の大学 1 年生の CEFR におけるレベルを調査した。彼ら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 レベルであった。ここで、Structure と Vocabulary を「言語知識」とし、Listening, Reading, Writing を「言語運用能力」とすると、被験者たちの「言語知識」は「言語運用能力」より 1～２レベル高い傾向にあると報告している。

教員作成によるプレースメントテスト

教員がそのプログラムに対応したプレースメントテストを作成するのは、骨の折れる仕事であるが、利点も多くある。まず、受験者のレベルはその教育プログラムによって異なるので、テストの難易度はそのプログラムに合わせて作成されるべきである。従って、そのプログラムの学生の習熟度を良く知る教員がテストを作成することが、そのテストの難易度を、ターゲットとする学生に最も適切に合わせることができる方法だと指摘している。また教員作成のプレースメントテストは、そのプログラムで学生たちが学ぶ内容に適した問題で構成されるべきで（Brown, 1990, 1995b, 2004; Brown & Hudson, 2002; Yamashita, 1995）、そのプログラムの授業到達目標に適合し、日程、テストの長さなどのプログラムの事情に合わせたテストにすることもでき、さらには、市販のテストを購入するコストが節約できるという利点もある（久保田, 2002; 斉田、小林＆野口, 2009）。

プレースメントテストの妥当性

量的研究に重点をおいたアプローチ

Brown (1989, 1990, 1992, 1995a, 1995b, 1996)は、より妥当性の高いプレースメントテストを開発するにあたって、まずテストは受験者の能力に合っていること、そしてその言語プログラムのカリキュラムの特徴や、教育目標に適応した内容をテスト問題に反映していなければならないと主張している。　また、量的分析に着目し、プレースメントテスト開発の手順としては、必ず予備テストを実施して結果を分析し、あらかじめ分割点（カットスコア）を設定しておくことの必要性も指摘している。これは、その能力が分割点に近い受験者たちの能力を最も正確に測定できるように、分割点の受験者能力に近い項目困難度の問題をできるだけ多く出題するべきであるからだ。

また、Fulcher (1999)は、イギリスの大学１年生に対して、エッセイ、文法 10 問、読解問題8問を45分で行うプレースメントテストを実施し、結果をラッシュモデルで分析した。このテストは上級と初級の２つのレベルに分けるプレースメントテストで、分割点はちょうど標準値（θ=０）に近く設定された。そのため、項目数が少ないこともあり、このプレースンメントテストの信頼性は低くなってしまったが、このテストのテスト情報関数（test information function: TIF）の値が、受験者の能力水準が-1.00 から 0.50 で最も大きくなっており、分割点に近い能力値をもつ受験者の能力を、より正確に測定できたと報告している。

また、Kondo-Brown and Brown (2000)は３パラメタＩＲＴモデルを使って、ハワイ大学のプレースメントテストをより妥当性の高いものにするように試みた。プレースメントテストのサブテストである、リスニング、文法、語彙、作文テストをそれぞれのセクションごとに分けて IRT により分析し、弁別力の低い項目を削除したあと、さらに吟味して適切ではない項目を削除し、難易度の低い項目から高い項目に並べた。このプロセスの結果、もともと70問あったテストを50問にし、テストを短くすることが可能になったにも関わらず、テスト信頼性はほぼ同じ水準を維持することが確認できた。

質的研究に重点をおいたアプローチ

Wall, Clapham, and Alderson (1994) は、イギリスでプレースメントテストの表面的妥当性、内容的妥当性、構成概念妥当性、併存的妥当性について調査した。この研究では 30 人の大学生が、文法、作文、聴解、読解のサブセクションがあるプレースメントテストを受験し、また、自己評価、チューター評価、教員評価も同時に実施して、総合的な観点からの妥当性の検証を行った。その結果、満足のできる妥当性が確認できたのは、表面的妥当性、内容的妥当性、構成概念妥当性で、併存的妥当性については、プレースメントテストのスコアと自己評価、チューター評価との相対関係は低く、妥当性の確認はできなかった。

Green & Weir (2004)は、イギリス、ギリシャ、日本で学ぶ1070人のEFL（English as a Foreign Language）の学生が Global Placement Test (GPT)という文法のプレースメントテストを受験した結果を IRT で分析した。彼らの注目点は、GPT が、どのくらいその英語プログラムに

入る受験者の文法能力を予測診断できるのか？というテストの内容妥当性に係る部分である。つまり、プレースメントテストによって学習者の英語の知識、能力、スキルの習熟度の違いを識別することができているのか？また、テストタスクはその授業コースごとに習得するべき英語の知識、能力、スキルの内容を反映しているのか？という疑問を解明しようとするものであった。Heaton (1988) は、プレースメントテストとは、受験者の得点をできるだけ大きくバラつかせるだけでなく、受験者に最も良くあっている授業内容や教材を調査する役割も果たすと主張している。また、過去完了形を学ぼうとしている受験者たちが、過去完了を学ぶ前に、過去形や現在完了が習得できているか確認する意味もあると指摘している。しかしながら、Green＆Weir（2004）では、項目困難度は出題された文法問題の、言語的知識としての文法の難易度よりも、問題のタイプによってより影響を受けるという結果となり、因習的に行われてきたプレースメントテストの実施の前提に疑問を投げかけることになった。これは、テストの一問ずつの内容や特徴によって、受験者が過去に習得してきた言語能力のレベルを推定することの難しさの片鱗を知る研究内容である。

コンピュータとプレースメントテスト

コンピュータテクノロジーの進歩・普及とともに、この約 15 年間、コンピュータベースのプレースメントテストが急激に普及しているが、1990 年代後半頃から、コンピュータベースのプレースメントテストと紙＆鉛筆のテスト（筆記テスト）の違いを調査する研究が発表されはじめていた。その中で Fulcher（1999）は、イギリスで学ぶ学習者たちを対象に 80 問の多肢選択式のテストを、オンラインによるコンピュータベースのテスト（CBT）と筆記テストの２つの方法で受験してもらい、その違いを比較した。CBT のほうが高いものの、信頼性は両方ともほぼ同じで、２つの方法でのテスト間の相関係数も高かった。また、コンピュータを使用する頻度の高い受験者と低い受験者の間で、そのテスト結果には有意差は認められなかった。この研究では、習熟度別クラス編成の意思決定について CBT のほうが、筆記テストと比べやや優位なことはあっても劣ることはなく、CBT が受験者にもたらす不利益は何もないという結論になった。

残された課題

従来のプレースメントテストの研究は、その難易度を対象とする学習者に合わせ、より正確な分割点を求め、効率の良く、能力が均等なグループを作ることに注目するものが多かった。しかし、近年、プレースメントテストの内容によって学習者の英語の知識、能力、

スキルの習熟度の違いを識別することに注目し、テストの内容的妥当性を追求することも多くなってきた。例えば文法によるプレースメントテストでは、受験者たちが、過去完了形を学ぶ前に、過去形や現在完了が習得できているかを調査し、さらに将来のパフォーマンスも予測診断するような役割をテストが果たしているのか検証するということである。しかし、テストスコアは受験者のさまざまな言語能力を複雑に結集したものなので、テストの一問ずつの内容や特徴によって、受験者が過去に習得してきたひとつひとつの言語能力を切り離して習熟度レベルの推定をすることは容易ではない。しかしながら当分野の最も大きな課題として、取り組むべき重要な問題だと思う。さらに、妥当性の高いコンピュータ版のプレースメントテストと、ライティングやスピーキングなどの発信能力を測るプレースメントテストを開発するための研究も今後の残された課題だと言える。

非常に多くの受験者の言語能力を、その英語プログラムが開始されるまでの短期間に測定して、習熟度別クラスに分けることが、ほとんどのプレースメントテストに付随する宿命である。このプロセスのなかで、どういう受験者を、何のために、どのように分けるのか？ということをあらかじめ決め､それに応じたテストを作成しておかなければならない。また、プレースメントテストは、受験者たちにとって大切な意思決定をもたらすことが多い。その規準設定には慎重を 期し、テストの妥当性検証を継続的に実施することは、テスト開発者、そのテストに係る関係者たちの義務である（Messick, 1989）。このような複雑で慎重な作業を要する工程を短期間で達成し、できるだけ正確に受験者たちを習熟度別クラスに分けるためには、コンピュータ版のプレースメントテストが今後ますます必要になってくると考えられる。Brown (1997) や Fulture（1999）のコンピュータによる言語テストの研究は 13～15 年前のものであるが､コンピュータを使用する頻度の高い受験者と低い受験者の間で、そのテスト結果には有意差は認められなかった。また、習熟度別クラス編成の意思決定についても、CBT のほうが筆記テストと比べやや優位なことはあっても劣ることはなく、CBT が受験者にもたらす不利益は何もないという結論になった。これらの研究をした頃から現在に至るまでに、想像を絶するほどの勢いでコンピュータ技術は進歩し、驚くほど速くすべての人に普及した。今では、コンピュータを使用する頻度の高い受験者と低い受験者の間で不公平を心配する必要さえなくなったと言えよう。

また、斉田（2008）は、日本人大学一年生に DIALANG というコンピュータベースのテストを実施して、その結果に基づき受験した学生が CEFR のどのレベルに入るのか判定する実験を行った。彼らのほとんどは、海外に渡航したことがない一般的な日本の大学 1 年生たちで、彼らの CEFR のレベルは､「言語知識」である文法や語彙のほうが､「言語運用能力」であるリスニング、リーディング、ライティング能力より、1～2 レベル高い傾向にあるという判定結果となった。以前から言われている日本人の英語能力の特徴である「知識偏重型」の傾向が未だに継続していることを示す一例だと考えられる。最近では、コミュニケーション能力、その中でもライティングやスピーキングなどの発信能力の養成に重

点を置く英語教育プログラムが増えてきた。そのようなプログラムでは、文法や語彙の問題より、ライティングやスピーキング能力を測る問題を多く出題することが必要になる。より多くの学生の言語発信能力を一度に測定することができるようなテストの開発を急がねばならない。

参考文献

Bradshaw, J. (1990). Test-takers' reactions to a placement test. *Language Testing, 7*(1), 13-30.

Brown, J. D. (1989). Improving ESL placement tests using two perspectives. *TESOL Quarterly, 23,* 65-83.

Brown, J. D. (1990). Where do tests fit into language programs? *JALT Journal, 12,* 121-140.

Brown, J. D. (1992). Classroom-centered language testing. *TESOL Journal, 1*(4), 12-15.

Brown, J. D. (1995a). Developing norm-referenced language tests for program-level decision making. In J. D. Brown & S. O. Yamashita (Eds.), *Language testing in Japan* (pp.12-19). Tokyo: JALT.

Brown, J. D. (1995b). *The elements of language curriculum.* Boston, MA: Heinle & Heinle.

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

Brown, J. D. (1997). Computers in language testing: Present research and some future directions. *Language Learning & Technology, 1,* 44-59.

Brown, J. D. (2004). Grade inflation, standardized tests, and the case for on-campus language testing. In D. Douglas (Ed.), *English language testing in U.S. colleges and universities* (pp.37-56). Washington DC: NAFSA.

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

Culligan, B., & Gorsuch, G. (1999). Using a commercially produced proficiency test in a one-year EFL curriculum in Japan for placement purposes. *JALT Journal, 21*(1), 7-28.

Culligan, B., & Gorsuch, G. (2000). Using item response theory to refine placement decisions. *JALT Journal, 22*(2), 315-325.

Fulcher, G. (1997). An English language placement test: Issues in reliability and validity. *Language Testing, 14,* 113-138.

Fulcher, G. (1999). Computerizing an English language placement test. *ELT Journal 53(4).* 289-299.

Green, A., & Weir. C. (2004). Can placement tests inform instructional decisions? *Language Testing, 21(*4), 467-494.

Heaton, J.B. (1988). *Writing English language tests.* Harlow: Longman.

Kondo-Brown, K., & Brown, J. D. (2000). *The Japanese Placement Tests at the University of Hawai'i: Applying item response theory.* Retrieved December 6, 2000 from

http://nflrc.hawaii.edu/NetWorks/NW20/default.htm/.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd edition, pp. 13-103). Washington, DC: American Council on Education & National Council on Measurement in Education.

Sugimori, M. (2002, September). Standardising the testing of English language proficiency at Risumeikan University with the GTELP. JACET 41st Annual Convention. Tokyo.

Yamashita, S. O. (1995). Monitoring student placement: A test-retest comparison. In J. D. Brown & S. O. Yamashita (Eds.), *Language testing in Japan* (pp. 48-56). Tokyo: JALT.

Wall, D., Clapham, C., & Alderson, J. C. (1994). Evaluating a placement test. *Language Testing, 11,* 321-344

久保田章 (2002) カリキュラム改革と英語検定試験. 川崎晶子編　筑波大学の英語教育 Institute of Modern Languages and Cultures University of Tsukuba: 『筑波大学現代語文化学系』　59 号. 150-154.

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.

斉田智里, 小林邦彦., & 野口裕之 (2009). 外部試験を活用した大学英語カリキュラム改革. 『日本テスト学会誌』, 5(1), 96-105.

杉森幹彦 (2003). 英語統一テスト・習熟度別クラス編成・到達目標の設定および測定に関する実態調査. 『政策科学』10 (3)、3-26.

4.3. The Use of the Rasch Model in Standard Setting

Ken Norizuki

Abstract

The Rasch one-parameter model is a sole IRT model, which can make direct comparisons between abilities and difficulties on the same latent interval scale. This unique quality, often criticized for its lack of sophistication, becomes a clear advantage over other measurement models when it comes to setting cut points on the test scale. The present report aims to review some major Rasch-based standard setting or maintaining approaches which have been conducted on its own or in conjunction with other techniques which are designed to classify students into several distinct proficiency categories. The report first exemplifies some clues as to how the model can facilitate complex and unstable standard setting procedures by integrating judgments and test data. After combinations of some non-IRT-based standard setting approaches (e.g., Angoff and borderline) with Rasch-based analysis (e.g., item-mapping and bookmark) are discussed, many-facet Rasch analysis is described, regarding how variables other than difficulty and ability facets can be statistically controlled in the analysis of a performance test data. Then two hybrid models called mixed Rasch model and Rasch formulation of Thurstone's paired comparison underlying the rank-ordering method are featured. The report ends with an overview of advantages and limitations of various Rasch-based analytic methods.

4．3． 課題規準設定におけるラッシュモデルの有用性

法月 健

あるテストの分割点を事前に設定するには、まず、一定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、厳密な規準設定を行っていない high-stakes テストが多く、Bramley (2010)は、プリテストが実施できない状況にある英国の GCE や GCSE 試験において、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。このような問題は我が国においても看過できない状況にある。

テスト（項目）に依存しない受験者の能力推定、受験者集団に依存しないテスト（項目）の難易度尺度化は、項目応答理論（IRT）の特性であるが、受験者能力と項目難易度を上下限のない共通の間隔尺度上で比較できるラッシュモデルについて、Bramley (2010) は、「ラッシュモデルに基づくシステムは最善で（恐らく唯一の）規準維持の状況を説明する方法である」(p.3)としている。

MCC を仮定する分析法とラッシュモデルの応用

合格・目標水準に達している最低限の能力を有している受験者 (minimally competent candidates: MCC) を仮定する際に、審査員は最低必須能力を適切に概念化し、それを基に規準設定の項目難易度を正確に推定することが求められる。しかしながら難易度推定は困難を極め、Wang (2003) は、Angoff法において、個々の項目にMCCが正答する確率を審査員が推定することの心理的な負担は、ほとんど限界の域を超え、ラッシュモデルに基づく項目マッピング (item-mapping) 法を応用することが改善策につながると唱えている。

Angoff 法と異なり、項目マッピング法では、各項目の分割点となる正答確率を.50 に設定し、審査員（教員）は自分が指導する典型的な MCC が.50 の正答確率に達しているか否かを判断し、項目の難易度と審査員が合意する MCC の能力水準とが一致する尺度地点をテスト全体の分割点とする。この方法では、すべての項目とその難易度がヒストグラム（項目マップ）で表示される。

Wang (2003) は、43～79 項目の4つの資格試験の分割点を Angoff 法と項目マッピング法で一部重複する6～13 人の審査員に審査させ、後者の方法のほうがより首尾一貫した規準設定を行うことができ、前者では実際の受験者の能力水準よりも高い規準設定をする傾向を確認した。

一方、MacCann & Stanley (2006) は、ラッシュモデルを使った Angoff 法の3種類の修正法（AR①～AR③）について説明している。

AR①では、まず審査員が推定する MCC の個々の項目への正答確率とラッシュモデルの正答確

率が一致する地点の能力推定値(素点)をマークする。たとえば、項目1に MCC が正答する確率を 90%と審査員が考えたならば、ラッシュモデルの正答確率で.90 になる能力推定値の素点を特定できる。このようなプロセスをいくつかの項目で繰り返し、ほぼ同じような素点に収束した場合、その素点がテストの分割点になるか、審査員が分割点を決定する参考データとして活用される。

AR②では、審査員が各項目に付与した推定正答確率に合致するラッシュ能力ロジット値を求め、ロジット値に相当する素点を各項目の分割点とする。目的に応じて各項目の分割点の中央値(median) や 25 パーセンタイルを審査員が最終的なテストの分割点として選ぶことになる。

AR③では、審査員が推定した各項目の推定正答確率とラッシュモデルで算出された正答確率が比較される。この方法は、個別に審査を行った最初の推定の後に審査員たちが最初の推定について議論を交わすため、2つの正答確率の差が大きかった項目についてはなぜ大きな差異が生じたのか吟味して、再度審査を行うときの参考として活用するのに有効な方法である。

OSS 法と境界グループ法とラッシュモデルの応用

Stone, Beltyukova & Fox (2008) は、多相 (multi-facet)ラッシュモデルを応用した OSS (objective standard setting)法と境界グループ法について、医療関係の試験を対象にして、比較分析を行っている。OSS 法では各項目が、「MCC が修得している必要不可欠な内容を表している」か否か、「特殊で重要な内容であるが、必要不可欠ではない」か否か、Yes/No で審査員が判断を下す。前者にYes、後者にNo と判定された項目の平均が分割点として定義される。境界グループ法では合格／不合格の可能性が4段階で評定される。分析の結果、OSS 法と境界グループ法ともに多相ラッシュモデルにおいて効果的に機能し、従来の古典的テスト理論に基づくモデルに比べてテスト過程の信頼性や合否の規準設定を平易にするモデルであると結論付けている。

Stone (2008) も、多相ラッシュモデルに基づく OSS 法と境界グループ法を高校生の作文の分析に運用して、その有効性と将来の研究の発展性に言及しているが、OSS 法の観点から多相ラッシュモデルの規準設定への応用を論じた Abu Kassim & Bond (2006)は、その利点として、審査員の評定の内部均質性の欠如や難易を誤って判定された項目の識別が容易にできるようになり、各分割点設定の際の審査員の評定に関する測定誤差を計算できることを挙げている。

Bookmark 法とラッシュモデルの応用

これまでに述べてきた規準設定法と異なり、Bookmark 法は最初から IRT を活用するように設計されている (MacCann, 2009)。また、この方法は 1990 年代に開発の進んだ前述の項目マッピング法の発展に起因するものだと考えられている (Cizek & Bunch, 2007)。

MacCann & Stanley (2006) によると、Bookmark 法には、2つの重要な概念があるとされるが、一つは分割点付近の能力の受験者が.67（3回のうち2回）の応答確率 (response probability: RP) で正解できる項目を見つけることで、もう一つはブックマークした難易度の位置（bookmark difficulty location: BDL）であり、ある項目に正答する確率が RP に相当するときに必要な受験者の能力水準を表している。

大友 (2008) によると、審査員には、順位付きテスト冊子項目 (ordered item booklets: OIB) が配布される。冊子には易しいものから難しいものへ順番に配置された項目内容と各項目の統計情報が掲載されている。MacCann & Stanley (2006) に紹介される OIB の例では、各項目の正答率とその項目にかろうじて正解する受験者の能力に相当する分割点が記されているが、審査員は冊子を検討して、分割点水準の受験者が 3 分の2の確率で正解する項目（あるいはそれ以上の正答確率の項目と未満の項目の間）のところにブックマークを置く。

特筆すべきは、RP= .5 の時、1パラメターのラッシュモデルでは BDL の値は項目難易度に一致する。これは、情報表示モードは異なるが、項目マッピング法の場合と同じである。また、弁別力パラメターが加わる 2 パラメター以上の IRT では、BDL は、必ずしも項目難易度と同じ順番にならない。能力の高い受験者にとっては弁別力の低い項目は他の項目に比べて相対的に難しくなり、能力の低い受験者にとってはその逆の現象が生じる。

MacCann (2009) は、応答選択型項目 (selected-response items)の場合、.5 の RP を Bookmark 法に適用することで OIB の解釈が容易になり、審査員の手間の軽減につながると主張し、項目マッピング法と OIB の概念の融合を唱えている。一方、応答構築型項目 (constructed-response items)の場合は、RP を変えることで項目の難易度が変わる可能性もあり、分割点の設定への影響について、今後研究を進めていく必要があることが指摘されている。

ラッシュモデルの応用

項目難易度と受験者能力を同じ潜在尺度上で比較できるラッシュモデルは、CTT や他の IRT モデルにない特性を有していて、その有用性が規準設定への応用において期待される。以降の議論では、以下の2つの観点から、ラッシュモデルの有用性を検討していくこととする。

（1）能力と項目難易度以外に審査に影響する主要なファセット（相）を統計的に調整することで、客観性を増すことは可能か
（2）順序尺度と間隔尺度の概念を融合する、ラッシュモデルを応用したモデルはないか

（1）については多相ラッシュモデル (many-facet Rasch model: MFRM)、（2）については混合ラッシュモデル (mixed [or mixture] Rasch model: MRM) と順位付け法 (rank-ordering method) を

論じる。

多相ラッシュモデル（MFRM）

評定者が採点するスピーキングやライティングテストが規準設定や規準維持に用いられる場合、受験者に割り当てられた評定者が他の評定者よりも厳しい評価を下すこともあれば、甘い評価をすることもある。つまり、そのままの評定値を基に規準設定を行うと、合格すべき受験者が不合格と判定されたり、本来の自分の習熟度よりも低いレベルのクラスに配置されたりすることもある。反対に、合格の資質のない受験者が合格となったり、本来の自分の習熟度よりも高いクラスが指定されることもある。

このような問題を未然に防ぐため、評定者トレーニングを行うことが古典的な対策法として考えられるが、どのようなトレーニングを行っても、評定者の厳格性・寛容性（severity / leniency）の差異を完全に解消することは期待できないとされる（Eckes, 2009 & 2011）。

近年、上記のような規準設定の問題点を克服する方法として、MFRM の有用性が活発に議論されている (Kozaki, 2004&2010; Eckes, 2009&2011; Kecker & Eckes, 2010 等)。

Eckes (2009&2011)は、外国語としてのドイツ語の high-stakes（利害関係の大きい）テスト、TestDaF のライティング部門の分析の中で MFRM を活用することで、上記のような評定者の寛容性・厳格性の問題だけでなく、経験の浅い評定者等が他の評定者の評定と隔たりが生じることを避けて、尺度の中央付近の評定に偏る中心的傾向 (central tendency) の事象を特定化し、効果的な統計的調整を行うことが可能であることを証明している。

Kaftandjieva (2004) や Thomas & Kantarcioglu (2009) は、クロンバックαのような古典的な評定者間信頼性は、数値が 1.00 を示しても評定者間一致が0の可能性もあるため、規準設定の枠組みの中では、評定者の一貫性 (consistency) と一致性 (agreement) が報告され、様々なファセットを基に算出されたスコアを見ることができる MFRM の有用性を指摘している。

Eckes (2011) は、評定者間の一致度が談合 (collusion) 等の要因によって過剰に高まることと、モデル化されていない誤差要因によって極端に低くなる問題点を指摘して、このような問題を解決するために開発された Rasch-Kappa index と呼ばれる統計指標を紹介している。

混合ラッシュモデル（MRM）

通常のラッシュモデルの分析では、「単一の構成概念が階層的な連続体を形成する項目の基盤となる」（Bond & Fox, 2007: 314）ことを仮定する一元性 (unidimensionality) が分析の前提条件となる。しかしながら、テストの分量が多いために、後半の問題が解けない受験者に不利な影響

を与えたり、それらの項目難易度が正確に推定できなくなる場合や、特定の社会背景に属する受験者グループに有利あるいは不利に働く項目が存在する場合は、一元性を仮定した分析が効果的に機能しないこともある。

このような心理測定的な制約に対応するため、MRM を追究した様々な研究が行われてきた (Rost & Langeheine, 1994; Cohen, Wollack, Bolt & Mroch, 2002; Jiao, Lissitz, Macready, Wang & Liang, 2011; Lee & Chen, 2011; Templin & Jiao, 2012 等)。

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 (Templin & Jiao, 2012: 387, 379) モデルである。LCA では、ある潜在クラス内の受験者がある項目に正答する確率は同一と見なされる。一方、MRM では、同一潜在クラス内の受験者たちの能力推定値は通常のラッシュモデルと同様に、項目難易度と同一の間隔尺度上に表現されるものの、異なるクラス間では、項目難易度パラメターは異なる数値を示し、相対的な難易度順位が異なることもある。

Jiao et al. (2011) は、大規模な言語熟達度テストを bookmark 法によって規準設定を行った結果に基づく1万人のシミュレーションデータを使って、MRM の複数の適合度指数を比較しながら、政策決定者が設定した熟達度水準の数の妥当性の検証を行っている。

Templin & Jiao (2012) は、境界グループ法や対照グループ法の審査結果を基にモデルを構築する「適応型 MRM」は、モデルがデータによく適合している場合は、Jiao et al. (2011) の「純粋 MRM」よりも正確な規準設定が期待できると主張している。

順位付け法

Bramley (2005) や Bramley & Black (2008) 等が提唱する「順位付け法」は、近年、新たな規準設定法として活発に議論されるようなった。Bramley 等 がこの方法に着目した背景には、共通項目と共通受験者集団を有さない異なる年度の GCSE や AS 等の試験を比較し、「規準維持」を図るのに有効な方法だとの判断に基づく。

「順位付け法」は、人間は絶対的な判断よりも相対的なほうがはるかに優れている、という心理学の概念に支えられたものであり、統計的にはサーストンの一対比較法に基づくラッシュ分析 (Rasch formulation of Thurstone's paired comparison) が使用されている。

Bramley (2005) は、2003 年と 2004 年に実施されたイングランドの全国カリキュラムテストの読解問題(短答式の応答構築型項目)を 12 人の専門家審査員が評価した結果を論じている。

各審査員は、得点層別に分けられた4セットの答案用紙を、得点を削除した後に渡されるが、各年度から5紙ずつ答案が含まれている各セットの合計 10 種類の答案用紙に、一番良くできているものからそうでないものまで、できるだけ異なる順位をつけるように指示される。重複して順位付けさ

れた答案データを基に、各年度の答案の得点 (mark) と審査員の評定のロジット値 (measure) から 2 本の最適合線 (best fit line) を描出し、前年度の分割点に相当するロジット値を見出すことで、今年度の分割点をどこに設定すれば規準が維持できるかが、わかるとされている。

Bramley (2005) の分析では、順位付けと実際の得点の相関も高く、順位付け法が有効な手法であることが確認されたが、Curcin, Black & Bramley (2009)の 2 種類の多肢選択式資格試験の分析では、項目難易度の順位付けと実際の難易度の相関はテストによって変動し、全般に低かった。Bramley (2010) は、能力よりも項目難易度の審査のほうが複雑だとし、その一つの理由として、審査員が能力に比べて項目難易度の審査に慣れていないためではないかと推測している。

残された課題

MacCann & Stanley (2006) の提唱した Angoff 法の修正法は、Angoff 法が最も普及した古典的規準設定法であることを考えると、実践的な価値の高い手法と言える。しかし規準設定に限られた日数しかかけられない状況においては、あまり現実的な選択肢にはならないだろう。

項目マッピング法や Bookmark 法が Angoff 法を補完する手法であることを、Wang (2003)や MacCann (2009) の議論を通じて確認したが、Wang (2003)が指摘するように、RP の値を変えることで、どのように分割点の設定に影響が出るのかについて研究することが必要である。RP については、MacCann (2009)も応答選択型項目では、ラッシュ以外の IRT モデルで、応答構築型項目 (constructed-response items) の場合は、ラッシュモデルにおいても値を変えることで、項目の配置に影響が出る可能性があることを問題点として指摘している。

OSS 法は、評定者の労力を軽減し、客観的に理解しやすい評価を導くが、Abu Kassim & Bond (2006) によると、元々、応答選択型項目 のために開発されたため、応答構築型が混在しているテストに応用することが難しく、評定者／審査員内誤差 (intra-rater variability) を規準設定の中でどのように調整するかが十分に研究されていないとされる。前者は他の手法にも共通する問題であるが、後者については Eckes (2011)が提唱する審査員内誤差を調整する計算式を使った研究が実践されることが望まれる。

Kecker & Eckes (2010) は、MFRM において、トレーニングを修了した審査員は独立した評定を下すことができる専門家と見なされるため、規準設定の審査員のトレーニングの前にはフィードバックは行うべきではないと主張している。これは、客観的なデータを見せると、審査員が人間的な審査プロセスを省略して同一の判定に収束し、項目応答理論の局所的独立性 (local independence) の仮定に反して、有効な分析を妨げる可能性があることを意味している。

MRM においても、審査員の規準設定の前に、研究者が探索的に潜在クラス数を求めてテストデータを分析することが仮にあったとしても、その結果を審査員が評定を下す前にフィードバックすることは、同様の理由で避けるべきであるかもしれない。

Kecker & Eckes (2010) の研究では、審査員の相互依存性 (rater dependence) の度合はトレーニング期間中は過度に高かったものの、規準設定の段階ではほとんど問題にならないくらいに小さくなっていたことが指摘されているが、現実的に十分なトレーニングが実施できない場合や適正な能力を持った審査員の数が確保できない状況下での研究も行われる必要がある。

順位付け法は、現在最も注目を集めている新しい規準設定法の一つであるが、Bramley 等の提唱者たちも認めている通り、実践的運用における最大の問題点は、複雑な答案用紙の配置デザインを組むプロセスを簡略化するアルゴリズムやマーキングした答案用紙のマークや書き込みをデジタル加工するシステムの開発である (Bramley & Black, 2008)。また、現時点では、統合的なパフォーマンス評価の規準設定法のイメージが強いが、多肢選択問題や多肢選択問題と短答形式の応答構築型問題にも妥当な手法と言えるのか、研究が行われる必要がある。

様々な主観的な審査に基づく規準測定法は、それぞれ独自の利点と欠点を有しており、単純に絶対的な優劣を論じることができないが、現実の言語テスト分析の中でモデル化されたデータと比較検証を発展的に継続していくことで、 将来に向けて、より正確で実践的な規準測定の手続き法を確立していくことが求められている。

参考文献

Abu Kassim, N.L., & Bond, T.G. (2006). *Using the many-facets Rasch model to resolve standard setting issues.* A paper presented at the 13th International Objective Measurement Workshop, University of California, Berkley, CA (April 5, 2006).

Bond, T.G., Fox, C.M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mhwah, Nj: Erlbaum.

Bramley, T. (2005). A rank-ordering method for equating tests by expert judgment. *Journal of Applied Measurement, 6,* 202-223. Retrieved from http://www.aliquote.org/pub/Bramley_2005.pdf

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

Bramley, T. & Black, B. (2008). *Maintaining performance standards: aligning raw score scales on different tests via a latent trait created by rank-ordering examinees' work.* A paper presented at the Third International Rasch Measurement Rasch Measurement conference, University of Western Austraila, Perth (January, 2008). Retrieved from http://www.cambridgeassessment.org.uk/ca/digitalAssets/171143_TB_BB_rank_order_Perth08.pdf

Cizek, G.J., & Bunch, M.B. (2007). *Standard setting: A practitioner's guide to establishing and evaluating performance standards on tests.* Thousand Oaks, CA: Sage Publications.

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

Curcin, M., Black, B., & Bradley, T. (2009). *Standard maintaining by expert judgment on multiple-choice tests: a new use for the rank-ordering method.* A paper presented at the British Educational Research Association annual conference,

University of Manchester, September 2009. Retrieved from http://www.cambridgeassessment.org.uk/ca/digitalAssets/184999_BERA_paper_Curcin_Black_and_Bramley.pdf

Eckes, T. (2009) Many-facet Rasch measurement. In S. Takala (Ed.), *Reference supplement to the manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment* (Section H). Strasburg, France: Council of Europe/Language Policy Division. Retrieved from http://www.coe.int/t/dg4/Linguistic/CEF-refSupp-SectionH.pdf

Eckes, T. (2011). *Introduction to many-facet Rasch measurement*. Frankfurt, Germany: Peter Lang.

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

Kaftandjieva, F. (2004). Standard setting. In . Takala (Ed.), *Reference supplement to the manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment* (Section H). Strasburg, France: Council of Europe/ Language Policy Division. Retrieved from http://www.coe.int/t/dg4/linguistic/Proceedings_CITO_EN.pdf

Kecker, G., & Eckes, T. (2010). Putting the manual to the test: The TestDaF－CEFR linking project. In W. Martynuick (Ed.), *Aligning tests with the CEFR: Reflections on using the Council of Europe's draft manual* (pp.50-79). Cambridge, UK: Cambridge University Press.

Kozaki, Y. (2004). Using GENOVA and FACETS to set multiple standards on performance assessment for certification in medical translation from Japanese into English. *Language Testing, 21*, 1-27.

Kozaki, Y. (2010). An alternative decision-making procedure for performance assessments: Using multifaceted Rasch model to generate cut estimates. *Language Assessment Quarterly, 7*, 75-95.

Lee, Y-H., & Chen, H. (2011). A review of response-time analyses in educational testing.

Psychological Test and Assessment Modeling, 53, 359-379. Retrieved from http://www.psychologie-aktuell.com/fileadmin/download/ptam/3-2011_20110927/06_Lee.pdf

MacCann, R.G. (2009). Standard setting with dichotomous and constructed response items: Some Rasch model approaches. *Journal of Applied Measurement, 10(4)*, 438-454.

MacCann, R.G., & Stanley, G. (2006). The use of Rasch modeling to improve standard setting. *Practical Assessment, Research & Evaluation, 11(2)*, 1-17. Retrieved from http://pareonline.net/pdf/v11n2.pdf

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

Stone, G.E. (2008). Establishing criterion measures on graded essays using objective standard-setting for judge mediated examinations. Retrieved from http://www.ieia.com.mx/materialesreuniones/1aReunionInternacionaldeEvaluacion/PONENCIAS19Septiembre/ColoquiodeRasch/ConferenciaMagna-GregoryStone.pdf

Stone, G. E., Beltyukova, S., & Fox, M. (2008). Objective standard setting for judge-mediated examinations. *International Journal of Testing*, 8, 180-196.

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.

Thomas, C., & Kantarcioglu, E. (2009). Bilkent University School of English Language COPE CEFR Linking Project. In N. Figueras & J. Noijons (Eds.), *Linking to the CEFR levels: Research perspectives* (pp.119-124). Arnhem, The Netherlands: Cito/EALTA. Retrieved from http://www.coe.int/t/dg4/linguistic/Proceedings_CITO_EN.pdf

Wang, N. (2003). Use of the Rasch IRT model in standard setting: An item-mapping method. *Journal of Educational Measurement, 40,* 231-253.

大友賢二 (2008).「Bookmark Method 再考」 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.

4.4. The Use of the Latent Rank Theory in Standard Setting – Comparison with Item Response Theory and Classical Test Theory

Ken Norizuki

Abstract

The Latent Rank Theory (LRT), also known as the Neural Test Theory, is a new test analysis model based on the latent ordinal scale. Although very little has been revealed about the practical utility of the model in standard setting operations, it appears that the process of assigning students to two or more latent ranks has a natural resemblance to that of setting one or more cut scores. The purpose of the present report is twofold: to compare similarities and differences between LRT ordinal and IRT or CTT interval scales; and to highlight some important features of the former which may serve as a new promising standard setting tool. The report starts with recent research findings about comparisons between LRT and IRT and/or CTT models. These findings are then corroborated by a similar hypothetical analysis, employing a real listening comprehension test data, as designed to explore solutions to potential problems. The report moves on to discuss different lines of LRT research which may be of some relevance to standard setting practices. It concludes by offering some perspectives on future research and applications.

4．4．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

法月　健

テストの分割点を仮に素点の 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。一方、受験者の能力に依存しない項目及びテスト難易度推定が可能な項目応答理論（IRT）についても、連続する間隔尺度上の値を基に判定を行うことに共通の問題が認識される。このような問題を解決するために、Shojima (2007) は、Neural Test Theory（ニューラルテスト理論、NTT）を提唱した。近年 Latent Rank Theory（潜在ランク理論、LRT）と呼ばれることが多くなっているため（荘島 2011a）、以降、LRT と統一標記する本理論は、一言でまとめると「学力を順位尺度上で段階評価するためのテスト理論」である（荘島, n.d.; 小泉・飯村 2010）。

小泉・飯村（2010）は、荘島 (n.d.) に基づき、LRT が順序尺度上に能力を表す理由を、統計学的な方法論的観点と教育社会学的観点から説明している。前者の理由としては、テストの解像度（信頼性）は間隔尺度上で表せるほど高くなく、10 段階程度の段階評価が現実的であるとする。後者としては、通知票・調査書などが段階評価で行われるため、初めから順序尺度上で推定するほうが分割点の設定や Can-do chart のような学力進度表の作成が容易になることを挙げている。

本節では、LRT の規準設定における有用性に関する主要な特徴を、古典的テスト理論 (CTT) や項目応答理論 (IRT) との比較も考慮しながら Test Reference Profile (テスト参照プロファイル、TRP)、Item Reference Profile (項目参照プロファイル、IRP)、Rank Membership Profile（ランク・メンバーシップ・プロファイル、RMP）の観点から論じ、LRT を使用した規準設定の方向性に示唆を与える近年の言語テストの研究について、考察することとする。

段階評価による有用性

山川・荘島 (2007)は、正規分布状の IRT による連続尺度と比較した場合、LRT は、得点密度が高い受験者をより細かく順序付けることができるとしている。小泉・飯村 (2010)は、CTT と IRT (ラッシュモデリング) とLRTを、受験者集団を能力別に3グループと5グループに分けて比較しているが、機械的に同数（但し、同得点者がいる場合は同グループに入るように調整）で分けている CTT や IRT と異なり、LRT（分布仮定なし）では5ランクのときに特にグループ間の人数差が顕著に表れ、同様に CTT と IRT ではほぼ一致するグループ分けが、IRT と LRT の一致度（カッパ係数）においては、3つのときは.90、5つでは.74 に下がっていることを示している。このことは、スコア分布の密度

が高いところでは狭い間隔で、低いところでは広い間隔でランクを設定する LRT の特性であると考えることができる（山川・荘島, 2007）。

小泉・飯村(2010)の議論から、分割点設定において、LRT は IRT や CTT とは異なる結果を生じる可能性が高いことが示唆されるが、一方で、LRT の RMP（各受験者がそれぞれの潜在ランクに配置される可能性がどの程度あるかを示す確率）と CTT の正答率と IRT（ラッシュモデル）の能力推定値の間に高い相関があったことも示されている。

小泉・飯村 (2010) のような分析プロセスを明示的かつ詳細に解析した実践研究は多くなく、今後、「他の文脈でどの程度一般化できるか」についての「適用研究が必要」になるだろう (小泉・飯村, 2010: 105)。

本格的な適用研究の方向性を探るために、筆者の有する言語テストデータを用いて、以下、TRP（ある潜在ランクに属する受験者が取ると予測される期待得点）、IRP（ある潜在ランクに属する受験者がその項目に正答する確率）、RMP の観点から、規準設定の問題に関連する可能性の高い LRT の特徴を整理していくこととする。

表1は、Norizuki, Ito & Shimatani (2011)で使用したある英語標準模擬テストのリスニングセクション 100 問（225 人）のデータを分析した際の基本統計量を示している。テストは、サンプルの平均的受験者には易しく、得点のピークは、平均や中央値と比べても、得点尺度のかなり高いところに位置していることがわかる。

表2は、Exametrika Version 5.3 （荘島, 2011b）を使って同テスト結果を分析し、 TRP の結果を、3、5、10 グループ数別にまとめたものである。TRP とは、「各潜在ランクに所属する受験者が、テスト全体で、どの程度の得点（ここでは正答数）をとることができるかという期待値」（荘島, 2010: 91）を意味している。テスト全体では、表2で提示したいずれのグループ数においても、ランクが上がるにつれて、期待値が上がる弱順序配置条件を満たしているため、分析として妥当であると言える。RMP 分析の複数の適合指数からはランク数2が最もモデル的には「効率的」であることがわかったが、荘島氏は、適合度が相対的に低くても実際の目的に応じたランク数で分析することを推奨している (荘島, personal communication, 2012 年 3 月 27 日、荘島, 2010 & 2011a 参照)。

表3は上記の TOEIC テストの項目 89 と 90 の IRP の5潜在ランクによる分析結果をまとめたものである。グラフにすると、項目 89 はランクの上昇に応じて正答確率は緩やかな右肩上がりを示すが、項目 90 ではランク間でほとんど変化がなく、平坦である。

表1
テストの基本統計量

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
225	100	67.8	78.0	68.0	12.4	98	39	.891

表 2
ランク数 3,5, 10 の時のテスト参照プロフィル（TRP）

	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5	Rank 6	Rank 7	Rank 8	Rank 9	Rank 10
LRT1	55.87	66.43	77.30							
LRT2	54.40	58.69	66.13	73.98	78.65					
LRT3	52.94	54.70	57.43	60.57	63.77	67.31	71.19	74.92	77.96	79.94

表 3
項目 89、90 の項目参照プロファイル（IRP）

	項目参照プロファイル (IRP)					項目参照プロファイル (IRP) 指標					
	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5	Alpha	A	Beta	B	Gamma	C
Item 89	.31	.38	.50	.59	.62	2	.12	3	.50	.00	.00
Item 90	.29	.32	.34	.32	.33	1	.03	3	.34	.25	-.02

小泉・飯村 (2010) は、表3の6つの指標の意味についても、詳しく説明している。Beta は、受験者の正答率が .5 に最も近い潜在ランクを示している。両項目とも3であるが、その時の受験者の正答確率の B には、かなり大きな差があることがわかる。Alpha と A の値は、項目 89 はランク2から3にかけて、項目 89 はランク1から2にかけて、正答率が最も大きく変化し、それぞれ.12、.03 の差だったことを示している。Gamma は正答確率が単調増加していないペアの割合で、C は単調増加していないペアの差の和である。項目 89 はすべてのランク間で単調増加しているため、Gamma と C は0であるが、項目 90 は4ペア中 1 ペア（1/4=.25）で減少していて、その和は、（ランク2から3の減少差のみの）-.02 であった。

このように一見まったく共通点がないように見える項目 80 と 90 は、Beta の数値で一致し、その他の数値においてまったく異なっていたことがわかる。このような項目の特性は、分割点設定法とともに教育課程及び指導の改善に向けての設定通過率の観点（大友, 2008）からも議論する価値があるだろう。

表4は、同じ正答率(.73)で同じ潜在ランク（4）に属する3人の受験者の、異なる RMP パターンを示している。受験者 A は、ランク4に属する確率が 56%あるが、ランク5に属する可能性も 42%ある。一方、受験者 B は、ランク4に属する確率が 75%に達し、ランク5に属する可能性は 21%にとどまっている。受験者 C については、ランク4の可能性が 61%、ランク3の可能性も 34%に及んでいる。同じ正答率であっても学習の発達状況が完全に一致しているとは言えない。RMP は、到達、未到達の領域を特定化し、教育診断情報のフィードバックに活用されることが期待される。

表 4 では、正答率と潜在ランクが一致する例を見たが、正答率が同じでもランクが全く異なる場合も多々見られた。上記の分析では、Exametrika の目標潜在ランク分布を「指定しない」にしたが、CTT の正答率やラッシュモデルの能力推定値を基に単純に等分に区切った結果とは全く異なり、潜在ランク数 10 では、ランク 1 が 100 点満点中 39 点から 63 点の受験者（225 名中 47 名）にまで及んだ。表 1 から、平均点がかなり高かったことが大きな要因であったと考えられるが、6 割以上の正答率でも最下位に分類されてしまうのは、通常の評価に慣れている学習者たちにとって、なかなか納得しがたいことであろう。

表4
受験者 A,B,C のランクメンバーシップ（RMP）

	正答率	潜在ランク	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5	
受験者 A	.74	4	.00	.00	.02	.56	.42	.
受験者 B	.74	4	.00	.00	.05	.75	.21	
受験者 C	.74	4	.00	.00	.34	.61	.04	

荘島氏は、ある全国的な習熟度調査テストを分析したとき、多くの学習者たちにとって易しい問題であったため、ほぼ半数近くが最高ランク（5）に位置づけられたが (Shojima, 2009 参照)、測定精

度が低い項目に多く正解している場合は、ランクの推定に影響しないため、ランクが低くなることもありえると述べている。(荘島, personal communication, 2012 年 3 月 13 日)。

LRT の順序尺度と間隔尺度との比較

上記の議論より、LRT の分割点設定が、CTT やラッシュモデルを使用する場合とは異なってくる可能性が示唆されたが、RMP の順序尺度が、根本的に間隔尺度とは全く性格を異とするものなのか、それとも設定条件を変えることや素点以外の他の間隔尺度と比較することで、対応関係を明確化できるのか、上記のデータを使って、検証することとする。

荘島 (n.d.) は、LRT の分析ソフトウェアである Exametrika で「事前分布」を選択すると、「正答数が高い受検者ほど、潜在ランクが高くなる傾向」があるとしている。そこで「事前分布」を指定したところ、ランク数 10、5では弱順序配置条件を満たすことができず、ランク数4でようやく有効な分析ができることがわかった。

次に、最初の分析で「指定しない」に設定した「目標潜在ランク分布」を、「一様分布」、「正規分布」にそれぞれ替えて、ランク数 10 の条件で、分析を行った。その結果、「指定しない」、「一様分布」、「正規分布」の順で、ランク1の比率は減り、「正規分布」ではランク1の受験者は 39 点から 44 点までの 4 名にとどまった。

どの分布様式を選択するかは、規準設定の目標にもよるが、素点をヒストグラムに表したところ、最頻値のピークは平均や中央値とは少しずれてはいたが、全体として正規分布に近い形状を示した。ちなみに素点と RMP のスピアマン順位相関は、「指定しない」で .947、「一様分布」で .981、「正規分布」で .976 で、いずれも高い数値を示した。

表5は、正答率とRMP(正規分布の場合)と1～3 パラメターIRT モデルによる能力推定値の順位相関係数を示したものである。正答率と RMP は Exametrika の LRT 分析、2&3パラメターIRT モデルも Exametrika の個別分析、1 パラメターラッシュモデルについては、Winsteps (Linacre, 2008) を使用している。いずれの間隔尺度の数値も RMP と高い順位相関を示しており、今回の分析データ・条件に関しては一様分布か正規分布の分析を指定することで、間隔尺度とも比較しやすい規準設定が可能になることが確認できた。

表5
正答率、RMP・1 パラメター(ラッシュ)、2 パラメター&3 パラメターIRT 能力推定値のスピアマン順位相関

	正答率	RMP	1P* IRT	2*IRT	3*IRT
正答率	1.000				
RMP	.978	1.000			
1*IRT	1.000	.978	1.000		
2*IRT	.991	.987	.991	1.000	
3*IRT	.993	.986	.993	.998	1.000

潜在ランク数が多くなるほど、間隔尺度に近づくため、今後、もっと少ないランク数で分割した場合も同じような結果になるか、様々なランク数、分布条件の組み合わせを検証する必要がある。

LRT と言語テストの規準設定に関する先行研究

近年、LRT と CTT や IRT との比較（山川・荘島, 2007; 小泉・飯村, 2010; 小山・木村, 2011; Koyama & Kimura, 2011）や英語の大学入試センター試験結果に基づく Can-do システムや学力構造の分析（Sugino et al., 2010; 杉野他, 2011）、学内実施の英語試験による入学者選抜、クラス配置や Can-do 記述の分析（木村, 2009a&b; 小泉・飯村, 2010; 小山・木村, 2011; Koyama & Kimura, 2011）、英語のコンピュータ適応型テストの分析（秋山・木村・荘島, 2011; 木村・永岡, 2011）等の LRT の実践的応用をテーマにした研究が活発に行われるようになっている。

Sugino et al.（2010）は、2004 年度の大学入試センターの英語テストデータから 4 万人のサンプルを抽出し、対象学生を 10 の LRT ランクに分け、各受験者が到達していると考えられる能力を記述した。杉野他 (2011) は、1990、1997、2004 年度の同問題データから 4 万人のサンプルを抽出し、各年の LRT ランク別の能力記述の比較を行った。このような大きな受験者サンプルを、等化手続きによって同じ潜在尺度上で経年比較することで、規準設定の基盤を築くことが期待される。

木村 (2009a&b)は、学内の英語プレースメントへの LRT 分析の応用の可能性を追究している。木村 (2009a) は、入試データに LRT を運用した場合の具体的なメリットについてもシミュレーションを使って、明示している。木村 (2009b) は、LRT の潜在ランクを使って順序尺度上で最初から段階評価を行うことで、クラスをどこで分けるかの判断が容易にできるようになることと、今後 LRT によって推定される能力（潜在ランク）と発達・習得段階の行動や態度を結びつけて考えることができるようになることに期待を寄せているが、このような研究及び実践のアプローチは、自ずと規準設定の議論につながることになるだろう。

小山・木村 (2011) は、LRT をラッシュモデルと併用することで、学内で開発された Can-do statements（CDS）の妥当化に効果があったことを指摘している。Koyama & Kimura (2011)は、LRT の限界や問題点を認識しながら、IRT モデルとともに CDS を言語テストと結びつけていくことの可能性を探求している。CDS や CD 表の作成は、LRT を用いてテストを標準化する上で最も重要であるとされる（荘島, 2010）。LRT に基づくレベルや統計数値情報を掲載した CD 表としては、松宮・荘島 (2009a)の国語や松宮・荘島 (2009a) の数学の事例がある。

LRT は2値（正誤）データだけでなく、IRT と同じように多値データにも活用することができる。（木村, 2009a; Shojima, 2009; 宇佐美, 2009）。多値データに対応する Graded Model （段階モデル）は、心理質問紙のようなリッカート型のデータの分析に適しているとされているが （Shojima, 2009）、宇佐美（2009）のような第 1 言語の小論文試験分析の研究事例もあり、今後、ライティングやスピーキングの constructed-response （応答構築型）のタスクの応用研究や実践が期待される。

残された課題

従前の大規模試験の中心的な役割は、受験者の習熟度段階を数字や記号で表したり、受験者がプログラムに参加するのに必要な能力を有しているかを合否で判断することであったと言える。なかには、段階的な指標を示さない試験もあるが、「70 点は合格、69 点以下は不合格」のような段階区分を行っている場合が多い。また、学期末のクラス評価においても「80 点以上は A、70～79 点は B …」といった間隔尺度的区分が示されても、テスト得点の 1 点区切りの情報だけでは、評価できないことが多いのではないだろうか。

能力は理念的には連続的な尺度の中で位置付けられるものであるが、総括的な (summative) 評価やそれに対応する規準設定においては、段階的に評価することが現実的な手段である。4.3 節で論じた「順位付け法」を 1990 年代に初めて提唱した Alastair Pollitt 氏は、総括的な評価における順位付けの意義を以下のように説いているが、LRT の研究課題にもつながる議論である。

> We are required to do two things: to sort the candidates into a rank order with sufficient precision and categorization to meet the needs that our national educational, economic and political systems place on the examination system, and (usually) to attach constant standards to that ordering. The first requirement ensures that those who will use the results to select some students rather than others are given enough information for their purpose; the second provides a system for interpreting the results, for giving meaningful 'reference' to each point on the scale and monitoring the standards of students' achievement over time and over different examinations.
>
> (Pollitt, 2004: ページ番号なし)

それぞれの教育システムのニーズに呼応した正しく安定した順位付けを行うことは、LRT にとっても重要な課題である。まだ人口に膾炙したとは言えない LRT に基づく決定を行うためには、他のテスト理論やモデルとの比較検証の研究（山川・荘島, 2007; 木村, 2009a&b; 小泉・飯村, 2010; 小山・木村, 2011; Koyama & Kimura, 2011）を継続・発展的に行い、適用の指針が現実の目的に合致することを明示しなければならない。

Pollitt (2004) の掲げた 2 番目の要件は、CEFR (Common European Framework of Reference for Languages) のような枠組みの構築や Can-do 表の作成を指すものと考えられるが、荘島 (2010: 108) は、「Can-do 表の作成は LRT を用いて「テストを標準化する上で最も重要なタスク」であり、LRT は Can-do 表の「作成支援ツールであると言っても過言ではない」としている。

小山・木村 (2011) や Koyama & Kimura (2011) は、既存の英語 Can-do 記述文への回答と英語クラス編成試験の結果を LRT やラッシュモデルで分析しているが、既存の規準に照応させる

「規準維持」(standard maintaining) の観点 (Bramley, 2010) とともに、松宮・荘島 (2009a&b) の国語や数学の事例のように、外国語テストにおいても、大規模な標準テストのLRT分析結果を基にCan-do 記述文を作成し、テスト項目を照合していく取り組みが、今後求められていくことになるだろう。また、あるテストにおけるCan-do記述文の体系が、別のテストにも適用できるのか、様々なテスト結果から、より汎用性のある能力記述へとつなげていくことができるかが、規準設定における大きな課題になるだろう。

また、応答選択型のテストだけでなく、応答構築型のテストへの運用も、4.3 節で述べたmany-facet Rasch model（Eckes, 2011）のような他の分析モデルと比較する価値があるだろう。

4.3 節で述べた潜在クラス理論とラッシュモデルを融合した mixed (or mixture) Rasch model (Jiao, Lissitz, Macready, Wang & Liang, 2011）のようなモデルへの発展は、LRTの追究する段階評価と他の評価情報との統合の理念から考えて現実的でないかもしれないが、Exametrikaのように、分析設定を簡単に変更して、瞬時に大量の（2パラメター以上の）IRT及びLRTの数量的、視覚的情報を容易に導き出すことが可能なソフトウェアも利用できることから、複合的な視点から新しい規準設定手続きの構築へと発展していく素地は、すでにできていると言える。

LRT はまだ多くの研究者に十分に理解されて、利用されているとは言いがたい。理論やその運用への理解が、国内外への研究者や教育関係者の間で広まっていくことが求められている。

謝辞

本節の執筆にあたって、貴重な識見を下さった大学入試センターの荘島宏二郎氏に感謝申し上げたい。

参考文献

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

Eckes, T. (2011). *Introduction to many-facet Rasch measurement.* Frankfurt, Germany: Peter Lang.

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

Koyama, Y. & Kimura, T. (2011). *Linking can-do statements with language tests using neural test theory*. A paper presented at the 50th Convention of the Japan Association of College

English Teachers, Fukuoka, Japan. Retrieved from http://www.rd.dnc.ac.jp/~shojima/ntt/index.htm

Linacre, M. (2008). *WINSTEPS Rasch measurement computer program* (Version 3.66.0). Chicago: Winsteps.com; New York: Kaplan Publishing.

Norizuki, K., Ito, A. & Shimatani, H. (2011). *Text and auditory processing characteristics affecting item difficulty in EFL listening comprehension - Analyzing TOEIC(R) short conversations and short talks.* A paper presented at the 15th JLTA Annual Conference, Momoyama Gakuin University.

Pollitt, A. (2004). *Let's stop marking exams.* A paper presented at the 30th Annual Conference of the International Association for Educational Assessment, June, Philadelphia, USA. Retrieved from http://www.cambridgeassessment.org.uk/ca/digitalAssets/113942_Let_s_Stop_Marking_Exams.pdf

Shojima (2007). Neural test theory. DNC Research Note, 07-02. Retrieved from http://www.rd.dnc.ac.jp/~shojima/ntt/index.htm

Shojima, K. (2009). *Neural test theory model for graded response data.* A paper presented at the International Meeting of the Psychometric Society. (July, 2009; Cambridge University) Retrieved from http://www.rd.dnc.ac.jp/~shojima/ntt/index.htm

Sugino, N., Yamakawa, K., Ohba, H., Shojima, K., Shimizu, Y., & Nakano, M. (2010). *Developing the can-do system based on the NCUEE Test results: An application of the Neural Test Theory.* A paper presented at The 4th Centre for Language Studies (CLS) International Conference (December, 2010, Orchard Hotel, Singapore. Retrieved from http://www.rd.dnc.ac.jp/~shojima/ntt/jindex.htm

秋山 實・木村哲夫・荘島宏二郎 (2011).「LRT モデルに基づく CAT の開発とシミュレーションによる特性解析」第 9 回日本テスト学会(2011 年9月:於;岡山大学) Retrieved from http://www.rd.dnc.ac.jp/~shojima/ntt/jindex.htm

木村哲夫 (2009a).「言語テストにおける段階評価の実際:入試とプレイスメントテストのデータ処理」第 13 回日本言語テスト学会全国研究大会発表. Retrieved from http://www.rd.dnc.ac.jp/~shojima/ntt/jindex.htm

木村哲夫 (2009b).「ニューラルテスト理論による英語プレイスメントテストの作成と評価」『関東甲信越英語教育学会研究紀要』, *23, 23-34*.

木村哲夫・永岡慶三 (2011).「潜在ランク理論に基づくコンピュータアダプティブテスト」 第 9 回日本テスト学会(2011 年 9 月:於;岡山大学) Retrieved from http://www.rd.dnc.ac.jp/~shojima/ntt/jindex.htm

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.

小山由紀江・木村哲夫 (2011).「Neural Test Theory を使った Can-do Statements の分析」統計

数理研究所共同研究リポート 254, 59-78. Retrieved from http://homepage3.nifty.com/yukie-k/publication/36.pdf

松宮功・荘島宏二郎 (2009a).「ニューラルテスト理論を利用した Can-do table 作成の試み」第 37 回 日本行動計量学会(2009 年 8 月 5 日;於:大分大学)Retrieved from http://www.rd.dnc.ac.jp/~shojima/ntt/index.htm

松宮功・荘島宏二郎 (2009b).「ニューラルテスト理論を利用して作成する Can-do table 作成」第 7 回 日本テスト学会(2009 年 9 月;於:名古屋大学)Retrieved from http://www.rd.dnc.ac.jp/~shojima/ntt/index.htm

大友賢二 (2008).「Angoff 法再考」 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.

荘島宏二郎 (2010).「学習評価の新潮流」 日本テスト学会・日本行動計量学会共催チュートリアル・セミナー(2011 年 9 月 11 日;於:岡山大学).

荘島宏二郎 (2011a).「学習評価の新潮流」 日本テスト学会・日本行動計量学会共催チュートリアル・セミナー(2011 年 9 月 11 日;於:岡山大学).

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

荘島宏二郎 (n.d.)「ニューラルテスト理論」 Retrieved from http://www.rd.dnc.ac.jp/~shojima/ntt/jindex.htm

杉野直樹・荘島宏二郎・清水裕子・大場浩正・中野美知子・山川健一 (2011)「英語学力構造の経年変化:潜在ランク理論を用いたセンター試験受験者データの分析」 第 37 回 日本教科教育学会発表 (2011 年 11 月 於; 沖縄大学).

宇佐美慧 (2009).「ニューラルテスト理論の応用可能性－方法論的課題の考察と多値型モデルの適用例－」『日本テスト学会』, 5(1), 65-79.

山川 修・荘島宏二郎 (2007)「項目応答理論とニューラルテスト理論の比較研究」『日本教育工学研究報告集』, 7(5), 223-226. Retrieved from http://www.rd.dnc.ac.jp/~shojima/ntt/yamakawaJSET07.pdf

5. Standard-Setting in CEFR and ELP

5.1. Ways of setting cut-offs of 6 levels of language proficiency in CEFR scale

Sukero Ito

Abstract

The CEFR is a document which describes i) the competences necessary for communication, ii) the related knowledge and skills and iii) the situations and domains of communication. The CEFR defines levels of attainment in different aspects of its descriptive scheme with illustrative descriptors scale. The CEFR divides learners into three broad divisions which can be divided into six levels:

A Basic User (A1 Breakthrough or beginner) (A2 Waystage or elementary)

B Independent User (B1 Threshold or intermediate) (B2 Vantage or upper intermediate)

C Proficient User (C1 Effective Operational Proficiency or advanced) (C2 Mastery or proficiency)

The CEFR describes what a learner is supposed to be able to do in reading, listening, speaking and writing at each level.

Standard setting aims to answer the question what a language learner can do at each level. The answer to this question depends, however, to a great extent on the choice of the standard setting method, since different standard setting methods yield different cut-off scores. And also, standard setting serves one function: to establish cut-off scores between different competence levels. In some contexts it may improve the validity of the cut-off scores if the link between items/tests and the CEFR is less opaque. There is yet a need for some clarification about the interpretation of the cut-off score itself.

５．CEFR と ELP における規準設定

５．１．CEFR における６レベルの規準設定の視点

伊東祐郎

欧州のスタンダード誕生の背景

欧州では、各国の EU 統合という動きの中で、現代語教育システムの統一的観点を構築することが求められていた。複言語主義（plurilingualism）に基づく欧州社会の実現を目指して、言語と文化背景の異なる市民の国境を越えた移動にともなう相互理解や共同作業などを円滑に推進することが必然的に生まれていたのである。そのために外国語学習と教授法、そして運用力評価を研究対象として、コミュニケーション能力を質的に改善することと、言語教育にかかわる政策面での整備・統合を実現する必要があった。具体的には学習者の学習ニーズ、学習動機、学習内容、学習目標などを包括的に捉え、それらを明示し、言語教育に従事する教師や学習者をはじめ、教育行政関係者、教科書出版社、試験問題作成者などに対して一般性を持たせようとしたのである。このような背景から、CEFR（前出）が誕生した。

一方、複合社会・複言語主義の下に教育の統合化が推進される中で、ALTE（The Association of Language Testers in Europe：ヨーロッパ言語テスト専門家協議会、外国語試験を実施する組織）では、欧州における外国語能力の共通認定証の促進を図るために、CEFR の尺度基準に一致するような能力評価の基準作りを行っていた。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必要性が生まれ、テストが測定する能力レベルを同一尺度で確定しようという試みがなされてきたのである。ALTE では独自の"Can-Do Statements"（以下「言語能力記述文」）を開発し、外国語能力の認定とその解釈を可能にしたのである。言語能力記述文の特徴は、目標言語を使って具体的に何ができるかを明文化し、パフォーマンスを基準にした記述となっていることである。

CEFR の生まれる背景には、欧州評議会の言語政策の共通化があり、一方、ALTE では、テストにおける測定尺度の統一化を目指していたのである。

概念・機能シラバスとコミュニケーション能力

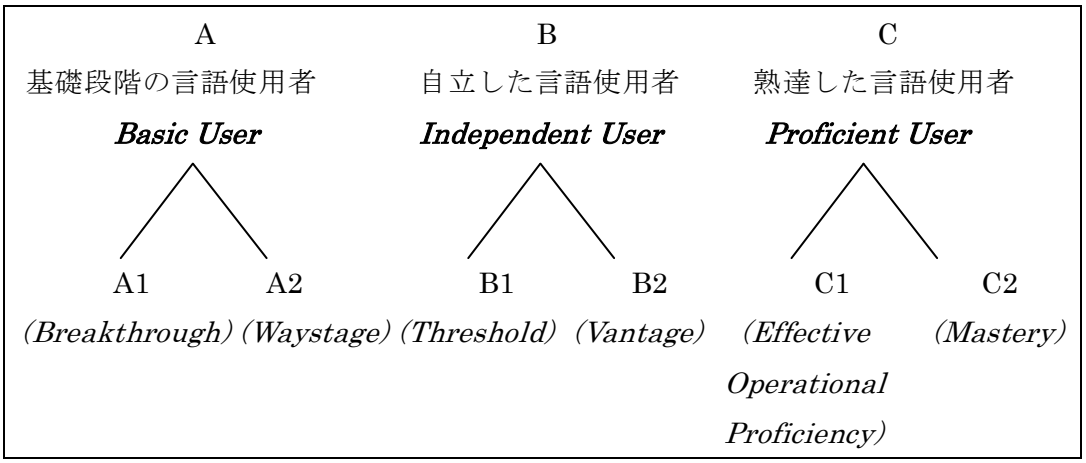
言語能力と言えば、Canale と Swain（1980）が提唱した、コミュニカティブな言語能力モ

デルがよく知られている。彼らの提唱したモデルは、①文法的能力（Grammatical competence）：目標言語の語彙、文型、構文をはじめ、音韻などを理解して、文を構成できる能力。文法構造、単語の意味に限らず、形態や統語などが含まれる。②社会言語学的能力（Sociolinguistic competence）：目標言語が使用される社会的文脈や状況を理解して、適切に言語を運用できる能力。また、言語の社会的機能を活かして円滑なコミュニケーションを行える能力。③談話能力（Discourse competence）：意味があり、結束性のある発話や文章を構成したり、理解したりする能力。意味と言語形式を結びつけ、発話や文章を機能的に伝達できる能力。④方略的能力（Strategic competence）：上記３つの能力不足を補うための能力で、言い換え、回避、転移、繰り返しなど。コミュニケーションを維持したり、補正したりするための対処できる言語的・非言語的能力。以上の４つを構成要素とした。

しかしながら、そのほとんどが言語を内的現象とみなす概念的な枠組みや心理学的な記述を中心に提唱されてきたものであって、決して言語が使用される場面や状況、また社会文化的な要因と関連づけられているわけではない。一般的に言われている言語能力の枠組みは、教育の現場における教授や評価への具体的な指針になるまでに至っていなかったことが指摘できる。言語能力の構成概念が観察可能なものとして定義されていないのである。

言語能力のスケール化への貢献

こうした統合的な言語能力観が議論される中で、実際にコミュニケーションのための言語教育をどのように推進すべきかを検討する上で大きな役割を果たしたのが、Wilkins によって提唱された、概念・機能シラバス（notional-functional syllabus）であった。これは、1970年初頭に欧州評議会（Council of Europe）がコミュニケーション能力を育成するための教授法の開発を目指して新たに提案したものである。欧州で働く成人学習者が表現する意味と実際の言語の機能をまとめて配列した項目一覧（シラバス）である。コミュニケーション能力を育成するために、実際の言語運用場面や状況を把握した上で、指導項目の基準を作成し、学習者のニーズに応じて、指導項目を選定できるように試みたものである。コミュニケーションを実現するために必要な意味を概念（notion）（例：時間、量、期間、位置など）と機能（function）（例：提案、要求、謝罪、拒否など）とに分類したことが特徴的である。このような貢献は、以下に示すような言語学習者の言語熟達度を垂直的に体系化した「Breakthrough（サバイバル）」「Waystage（スタートライン）」「Threshold（入門段階）」「Vantage（中等教育後期）」「Effective Operational Proficiency（効果的に操作可能な熟練レベル）」「Mastery（熟練レベル）」の基礎となっている。これは、欧州の言語学習者の言語学習と熟達度を大まかに６レベルに分類していて、言語能力を具体的な形で記述している。CEFR における言語能力記述文はこのような背景から生まれたものである。



言語運用能力のスケール化は、最近の言語教育内容の透明性や説明責任に対する関心の高まりとともに広まってきたと言える（North 2000）。1950年代に開発された US Foreign Service Institute (FSI)のスケールは大変珍しいものであった（Wild 1965）。1990年代には、the British National Language Standards (Language Lead Body 1992)、the Eurocentres Scale of Language Proficiency (North 1993)、the Finnish Scale of Language Proficiency (Luoma 1993)、the ALTE Framework (Association of Language Testers in Europe 1994)が開発されてきた。これらに共通する特徴は、特に言語能力の運用力の評価においては、「現実世界」で何ができるかどうかを測定することが重要になってきたことと深く関わっている（Bachman 1990:325-330)。言語を使っての双方向のやりとり(interactive-ability) に焦点をあてたパフォーマンス・テストの評価におけるスケール化が必然的に出てきたことも見逃せない。

言語能力記述文("Can-Do Statements")とは

ALTE が、欧州における外国語能力の共通認定証の促進を図るために開発したものがよく知られている。以下の表は、その概略である。

"ALTE 'Can Do' statements"（ALTE 能力記述文概略）

ALTE レベル	聞くこと／話すこと	読むこと	書くこと
ALTE レベル5	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。
ALTE レベル4	自分の仕事の範囲内で会議やセミナーに効果的に貢献し、抽象的な表現に対処しながらかなりの流暢さでうち解けた会話を維持することができる。	学習コースに十分対応できるほどに早く読み、情報を得るために媒体を読み、非標準的な通信文を理解することができる。	職業上の通信文を下書きしたり作成したりし、会議で適度に正確なメモを取り、コミュニケーションできる能力を示すエッセイを書くことができる。
ALTE レベル3	よく知っているトピックを題材に会話ができ、話についていくこともでき、またはかなり幅広い話題について会話を維持することができる。	関連する情報を得るために文章を検索して、細かい指示や助言を理解することができる。	人が話している間にメモを取り、あるいは非標準的な依頼を含む手紙を書くことができる。
ALTE レベル2	限られた方法で抽象的・文化的な事柄について意見を述べ、あるいは周知の範囲内で助言をし、説明・指示や公示を理解することができる。	日常的な情報や記事を理解し、精通している分野内の非日常的な情報について全般的な意味を理解することができる。	よく知っている事柄またはありきたりの事柄について、手紙を書きメモを取ることができる。
ALTE レベル1	慣れた環境の中で、単純な意見や要求を表現することができる。	周知の範囲内で率直に書かれた情報、たとえば製品に関する情報や、標示、簡単なテキストブック、またはよく知っている事柄に関するレポートを理解することができる。	用紙に記載し、個人情報に関係する短い簡単な手紙やハガキを書くことができる。
ALTE Break-through レベル	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）

ALTE の言語能力記述文は６レベルを枠組みとしているが、実は、CEFR の６レベルが拠り所としているところと基盤は共通である。

各レベルにおける定義付け

それでは、それぞれの言語熟達度はどのような規準をもとに定義づけられているのだろうか。項目応答理論により、調査の結果厳選された 212 の言語能力文の困難度が計

算された。推定された困難度-5.68 から 4.68 の範囲に分布した言語能力文を一つずつ分析し、内容性、類似性、わかりやすさの点から検討し修正を加えた。その後、目標基準をどのように設定するかを検討しながら、主観的に cut-off ポイントを設定している。基本となった考え方は、(1) スケール化を決める際には logit 値を参照、(2) Threshold (B1) を基本軸にそれぞれの logit 値との差を基に傾向を分析、(3) 従来からのレベルとの比較から検討、という手法であった。

以下の表 (North 2000) は、各レベルの構造を示したものであるが、等間隔によるレベルが設定され、各レベルおおよそ 1 logit の間隔が保たれている。表に続いて、各レベルの定義を簡略にまとめておく。

Equal Interval Levels and Common Reference Levels

	Common Reference Levels	Finer Level (Swiss)	Abbrev	Cut- off	Range on Scale
C2	Matery	Matery	M	3.90	
C1	Effective Operational proficiency	Full Effectiveness	EOP	2.80	1.10
	(*Vantage Plus*)	Effectiveness	V+	1.74	1.06
B2	Vantage	Ful independence	V	0.72	1.02
	(*Threshold Plus*)	Independence	T+	-0.72	0.98
B1	Threshould	Threshould	T	-1.23	0.97
	(*Waystage Plus)*	Waystage Plus	W+	-2.21	0.98
A2	Waystage	Waystage	W	-3.23	1.02
A1	Breaktrhrough	Breakthrough	B	-4.29	1.06
	-------	Tourist	Tour	-5.39	1.10
	-------	Smattering			

(出展：North 2000,p.274)

Smattering：レベルと呼ぶにはあまりにもわずかな言語能力。ごく簡単で一回限りの発話。例えば、ごく簡単な挨拶や、はい／いいえ、すみません、ありがとう、程度の発話。

Tourist：旅行者が行うような宿帳への記入、絵はがきを書く、時間を聞いたり教えたりする、簡単な買い物が能力規準となっているところからこの名称がつけられた。

Breakthrough (A1)：一般的な言語使用における最下位に相当するレベル。学習者はごく簡単な方法でやり取りするレベル。住んでいる場所や知っている人物、自分が所有している物など自分自身についての質問や応答ができるレベル。場面限定の決まり切った Tourist レベルより高いレベルで、A1 として位置づけられる。

Waystage (A2)：欧州評議会が定めている初級レベル。多くの能力記述文は、日常生活における挨拶やどんな仕事をしているか、暇なときには何をしているかを尋ねたりする程度の会話、また誘われた時に返答したり、約束の時間や場所を決めたり、申し出を了承したりするような会話レベル。

Waystage Plus：このレベルは Waystage の得意とする部分と次のレベル Threshold の弱点の

ちょうど中間レベル。共通参照枠の A＋相当レベル。顕著な行動としては、会話に積極的にかかわろうとする様子が見られる。対面会話を自ら始めたり、決まり切ったやり取りを努力なしに続けたり、日常の話題について質問したり、ごく簡単な会話を維持させられるレベル。

Threshold（B1）：欧州評議会では外国訪問者に該当するレベル。二つの特徴が見られる。ひとつは、やり取りを維持したり、ディスカッションで話の筋道が理解できたり、個人的な見解を述べたりできるレベル。もうひとつは、日常生活の予期せぬ事態や課題に対応できるレベル。例えば、旅行会社と旅行の計画を立てたり、苦情を述べたり、相手の発言したことを明確に述べてほしいと要求したりできるレベル。

Independence：Threshold レベルよりもより独立した言語行動ができるレベル。問題について説明したり、インタビューでの質問にわかりやすく答えたり、詳細とまではいかないまでも医者に症状を説明したり、理由や根拠を要約できるレベル。事前に準備したインタビューをしたり、入手した情報を確認したりできるレベル。

Vantage（B2）：Full Independence とも呼ばれるレベル。学習者は新たな段階に来たことを実感できるレベルで、これまでとは異なる展望を感じるようになるレベル。自分の意見を主張し、そのための説明ができるレベル。仮説を述べたり、提案をしたりする中で、長所や短所も説明できるレベル。談話能力を発揮し、自然にしかも流暢且つ効果的に会話を実現できるレベル。

Vantage Plus：Effectiveness と形容されるレベル。幅広いジャンルにおける話題に対応でき、その上で、言語使用においても流暢でかつ正確、効率のよさが現れるレベル。

Effective Operational Proficiency（C1）：Full Effectiveness と位置づけられているレベル。このレベルは、流暢さに加え、コミュニケーションに即時性（spontaneous）が求められる。どんな場面においても言語的な限界を感じることはなく、語彙や統語面においても難なく対応できるレベル。流暢さというのは、どのレベルでも認められれば、加えて、このレベルでは、談話の一貫性、すなわち洗練された会話の構成と流れ、などの管理能力に優れているレベルと言える。

Mastery（C2）：このレベルはネイティブあるいはネイティブに近い言語使用レベルを意図していない。言語運用における正確さと適切さにおいて優れているレベルである。

CEFR の 6 レベルにかかわる研究動向

CEFR の 6 レベルの能力規準に関しては、運用面や応用面における一貫性や透明性、そして信頼性の確保及びその保証のために各種の研究や調査などが行われ、研究報告書という形で公開されている。その中で、Cito（2009）から出版されている Linking to the

CEFR levels: Research perspectives は、理論面と実践面から意味のある示唆を得ることができる。この報告書は 2 部構成になっており、第 1 部は理論面からの考察、第 2 部は実践面からの論証と課題をまとめている。第 1 部で、Mark D. Reckase は、第 1 章 Standard Setting Theory and Practice: Issues and Difficulties の中で、「スタンダード」を定義した上で、ラッシュモデルや３パラメターを使っての統計データを示しながら解説を加え、Standard Setting の作業段階と妥当化について詳述している。課題等についても言及してる点が面白い。第４章の Norman Verhelst は、Linking multilingual survey results to the Common European Framework of Reference　の中で、テスト開発を取り上げ、言語能力によって異なる視点から記述する必要があると述べている。まず、テストの細目表（Test specification）は基礎基本であることを述べ、関係者がテストと CEFR の内容に精通（Familiarisation）していることを強調している。産出スキルであるライティングの評定については、明確なサンプル規準に基づいて判定することが重要であると述べている。加えて、受容能力であるリスニングとリーディングについての Standard　Setting の方法を解説している。第２部第 13 章 Standard Setting for Listening, Grammar, Vocabulary and Reading Sections of the Advanced Level Certificate in English (ALCE)の中で、N. Downey と C. Kollias は、実際の大規模テストを取り上げて、CEFR との関連づけを通して、Standard Setting の実践例を紹介している。この実践研究を通して、CEFR は言語的側面である文法や語彙にかんするスケールが明示されていないためにテスト開発が困難である点を述べている。また、手法については Angoff 法が好ましいとの言及もある。第 10 章では、Jamie Dunlea and Tomoki Matsudaira による Investigating the Relationship Between the EIKEN Tests and the CEFR も掲載され、英検にかかわる報告もある。

残された課題

言語能力のスケール化は、多様な名称で説明されている。例えば「バンドスコア（band scores）」「バンドスケール（band scales）」「能力レベル（proficiency levels）」「能力スケール（proficiency scales）」（Alderson　1991a:71）や「ガイドライン（guidelines）」「スタンダード（standards）」「レベル（levels）」「スケール（scales）」（De Jong　1992:43）などである。これらは総じて言語運用力の一連の段階的記述を試みようとするものである（North et al 1992）。また、統合的言語能力のヒエラルキーとも説明されている（ACTFL 1986）。このように、言語能力のスケールにかかわる定義は、それを命名した筆者の能力に対する見解や議論、また根拠を反映したものであることがわかる。一般的に、このようなスケールは、言語学習や言語習得の段階的到達点をイメージ化したものであることが多い。このスケールを使用したり参照したりする者は、達成の度合いや程度を認識するために活用すること

が多いが、スケールが必ずしも第二言語の習得の成功を正確に明示しているかどうかの妥当性の根拠があるとは限らないものである（Brindley 1998）。

最後に、残された課題を以下にまとめておきたい。

(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

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.

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

Bachman, L. F. (1990) *Fundamental Considerations in Language Testing*. Oxford University Press.（池田央・大友賢二監修 (1997)『言語テスト法の基礎』C.S.L.学習評価研究所）

Bachman, L. F. & Palmer, A. S. (1996) *Language Testing in Practice : Designing and Developing Useful Language Tests*. Oxford University Press.（大友賢二他監訳（2000）『<実践>言語テスト作成法』大修館書店）

Brindley, Geoff (1998): Describing Language Development? Rating Scales and Second Language Acquisition. In: Bachman, L.F. and Cohen, A.D. (eds.): *Interfaces between SLA and Language Testing Research*. Cambridge: University Press.

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

Canale, Michael and Swain, Merrill (1980): Theoretical bases of communicative approaches to second language teaching and testing. In: *Applied Linguistics* 1/1. 1-47

Cito（2009）Neus Figueras & Jose Noijons (eds.) *Linking to the CEFR levels: Research perspectives*. Cito, Institute for Educational Measurement, Council of Europe, European Association for Language Testing and Assessment (EALTA)

Council of Europe (2001) *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge Universityo Press.（吉島茂他訳 (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）

Language Lead Body (1992) *National Standards for Languages: Units of Competence and Assessment Guidance*. UK Language Lead Body, July 1992.

Luoma, Sari (1993): *Validating the (Finnish) Certificates of Foreign Language Proficiency*. Paper presented at the 15th Language Testing Research Colloquium. Cambridge / Arnhem, 2-4 August 1993.

McNamara, T. F. (2000) *Language Testing*. Oxford University Press.（伊東祐郎他監訳（2004）『言語テスティング概論』スリーエーネットワーク）

North, Brian (1993). Transparency, Coherence and Washback in Language Assessment. In: Sajavaara, K., Lambert, R., Takala, S. and Morfit, C. (eds.), 157-193

North, Brian (2000). *The Development of a Common Framework Scale of Language Proficiency*. (Theoretical Studies in Second Language Acquisition. Vol. 8) New York: Peter Lang Publishing, Inc..

North, Brian, Page, Brian, Porcher, Louis, Schneider, Gunther and Van Ek, Jan A. (1992): A *Preliminary Investigation of the possible Outline Structure of a Common European Language Framework*. Strasbourg: Council of Europe, Language learning for European citizenship CC-LANG(92)12.

伊東祐郎（2006）「日本語能力"Can-do"記述文作成の試みーテスト得点の妥当化をめざしてー」『高見澤孟先生古希記念論文集』

伊東祐郎（2005）「これまでの評価／これからの評価」『AJALT』第 28 号、国際日本語普及協会

金谷憲編著（2003）『英語教育評価論』桐原書店

国際交流基金編（2003）『日本語能力試験企画小委員会口頭能力試験調査部会報告ー口頭能力試験科目の創設に向けてー』日本語能力試験企画小委員会口頭能力試験調査部会

国際交流基金・日本国際教育協会（2004）『日本語能力試験　出題基準[改訂版]』凡人社

三枝令子他 (2004)『日本語 Can-do-statements 尺度の開発』（平成 13 年度～平成 15 年度 科学研究費補助金・基盤研究 B1　研究成果報告書）

靜哲人他 (2002)『外国語教育リサーチとテスティングの基礎概念』関西大学出版部

東京外国語大学（2006）『JLC シンポジウム報告書ー日本語スタンダードを考えるー』

日本語教育学会編 (1999)『Can-do-statements 調査報告』国際交流基金

日本語教育振興協会（2001）『運用能力獲得のための基礎日本語能力』

5.2. The social backgrounds for the origin of the CEFR: Plurilingualism and Pluriculturalism

Kenji Ohtomo

Abstract

It is essential to consider the social background of the origin of the CEFR when discussing it. In order to fully understand the background, attention has to be paid to the Council of Europe Language Education Policies. The significant points of the policies are an understanding of (1) plurilingualism, (2) linguistic diversity, (3) mutual understanding, (4) democratic citizenship, and (5) social cohesion.

What is meant by "plurilingualism"? The policy states that all citizens are entitled to develop a degree of communicative ability over their lifetime in a number of languages, and in accordance with their needs. In particular, it is necessary to understand the importance of plurilingual and pluricultural competence as described in section 6.1.3.1: An uneven and changing competence in CEFR, published by Cambridge University Press in 2001. The essential features here are: plurilingual and pluricultural competence is generally uneven in one or more ways.

The report concludes by summarizing the differences between the educational backgrounds of Europe and Japan and by proposing that further research should be explored.

５．２．CEFR 誕生とその背景：複言語・複文化主義など

大友賢二

CEFR 誕生の意味

CEFR 誕生とその背景を検討するにあたって、それと同時に、わが国の言語教育の背景はどうなってきたのか、その歴史を振り返ってみることの意味は大きい。

> 「江戸時代から３回目の世紀を迎えて間もない１８０８年（文化５）の８月１５日早朝、長崎湾の沖合に３本マストの不審船が 1 艘、前触れもなく姿を現した。その異国船は（当時日本との貿易を許されていた中国・オランダ以外の国の船を指す）は、日が傾きかけた頃、オランダ国旗を掲げて、港内に侵入して投鎖した。」(p.4)

このオランダ国旗を掲げた船は、あとで、英国軍艦フェートン号で、あることが判明した。これがフェートン号事件の最初をしめす伊村（2003）からの一節である。これは、日本の外国語教育を語るとき、最初に取り上げられる事項のひとつである。1851 年の中浜万次郎の帰国、1898 年の斎藤秀三郎『実用英文法』、1912 年の市河三喜『英文法研究』、1922 年の H. E. Palmer 文部省英語教授顧問の来日と Oral Method、1945 年の敗戦、米軍進駐、英語ブーム、1956 年の C. Pries 来日、ＥＬＥＣ設立と Oral Approach、Robert Lado, N. Chomsky, Communicative Approach, Focus on Form, Content and Language Integrated Learning, 小学校外国語活動、など枚挙にいとまがない。

一方、ヨーロッパに関する詳細は省くが、わが国との大きな違いは、ヨーロッパの言語教育方針の基本は、「不戦共同体」の建設に向かったことである。大谷（代表）（2010）では、その動向を、次のように述べている。

> ヨーロッパの大国のドイツとフランスは、19 世紀後半から 20 世紀前半までのわずか 80 年足らずの間に,普仏戦争、第 1 次世界大戦、第 2 次世界大戦と、実に 3 度も戦火を交え、殺し合い、憎みあった。そして、周辺の国はそのたびごとに大変な被を被り続けた。３度にわたって相戦った悲惨な体験をもつドイツ、フランスとさらにその周辺国は、戦後の国連の成立をもってしても、なお国際的な平和と安全の確のための十分な条件とは考えなかった。彼らは、単なる理念でなく、ドイツとフランスの「和解」、しかも、「恒久的和解」のための具体的な政策としての「不戦共同体」の建設を構想し、それをヨーロッパの新たな国

際的秩序とする以外にはないと考えた。(p.9)

この動向は、「欧州評議会」（Council of Europe）の設立、2001 年を「欧州言語年」（European Year of Languages）と定める原動力となった。

複言語・複文化主義の開発（１）

さきに述べたヨーロッパ諸国の背景をみると、彼らが作り上げた欧州連合は、まさに、むかしの大戦終結時の状況をみごとに乗り越えてしまっていることが分かる。「欧州連合」（European Union: EU）加盟国は 27 カ国であり、その中で、23 の公用語のすべてをひとしく、ＥＵの公用語として認めている。この 23 の公用語を持つＥＵの動向は、われわれ日本人には想像を超えた力を持っていると言えよう。2004 年の時点での調査によると、約 7000 人の職員をかかえ、そのための費用として、年間約８億 8000 万ユーロというお金をつぎ込んでいたと伝えられている。それでも、これほどの費用は、「不戦共和体」ＥＵの維持のための費用としては、安価であると考えているとのことである。じつに驚くべき動向である。

こうした動向の背景とも言われる複言語、複文化主義の詳細に関しては、あとで触れることとする。しかし、その基盤となっているのは、欧州評議会の「言語政策局」の「ヨーロッパの複言語教育：半世紀にわたる国際協力」（L’education plurililengue en Europe: 50 ans de cooperation international)と題された文書の中で見ることができる。以下、その内容を、細川・西山編（2010)に基づいて示すこととする。

欧州評議会の言語政策は、以下の点の促進を目標としている：

複言語主義：欧州のすべての市民各自の必要に基づき、生涯にわたって、複数言語による、ある程度のコミュニケーション能力を獲得する権利を持つ。

言語の多様性：ヨーロッパは多言語の大陸であって、その言語はすべて、コミュニケーションの手段ならびにアイデンティティの表現手段として同等の価値を持つ。欧州評議会の協定は、これらの言語を使用し、学ぶ権利を保障する。

相互理解：異文化間コミュニケーションや文化的差異の受容は、多言語を学ぶ可能性に基礎づけられている。

民主的市民性：多言語社会において、市民それぞれの複言語能力は、民主的かつ社会的なプロセスへの参加を容易にする。

社会的結束：個人の成長、教育、雇用、移動、情報へのアクセスと豊かな文化的な生活をおくる機会の均等は、生涯にわたる言語学習の可能性にかかっている。」(p.3)

このように、言語の多様性と言語権を促進していることは見逃すわけにはいかない。

複言語・複文化主義の開発（２）

「外国語の学習、教授、評価のためのヨーロッパ共通参照枠」（Common European Framework of Reference for Languages: Learning, teaching, assessment）は、2001 年に刊行され、35 の言語にも翻訳されていると言われている。わが国では、英語版を原点として 2004 年に吉島　茂・大橋理枝他によって、朝日出版社から出版されている。以来、多くの言語教育者によって関心が寄せられている。とくに、can-do statements は、多くの中学校、高等学校、大学等で取り上げられている。また、最近の文部科学省では、平成 23 年 6 月 30 日、「外国語能力の向上に関する検討会」による「国際共通語としての英語力向上のための 5 つの提言と具体的施策」を発表し、そのなかの「提言 1：生徒に求められる英語力について、その達成状況を把握・検証する」に見られる“can-do リスト”などの言及のなかにもその関心が見られる。

しかし、この動きに対する注目の声も聞かれる。たとえば、西山教行（2010）で示されている序における以下の 3 点は、その一例である。

> 受容はおよそ 3 種類に類型化できる。第一に研究者による紹介があげられるが、その多くは「共通参照レベル」の紹介に費やされ、どのようなヨーロッパの社会的文脈から、とりわけ、ヨーロッパ統合の動きの中からなぜ何を目的としてこの教育装置が出来上がったのかについてはほとんど検討が行われていない。第二に、『参照枠』の「共通参照レベル」を無批判に適応する手法が認められる。・・・・・第三に挙げられる傾向は、『参照枠』から着想を得て、日本独自の「言語共通参照枠」を作成しようとするもので・・・・・。しかし、『参照枠』の揚げる複言語、複文化主義という言語教育思想を考慮することなく、単一言語主義の発想にとどまり、「共通参照レベル」をほぼ唯一の着想原としている点に限界がみとめられる。（p.v）

こうした批判の声に耳を傾けながら、さらにその究明を進めるにあたって、常に、心しておきたいごく当たり前の疑問は、大木・西山編（2011)と同様に持ち続けたいものと考えている。たとえば、１．ネィティブは、もはや外国語学習のモデルではないのでは

ないか、２．そもそも、なぜ外国語を学ぶ必要があるのか、3．日本人も複数の言語を学習して、相互理解、民主的市民性、社会的結束を推進する必要があるのか、４．多文化共生のために必要な外国語は、英語だけでいいのではないか、５．なぜ英語だけでは「文化的アイデンティティと多様性をさらに尊重する」ことも、「よりよい相互理解」も十分できないのか。

複言語・複文化主義の発展（１）

複言語主義と多言語主義

いわゆる「複言語主義」（plurilingualism）と呼ばれているものは、それと同じような意味合いを持っている「多言語主義」（multilingualism）とは異なる概念である。ここで、multi に対する pluri という言葉は、大きな辞書にしか出ていないが‘several, many の意を表すラテン語起源の造語要素’（小学館ランダムハウス英和大辞典 (1979)である。それを、まず明確にしておくことが必要である。「多言語主義」、または、多言語使用というのは、個人のレベルでは複数の言語を使用していること、あるいは、国または地域集団で複数の言語が使用されていることを意味している。　しかし、「複言語主義」が強調しているのは、複数の言語や文化を完全に切り離して、心や頭の中の別々の部屋にしまっておくわけではないということである。その複数の言語や文化を一つの部屋にしまいこみ、その中で新しいコミュニケーション能力を作り上げているという点である。この視点は、Council of Europe (2001) でも次のように示されている。

> “ he or she does not keep these languages and cultures in strictly separated mental compartments, but rather builds up a communicative competence to which all knowledge and experience of language contributes and in which languages interrelate and interact. “ (p.4)

> “ Plurilingual and pluricultual competence refers to the ability to use languages for the purpose of communication and to take part in intercultural communication, where a person, viewed as a social agent has proficiency, of varying degree, in several languages and experience of several cultures.”(p.168)

つまり、複数の言語を知っているだけでは、多言語と言えても、複言語とは必ずしも言えないということである。複言語と言えるためには、個人の心や頭に入っている複数の言語間の境界があまりなく、たえず出入り可能な状態になっていることが必要である。

だからこそ、たとえば個人の中で、２つの言語や文化の共通点、相違点を意識でき、個人の言語や文化を相対化できるのである。そのことは、大木・西山編（2011)でも明記されている。

このことに関するきわめて重要な資料としては、Plurilingual Education in Europe: 50 Years of International co-operation (Council of Europe（2006)）などがある。これは、 Part 1: Council of Europe Language Education Policy, Part 2: History and Current Developments, Part 3: Policy Instruments and Initiatives からできている。

複言語・複文化主義の発展（２）

理想的母語話者

大木・西山編（2011: 4）で述べている「ネイティブはもはや外国語学習のモデルではない。学習者の目標は、ネイティブ・スピーカーのようになることではない。」という視点は、従来の言語教育の目指すものとは、基本的に異なっているということにも注目しなければならない。この視点を採るならば、言語教育の究極目標としては、以下のように、「理想的母語話者」（ideal native speaker)とすることは再考を要することになる。

> It is no longer seen as simply to achieve ‘mastery’ of one or two, or even three languages, each taken in isolation, with the ‘ideal native speaker’as the ultimate model. Instead, the aim is to develop a linguistic repertory, in which all linguistic abilities have a place. (Council of Europe (2001: 5))

部分的言語能力の意味

複言語と複文化の能力には何らかの偏りがあるのは、全く自然な現象である。つまり、一つの言語の熟達度は、他の言語の熟達度と同じでなければならないと言う必要はない。このような不均衡は、全く普通である。むしろ当たり前であると言える。このことは、**6.1.3.1.** An uneven and changing competence : Plurilingual and pluricultural competence is generally uneven in one or more ways.… Such imbalances are entirely normal.(Council of Europe (2001:133))でも明らかである。この「不均衡」に関する言及は、その後の議論にも多くみられるが、主だったものの一つには、Coste, D., Moore, D., and Zarate, G.(2009) が示した. *Plurilingual and Pluricultural Competence* がある。たとえば、その **2.1.3.** Partial competence and plurilinguistic competence (p.12) や **6.3.4.** Return to the concept of partial competence (p. 28) などはそれである。この原本は、ＣＥＦＲ出版の前にフランス語版で出版されている。英語版は、2009 であるが、フランス語版は 1997 年にすでに出版されて

いるので、2001 に出版されたＣＥＦＲにこの視点が反映されていることは確かである。

このような部分的言語能力の持つ意味は、外国語学習と外国語教育の領域でさらに深く検討しなければならない課題である。この点に関しては、「「部分的能力」は、複合的能力の一部であり、同時に、具体的な限定的目標との関係で「機能的能力」である」と述べていて、さらには、「このCEFRは、部分的能力という概念を、第３章で示したモデルの各構成要素や、さまざまな目標との関連から見ようとしているもの」という重要な意味を持っている **6.1.3.4.** partial competence and plurilingual and pluricultural competence（Council of Europe(2001: 135）の考え方に目を転じなければならない。

残された課題（１）

複言語・複文化主義と異文化理解

たとえば、「英語教育と異文化理解」いう言葉をよく耳にする。その場合、最も注目しなければならないのは、異文化理解とはどんなものであるか、という疑問である。しかも、「異文化」という場合、「文化」とは何であるかという基本的課題がある。このように異文化と言う場合、異なった文化があるのが前提であろうが、その場合、「文化」と「文化」の境界はどこに求めるのが適切か、という課題にも直面する。このようにして、残された課題は、「文化」とは何か、そして、「文化」と「文化」との境界をどこに、どのような形で求めることができるか、ということが課題である。

たとえば、わが国の外国語教育として「英語」をとりあげた場合、これを、どう捉えるのが適切であろうか。わが国の中学校での評価の４つの観点には、「コミュニケーションへの関心・意欲・態度」や、「言語や文化についての知識・理解」が含まれている。これと「文化」と「文化」との境界をどう取り上げるのが適切かは、検討されなければならない課題である。その解決がなければ、「異文化理解」は存在しないからである。

複言語・複文化主義は、社会におけるというよりも、個人における複数の言語と複数の文化との関わりを重視するものであると理解する。そして、異なった言語や異なった文化を理解することによって、言語と文化の境界を越え、複数言語で、双方向的な行動ができる人間育成が目的である。こうした極めて広大な目標に向けて、わが日本の学校でも進むことが必要であれば、それにどう対応することが必要であろうか？細川 (2012) でのべていることは、こうした異文化理解教育と複言語・複文化主義の課題究明のための一つの糸口となるであろう。

複言語・複文化主義と言語教育

欧州評議会による文書、Council of Europe (2007)は、きわめて貴重な資料である。われわれが、言語教育、とりわけ、ヨーロッパにおける複言語教育に関する視点をどこに求めたらよいかは、その究明手段を見出すのが困難である。そのひとつの糸口は、この資料にあると考えることができる。残された課題の内容としては、この「ヨーロッパの複言語教育」を以後、さらに検討し、その本質を探らなければならない。ここでは、その意味で、欧州評議会の作成している言語教育のための視点として挙げている３つのことがらの題目だけを示すこととする。これを基に、複言語・複文化主義と言語教育との関連をさらに究明することが残された課題の一つである。その３点は、（１）Language Education Policies（２）Data and Methods for the Development of Language Education Policies（３）Organizational Forms of Plurilingual Education である。この具体的な内容には、＜http://www.coe.int/t/dg4/linguistic/default_EN.asp＞ でアクセスすることができる。

残された課題（２）

わが国の生徒指導要録

ヨーロッパのこれまで述べた複言語・複文化主義の動向は、わが国の異文化理解の動向とは、大きな距離が見られる。その距離はどんなところに見られるか、その距離を短縮するのが適切であると考えられるのか、こうした課題は残されたままであろう。

わが国の学校教育における外国語教育の実態を考えてみよう。その指導の目的として挙げている内容は、一般的には、その評価の内容ともなっている。たとえば、現在の中学校における生徒指導要録では、４つの観点をもとに評価するように示されている。「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」というのがその観点である。こうしたそれぞれの観点に対する評価は、「十分満足」をＡ，「おおむね満足 」をＢ，「努力を要する」をＣとして行われ、そのすべてを総合した「評定」では、５段階の評価をすることになっている。また、これまで、絶対評価を加味した相対評価で行っていた「評定」は、これを、絶対評価で行うことに改訂した。

この状況に関するわが国には、さまざまな意見がある。そのひとつは、平成２２年の報告をまとめた中央教育審議会のワーキンググループ（第９回議事録）では、「関心、意欲、態度の評価結果を評定に入れることで、評定は信頼のおけないものだという批判が付きまとう」という指摘もある。これに関連して、松沢（2011）では、中学校で、平成２４年度よる始まる評価を「新しい評価２」とすれば、「「新しい評価２」では、（新設の）「外国語学習への関心・意欲・態度」と「言語や文化についての知識・理解」の観点の成績は、通知書や生徒指導要録に記録して指導に活用するが、学力証明としての評定に

は含めない、としたい。」(p.12) という意見もある。この種の課題は、やはり、残された課題の一つでもある。

統合的英語能力と partial competence

複言語・複文化主義では、人間の複言語能力全体は、不均衡な状態であることを認めている。外国語教育の目標としては、わが国においては、「理想的な母語話者」を設定し、それに向かって日々努力を重ねるのが従来の動きである。しかも、ごく最近の新学習指導要領下の英語では 4 技能の統合が課題になっている。例えば、新学習指導要領では、「中学校・高等学校では、 4 技能をバランスよく育成することを目指している。この「バランスよく」が「統合的」ということであり、特定の技能に偏った指導にならないようすることを意味している。」（向後（2011)。

こうした議論の行方が気になる。一体、わが国の「統合的英語能力」を目指す外国語能力向上の動向と複言語学習における外国語能力向上のための動向の接点はいずこにあるのであろうか、これも残された課題である。

参考文献

Coste, D., Moore, D., and Zarate, G. (2009). *Plurilingual and Pluricultural Competence*, Language Policy Division, Council of Europe.

Council of Europe (2001)*Common European Framework of Reference for Languages: Learning, teaching , assessment.* (p.5) Cambridge University Press.

Council of Europe （2006), *Plurilingual Education in Europe: 50 Years of International co-operation.* Language Policy Division, Strasbourg.

Council of Europe (2007), *From Linguistic Diversity to Plurilingual Education* (main version), Language Policy Division

伊村元道（2003）『日本の英語教育 200 年』(p.4). 大修館書店

岡戸浩子（2003）『「グローカル化」時代の言語教育政策』くろしお出版

嘉数勝美（2011）『グローバリゼーションと日本語教育政策』株式会社ココ出版

吉島　茂 (2007) ヨーロッパの外国語教育を支える考え方――複言語・複文化主義、行動主義、4 つの Savoirs、部分的能力、European Language Portfolio (Can Do Statement) 『英語展望』2007 年増刊号(114). 49-54.

向後秀明(2011)、「４技能統合におけるスピーキング指導はどうあるべきか」『英語教育』（July, 2011, Vol.60,)(p.10) 大修館書店

細川英雄・西山教行　編（2010）『複言語・複文化主義とは何か――ヨーロッパの理

念・状況から日本における受容・文脈化へ――』(p.3) くろしお出版.
細川英雄（2012）「異文化理解教育と複言語・複文化主義」ジョイント研究大会予稿集: ２０１２年３月１０日、早稲田大学におけるフランス語教育学会、日本独文学会ドイツ語教育部会、ＪＡＣＥＴ教育問題研究会、西山教行科研、神保尚武科研の大会
小学館ランダムハウス英和大辞典編集委員会（1979）『小学館ランダムハウス英和大辞典』(p.1987)（株）小学館
松沢伸一（2011）、「「新しい評価」は定着したのか」『英語教育』（May, 2011, Vol.60 No.2）(p.12) 大修館書店、
大谷泰照（2007）『日本語にとって英語とは何か』2007、大修館書店
大谷泰照（編集代表）（2010）『ＥＵの言語教育政策 : 日本の外国語教育への示唆』(p.9) くろしお出版.
大木　充・西山教行編（2011）『マルチ言語宣言』(pp.4-5, pp.13-14) 京都大学学術出版会

5.3. The role and function of a Language Passport, Language Biography, and Dossier in the European Language Portfolio（ELP）

Sukero Ito

Abstract

The European Language Portfolio (ELP) was conceived along with the Common European Framework of Reference for Languages (CEFR) in 1991. The ELP consists of 3 parts: 1. The Language Passport: to help learners assess their competencies in the language(s) being learned as well as the growth in these competencies. Learners may also record learning and intercultural experiences here. 2. The Language Biography: to set learning targets and regularly assess progress in order to develop the learner's sense of responsibility for the learning process. 3. The Dossier: to keep samples of work so as to show evidence of language learning competencies and progress to the learner, teacher and others.

The guiding principles of the ELP include: it is the property of the learner; it values competence in a positive way; it promotes learning inside and outside the classroom; it takes a lifelong perspective on learning of languages; and it is based on the CEFR.

Using the ELP, language learners are expected to assess their language and intercultural skills, as well as their approaches to learning. However, it is not easy for anyone to become reflective one's learning or for teachers to assess learning processes and learning outcomes.

５．３．ELP（European Language Portfolio）の中の言語パスポート、言語学習履歴、資料集の役割と機能

伊東祐郎

CEFR（Common European Framework of Reference for Languages: Learning, teaching, assessment）とは

CEFR とは、欧州評議会の言語教育に対する理念に基づいて、ヨーロッパの言語教育のシラバス、カリキュラムのガイドライン、試験、教科書等の向上のために一般的基盤を与えることを目的としている（吉島・大橋他 2004)。それは、ヨーロッパの言語教育及び学習の場で共有される枠組みとなっている。CEFR はその基本理念の中に、複言語主義を軸として、学習者中心の言語教育、自律した学習者の育成、生涯を通しての言語学習、行動中心の考え方を推進しており、学習者を社会的な存在と見なす立場を取っている。

CEFR はさまざまな教育実践の内容を分かりやすく開示できる透明性と、教育実践の比較や連携を実現できる一貫性を特徴としている。具体的には、言語能力の熟達度を、A1 から C2 までの６段階で構成される共通参照レベルで提示し、例示的言語能力記述文を示している。これは、コミュニケーション言語活動、コミュニケーション方略、コミュニケーション言語能力の３つの視点に立脚していて、この枠組みを活用することによって教育関係者が自らの実践をふり返ることを可能にしている。CEFR は、言語教育と言語学習のあり方や、言語学習者や言語使用者の言語運用能力を具体的に明示することにより、言語教育をより広い視点から捉えられるようになっている。関係する者同士の「対話」と「内省」を促すツールとなっている点は見逃せない。

ELP（European Language Portfolio）とは

ELP は、言語能力の熟達度（language proficiency）をヨーロッパ共通の尺度で示し、多様な背景をもつ学習者の学習成果を、学習者のおかれた状況や、学習の必要性、また学習者自身の属性や性質、また特徴を必要に応じて記述、記録するためのツールである（Website for the EUROPEAN LANGUAGE Portfolio)。ELP は、以下の 3 つから構成されている。①「言語パスポート（Language Passport)」: 自己評価による言語能力と異言語・異文化体験履歴を簡略に示すもの。②「言語学習履歴（Language Biography)」: 言語学習の記録、言語能力自己評価チェックリスト、言語学習の目標設定などを記述する。③「資料集（Dossier)」:

学習に関係する資料や学習成果物、証明書等を保存する。

欧州内での人の流動性が高まる中で、個人の言語能力や資格を共通の尺度でわかりやすく示し、理解解釈が可能となるものが求められる。また、言語学習は生涯にわたって行われるもので、言語学習は人間発達や人格形成に少なからぬ影響を与えるものとして教育的機能の重要性を強調している。これまでの学習観から脱却し、人の社会生活での異言語・異文化体験を記録し、自分の言語学習状況を内省的に自己評価することの大切さもアピールしている。これは、欧州内で生活する人の中に存在する複数の言語や文化の価値を認めることは異文化間能力の育成につながり、自らの言語学習の過程を計画・モニター・評価することは、学習者の自律性を促進するものとして欧州市民の育成の推進が掲げられているところからも読み取れる。その意味で、ELP は、教師にとってはこれまでの教育実践を見直す契機となり、さらにカリキュラムやシラバス改善につながる可能性のあることも認識されている（Little 2002）。

ELP 導入の目的

ELP の導入に際しては、以下の二つが主な目的とされている。

(1)学習者が、学習段階のどのレベルにおいても、自己の言語的技能を多角的に伸ばしていることを認識し、学習動機が高められるようにする。

(2)学習者が習得した言語的技能、文化的技能の記録ができるものを提供する。

ELP の機能

ELP は、「言語パスポート（Language Passport）」、「言語学習履歴（Language Biography）」、「資料集（Dossier）」の 3 部構成を基本としている。また、ELP が担う機能には、「教育的機能（pedagogical function）」と「報告的機能（reporting function）」があり、以下のようにまとめられる。

＜教育的機能＞

・複言語主義、複文化主義を促進する

・ELP 所持者に言語学習過程をより分かりやすく示し、自律学習（learner autonomy）を育成する。

＜報告的機能＞

・ELP 所持者の言語学習経験、外国語の熟達度、到達度を具体的に示す。また、公的試験で与えられる言語に関する資格を補足する。

・学校教育内、学外両方の言語学習を記録する。

上記の目的を達成するために欧州評議会の多くの加盟国が積極的な導入を試みているが、欧州評議会が認定する ELP は、上記の二つの機能を担うものでなければならない。

ELP の構成

学習者は、異なる年齢や異なる環境において、さまざまな目的で言語を学習している。それゆえに、ヨーロッパ各地で開発されている ELP は、背景の異なる多様な学習者グループそれぞれに適したものとして、上記の二つの機能を担うものでなければならない。そして、「言語パスポート（Language Passport）」、「言語学習履歴（Language Biography）」、「資料集（Dossier）」の三つから構成されることが原則とされている。

言語パスポート（Language Passport）

言語パスポートは、ELP 所持者の言語に関する資格、CEFR 参照レベルを基盤とした自己評価表（self-assessment grid）に基づいた言語技能の熟達度、そして言語学習経験、異文化経験等を簡潔に記入するページである。

ELP 所持者の言語に関する能力が把握できるように、学習中あるいは学習済みの言語の自己評価を、CEFR 参照レベル（A1～C2 の 6 レベル）に基づき、技能別に記入できる。

技能は、聞くこと（listening）、読むこと（Reading）、やり取り（spoken interaction）、表現（spoken production）、書くこと（writing）の五つのコミュニケーション技能に分かれていて、学習言語の技能別レベルを、言語パスポートに記入することができる。

また、自己評価だけでなく、教師による評価、言語関係の学位や資格、教育関係機関による評価も追記することができる。一方、言語学習経験、異文化体験に関しては、学校における言語学習経験、それ以外の経験、職場における言語使用、その言語の母語話者との接触等も記録できる。

言語パスポートは、ELP 所持者の言語能力にかかわるあらゆる情報を、誰でもいつでも簡単に確認できるもので、まさに、言語能力を証明するパスポートの役割を担っている。15 歳以上の学習者に対しては、成人用標準パスポート（Standard Adult Passport）がある。

言語学習履歴（Language Biography）

言語学習履歴は、ELP 所持者が学習目標を設定し、自己の学習過程を観察し、重要な言

語学習、異文化経験を記入していくページである。

言語パスポートとの違いは、言語パスポートが言語コース開始時および終了時に記入するもので、ELP 所持者の言語技能の熟達度、言語学習、異文化体験を簡潔にまとめるものであるのに対し、言語学習履歴は、学習者が学習計画を立て、それを実行していく上で、自己の学習過程、学習進度を観察し、自己評価を行う際の助けとなる言語学習ダイアリーのようなものである。そして、学習期間中に毎週、毎月といった頻度の高い割で記入していくものである。

言語学習履歴は、内省学習を促進させることを使用目的の一つとしているので、教育的には ELP の中心的な役割を担っている。言語パスポートがある時点までの言語技能の熟達度、言語および異文化学習の記録を示す総括的評価（summative assessment）であるのに対し、言語学習履歴は現在進行中の言語技能の熟達度、言語および異文化学習の記録を示す形成的評価（formative assessment）である。したがって、言語学習履歴は、学習者が目標言語の各技能にかんして何ができるのかという自己評価を行う。自己評価は、CEFR 参照レベルの自己評価表に対応したチェックリストを使って行う。チェックリストは、自己評価表にある各レベルに対し、技能別に、達成すべき目標が能力記述文（descriptor）で詳細に書かれている。その記述は、CEFR 参照レベルに反映しているため、「～ができる」という例示的能力記述文（illustrative descriptor）となっている。

チェックリストは、各レベルの各技能に、いくつかの項目が用意されているので、学習者は、明確に自己の言語学習を省みることができる。そして、今の自分に必要なことは何かを考え、次の目標を立てていくことができる。自己の言語学習を計画、観察、評価することは、自律学習を育成、促進する上で非常に重要である。また、言語学習履歴では、学習者が学校教育内、学外の両方で得た言語的、異文化的知識を記入するページもある。これは、いくつかの言語および文化能力を伸ばし、複言語主義および複文化主義を促進することを狙っているものである。

資料集（Dossier）

資料集は、ELP 所持者が、自分自身の学習言語の熟達度を示す重要だと判断する資料をまとめて保管しておくものである。言語パスポート、言語学習履歴に記入してある言語学習、文化学習において達成したこと、経験したことの実際の記録を保管するためのものである。

具体的には、学習内容をまとめたもの、プロジェクトワーク、言語に関する資格の認定書、教師からのフィードバック等が対象となる。子どもたちにとっては、手作りの補助教材をまとめたスクラップブックのようなものとなりうる。また、ある学習者にとっては、

公的試験で問われる技能と関連したプロジェクトワークの結果なども対象となる。そして、成人の学習者にとっては、目標言語を使用し、実際の生活で達成できる能力を証明できるもの、例えば、手紙のサンプル、メモ、レポートのようなものも保管対象となる。話し言葉の能力を示したい場合は、カセットテープ、ビデオテープを資料集に入れることができる。

ELP の使用

ELP 使用者は大きく分けて、学習者、教育機関および教師、そして雇用主に分けることができる。具体的な使用実態は、それぞれの立場によって異なってくる。

学習者は、言語学習を記録することができる。言語学習履歴を記録することは、学習計画、学習過程の観察、学習成果の自己評価という一連の能力開発につながる。それは、自己の言語的技能および重要な異文化経験の記録を示す証拠にもなる。結果的には、総括的な言語学習の記録となる。そして、転校、進学、就職など人生の節目となる機会に、ELP を提示して自己の言語能力を示すことができる。

教育機関および教師にとっては、欧州評議会が制定する CEFR 参照レベルと照らし合わせ、その教育機関のコースおよび資格等をより明確なものにできる。また、新入生や転校生、留学生を受け入れる際、その学習者が学習言語を使用し、何ができるかということが分かるので、教育機関、教師にとってもレベル等の対応が容易になる。

雇用主にとっては、社員を採用する際に、ELP に記録されている情報からその候補者の言語的技能を把握、特定することができる。また、ある仕事に要求される言語的技能の有無を判定することも可能である。

アセスメントの最近の動向

最近、外国語教育における評価において、ポートフォリオを用いた評価法が注目されるようになってきた。その背景には、従来型のペーパーによるテストでは、出題形式による制約から実際の運用力が測定できないばかりか、テスト結果から得られる点数からは具体的なパフォーマンス力の解釈に限界のあることが指摘されている。それとともに、それまでのテスト中心の評価への批判も挙げられている。すなわち、テスト中心の評価では、学習者の本来の学びの結果が見えないばかりか、授業や学習を計画する際に必要なフィードバック情報が提供されないという実情もある。そこで、提唱され始めたのが「代替評価法（alternative assessment）」や「オーセンティック・アセスメント（authentic assessment）」と

してポートフォリオ評価が注目を集めている。「オーセンティック」とは,「本物の、真正の」を意味するが、社会生活の中で学習者が学習の到達度や動機づけ、態度を振り返る様々なやり方を用いた評価法のことを意味する（O'Malley & Valdez-Pierce，1996）。ポートフォリオは、学習者の成長や努力の結果としての達成度を表す集合体として捉えられており、学習への参加の記録を包括したものであると定義づけられる。学習者と教師が共同で展開する、学習データを継続的かつ時系列的に集めたもので、多様な学習者と学習の多様性に対応するための評価方法であると言われている。

ポートフォリオとは，もともと芸術家や写真家たちが自分の作品を入れて持ち歩いていたフォルダやファイルを意味していた。芸術家たちは、自分の作品の中から優れたものを選び出し、フォルダに放り込んでいた。教育においては、学習者の成果物や達成の証しとなるものを自己評価し、教師や他の学習者からのコメント、また自分のパフォーマンスのプロダクトを体系的に評価する方法として捉えられている。学習を社会的なものとして考えているのである。ポートフォリオは、学習者が自発的に学びの伸びや変容を多面的多角的、かつ長期的に評価し、新たな学びに生かすためのもので、反省と自律学習のサイクルの促進をねらいとしたものと言えよう。

残された課題

自律学習を促進するために開発され、導入された ELP であるが、実際の現場では運用面で解決すべき課題が出てきている。まず、従来の言語能力にかかわる情報収集と評価方法のあり方が異なるために、現場教師の間で戸惑いが生まれていることである。言語パスポートにおける自己評価については、学習者が自らの言語能力を過大評価してしまう者や、逆に過小評価してしまう者がいたりして、評価結果が必ずしも実情とあっているかどうか確信がもてないことである。ELP には報告機能が盛り込まれているが、学習者個人や教師によってこの機能に対する認識が様々で、報告機能そのものがある水準を満たしていないこともあって、疑問をなげかける声も聞かれる。これは、国境を越えて言語パスポートとして活用される場合には、その扱いや使用方法が国によって異なることが少なくなく、普及推進の取り組みが必要となっている。

また、勉学や就労にかかわる手続きの際や第三者が何らかの目的のために ELP を参照する場合、言語学習歴や資料集は、情報そのものが多すぎて活用し切れていない状況があることである。短時間に効率よく内容を把握するためには、わかりやすくまとめられた形式で提示することが期待されている。特に学習者の言語熟達度を明示するためには、CEFR の Can-Do Statements と照合してレベルを決めることになるが、正確な評価を行うには、学習者のみならず運用にかかわる教師や教育関係者に対する研修の必要性が出されている。

また、従来型の試験による評価とポートフォリオによる評価のそれぞれの結果の解釈において正当に評価してもらえるかどうか不安を感じる者もいて、さらなる推進活動が必要となっている。駒形（2008）は論文の中で、ELP そのものについて寄せられたフィードバックでは次の３つの改善点が望まれたことを報告している。

- 年齢層別に適応させたモデルの開発
- 国あるいは学校のカリキュラムに明確にリンクしたもの
- 情報や指示文は学習者の母語で明記

課題や解決すべき問題は残るものの、ELP は確実に普及の段階に入ってきている。欧州連合が 2004 年から実施している「ユーロパス (Europass)」は、ヨーロッパ市民が特に自分の居住地域を離れて就職や職業研修の機会を求める場合に、自分の持つ言語能力や職業能力の一覧を作成することを目指して開発された証明書である。ユーロパスは「ユーロパス履歴書 (Europass CV)」と「ユーロパス言語パスポート (Europass Language Passport)」の二部構成で、後者に ELP の言語パスポートが応用されている。欧州評議会はヨーロッパ言語検定協会 (Association of Language Testersin Europe, ALTE) およびヨーロッパ言語教育質的保証協会 (European Association for Quality Language Services, EAQUALS)と共同して、インターネット上で作成できる「ePortfolio」を展開し、欧州評議会の ELP ウェブサイト上で公開されているものである。

参考文献

Little, David（2002）"The Europena Language Portfolio: structure, origins, implementation and challenges", *Language Teaching*, 35, pp.182-189

O'Malley, J.M. & Valdez-Perce, L.（1996）*Authentic assessment for English language learners*. Addison-Wesley.

国際交流基金 (2009)『日本語教育スタンダード』（試行版）

小玉安恵・木山登茂子・有馬淳一 (2007)「外国人日本語教師教育へのポートフォリオ評価導入の試みー17 年度長期研修 B コース教授法クラスにおける実施報告ー」『日本語教育紀要 』３，国際交流基金、pp. 95-111.

駒形千夏 (2008)「ヨーロッパ言語ポートフォリオー開発と導入に関する一考察ー」『現代社会文化研究』42、新潟大学、pp. 63-79.

駒形千夏 (2009)「ヨーロッパ言語ポートフォリオにおける言語バイオグラフィーの意義」『フランス文化研究』２、新潟大学大学院 現代社会文化研究科、pp. 119-131.

櫻井直子 (2010)「言語教育機関における CEFR 文脈化の意義－ベルギー成人教育機関での実践例からの考察」細川英雄・西山教行編『複言語・複文化主義とは何か－ヨーロッパの理念・状況から日本における受容・文脈化へ』くろしお出版、pp. 65-79.

柴原智代 (2007)「各国スタンダード作成の意義と日本の課題 ―ヨーロッパ，米国，オーストラリア及び中国，韓国の比較・分析―」『国際交流基金日本語教育紀要』3、pp. 113-122.

塩澤真季・石司えり・島田徳子 (2010)「言語能力の熟達度を表す Can-do 記述の分析－JF Can-do 作成のためのガイドライン策定に向けて－」『国際交流基金日本語教育紀要』6、pp. 23-39.

島田徳子 (2010)「国際交流基金レポート (8) JF 日本語教育スタンダード (第 2 回) JF 日本語教育スタンダードの内容と活用方法」『日本語学』, 29 (8) , pp. 76-91.

島田めぐみ・三枝令子・野口裕之(2006)「日本語 Can-do-statements を利用した言語行動記述の試み：日本語能力試験受験者を対象として」『世界の日本語教育』16、pp. 75-88.

島田めぐみ (2010) 「自己評価 Can-do statements に関する一考察－客観テストとの比較を通して－」『東京学芸大学紀要 総合教育科学系 II 』61、pp. 267 - 277.

田中和美 (2007)「ヨーロッパの現状とイングランドの例－学習基準と文化・連結・コミュニティ」『日本語教育』日本語教育学会、133、pp. 5-10.

根岸雅史（2008）「CEFR の日本人学習者への適用可能性」『応用言語学研究 : 明海大学大学院応用言語学研究科紀要 』10、pp. 45-54.

真嶋潤子（2007）「言語教育における到達度評価制度に向けて－CEFR を利用した大阪外国語大学の試み」『間谷論集』１、大阪外国語大学日本語日本文化教育研究会、pp. 3-27.

真嶋潤子(2008)「ヨー ロッパにおける移民への言語施策と Common European Framework of Reference (CEFR) に基づく自国語教育－フランス・デンマーク・イギリス・ドイツ・オランダ・オーストリア・アイルランドとカナダのケベック州を中心に－」『平成 19 年度文化庁委嘱事業 生活者としての外国人のためのモジュール型カリキュラムの開発と学習ツールの作成』コミュニカ学院発行、pp. 75-91.

真嶋潤子 (2010a)「日本の言語教育における「ヨーロッパ言語共通参照枠 (CEFR)」と「能力記述(can do statement)」の影響－応用可能性に関する一考察」マリア・ガブリエラ シュミット他編『日本と諸外国の言語教育における Can‐Do 評価－ヨーロッパ言語共通参照枠(CEFR)の適用 』朝日出版社、pp. 49-65.

真嶋潤子 (2010b)「CEFR における評価とアセスメント」佐藤慎司・熊谷由理編 『アセスメントと日本語教育－新しい評価の理論と実践－』くろしお出版、pp.19-43.

森本由佳子、塩澤真季、小松知子、石司えり、島田徳子 (2011)「コミュニケーション言語活動の熟達度を表す JF Can-do の作成と評価 －CEFR の A2･B1 レベルに基づいて－」『国際交流基金 日本語教育紀要』7、pp. 25-42.

山本弘子（2008）「日本語学校から見た評価の観点の見直しーヨーロッパ共通参照枠の視点からー」『日本語教育』136 号、pp. 38-48.
横溝紳一郎 (2000)「ポートフォリオ評価と日本語教育」『日本語教育』107 号、pp. 105-114.
ヨーロッパ日本語教師会（2005）『ヨーロッパにおける日本語教育と Common European Framework of Reference for Languages』国際交流基金.
吉島茂 (2007)「ヨーロッパの外国語教育を支える考え方ー複言語・複文化主義・行動主義、４つの Savoirs、部分的能力、ELP（Can Do Statement）」ELEC『英語展望』増刊号、pp.47-53.
吉島茂・大橋理枝他（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』朝日出版社.
萬美保 (2009)「言語共通参照枠」を参考にしたプログラムスタンダードの構築-香港大学日本研究学科必修日本語カリキュラムの例」萬美保・村上史展編『グローバル化社会の日本語教育と日本文化』ひつじ書房、pp. 72-94.

参考・引用ウェブサイト

Website for Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR）：http://www.coe.int/t/dg4/linguistic/cadre_en.asp
Website for the EUROPEAN LANGUAGE Portfolio：http://www.coe.int/t/dg4/education/elp/

5.4. Means and methods of standard setting between CEFR and tests

Kenji Ohtomo

Abstract

This article reports on the earlier literature on standard setting in relation to the Common European Framework of Reference for Languages. With the aim of exploring present problems with aligning tests with the CEFR, attention is focused on three areas of standard setting in language testing discussed in the literature.

The first area recognizes key issues which have to be considered with the visual representation of procedures of relating examinations to the CEFR. The issues are as follows: (1) familiarization with CEFR, (2) specification, (3) standard setting, (4) validation. Second, the structure of the Can-do Statements (CDS) need to be carefully investigated in order to avoid confusion with the construction of standard setting and language proficiency scale development. Third, in order to retain sound scale development for particular tests and the CEFR, attention must be paid to the results of alternative research, such as the "ranking approach" developed by a number of researchers in Europe.

These three points of discussion will, it is believed, lead to new international explorations on standard setting in relation to the CEFR.

５．４．CEFR と担当するテストの比較：その手段と方法

大友賢二

CFER との連係：その意味

CEFR 以外のテストを実施している者の中には、われわれのテストはこの CEFR と比較すると、どうなっているのだろうか、という疑問を持つものも多い。そうした状況にあって、それでは、どんなことに注目して、どんな事を検討したら、より適切な比較が可能なのだろうか。ここでは、CEFR と担当するテストの比較、それを行う場合検討しておかなければならない重要な課題は何かを考えることとする。

CEFR との比較検討を行っているテストのデータは、多くみられる。開発初期の段階ではないが、たとえば、2004年の時点で、すでに、この CEFR と TOEFL や TOEIC との比較検討はなされ、その結果も発表されている。Tannenbaum, R.J. &Wylie, E.C. (2004 : 15) で示されている次のような結果は、きわめて興味深い。CEFR では、英語能力の尺度を basic user (A1, A2), independent user (B1, B2), proficient user (C1, C2) と６段階に設定している。この中のＢ1，Ｃ1の段階に該当するのは、Test of Spoken English では 45，55 ; Test of Written English では 4.5, 5.5 ；（ paper-based)TOEFL では 457, 560 ; TOEIC では 550, 880 となっている。

また、（財）日本英語検定協会と CEFR との関連に関しては、Dunlea, J.(2009) での英検各級の内容と CEFR の各レベルを比較した結果は、次のようになっている。英検１級は CEFR の C1，英検準１級は CEFR の B2, 英検２級は CEFR の B1, 英検準２級は CEFR の A2, そして、英検3，4，5級は CEFR の A1 に該当するとして、発表している。

CEFR と CEFR 以外のテストを比較検討する際に、注意しなければならないことは、検討結果の十分なデータの裏づけのある検証に立った結果を利用することである。その検証手順のひとつとしては、Council of Europe (2009), がある。

Council of Europe (2009) の Language Policy Division, Strasbourg によって出版されている *Relating Language Examinations to the CEFR of Reference for Languages: Learning, Teaching, Assessment : A Manual* の 15 頁では、その Figure 2.2: Visual Representation of Procedures to Relate Examinations to the CEFR が示してある。ここに示されている手順は、CEFR との関連性を検証するために作成されたもので、その中のどれかひとつが決め手になるわけではなく、最終的には、すべての観点から関連性を検証するデータを集める必要があると言われている。その検討すべき事項は、（１）Familiarisation with CEFR, (2) Specification, (3)

Standard Setting, (4) Validation であり、それぞれのデータ―で検証されなければならない。（１）は、CEFR のことについてよく理解しておくこと。（２）では、自分のテストの内容を検討し、CEFR の能力説明文（descriptor）の内容とレベルを比較検討すること。（３）は、規準設定と呼ばれるテスト分割点の設定を試み、その検討をすること。（４）は、両者のテストの妥当性を検討することである。この中では特に、（３）standard setting においては、Standardisation of judgments, Judgment session, Establishing cut-off scores、（４）validation では、test validity, standard setting validity に注目しなければならない。

CEFR との連係：その開発（１）

CEFR と CEFR 以外のテストを比較検討するのに役立つ論文：North, B. (2010)には、The purpose of the CEFR, The common reference levels, Relating language examinations to the CEFR—the purpose of a Manual, The approach adopted, など CEFR と比較検討するテストで検討しなければならない重要な視点が示されている。この中で、さきに述べた４つの視点：familiarisation, specification, standardization, empirical validation：を再度説明しているが、その最も重要な validity と consistency and stability を、次のように述べている。

> Therefore, before an examination can be linked to an external framework like the CEFR (external validity), it must demonstrate the validity of the construct, and the consistency and stability of the examination (internal validity).(p.8)

CEFR との比較検討手順で必要とされる視点としては、以上であるが、我が国における最近の英語教育の現場では、Can-do statements の話題は絶えることがない。しかし、その中の要素に関する考察は、やはり不十分である。Can-do statements の価値を高める為には、その Can-do statements の要素を十分検討することが必要であろう。この課題を開拓するためのひとつの道は、North, B. (2000) で示されている次のような要素の分類である。

例えば、「適切に話される話を聞いて理解できる」が Can-do statements で示されている descriptor の一つとする。しかし、これでは、評価基準が不明であり、これを改善し、評価規準を明確にするには、何が必要であるかを検討しなければならない。たとえば、これは、「はっきりとゆっくり話してもらえば、自分の周りでの料理の話は普通に理解することができる。」とした場合の要素を考えると、以下のようになる。

（条件）「はっきりとゆっくり」、（場面）「自分の周りでの」、（対象）「料理の話」、（活動）「普通に理解できる」という要素が含まれていれば、評価のデータとしては、必要条件を満たしていると考えられる。North は、これを、condition, setting, object, action の４つの

要素としている。この要素が含まれているのであれば、一定の評価は可能であろう。こうした４つの要素は、言語能力を‘*ability –in –language user –in –context*‘ の視点から捉えている Chalhoub-Deville, M. & Deville, C. (2006) の考えと相通ずるものがある。こうした視点は、受験者のアンケートのみを基盤とした Can-do statements や descriptor から生まれる問題点を解消させてくれるものと期待している。

CEFR との連係：その開発（２）

CEFR との連係を考える場合に、過去において CEFR 作成の過程で行われた数多くの議論の跡を究明することは、極めて重要であり、かつ意味のあることである。現在、わが国で最も広く利用されている参考文献は、１９７１年来続けられた研究作業の最新の成果を示す Cambridge University Press から出版された Council of Europe (2001), *Common European Framework od Reference for Languages: Learning, Teaching, Assessment* である。また、これの日本語訳は、(訳・編)吉島　茂/大橋理枝（ほか）(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』朝日出版社である。

この中で、2001 の出版以降、多くの関係者によって、この内容が議論されてきているが、その議論の跡は、MANUAL に見ることができる。筆者の手持ちでは、September 2003 に出版されている　Council of Europe (2003), *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching Assessment (CEFR), Manual: Preliminary Pilot Version* と、もう一冊は、January 2009 に出版されている、 Council of Europe (2009), *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching Assessment (CEFR),A Manual* がある。

この中で、2003 に出版された Pilot Version の内容は、Introduction, The Common European Framework, Familiarisation, Specification, Standardisation of Judgement, Empirical Validation, Guidelines for Reporting である。また、その「はしがき」Preface to the Pilot Version に示されている、Charles Alderson 教授に関する次の一節は、きわめて興味あるものである。

> This seminar had been organized as a response to the complex issue which member states and examination/ certification bodies have to address, and which was succinctly summarized by Professor J. Charles Alderson as:**“How do I know that my Level B1 is your Level B1?”** (ix)

CEFR との連係：その発展（１）

すでに述べた、Council of Europe (2009) のなかの Figure 2.1. Validity Evidence of Linkage of Examination/Test Results to the CEFR (p.8) は linking process の大枠を示すものであるが、これをさらに、明確に示すために、Council of Europe (2009) では、以下のような図： Visual Representation of Procedures to Relate Examinations to the CEFR を示している。この図は、多くの場所で、多くの研究者に示されているので、内容としては、あえてここに示すこととする。原文では、英語で示されているが、その内容をより明確に多くの読者に理解していただくよう、日本語で、示すことにする。

図１　担当するテストが CEFR とどのように関係しているかを検討するための手順に関する説明

（検討のための論点）

（CEFR の精通）

（テスト様式）　　**（規準設定）**　　**（妥当性検討）**

テストの質の説明と分析
＊テスト内容の概要
＊テスト開発の手順
＊採点、評価、結果
＊テスト分析とテスト後の検討

判断の規準
ＣＥＦＲの水準に関する訓練
ＣＥＦＲの項目に関する項目困難度判断の訓練

判断の訓練
担当テストの行動例とＣＥＦＲ水準との検討
担当項目をＣＥＦＲ水準に一致させる作業

テスト妥当性
＊内容妥当性
＊予備テストの作業諸相
＊心理測定の諸相

規準設定妥当性
＊手続関連妥当性
＊内的妥当性
＊外的妥当性

ＣＥＦＲとの関連についての説明

分割点の設定

＊テストの水準に関する推定
＊測定されている意思伝達活動
＊テストされている意思伝達言語能力の諸相
＊テストとＣＥＦＲの関係を示す図表

CEFR との連係：その発展（２）

CEFR との比較を目的とした研究は、年々その数を増している。最近の研究の跡は、Waldemar Martyniuk (Ed.)(2010), *Aligning Tests with the CEFR: Refections on using the Council of Europe's draft Manual*, Cambridge University Press に見出すことができる。この文献の基盤

になっているのは、２００７年１２月に University of Cambridge で行われた２日の Colloquium である。それは、２００３年に公開された *Manual for Relating Language Examinations to the Common European Framework of Reference for Languages (CEFR)* の Preliminary Pilot Version の最終検討に関連する会議である。内容は、 Linking single tests to the CEFR, Linking a suite of exams to the CEFR, Large-scale multilingual assessment frameworks at national level の３部で論ぜられている。

CEFR との比較検討はわが国の多くの期間などで行われてきているが、その課題の多くは、３つ挙げることができる。第 1 は、CEFR との比較検討法である。第 2 は、Can-Do statements の構造に関する課題である。大方の Can-Do statements は、受験者の自己評価に基づいて作られているが、それは、受験者の実際の能力をどの程度表わしているかという課題である。そして、第３は、CEFR で設定されている６段階の水準との比較検討である、「尺度設定」（scale development） と「規準設定」（standard setting)であろう。

第 1 の CEFR との比較検討法は Figure 2.2.「担当するテストが CEFR とどのように関係しているかを検討するための手順に関する説明」で述べたとおりである。

第２は、can-do statements の構造に関する課題である。できるだけ簡潔にという方針があるけれども、簡潔すぎて、その行動が達成できたのかどうかが、不明であるという傾向は多くの場合、見出すことができる。設定された行動は、達成されたのかどうかという大きな課題に関しては、それは、達成目標ではなく、「向上目標」や「体験目標」であるので、議論の対象として考える必要ではないのではないかという議論もある。教育目標を達成目標、向上目標、体験目標分類する見方は、梶田叡一（2005）『教育評価（第２版補訂版）』有斐閣、に見出すことができるが、こうした考えにどう対応するのかも、残された課題である。これにたいしては、さきにあげている Brian North (2000), The Development of a Common Framework Scale of Language Proficiency, Peter Lang 等の Can-Do statements の４つの要素とどう対処するのであろうか？この課題は、塩澤、石司、島田(2010)、「言語能力の熟達度を表す Can-do 記述の分析」『日本語教育紀要 第６号』との関連も考察しなければならない。

残された課題（１）

CEFR との連係に関する研究で残された課題をもう一度、整理してみることにする。これまでは、課題 1. CEFR との比較検討法、課題 2. Can-do statements の構造について述べてきている。以下、課題 2 の後半として、Can-do statements の妥当性検討について考えてみることとする。

Can-do statements の作成の基盤となる自己評価と実際の能力との関係に関しても、多く

の検討がなされてきているのは事実である。わが国で開発されている言語テストの中には、Can-do statements 作成のための基盤となっている受験者の自己評価と実際の能力との検討が必要であろうとの声は大きい。そのなかで、どのようにしたら、自己評価と Can-do statements と実際の能力との関係を検討することができるのかも、残された課題の一つである。たしかに、この課題は、「テストの得点がよければ何かができる、というのではなく、何かができるようになった人は、テストで何点とれるということなのである。」「英語で何ができるようになれば、Ｂ１が取れる確率が高くなる」という言語能力の運用目標の記述に関する適切性も問われているのが現状であろう。この課題についての議論は、たとえば、Takanori Sato (2010), Validation of the Eiken Can-do Statements as a Self-assessment Measure Using Rasch Measurement, *JLTA Journal No.13*、または、Steven Ross (1998), Self-assessment in second language testing: a merta-analysis and analysis of experimental factors, *Language Testing,* Arnold, 1998, 15(1) などをもとに、さらに究明を必要とする課題であろう。さらに、川畑（2012)、自己評価データに基づく can-do 尺度構成法の改善、『新領域融合プロジェクト：人間・社会システム　データ―中心人間・社会科学の創生』情報・システム研究機構　新領域融合研究センター、などへの注目も必要であろう。

第３の課題は、CEFR で設定されている６段階の水準との比較検討である、「尺度設定」（scale development） と「規準設定」（standard setting)であろう。この尺度設定に関しては大きな課題であるが、それになかの規準設定の課題に関しては、先の課題１．（１）「海外における規準設定法の研究とその動向」で述べてあるので、ここでは、尺度設定にその課題の範囲を限定することとする。

残された課題（２）

CEFR との比較や連係を考える場合、その検討には、この CEFR の尺度設定は必ず必要であり、その設定方法は見逃すことのできない要素である。ここでは、Council of Europe (2001) の Appendix　のなかの Scale Deveplopment Methodologies　に関して、その概略を述べ、わが国に残された課題として、今後の発展を願うものである。CEFR の尺度開発の方法に関しては、さまざまな手段が取られているが、大きく分類すれば、（１）直観的方法（intuitive method)（２）質的内容的方法（qualitative method）（３）計量的方法(quantitative method) ３つがあげられる。

このうち、（２）と（３）の方法がとられた場合は、それを開始する手段として２つの方法が考えられる。そのひとつは、「能力記述文」(descriptor)から始める場合、もう一つは、「言語行動」（performance)から始める場合が考えられている。「能力記述文」の準備では、その草稿、収集、編集が必要である。それによって、尺度の質的、内容的要素を検討する

準備を行うことになる。具体的には、以下の作業分類番号の（No.4）, (No.9)がその例にあたる。「言語行動」の例から始める場合は、典型的な言語行動の例から始めるのがよい。
その行動を評価するにはどんな点に注目すればよいかを検討することになる。以下の作業分類番号では（No.5）（No.6）(No.7) (No.8) がこれにあたる。（No.10）（No.11）は、試験官に行動を評価させ、統計的分析を行って、その特徴を設定していくものである。また、（No.12）は、数理的に、能力記述文を尺度化する手段として、テスト理論の一つ：項目応答理論を使うというものである。

上に述べた作業分類の柱は、以下のように示すことができると言われている。

Intuitive Methods

No.1 Expert　　No.2 Committee　　No.3 Experiential

Qualitative Methods

No.4	Key concept: formulation	No.5	Key concepts: performances
No.6	Primary trait	No.7	Binary decision
No.8	Comparative judgments	No.9	Sorting tasks Quantitative Methods

Quantitative Methods

No.10	Discriminant analysis	No.11	Multidimentional scaling
No.12	Item Response Theory (IRT)		

こうした「尺度設定」のさらなる検討は、特にわが国では必要であり、残された大きな課題の一つである。この課題は、descriptor と最も関係が深く、scale と scaling descriptors の関係を究明する貴重なデータは、Brian North and Gunther Schneider (1998), Scaling descriptors for language proficiency scales, *Language Testing,* 1998, 15(2), Arnold, や、Branley, T. (2005), A Rank-Ordering Method for Equating Test by Expert Judgment がある。 最新のものとしては、Neil Jones (2009), A comparative approach to constructing a multilingual proficiency framework: constraining the role of standard setting, *Cambridge ESOL: Research Notes: Issues 37.* に見いだすことができる。また、Bechger, Kuijper and Maris (2009), Standard Setting in Relation to the Common European Framework of Reference for Languages: The Case of the State Examination of Dutch as a Second Language, *Language Assessment Quarterly*, 6 (2) に対する検討：Fulcher, G. (2010), *Practical Language Testing*, Hodder Education は興味ある参考論文である。

参考文献

Bechger, Kuijper and Maris (2009), Standard Setting in Relation to the Common European Framework of Reference for Languages: The Case of the State Examination of Dutch as a Second Language, *Language Assessment Quarterly*, 6 (2), 126-150.

Bramley, T. (2005). A Rank-Ordering Method for Equating Test by Expert Judgment, *Journal of Applied Measurement*. 6/2. 202 -223.

Brian North (2000), *The Development of a Common Framework Scale of Language Proficiency,* Peter Lang.

Brian North and Gunther Schneider (1998), Scaling descriptors for language proficiency scales, *Language Testing,* 1998, 15(2). 217-263.

Chalhoub-Deville, M. & Deville, C. (2006). Old, Borrowed and New Thoughts in Second Language Testing. In Brennan, R.(ed.)(2006), *Educational Measurement (Fourth Edition)*, (pp.517-530) . NCME & ACE.

Council of Europe (2001), *Common European Framework of Reference for Languages: Learning, Teaching, Assessment,* Cambridge University Press

Council of Europe (2003), *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching Assessment (CEF), A Manual, Preliminary Pilot Version, (ix).* Language Policy Division, Strasbourg

Council of Europe (2004), *Reference Supplement to the Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching Assessment (CEF), A Manual, Preliminary Pilot Version,* Language Policy Division, Strasbourg

Council of Europe (2009) *Relating Language Examinations to the Common European framework of Reference for Languages: Learning, Teaching, Assessment (CEFR): A Manual* ,(p.8, p.15). Language Policy Division, Strasbourg

Dunlea, J. (2009), The EIKEN Can-do List: improving feedback for an English proficiency test in Japan. In Taylor, L. and Weir, C.J. (eds.) *Language Testing Matters,* (pp. 245-262). Cambridge University Press

Dunlea, J. (2009),「英検と CEFR との関係について：研究プロジェクト報告」『EIKEN-Times 2009 特集』

Fulcher, G. (2010), *Practical Language Testing*, (pp.246-248). Hodder Education

Neil Jones (2009), A comparative approach to constructing a multilingual proficiency framework: constraining the role of standard setting, *Cambridge ESOL: Research Notes: Issues 37.* 6-9.

North, B. (2000) *The Development of a Common Framework Scale of Language Proficiency*, Peter Lang.

North, B. (2010), Introduction: The manual for *Relating language examinations to the Common European Framework of Reference for Languages* in the context of the Council of Europe's

work on language education. (pp.1-17). In Martyniuk, W. (ed.) *Aligning Tests with the CEFR*, Cambridge University Press

Steven Ross (1998), Self-assessment in second language testing: a merta-analysis and analysis of experimental factors, *Language Testing,* 1998, 15(1) .12-20.

Takanori Sato (2010), Validation of the Eiken Can-do Statements as a Self-assessment Measure Using Rasch Measurement, *JLTA Journal No.13,* 1-20.

Tannenbaum, R.J. &Wylie, E.C. (2004), *Mapping test scores onto the Common European Framework: Setting standards of language proficiency on the Test of English as a Foreign Language (TOEFL), The Test of Spoken English (TSE), The Test of Written English (TWE), and The Test of English for International Communication (TOEIC),* (p.15). Educational Testing Service

Waldemar Martyniuk (Ed.) (2010), *Aligning Tests with the CEFR: Reflections on using the Council of Europe's draft Manual*, Cmbridge University Press

塩澤真季、石司えり、島田徳子(2010)、「言語能力の熟達度を表す Can-do 記述の分析」『日本語教育紀要　第６号』 (pp. 23-39). 国際交流基金 .

梶田叡一（2005）『教育評価（第２版補訂版）』(p.82). 有斐閣 .

吉島　茂 / 大橋理枝（ほか）(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 .

川畑一光（2012)、自己評価データに基づく can-do 尺度構成法の改善、『新領域融合プロジェクト：人間・社会システム　データ―中心人間・社会科学の創生』(pp.158-163).情報・システム研究機構　新領域融合研究センター.

おわりに

この報告書は、財団法人 日本英語検定協会　英語教育研究センターとの委託研究契約に基づいて行われた平成２３年度研究課題「言語テストの規準設定」に関する研究成果をまとめたものである。研究期間は、平成２３年４月１日から平成２４年３月３１日の１か年で、研究員は、代表　大友賢二（筑波大学名誉教授)、副代表　渡部良典（上智大学教授)、伊東祐郎（東京外国語大学教授)、法月　健（静岡産業大学教授)、藤田智子（東海大学教授）の５名である。その内容としては、１．規準設定の意味と歴史、２．内容言語統合型学習における規準設定、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と規準設定という５本の角度から、それぞれ言語テストという視点で検討したものである。

時間的な理由で、この報告書では触れることができなかった研究も少なくない。それぞれの研究者が執筆を開始した後に国内外で行われた研究結果がそれである。例えば、言語テスト関係での国際的会議では、ILTA（International Language Testing Association）主催の LTRC（Language Testing Research Colloquium ）がある。University of Michigan での LTRC2011、Princeton の ETS が中心で準備された LTRC2012 がある。また、桃山学院大学で行われた JLTA（Japan Langugage Testing Association）2011 がある。さらには、大学入試センター主催の「2011 国際シンポジウム　教育テストの可能性」（２０１１年１１月１８日（有楽町朝日ホール))、東京外国語大学科研費研究チーム（代表者：投野由紀夫）主催の「新しい英語能力到達度指標 CEFR-J 公開シンポジウム」（２０１２年３月９－１０日（明治大学))」などは、注目に値する動向のひとつであろう。

CEFR に関連した研究のひとつ：Council of Europe (2006). *Plurilingual Education in Europe* などは、その副題にあるように、50 Years of international co-peration という長い歴史を背景としている。その流れは、わが国の外国語教育にも大きな影響を与えてきているが、この「規準設定」という課題は、現在最も注目されている研究分野の一つである。本報告書は、その課題に関する研究の現状、また、いま残された重要な課題は何であるかを究明しようとしたものである。先行研究のひとつとしてご覧いただければこの上ない幸いである。この先行研究に見られる残された課題を基盤とした実験・研究が重ねられて、規準設定の新しい開発が行われることを切に願っているものである。

２０１２年3月３１日

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
２０１２年3月３１日

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
