論文 / 著書情報
Article / Book Information

題目(和文)	大学英語教育プログラムの客観的評価手法の開発に関する研究
Title(English)	Study of a New Model of Program Evaluation for University Language Programs Using Objective Methods
著者(和文)	藤田智子
Author(English)	Tomoko Fujita
出典(和文)	学位:博士（学術）, 学位授与機関:東京工業大学, 報告番号:甲第9908号, 授与年月日:2015年3月26日, 学位の種別:課程博士, 審査員:前川　眞一,中川　正宣,室田　真男,中山　実,山元　啓史
Citation(English)	Degree:, Conferring organization: Tokyo Institute of Technology, Report number:甲第9908号, Conferred date:2015/3/26, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Type(English)	Doctoral Thesis

大学英語教育プログラムの
客観的評価手法の開発に関する研究

藤田智子

東京工業大学社会理工学研究科

人間行動システム専攻

2015 年 2 月

目次

第 2 章　日本の大学英語教育プログラム

第 3 章　Can-Do 自己チェックリストによる自己評価

第４章　自己評価のためのツールの作成

第５章　事前事後テスト

図目次

表目次

第 1 章　プログラム評価

1．1．はじめに

政府統計の e-Stat によると 2014 年度調査では、日本にある大学数は 781 校で、このほとんどに大小に関わらず英語教育部門があると考えられる。つまり、ほとんど全ての大学生は英語を履修しているということになる。また最近では大学でのグローバル人材育成が急務だと言われ、大学英語教育プログラムに対する関心は高い。しかし、これらのプログラムは各大学が自由な内容で運営し、文科省の指導要領のような統一された教育の基準は無く、実態は良く分からないのが現状である。今後、プログラム評価を実施して実施内容を確認し、改善を進めなければならない分野だと思われる。

しかし、欧米中心に発達したプログラム評価は、日本での導入はまだ進んでいない。体系的なプログラム評価を実施するには、複雑で様々な分野の専門知識に立った判断が長期的に必要となり、敬遠されて実施には至らないこともある。また、学校においてプログラム評価を実施する時、評価者が教員になる場合が多く、普段の業務に加えて評価者を担当することになった教員に過大な負担がかかることになる。さらに、授業の結果を数値化して優劣をつけるようなことをすると、職場のハーモニーが壊れる、というような風土が、一部ではまだ残っているらしい。このように、さまざまな要因が大学英語教育プログラム評価の実施を妨げているようである。

大学英語教育プログラムに携わる 1 人として、実際にプログラム評価を実施するとしたら、どのようなデータをどのくらい収集してプログラム評価のエビデンスとするべきか、なるべく多角的なエビデンスを数多く収集するべきであろうが、時間と人的リソースには限界がある。何をどこまで行えば、プログラム評価をしたことになるのかさえはっきりしない。例えば、スピーキング能力を伸ばすことを授業の到達目標としながら、その授業とあまり関連がない TOEIC テストを事前事後に実施して、その平均値の伸びを示すことがプログラム評価だと考えているのも稀ではないのが現実である。これらを総合すると、大学英語教育プログラムにおいてプログラム評価をもっと普及させるには、統一された客観的基準や評価手法を取り入れ体系的にプログラムを評価するためのモデルが必要なのではないかと考えた。

本論は、大学英語教育プログラムの評価について、教員たちが項目反応理論（IRT）の手法を駆使し、客観的な証拠を示して体系的にプログラム評価を実施するための新しいモデルを開発することをめざすものである。

1．2．プログラム評価とは

1. 2.1.　プログラム評価の定義

プログラム評価が対象とする「プログラム」とは、特定の社会的・教育的目標を達成するために、人が中心となって介入やアクションを行う事業のことである（安田、2010）。例えば、英語教育プログラムとは、この事業の内容が英語教育に限定されたものを言うことになる。また、このような数々の事業や活動を、プログラムという枠組みから捉えなおす必要があるのは、評価を行いやすくするためである。つまり、評価を体系的なものとして実施するためには、対象とする事業や活動を、評価対象として限定して捉えなければならないからである。

プログラム評価の研究者たちは（eg. Weiss, 1998; Rossi, Freeman, and Lipsey, 1999; Patton, 1997)、その著書でそれぞれに「プログラム評価」の定義をまとめている。その中でも、この分野で最も幅広く普及している Rossi, Freeman, and Lipsey (1999, p.4)によると、「プログラム評価とは、社会調査の方法を活用し、社会プログラムなどによる介入の効果を体系的に研究することである。」と言っている。これらのプログラムは、政策的・組織的な文脈・環境において用いられるものであり、社会状況を改善するためのソーシャル・アクションの情報源となるものである。また、渡辺（2000, p.147）は、「プログラムが、本来想定した目的をどの程度実現しているかを検証し、その結果を関係者に報告する一連の活動をいう」と定義している。これらを踏まえて安田・渡辺（2011, p.5）にまとめられた定義は、「特定の目的を持って設計・実施される様々なレベルの介入活動およびその機能についての体系的査定であり、その結果が当該介入活動や機能を付与するとともに、後の意思決定に有用な情報を収集・提示することを目的として行われる包括的な探求活動」である。

この定義からも「目的」は、プログラム評価にとって重要な出発点であると言える。評価の目的は、改善・発展のためなのか、説明責任（アカウンタビリティー）のための評価なのか、このように目的によって、次の手段を考える必要が生じる。さらに、目的に応じてエビデンス（客観的根拠）の提示が必要なのかどうかを考えることが次に繋がっていく。そして、プログラム評価の本質に関わることは、何を評価するか決めることである。Weiss (1998)は、評価の焦点は「プログラムの働きや機能」と「プログラムの結果と効果」の2つであると述べている。つまり、プログラムの働き（preparation）や機能（function)：プログラムの運営状況、参加率、どのような内容のプログラムなのかなど、そして、プログラムの結果(outcome)や効果(effectiveness)：プログラムに参加することによる効果などに着目することである。

1. 2.2.　プログラム評価と客観的エビデンスの提示

プログラム評価の結果として、「何となくその効果が上がっている」と結論づけた報告をしても誰も納得しないはずである。プログラムが確実に効果をあげているのなら、それを客観的に裏付ける証拠、つまり客観的エビデンスが求められる。その客観的エビデンスを提示するためにプログラム評価が必要とされているとも言える（安田、2010）。そこで、プログラム評価の手法が客観化されるためには、まず比較可能な統一基準の設定が必要となる。また、プログラム評価とは、突き詰めると理論的に良いとされる介入の内容が実際に良かったということをエビデンスとして示すことである（Weiss, 1998）。明確かつ客観的なエビデンスによって、あるプログラムの効果が実証されれば、同様のプログラムの作成や実施に関して試行錯誤して最善の方法を探す必要が低くなる。これによって実施者の負担やコストの削減につながり、効率が上がると言えるのである（Rossi et al., 1999）。

1．3．プログラム評価の実施

1. 3.1 評価可能性アセスメント

評価を行う前に、そのプログラムが評価可能かどうか検討するための包括的な査定をすることで、有用な評価を見極めることができる（Wholey, 2004）。これを評価可能性アセスメントと呼び、これによりプログラム評価をする前に評価ができるかどうか検討し、プログラム評価に投じる時間、経済、人的資源が無駄にならないようにする。評価可能性が高いプログラムとは、（1）プログラムゴールがしっかり定まっている。（2）介入手順・方法が明確で、安定性・整合性がとれている。（3）関連データの入手が容易である。（4）評価結果の利用目的がしっかりとしている。（5）プログラム実施現場と評価チームの意思疎通がうまくとれている。（6）評価への合意が得られている。などの条件が揃っている。これらを踏まえ、ここで収集すべき情報は、プログラム実施の目的、プログラムが達成しようとしているもの、ゴールに対して投入された人的経済的資源、実際に行われている活動、プログラムへの参加率やサービス利用率などがある。プログラムにうまく行っていない点があれば、どんな内容や種類のものか、これらを見極めることがまず第 1 歩である。

1. 3.2　評価クエスチョンとオーディエンス

評価を行う上での核となるべき評価クエスチョンを設定するのは、そのクエスチョンに関する情報を欲している相手であるオーディエンスに大きく依存していてこれを見極めるのが重要な

作業である（Rossi et al., 1999）。評価クエスチョンの数を設定するときは、評価者はデータ収集との兼ね合いを考えて適切な数のクエスチョンを設定しなければならない（Weiss, 1998）。必要以上に多くの評価クエスチョンを設定し、データを収集しようとすると、そのクエスチョンの数だけプログラムを実施している関係者や参加者への負担が増える。また、評価者も多くのクエスチョンがあると、そのぶん多くのデータを収集し、分析することを計画に入れる必要がある。そして、誰が何を知りたいのか、オーディエンスの見極めと評価結果をどのように使うかについて検討しなければならない。 例えば、オーディエンスがプログラムの運営スタッフである場合、その評価クエスチョンは自ら決まってくる。ターゲットに対してプログラムがしっかりと行き届いているか？プログラムへの参加者は満足しているか？プログラムは効率的に運営されているか？どのようにプログラムを改善できるか？などである。また、英語教育プログラム評価においては、教員がプログラム評価の運営者であり、オーディエンスとなることが多い。このような参加型評価においては、自分たちに都合の良い評価結果を誘導しないように、しっかりした評価項目やモデルを構築し客観的な評価に徹しなければならない。

1.3.3 プロセス評価

プロセス評価とは、プログラムが対象者に意図された通りに提供されているか、体系的に判断する作業のことである。もし、あるプログラムがうまく行かず悪い結果が出たとしても、そのプログラムがなぜ失敗したのか明確化できないことがある。これを「プログラムのブラックボックス化」と呼ぶが、このようにアウトカム評価の結果にどのようにして到達したのか分からない状況を避けるために必要なのが、プロセス評価である（安田、2010）。これは、対象となるプログラムの成熟度により異なるが、Chen (2005)によるプログラム評価の枠組みは以下の3つのフェーズからなる。

導入フェーズ：現実にはパイロットテストは行われず、見切り発車されることが多い。プログラムの実施・運営上の問題・課題にターゲットを絞り、それらを解決・改善し、より効果的なプログラムにしていく。形成的（Formative）評価と呼ばれる。

① プログラムがなぜ期待される結果や効果を生み出すことができるのか。
② プログラムの主な問題点は何で、それを解決していくためにはどれくらいの期間やコストがかかるのか。
③ プログラム効果を助長する要因は何か。
④ ステークホルダーは、プログラムのパフォーマンスについてのどんな情報を必要としているのか。

展開フェーズ:いかに効率的にプログラムの進行度合いのチェック(モニタリング)ができるかが大切となる。プログラム参加者の属性、基本情報、プログラムの提供時期や回数、参加率、参加者のニーズなどが必要。プロセス・ユース(process use)評価の実施を通じた評価者やステークホールダーなどのプログラムに対する考え方・態度・行動の変化、プログラム運営方法についての変化。

効果健在フェーズ: 次の段階であるアウトカム評価に進んで良いか判断し、当初の計画通りにプログラムが実施され肯定的な変化の兆しがあるかどうか評価する。

1.3.4 アウトカム評価

プログラムによって生じた利益や恩恵について、どれだけ効果をもたらすことができたか評価することがアウトカム評価である。アウトプットは、プログラム活動と直結しているが、アウトカムはそれとは異なり、プログラム参加者など対象となる人々への「影響」、つまり「プログラム介入」によって生じる「参加者の変化」のことを言う。この中には、目に見えず、把握するのが困難な構成概念を定義して測定する場合が多い。量的なアウトカム評価では、プログラムが説明(独立)変数が原因となり、アウトカムという目的(従属)変数を生み出したと考える。

アウトカム総計(gross outcome)は、評価デザインや測定方法の不具合まですべてを含むアウトカムのことを言う。これを式で表すと(1.1)式のようになる。

【アウトカム総計】=【実質効果】+【外生要因の総計】+【デザイン効果】　　(1.1)

デザイン効果:各種の統計的効果およびアウトカム指標の測定による影響

外生要因:外から影響を及ぼすこと。例)景気の動向など

1.4. 参加型評価

1.4.1 参加型評価とは

参加型評価とは、評価対象となるプログラムに関係する人たちが評価者や協力者となり、自らのプログラムを自己評価して修正改善することを目指す評価の総称である。三好・田中(2001)は、参加型評価とは、プログラムの参加者、受益者や実践者が専門家主体の評価(介入型)を代替えする評価概念であると述べている。また、Patton (1997, p.98-100)は、外部評価者が効果検証を行う介入型と、この参加型では問題の捉え方から評価のし方まであらゆる点で異なっていると述べている。参加型では評価への参加者が評価の考え方や技法を修得し、評価者で

はなく、参加者が重要だと思われるプロセスやアウトカムを中心に協働して評価を行う。外部より内部へのアカウンタビリティーが優先される。このように評価を参加者である英語教員が自らのプログラムを評価するのは、参加型評価と言える。このタイプの評価はどうしても自己評価的な要素が強くなる傾向があるが、これを防ぐにはできるだけ多角的なデータを使って、客観性を重視することが大変重要となる。

本論で対象とする大学英語教育プログラム評価モデルとアクションリサーチを比較してみると、カリキュラムや授業の改善を目指し、教員たちが協働的に自ら実施する点は共通する。しかし、プログラム評価モデルで対象とするのは、プログラムそのものであり、全学年や大学全体であるが、アクションリサーチでは、一人ひとりの学習者であることも多い。また、本論では、CEFR などの世界的外部基準などを取り入れて、日本の英語教育プログラム間で比較できる評価の共通基準を構築し普及させることを目指している。しかし、アクションリサーチは必ずしも他のプロジェクトと比較したり、統一化する必要はない。さらに、中村（2008）によれば、アクションリサーチの定義には、実践的な「研究」「研究活動」「研究にのぞむ態度」というように「研究」が必ず関連する。しかし、プログラム評価の定義（1. 2.1）には、「探究活動」「体系的な査定」のように「評価」そのものが「研究」と繋がるわけではない。

1.4.2　エンパワメント（empowerment）評価

参加型評価の代表的なものであるエンパワメント評価は、Fetterman ら（1996, 2001）により提唱されたプログラム評価の方法で、「当事者が自分のプログラムを計画し、実行し、評価する能力をもつことでプログラムが成果をあげることをめざす評価方法」と定義されている。もともとは、社会的に権利をはく奪されている状況からエンパワメントによって脱出させるプログラム評価のモデルとして知られるようになった。プログラム参加者が、プログラム評価について基本を学び、プログラムの理念やゴールに向けて、プログラムそのもの、その環境、実施内容などを改善していくことを目的とするものである。また、Wandersman (2003)　によれば、エンパワメント評価は特に発展のための評価に適している。

エンパワメント評価には5つの主要概念がある。training:　プログラム評価の専門知識を学び、評価能力を強化する、facilitation：評価者がステークホルダーとの潤滑油となり、プログラム評価が円滑に進むように努力する、advocacy:　評価の結果に関わらず、改善のための提言に持ち込む、illumination:　評価を通じて発見や経験をする、liberation:　自らのプログラムを新しい価値観に開放し自己決定能力に繋げる、などである。

さらに、Fetterman ら（1996, 2001）は、エンパワメント評価には4つのステップ（1. ミッションの構築、2. 実態把握、3. 今後のプラン策定、4. 戦略の構築と記録）を発表したが、その後 2005 年には以下の3つのステップに改められた（Fetterman and Wandersman, 2005）。

ステップ1：ミッションの構築では、ステークホルダーの意見も聞きプログラムの使命について確認する。そのプログラムが何を達成しようとしているのかを多角的に定義・構築する。

ステップ2：現状把握においては、プログラムの活動内容、実施状況、予算や人的などの資源などの実態を掴み、列挙されたそれらの情報に優先順位をつける。

ステップ3：将来計画は、実態把握を受けて、プログラムの理念やゴールにより近づけるための方策を検討する。

日本でエンパワメント評価を教育プログラムに適用した例は、いまだかつて無かったが、鎌田他（2012）が理科系留学生の小規模日本語プログラムに導入したのがはじめての例となった。

1. 4.3 エンパワメント評価への批判

Fetterman and Wandersman （2005）は、伝統的な評価と比べると、「エンパワメント評価は評価と呼べるのか」という批判があることを認めている。エンパワメントという概念は、社会正義の実現をめざすものなので、本当は「評価」ではなく「社会運動」ではないのかとの疑念を抱く人たちが居る。確かにエンパワメント評価の根底には、社会活動における「エンパワメント」の価値を高く評価する価値観がある。これは、プログラムの中に良い環境を整えること、人々により多くの権利を与えることが、すべての社会活動の改善に繋がるという考え方である。しかし、プログラムの中にエンパワメントのための方法論を取り入れたり、教員たちがエンパワメントによってプログラムの改善を目指しているのは、明らかに改善を目指す「評価」のためであって、社会運動そのものではない（鎌田他、2010）。

エンパワメント評価の画期的なところは、プログラムとプログラムの当事者を、評価の対象とし、同時に、評価の主体にしたという点にある。しかし、そこには「自己評価の客観性の確立」という越えなければならないハードルがある（鎌田他、2010）。Fetterman ら（1996, 2001）は、評価原理や手順をできる限り細かく定め、科学的なツールを用いることで客観性の保持を担保することを提唱している。しかし、エンパワメント評価は、基準を設定し比較可能な客観的なエビデンスを提示することに重点をおいた評価モデルではない。

1.5 プログラム評価とテスト理論

1. 5.1. テスト理論の活用

安田(2010, p. 12-14)は、プログラム評価とテスト理論の接点は非常に多く、テスト理論はプログラム評価に対して多くの貢献ができると考えている。特に客観的エビデンスの提示を目的としたプログラム評価への貢献は高く、その中核となる作業ではテスト理論における考え方や専門性を大いに生かせるとしている。

しかし、焦点の違いは存在していて、プログラム評価は現実性や実践的な課題を重視するのに対して、テスト理論は測定を重視する傾向がある。プログラム評価においては、測定の知識は必要要件であるが、十分条件ではない。プログラム評価を実行するには、ある程度の曖昧さを容認しなければいけない場面がある。たとえば、評価に於いては理想論の追求だけに偏らず、ある程度の段階で、評価実施に踏み切るような状況もあるはずだ。つまり、テスト理論の専門家がプログラム評価に関わる際には、測定の計画・実施に柔軟に適合させたような見方が要求される。プログラム評価には様々な専門性を持った人が関わることが多いが、これらの専門家から学び、さらに協調関係を築くことが、非常に貴重な経験となっていくと思われる。

具体的にテスト理論の専門性をどのようにプログラム評価に活かすことができるかは、(1)プログラム評価についてのエビデンスの提示に関して統計学的なサポートができる。(2)実験的アプローチを用いた評価デザインの構築、アウトカム評価における指標作成やアセスメント法の開発などデータ分析能力が活かされる。(3)プログラム評価についての教育(評価教育)を量・質ともに充実させる方向に導くことができる、というようにまとめられる。

1. 5.2 規準設定のための項目反応理論

規準設定は、教育プログラムでの履修者に関する「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く、規準設定の正確さこそが、「意思決定」の妥当性の確立の中核となる(Plake and 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。古典的テスト理論(CTT)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、CTT 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、CTT 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 CTT と比べ、項目反応理論(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。

1. 5.3 項目反応理論(IRT)と古典的テスト理論(CTT)によるテスト分析結果の比較

Culligan and 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、CTT 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。

IRT にはいくつかの前提や条件はあるが、テストの規準設定をする上で、CTT に比べ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。実際の教育現場において、IRT を一般に普及させ、教育の現場でより正確な数値的データの分析ができるようになる(Fujita, 2005)。

1.6. プログラム評価の妥当性

プログラム評価と、テスト理論はその妥当性についても重なる部分が大きい。共通点の一つは、プログラム評価を始める前に、なぜプログラム評価が必要なのか？何のため、誰のための評価なのかを明確に意識しなければならない点である。これは、テストを作成する前にそのテストの対象と目的を明らかにし、いかにその目的にかなったテストであるか検証することを前提とするテスト妥当性の理論と一致する。

Messick (1980, 1988)は、妥当性を1つの単一の概念として認識する場合、得点を基礎とするすべての推測の根本には構成概念妥当性に統一された得点の意味が存在するとしている。しかし、完全に統一された妥当性の見解に至るには、得点に基礎をおく推測の適切性、有意義性、有用性と、同時に社会的な価値や、社会的結果としての妥当性は一緒に考慮しなければならないと述べている。彼は、社会的な価値とともにテストスコアの機能的な価値を推定することで、テスト妥当性の意義を確立することができると指摘している。Messick (1989, 1994) は、テスト結果の使用と解釈についての妥当性の判断は、そのテストごとに結果の使用と解釈としての評価をしなければならないと言及している。Cronback (1988, 1989)は、妥当性についての議論は、概念、証拠、社会と個人の結果、そして価値、これらの関連を基盤に実行されなければならないと提唱している。テスト開発者にとって、実施されているテストが個人と機関にとって適切な結果をもたらしているか妥当性の確認をすることは、選ぶ余地のない義務であり、これと同様のことが、プログラム評価を実施する人たちにも当てはまると思われる。

Popham (1997)は、研究者の中には、テストから導き出されるスコアをもとにした推論の正確さを妥当性だと思っているが、それは教員たちが、ある特別な知識に関するスキルや修得が学習されたかを正確に推定しようとする傾向があるからだと述べている。彼は、教員たちはテストにのみ注目した妥当性というよりも、導き出された推論に焦点を当てた妥当性の重要性に気付くべきだということを強調している。このようにして Popham（1997）はテスト妥当性に関して、広く社会的な視野を持つことの重要性を強調している。そして、テストに比べプログラム評価は包括範囲が広く、テストはプログラム評価の一部である。プログラム評価においては、広く社会的な視野を持って評価することがさらに重要になる。

1．7. 英語教育に関するプログラム評価

1. 7.1 英語教育プログラム評価の変遷

初期の英語教育プログラム評価は、カリキュラムデザインや教育的な手法の中にどのような革新的なものがあるかということを、まるで万能薬でも探すように求めていた（Ross, 2003）。 これらの初期の英語教育プログラム評価法は、アウトカムに焦点を合わせたものであり、特別な条件の下でしか有効でない教育方法についての評価に限られる傾向があった。その為、その特別なケースでは成立しても他の場合には成立しないプログラム評価方法だという批判を受けていた。従って、一般化するには及ばず、英語教育プログラム評価研究の成果として顕著なものはあまり多く無かったと言える。しかし、その中で疑似実験的手法（quasi-experimental

evaluation method）が使われるようになったことだけでも意義があったと言われている（Alderson and Beretta, 1992; Lynch, 1996）。この疑似実験的な手法とは、計量心理学（psychometrics）的に統計処理をすることによって、プログラムの中で実験に参加したグループ（treatment group）と実験に参加しなかったグループ(control group)を作って、この2つのグループを比較することにより、実験の効果を統計的に推定する評価手法のことである。

このようなプログラムアウトカム集中型の評価が過剰な注目を浴びたあとは、1990 年代後半からは、もっとプロセスにも注目した評価を求める動きが隆盛になってきた。そして、treatment とcontrol の 2 つのグループを比べる実験的手法でプログラムの効果を求めるような手法では不充分であるという声が上がった（Rea-Dickens and Germaine, 1998; Kiely and Rea-Dickens, 2005）。なかでも、Lynch (1996) は、実験的・疑似実験的手法、計量心理学を用いた量的研究と、対局にあるように見える民俗誌学・記述人類学（Ethnography）や自然主義的研究デザイン（Naturalistic design）を用いた量的研究を合体させた言語教育プログラム評価を提唱した。この多角的な方面からのデータを分析してプログラム評価に取り入れる折衷主義的なアプローチにより、プログラムプロセスとアウトカムの両方についての評価の根拠が豊かに広がった。加えて、この形成的アプローチには、たとえ一つの誤ったデータがあったとしても、それを理由にプログラム評価全体が、決定的な欠陥評価になる危険性が薄らぎ、積み重ねられた幾つもの他の根拠によって補えるという安定感がある。

1. 7.2 普及しない日本の英語教育プログラム評価

日本では、プログラム評価の概念そのものが欧米に比べて相対的に理解されていない（金谷他編 2003; 山中 2007 p.23）それゆえに、プログラムを評価する基準が明確にはないのが実態である（山中、2007）。また、これまでの日本における教育評価は大学基準協会によるものなどがあるが、一貫性などの点で問題がある可能性があると言われていて、現状では教育成果による評価を体系的に行うには不便がある（串本、2006）。プログラム評価は、欧米では関心がはるかに高い。例えば”program evaluation”をインターネットで検索すると、アメリカ心理学会のデータベースでは 2000 件以上の関連文献がヒットするが、日本のデータベースでは 135 件であった（安田・渡部、2011）。

このようなプログラム評価に対する無関心が英語教育プログラム評価に対しても存在している。これではアカウンタビリティーが保証されない可能性があり、受益者である大学生の、大学選択の理由に繋がる可能性は低く、悪い英語プログラムの淘汰が起こることもないと思われる。

これは、国家の言語政策の問題点と言えるかもしれず、日本の英語教育が成果を上げていると断言できない原因となっている（山中、2007）。また、ほとんどの日本の大学では、プログラム評価の習慣が無く、知識が欠如している場合が多く、専門知識を有した担当者が不在というのが現実である（山中・鈴木、2006）。

しかしながら、今後英語教育におけるプログラム評価は、国家政策が掲げる言語教育の理念に対して各教育機関が果たすべきアカウンタビリティーとして、教育の「質の保証」を強調するのなら、当然のようにもっと実施されるべきなのである。そこで次に必要になることは、「英語教育プログラム評価」の専門性を持った人材確保と育成である。この専門性をもった人材は、単に統計学を専門としているだけでは不充分で、英語教育を適切に評価できるとは言えない（山中・鈴木、2006）。英語教育のパフォーマンスやアウトカムを知るには、統計学に通じているだけでは十分ではなく、英語教育学を専門とし、実際に教育を実践している人材が最も適任だと思われる。

１．８．日本の英語教育プログラム評価の実施に関する論点

プログラム評価の目的

プログラム評価のさまざまな目的は、評価を要請する関係者・利害関係を持つ人々であるステークホルダー（stakeholders）側の様々な要素にかかってくる。また、ステークホルダーたちも評価結果によって影響を受ける。この stakeholders のニーズを的確に評価に反映させることが重要なポイントである。考えられるプログラム評価の目的は、主に、（1）改善・発展のための評価（evaluation for development）、（2）アカウンタビリティーのための評価（evaluation for accountability）、（3）知識習得のための評価（evaluation for knowledge）、（4）価値判断と意思決定のための評価（evaluation for judgment）、（5）宣伝活動のための評価（evaluation for public relations）などがある（eg. Rossi et al., 1999; Weiss, 1998）。これに加え Bachman（1990）は、「評価：evaluation」にはプログラム自体に対する評価や改善のニュアンスがふくまれていて、評価とは本来はこのような視点を含んだ包括的なものであるべきであると述べている。

評価の実施者

三好・田中（2001）は、参加型評価を従来型の評価と比較して説明している。参加型評価とは、プログラム評価を外部に委ねる（従来型）とは異なり、プログラムの参加者、すなわち、受益者や実践者が専門家主体の評価を代替えする評価概念である。これを英語教育プログラム評

価で言えば、英語教員が自らのプログラムを自分たちの手で評価すること意味している。参加型評価は、外部介入型評価と比べ当然ながら自己評価的な要素が強くなる傾向がある。英語教員や履修学生のプログラムへの見方・考え方とプログラムに対する判断・意見が重視される傾向があるからだ。また、評価の焦点はプログラムのアウトカムに当てられる傾向があり、そこには参加者の見方や考え方が反映されがちとなるため、主観的な要素が強くなる。これを防ぐには、多様なデータを用い、評価の偏りがないようにすること、また、基準の客観性が重視することが必要である。Fetterman (2001)によると、参加型評価の１つである「エンパワメント評価」とは自己改善を促進させて、自己決定を促すものと定義している。この説では、自らのプログラムを自己評価することは、自己修正、すなわち改善をもたらすものとして肯定的に捉えていることが特徴である。エンパワメント評価においては評価者が参加者と協調して主体的に参加することで、より現実に則した実用的な評価が行われ、その結果として積極的に行動が伴うことが強調されている。

評価の重点

プログラムを評価する際、まず初めに考えなければならないことは、いったい何を評価するのか？という根幹的なことである。これは評価の重点を何に置くかという非常に重要な問題で、大きく分けると「プログラムの働きや機能」と「プログラムの結果と効果」になる（Weiss, 1998）。同じことを英語教育プログラムに適用すると、評価の重点は、「英語教育プログラム実施・運営の状況」と、「英語教育プログラムの効果」ということになる。

第 1 章で既に述べた山中（2007）が提唱した EPEU（English Program Evaluation of University）モデルでは、理論の評価⇒実施の評価⇒結果の評価という流れがあった。これは、プログラムの理論が念頭にあり、それが実現するための遂行方法としてカリキュラムがあり、そして遂行されたかどうかの成果がもたらされている、ということを表している。このモデルの中にあって、「理論の評価」が評価するに値するという結果であれば、次に進むことができる。「実施の評価」は、プログラム理論・理念のプログラムでの実現度を評価する部分である。また「結果の評価」は、その一部に外部試験（TOEFL, TOEIC, 英検など）の結果を取り入れることは、それらの妥当性が確保できる保証があってはじめて可能になると説明している。

量的研究アプローチ

プログラム評価の心理学的諸領域において実験的手法の基礎を築いた Campbell は、その

著書「Reforms as Experiments」(Campbell, 1969)のなかで実験的手法の重要さを説き、プログラム評価には、客観的根拠を示す必要があることを強調した（安田・渡辺、2011）。Campbell とその同僚たちが基礎を築いた分析方法は今も多くの量的研究の研究・分析方法として利用されている(Ross, 2003)。また、同じく量的研究アプローチの、その中でも言語教育評価学に大きなインパクトを与えた Popham (1978, 1981) は、最も関心を寄せるべき、プログラムアウトカムに対する考え方について見解を述べている。彼によると、そのプログラムの目的に対して履修者の能力がいかに変化したかを評価対象とするべきだと指摘している。つまり、その英語教育プログラムの目的が「リスニング and スピーキング能力が、CEFR の A レベルから B になる」という目的であれば、履修者のリスニング and スピーキング能力に的を絞り、その能力が CEFR の A から B レベルになったかどうか測れるテストを使って客観的根拠を提示して評価するべきという事である。

日本の大学英語教育プログラム評価での傾向は、TOEFL、TOEIC のスコアや英検の合格率を基準として教育の価値を決定しようとするところにあると言われている。実際に日本の大学が独自の英語教育プログラムについてまとめた文献が複数あるが、これらはいずれも正式なプログラム評価の手順を踏んだものではなく、その評価の仕方も一部の外部テストの事前・事後の比較であることが多く、断片的かつプログラムの到達目標が達成されたか評価されていない（山中、2007）。

外部テストと教員作成テスト

伊東（2008）は、指導と評価は一体であるべきで、評価基準や評価方法は、教育目標やそれを達成する教育内容や指導方法と表裏一体でなければならないと主張している。そして、評価によってプラスの波及効果がもたらされなければならないと述べている。このように指導と評価を表裏一体と見た場合、指導した内容を学生がどれくらい習得しているかを測定するための到達度テスト（achievement tests）が実施されるべきである（Brown, 1996）。さらに、そのプログラムで学生たちが学んだ内容をテスト領域として、そのプログラム内容を良く知る教員が作成したテストを使って事前事後の能力変化を推定し、そのプログラム評価の根拠の一つとすることにより妥当性が保たれる。また、Kiely and Rea-Dickins (2005) の著書 “Program Evaluation in Language Education” のなかの、教師主導の評価についての章では、教師作成テストは、（1）学生のニーズや望みを満たすことができる。（2）指導と学習の両方を交えて実施できる。（3）学生が自己評価をしながら学習ができる。（4）ある特定の内容に焦点を当てて評価できる。（5）

幾つかの領域にまたがって評価できる。このように多くの利点を並べている。さらに、(6) 学生のレベルに適応しやすい。(7)授業の到達目標に適合している。(8)その英語プログラムの日程やシステムの状況に合わせやすい。(9)市販のテストを購入するコストがかからない。などが挙げられる(e.g., 久保田, 2002; 斉田・小林・野口, 2009)。

しかし、教員作成テストによって、担当者となった教員の負担は計り知れないものがある。また、信頼性の高いテストを作成するにはさらに莫大な時間とエネルギーを要する。このような負担をかけずに、教員作成テストの利点を生かす方法でテストを実施することができることが求められている。

1．9. 大学英語教育の為の新プログラム評価モデル EEP-J(Evaluation of English Programs in Japan)の提案

1. 9.1. EPEU モデル(English Program Evaluation of University Model)

山中(2007, p22)は、日本ではプログラム評価の概念そのものが普及していない現状にあり、「英語教育のプログラムそのものに対する評価」を実行するための明確な基準が欠けていると述べている。その為に、TOEFL 等の市販テストの平均点など断片的かつ不適切なデータで、その英語教育プログラムの評価を短絡的に評価しようとする傾向があるのではないだろうかと推察している。彼は、プログラム評価において、評価される対象は学生だけではなく、プログラムの理念で、その理念を実現させるカリキュラムと遂行方法、そして遂行されたかどうかの成果であると主張している。

そして、この概念を大学英語教育におけるプログラム評価の具体的なモデルとして English Program Evaluation of University Model (EPEU)を提唱した。このモデルには3つのフレーム(理論の評価 ⇒ 実施の評価 ⇒ 結果の評価)があり、プログラムが生み出した「アウトカム(成果)」を可視化する評価を行うことをめざそうとしている。つまり、このモデルに従ってプログラムを評価することによって、各大学が、その英語教育政策における説明責任を果たす有力なツールとなることを目標としている。

理論の評価:学生、教員、関係者、ステークホルダーによるニーズアセスメントとその検討、理論・理念の評価とカリキュラムへの反映の実現度、プログラムに対する介入の戦略の検討を行う。

実施の評価:理論、理念のプログラム実現度、プログラムの実現度(プロセス評価)

結果の評価:アウトカム評価、インパクト評価、結果の普及

山中（2007）は、TOEFL, TOEIC のスコアや事前事後テストなどを指標としてアウトカムの基準に用いて英語教育プログラム全体を評価することは危険であると指摘している。その 1 つの根拠は、テストスコアは「測定」であるのに対して、「評価」は「測定」を包括し、さらに全人格的なより大きな概念として存在しているからだと説明している。金谷他編（2003）は、教育における「測定」と「評価」の違いを述べている。「測定」とは、能力や技能を量的・客観的に表すことであり、「評価」とは、一部ではなく全てを対象として、質的、主観的な判断やフィードバック機能や意思決定機能を含むものだと言っている。評価と測定の関係は、測定が数値的で客観的に量化が可能な領域であるのに対して、評価は測定を包括し、さらに主観的、全体の要素を付け加えた、より大きな概念であると主張している。さらに、Bachman (1990, p.18 - 24) は、その著書で measurement, test , and evaluation について言及している。測定（measurement）が情報提供機能を持つのに対し、評価（evaluation）は、プログラムに関する意思決定のために情報提供機能を持つ。彼は、評価をプログラムに関する意思決定のために情報を組織的に収集することと位置づけている。そして、必ずしも測定やテストが評価とは限らず、テストの結果が意思決定の基礎として用いられる時だけ評価が関わると指摘している。言わば、「評価」は1～2回のテストで測定された結果のように単純なものと同等に扱われるべきではないと言っているのである。これらを取り入れて、山中（2007）の EPEU モデルでは、英語教育における「評価」の定義を「テストを含む「測定」による客観的かつ質的な判断を含む。」と限定している。

そのプログラムで学ぶものが TOEFL や TOEIC などのテストを事前事後テストとする場合、授業の内容も TOEFL や TOEIC テストに連動したものでなくてはならない。プログラム評価における事前事後テストは、そのプログラムで学んだことが、履修者にどれくらい定着したかを測らなければいけない。

1. 9.2.　EPEU とエンパワメント評価モデル

日本の大学英語教育プログラム評価のために独自に作成された山中（2007）の EPEU モデルには3つのフレーム（理論の評価　⇒　実施の評価　⇒　結果の評価）があり、プログラムが生み出した「アウトカム（成果）」を数値的な結果のみにこだわることなく可視化する評価を行うことをめざそうとしている。つまり、EPEU モデルは各大学の英語教育プログラムの教育政策について説明責任を果たす有力なツールとなることを目指している。

EPEU モデルの第1フレームである「理論の評価」について山中（2007）は、英語教育のプログ

ラム評価において、評価するべきものは、プログラムのゴールや目標であり、これらが失敗している場合は、そのプログラムは評価するに値しない可能性が高いと指摘している。確かに、英語教育プログラムに於いてその目指すゴールや目標が「理論上の失敗」であった場合は、次の実施の評価や結果の評価に進んだとしても、あまり成果はないことになる。

EPEU モデルは、説明責任を果たす有力なツールとなることを目指し、第3フレーム：結果の評価で終了している。エンパワメントモデルのように、改善計画を第３ステップとしておらず、第１ステップに戻りサイクルを繰り返す循環型モデルであることは特に強調されていない。

Fettermanら（1996, 2001, 2005）が提唱したエンパワメント評価（1. 4.2）の特徴は、当事者が自分のプログラムを計画し、実行し、評価する能力をもつことでプログラムが成果をあげることをめざすことであり、参加型評価の代表的なモデルである。本論で提案したい英語教育プログラムのための評価モデルは、英語教員が自らのプログラムを評価することを前提としている点では「参加型評価モデル」であると言える。これは自己評価的な要素が大きいプログラム評価方法であるため、客観性を欠かないように留意する必要があるが、自己評価的評価モデルの特徴とも言える「修正」のしやすさを備えている。エンパワメント評価モデルの 3 つのステップのうち、ステップ３：将来計画（Planning for the future）は、プログラム参加者が、プログラム評価の実態を学び、プログラムの理念やゴールに向けて、プログラムそのもの、その環境、実施内容などを改善していくことを目指すエンパワメント評価の真髄であり、「改善のためのプログラム評価」と呼ばれる所以でもある。これは、本論で提案したい、「英語教員が自分たちのプログラムの実態を自ら評価し、改善を目指す」というプログラム評価の目的と非常に良く適合する。いかに人的・物質的・時間的資源を傾けてプログラム評価を実施しても、その評価結果の報告を知るだけで終了すると、結果を受けて改善に向かう新たな進展は期待できない。そこで、エンパワメント評価のステップ３である「将来計画」を踏襲して、「改善計画：改善に向けての示唆と実施」を新プログラム評価モデルに取り入れることは重要である。

1. 9.3.　新しいプログラム評価モデル EEP-J（Evaluation of English Programs in Japan）

大学英語教育プログラムのための評価モデルである EPEU モデルは、各大学がその英語プログラムに対する説明責任を果たすために開発されたものである。エンパワメント評価のようにそのプログラムの英語教員たち自らが英語教育プログラムそのものを評価することを強調しているものではなかった。また、様々な角度からのプログラム評価のためのエビデンスを求めることは想定してはいるが、その中には具体的手段として学生の声を聞く機会である「学生の自己評価」

に関する言及はなく、重視しているようではない。そして、市販テスト（TOEFL, TOEIC 等）は、そのプログラムうの授業内容と関連が強くない限り、プログラム評価のエビデンスとはならないとしているが、テストに代わる体系的・客観的なエビデンスを何にするべきかを具体的に提示していない。

これらのプログラム評価に関する先行研究や先行モデルを参照し、客観的評価手法を中心にした大学英語教育プログラムの新しい評価モデル：Evaluation of English Programs in Japan（EEP-J）を提案する（図 1.1）。このプログラム評価モデルの特徴は、Can-Do 自己チェックリストによる学生自己評価と事前事後テストを利用した教師による評価という、2つの異なる角度からの客観的エビデンスを追求してプログラムアウトカム（効果）を評価する点である。この評価モデルによって、教員が学生に自己修正能力を意識させながら、プログラムの改善に結び付けていくことができることを期待する。本評価モデルには、次の 3 つの段階がある。

実施内容確認

プログラム理念やゴールの実現度、プログラムの実現度。実施内容が、プログラム理念・ゴールが達成できるように計画され・実施されているか確認する。ここでは、授業の到達目標・評価基準として用いられている Can-Do statement (CDS) が、プログラムで学ぶ学習内容、学習者の習熟度レベルとズレがないかなどを確認する。またニーズ分析をして、学習者や教員が求めるニーズが反映できているかを教員間で話し合い、協調しながら次の段階へ進めていく。

アウトカム

結果の評価のためのデータとしては、数値的な量的データと質的データの両方があることが望ましい。そしてそれらのデータは異なった角度からのエビデンスであるべきである。データの 1 つは、（1）そのプログラムを履修する学生に適合した CDS を設定し、それを授業の到達目標、評価基準などにする。その CDS を自己評価の形にした Can-Do 自己チェックリストを実施して学生に回答してもらい、学生側から出てきたデータとする。もう一つは教員が作成した（2）事前事後テストの結果データである。これは学生の英語能力の伸びを教員が作成したテストで測定した「教師による評価データ」である。この段階では、これらの2つの異なった角度からのデータを、できる限り正確性の高い、客観的なエビデンスにすることに注力する。

改善計画

Can-Do 自己チェックリストを実施・分析して変更すべき部分が見つかれば、CDS を改訂する。事前事後テストの IRT 分析の結果から、習熟度レベルごとの反応を調べ、他のレベルと比べ能力値 θ の伸びが非常に低いレベルがあれば、その原因を分析して対応を考える。評価結果か

ら、必要に応じて教材や教授法の変更をするためのアクションを起こす。「改善計画」の段階に達したら、最初のステップである「実施内容確認」へ立ち返り、例えば、改訂した Can-Do 自己チェックリストが実態にあっているか確認する。このサイクルはプログラムが実施されている限り繰り返されるべきである。

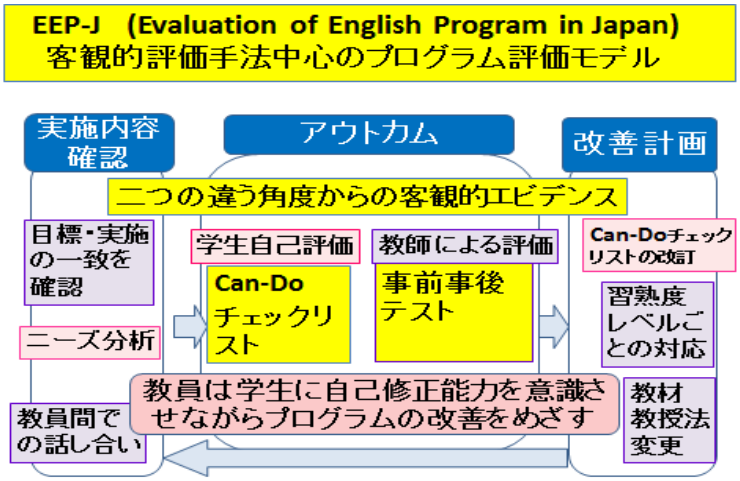


図　1．1　　大学英語教育プログラム評価モデル（EEP-J モデル）

1．10．本論の目的

1．10.1　　異なる2つの角度からの客観的エビデンスの提示

学生自己評価

本論では、まず第 2 章において、本論の背景である日本の大学英語教育プログラムの最近の傾向を紹介する。ここでは言語レベルの世界基準である CEFR をはじめとする Can-Do statement（CDS）に着目する。その英語教育プログラムの学生たちが「英語でできる能力」を文章の形で表したものを Can-Do statement (CDS)と呼び、これを自己評価の形にしたものを Can-Do 自己チェックリスト（今後 SCL と略す）と言う。このように CDS をカリキュラムの根幹、つまり授業の到達目標、評価基準、習熟度別クラスのレベルとして導入することによって、英語授業の流れの中で、学生が自己修正へのきっかけを得る機会とするためである。SCL の背景に CEFR のような世界的に認められた外部基準を取り入れることによって、SCL の信頼性を高めることが可能となる。しかし、ヨーロッパの言語学習者のために開発された CDS を日本人学習者向きに、さらにはその英語教育プログラムの対象学生向きにカスタマイズしていくことによって、その学習者に合った自己評価ができるようになる。Can-Do 自己チェックリストによる自己評価の

結果を IRT で分析することで、学生の実態により適した SCL に改訂することができると思われる。

(1)本論の目的1は、IRT を利用して Can-Do 自己チェックリストを、客観的で正確性の高い自己評価ツールにする方法を探ることである。

事前事後テスト

教員が作成した客観的テスト(リスニングやリーディングの筆記テスト)による事前事後テストを実施し、その結果を IRT を用いて分析するシステムを構築する。第 5 章では、IRT を用いて共通の尺度を作ることで、教員に大きな負担をかけず、かつテスト結果の有用性を高める例を示す。IRT によって事前テストと事後テストが同一項目でないテストであっても、受験者の能力が比較可能となる。また、事後テストは事前テスよりはるかに少人数の受験者数であっても比較が可能となった。また IRT によって項目分析済みの過去問題を使用したため、新項目を数多く作成することなくテストが実施できた。そして、事前事後テストを等化するとき、どの等化法が最も適切であるか確認する。**(2)本論の目的2は、IRT を利用して、教員作成の事前事後客観テストシステムを構築する方法を明らかにすることである。**

1. 10.2　大学英語教育プログラムにおける新しい評価モデル:EEP-J の提案

大学英語教育プログラム評価において、教員たちが、自らのプログラムを自分たちの手で評価する(参加型プログラム評価)場合、自己評価的な要素が非常に強くなり、主観的な評価になってしまう危険がある。これを回避するためには、できるかぎり多角的、かつ客観的なデータを使って、教員たちが自分自身の手でエビデンスを提示しなければならない。本論の研究目的である Can-Do 自己チェックリスト①と事前事後客観テスト②を組み込むことで、本論 1. 9.3 に示した大学英語教育プログラムにおける EEP-J モデル(図 1.1 参照)を開発する。

(3)本論の目的3は、大学英語教育プログラムにおける新しいプログラム評価モデル(EEP-J)を開発し、学生が自己修正能力を高めることを認識しながら、教員はプログラムの改善をめざす。そしてその他大学での利用の可能性を示すことである。

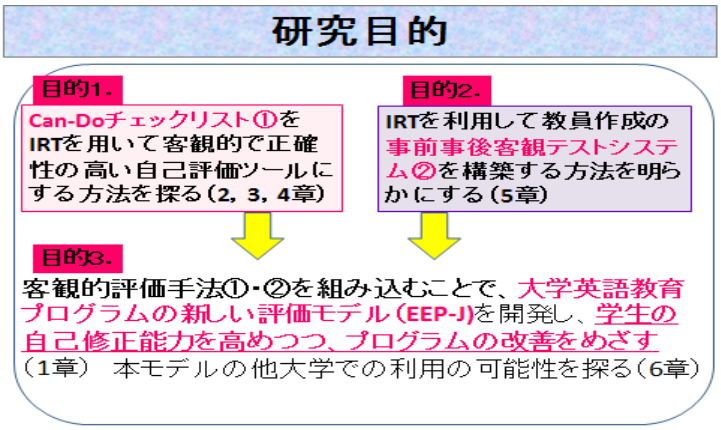


図 1.2　3つの研究目的

1. 10.3　本研究の特色と意義

欧米に比べ、日本ではプログラム評価の概念が普及していない。そのため、「英語教育のプログラムそのものに対する評価」を実行するための明確な基準がない（山中、2007）。日本の大学生のほとんどが履修している大学英語教育プログラムの質の保証は、国の政策として進めるべきインパクトのある問題であるにも関わらず、対策が打たれた形跡はない（山中・鈴木、2006）。プログラム全体を体系的に査定するために、客観的で明確な基準を持った評価モデルを導入して普及させるべきである。しかし、日本の英語教育のために構築されたプログラム評価モデルは、今まで EPEU モデルしかなく、このモデルの英語教育プログラムでの実践も 1 例のみであった。また、エンパワメント評価モデルにおいても、他の分野も含め、日本で教育プログラムに適用して体系的に実施したのは1例にとどまる（鎌田他、2010, p.89; 2012. P.45）。従って、日本の大学英語教育プログラムにおいて、EEP-J 評価モデルの導入を提案することの意義は大きいと考える。

１．１１．　本論の構成

本論の構成は図 1.3 のように 7 つの章より成っている。

図　１．３　本論文の構成

第 1 章では、プログラム評価とは何かについて述べ、それをさらに大学英語教育プログラム評価へ限定して説明する。そして、大学英語教育プログラムのための新しい評価モデル（EEP-J）を提案する。

第 2 章では、本論の背景となる日本の大学英語教育プログラムについて述べる。Can-Do statement（CDS）とは、到達目標や評価基準をテストスコアのような数値ではなく、現実の状況の中でコミュニケーションとして何ができるか記述したものである。その中でも Common European Framework （CEFR:ヨーロッパ共通参照枠）は、20 年以上の研究を経て今では世界で最も普及している CDS である。CEFR を日本の大学英語教育に導入する取組と、これを日本人学習者に適用させるための研究についてまとめる。

第 3 章では、ある大学英語教育プログラムにおいて事前事後 Can-Do 自己チェックリストによる自己評価を行い比較した。IRT を使って学習者たちの英語の能力値（θ）が事前事後でどのように変化したか推定する。また、教員が作成した SCL と学習者の回答から推定した項目困難度の違いを調査し、学習者の声を取り入れた SCL にするための試みについて調査する。

第 4 章では、より正確に自己評価を測定できる Can-Do 自己チェックリストの質問順序効果について、どのような項目順にすると順序効果を減らし、使いやすい SCL を作成できるか調査する。

第 5 章では、ある大学英語教育プログラムで事前事後リスニングテストを実施し、そのデータを、項目反応理論（IRT）を用いて分析するとき、どのモデルと等化法が最適であるのか、10 通りの組み合わせの中から判定を行った。適切な等化法をシミュレーションで判定する先行研究は存在するが、（1）2つのテストの受験者数が、事前が約 6000 人で事後が約 700 人というような大幅に異なる受験者数で実施した研究はほとんどない。（2）等化法の判定基準として θ を用いている。（3）実データを使用している。これら（1）（2）（3）の条件で最適等化法のシミュレーションを実施した点で新規性があると言える。

第 6 章では、日本の大学で英語教育プログラムを担当する教員 40 名に対して大学英語教育プログラム評価についてのアンケートを実施した。この結果により、本論で提案する大学英語教育のための EEP-J モデルが他大学で利用される可能性があるか考察する。

第 7 章では、研究目的に対応する結論1，2，3をまとめ、本論で得られたプログラム評価についての知見を概要し、今後の展望を示す。

第2章　日本の大学英語教育プログラム

２．１．はじめに

日本の大学では、English as a foreign language (EFL)として、ほとんどが母国語を日本語とする学習者たちが外国語として英語を学んでいる。従来、日本の大学では「知識としての英語修得」「大学の教室の中だけでの英語教育プログラム」であることが多かった。そこには統一の到達度目標や評価基準が無く、習熟度別クラス編成もされていないことも珍しくない。しかし、2000 年代後半頃から、Common European Framework (CEFR：ヨーロッパ共通参照枠)に代表されるCan-Do statement（CDS）を言語教育の授業に取り入れる動きが活発になり始めた。CDSとは、テストスコアなどの数値による能力の点数化ではなく、実際の状況の中で「コミュニケーションとして何ができるかを質的に記述したものである。例えば、その大学英語教育プログラムで学ぶ学生に、「英語で何ができるのか」を記述したCDSを作成し、それをプログラムの到達度目標、評価や習熟度別クラスレベルの基準として導入して、システマティックな教育プログラムを運営するところが出現し始めた。そして、これらのプログラムの目指すところは、生涯学習を前提とする自律的学習者の養成である。従来は教室の中だけの英語教育であったものが、大きく変貌を遂げつつある。しかし、変革を行うには従来型に留まるより、はるかに大きな負担が、特に教員たちにのしかかることになる。そして、その英語教育プログラムの関係者たちは、費やした負担に見合った効果が出ているのか確かめなければ変革の続行を望まなくなるであろう。客観的なエビデンスによる体系的プログラム評価を、現在最も必要としているのは、特に、このCDSを基盤にした最近型の大学英語教育プログラムである。本論で扱うCan-Do 自己チェックリスト（SCL）はCDSを自己評価の形にしたもので、最近型の大学英語教育プログラム評価の客観的エビデンスとして利用することをめざしている。

２．２．従来型と最近の英語教育プログラム

日本の英語教育プログラムは、徐々に進化を遂げている。いわゆる従来型と呼ばれる英語教育プログラムは、知識の習得としての言語学習であり、文法や単語の暗記に始まり、長文読解や和訳が、典型的な授業内容であった。これらは、大学受験を目的とする「受験英語」とか「入試の為の英語」と呼ばれ、中学高校時代に「受験英語」を経験した大学生が、違和感なく学ぶのが、このタイプの英語教育プログラムである。そして従来型の場合、到達度目標、評価基準

等が無く、習熟度別クラス編成も実施されないことも多い。また、プログラムゴールが設定されていないことも稀ではなく、言語を使用する場所として想定しているのは、主に教室内である。

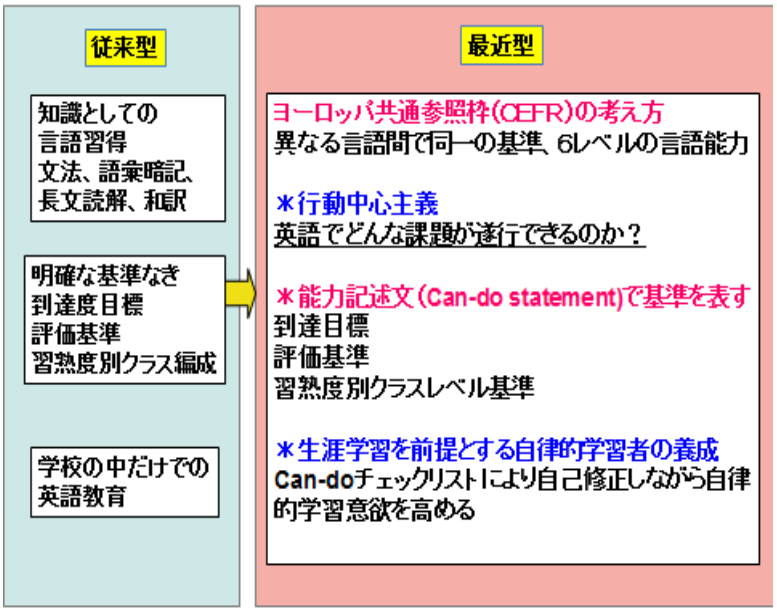


図2．1　従来型と最近の英語教育プログラムの比較

しかし、最近の英語教育プログラムは、ヨーロッパ共通参照枠：Common European Framework(CEFR)のポリシーの一つである、「その言語でどんな課題が遂行できるのか？」という行動中心主義の考え方を取り入れて、実践的な授業を展開しているところも増えてきた。このような最近型では、到達度目標、評価基準、習熟度別クラスレベルなどが CEFR に代表される能力記述文（Can-Do statement：CDS）で統一された基準などとリンクして表現される。また、プログラムゴールも CEFR の目指す「生涯学習を前提とする自律的学習者の養成」であり、Can-Do statement を自己評価の形にした Can-Do 自己チェックリストにより、自己の効力感や自律的学習意欲を高めることを目標としている（図 2.1）。

2．3．ヨーロッパ共通参照枠

ヨーロッパ共通参照枠（CEFR）の目的は、もともとはヨーロッパ言語教育の内容、試験、教材などの向上のために一般的・基本的なガイドラインを与えることである。そして、学習者の習熟度レベルを能力記述文（CDS）として明確に表し、それぞれの習熟度レベルの段階で生涯を通して学習の進度が自律的に分かるように考えられている（吉島・大橋　2008: p1）。この参照

枠のなかには、様々なポリシーが包括されている。そのなかでも、複言語主義と行動中心主義は根幹をなすものである。

言語は文化の主要な側面であるばかりではない。例えば、ある個人には、それまでに接した様々な文化が比較・対比されながらさかんに作用しあって、統合された複分化能力（pluricultural competence）を形成している。複言語主義とは、この能力の中の一部として複言語能力（plurilingual competence）が存在することを言う。そして、行動中心主義とは、言語の使用者と学習者を基本的に「社会的に行動する者・社会的存在」とみなすことである。従って、一定の条件や特定の行動領域の中で、認知的、感情的、意志的、社会的存在としての個々人の能力を考慮することになる（吉島・大橋　2008: p9）。

これらの考え方をもとに、多くの学習者を対象にデータを収集し、分析して尺度化したものが図 2.2 にある6つのレベルである。習熟度が低いレベルから、基礎段階の言語使用者（Aレベル）は、その中でも低いレベルのA1と高いレベルのA2に分かれている。真ん中のレベルである自律した言語使用者（Bレベル）も同様にB1とB2に分かれる。さらに、最も高いレベルである熟達した言語使用者（Cレベル）もC1とC2に分かれ、これらすべて合わせると 6 レベルとなる。CEFR

は、主な言語使用のカテゴリーと6つのレベルをそれぞれ縦軸と横軸にした表の形で示すことができ、それぞれのレベルに基づいた自己評価の表に連結している。この参照枠は、実際的な言語教育の目的・方針・事情に合わせてより詳細な一覧を、その教育現場に応じて作ることが必要となってくる。このように現状に合わせた参照枠があれば、学習者は自分たちの言語技能のおおよその状態が分かり、SCL を使って習熟度レベルを自己評価するようになる。

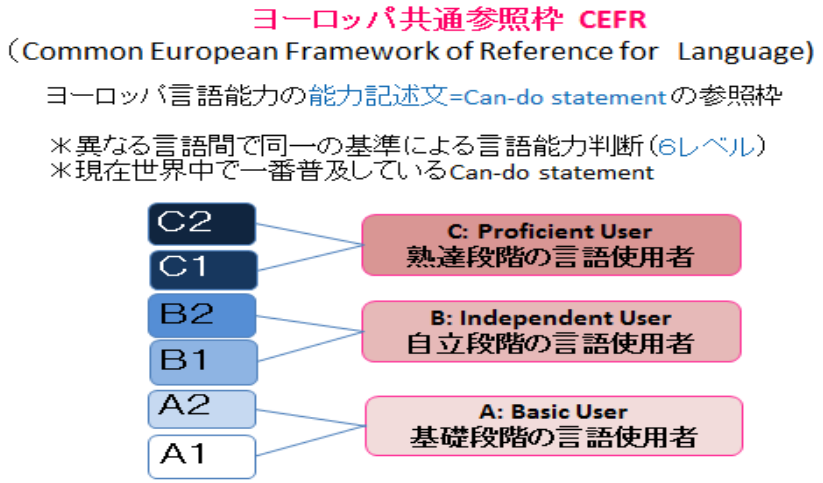


図2．2　ヨーロッパ共通参照枠

２．４．CDS による規準設定：2つの異なった目的

CDS には、Common European Framework (CEFR) (Council of Europe, 2001)や、英検、GTEC for STUDENTS, TOEFL, TOEIC などがそれぞれの規準で、「どのようなテスト結果を得た学習者は何ができる」という Can-Do statements (CDS)を設定している（テストスコアの解釈規準としての CDS）。そしてまた、European Language Portfolio (ELP)のように学習者が自己評価として自分の英語能力を診断し、また教員も学習者のレベルを判断する手段として利用可能な自己評価としての CDS（Can-Do 自己チェックリスト）がある。

2.4.1 CDS の妥当性

しかし、これらの CDS の妥当性の検証は充分に実施されていると言って良いのであろうか。Weir (2005)は、もっと慎重に CEFR の妥当性検証を行い、多言語共通参照枠として完成度をより高いものにするべきだと述べている。また彼は、CDS はそれを使用する国ごと、さらに教育機関ごと、言語カリキュラムごと、テストごとに、その学習者や受験者に適した CDS として誂える（Tailor made）必要があると主張している。例えば、文化や言語環境が異なるヨーロッパの言語学習者のために作られた枠組みである CEFR を日本の言語学習者にそのまま適用させるには無理があり、変更や工夫をする必要がある。CEFR の枠組みを参照してもらい、その言語学習の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）。そこで、これを日本人に適用した CDS にする必要性が強調されている（境、2009；根岸、2006a）。しかし、これも日本人に適用することだけでは十分ではなく、日本人学習者の中にも子供、大人、学生、社会人など、より細かく識別して、本来は、そこで学ぶ学習者に対応した CDS を作成するべきである。

2.4.2 テスト解釈と自己評価

このように妥当性が高く、その英語教育プログラムの履修者の英語能力に可能な限り適応した CDS を作成する試みが行われているが、その典型的なものが、第 3 章で説明する項目反応理論（IRT）を用いて困難度パラメタを推定し、CDS の規準設定をする方法である。North and Schneider (1998)、Sato (2010)、筒井、近藤、and 中野 (2007) は、CDS を自己評価のツールとして、あるいは学習者のレベルを判断する教師評価の手段として実施し、IRTを用いた分析を行ってその妥当性を確認した。これらの研究は主に Can-Do 自己チェックリストの結果を、IRT1 パラメタまたは 2 パラメタモデルを利用して、各 CDS の困難度パラメタを推定して難易度の

規準設定の目安にする方式を採用している。今後、妥当性の高い CDS を日本の言語教育の現場に普及させるにあたって、その重要なカギとなるのは、充分に多くの事例研究を実施して、その英語教育プログラムにできるかぎり適応した CDS の設定を追求することである。Green (2010) も、研究者、教員と学習者が実際に使っている教材や言語運用の実践的な例を持ち寄って、より妥当な CDS のレベルの設定のために意見交換し、積極的に協力し合うことの重要性を強調している。

2.4. 3 テストスコアの解釈のための CDS

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

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

次に DIALANG は、CEFR をもとにした言語能力診断をオンラインで実行できるように開発された言語能力テストである（Alderson and Huhta, 2005）。ヨーロッパの 14 言語に対応でき、受験者がどの言語のテストを受けるか選択できるようになっている。はじめに、どの言語でテストを

受けるか決め、そのあと語彙テストを受けるかどうか、自己評価をするかどうかなどは、受験者が決めることができる。次いでリーディング、リスニング、ライティングの能力テストを受ける。

DIALANG は、言語能力を測定することだけを目的にしているわけでなく、言語能力診断をして今後の学習に役立てるために開発されたものである。またテストの結果が、素点ではなく受験者が CEFR の A1～C2 のどのレベルに該当するかで判定されるのも特徴である。そして受験者が自己評価をすることにより、自己のテスト得点と自己評価の相関を知ることもできる。結果レポートには、自分のそれぞれの技能が CEFR のどのレベルであるか判定されたものと、そのレベルの学習者は典型的にどういうことができるかを通知してもらえる欄がある。これは CEFR のポリシーである、「学習者が自律的に自己修正しながら学習を進めることのサポート」を提供することに対応している。DIALANG で判定する CEFR　A1～C2 のレベルの規準設定は、14 言語のそれぞれの専門家たちを集め、各技能に対して大がかりに実行された。専門家たちは、CEFR を熟知するためのトレーニングを受け、「CEFR で記述されている、あるレベルの能力を持つ受験者が、そのテスト問題に正解できるかどうか。」を判断基準にし、一つずつのテスト問題にレベル判定を下していった。 さらに、評価者間信頼性や予備テストの結果との相関係数など、量的分析の結果も踏まえ最終的に CEFR 判定レベルの分割点（カットポイント）を決めている。

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

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

なことができるのかを質的能力指標として示すことができるようになったと指摘している。

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

２．５．　自己評価としての **CDS**　**：Can-Do** 自己チェックリスト

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

North (1995, 2000), North and 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 CDS を Rasch モデルを利用して分析検証した。彼らは、言語能力を communicative language activities, strategies, qualitative aspects of language proficiency のようにカテゴリーに分ける大枠を作り、さらにその中で細分化してからそれぞれにあてはまる CDS を作成した。次に、その CDS を利用して教師が学習者を評価し、その結果を同一尺度化するために Rasch モデルによる項目バンク作成手法を用いて分析した。その後、

Lenz and Schneider (2004)は、作成した英語の CDS の項目困難度を、CDS 項目バンク (Bank of Descriptors) としてウェブ上で公開している。

また、Sato (2010)は、英検 CDS を自己評価ツールとしてその妥当性の確認をする研究を実施した。彼は、英検 CDS のうち、5 級～準 2 級までの 16 項目の CDS を、2571 人の日本の中学 1～3 年生に自己評価として回答してもらい、そのデータを Rasch モデルを使って分析した。その結果、16 項目すべてが受験者の中学生たちにとって、困難度が比較的低めで、また、16 項目に対する自己評価による項目困難度と 5 級～準 2 級までの設定されたレベルは、ほぼ一致した。さらにまた、この受験者の自己評価結果と彼らの英語能力のレベル、さらに英語学習に費やした時間とも比例した。しかし、研究対象とした 16 項目は、英検 5 級～準 2 級までの CDS の限られた一部であるため一般化することは難しい。しかし、少なくともこれらの 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 and Schneider (1998)の研究で開発された CEFR の発話能力に関する 99 の CDS を用いて、ある日本の大学でのスピーキング能力の自己評価と教師評価を比較したものである。約 2600 人の学生たちは、プレースメントテストの結果により CEFR の 6 段階（A1 から C2 まで）に対応した習熟度別レベルに編成され、英語コミュニケーション能力を養成するコースを履修している。学生たちが自己評価している間に、教員も学生たちを評価した（教師評価）。ここでは、4 件法でなされた評価を「できる」「できない」の 2 段階にした後、BILOG-MG3.0 を使用して、2 パラメタ IRT モデルでこのデータを分析している。その結果、学生の自己評価と教師評価の項目困難度の相関はかなり高い（r = .83 ～.97）が、学生の自己評価と教師評価の能力値の相関は低い（r = .18～.28)ことが分かった。また、この研究で使用した A1～C2 まで 6 段階の CDS 項目群を、①学生自己評価による特性曲線と②教師評価による特性曲線としてそれぞれ図にしたところ、①②の図ともに各群の特性曲線が、CEFR が設定した通りの 6 段階に分かれた。

２．６．日本人学習者に適応する CEFR へ

ヨーロッパの言語学習者のために作られた CEFR は、日本人学習者にそのまま適用するには無理があり、修正や工夫をしてより日本人学習者に適応させる必要があると言われている（境, 2009; 根岸, 2006b）。例えば、中島・永田（2006）は、CEFR 準拠の自己評価アンケートである DIALANG self-assessment (SAS) を使用して、CEFR がどのくらい日本人学習者に適用

可能かを検証した。彼らは日本人学習者たちが、各 CEFR の能力記述文に対してどのような困難度レベルとして認識しているかを調査した。さらに根岸（2006b）は、この研究の中で、日本人学習者たちが答えた困難度レベルと CEFR の設定している困難度レベルの間にはっきりとした相違があった項目に注目した。例えば、CEFR の Reading の A1 レベルの項目にある「葉書などに書かれた、短く簡単なメッセージを理解することができる。」に対して、日本人学習者はより困難である A2 レベルと判定した。これは、おそらく CEFR の基準では、カードに Happy Birthday! や Congratulations! にプラスして、とても簡単な短いメッセージを付け加える程度を設定していたと思われる。ところが日本人学習者たちが、「post card = 葉書」に書かれたメッセージとして連想する内容が、もっと長い情報量であったからだと思われる。そしてまた、「お店や郵便局、銀行で簡単な用事を済ませることができる。」という CEFR Listening A2 の項目に対して日本人学習者たちは、CEFR 設定より困難度ランクが 1 つ上の B1 レベルと判定した。これは日本人学習者が英語でこれらの経験をしたことがほとんど無いために、困難度が高いと思ったからだと推測できる。このように、学習者が自己評価するとき、彼らが体験したことがない内容を自己評価のための質問にしても、その回答はあまり正確ではないと言われている（伊東・川口・太田、2008）。

Negishi (2005)や根岸（2006a）では、このように CEFR レベルと日本人学習者の判定が異なった項目に、学習者が具体的に内容を理解するための工夫として参考資料を付けることで成果をあげたと報告している。例えば、前述した Reading A1 レベルの項目には、参考資料として具体的なカードの見本を示し、Listening A2 レベルの項目には、銀行や郵便局での簡単なやりとりの例を示した。両方とも改良後の項目の困難度は、ほぼ CEFR 設定どおりの順序となった。

さらに、CEFR をもっと日本人学習者に適用させる動きのなかで、日本版 CEFR（CEFRjapan）のフレームワークを構築しようとする取り組みも行われている。ここではまず、一般的な日本人学習者のレベルは、CEFR の下位レベルをさらに細かく分ける必要があると認識し、ヨーロッパで CEFR の下位レベルをより細かく分けている CEFR フィンランド版を参考にして、A1 を3つに、A2，B1，B2 はそれぞれ2つに分ける日本人学習者向きレベルの設定を提唱している（岡、2008）。そしてこの動きと符合する研究として、2.4 で述べた斉田（2008）によると、日本人大学 1 年生のリスニング能力は CEFR の A1 レベル、リーディング能力は A2 レベル、ライティング能力は A2 レベル、文法能力は B1 レベル、語彙能力は A2 から B1 レベルという結果であり、大多数が A1～A2 という非常に狭いレベル範囲に入るという可能性を示している。これは、

CEFR を日本人学習者に適用させるようにレベル設定をするには、やはり A1、A2、B1 の3レベルのなかに、より詳細なレベルを設定したほうが現実的であるという方向性をサポートしている。

２．７．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 自己チェックリスト、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 自己チェックリストは、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。

これらの問題を解決し、より正確な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チェックリストも変えなければならない。このように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。

２．８．CDS を導入する日本の大学英語教育プログラム

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 2001 年に発表し、その日本語訳『外国語の学習、教授、評価のためのヨーロッパ共通参照枠』(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 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) が開発されるようになった。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。そして、日本の大学英語教育プログラムにおける英語の授業に Can-Do statements (CDS) を導入する動きは緩やかに広まりつつあるが、2012 年には、文部科学省に「外国語教育における「Can-Do リスト」の形での学習到達目標設定に関する検討会議」が設置され、その動きはますます本格的に拡大しつつある。

英語教育プログラムに於いて、CDS 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 and 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。

さらに、ここで CEFR に基づいた CDS を日本の大学英語教育プログラムに導入する動きを、個々の大学ごとに焦点を当てつつ説明する。まず茨城大学では、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 and 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、教育原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）。

２．９．今後の CEFR ベースの英語教育プログラム

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 and 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。しかしながら、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、その言語学習の現場に適用する形に修正しなければならない。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到達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。しかし、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ことになるので、今後このような取り組みを実践した事例研究を実施していくべきだ。そして、この一環した教育プログラムを教員たちが参加型プログラム評価することで、改善へのサイクルが廻り始める。

2．10　日本のある大学での CDS を基盤にした英語教育プログラム

CEFR ベースの英語教育プログラムをめざして、2010 年にカリキュラム改革を実施したのが日本の大学英語教育プログラムとしては最大級の規模の T 大学である。1学年約 5500 人の学生数で2年間にわたり必須英語教育プログラムを実施している。新しいカリキュラムは CEFR をもとにして独自に開発した「T 大学 CDS」を授業の到達目標、評価基準、習熟度別クラスの基準、として授業の中心に据える。実際の場面でコミュニケーションとして英語で対応できるようになることを目標にしている。これは、日本の中学・高校の英語指導要領にも4技能統合型やインプットとアウトプット統合型授業が強調されるようになってきたことによる。新カリキュラムでは、インプットとアウトプットを合わせた 2 技能合体型（リスニング&スピーキングコース、リーディング&ライティングコース）を設置した。

T 大学 CDS を中心にした新英語教育プログラム

CEFR に準拠して作成された T 大学 CDS は、授業の到達目標であり、その目標に到達しているかどうか、学生が自己評価をする基準でもある。また、教員は T 大学 CDS に基づいた教材を使って授業を実施し、学期の中間と期末に教員が作成したテストを実施して学生を評価する。1 学年 5500 人が 2 年間に渡って受講する必須英語教育プログラムの壮大な構想を実現するために約 40 名の日本人・ネイティブスピーカーの専任教員が中心となり、約 120 名の非常勤教員とともにこのカリキュラムを実施している。

この英語教育プログラムでは、1年次入学直後に事前テストを実施し一週間以内に採点してそのスコアをもとにクラス分けを実施している。クラスのそれぞれのレベルは、CEFR の4つのレベルと連動している。1年生のクラスは、初級が CEFR のA1レベル、中初級がA2、上級がB1で、2年生のクラスは、初級がA2，中級がB1，上級がB2となり、1年間2セメスターの授業を受けたのち、全体的にレベルが1つ上がり、さらに 1 年間2セメスターの授業を受けて 2 年間で合計4セメスターの授業が完結することになる。

Brown（1995）は言語カリキュラムの要素に、ニーズ分析、授業の到達目標、テスト、教材、教授法を入れている。それらを総合的にとらえてプログラム評価をし、その評価のプロセスは終了することなく循環するべきであると述べている。T 大学の新しい英語教育プログラムは、それぞれの要素に T 大学 CDS を埋め込んだ形となっている。

図 2.3　T 大学 CDS を基にした授業の流れ

リスニング筆記テストと Can-Do 自己チェックリスト

図 2.4 では、リスニングの筆記テストと、テストによって測る能力と直接的に関わる授業内容を黄色にし、Can-Do 自己チェックリストで自己評価する能力は黄色に加え、赤で示した。これらの能力を区別することは非常に難しいが、リスニング筆記テストで測る能力（黄色）は θ o、Can-Do 自己チェックリストで自己評価する能力（赤と黄色の両方）は θ s で表すことにする。

全員が受験する事前テストの結果によって習熟度別クラスに分けられた学生たちは、学期初めに第 1 回 Can-Do 自己チェックリストに答え、自分の英語リスニング＆スピーキング能力を確認する。その後は、レベルごとに異なったリスニング＆スピーキングのための教科書を使用して学習を進める。教科書はそのクラスのレベルにあった CEFR レベルのテキストを使い、リスニングとスピーキングのアクティビティー（θ o と θ s）は、できるだけ偏らず半分ずつになるようにする。リスニングに関しては、ほぼ毎回、教科書の各テーマに関する説明文や会話を、CD を使って聴き、聴解問題に答える練習をする(θ o)。リスニングで使用したのと同じテーマについて 2 人組で会話練習する。会話練習はスピーキング能力だけでなく、相手が言っていることを理解して初めて会話が成立するのでリスニング能力の練習でもある。また、実際のコミュニケーションとしての英語能力は、ジェスチャーや会話方略等も包括した能力（θ s）である。このアクティビティーを基本にして、各教員が扱われているテーマに関連する素材を見つけ補足をしたり、グループでディスカッションしたり、様々なバリエーションのある授業を展開する。

第 2 回 Can-Do 自己チェックリストは、中間リスニングテストと中間スピーキングテスト実施直後の授業中に行う。中間と期末の両方のスピーキングテストとも学生が 2 人一組で、与えられたテーマについて 5 分間会話し、その会話を共通の評価表を用いて各担当教員が一人で評価するものである。このスピーキングテストも、先に述べた会話練習と同じくリスニング能力も同時に測っている。学生は Can-Do 自己チェックリストの一つひとつの項目について自己評価しながら、どの項目が「できる」ようになったか、「あまりできない」ままなのかなど、よく考えて計画的に残り 15 回の授業を使って自己修正できるようにする。この時、授業内で一人ずつの学生にアドバイスする時間をとるように教員に対して要請している。学期前半と同じ要領で後半の授業が実施され、期末リスニングテストと期末スピーキングテスト実施直後に第 3 回 Can-Do 自己チェックリストを実施する。中間・期末のリスニングテストは、会話や説明文を聴いてその答えを4つの選択肢から答える多肢選択問題で、事前・事後テストと同じテスト形式である。

授業における Can-Do 自己チェックリストの活用

これからも CDS を到達目標とする英語教育プログラムにおいて、妥当性・信頼性の高い Can-Do 自己チェックリストの重要性は高まるであろう。本研究は、1万人以上の学生が履修するある日本の大学英語統一プログラムに於いて、大学独自の Can-Do 自己チェックリストを作成し、それを到達目標として設置するプロジェクトの一環として実施された。このような大規模な統一英語カリキュラムを維持するのに最も重要なのは、教員と学生の間に「対話」を作ることではないかと思う。この統一英語カリキュラムでは、学期のはじめ、半ば、おわりの 3 回、学生による Can-Do 自己チェックリストを使った自己評価を実施して、教員と学生の対話の機会を作っている。これら 3 回の Can-Do 自己チェックリストは、学生にとっては、これまでの学習を「振り返る」機会であり、教員にとっては学生の自己評価と教員評価をすり合わせて、学生を自己修正に導く機会となる。そして重要なことは、教員たちが僅かな時間であっても、学生ひとり一人の Can-Do 自己チェックリストの回答を見てフィードバックを与えるなど、Can-Do 自己チェックリストを通じて学生と「対話」するようにするところである。このように「対話」のきっかけとなることも、Can-Do 自己チェックリストの大切な役割だと思う。

第3章　Can-Do 自己チェックリストによる自己評価

３．１．　はじめに

本章には、Can-Do 自己チェックリスト（以下 SCL と略す）を、項目反応理論（IRT）を用いて分析することに関する、以下の2つの目的がある。1 つは、プログラム評価の客観的エビデンスにするために、CEFR を日本人学習者に適応させたものを、さらにそのプログラムの学習者に適応したものにカスタマイズしていくことで、より学習者の声を反映させた実態に即したものに修正する試みである。もう1つは、学生が CDS を到達目標として意識しながら学習し、学習過程で各 CDS に対する自らの能力がどのように変化したか認識し、自己修正することをサポートするためである。

本章では、教員が難易度を設定した SCL を学習者に回答してもらい、IRT を用いてその回答結果の困難度パラメタを推定し、その推定値を使って改訂した SCL を作成する。このように、学習者の声を反映させて規準設定した SCL は、教員だけで作成したものより学習者の実態に近づいた自己評価が可能になるはずである。これにより、学習者の大切な「振り返り」「自己修正」のチャンスをより有効なものにするだけでなく、プログラム評価のエビデンスの客観性や、信頼度も高まるという好循環が生まれることを期待する。

３．２．　英語教育プログラムに適用する CDS

この約 10 年、日本の高等英語教育プログラムの授業に Can-Do statements (CDS)を導入する動きが穏やかに広まりつつあったが、2012 年には、文部科学省に「外国語教育における「Can-Do リスト」の形での学習到達度目標設定に関する検討会議」が設置された。

Common European Framework (CEFR) (Council of Europe, 2001)　や、英検、GTEC for STUDENTS, TOEFL, TOEIC などがそれぞれの規準で、「どのようなテスト結果を得た学習者は何ができる。」という CDS を設定している（テストスコアの解釈規準としての CDS）。そしてまた、European Language Portfolio (ELP)のように学習者が自己評価として自らの英語能力を診断し、また教員も学習者のレベルを判断する手段として利用可能な CDS として、SCL がある。例えば、英語リスニング能力についての SCL には、会話について「丁寧にゆっくりと話された短い簡単な会話なら理解できる」、また、音声素材を聞くことについて「テレビのニュースのトピックや天気予報、商品の宣伝などの要点を理解することができる」などがある。

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

このように妥当性が高く、その英語教育プログラム履修者の英語能力に可能な限り適応した CDS を作成する試みが行われているが、その一つが、項目反応理論（IRT）を用いて困難度パラメタを推定し、CDS の規準設定をする方法である。North and Schneider (1998)、Sato (2010)、筒井・近藤・中野 (2007) は、CDS を自己評価のツールとして、あるいは学習者のレベルを判断する教師評価の手段として実施し、IRT を用いた分析を行ってその妥当性を確認した。この時、英語の 4 スキルを一つずつ 1 元データとして扱うことはあるが、1つのスキルのなかの詳細（例えば、リスニングスキルのなかで、聞く素材、会話、レクチャー、テレビ・ラジオなど）に細分化して分析していない。さらに、これらの研究は主に SCL の結果を、1 パラメタ・ロジスティック・モデル（1PL）、または 2 パラメタ・ロジスティック・モデル（2PL）を利用して、各 CDS の困難度パラメタを推定して難易度の規準設定の目安にする方式を採用している。

今後、CDS の設定が日本の英語教育の現場に浸透していくであろうが、いかに個々の英語教育プログラムに適した妥当性の高い CDS を設定することができるかが、CDS が普及する重要な鍵となるはずである。その為には充分に多くの事例研究を実施して、その英語教育プログラムの学習者にできるかぎり適応した CDS の設定を追求することが必要である。Green (2010) も、研究者、教員と学習者が実際に使っている教材や言語運用の実践的な例を持ち寄って、より妥当な CDS のレベルの設定のために意見交換し、積極的に協力し合うことの重要性を強調している。このような背景から、本研究が、日本の大学英語教育プログラムにおいて、対象となる学習者に適応した CDS を設定し、妥当性の高い規準設定を実施しようとするときの、資料の一つとなれば幸いである。

３．３．　Can-Do 自己チェックリスト（SCL）についての先行研究

3．3.1　自己評価としての Can-Do 自己チェックリスト（SCL）

CDS にもとづいて、学習者が自己の能力を診断し、また教員も学習者のレベルを判断する手段として利用するための自己評価チェックリストを Can-Do 自己チェックリスト（SCL）と言う。こ

の代表的なものが、CEFR に基づく SCL として開発された European Language Portfolio (ELP) である。ELP は、技能ごとに CEFR の 6 段階（A1, A2, B1, B2, C1, C2）の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）をチェックリストにしたものである。このリストを学習者が自己評価としてチェックすることによって、自分の能力レベルを診断することができる。このようにして ELP は、能力と目標の 2 つの面から学習プランを立て、学習者が自ら目的をはっきりと持って学習できるようにし、最終的には、学習者の自律的学習を促進することをめざしている。そしてまた、学習の記録を残すことができるようにするために、ポートフォリオのスタイルをとっている。ELP は、CEFR の 6 段階のレベルごとに、領域、場面、状況に合うように CDS を設定している。

North (1995, 2000), North and 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 CDS を Rasch モデルを利用して分析検証した。彼らは、言語能力を communicative language activities, strategies, qualitative aspects of language proficiency のようにカテゴリーに分ける大枠を作り、さらにその中で細分化してからそれぞれにあてはまる CDS を作成した。次に、その CDS を利用して教師が学習者を評価し、その結果を同一尺度化するために Rasch モデルによる項目バンク作成手法を用いて分析した。その後、Lenz and Schneider (2004)は、作成した英語の CDS の項目困難度を、CDS 項目バンク (Bank of Descriptors) としてウェブ上で公開している。

また、Sato (2010)は、英検 CDS に着目し、その妥当性の確認をする研究を実施した。彼は、英検 CDS のうち、5 級～準 2 級までの CDS16 項目を、2571 人の日本の中学 1～3 年生に自己評価の形で回答してもらった結果を、Rasch モデルを使って分析した。これら 16 項目すべてが受験者の中学生たちにとって、比較的低めの困難度となったが、16 項目に対する自己評価による項目困難度と5級～準2級までの設定されたレベルはほぼ一致した。さらに、この受験者の自己評価結果と彼らの英語能力のレベル、英語学習に費やした時間とも比例した。しかし、研究対象とした 16 項目は、英検 5 級～準 2 級までの CDS のなかの限られた一部であるために一般化することは難しいが、これらの項目の妥当性が高いと言うことはできる。

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 and Schneider (1998)の研究で開発された CEFR の発話能力に関する 99 の CDS を用いて、ある日本の大学でのスピーキング能力の自己評価と教師評価を比較したものである。約 2600 人の学生たちは、プレースメントテストの結果により CEFR の 6 段階（A1 から C2 まで）に対応した習熟

度別レベルに編成され、英語コミュニケーション能力を養成するコースを履修している。学生たちが自己評価している間に、教員も学生たちを評価した（教師評価）。ここでは、4 件法でなされた評価を「できる」「できない」の 2 段階にした後、BILOG-MG3.0 を使用して、2 パラメタ IRT モデルでこのデータを分析している。その結果、学生の自己評価と教師評価の項目困難度の相関はかなり高い（r = .83 ～.97）が、学生の自己評価と教師評価の能力値の相関は低い（r = .18～.28)ことが分かった。また、この研究で使用した A1～C2 まで 6 段階の CDS 項目群を、①学生自己評価による特性曲線と②教師評価による特性曲線としてそれぞれ図にしたところ、①②の図ともに各群の特性曲線が、CEFR が設定した通りの 6 段階に分かれた。

これらの Can-Do 自己チェックリストに関する先行研究は、CEFR や英検 CDS などの既存の CDS を使用している。CEFR や英検 CDS の項目困難度が、どれくらいターゲットとする学習者の示した項目困難度と一致しているかを検証したものである。

3. 3.2.　日本人学習者のための CDS

根岸（2006b）は、ヨーロッパの言語学習者のために作られた CEFR は、日本人学習者にそのまま適用するには無理があり、修正や工夫をしてより日本人学習者に適応させる必要があると述べている。例えば、中島・永田（2006）は、CEFR に準拠した自己評価アンケートである DIALANG self-assessment (SAS) を使用して、CEFR がどのくらい日本人学習者に適用可能かを検証した。彼らは日本人学習者たちが、各 CEFR の CDS に対してどのような困難度レベルとして認識しているかを調査した。さらに根岸（2006b）は、この研究の中で、日本人学習者たちが答えた困難度レベルと CEFR の設定している困難度レベルの間にはっきりとした相違があった項目に注目した。例えば「お店や郵便局、銀行で簡単な用事を済ませることができる。」という CEFR Listening A2 の項目に対して日本人学習者たちは、CEFR 設定より困難度ランクが 1 つ上の B1 レベルと判定した。これは日本人学習者が英語でこれらの経験をしたことがほとんど無いために、困難度が高いと思ったからだと推測できる。また、このように学習者が自己評価するとき、彼らが体験したことがない内容を自己評価として質問にしても、その回答はあまり正確ではないと言われている（伊東・川口・太田、2008）。

また、忘れてはならないのが日本人の自己評価に対する一般的な傾向である。日本の大学英語教育プログラムにおけるライティングのクラスで、学生の自己評価、学生同士の相互評価、そして教員評価の3つを比較した研究では、相対的に教員からの評価に比べ自己評価はやや低くなる傾向が見られた。（Fujita,2001, 2002; Saito and Fujita, 2004; Saito and Fujita, 2009）。

これは、アジア系言語の母語話者は、中南米を含む、ヨーロッパ系言語の母語話者に比べ、自己評価を低く推定する傾向があるという研究結果（Yu and Murphy, 1993）とも符合する。

Negishi (2005) や根岸 （2005、2006a） では、このようにCEFRレベルと日本人学習者の判定が異なった項目に、学習者が具体的に内容を理解するための工夫として参考資料を付けることで成果をあげたと報告している。前述したListening A2レベルの項目には、銀行や郵便局での簡単なやりとりの例を示したところ、改良後の項目の困難度は、ほぼCEFR設定どおりの順序になった。

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

この動きと符合する研究として、斉田 （2008）は、CEFRのレベルでテスト結果が判定される言語能力テストDIALANGの英語版（Alderson and Huhta, 2005）を使って調査し、日本人大学1年生のリスニング能力はCEFRのA1レベル、リーディング能力はA2レベル、ライティング能力はA2レベル、文法能力はB1レベル、語彙能力はA2からB1レベルという結果を得た。これは、日本人大学生の大多数がA1～A2という非常に狭いレベル範囲に入るという可能性を示していて、CEFR を日本人学習者に適用させるようにレベル設定をするには、やはり A1、A2、B1の3レベルのなかに、より詳細なレベルを設定したほうが現実的であるという方向性をサポートしている。

3. 3.3. 日本の大学英語教育における CDS

ここで、CEFRに基づいたCDSを日本の大学英語教育に導入する動きに焦点を当てる。主なところでは、茨城大学、大阪大学、慶応義塾大学、東海大学で、これらの大学ではCEFRやCEFRをベースとしたCDSを英語教育プログラムに導入している。茨城大学では、CEFRのレベルを基準にして習熟度別クラスを編成し、また総合英語プログラムを開発し、自律的に英語学習ができる人材養成をめざしている（Ano et al., 2007; Fukuda, 2009; Nagai and Fukuda, 2004）。大阪大学では、25の専攻語すべてにおいて、到達目標をCDSで表して公開し、「透明」「共通」「強制しない」姿勢を原則にしてCDSを中心としたカリキュラム改革を行ってきた（真嶋、

2010）。さらに、Majima (2010)では、日本で CEFR を取り入れた言語教育を行っている事例を 7 つの活用分野に分けて紹介し、そのうちの一つが「CEFR のレベルと教育機関の言語プログラムの到達目標を関連づけたもの。」である。さらに慶應義塾大学では、小中高大一貫教育の中に、CEFR を基にした英語教育を実現しようとしている。この中心的な取り組みの一つとして、English Language Portfolio (ELP) の日本版と言える慶應 ELP を開発、試行している (Horiguchi, et al., 2010)。

最後に、東海大学では、2010 年から CEFR を基にして独自に開発した東海 CDS をカリキュラムの根幹に置いた英語統一必須プログラムを、全学部の 1, 2 年生、約 11000 人を対象にして実施している。ここでは、3 つの習熟度レベルごとに、東海 CDS を自己評価できるチェックリストを作成した。そして、1 学期間に、授業の初回、中間、期末の 3 回にわたり、全学生がそのチェックリストを使って自己評価をしながら、英語学習の進捗状況を確認、振り返り、今後の学習に役立てるシステムを実施している。

3. 3.4. 項目反応理論（IRT）

本論で挙げた自己評価としての CDS に関する多くの先行研究で研究結果の分析に使用されている項目反応理論（IRT）についてここで説明を加えたい。IRT の代表的なモデルとして、1 パラメタ・ロジスティック・モデル (1PL)、2 パラメタ・ロジスティック・モデル (2PL)、3 パラメタ・ロジスティック・モデル (3PL)の 3 モデルがある。1～3PL モデルはそれぞれに違う式で表され（式 1～3）、それぞれの特徴や、項目パラメタを安定して推定するために必要な受験者数もモデルによって異なる（Bond and Fox, 2001; Brown and Hudson, 2002; Hambleton, Swaminathan, and Rogers, 1991; McNamara, 1996; 大友, 1996; 芝, 1991）。

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、安定した推定に必要な受験者数は 500 人以下と言われている。

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

2PL モデル（（2）式）は、*b* パラメタに加えて *a* パラメタ（項目弁別力）の推定もできる。また、受験者数は 500 人から 1000 人であれば安定した推定ができる (eg, Ayala, 2009)。

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

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）。

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

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

3. 3.5. 本章で解明しようとすること

ここでは、特にリスニング能力に注目し、その英語プログラムを履修する大学生の自己評価による困難度に適応したリスニング能力の SCL を作成するための調査を行うことにした。学生の SCL への反応、事前事後の変化を、学生の習熟度レベルによる違いとリスニングテストのスコアとして測定された能力との相関性を調査して、CDS の規準設定にあたって、どのような点に留意することが必要なのか検討する。

（1）事前事後（学期初めと終わり）で実施した日本人大学 1 年生の自己評価としての SCL の結果は、事前事後どのように変化するのか、さらにその変化の度合いは 3 つの習熟度別レベルごとに違いがあるかを調査する。そして、（2）SCL の結果とリスニングテストスコアの相関係数が、リスニングに加え、リーディングや文法も含まれるプレースメントテストで決定した 3 つの習熟度レベルごとにどのように異なるのか比較する。最後に（3）SCL を作成した時、教員たちが想定した項目の困難度と、IRT を用いて分析した項目困難度推定値を比較し、大きな違いがある項目についてその原因を考察する。

3．4． 研究方法

3. 4.1. 受験者

ある日本の大学で必須英語教育プログラムを履修する 1 年生 445 人（全体の約 8%）が本研究に参加した(表 3.1)。このプログラムには、スピーキングコース、ライティングコース、リーディングコース、リスニングコースというように、英語の 4 スキルごとにコースがあり、学生たちは、入学時にプレースメントテスト（文法 30 問、リスニング 30 問、リーディング 40 問の合計 100 問を所要時間 90 分で受験する）のスコアによって 3 つの習熟度別レベル（初級レベル：Basic、中級レベル：Intermediate、上級レベル：Advanced）に分けられる。2 年間で 4 つのコース、合計約 168 時間の英語の授業を履修する。レベルによって使用する教科書も異なっていて、その習熟度レベ

ルに適応した授業内容を実施することになっている。

その中で、リスニングコースを 1 学期間履修する 1 年生の中で、できるだけ全体の比率と近くなるように、各レベルからアトランダムに選んだ学生たちに SCL を実施してもらった。2 回の自己評価回答者の習熟度レベル別の内訳は、表 1 のようになっている。445 人が学期開始後すぐ（4 月）に、本研究における一回目の Can-Do 自己チェックリスト 1（今後 SCL1 と略す）に回答した。そのうちの、331 人が約 3 か月半後の学期末（7 月末）に 2 回目の Can-Do 自己チェックリスト（今後 SCL2 と略す）に回答した。

表 3. 1　*SCL1 とSCL2 の回答者レベル別人数*

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

3. 4.2.　リスニング Can-Do 自己チェックリスト（SCL）

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

SCL をより妥当なものにするために、この英語教育プログラムのリスニングコース担当教員のうち 10 人にご協力いただき、アドバイスやフィードバックをいただいた。各担当教員には、(1)それぞれの SCL の難易度レベルが想定したレベルと合っているか、また、(2) SCL の内容が学習者に問題なく理解できるような表現になっているか、(3) CDS の表現に誤りがないかなど、これら

3 点を中心に修正・変更したほうが良いと思われる点に赤を入れてもらい、また書き込んでもらった。これらを回収して、修正、変更、削除を行い最終版の 28 項目からなる SCL を作成した。

本論では、このようにして作成した SCL を 3 種類（初級、中級、上級）、それぞれ 3 レベルに到達目標として 14 ずつの CDS になるように設定した。以下に示すように、合計 28 の CDS が 7 文ずつ別のレベルと重なる構成になっている（図 3.1 参照）。

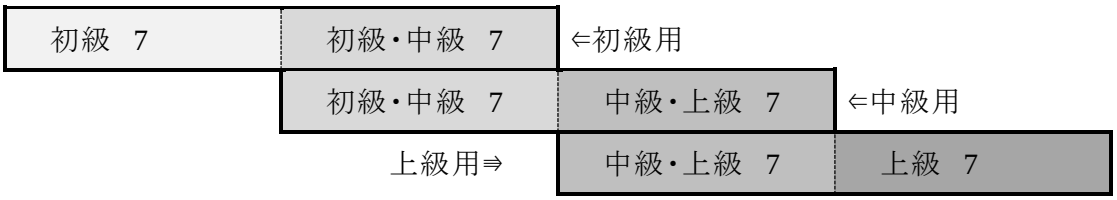


図 3. 1　3 種類（初級、中級、上級）の Can-Do 自己評価チェックリスト(SCL)

14項目の3つのレベル別 SCL

図 3.1 のように 3 つの SCL にしても、IRT によって分析するとこれら3つの SCL を共通尺度化することができる。つまり、被験者が自分のレベルだけの SCL に回答しても、他のレベルとの差も分かることになる。このような 3 種類の SCL にしたことで、1 人の学習者が応える項目数は 14 項目となり得られる情報量は少なくなってしまった。しかし、以下のような利点も挙げられる。

（1）学生が一度に多くの質問に答える必要がなくなり、少ない質問に真剣に回答する可能性がある。

（2）初級の学生が上級の SCL に答える必要がなくなり、回答者の習熟度に適応した質問が可能となる。

（3）自分のレベルの SCL を記憶・認識して学習しやすくなる。

（4）短時間で回答できるため、学期のはじめ、中間、期末のように授業中、容易に3回実施することができる。

3. 4.3. テストスコア

本論で扱うテストは 2 種類あり、一つは、学期初めに実施する学生の英語能力を判定するための実力テストで、所要時間は 90 分間、そのうちリスニング 30 問（これをリスニング 1 とする）、文法 30 問、リーディング 40 問の合計 100 問に多肢選択（4 択）で解答するテストである。表 3 と表 3.4 の文法とリーディングテストは、このテストのサブテストのことである。また、もう一方のリスニングテストは、学期末に実施するリスニング能力の到達度を測る 70 問の多肢選択（4 択）問

題のテストである。これをリスニング 2 とする。

これらのテストは、いずれも CEFR や他の CDS に基づいて作成されたものではなく、リスニング 1 は項目困難度のバランスを重視して作成され、リスニング 2 は、リスニングコースのテキストをもとに問題が作成された。

3. 4.4. IRT による分析

本研究のデータは IRT を利用して、受験者の能力値 θ が変化し、項目パラメタは同じという前提のもとに、統計ソフト BILOG-MG3.0 を使用して 1 パラメタ IRT モデルを用いて分析した。2 パラメタモデルを使用しなかった理由は、受験者数が 500 人以下なので、パラメタの推定が不安定になることを避けるためである。

アンケートの回答は、できない 0、あまりできない 1、まあできる 2、できる 3 の 4 件法で実施した。本来は、これを多値型データとして扱うべきであるが、筒井・近藤・中野（2007）の分析方法に習い、回答のうち 0 と 1 を 0 に、2 と 3 を 1 にして 2 値データとして BILOG-MG3.0 を用いて分析した。ここで SCL1 に応えた受験者を、それぞれ初級 B、中級 I、上級 A という習熟度レベルごとに初級から順に G1-B、G2-I、G3-A とし、SCL2 に応えた受験者も同じように G4-B、G5-I、G6-A と名付け、6 つのグループを同時に分析した（図 3.2 参照）。

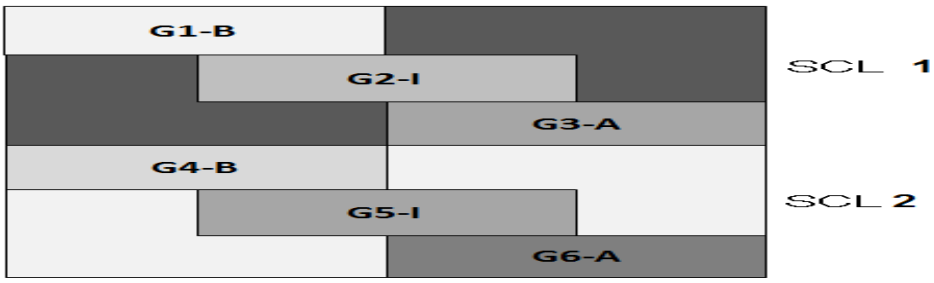


図 3.2.　分析方法：SCL 1 と SCL 2 に応えた 3 習熟度別レベル

3. 5. 結果

3. 5.1.　Can-Do 自己チェックリスト 1 と 2 の変化

前述した分析方法で、6 つのグループの平均能力値（θ）を比較したのが表 3. 2 である。これによると、SCL1 から SCL2 への変化は、全体的に上昇しており、学期初めより学期末の方が平

均で θ = 0.322 の上昇となっている。また、習熟度別の θ の平均値は、学期末のほうが学期初めより、初級レベルは 0.310、中級レベルは 0.315 上昇した。ところが、上級レベルだけ 0.172 の上昇にとどまり、他のレベルの θ に比べ約半分の上昇となった。これは、上級レベルの学習者の特性によるものかもしれない。

表 3. 2.　*SCL 1 とSCL 2 の習熟度別能力値(θ)の平均変化*

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

**有意確率　p＜0.05

3. 5.2.　Can-Do 自己チェックリスト結果とテストスコアの比較

表 3.3 は、SCL1 と英語テストスコアの相関関係を表している。SCL 1 とすべての英語テストは、p<0.01 で相関係数が有意だと認められたが、その相関係数は 0.275 < r < 0.329 と全体的にあまり高い相関関係だとは言えない。自己評価である SCL 1 と、各英語能力テストとの相関関係に比べると、英語能力テストどうしの相関関係（例えば、リスニング 1 とリスニング 2 はr = 0.752）は高く、リスニング能力を測るテストとリーディング能力を測るテストであっても英語能力テスト間では、高い相関係数を示している（リスニング 1 x リーディング、0.746 < r < 0.753）。

相関が低い原因としては、SCL が各レベル 14 項目しかない英語リスニング能力を抽象的に表したものであること、そして本研究で使用したリスニングテストはいずれもここで扱う SCL に基づいて作成されたものではないこと等が考えられる。また、SCL は自己評価であるため、できないのかできるのかを本人の主観で回答するものであり、過大評価や過小評価が起こる可能性がある。これに対して、リスニングやリーディングの英語能力テストは、どの問題に正答したのかを判定する客観的評価であるため、これらの相関が小さいのは、テストタイプの違いからくるのではないかと推測される。

表 3. 3 *SCL1 と英語テストスコアの相関関係（Peason）*

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

Note. すべての相関係数は有意 $p<0.01$. $N = 426$

さらに、表 3. 4 は SCL 2 と英語テストスコアの相関関係を示している。SCL 2 と英語テストのスコアも、$p<0.01$ で全ての相関係数が有意であることを示しているが、その相関係数は $0.210 < r < 0.327$ で、SCL 1 とほとんど大きく変わらずあまり強い関係があるとは認められなかった。

表 3. 4 *SCL2 と英語テストスコアの相関関係（Peason）*

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

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

表 3. 5 は、習熟度レベル別に SCL 1、SCL 2 と、リスニングテスト 1、リスニングテスト 2 との相関関係を集計して比較したものである。ここでの習熟度レベルは、リスニングテストだけでなく、リーディングと文法テストも含むプレースメントテストの結果によってクラス分けされたものである。SCL1 は、習熟度レベルによって大きな変化はなく、全体的に良く似た相関係数を示している。しかし、SCL2 には、特徴があり、リスニングテスト 1、リスニングテスト 2 ともに初級レベルが最も相関係数が高く（ $r = 0.414$、$r = 0.395$）、次に中級レベル（ $r = 0.175$、 $r = 0.230$）、上級レベル（ $r = 0.211$、 $r = 0.115$ ）の順になっている。これは SCL2 は習熟度が低いほど、リスニングテストとの相関係数が高いという結果を示している。

表 3. 5. *習熟度別 SCL1 とSCL2 と英語テストスコアの相関関係*

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

Note. すべての相関関係は有意 p <0.01.

3. 5.3. 教員が想定した困難度と IRT による項目困難度推定値

SCL の 28 項目を IRT によって分析し、項目困難度（*b* パラメタ）順に並べたものが Appendix A である。ここでは、項目困難度パラメタの値が低いものから高いものへ上から下へ並べたのが「パラメタ順」の列で、その次の列には項目困難度の値が示されている。さらに、次の列には「想定順」として、SCL を作成したときに教員グループが想定した困難度の順位が、同じように低いものから表示してある。この中で全く順位の変わらなかったのは、項目（3）「ゆっくりなら日常生活に関する簡単で短い話（家族、趣味、大学、週末など）の大筋やキーワードを理解することができる。」、項目（24）「幅広い、成句（例, give up/ hold out） イディオム （例,be in the same boat / break somebody’s heart）、口語表現（話し言葉にしか使われない表現）を理解することができる。」、項目（26）「多様な内容であっても電話で問い合わせ、クレーム、交渉などを行い、相手の話を理解することができる。」これらの3項目であった。しかし、「パラメタ順」と「想定順」、2 つの順位が 6 位以上くい違っている SCL の項目は表 3. 6 のようになっており、想定順よりパラメタ順が上位となったのは 3 項目、そして想定順よりパラメタ順が下位となったのは、2 項目で、合計 5 項目あった。

まず、想定順よりパラメタ順の困難度が低くなった項目では、想定順 28 位だが、パラメタ順が 16 位となり 12 位も困難度が低くなった項目（16）は、「いろいろな種類のドラマ、ドキュメンタリーや映画などを楽しみながら理解することができる。」である。これは、対象となるドラマ、ドキュメンタリーや映画が、そのテーマや内容によって難易度は大きく異なること、また「楽しみながら理解する」のは、どの程度の深い理解であるのか、など限定しづらいところが順位を大きく変えた理由ではないかと考えられる。次に、想定順 20 位であったが、パラメタ順 13 位になった（13）「買い物に行った場合、商品について店員からの情報（サイズ、機能、割引、在庫など）を聞いて理解することができる。」は、店員とのやりとりが買う商品によって、また会話の内容の奥の深さによって大きく変化することが考えられる。例えば、その商品がパソコンで、機能についての詳細な

内容のやりとりになる場合と、商品がTシャツでサイズや色についての単純なやりとりになる場合とでは、難易度が大きく変わるはずである。

表 3. 6. *想定順とパラメタ順のくい違い*

想定順よりパラメタ順が 6 位以上易しい		想定順よりパラメタ順が 6 位以上難しい	
想定順	パラメタ順	想定順	パラメタ順
12 位	6 位	8 位	14 位
20 位	13 位	15 位	21 位
28 位	16 位		

最後に、想定順 12 位であったがパラメタ順は 6 位であった(6)「自分の良く知っている話題(趣味や好きなこと)で、簡単な内容であれば、話の要点を理解することができる。」で、これも先に述べた項目(16)項目(13)と同じく、話題の種類によって難易度が異なるうえ、「簡単な内容」の簡単さが受け取る側に個人差がある。

反対に、想定した順よりパラメタ順の困難度が高くなった項目を見てみると、想定順 15 位がパラメタ順が 21 位となった(21)「テレビで政治、社会、経済などに関するニュースを見て、映像を見ながらその要点を理解することができる。」がある。これは、ニュースの話題である、政治、社会、経済が、大学 1 年生にとってはあまり身近でなく、関心もない場合が多い。このような話題は、母国語でも難しいイメージがあるかもしれない。また、想定順 8 位でパラメタ順が 14 位の(14)「テレビのニュースのトピックや天気予報、商品の宣伝などの要点を理解することができる。」については、英語放送のテレビでニュース、天気予報、商品の宣伝を見たことがないという経験の無さが、想定した順位よりも学生たちが難しいと考えた原因ではないかと思われる。英語放送の映画やTVドラマは、日本でも2か国語放送などで容易に見ることができるが、ニュース、天気予報、商品の宣伝などは現地でしか見ることができないことを認識するべきであった。

3. 6. 考察

SCL で測る尺度とリスニングテストの能力尺度 （θs と θo）

本論では、CDS の項目に対して「できない」「あまりできない」「まあできる」「できる」の 4 段階で回答する SCL の結果を IRT により分析し、得られた自己評価の尺度を subjective の s を付けて「θs」とし、リスニング能力を筆記テストで測った能力値を objective の o をつけて「θo」と

呼ぶことにする。既に本章 3.5. 2. で両者の相関係数が低い(0.210 < r < 0.329) ことは述べた。θs と θo を別々に扱うことにしたのは、ここで提案するプログラム評価モデル EEP－J において θs と θo の両方を測る必要があると思うからである。プログラム評価が追求するのは、異なる別の角度からのエビデンスである。リスニングテストは学生の英語能力のすべてを測っているわけではない。英語の「リスニング能力」と言っても、テスト問題だけでは到底測れないものが沢山あり、それらの一部は自己評価によって測ることができるかもしれない。また、教員による評価では、伸びが認められない場合でも、学習者自身は能力が伸びたと考えている場合もある。そして、自己評価に伸びがあれば 1 年から 2 年生になるとき、学習者が現在のレベルより 1 つ上のレベルに上がることへの肯定的な判断の根拠の 1 つとすることができる。また、θs が伸びない場合は、そのプログラムに問題がある可能性がないとは言えない。最後に、プログラムを履修したことで、学生自身がどれくらい自分の能力変化を認識しているのかを示す指標は必要だと考える。

本章では、学期初めと終わりに同じ SCL による自己評価を行い、その平均能力値（θs）の変化を比較した結果、受験者たちは学期初めに実施した SCL1 より、学期末に行った SCL2 のほうが、平均で θs = 0.322 上回る傾向にあった。そして習熟度別による違いは、中級と初級レベルの学生の平均 θ が、ほぼ同じくらいの伸び（θs = 0.310、0.315）を示したのに反し、上級レベルのみが 0.172 の伸びにとどまった。コースを履修した事前事後で学習者たちが自己評価を行うと、自分の能力が「上昇した。」と答える人が「下降した。」と答える人を上回るのは自然である。したがって、SCL1 と SCL2 の間で θs が 0.322 伸びることは充分あり得る。しかし、習熟度別の θs の変化では、上級レベルの θs の伸びだけが、他のレベルの半分であったことは、「上級者ほど自分の能力のさらなる伸びを実感しづらい。」のに反して、「初級者ほど能力が伸びる余地を多く残している。」などの理由が推測できる。

IRT 分析の新規性（項目困難度による教育プログラムの SCL 修正）

3. 3.1 で述べたように、IRT を用いて SCL の難易度を分析する先行研究は、CEFR や英検 CDS などの既存の CDS を使用して、それらが実際に SCL に回答した学習者の示した項目困難度と違うのかを検証したものである。先行研究で使用したのは、既に広く使われている CEFR や英検 CDS で、これらの妥当性を検証するものである。IRT 分析によって得られた結果によって修正・変更して教育プログラムで利用するものではない。本研究の新規性は、この IRT 分析で得られた項目困難度によって、実際の教育プログラムで授業を受ける学生たちの実態に合わせ、SCL を修正して教育プログラムで使用するところである。

SCL 作成時に教員たちが想定した項目の困難度による順序（想定順）とそれに学生が応え

た結果を、IRT を用いて分析して得られた項目困難度推定値パラメタによる順序（パラメタ順）のくい違いを調査した結果、くい違いが 6 位以上のものが 5 項目あった。この 5 項目の問題点を総合的に考察すると、まず、聞く対象となる英語、つまり会話、テレビ、映画、店でのやりとりなどの「話題」をシステマティックに限定して、難易度を設定しなければならない。話題の難易度に関する規準設定は、CEFR に示されているが、その設定の規準はヨーロッパの言語学習者であって日本人学習者ではない。日本人学習者の環境や条件を加味した規準設定をする必要がある。例えば、想定順 28 位の項目（16）にある「いろいろな種類のドラマ、ドキュメンタリーや映画」は、日本では字幕放送や吹き替えで見る機会も多い。日本人学習者が、これらを見た経験があり、あまり難しくないと感じたのかもしれない。それに反して、項目（14）や項目（21）のように、英語圏で放送されているような、ニュース、天気予報、TVCM は、実際に海外に行ってテレビ番組を見る機会がない日本人学習者にとって、経験がなく難しいと感じるのは自然な反応であるように思われる。このように、「経験したことがない話題」に関して学習者が自己評価するとき、あまり正確な判断ができないことは、先行研究結果と一致する。そして、会話、映画、テレビ、ラジオなど、英語をどのような媒体で聞くかについても、日本人向けの規準設定に配慮をする必要がある。特に英語のラジオ放送を聞く機会は日本ではとても限られている。

最後に、SCL の項目内容を日本語で表現する方法には、細心の注意を払う必要がある。項目（16）の表現に使われた「楽しみながら理解する」は、どの程度の深い理解であるのか限定しづらい。また、項目（6）の「簡単な内容」も、その簡単さに個人差がある。このような日本語表現に関することは些細なことではある。しかし、あいまいな表現方法を使わないようにすることは、SCL による自己評価をより正確なものに近づけるために、とても大切な方策の一つであると思う。

３．７．　結論

今後も日本の大学英語教育において、CDS を授業に取り入れる動きは進んでいくであろう。しかし本来 CDS は、その英語教育プログラムのニーズに適応するように作成されるべきで、例えば日本人学習者、大学 1, 2 年生というように、対象者を絞り込んで、その英語教育プログラムの学習者に可能な限り適応した CDS を作成することが理想的である。

本論では、ある大学英語教育プログラム独自の CDS を作成するにあたって、どのような留意点があるのか探るため、その CDS に基づいた SCL を作成し、学習者に回答してもらって習熟度レベルの規準設定をする方法を検討した。そして、英語リスニング能力についての SCL に注目

し、その英語プログラムを履修する大学生が SCL に答えた結果を分析し、その困難度パラメタを規準設定の材料として SCL を作成することを目指した。学生の SCL への反応と、学生の習熟度レベルやリスニングテストのスコアとして測定された能力との相関性を調査したところ、上級レベルの学生よりも、初級レベルの学生ほど、自分たちの能力が伸びたと実感しやすいと言う結果が得られた。また、教員たちの予想に反し、初級レベルの学生たちも正確に自己評価をしている可能性があるという傾向も示された。これらの結論は、その英語教育プログラムの履修者に適合した CDS を設定しようとするとき、履修者たちに作成した SCL の項目を自己評価してもらい、その結果を IRT を用いて分析して困難度パラメタを推定することは、規準設定に役立つ可能性があることを示唆している。

また、本論では教員たちが想定した項目の難易度による順位（想定順）と、IRT によって推定された学生自己評価による項目困難度の順位（パラメタ順）のくい違いから、今後のリスニング CDS 作成時に留意すべき、幾つかの点に気付くことができた。日本人大学生たちが聞く英語の「話題」と「媒体」（例：会話、映画、テレビ、公共のアナウンス）による困難度の違い、また CDS の日本語の表現についても、もっとシステマティックに、そして慎重に規準設定しなければならない。これから CDS を導入したカリキュラムが普及するにつれ、その英語教育プログラムに適合し、学習に役立つ CDS を設定するための研究をもっと活発に実施するべきだと思う。そして、CDS の規準設定に対する、さまざまな角度からのデータを収集し、研究者どうしの連絡や情報交換を密接に行う必要がある。

このように、リスニング能力 CDS の規準設定を、実際の学習者の声を聞いて実態に近づけるのは、非常に地道な労力を要する作業である。しかし、リスニングコースの授業をシステマティックに確実に向上させるものである。これはリスニングコースに限らず、他の技能にも同様のことが適応でき、英語教育プログラム全体に適用することができる。本論で対象とする大学英語教育プログラムの参加型評価は、ステップ３で「改善」のためのアクションを起こさなければならない。英語教育プログラムの基盤として設けた CDS を最初に一度設定したら、２度と変えないのではなく、「改善」に向けて変更・修正していく努力を惜しまないことこそが重要だと思う。

第4章　自己評価のためのツールの作成

4．1．はじめに

Can-Do 自己チェックリスト（これ以降 SCL と略す）を、プログラム評価のための重要な客観的エビデンスの 1 つにするために、第 2 章では、CEFR などの言語能力の世界基準を授業の基盤に置くことの重要性について説明した。そして、第 3 章では IRT を利用した分析により学習者の声を取り入れて実態に近づけるための研究を実施した。本章では、SCL の question order effects（質問順序効果）に注目し、質問項目の順序が結果に与える影響を明らかにする。

CDS には Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001)に準拠した European Language Portfolio (ELP)があるように、学習者が自己評価として自分の英語能力を診断し、また教員も学習者のレベルを判断する手段として利用可能なものがある。この ELP のように Can-Do statements (CDS) を基に編集し、そのまま自己評価としてのツールにしたものが、SCL である。これは、CEFR の目的の 1 つ、「自律的学習者を養成すること」に由来する。つまり学習者が、学んだことがどのくらい身についたか自己評価し、それが十分でなければ何が良くないのか自分で考え、自己修正する「振り返り」の機会を持つことができるようにすることを意図している。いわば、SCL は、自律的学習者のための重要なツールの 1 つである。

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

さて、SCL は、その質問項目にスムーズに回答してもらうため、一般的に項目を難易度の低いものから高いものへ並べる配置になっていることが多い。そのため、学習者が、その項目に対して、「できる/できない」などと回答しているとき、真剣に質問の内容を読まず、SCL のその項目

の位置で難易度を推定していることが考えられる。「英検 Can-Do リスト」を自己評価のツールとして活用する場合も、問題となるのは学習者が上の級の能力記述文を、下の級の能力記述文と比べ、それらの内容ではなく、その項目の書かれた場所から、「できる/できない」の判断をする可能性がある。例えば、準 2 級と 2 級の Can-Do リストを比べ、2 級の Can-Do リストの内容を良く読んで理解することなく、準 2 級の Can-Do リストより高度な内容だと判断することが問題となる。

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

4．2． Can-Do 自己チェックリスト（SCL）に関する先行研究

4. 2.1 日本人学習者に適応する CDS と SCL

ヨーロッパだけでなく他の地域にも影響を与えるようになった CEFR を、その国や地域に適合させたものを、国・地域言語参照レベル記述 (Reference Level Descriptions for National and Regional Languages: RLD)という。これは、ヨーロッパの言語学習者のための CEFR を、世界中の言語学習者にそのまま適用させるには無理があり、大きな変更や工夫をする必要性から出てきた動きである。そして、CEFR の枠組みは参照のためであり、その言語学習の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 のポリシーである。Weir（2005）は、CDS はそれを使用する国ごと、さらに教育機関ごと、言語カリキュラムごと、テストごとに、その学習者や受験者に適した CDS として誂える（Tailor made）の必要があると言っている。 この CEFR を日本人学習者に適合させた RLD にする試みが次に述べる CEFR-J やジャパン・スタンダード（Japan Standards for Foreign Language Proficiency based on CEFR:JS）である。このとき CEFR を、レベルを示す尺度としてのみ日本人学習者に合わせるのではなく、CEFR の理念もともに盛り込む必要があるが、日本版は CEFR は、これらをうまく RLD 化させている「フィンランド版 CEFR」を踏襲している (笹島、2013)。

まず、CEFR を日本人学習者に適応させる RLD の動きのなかで、CEFR-J のフレームワークを構築しようとする取り組みが行われ、2012 年に公開された。日本人の平均的な英語能力を CEFR の 6 段階にすると、中学 3 年間はすべて A1, 高校の 3 年間から大学生は、すべて A2

になる可能性がある。日本人の英語学習者の 8 割が A レベルであり（投野、2013）、日本人学習者全体のほぼ全てが 6 段階の下 4 レベル（A1, A2, B1, B2）を占めているので、これらの下位レベルをさらに細かく分ける必要があるとの認識がなされた。そこで、CEFR の下位レベルをより細かく分けている CEFR フィンランド版を参考にして、まず、A1 を3つに分ける(A1.1, A1.2, A1.3)。さらに、A2, B1, B2 はそれぞれ2つに分ける（A2.1, A2.2, B1.1, B1.2, B2.1, B2.2）方法をとって、日本人学習者に適応したレベルの設定を提唱した（岡、2008）。

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

4. 2.2. 日本人学習者に適応する CDS へ

日本人学習者の英語能力の特徴をより理解するための研究としては、CEFR のレベルでテスト結果が判定される言語能力テスト DIALANG の英語版（Alderson and Huhta, 2005）を使って調査したものがある（斉田、2008）。DIALANG の結果、ある日本の大学の 1 年生のリスニング能力は CEFR の A1 レベル、リーディング能力は A2 レベル、ライティング能力は A2 レベル、文法能力は B1 レベル、語彙能力は A2 から B1 レベルとなった。これは、文法や語彙能力は高いが、リスニング能力を含む英語コミュニケーション能力は低いという、日本人英語学習者の一般的な特徴と符合している。また、日本人大学生の大多数が A1～A2 という非常に狭いレベル範囲に入るという可能性を示唆していて、CEFR を日本人学習者に適用させるようにレベル設定をするには、A1、A2、B1 の3レベルではなく、より詳細なレベルを設定したほうが適切であるということを方向づけている。

また、中島・永田（2006）は、この DIALANG の自己評価アンケート、DIALANG self-assessment (SAS) を使用して日本人学習者たちが、各 CEFR の能力記述文に対してどのような困難度レベルとして認識しているかを調査した。この研究を踏まえ、根岸（2006b）は、日本人学習者たちが答えた困難度レベルと CEFR の設定している困難度レベルの間にはっきりとした相違があった項目に注目し、その原因を推測した。そして、例えば学習者が自己評価するとき、体験したことがないことについて質問しても、回答はあまり正確ではないのではないかということを示した（伊東・川口・太田、2008）。このように CEFR レベルと日本人学習者の判定が異

なったCDSに、参考資料を付けることで、学習者が具体的に内容を理解するための工夫として成果をあげ、もともとCEFRが想定していた難易度レベルに近づけることができることがあると報告されている（Negishi、2005； 根岸 2006）。これらの実証研究の結果から、英検 Can-Doリストの能力記述文には、その記述に説明を加えるための典型例を（ ）を用いて説明している。前述の例のように、説明を加えることで、より内容の本質を理解できることもあるからだ。しかし、反対に、能力記述文の内容が特定のことに限定されすぎるという面もある（柳瀬、2013）。

４．３． 英検 Can-Do リスト

4．3.1. 英検 Can-Doリストのなりたち

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

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

柳瀬（2013）によると、英検 Can-Doリストの作成は2003年から、各級、約2000人の任意抽出した合格者を対象に作成のための調査を開始した。4技能別になった能力記述文を被験者に自己評価として5段階（1. ほとんどできない。2. 少しできる。3.ある程度できる。4.だいたいできる。5. よくできる。）で回答してもらう方法で実施した。この調査に先立ち能力記述文の作成は、中学校・高等学校学習指導要領、各種英語検定教科書、英検のテスト課題、さらにACTFL, ALTE, Canadian Language Benchmarks, CEFR, DIALANG Self-assessment List, TOEIC Can-Do Guideなどを参考にして作成された。4技能別に7つの級に分類した能力記述文を作成し、被験者の回答する項目をよりレベルにあった、限定したものにするために、隣接す

る級の項目を共通項目とした 5 フォームの質問紙を作成した。例えば、フォーム1は 1 級、準 1 級、2 級の項目が含まれ、フォーム2には準 1 級、2 級、準 2 級の項目が含まれる。フォーム1とフォーム2の共通項目は、準 1 級と 2 級の項目となる。これにより、IRT を利用して 5 フォームすべての項目を同じ 1 つの尺度に載せることができる。しかし、最終的に採用する各級の能力記述文は、選択肢 3（ある程度できる）以上を選んだ回答者の割合が 80％以上、かつ選択肢 4（だいたいできる）以上を選んだ回答者の割合が 50％以上という条件を設定して選択した。1 つ 1 つの能力記述文が、どの級のものとして採用するかは、この条件を満たしていることが基準となった。

4. 3.2. 英検 Can-Do リストのインパクト

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

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

1. Can-Do リスト作成と選別の過程に専門家グループが関わり、検証する。
2. Can-Do リストの作成過程は、どのように行われたのか公開する。例えば JS ディスクリプターは、4 つの構成要素を決めて 1 つずつの能力記述文をシステマティックに作成し、言語材料参照表で非常に細かく全体の整合性を測った作成過程をウェブで公開している。
3. 調査の質問紙に対する回答方法は 5 件法でなく、4 件法にして真ん中に答えやすい傾向を排除する。
4. Can-Do リストを選抜するとき、IRT で推定された各項目の困難度 θ を基準にする。
5. 調査用の質問紙を作成するとき、Question order effects に配慮する。

４．４． Can-Do 自己チェックリストの Order Effect

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

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

Schuman and Presser (1996) は、アンケート用紙に書かれる質問項目の順番や位置によって起こる回答への影響を order effects（順序効果）と呼び、アンケートを作成するときの重要事項としている。ある順番で質問したとき、その順番が回答に影響を与えることがあるような場合は、そのアンケート結果を一般化することが難しくなる。そして order effects は、内容が似た質問の間で発生する可能性が高い。この影響を数値化して報告している研究はほとんどないが、Duverger (1964) が行った、フランスのある世論調査では order effects による結果への影響は 6％であった。彼らによると、order effects は避けるべきではあるが、似た内容の問題をまとめて質問するほうが、潤滑に都合よく質問できることが多い。従って、order effects を警戒しつつスムーズに質問が進むようにバランスを考えて質問を配置することもできると述べている。

また、Knowles et al (1996) は、質問紙の最初のほうに位置する質問項目は、後に続く項目の背景となって影響を与えると報告している。そして内容が関連のある項目をまとめて一連の項目として質問すると、項目どうしの相互に与える影響は増す。項目がひとまとまりにされることで、回答する側に項目どうしの関連性をより強調して受け取る傾向があるからだ。Roberson and

Sundstrom (1990) の雇用者に対するアンケートの研究では、この order effects は内容によって結果への影響の与えかたが変化し、その中でも給与収入に関するものに最も影響が大きく表れたと述べている。これはおそらく、他の項目よりも回答者に強いインパクトがあるからだと考えられている。最後に、Couper, Traugott, and Lamias (2001) は、実施したウェブアンケートの実験のなかで、関連項目を1スクリーンに納めた場合と、1 スクリーンに1項目にした場合を比較した。1スクリーンに 5 問を載せた時、1スクリーンに 1 問ずつにした場合と比べ 5 問の平均点は、1 スクリーンのほうが低い場合と高い場合があった。しかし、他にも何回か同様の実験をしたが、はっきりとした違いは見られなかった。

本章で解明しようとすること

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

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

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

（3） 被験者にとっては、どのフォームが回答しやすいのか調査する。

4．5．研究方法

4．5.1．被験者

ある日本の大学で必須英語教育プログラムを履修する1年生 627 人（全体の約 8%）が本研究に参加した。表 4．1の通り、被験者らは入学時に、英語プレースメントテスト（リーディング、リスニング、文法）のスコアによって3つの習熟度別レベル（初級レベル：Basic、中級レベル：Intermediate、上級レベル：Advanced）に分けられ、2 年間で約 168 時間の英語の授業を履修する。レベルによって使用する教科書も異なっていて、その習熟度レベルに適応した授業内容を実施することになっている。リスニングコースを1学期間履修する 1 年生の中で、できるだけ全体の比率と近くなるように、各レベルからアトランダムに選んだ学生たちに SCL を実施した。

表 4.1.各フォームのレベル別回答者数

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

4. 5.2. 3 フォームの Can-Do 自己チェックリスト

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

このプログラムのネイティブ教員、日本人教員合計 8 人からなる委員会で、どのような CDS がプログラム履修学生たちに相応しいか話し合った。CEFR の日本語訳（吉島・大橋、2004）を基にして、英検 CDS、清泉アカデミック Can-Do framework（Naganuma and Miyajima, 2006）などを参考にしながら、この大学の履修学生に適応するように CDS を作成した。この作成の過程で、CDS を 3 名の日本人の担当教員が読み返して、日本人大学生に理解できるように書かれているか確認した。

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

次に、SCL の記述文については Negishi (2005), 根岸(2006)での報告に従って、できるだけ具体的な例を示すように努力した。英検 CDS を前例として倣い、典型例を(　　) を用いて説明

する手法（柳瀬、2013）も導入した。最終的に日本語で書かれたリスニングの能力記述文を 30 作成し、リスニングコース担当の教員 3 名が、言葉の言い回しが学習者にうまく理解されるかどうかチェックして完成版とした。被験者たちには、これら 30 問に対して「できない」、「あまりできない」、「まあできる」、「できる」の 4 件法で答えてもらう形式をとった。4 件法を採用したのは、柳瀬（2013）での英検 CDS 作成時などで生じたように、5 件法にして真ん中の選択肢　3「ふつう」を入れると、3 と答える日本人被験者が多くなる傾向があるためである。これらの 30 能力記述文を、作成者側で想定した難易度低～高の順序に並べたフォームLを最初に作成し、フォームC、フォームRを次に作成した。各フォームについては、以下に詳細を示す。また、項目番号はすべてフォーム L のものとする。

フォーム R: アトランダムに 30 項目を質問紙に配置した。内容が同じものをまとめることなく、難易度にも関係なく配置された。

フォーム C: 以下の 10 種類の内容があり、それぞれの種類につき 3 つずつ項目がある。

1. 話す速度と内容、 2. メインアイディア、 3. 指示と説明、 4. 音声教材、 5. 会話、
6. スピーチとレクチャー、 7. クラスルームの英語、 8. ビジュアル教材、 9. 語彙、
10. 文の複雑さ

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

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

4. 6. 結果

4. 6.1. フォームの違い

3 種類それぞれのフォームの 30 項目に対する 4 件法による回答を、1～4で点数化して項目 1から 30 までの平均値を計算した。表 4.2 は、それぞれのフォームごとの信頼性と記述統計で

ある。信頼性は、どのフォームも非常に近い値を示している（0.916 ≦α≦0.932）しかし、この中で最も高かったのは、フォームL、次いでフォーム C であった。

表 4.2.　3 フォームの信頼性と記述統計

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

図 4.1.は、3 つの異なったフォームごとの各項目平均値を表したものである。どのフォームも似た傾向になっていて重なりが多いように見える。次に、これらのデータを、SPSS を使用して一元配置分散分析にかけ、Bonferroni による多重比較による検定を行って、どの水準間に有意差があるかどうか調査した。表 4.3.に示した項目は、5%水準で有意差があるものである。フォーム R と L が最も有意差がある項目が多く 10 項目で、次にフォーム R と C が 6 項目であった。これは、フォーム R だけ他2つのフォームと差があり、フォーム L とは1番違いが大きく、フォーム C とは 2 番目に違いがあるということを示している。

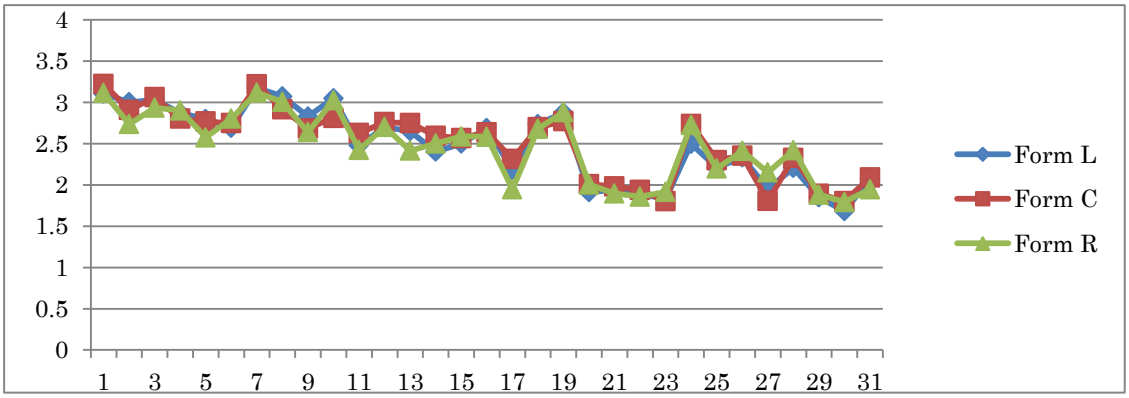


図 4.1.　3 種類のフォーム 30 問に対する回答の平均値

表 4.3.　フォームの違いによる多重比較

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

Bonferroni ,　$p<0.05$ で有意差があるもの＊

4.6.2. フォームごとの習熟度レベルによる違い

次に、フォームごとの習熟度レベルによる違いを調べる前に、学習者の習熟度レベル　初級（B）、中級（I）、上級（A）の違いによって回答がどのように異なるのか確認した。図 4.2. は、それぞれのレベルの回答の平均値をグラフにしたものである。順当に、習熟度レベルが高い　A、I、B の順に「できる。」と答えた人が多かったことが示されている。

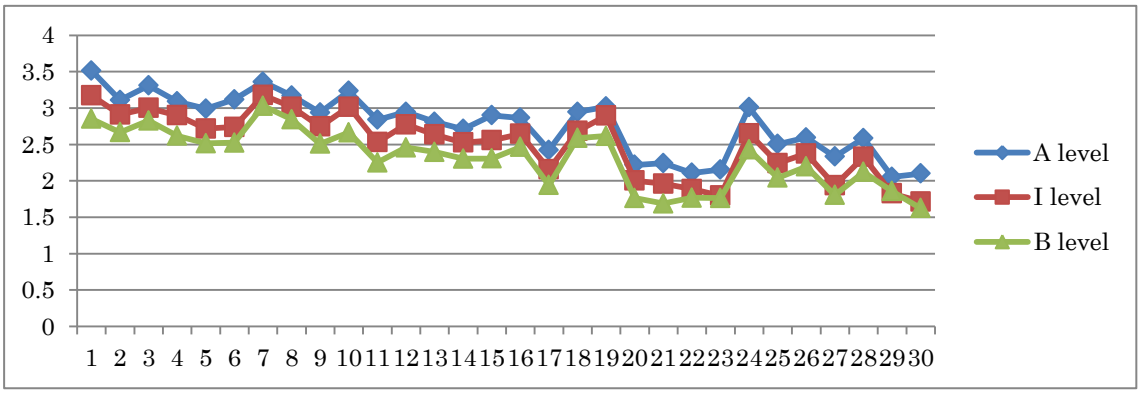


図 4.2. 習熟度レベル毎の回答平均値

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

4. 6.3. フォーム R

さらに、フォームごとに習熟度別レベルの違う被験者の反応を調査した。まず、フォームRに関しては、図 4.3.によると、レベル A と他の 2 レベル（B と I）は違いがあるように見える。これに反し、レベル I と B は非常に似た結果になっている。

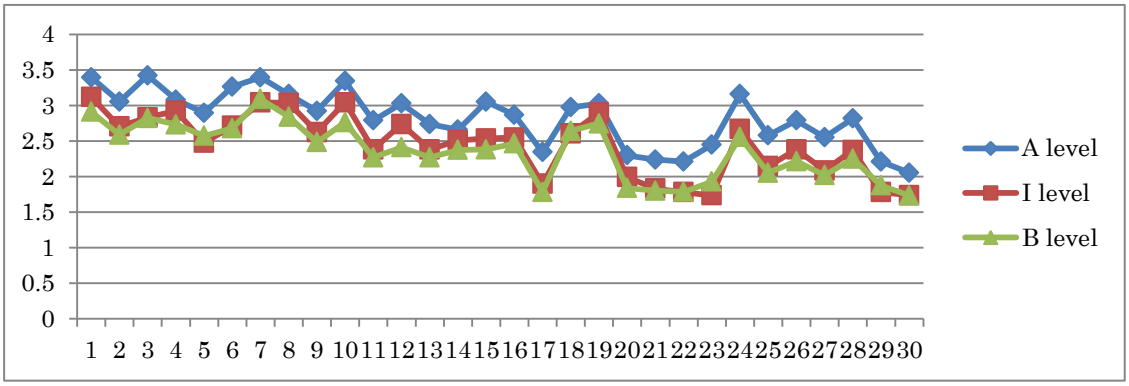


図 4.3. フォーム R に対するレベルごとの回答平均値

これらのデータを SPSS を使用して一元配置分散分析にかけ、Bonferroni による多重比較による検定を行ってどの水準間に有意差があるかを調査した。表 4.4.に示した項目は、5％水準で有意差があるものである。フォーム R の受験者のうち、A と B レベルの受験者間では 20 項目、A と I レベルでは 15 項目、I と B では 4 項目有意差があることを示している。項目がアトランダムに配置されたフォームでは、A レベルの受験者と他の 2 レベルの受験者の回答のしかたに違いがある。

表 4.4.　フォームRの習熟度レベル A, I, B による多重比較

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

Bonferroni ,　$p<0.05$ で有意差があるもの＊

4.6.4. フォームC

図 4.4.では、フォーム R とは反対に、フォーム C では、習熟度レベルが低い B レベルだけ、他の 2 レベルと異なる回答のしかたをしているように見える。

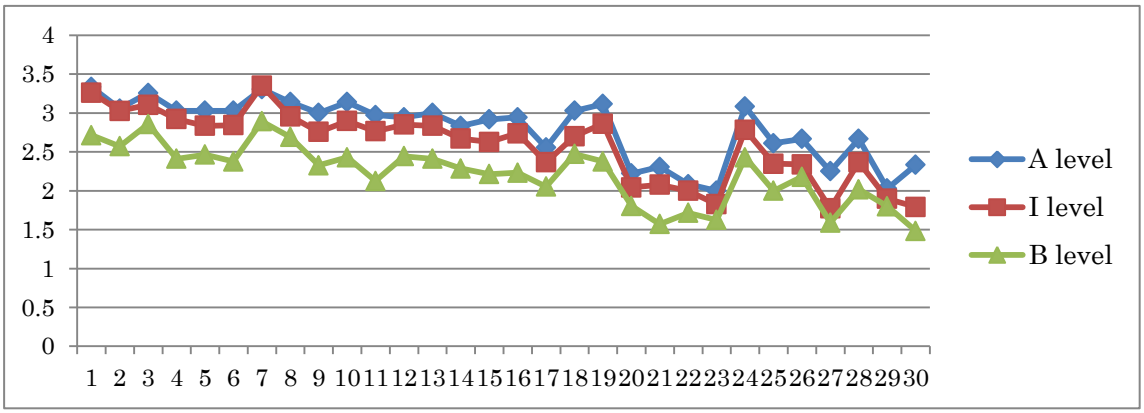


図 4.4. フォーム C に対するレベルごとの回答平均値

これらのデータを一元配置分散分析にかけ多重比較による検定を行った結果、5%水準で有意差があるものは、表 4.5.によると、A と B レベルが 29 項目、 B と I レベルが 20 項目もあるのに対し、A と I レベルでは 2 項目だけであった。 フォーム C では、A とIレベルは、どの項目でも非常に近い平均値を示している。項目が内容別に配置されたフォームでは、B レベルの受験者と他の 2 レベルの受験者の回答のしかたに違いがある。

表 4.5.　フォームCの習熟度レベル A, I, B による多重比較

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

Bonferroni , $p<0.05$ で有意差があるもの＊

4.6.5. フォームL

図 4.5.によると、フォーム L では、どの習熟度レベルも似たような平均値を示しているようである。

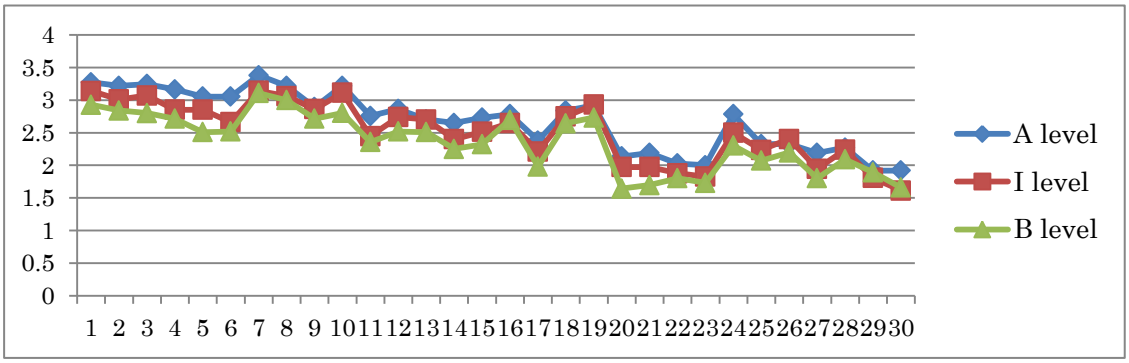


図 4.5.　フォーム L に対するレベルごとの回答平均値

表 4.6.　フォームLの習熟度レベル A, I, B による多重比較

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

Bonferroni ， $p<0.05$ で有意差があるもの＊

表 4.6. に示す一元配置分散分析の結果では先のフォーム R,C に比べ、全体的に、どのレベル間においても有意差がある項目が少ない。最も有意差がある項目が多いのは、レベル A と B で 14 項目あるが、レベル I と B では 5 項目、レベル A とIでは 2 項目だけであった。特に困難度が高いと推定される後の方の項目（22～30）ではどのレベル間においても有意差がなかった。これは、フォーム L を使用すると、どの習熟度レベルの被験者も、大きな違いがなく回答する可能性を示唆している。

4.6.5. 使いやすさ

最後に、表 4.7.は3つのフォームごとに、どれが最も使いやすいか尋ねた時の回答の平均値である。大きな差ではないが、フォーム L、フォーム C、フォーム R の順に使いやすいという回答を得た。しかし、これらのデータを一元配置分散分析にかけて多重比較した結果、どのフォーム間においても有意差はなかった。回答を習熟度レベルごとに調査した結果は、初級（Basic）、中級(Intermediate)、上級(Advanced)レベルの学生の順に使いやすいと回答している。

表 4.7. フォームごとの使いやすさ(レベル別)

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

4．7．考察

4. 7.1. 3 種類のフォームによる違い

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

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

4.7.2. フォームごとの習熟度レベルによる違い

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

4.7.3　フォームごとの使いやすさ

非常に少ない差ではあるが、3つのなかではフォームRについて「答えにくい」と回答した被験者が多いという結果になった。最も「答えやすい」のは、フォームCで、僅差であるが次いでフォームLである。フォームRが「答えにくい」のは、フォームCやLに比べて追加の情報がないからであろう。フォームCのように、同じ内容の項目がまとまっていると、1つずつ項目に答えるより、関連づけることで同じ内容の他の項目に答えやすくなると考えられる。また、SCLの場合は、できる/できないで答える性質上、フォームLのように難易度が予測できると「答えやすい」と感じるのも道理である。

4.8. 結論

本研究において、「項目配置順の異なる SCL3種類のフォーム」の分析結果から、難易度順や内容別のフォームを使用した被験者は、アトランダム順のフォームを使用した被験者と異なる結果を示し、フォームの項目配置順序が結果に与える影響（question order effects）がある可能性が高いことが分かった。また、習熟度レベル別の被験者の反応を、フォームごとに比べた場合、習熟度の高い被験者ほど低い被験者に比べ、フォームの違いによる影響を受けにくい可能性が示唆された。フォームは、内容別、難易度順フォームがほぼ同じくらい使いやすいが、アトランダム順のフォームは他のフォームより使いにくいと被験者たちが感じる傾向があることも判明した。これらを総合すると、項目を難易度順に並べたり、内容ごとにまとめると、question order effects の影響は受けやすいが、スムーズに回答しやすくなると考えられる。つまり、本研究の結果は、order effects に対する考慮と被験者が円滑に回答しやすいこと、この両面のバランスを考えて SCL の項目を配置するべきであることを示唆している。

Question order effects を回避するために考えられる対処としては、難易度順や内容別に項目を並べる場合、できるだけ難易度順や内容別であることが一目瞭然にならないフォームを作成する、また、order effects を避けるために項目はアトランダムに並べるが、一つ一つの項目を簡単明瞭にして回答しやすくする、等が考えられる。そして、より妥当性・信頼性が高い SCL フォームを作成するには、対象とする学習者の習熟度レベルに合わせたり、その学習内容や到達目標に合わせたり、千差万別の対応をその都度考える必要があるが、今後 SCL を作成する際は、order effects と使いやすさのバランスを考慮することを付け加えたい。

短い SCL の利点

本章の結果から、明らかに難易度の異なる項目を並べるよりも、難易度の異なる項目をできるだけ少なくするほうが、order effects の影響を回避することができることが分かった。これは、第 3 章、3. 4.2 の「14 項目の 3 つのレベル別 SCL」で述べたこととも繋がる。IRT を用いて3つの難易度別 SCL を同一尺度化することで、回答者のレベルにあった 14 項目しかない SCL であっても他のレベルの SCL との差が分かる。学習者は自分のレベルだけの 14 項目だけの SCL に答えることで、多くの項目の SCL に答えることに比べ得られる情報量は減るという弱点もあるが、次の4点（1）学生の負担減、（2）回答者のレベルに合った SCL に回答できる、（3）授業の到達目標として記憶・認識しやすい、（4）授業中に複数回実施しても支障が少ない、の優位性が考えられる。加えて、難易度が均一の SCL であるので question order effects の影響は少なくなる。

第５章　事前事後テスト

５．１．はじめに

大学教育の「質の保証」の必要性が緊急課題である昨今、受講生たちにどれだけ教育効果をもたらすか検証すること（プログラムアウトカム評価）（安田, 2010）への注目度が高まりつつある。このアウトカムがわかりやすい形で表わせるものの一つとして、プログラムを履修する前と後でどのように受講生の能力が変化したのか測る事前事後テストがある。大学英語教育プログラム評価のなかでも、テスト理論の利用によってプログラム評価の証拠として提示しやすいのが、リスニングやリーディングなどの客観的テストによる事前事後テストの結果だと言える。プログラム評価についてあまり関心のない大学上層部は、この事前事後テストの結果がプログラム評価の全てだと思っていることさえある。

しかし、日本の大学教育の中では、この事前事後テストシステムを構築する試みがあまり実施されていない。おそらくその理由には、大規模なテストを 1 学年全員に事前と事後の2回実施することで大きな労力とコストがかかること、また、テストデータに最適な方法を選択してテストの等化や結果分析をするための専門知識が必要とされること。そしてさらに、教員作成によるテストの場合は、質のコントロールが難しいことなどが挙げられる。

本研究で対象とする大学英語教育プログラムでは、履修学生全員である約 5,000 人～6,000 人が事前テストを受験する。これには約 150 人の試験監督と100 室以上の教室を要する。事後テストにおいても、これと同じ規模で実施することは非常に難しい場合が多い。しかし、項目反応理論（IRT）を利用することによって、事前テスト受験者の約 14％にあたる約 700 人のみが事後テストを受験するとしても、事前事後両テストの受験者の能力値(θ)の変化は推定でき、プログラムアウトカム評価の証拠のひとつにすることは可能である。

本研究では、このように受験者数が事前事後で大きく異なる2つのテストを共通項目デザイン（Angoff, 1984; Ayala, 2009; Kolen and Brennan, 2004）で等化するとき、最も適切な IRT モデルと等化法を選択する方法を提案する。2通りの項目反応モデルと5通りの等化法を組み合わせて 10 種類の等化の方法（ここでは項目反応モデルと等化法の組み合わせを、等化の方法と呼ぶことにする）を比較する。この中には、教育現場の教員が比較的容易に計算できる等化法も組み入れた。

また、ここで使用した事前事後テストは、実際に教えている内容をもとに教員が作成した英語

リスニングテストである。これらのテスト問題は中間テストや期末テストとして英語プログラムで過去に使用され、既に分析されて項目パラメタも分かっている。これらの問題の中から適切な項目を選び、新たに問題を作成せず編集した2つのテストフォームを事前事後リスニングテストとして使用している。これらの事前事後テストを通じて推定された受験者の能力値(θ)を利用して、学年ごとの能力値平均の推移を観察し、また、プログラムニーズに関するアンケート結果と比較して、能力値を伸ばした学生のニーズを理解することにより、有効なプログラム評価へ繋げることをめざしている。

このように、本研究は、日本の大学教育機関での事前事後テスト実施を妨げていると考えられる多大な労力とコスト、高度に専門知識を要するテスト分析、困難なテストの質のコントロールなどの問題点を少しでも改善し、より有効性が高く効率の良い事前事後テストシステムを構築し、プログラム評価のための重要な手がかりとなることを目的としている。

５．２．　先行研究

5. 2.1　IRT モデルと等化法の選択

項目反応モデルの選択

本研究で使用した事前事後英語リスニングテストは、ともに 4 択の多肢選択問題である。このことだけに着目すると、あて推量パラメタと呼ばれるCパラメタを含む 3PL を用いて分析するのが一般的である（e.g., 大友, 1996; Hambleton, Swaminathan, and Rogers, 1991）。しかし、受験者数を見ると、本研究の事前テストの受験者は約 5100～6600 人であるのに対し、事後テストは約 700 人と大きく異なるため、どの項目反応モデルを使うか慎重に吟味する必要がある。項目パラメタを安定して推定するために必要な受験者数は、項目反応モデルによって異なり、普通は 2PL モデルで 500 人から 1000 人（e.g., Ayala, 2009）、3PL で 1000 人（Lord, 1968）と言われている。3PL モデルのほうがより多い受験者数を必要とする理由として、項目ごとのパラメタ数が多いということに加えて、3PL モデルは、あて推量パラメタの影響を受けて、能力推定値が高い受験者の情報量を過大評価し、能力推定値が低い受験者の情報量を過少評価する傾向が挙げられる。従って、2PL モデルより 3PL モデルの有効サンプル数は少なくなるので、受験者数がかなり多い場合を除き 3PL の項目パラメタ推定値のほうが不安定になる傾向がある（張, 2009）。

本研究のデータは、事後テストの受験者数である約 700 人が、3PL モデルが要求する 1000 人に満たないために 2PL モデルの方が安定した推定値を得られる可能性が強い。しかし、すべ

て4択の多肢選択問題なので c パラメタの情報も得られれば、近い将来、項目バンクを構築する場合に利用できる可能性がある。また、今後継続的に事前事後テストを実施する予定があるので、ここでどの項目反応モデルを採用するかによって、分析結果の精度が大きく変わることがあるので、データや目的を良く考慮して、どの項目反応モデルが最も適応しているか慎重に吟味する必要がある（e.g., Choi and Bachman, 1992；Kolen and Brennan, 2004)。従って、本研究では2PLと3PLの2種類の項目反応モデルを使用して、どちらが本研究で扱うデータに適合するか検討する。

等化法の選択

本研究で使用する事前事後リスニングテストの等化は、事前テストと事後テストに同じ問題が10問ずつ含まれる共通項目デザインである。つまり、共通の10項目を繋ぎとして、2つの異なったフォームの事前事後テストを等化し、比較可能な統一尺度に乗せて、受験者の能力値変化を推定する。本研究の事前事後テストは、プログラム評価を目的としているので、等化するとき、共通項目のパラメタは変化しないが、受験者の能力は変化するとした解釈を前提にする。

この共通項目デザインで項目パラメタ値を等化するとき、事前テストと事後テストを独立に分析したとすれば、等化前の受験者の尺度値を θ、項目識別力パラメタ値を a、困難度パラメタ値を b、として等化後の尺度値 θ、パラメタ値 a、b にアスタリスク*をつけ θ^*、a^*、 b^*で表すとすると、これらの関係は

$$\theta^* = k\theta + l \qquad (1)$$

$$a^* = a/k \qquad (2)$$

$$b^* = kb + l \qquad (3)$$

となることが知られている（e.g., 芝, 1991; 大友, 1996）。等化をするには、この等化係数 kと lを推定する必要があり、多くの方法が提案されているが、本研究では① mean-mean (MM), ② mean and sigma (MS), ③ Haebara (HB) (1980), ④ Stocking and Lord (SL) (1983) ⑤ Calr (CR)（前川, 1991; Arai and Mayekawa, 2011；光永・前川, 2012）の5種類を選び本研究のテストデータを使用して、どの方法が最も適切かを評価することにした。

これらの等化法は、2通りに分けることができる。まず、上記の① MM や ② MS のように、項目困難度の推定値の平均値や標準偏差を利用する方法で、項目困難度の推定値の平均

値mbや標準偏差sbを使って等化係数の推定値 $\hat{k}$, $\hat{l}$を導き出す方法である(Marco, 1977)。

次に、③ HB, ④ SL, ⑤ CR, のように識別力パラメタと困難度パラメタ推定値から得られる項目特性曲線の情報を利用して等化係数を推定する方法もある。これらは、項目パラメタ推定値$\hat{a}$, $\hat{b}$を前述の式(2)や(3)で変換した値で描かれる項目特性曲線と、項目パラメタの推定値$\hat{a}*$ $\hat{b}*$で描かれる項目特性曲線をできる限り近づけるようにして等化係数を求める方法である。

等化に関するこの 2 通りの方法のうち、前者は計算が容易であるが、推定精度の良くない項目の影響を受けやすく、項目特性曲線に基づいた方法のほうが正確な推定ができると言われている（e.g., 芝, 1991）。しかし、本研究においては、項目困難度の推定値の平均値や標準偏差を利用する方法を敢えて比較対象に含めることにした。その理由は、Saida and Hattori (2008) などの日本の教育現場での等化に関する先行研究に MS 法が使われていることや、実際の教育現場ではテストの最新専門知識を特に持たない教員がテストの分析をすることが多いことを考慮したからである。

等化法の比較の研究

IRT によるテストの等化法を比較するシミュレーション研究のなかで、共通項目デザインによる等化では、その共通項目の数や質、テスト全体の中での共通項目の困難度、そしていかにテスト全体をその共通項目が代表しているか、などが等化結果に影響を及ぼす要因だと言われている（Kolen and Brennan, 2004)。

Kim et al. (2008) はランダムグループデザインと、共通項目デザインというデータ収集方法の異なる4通りの等化法を、実際に実施された 14 種類の異なる内容のテストで比較した。その結果、4 つの等化法による結果の違いには、受験者数の違いが影響していると指摘している。また、共通項目デザインの等化では、共通項目の数や質が等化結果に大きく影響することも報告している。

Zu and Liu (2010) は、共通項目デザインによってテストを等化するとき、共通項目に含まれる 2 種類の異なるタイプの項目の配分が、テスト本体と類似していることが等化結果にどのような影響を与えるかシミュレーションした。1問ずつ独立した項目と1つの長文問題に付属する一連の項目からなるテストを等化するとき、従来から言われてきたように、共通項目がテスト全体のミニチュア版であるべきであるという定説に従えば、独立した項目と一連の項目が、テスト本体と同じくらいの比率で共通項目に含まれるべきである。そこで、独立した項目と一連の項目の配分率を変えた5通りの共通項目によって、テストを等化するシミュレーションを実行した。その結

果、共通項目とテスト本体の類似性にかかわらず、より多くの一連の項目を共通項目に用いた場合ほど、等化誤差が大きくなることが示された。これは、適切な等化を行うには、共通項目がテスト全体のミニチュア版であることよりも、独立した項目をより多く共通項目に含めるほうが、等化誤差を小さくする可能性があることを示唆している。

さらに、異なる等化法による誤差の違いを比較した研究としては、Lee and Ban (2010) がある。彼らは項目特性曲線に基づいた 2 種類の等化法である SL 法と HB 法によって等化を行い、その結果を比較したところ HB 法のほうが SL 法より等化誤差が少なくなった。また、Jodoin, Keller and Swaminathan (2003) では、MS 法と共通項目パラメタを固定した等化法を実際のデータを用いて比較したところ、受験者の習熟度による影響はあるものの、2種類の等化法の結果に大きな違いがなかったと報告している。そして Keller, Keller, and Parker (2011)は、受験者を習熟度別に分けて MM 法, MS 法, HB 法, SL 法, その他2つの方法を含む、6 通りの等化法を比較したが、やや誤差が大きかった MM 法以外はほとんど同じ結果になった。

このように IRT による共通項目デザインの等化法を比較する研究は、共通項目の違いが等化結果に及ぼす影響に着目するものがほとんどである。これらの研究はテスト開発者や研究者主導で行われていて、実際のデータを使わず仮定されたモデルのもとで乱数を発生させて得られたシミュレーション研究も多い。これらの研究者たちの多くは、結果の考察として、シミュレーション研究は方向性を示唆するが、すべての事例にもれなく適応するものとも限らないとしている。そして結論として、ガイドラインとしてのシミュレーション研究に、実際のデータを使う事例研究を合わせて実施する必要があることを強調している（e.g., Kim et al., 2008; Keller, Keller, and Parker 2011)。

5. 2.2. 英語教育プログラム評価

安田（2010）は、プログラムとは特定の社会的・教育的目標を達成するために、人が中心となって介入やアクションを行う事業と定義している。本研究におけるプログラム評価クエスチョンは、「学生の英語能力育成が成果をあげているか？」であり、これを裏付ける客観的根拠が求められる。また彼はプログラム評価とテスト科学の専門性の関係について、テスト科学は、アウトカムの特定や指標の作成、実験デザインの選定、データ解析などプログラム評価の中核となる部分での証拠の提示において貢献度は非常に高いと指摘している。しかし、プログラム評価において、テスト科学の焦点である測定の問題は必要条件であるがそれが十分条件では必ずしもないと言及している。

また、プログラム評価のモデルについて植山（2004）が紹介した Context-Input-Process-Product (CIPP)モデルは、プログラムの目標、計画、行動、結果の評価という一連のプロセスを基に考え出された評価方法である。CIPP の中には、Context, Input, Process, Product という4タイプのモデルがあり、目的に応じて使い分けられている。この中で Context, Input, Process の3モデルは、学習・教授・運営の計画実行プロセスから得る情報を形式的評価するものである。しかし、Product モデルだけは、学習・教授・運営が終了した段階で、期待した成果が得られたかどうかを総括的に評価することが特徴である。したがって Product モデルについては、実施過程で得られた結果をデータの質的・量的分析を行って的確に測定することが求められる点で、テスト科学の貢献度は大きいと考えられる。さらに分析結果をプログラムニーズや目的などと照合する方法も用いられている。

本研究では、事前事後テストを継続的に実施することで、そのプログラムで学んだ学生の能力変化を学年ごとに比較し、プログラムの改定による影響を、質的・量的分析を実施して検証することも視野にいれている。さらに、事前事後テストで得られた学生の能力変化を、クラスサイズ、教員の属性（例：経験年数、学位の種類など）、や教材の違いなどを変数として比較したり（Boldt and Ross, 2005)、プログラムニーズアンケートと照合することも可能である。このようなプログラム評価のための試みは、従来の教育プログラムを漫然と実施していくのではなく、より学生のニーズに合った有効性の高い教育プログラムに改善していく手掛かりを探るという大学側の姿勢を広く社会に示すものでもある。

5 .2.3.　教員作成の事前事後テスト

教員作成の事前事後テストには、1．8で述べたような多くの利点がある。特に、プログラム評価のための事前事後テストには、その妥当性を考慮すると、その英語教育プログラムで学習した内容を、教えた教員が作成する事前事後テストが理論的にもっともふさわしい。しかし一方で、教員が一からテスト問題を作成することは、非常に大きな労力と時間を要する。そのうえ、予備テストとして実施することもなく教員作成のテストを使用することは、テストの質の保証がなく危険である。そこで本研究では、教員作成による過去に使用されたテストを利用することにした。過去に使用され、既にそのテストの困難度や識別力が分かっている問題の中から、適切な問題を選び事前事後テストとして実施する。

5．2.4　プログラムニーズのためのアンケート

Busch, et al.（1992）は神田外語大学のカリキュラム改革の一環として学生 348 人にニーズ調査を実施した。これは教師中心ではなく、学生中心のカリキュラムに近づけるための改革の一つとして、大学側がもっと学生の声を聞くために実施された。全部で 112 項目の質問に対して、40 分間で回答する大がかりな調査であった。学生たちが賛成するかどうかを 9 段階で答えるもので、個々の回答の平均値で結果を表している。外国語を学ぶ日本の大学生たちは、全般的に学習者中心のコミュニケーションに重点を置いた授業を望んでいるという傾向が見られた。

Kikuchi (2005)は、日本大学国際関係学部の 434 人の学生と 53 人の教員にアンケートに答えてもらいニーズ分析した結果を報告している。彼の学生に対する調査表は、前述した Busch, et al. (1992) の、学生たちの学習スタイル、教員や評価に対する希望に関するニーズアンケートの質問の中から、対象とする日本大学の英語教育プログラムに適さない質問を取り除いて作成された。このアンケートは 100 問の質問が日本語で書かれ、回答者は 5 段階で回答するものである。これらの質問の中で最も同意を得たのは、「先生が授業を楽しくしてくれるとよく勉強できる。」であった。また、彼はこの結果を3つの習熟度別グループに分けて集計した。その結果、ひとつだけグループ間で大きく違った質問結果は、初級レベルの学生だけが外国人より日本人教員を好んだ点であると報告している。

5．2.5　本章で解明しようとすること

事前テスト受験者は 5000 人以上だが、コストや労力を抑えるために、事後テスト受験者をその 14％である約 700 人で実施するようなとき、2PL モデル、3PL モデルと5つの等化法合わせて 10 通りのうち、どの方法を利用するのが最適であるのか決定する方法を検討する。また、最適な等化の方法と、最も適さないと判断された方法で、実際のデータを等化して受験者の θ を推定したときに生じる差について調査する。

この事前事後テストによって推定された受験者の θ を利用して、どのようなプログラム評価の証拠を提示することができるのか検討し、習熟度別に集計した受験者の平均能力値変化や、受験者 θ の伸びとプログラムニーズアンケート結果の関係が、プログラム評価の証拠になる可能性はあるのか探る。

日本の教育機関で実施しやすいシンプルで、労力やコストを省いた「プログラム評価のための事前事後テストシステム」の実現可能性を念頭に置き、θ を利用してどのようなプログラム評価の証拠を提示できる可能性があるのか検討する。

５．３．　研究方法

5. 3.1　受験者

ある大学の同じキャンパスにおいて2年間の必須英語教育プログラムを履修する1年生のほぼ全員（（注1）年度 01 は 6630 人、年度 02 は 5157 人）が事前リスニングテストを受験した。このプログラムでは、新入生は事前テストの結果で上級(A)　中上級(U)、中初級(L)、初級（B）の4つのレベルに分けられ、それぞれの習熟度に適した教材を使用して2年間で 168 時間の必須英語プログラムを履修する。クラス分けの為のテストと言うと、学生たちが低いレベルのクラスに入って楽をしようとする傾向がある。そのために、この事前テストを、「基礎学力テスト」と称し、英語、国語、数の3科目からなる入学時の実力を測定するためのテストのひとつと位置づけている。基礎学力テストのなかの英語テストが、英語必須科目のクラス分けテストであることは学生には告げていない。

この事前テストを受験した全学生のうち、年度 01 は 688 人、年度 02 は 772 人が事後テストをプログラム終了時に受験した。上記4つの習熟度レベルそれぞれの約 10～20％ずつの学生をクラス単位で抽出し、標本の代表性を保つように考慮した。いずれの年度も、事後テスト受験者の事前テストの平均能力値は、事前テスト受験者全体の平均能力値とほぼ類似している。

5．3.2.　テスト

事前事後リスニングテスト

最初に、6630 人が年度01に事前リスニングテスト30項目を受験し、翌年この事前テストの分析結果を見て、以下のような点に配慮して共通項目を10項目選んだ。まず、テスト全体の内容の代表性を表す項目であるか、そしてひとつの長い会話に続く一連の項目より、独立した 1 問ずつの項目であるかどうかに配慮した。また、事前テスト 30 項目の識別力パラメタ(a)と困難度パラメタ(b)をできる限り全問題の平均値（a = 0.390、b = 1.123）に近くなるように共通項目を選んだ結果、10 項目の a パラメタの平均値は 0.431、b パラメタの平均値は 0.077 となった。過去に実施した教員作成のリスニングテストの中から 20 問をこの 10 問に加え 30 問の事後テストを作成した。この教員作成テストの 20 問は事前テストの内容と重複がなく、事前テストの項目より困難度がやや高くなるように選択した。

図 5.1. は、事前事後テストを図解したもので、縦が受験者、横がテスト項目を表している。事前事後テストともに 4 肢からの選択、30 問のリスニングテストで共通項目 10 問が含まれている。

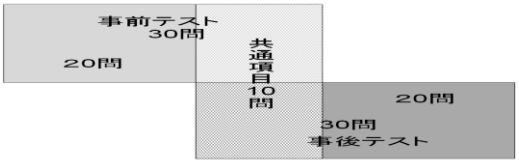


図 5.1. 事前事後テスト

年度01生に実施した事前事後リスニングテストと年度02生に実施した事前事後リスニングテストは同じテストである。また、この事後テストの 20 問は既存の教員作成によるテストの一部であるが、これを共通項目 20 問として等化して項目プールを作成することを計画している。

項目反応理論を用いて異なるテストに含まれる項目パラメタの等化をする際には、そこに含まれる共通項目のパラメタの値が不変であること、もしくは、各テストを受験した受験者の能力値が不変であることを仮定する必要がある。本研究においては、事前テストと事後テストの実施の間に2年の隔たりがあり、しかもその間に受験者が英語を学習していることから、2時点における受験者の能力が不変であると仮定することには無理があると考えられる。他方、共通項目に関しては、対象がリーディングと比べ比較的記憶に残りづらいと考えられるリスニングテストであることや2年の間隔の長さから、両方のテストに含まれる 10 項目の共通項目の項目パラメタは事前テストの時点から変化していないものと仮定することに、それほど無理はないと考えられる。

このような観点から、本研究では、事前テストと事後テストの差異を、それを受験した受験者の能力値の変化として捉える立場を取った。しかし、事後テストから推定された能力値に関しては、この共通 10 項目の記憶効果が加算的に作用している可能性はゼロではない。

プログラムニーズアンケート

このアンケートは、事後テスト終了後、同じ教室で授業時間内に実施する予定であったため、回答する時間は 10 分以内と考えられた。そこで、Busch, et al.（1992）をもとにして作られた Kikuchi (2005)の学生のニーズ調査のための 100 問のアンケートから 19 問と、独自の問題 6 問を合わせて 25 問のプログラムニーズアンケートが作成された(Appendix C 参照)。学生は、自分たちの授業での学習方法(10 問)、先生の指導方法(7 問)や評価(7 問)に関する3つのカテゴリーの質問に対して、同意するかどうかのレベルを4段階で回答した。そして、最後に1問、TOEIC, TOEFL, 英検の受験の有無を聞く質問 25 を新たに設けた。Kikuchi (2005)のアンケートからの 19 問は、3つのカテゴリーからほぼ同数になるように選んだ。アンケートを実施する英

語教育プログラムの現状に適合し、そこで最近実施されたプログラム改革に関連があるものを優先的に選択した。

新たに作成した6問のうち前述した質問25以外の5問は、その英語教育プログラム独自の内容（質問 10）や、プログラム改革に関連ある質問である。例えば、質問8「ひとつのスキルだけ（「リーディング」ならそれだけ学ぶ）の授業がよい。」、質問9「聞く、話す、読む、書くあわせた総合英語の授業がよい。」は、プログラム改革後、2スキル合体の授業になったことを反映した質問である。さらに質問 17 と 23 もプログラム改革後、日本人と外国人教員が同じコースを担当することになったこと、テストが一学期に1回から2回に増えたことを反映させた。また、このアンケートは、年度01、02ともに事後テストの解答用紙に印刷したので、回答者数は両年度とも事後テスト受験者数とほぼ同じである。

5. 3.3.　最適な IRT 項目パラメタモデルと等化法の選択

本研究のデータを分析するのに最適な等化の方法を探るため、2通りの項目反応モデル（2PL, 3PL）と、5通りの等化法：①MM, ②MS, ③HB, ④SL, ⑤CR これらをそれぞれに組み合わせて、合計 10 種類の方法の中から最適な方法を選択するための評価を実施した。項目反応モデルの選択や等化法の比較の基準としては、項目パラメタの値の安定性を用いることも可能であるが、本研究では、教育機関のプログラム評価における θ の果たす役割を考慮して、その基準として θ に準拠したものを採用することにした。

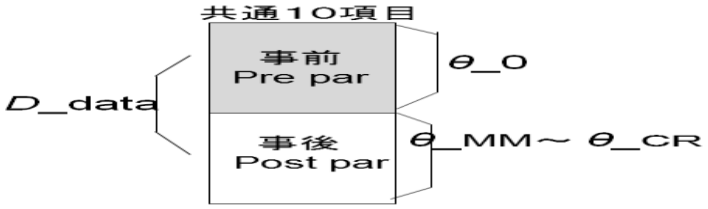


図 5.2. 等化の方法の評価

まず図 5.2. にあるように、評価の前提として共通項目に着目した。事前テスト 30 項目のデータを BILOG-MG3（2003）で2つの項目反応モデルを使ってそれぞれ分析し、30 項目のパラメタの中から、共通 10 項目のパラメタ（これらを Pre-par とする）を取り出す。また、同様に事後テスト

30 項目のデータを2つの項目反応モデルによって BILOG-MG3 で分析し、推定された 30 項目のパラメタから共通 10 項目のパラメタ（これらを Post-par とする）を取り出す。これらの2種類のパラメタ（Pre- and Post-par）を使って前述の①～⑤の 5 通りの等化法により事後テストを事前テストに等化する。

ここで事前テストと事後テストのデータから共通項目 10 項目分の項目反応を抜き出し、これらを縦に繋いで結合したデータファイル（D-data）を作る。この D-data を用いて、事前テストのパラメタの共通項目部分（Pre-par）を使用して、すべての受験者の2通りの θ を計算し、これらを 2PLθ_0 と 3PLθ_0 とする。

さらに、①～⑤の方法で等化した事後テストのパラメタの共通項目部分(Post-par)を使い、D-data のすべての受験者 θ を計算した。そして、分かりやすくするために使用した等化法の名前を θ に付け 2PLθ_MM, 2PLθ_MS, 2PLθ_HB, 2PLθ_SL, 2PLθ_CR とし、3PL モデルの5つの θ にも同様に 3PLθ_MM～3PLθ_CR のように名づけた。もしもモデルの選択が正しく、しかも等化が適切に行われていれば、これらの θ の値、例えば、2PLθ_0 と 2PLθ_MM~2PLθ_CR は、理論的には等しくなる値である。従って、2つの項目反応モデルごとのすべての受験者の θ_0 を真値（ θ_0 ）とし、これらのモデルの違いと等化法の違いからなる10 通りごとの受験者 θ($\widehat{\theta_i}$)との平均 2 乗平方誤差（RMSE：Root Mean Square Error）(4) 式を計算した。10 通りの RMSE の中で、最も 0 に近いものが本研究データの分析に最適な方法だと考えることができる。

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} \Sigma(\hat{\theta}_i - \theta_{0i})^2}{n}} \qquad (4)$$

5．4． 結果

5．4.1． IRT モデルと等化法の決定

RMSE による評価

表1は年度01、年度02の事前事後テストの等化係数を、共通項目デザインを用いて2種類の IRT モデルごとの5つの等化法によって算出したものである。まず、BILOG-MG3 (2003) を用いて、2PL と 3PL モデルそれぞれの共通 10 項目のパラメタを推定した。次に、*R*を用いて PLINK Version 2 (Purcell, 2007) パッケージを使用し、表 5.1. にあるように、合計 10 種類の等化係数を計算した。

表 5.1. 年度01と年度02の事前事後テストの等化の方法10種類の係数

モデル	等化法	年度 01		年度 02	
		k	***l***	***k***	***l***
2PL	① MM	1.293	0.656	1.386	0.451
	② MS	1.112	0.575	1.359	0.442
	③ HB	1.224	0.681	1.371	0.466
	④ SL	1.251	0.622	1.376	0.482
	⑤ CR	1.230	0.613	1.375	0.491
3PL	① MM	0.993	0.473	1.171	0.334
	② MS	1.079	0.461	1.195	0.328
	③ HB	1.043	0.455	1.172	0.352
	④ SL	1.069	0.501	1.187	0.377
	⑤ CR	1.036	0.502	1.157	0.389

表 5.2 は、表 5.1. の等化係数を使用して、5. 3.3 で前述した等化の評価方法によって、2PL モデルと 3PL モデルごとに θ_0、ならびに θ_MM～θ_CR の値を計算して RMSE を求めたものである。この表の中で、RMSE が最も0に近い方法が最適の方法であると考えられる。まず、年度01と02の項目反応モデルによる違いについては、年度01，02ともに 2PL が 3PL より小さい RMSE を示している。年度01での 2PL による5等化法すべての RMSE 平均は 0.065、年度02では 0.090、2PL 両年度の平均は 0.078 となるのに対して、3PL での年度01における5等化法 RMSE 平均は、0.105、年度02では 0.101、3PL 両年度の平均は 0.103 となり、2PL モデルのほうが平均で 0.025 小さい RMSE を示すことが分かった。さらに、5種類の等化法については、MS法の RMSE が一番高く、その平均値は 0.093、次に MM 法と HB 法で 0.091、SL 法は 0.089、最も RMSE が小さい等化法は CR 法の 0.088 であった。

これらの結果から、本研究のデータにおいて 10 種類の等化の方法のなかで、RMSE が最も小さくなるのは、2PL モデルで CR 法によって等化した場合（2PL－CR）で、RMSE が最大となるのは 3PL モデルで MS 法によって等化した場合（3PL－MS）であることが分かった。しかし、5種類の等化法間での RMSE の差は非常に小さく、最小の CR 法と最大の MS 法との差はわずか 0.047 であった。

表 5.2.　年度01と年度02の 10 通りの等化の方法の平均値 θ RMSE 比較

モデル	等化法	年度01	年度02
2PL	① MM	0.065	0.092
	② MS	0.066	0.092
	③ HB	0.065	0.090
	④ SL	0.065	0.089
	⑤ CR	0.064	0.089
3PL	① MM	0.104	0.102
	② MS	0.111	0.103
	③ HB	0.110	0.100
	④ SL	0.102	0.099
	⑤ CR	0.100	0.099

なお、(4) 式に現れる二つの θ の差の2乗の、モデルと等化法による2元配置の分散分析の結果（表 5.3. 表 5.4. ）は、各年度とも、モデルも等化法もきわめて小さな p 値で有意な結果となった。しかし、等化法の MS 値はモデルの MS 値の 50 倍以上となり、各年度とも受験者数が大きいこと (N=6630, 5157) を考慮すれば、D-data に関する限り、RMSE に及ぼす等化法による違いは実用的に無視できるほどの大きさであると考えられる。

表 5.3. RMSE の分散分析表　年度 01

	平方和 *SS*	自由度 *df*	*F*
モデル	0.874	1	6124.268**
等化法	0.017	4	29.896**
モデル×等化法	0.014	4	24.049**
残差	10.434	73100	

**有意確率　$p < 0.01$

表 5.4. RMSE の分散分析表　年度 02

	平方和 *SS*	自由度 *df*	*F*
モデル	0.0600	1	322.6468
等化法	0.0043	4	5.7597
モデル×等化法	0.0001	4	0.1018
残差	11.1165	59750	

**有意確率　$p < 0.01$

2 通りの方法での学生平均能力値 θ

最適とされたモデルと等化法の組み合わせである 2PL-CR と、最も適切でないと評価された 3PL-MS により計算された等化後の 30 項目の項目パラメタを用いて、リスニング事前事後テストを両方受験した年度01の 688 人と、年度02の 772 人分のデータを分析し、それぞれの θ を計算した。表 5.5 は、2PL-CR 法で推定した年度01と年度02の事前と事後の間の学生平均能力値 θ の変化である。また表 5.6. は、3PL-MS 法を用いて同じように2セットの事前事後テストによって推定された受験者能力値変化の平均値である。同じデータであるにもかかわらず、使用するモデル・等化法によって結果がかなり変わることが確認された。

表 5.5. 2PL-CR による能力変化

年度	θ の変化平均値	*SD*
01	0.411	0.822
02	0.286	0.881

表 5.6. 3PL-MS による能力変化

年度	θ の変化平均値	*SD*
01	0.269	0.767
02	0.135	0.808

これによると、より適切な等化の方法である 2PL-CR 法で推定した事前事後テストの θ の変化は、年度01で θ=0.142、年度02で θ=0.151　3PL-MS 法による推定値より大きい伸びを示した。また、予備テストとして、ほぼ同人数に対して年度00（年度01の 1 年前）に実施した事前事後予備テストでも、やはり 2PL-CR のほうが 3PL-MS より θ=0.183　大きい伸びを示した。年度01，02ともに、より適切な等化の方法のほうが、あまり適切でない方法より大きな伸びを示す傾向があることが確認できた。この結果は、等化後の θ の尺度の分散が 1 であることを考慮するとかなり大きいと考えられる。また、すべての等化法とモデルの組み合わせに於いて、θ の伸びに関する t 検定の結果は極めて低い p 値で有意であった。

5. 4.2. プログラム評価のための要素

4つの習熟度別レベルごとの事前事後における平均能力値 θ の変化

表 5.7.は、年度01、年度02、それぞれの年度ごとの 2PL-CR 法で推定した学生の θ の伸びの平均値を、習熟度レベル別（Aが上級、Uが中上級、Lが中初級、Bが初級）に表にしたものである。

表 5.7. 4つのレベルと全体の平均能力値 θ の事前事後での変化

	年度01			年度02		
	人数	平均θ伸び	*SD*	人数	平均θ伸び	*SD*
A level	80	0.279	1.049	105	0.530	0.867
U level	280	0.569	0.836	312	0.322	0.880
L level	177	0.184	0.710	191	0.061	0.865
B level	151	0.437	0.755	164	0.324	0.860
全体	**688**	**0.411**	**0.881**	**772**	**0.286**	**0.881**

年度01では、平均 θ の伸びが最も大きいレベルはU（中上級）、年度02では最大の伸びを示したのはA（上級）レベルである。また、平均 θ の伸びが最低なのは、両年度ともにL（中初級）レベルで、特に年度02では、他のレベルに比べ極端に θ の伸びが小さいことは特筆に値する。θ の伸びの平均値に関して検定の結果は有意であった。

プログラムニーズアンケート

事前事後テストで推定した受験者の能力値 θ の変化量を従属変数とし、プログラムニーズを測るためのアンケート 25 問（Appendix C 参照）の回答結果を独立変数とした強制投入法に基づく重回帰分析を行った。その結果、年度01の学生は質問 13、15、19、25 で有意確率 p が0.05以下になり、年度02の学生の結果では、質問6、7、8、13、17で両側5%水準で有意であった。また、両年度において両側 5%水準で有意であったのは、質問 13「先生が授業を楽しくしてくれるとよく勉強できる。」だけで、この質問に対する偏回帰係数は、年度01で 0.142、年度02では 0.129 で、25 の質問の中で両年度とも最大であった。これは、能力値 θ が事前事後で伸びた学生ほど、25 問の中では質問 13 に対して同意した傾向を示しているが、偏回帰係数の推定値の絶対値は目立って大きいものではないと言える。

次に、各質問項目からの説明力を評価するというよりは，これらの中で特に重要なものを選別していくという観点から，受験者の能力値θの変化量を従属変数として、同様の 25 個の独立変数に基づく変数増加法による重回帰分析を行った。その結果、表 5.8.にあるように、年度01では、質問 17「外国人の先生から英語を学ぶのが好きだ。」、質問 25「今まで TOEIC, TOEFL, 英検のどれかを受けたことがありますか？」、質問 13「先生が授業を楽しくしてくれるとよく勉強できる。」これら 3 つの質問において両側 5%水準で有意であった。また表 5. 9.で示したように、年度02では質問 17, 13 以外に、質問 15「先生がテストをしてくれたり、宿題をだしてくれるとよく勉強できる。」、質問8「ひとつのスキルだけの授業がよい。」、質問2「ひとりで(ペアやグループではなく)学習するほうがはかどる。」の計 5 つの質問において両側 5%水準で有意であった。

表 5.8. 年度 01 学生の能力値変化 θ とプログラムニーズに関するアンケート回答の回帰分析（変数増加法）

	偏回帰係数	標準誤差	標準偏回帰係数	*t*	*p*
切片	-.391	.193		-2.027	.043
質問17	.106	.046	.102	2.325	.020
質問25	.076	.027	.118	2.839	.005
質問13	.103	.052	.086	1.978	.048

Note. 従属変数: θ $R^2 = 0.040$ 調整済み $R^2 = 0.035$

表5.9. 年度02学生の能力値変化θとプログラムニーズに関するアンケート回答の回帰分析（変数増加法）

	偏回帰係数	標準誤差	標準偏回帰係数	*t*	*p*
切片	-.592	.223		-2.649	.008
質問17	.116	.040	.115	2.891	.004
質問13	.119	.045	.105	2.654	.008
質問15	.077	.039	.075	1.953	.051
質問8	-.085	.036	-.090	-2.337	.020
質問2	.074	.038	.076	1.981	.048

Note. 従属変数: θ $R^2 = 0.055$ 調整済み $R^2 = 0.047$

5.5. 考察

5.5.1. 等化の方法10通りの比較

項目パラメタを安定して推定するために必要な受験者数は、3PL では 1000 人と言われているとおり、事前テストは 5,000 人以上であっても、事後テストの受験者数が 1000 人に満たない本研究のようなケースでは、やはり 3PL より 2PL モデルが相応しいことが確認できた。共通項目の θ を利用し、RMSE を基準とする評価は、モデルの違いが判定結果に与えるインパクトは大きく、本研究のデータには、2PL が 3PL よりも適合しているという結果を示唆している。今後、繰り返し継続的に本研究と類似したデータで事前事後テストの等化を行う場合は、2PL モデルを使用することが適切である。

しかし、実際の教育現場では、IRT を利用してテスト結果を分析するとき、そのテストの受験者数、項目数、テストの目的などを考慮して、どの IRT モデルを選択するか十分な検討が必ずしも行われずに分析を始める可能性がある。受験者数の点では、事前テストは 1200 人だが、事後テストは 900 人が受験するようなケースでは直ちに判断し難い。また、受験者数問題に加えて、将来項目バンク構築に備え c パラメタの情報も得たい場合もある。このように、どのような方法で等化するか即座に決定できないときは、等化する方法を決める何らかの基準があることが望ましい。本研究では教育現場での θ の果たす重要性を考慮して、共通項目の θ を利用して RMSE による評価方法を判断基準に採用し、データによりふさわしい等化の方法を決定することを提案した。

また、このとき IRT モデルの違いによる RMSE へのインパクトは、等化法のインパクトより大きいことが確認でき、等化の方法を選択するときは、まずは IRT モデルの選択を優先するべきだと考えられる。さらに、等化法の選択に関しては、本研究で用いた実際のデータにおいて、5つの等化法の RMSE に大きな違いはなかった。従って、本研究の目的の一つである、高度な専門知識を必要としないテスト分析を行うには、複雑な計算を要する項目特性曲線に基づいた HB 法、SL 法、CR 法よりも、計算がシンプルな項目困難度の推定値の平均値や標準偏差を利用する MS 法や MM 法のほうが適切だと言える。

しかしながら、最適と思われる 2PL-CR 法と最も適さないと判定された 3PL-MS 法、この2通りの方法を用いて、年度01と年度02の事前事後テストのデータを分析したところ、同じデータであっても用いたモデルと等化法によって結果に相違が生じた。そして、最適な等化の方法によって導き出された事前事後テストにおける受験者の平均能力値の伸びのほうが、不適切だと判定された方法による能力値の伸びより高い傾向を示すことが確認できた。これは、最適と思わ

れる等化の方法を用いずに θ を推定することは、実際よりも θ の伸びを過少評価して報告する可能性があることを意味している。この理由は定かではないが、5.4.1 における RMSE の比較に際しては、共通 10 項目のみを用いて推定された θ の安定性・不変性を調べたのに対し、最適とされた 2PL-CR による推定に於いては、事前・事後テストのそれぞれに含まれる全ての項目（30 項目）を用いて θ の推定を行いその伸びを調べている。したがって、共通項目以外の部分に等化法の違いの影響が出たと解釈することも可能である。

今回の本研究の実データを基にした RMSE 比較結果では、大きな違いはないものの5種類の等化法のなかでは、CR 法が安定して最も誤差の少ない推定を行える可能性を示した。さらに、最近のシミュレーション研究（eg. Arai and Mayekawa, 2011； 光永・前川, 2012）でも、CR 法による等化の優位性を示す結果が得られている。計算の煩雑さはあるが、PLINK (2007) などのパッケージを用いると比較的簡単に等化係数を求めることはできるので、教育機関において、項目バンクを作成したり、繰り返し等化を行う場合はオプションの一つとして考慮する価値はあると考える。

本研究の等化法の比較における新規性

本章の 5. 2.1　IRT モデルと等化法の選択、等化法の比較の研究のところで述べたように、IRT の共通項目デザインによる等化法を比較する研究は、共通項目の違いに着目するものに関しては数多く存在する。しかし、（1）等化する2つのテストの人数に大きな違いがある場合の等化法を比較する研究である。また、（2）等化法を比較する基準をパラメタではなく θ で実施している。さらに、（3）仮定されたモデルのもとで乱数を発生させて得られたデータではなく、実データを使用している。これらの3つの点が、等化法比較における本研究の新規性である。

等化法の先端的研究はテスト開発者や研究者主導で行われていることが多いため、教育現場で授業の一環として実施されたテストデータを使うことはあまりない。ところが、実際の教育現場では事前テストは全員でも、事後は少数の受験者であれば実施可能となる場合も多く、この条件での等化法を比較する研究が必要となる。また、等化法を比較する基準にはパラメタを使う方法が一般的であるが、パラメタによる推定値で等化法を比較した場合、θ による結果と異なる可能性がないとは言えない。従って、実際にその教育プログラムの学生の能力値である θ を基準にして等化法を比較するほうが、実態に適合した結果となると考えられる。

さらに、本章で用いた最適な等化法の判定方法は、この事例研究だけに関わらず他の教育プログラムでも利用することができる。例えば、ある教育プログラムで事前事後テストを実施し、その事前事後テストを等化する必要が生じた場合、本章と同じように θ を基準にして θ_0 とそ

れぞれの等化法との差の RMSE を求める方法を使うことで、その教育プログラムにおける最適な等化法を見つけることができる可能性が高い。

5.5.2. プログラム評価の証拠

習熟度レベル別 θ の伸び

両年度通して θ の伸びが最も低かったレベルはL（中初級）であった。特に年度02では、他のレベルと比べ中初級レベルの伸びの落ち込みが激しく、このレベルの θ の伸びが非常に低いために、他のレベルの θ の伸びも影響を受けて年度01に比べ、年度02が全体的に低い θ の伸びを示す結果となっている。年度00の結果もやはり中初級レベルの平均 θ が最も低かった。事前事後テストによる調査結果により、中初級レベルだけ θ の伸びが3学年続けて最も低いことが確認できた。これはプログラムの現状を表すプログラム評価の証拠の一つとなり、原因を解明する価値がある。大規模な英語教育プログラムの中で、B（初級）レベルは、他のレベルよりややクラスサイズが小さく平均 25 人から 28 人になるなど特別な対応を受けている。しかし中初級レベルのクラスは、中上級、上級レベルと同じく、クラスサイズも平均 33 人で対応に差が無い。この他に何か要因があるのか、学生・教員へのインタビューも含めて詳しく調査する必要がある。

プログラムニーズアンケート

事前事後テストで推定した受験者の能力値 θ の変化量を従属変数とし、プログラムニーズアンケートの回答結果を独立変数とした強制投入法に基づく重回帰分析を行った。偏回帰係数の推定値の絶対値は、全体的にあまり顕著なものではなかったが、年度01、02ともに、質問 13「先生が授業を楽しくしてくれるとよく勉強できる。」については、能力値が伸びた学生ほど、この質問に同意している傾向を示した。この傾向は、先行研究　Kikuchi (2005)の結果とも一致する。また、変数増加法による独立変数の選択を行った結果、質問 17「外国人の先生から英語を学ぶのが好きだ。」も両年度を通じて能力値の伸びを予測する傾向を示した。

年度別に見ると、質問8「ひとつのスキルだけ（「リーディング」ならそれだけ）を学ぶ授業が良い。」に対して、年度02では負の偏回帰係数の推定値を示しており、θ の伸びが大きかった学生ほど、質問8に対してネガティブな反応を示している可能性がある。これは年度01と年度02の間でカリキュラムを変更し、以前はスキル別であった授業の形態を、年度02から 2 スキル合併型（例:リスニング and スピーキングコース）に変えていることと関係があるかもしれない。

今回のプログラムニーズアンケートは、回答のための時間が 10 分以内であったこともあり、十

分な情報が得られなかったが、そのプログラムを履修して能力値が伸びた学生の声を聴くための手段としてプログラム評価上意味のある証拠となる可能性はある。

5.5.3. まとめと今後の研究

本研究では、2 学年分の実際の事前事後リスニングテストのデータを使ったために、これをすぐに一般化して結論づけることは難しい。しかし、テストの等化を実行するとき、どの項目反応モデルを選択するかは分析結果に影響をもたらすと言える。この結果からも、テスト分析者は安易に分析を実行せず、データと分析の目的を良く確認してから、用いる項目反応モデルと等化法を選択するべきだと言うことを再認識した。しかし、先行研究にもあったように等化法の比較研究は、シミュレーション研究、事例研究、どちらか一方だけでは不十分で両方の研究を合わせて実施するべきである。最後に、本研究が労力を省きながらも少しでも有効性が高く、効率の良い「事前事後テストシステム」を、改善のためのプログラム評価に繋げるひとつの例となれば幸いである。

（注1）個人情報と関連性があるため具体的な学年度を避け、予備調査を行った学年を年度00、本研究データの初年度を年度01、次年度を年度02とした。

第6章　EEP–J モデルの他大学での利用の可能性

６．１．はじめに

日本の大学のどこかの部局でプログラム評価を開始するとしたら、英語教育プログラムから始めるところが多いであろう。2014 年度には、781 校ある日本の大学のほとんどで英語教育は実施されている。また、最近の大学を取り巻く時流では、トップレベルの高校生は、もはや日本の大学ではなく海外の大学を目指す動きもあり、グローバルな視野を持った人材を日本の大学で育成できるようにすることが急務であると言われている。また、5, 6 年前から CEFR を到達目標として授業に導入する大学も増加し、その教育プログラムの効果を検証する動きも出てくる頃だと思われる。このような状況のなかで、各大学の英語教育プログラムでは改革を求められ、プログラム評価を計画しているところも多いのではないかと推察する。

本章では、日本の大学英語教育プログラムの評価について、現在の実態や傾向を少しでも知るために、英語教育プログラムの全体像と内容を知る管理職的な立場の教員に、アンケート調査を実施することにした。現時点ではあまり実施されていないと思われる英語教育プログラム評価であるが、実施されていない場合はその原因を解明し、将来、プログラム評価をより広く、効率よく実施する方法を提案できるように、日本の大学における英語教員たちの声を少しでも多く聞きたいと考えた。そして、第 1 章 1. 9.3 で示した、「大学英語教育における新しいプログラム評価モデル：EEP-J モデル」の他大学での利用の可能性を明らかにしたい。

６．２．　日本の英語教育プログラム評価の論点

6.2.1　EEP-J モデルの他大学での利用

日本の英語教育プログラムにおいて、他大学で本 EEP-J モデルを利用する可能性があるとしたら、論点となるであろうプログラム評価の目的、評価の実施者、評価の重点、量的研究アプローチ、外部テストと教員作成テストについては本論 1 章の1．8にまとめた。本章では、主に、これらの論点を中心に「英語教育プログラム評価アンケート」で教員に尋ねる 10 項目の質問を作成する。

これらの論点のなかでも、本論で提案する EEP-J モデルの他大学での利用について、特に関係が深い点がある。例えば、英語教育プログラムにおいて、EEP-J モデルの評価者は、英語教員となる可能性が極めて高いが、果たしてどのくらいの割合で参加型（教員が実施）、あるい

は介入型（外部に委託する）評価を実施しているのか。また、教員が実施するプログラム評価は、自己評価的要素が強く、評価の焦点をプログラムアウトカムに当てる傾向があるため、できるだけ多様なデータを用いて、基準の客観性を重視しなければならない。つまり、客観的エビデンスの厳密さが求められ、より多くのデータを収集し、厳密に分析するほど教員の負担は増えるが、これは問題にはならないのか。EEP-J モデルの長所である「協調して自己修正できること」は大いに活用したいが、その可能性はあるのか。そして、本論で提案するモデルは、山中（2007）の EPEU モデルのうち、「理論の評価」を省略しているが、英語教育プログラムにおいて、教員たちは「理論の評価」に対してどのように考えているのか、などである。これらの点について質問することが他大学での利用の可能性を探るための教員へのアンケートの目的である。

6.2.2　本章で解明しようとすること

実際の日本の英語教育プログラムにおいて以下の点を中心に、プログラム評価の実施状態の調査を行う。まず、（1）どのくらいの大学英語教育プログラムにおいてプログラム評価が実施されていて、また実施されない場合はその理由、また、実施されている場合は、プログラム評価の目的、（2）日本での言語教育プログラム評価のほとんどが教員主導型の評価だと推測されるが、外部の専門家に委託して評価しているところとの比率、（3）プログラム評価結果を報告する相手は主に誰または何なのか、（4）プログラム評価を行う重点はどのようなものに置かれ、（5）評価によってもたらされたものは何か、（6）評価の際に客観的証拠をもたらす（統計を含む）テスト科学はどのくらい重要視されているのか、（7）日本の言語プログラム評価における問題点は何か、（8）事前事後テストを実施するとしたら市販テストと教員作成テストのどちらが適しているか、これらの疑問に対する手がかりを得るための調査を実施した。

6．3．　研究方法

アンケート対象者

日本の高等教育における言語教育プログラムに携わる教員で、しかもそのプログラム全体を統括するような立場の方、48 名にお願いして、「言語教育プログラム評価アンケート」にお答えいただくように手配した。そのうち、約 3 分の2は先生方に郵送し、残りは E-メイルにアンケートを添付して送付した。依頼してから 3 週間以内に 40 人からアンケート結果が届いた。

実施手順

プログラム評価をしたことがあるかないかで分け、2 種類のアンケートを作成した。 プログラム

評価をしたことのある教員は、黒字で白のアンケート用紙（Appendix D）を、また、プログラム評価をしていない教員には、青字でブルーの質問用紙 (Appendix E) に答えてもらうようにお願いした。 2 種類のアンケートは両方とも 10 問ずつあり、質問は似た内容で、一部は同一問題である。まず、評価をしたことがある教員用のアンケートは、プログラム評価をしたことが前提となっているため、例えば、質問1は「プログラム評価をされた目的は何ですか？となっているのに対し、評価したことがない教員用のアンケートの質問1は、「プログラム評価をしない理由は何ですか？」となっている。このように、プログラム評価をしたことと、していないことが、それぞれの質問紙の前提となっている点が 2 種類の質問紙の大きな違いである。 表 6.1 は、1. プログラム評価を実施した方への質問と、2. プログラム評価を実施していない方への質問を、実際の質問文を簡略化して記したものである。

表 6.1　　プログラム評価実施に関するアンケート

問1	1	プログラム評価を実施した目的は何か
	2	プログラム評価を実施しない理由は何か
問 2	1＆2	プログラムのゴールや目標は設定しているか
問 3	1	プログラム評価の実施をしたのは誰か
	2	プログラム評価をするとしたら誰が実行するか
問 4	1	プログラム評価の結果は誰又は何に対して報告したか
	2	プログラム評価をするとしたら誰又は何に対して報告するか
問 5	1	プログラム評価の重点は何か
	2	プログラム評価するとしたら、重点は何か
問 6	1	プログラム評価によって何がもたらされたか
	2	プログラム評価をするとしたら何がもたらされると思うか
問 7	1	統計やテスト科学はプログラム評価にどれくらい役立ったか
	2	統計やテスト科学はプログラム評価にどれくらい役立つと思うか
問 8	1＆2	プログラム評価のためのデータで重要なものは何か
問 9	1	プログラム評価を実施して、どのような問題点があったか
	2	プログラム評価を実施するとすればどのような問題点が予測されるか
問 10	1＆2	事前事後テストは、市販テストか教員作成テストか

6．4． 結果

集計の結果、集まった 40 人のアンケートのうち、22 人がプログラム評価を実施したことがあり）、18 人がプログラム評価を実施したことがないことが分かった。10 問の質問はそれぞれ全く同じか、非常に似た主旨の質問をしているが、プログラム評価を（1）実施あり / （2）実施なしに分けて 10 問の質問に対する回答を以下にまとめる。

質問1．

（1） プログラム評価を実施した目的は何か

カリキュラムの機能が果たされているか、そして到達目標に対する学生の英語能力の伸びを測るため、現状把握して改革が必要なのかを調べる。

（2）プログラム評価をしない理由は何か

学生の英語力向上に対する必要性が低く、あまり評価する重要性を感じない。評価や成果を測ること、現実をはっきりさせることへの反発、恐れ、そして不安がある。プログラム評価を担当する専任教員に負担がかかりすぎ、また専門知識もなくてできない。トップダウン的な発想を嫌う文化がある。

質問2．

（1）＆（2） プログラムのゴールや目標を設定しているか

プログラムのゴールや目標は、プログラム評価実施校（1）では 22 件中 18 件が（全体のうち 82％）、プログラム評価実施なし校（2）では 18 件中 12 件（全体の 67％）が、合計 40 件中 30 件（全体のうち 75％）が設定している。

設定している目標の主なものは、学部や大学全体と連動した目標、TOEFL、CASEC のスコアや CEFR のレベル、語彙レベル（語彙数、例　8000 語）などがあった。

質問3．

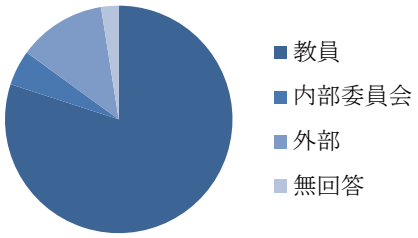
（1）プログラム評価を実施したのは誰か

外国語教員（＝同僚の教員）という回答が圧倒的に多かった（18 人）。学内の委員会が 1 人、3 人は外部の専門機関に委託したと回答している。

(2)プログラム評価を実施するとしたら誰が実行するか

外国語教員が 14 人、1 人が内部委員会、2 人が外部専門機関に委託するであろうと回答している。

(1)と(2)を合わせると、32 人が外国語教員、2 人が大学内部の委員会、5 人が外部の専門機関に委託ということになり、80%が外国語教員の参加型評価として実施した/する可能性が高いという結果である。



質問4.

(1)プログラム評価の結果は誰又は何に対して報告したか

理事会、学長、副学長、大学執行部、学部長、教務委員長、などの大学内の上層部や上司に報告すると回答したのは 6 人(27%)。他学部教員、教授会、同僚など大学内の同列関係の相手に報告すると回答したのは 14 人(64%)、また、学生、一般公開と回答したのは 2 人(9%)で、さらに、外部の機関と答えたのは 3 人(14%)であった。

(2)プログラム評価をするとしたら誰又は何に対して報告すると思うか

理事会、学長、副学長、大学執行部、学部長、教務委員長、などの大学内の上層部や上司に報告すると回答したのは 9 人。他学部教員、教授会、同僚など大学内の同列関係の相手に報告すると回答したのは 6 人、また、学生、一般公開と回答したのは 2 人で、さらに、外部の機関と答えたのは 2 人であった。

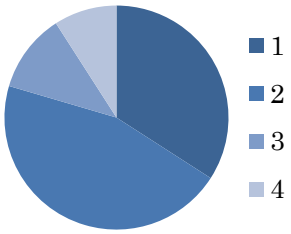
(1)と(2)を総合すると、

1　大学内の上層部や上司に報告する　⇒　15 人(34%)

2　同列関係の相手に報告する　⇒　20 人(45%)

3　学生、一般公開と回答した　⇒　4 人(9%)

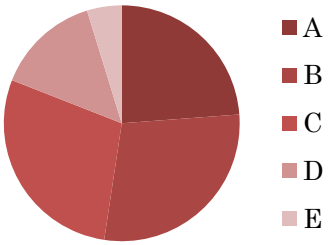
4　外部の機関と回答した　⇒　5 人(11%)であった。



質問5.

(1)プログラム評価の最重点は何であったか

（A）プログラム理念　　5 人
（B）カリキュラムの内容　　6 人
（C）プログラムによる効果　6 人
（D）学生の反応　　3 人
（E）教員の反応　　1 人



Ｂが最重点の理由

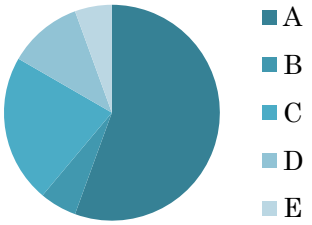
プログラム評価の結果を多くの要素を含むカリキュラム運営に生かすべきである。また、習熟度別クラス編成を実施した効果を測る必要があり、学生のためにより良いカリキュラムを提供する必要があった。

Ｃが最重点の理由

学生の言語能力がどう変化しているか調査することが先決で、学生の学習態度や語彙レベルが向上しなければ意味がない。大学の上層部が最も関心があるのが「効果」であり、学生の言語能力について効果を検証するように指示された。授業内容・方法を改善するためにその効果を調査するべき。

(2)プログラム評価するとしたら、最重点は何か

（A）プログラム理念　　10 人（56%）
（B）カリキュラムの内容　　1 人　（6%）
（C）プログラムによる効果　4 人　（22%）
（D）学生の反応　　2 人　（11%）
（E）教員の反応　　1 人　（6%）



Ａが最重点の理由

理念が最も重要な基本で重視するべきである。理念がないとカリキュラムがバラバラになる。何を学ぶかゴール設定し、それが達成できているか調査して示す必要がある。

（１）と（２）を総合すると、

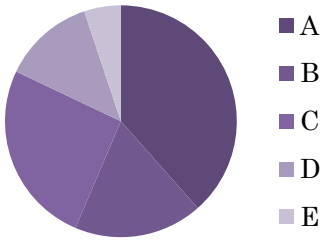
（A）プログラム理念　15 人（38%）

（B）カリキュラムの内容　　7 人（18%）

（C）プログラムによる効果　10 人（26%）

（D）学生の反応　　5 人（13%）

（E）教員の反応　　2 人（5%）



質問 6.

(1) プログラム評価によって何がもたらされたか

教員

教員の意識改革、授業改革、協力体制、自身の授業に対する振り返りのチャンスとなった。また、授業のゴール設定や方法を各教員が認識できるようになった。

教員間の情報共有、課題の発見、改善の話し合いの機会が得られた。

学部教員の言語教育に対する意識の変化があった。

教員に対して非常に荷重がかかっていることが分かった。

学生

学生の意識の変化、自己評価することで「振り返り」の機会を得た。

学生の要望を教員が知ることができた。学生の意外な反応が分かった。

学生はそのテストスコアから導かれたフィードバックで自己修正がしやすくなった。

プログラム

プログラム・カリキュラムの変更・改善すべき点が明確になった。問題点に対して具体的な対策を立てやすくなった。教材の難易度をより適切なものに変更するなど、具体的な改善が可能となった。段階的な授業内容の変更など、改善策が具体化した。

否定的

専任教員は現状把握できるようになるが、非常勤教員に対しては情報共有されず。教員が学生のテストスコアを気にしすぎて、テスト偏重授業になる恐れがある。

学生の本音の反応は分かりにくい。

調査項目のなかにもっと必要な情報が組み込まれていず、不十分である。

(2) プログラム評価をするとしたら何がもたらされると思うか

教員

教員に対する評価が数値化してもたらされる。教員・学生間の連帯感が深まる。教員の責任感・意識を向上させる。

学生

学生の英語力を向上できる。仕事で使える英語力をもった学生の輩出

プログラム

プログラムを修正し、教育方法を検討する為のFDに結びつけ、より良いプログラムへの指針が得られる。プログラムの強みをさらに強化し、弱点を変更改善する。カリキュラム・授業内容の見直し・改善。言語教育改革のスピードアップ。理念に沿ったカリキュラム内容であるか検討できる。学生の反応を分析できるため、効果的なプログラムにできる。一貫性をもった言語教育プログラムにできる。PDCAサイクルを形成できる。言語教育の効率を上げることができる。

否定的

望まないが授業の統一・均一化が進む。テストスコアだけを重視し、深い思考の育成を伴わない教育になる可能性がある。共通教科書の授業になり、教員の個性が失われる。

質問 7.

(1)統計やテスト科学はプログラム評価にどれくらい役立ったか　　50.5%　役立った

(2)統計やテスト科学はプログラム評価にどれくらい役立つと思うか　62.2%　役立った

(1)と(2)

平均値　57.3%　役立った

質問 8.

(1)と(2)プログラム評価のためのデータで重要なものは何か

表 6.2 は、(1)と(2)の合計を円グラフにしたものである。

プログラム評価のためのデータとして重要なものは、2. 学生の事前事後テストのスコア、1. 学生の市販テストのスコア、3. 学生の自己評価の順に重要という回答結果である。

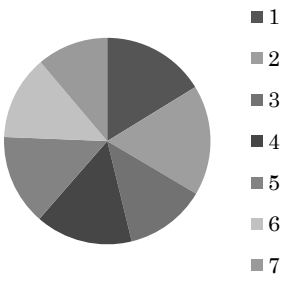


表 6.2　データとして重要なもの

	(1)	(2)	total
1 学生の市販テストのスコア　例、TOEIC、GTEC、CASEC など	16	16	32
2 学生の事前事後（入口出口）テストのスコア	19	15	34
3 学生の授業内容に対する到達度を測る中間テストや期末テストのスコア	13	12	25
4 学生の自己評価　例、Can-Do 自己チェックリスト	16	14	30
5 学生の授業に対する評価　例、学生による授業アンケート	16	12	28
6 教員の学生に対する評価　例、最終総合成績、授業中の学生観察	14	12	26
7 教員たちのカリキュラムに対する評価　例、アンケート	16	6	22

質問 9.

（1）プログラム評価を実施して、どのような問題点があったか

教員

関わる教員への負担が大きく、また、実施した教員の専門知識の欠如を実感した。

評価結果に対応するべき非常勤と常勤教員の交流が無い。小規模では問題なし。規模が大きくなるほど時間的や協力関係に問題が発生する。頻繁にアンケートに答える教員にとって煩わしさが問題となる。学生による評価に教員がとりわけ影響を受けやすくなる傾向があった

教員から「自分たちが評価されている」として反発がでた

データが示す現実を直視せず、勘や経験を優先にする傾向があった

学生

学生評価は（過少評価傾向があり）信頼できない

プログラム

実際のカリキュラム改革に容易に結びつかない。報告書の作成に労力がかかる

習熟度別クラス編成に問題が発覚した。限定されたスケジュールの調整をしながらプログラム評価をすることが困難であった

大学上層部

TOEFL のスコアで一喜一憂し、数値的データにしか関心を示されなかった。また、全体の効果の有無にのみ関心を示し、評価結果の詳細は無視した。報告しても説明をじっくり聞いてもらえなかった

大学が主導権をとってデータを収集し、教員には参画する機会を与えられなかった
大学上層部は結果を受けて、改善へ向けての対策は無かった

(2)プログラム評価を実施するとすればどのような問題点が予測されるか

教員

評価結果を意識し過ぎ、テストの為の授業になってしまう。そして、評価に参加した関係者の時間的負担が大きい。同僚の授業を評価したり、問題を指摘したりしづらい。誰がどのような目的でデータを収集するのか決定が難しい。専任教員の協力体制ができるかどうか不明。また、非常勤教員の協力が得られるかどうか、語学教員全部が一致団結し、意思統一を測れるか不明。毎年実施していける体力があるかどうか不明。
統計的数値によって本当に評価したことになるのか疑問。良いテストを作成して正確なデータを収集することは難しい。評価に要する費用の問題。
プログラムの理念やゴールと一致しない授業内容を実施している教員も居る。また、専任教員同士の意見のすり合わせに時間がかかりそうである。同僚教員・学部教員や学生の消極的反応が予想される。教員の勤務評価と結び付けられ大学から監視されている気になるなどの不満を言う教員が居る。教員の授業に対する自由度(教材選択・授業内容)が奪われそうだ。

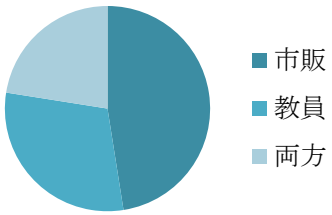
大学上層部

プログラム評価に対する理解があるか疑問。

質問 10.

(1)と(2)事前事後テストは、市販テストか教員作成テストか

市販テスト　19 人　(48%)
教員テスト　12 人　(30%)
両方　　9 人　(23%)



市販テストが良い理由

他大学と比較ができ、信頼性、客観性、妥当性が高い。大規模なプログラムに相応しく、外部基準に合う。また、経年比較ができ、精度が高い。また、社会に対するアカウンタビリティーが確立しやすく、教員の問題作成の手間が省ける。

否定的

費用がかさむ

教員テストが良い理由

授業内容の直接的評価につながる。テストはカリキュラムや指導内容に応じて、学生のレベルにあわせて作成されるべきである。また、費用がかからず、学内のニーズに合わせ柔軟な評価ができる。教育内容と目的に沿ったテストができる。

テスト内容のコントロールができ、プログラム内容を充分に反映させられる。プログラムの内容を熟知する教員がテストを作成するべきで、そのプログラムでどのような教育をうけ、学習したかを測る必要がある。

否定的

教員の負担が非常に大きい

両方が良い理由

教員作成は無料でできるが市販テストは妥当性に問題あり。プログラム成果の確認には教員テスト、総合的英語能力の確認には市販テストが良い。

教員作成テストはプログラムの達成度を市販テストは客観的な位置づけを知るために実施すべきだ。市販テストで測定できない内容は教員作成テストを実施。

市販テストは客観性が高く、教員テストは到達度がより明確だが、適切なテストが開発できるなら教員テストだが、人的・時間的・予算等資源がかかりすぎる。

６．５．　考察

6. 5.1 プログラム評価の実施と形態

今回のアンケートは限られた人数（40 人）の先生方にお答えいただいたもので、これが一般論だと考えることは難しい。しかし、ある程度の傾向を知るための手掛かりの 1 つにすることは可能であるかもしれない。まず、日本の大学英語教育プログラム 40 件のプログラムのうち 22 件、約半数のプログラムが“何らかの”プログラム評価を実施していた。22 件がプログラム評価を実施した目的は、プログラムの実施状況、効果を調査して改善するためであるとほとんどの回答が示していた。しかし、、「TOEIC を 1 学年全員に実施した」 また、「学生に授業評価を実施した」などの理由で“プログラム評価を実施した”と回答された方が多かったのは、アンケート作成時

に、“何をもってプログラム評価を実施したことになる”という定義を提示しなかったためである。また、18 件のプログラム評価をしない理由には幅があり、(1)評価を担当する教員の負担が大きいこと、専門知識の欠如、(2)評価を実施し成果を測定するというような、トップダウン的な発想に対する教員の反発、(3)プログラムの運営や効果への無関心、(4)プログラム評価を実施してプログラムの改善をする必要を感じていないなどがあった。

40 校中 32 校において、プログラム評価を実施するのは英語教員であった。これは、80%の大学でプログラム評価は教員主導型で行われている可能性が高く、教員たちが自らのプログラムを自分たちで評価しているということになる。このために、自己評価的な要素が強くなることがある。傾向としては、プログラムの効果に評価の焦点を当てすぎること、英語教員の考えや意見が反映され、客観性を欠き主観的な要素が強くなることが指摘されている(三好・田中, 2001)。この傾向を避けるためには多様なデータを用い、評価の偏りがないように注意し、基準の客観性を重視する必要がある。しかしながら、教員が主導する評価の自己修正の容易さ、評価者がプログラムに直接関わり、参加者と協調して実用的な評価ができる点など、肯定的な部分も多い。プログラム評価を報告する相手は、理事会、学長、副学長、大学執行部、学部長、教務委員長など大学内の上層部や上司の関係にある相手(機関)、また、他学部教員・教授会・同僚教員ら同列関係の相手、これらを合わせると 80%を占める。教員が主導する評価の特徴は、自らのプログラムを自ら評価して同僚に報告し、自己改善を促進させる機能を働かせやすい点である(Fetterman, 2001)。この特徴を生かし、自分たちで評価した結果を尊重して積極的に改善のための行動に繋げることができるという利点を惜しみなく発揮するべきであろう。アンケート結果は、5 件のプログラムだけは外部の専門機関に委託してプログラム評価を実施した／したいとしている。報告書を精読し非常に役に立つフィードバックをもらったとしても、外部の第 3 者専門機関が評価した結果をすべて理解し、迅速にそれらすべてを改善へ導くためのアクションを起こすことは、容易ではないと思われる。

6. 5.2 プログラム評価内容

プログラム評価を実施した大学では、その評価の最重点がカリキュラムの内容やプログラムによる効果であった。Weiss (1998) は、評価の重点は、「プログラムの働きや機能」と「プログラムの結果と効果」の2つに大きく分けられると述べていて、本論の結果と相反するものではなかった。また、今回のアンケート回答者の 8 割が、教員がプログラム評価を実施することを支持していて、彼らはプログラムの実施と効果に関心が集まる傾向がある(三好・田中、2001)。

また、プログラム評価を実施していない大学で、プログラム理念に重点に置くところが多かったが、その理由は、理念を設定してその到達目標に向かって学習し、目的が達成されているか効果を測定するなかで、根源であるプログラム理念が最重要であるからというものである。一方、プログラム評価を実施した大学では、理念に重点を置いているところはそれほど多くない。これは、評価者である教員が自ら設定した理念を評価することはあまりなかったということかもしれない。実際、「理念の評価」において評価対象のプログラムの大きな失敗を発見した場合、次の段階（実施の評価）に行くことが難しくなるからであろう。

次に、プログラム評価によってもたらされたものは何か？という質問については、プログラム評価を実施した大学のほうが当然ながらより具体的な回答をしている。教員に対しては、意識改革・現状把握・協力体制がもたらされ、カリキュラムの改善に有用だったという声が多い。また、学生にとっても自分たちの声をプログラムに反映させる機会を与えられ、自己の振り返りや修正ができ、意識の変化が起きたようだという回答もあった。しかしながら、否定的な回答は約1割あり、プログラム評価の実施方法についての問題点や学生の自己評価の信憑性、教員のテストスコア偏重主義を心配する声もあった。プログラム評価を実施していない教員に対して、もし実施するとしたら、何がもたらされるかという質問に対して、否定的な回答のなかに、テスト偏重主義になり、多肢選択型テストが尊重されるあまり、深く熟考する必要がない教育にシフトするのではないかとういう危惧があった。また、プログラム評価によって、教員の自由と個性が奪われ、統一共通教材や統一カリキュラムに向かうきっかけになるのではないかと心配する声もあった。

6. 5.3 プログラム評価のためのデータ

質問8「評価のために重要なデータは何か？」について、プログラム評価を実施した教員と実施していない教員のあいだに生じた回答の違いは、評価を実施した教員のほうが、「教員のカリキュラムに対する評価」が重要なデータとしていたところである。評価を実施するのがほとんどの場合英語教員であるので、プログラム評価を実施した教員にとっては、同僚や非常勤教員がカリキュラムに対して抱いている思いを知ることは重要であったのではないかと思われる。そして、評価結果を受けてカリキュラムを改善することになった場合に、最初に協力をお願いするのも同僚や非常勤教員たちであったと推測する。

プログラム評価の実施の有無に関係なく重要だと回答されたデータについて、2. 学生の事前事後テストのスコア、1. 学生の市販テストのスコア、4. 学生の自己評価の順になっている。重要なデータの1位と2位がどちらもテストスコア（1位：事前事後テスト、2位：市販テスト）で、

回答者たちにとってプログラム評価の重要なデータが英語の客観的テストだと考えられている可能性が高いことがわかる。そしてその次が学生の自己評価である。

また、質問7「統計やテスト科学はプログラム評価にどれくらい役立ったか/役立つと思うか？」という問いに対し、パーセンテージで表してもらった回答の平均は57％であった。

6. 5.4 プログラム評価の問題点

プログラム評価を実施することで発生した問題点には、実施にかかる教員の負担、専門知識の欠如、参加者全員の時間的負担、財務的負担、そしてまた同僚や非常勤教員から「自分たちの授業が評価され、監督されている。」と受け取られ反発をかったことなどである。評価の実施に関する問題点は、学生自己評価の信憑性の無さが挙げられ、さらに大学側とのズレ・摩擦もあった。例えば学長がTOEICスコアなど数値的データにだけしか関心を示さず、そのスコアの上下に基づいて一喜一憂したことが回答されている。また、別の大学では学長が評価結果報告書のなかで、プログラム効果の有無だけ確認し、あとのことには目もくれなかったというのもあった。その上、評価の結果が生かされず、何の改善策も打たれていない場合が多いのは非常に残念なことである。

次に、プログラム評価を実施していない教員に対して、実施したとしたらどのような問題点が予測されるか？という問いに関して、まず教員の負担増を心配するもの、教員間での協力体制が確立できるか、また毎年プログラム評価を実施することが経済・人的資源的に可能かどうか心配する声や、プログラム評価の結果を意識しすぎてテストのための授業になることを憂う声なども聞かれた。

これらの中でも深刻なのは、プログラム評価結果が示す現実を直視せず、自分の勘や経験を優先する傾向である。プログラム評価をすることによって、個々の教員の授業に対する（教材選択・授業内容）などの自由度が奪われそうだという懸念もあった。

6. 5.5 市販テストか教員作成テストか？

40人中19人が市販テスト、12人が教員作成テスト、9人が両方のテストを事前事後テストとして支持した。市販テストが良い理由は、信頼性・客観性が高く、他大学と比較、経年比較も可能で、外部基準として使用できることが挙げられた。そして、社会に対するアカウンタビリティーが確立しやすく、教員の問題作成などの負担が省略できるという声もあった。これに対して教員テストが良い理由は、プログラム評価のための事前事後テストとして考えるなら、履修した授

業の内容がどれくらい習得できたか測るには、授業内容に合わせて教員が作成したテストを使用することで、テスト妥当性が確立できることが挙げられる。また、学生のニーズ、レベル、教育内容にあったテストが作成でき、費用がかからないことも理由になっている。さらに、市販・教員作成の両方が良い理由は、市販テストでは測れない内容は教員が作成したテストで測れる。そして、適切なテストが開発できるなら教員作成テストだが、それにかかる人的・時間的資源とのバランスを考えて決める必要があるという意見もあった。

６．６．　まとめ

今回の調査に参加していただいた先生方の英語教育プログラムにおいて、80％が教員によってプログラム評価が実施された/される予定だという回答を得た。したがって、これらの評価は教員主導型評価（三好・田中、2001）としての特徴を持つと考えられる。つまり、自己評価的な評価の否定的・肯定的な特徴を有することになる。否定的な面では、自らのプログラムを自分たちで評価するということで主観的になり、評価にバイアスがかかる可能性がある。そして、「プログラムの効果」に評価の重点を置く場合は、特に客観性を保つように注意しなければならない。しかし、教員主導型で評価を行うときの肯定的な特徴としては、自己修正の容易さが挙げられ、同僚と協力関係を確立し、結果を踏まえて改善に結びつきやすいという特徴を生かせば、プログラム評価の実施そのものが自己修正に発展しやすくなる。同質問に対して、13％が外部の専門機関に評価を委託した/する予定であるが、こちらの方は、客観的な見地に立ったプロフェッショナルな評価報告書を受け取れることはもちろん、教員自身では思いつかなかった事実や新情報を多く得られるメリットがあると推測できる。しかし、一方で多くの問題を指摘された教員たちが、改善策を容易にすぐ実行に移すことができない可能性が高いと思われる。

プログラム評価の重点は、一般的にプログラムの「働きや機能」と「結果と効果」の2つに分けられる（Weiss, 1998）。山中の EPEU モデル（2007）では、これに理念の評価が加わり、理論の評価⇒実施の評価⇒結果の評価という流れになる。しかし、今回のアンケート調査では、プログラム評価を実行した教員のなかでは「理念の評価」は 24％が重要と指摘したが、とりわけ重要性が高いものではなかった。プログラム理念がいかに大切かは周知されていて、今回の調査でもプログラム評価を実施していない教員の回答では、プログラム理念に最重点を置く教員が最も多かった。しかし、実際にプログラム評価を実施することになり、既に設定されたプログラム理念の評価から始めると、もし、理念の評価結果が適切でないとなった場合、次の段階の実施の評価に進めなくなる。また、理念の評価はかろうじて適切だと判断され、次段階に進めたとし

ても、評価可能性アセスメント（安田・渡辺、2011）をしなければならず、予定外の負担が増える可能性もある。

次に、プログラム評価のために必要とされるデータについての回答は、1 位が学生の事前事後テストのスコア、2 位：市販テストのスコア、3位：学生の自己評価の順となった。必要なデータの 1 位も 2 位もテストスコアという点と、プログラム評価のなかで統計やテスト科学の有用性が平均で 57％であった点も併せ、プログラム評価における客観的エビデンスの割合は決して低くないと言える。そして、3位の学生の自己評価は、履修者の声を聞く上で非常に重要な、異なる角度からのデータの1つであり、また実施することで学生の自己修正の機会にもなる。

今回の調査を通じて、プログラム評価を実施した/していないにかかわらず、大多数の教員がプログラム評価の実施に肯定的であった。プログラム評価によって、教員の意識改革・現状把握・協力体制がもたらされ、カリキュラムの改善に役立ったという回答も多々あった。また、学生の自己評価や授業評価によって教員だけでなく、学生も自分たちの声をプログラムに反映させる機会が与えられ、振り返りや自己修正の機会が与えられることは非常に重要である。

プログラム評価の問題点としては、評価結果を報告する相手である大学上層部との摩擦や、評価する側/される側のズレが報告されている。特に、学長をはじめとして大学上層部はどんな状況であっても、数値的なデータや効果の有無のみで全てを判断せず、収集されたデータすべてとその分析結果、教員・学生の声を総合的に理解することを疎かにしてはならない。また、評価者側も、プログラム評価結果を万人が理解できるように報告しなければならない。

最後に、今回のアンケートからは、客観的エビデンスを収集して総合分析する体系的な評価方法が受け入れられにくい日本的風土とも思える傾向が垣間見られた。それは、「はっきりと数値化して評価し、物事に優劣をつけて公開することへの反感」、「評価は教員に順位づけするためなのか？」また、「テスト偏重主義を強め多肢選択テストには長けているが、深い思考ができない学生を増やすのか？」などの反発を受けたという報告があったことである。このような反感を持つ人々とも上手くコミュニケーションをとり、“あいまいさ”を許容する柔軟性を持つことが評価者の資質の 1 つでもある（安田・渡辺、2011）。このように“反感”を持つ人たちも含め、異なった角度からの様々な意見やデータを収集する努力を惜しまないことも評価には要求される。そして、何よりも肝心なことは、評価結果を出すことをゴールにせず、結果を見て「改善する」ことを最終ゴールにすることである。

第７章　総合考察

7.1. まとめ

本論では、大学英語教育プログラムに焦点を合わせたプログラム評価のモデル EEP-J を提案している。このモデルの中核となるのは、Can-Do 自己チェックリストによる学生の自己評価と事前事後テストによる評価という2つの異なる角度からの客観的なエビデンスである。IRT を利用することでこれらのエビデンスの客観性のみならず有効性・妥当性も高め、さらに学生の自己修正能力を高めつつプログラムの改善をめざしている。次に述べるのが本論の3つの結論である。

7．1.1．　結論1：Can-Do 自己チェックリスト（SCL）による自己評価

信頼性の高い外部基準を基にしてカスタマイズする

最近の英語教育プログラムで関心が集まっている Common European Framework (CEFR)をはじめとしたCan-Do statement (CDS)とは、実際の状況のなかでコミュニケーションとして何ができるかを自然語で記述したリストである。SCL は CDS を自己評価の形にしたもので、学習者たちは「できる」「まあできる」「あまりできない」「できない」等のように、記述された言語能力について自己評価するものである。SCL の客観性を高めるために、まず世界的に普及している外部基準である CEFR の Can-Do statement (CDS)を日本の大学英語教育プログラムに取り入れる為の研究から始めた。CEFR は 2001 年に一冊の本として Council of Europe　が出版して以来、世界中の多くの研究者が関連の研究を進め今も進化を続けている。しかし、CEFR はもともとヨーロッパの言語学習者のために開発されたものなので、日本人学習者に適応するように変更を加える必要があった。日常的に英語を使わず外国語として英語を学ぶ環境や文化的な違いを配慮をして日本人学習者のために書かれた CEFRjapan などの先行研究も多い。それでも日本人学習者への適応だけでは十分ではなく、さらに、その英語教育プログラムの履修者に合わせた CDS を作成することによって、その CDS の信憑性は高まり、また、この CDS を授業に取り入れ、学生たちが一つ一つの項目を密接に意識しながら学習することで、その自己評価としての客観性が高まると考えられる。

IRT による項目困難度の利用

第 3 章では、T 大学のリスニングコースを担当する教員 10 名が吟味した T 大学のリスニング

CDS を、より学生の実態を反映したものに改訂するために約 330 人の学生に対して SCL として実施した。この結果を IRT によって分析し項目困難度を推定した、教員が想定した難易度と比較して大きな違いがある場合は原因を考察したあと、変更や改訂を加えた。このような IRT を利用して SCL の難易度を分析する先行研究は、既に広く使われている CEFR や英検 CDS の妥当性を検証するもので、IRT 分析による項目困難度によって修正・変更して教育プログラムで利用したものではない。本研究の新規性の一つは、実際の教育プログラムで授業を受ける学生たちの実態に合わせるために IRT による項目困難度を利用して教育プログラムで使用するところである。

自分のレベルに合った Can-Do 自己チェックリスト

第 4 章では、SCL の question order effects（質問順序効果）に注目し、質問項目の順序が結果に与える影響を明らかにするための調査を実施した。その結果、質問順序効果を小さくするには、回答者に適した難易度の差のない項目を質問することで回避される可能性が高いということを示唆している。これは、3. 4.2 の「14 項目の 3 つのレベル別 SCL」で述べたことと重なる。ここでは IRT を用いて全てのレベルの SCL を同一尺度化することで、回答者のレベルの項目しかない SCL であっても、他のレベルの SCL との差が分かるようになる。従って、学習者は難易度別になった 14 項目ずつの SCL に回答している。これは、多くの項目の SCL に答えることに比べ、回答者から得られる情報量は減るという弱点もあるが、本論では、学生のレベルに応じた短い SCL を教育プログラムに導入することに対する以下のような利点を挙げている（4.8）。（1）学生の負担減、（2）回答者のレベルに合わない SCL に回答する必要がない、（3）授業の到達目標として記憶・認識しやすい、（4）授業中に複数回実施しても支障が少ない、（5）質問順序効果の影響が少なくなる。

結論1：

CEFR のような世界的外部基準にもとづく SCL を英語教育プログラムに導入し、授業の到達目標、評価基準、習熟度レベルの基準として活用する。また、IRT で分析した項目困難度を用いてより学生の実態を反映した CDS に改訂する。自分のレベルだけの SCL を使用する。このようなプロセスを経て SCL を英語教育プログラムとその学生の実態に近づける努力を続けることで、SCL を客観性の高いエビデンスにすることができることが示された。

7. 1.2 結論2:事前事後テスト

第 5 章において、大学英語教育プログラムの 1 学年約 6000 人が事前リスニングテストを受験し、2 年後に、約 700 人が事後テストを受験した 2 学年分のデータを IRT によって分析した。この研究によって、事前事後テストを IRT によって分析することで、プログラム評価の信頼性や客観性の向上につながることを本論において述べてきた。

本研究の等化法の比較における新規性

5. 5.1「等化の方法 10 通りの比較」のところで述べたが、IRT の共通項目デザインによる等化法研究では、共通項目の数や質の違いに着目するのが一般的である。しかし、本研究では(1)等化する2つのテストの人数に大きな違いがある、(2)等化法を比較する基準をパラメタではなくθをにしている、(3)シミュレーション研究ではなく実データを使用している。これらの3つの点が、等化法比較における本研究の新規性である。普通、テスト等化法の先端的研究はテスト開発に関わる研究者が行うことが多いため、教育現場で授業の一環として実施されたテストデータを使うことはあまりない。ところが、実際の教育現場では事前テストは全員でも、事後は少数の受験者であれば実施可能となる場合も想定される。また、等化法を比較する基準にはパラメタを使う方法が一般的であるが、パラメタによる推定値で等化法を比較してθによる結果と異なる可能性がないとは言えない。従って、実際にその教育プログラムの学生の能力値であるθを基準にして等化法を比較するほうが、実態に適合した結果となると考えられる。これは、教育現場でこの条件での等化法を比較する研究が必要となった必然から生まれた新規性であるとも言える。

IRT による事前事後テスト分析による他のアドバンテージ

IRT を利用することで項目を同一尺度化できるため、同一項目でない事前事後テストであっても受験者の能力値θの変化が比較可能となる。そのため全員が事前事後両方のテストを受ける必要はない。例えば事前テストの約 14%が事後テストを受験しただけであっも事前事後両テストの受験者の平均能力値θ変化の推定に問題はない。また、ここで使用した事前事後テストはテスト妥当性を考慮して、実際に教えている内容をもとに教員が作成したものである。これらのテスト項目は、その英語教育プログラムで過去に使用され、IRT による項目パラメタも分かっている。従ってパラメタを考慮して適切な問題を選ぶことができる。このようにテスト問題を再利用することでテストの妥当性は高まり、教員の問題作成の負担も減る。さらに、等化を繰り返すことで項目バンクを作ることができるようになり、ひいてはテストの質のコントロールもできる。

また、第 5 章において示したように、習熟度レベル、学部、教員、学年ごとのθの伸びと、こ

れらの経年度変化も示すことができる。また、ニーズアンケートを実施してその結果と能力値 θ の変化を関連づけることも可能となる。

結論2:

事前事後テストは、教員の手による IRT 分析でプログラム評価の客観的なエビデンスとなることが明らかになった。

7. 1.3　結論3:大学英語教育プログラム EEP-J モデルの提案

本論で提案する EEP-J モデルは、大学英語教育プログラムにおけるプログラム評価の新しいモデルである。1. 9.3 で示したように、実施内容確認⇒アウトカム（Can-Do 自己チェックリスト、事前事後テスト）⇒改善計画という構成になっている。モデルの中心はアウトカムで、ここで2つの異なる角度（学生と教員）からの客観的エビデンスを提示する。これら2つのエビデンスのうち、既に Can-Do 自己チェックリスト（SCL）は結論1で、事前事後テストは結論2で、幾つかの過程や分析を経て客観性の高いエビデンスとする可能性について明らかにしてきた。EEP-J モデルのねらいは、学生は SCL に回答することによって、実際の状況の中でコミュニケーションとして「できない」や「あまりできない」SCL 項目を「できる」になるように修正能力を高める意識を持つこと、それと同時に、教員がプログラム評価のエビデンスとして学生の自己評価データを得ることである。さらに事前事後テストを IRT で分析することで得られる客観的エビデンスも併せ、EEP-J は教員がプログラムを多角的かつ体系的に評価して改善をめざすためのモデルである。

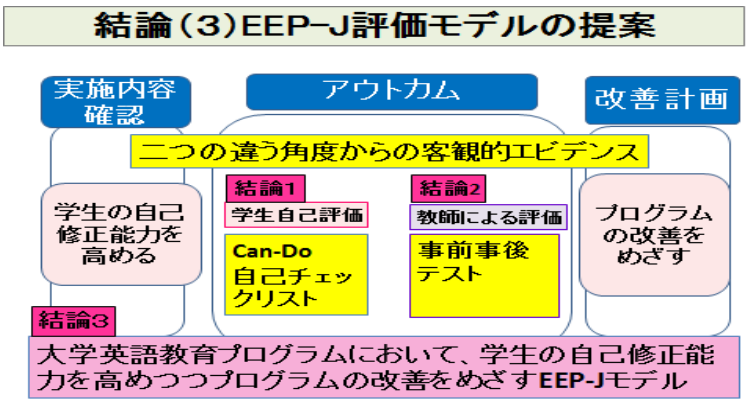


図 7.1.　EEP-J 評価モデル

EEP-J モデルの他大学での利用の可能性

第 6 章において大学英語教育プログラムそれぞれの中心的存在として従事する 40 人の教員に「プログラム評価アンケート」に回答してもらった。その結果 40 人中 32 人（80%）が、プログラム評価を実施するのは「教員である。」と回答した。この結果から、大学英語教育プログラムでは、教員がプログラム評価者となることが想定できる。また、プログラム評価によってもたらされるものは、プログラムの改善という声が多く、改善をめざす EEP-J モデルの目的とうまく適合すると思われる。そして、プログラム評価のエビデンスとして重要なものは 1 位が事前事後テスト、3 位が学生自己評価という回答結果であったことも、この2つをアウトカムの中心に据えている本モデルが他大学でも受け入れられることを暗示していると思われる。以上のように EEP-J モデルと教員へのプログラム評価アンケートの結果は大きく相反するところがないことから、本モデルが他大学でも利用される可能性は十分あるのではないかと思われる。

結論3:

大学英語教育プログラムモデルのアウトカムに、SCL と事前事後テストという2つの客観的エビデンスを設定することで、教員は、学生に自己修正能力を意識させながら、学生自己評価と事前事後テストによる客観的エビデンスを得ることができる。そして、プログラムの改善をめざす EEP-J モデルの開発ができた。また、英語教育プログラムアンケートの結果から、本モデルの他大学での利用の可能性は十分あると考えられる。

7．2． 今後の展望

SCL で測る尺度とリスニングテストの能力尺度 （θ s とθ o）

本論 3.6 では自己評価である SCL の結果を IRT により分析し、得られた自己評価の尺度を「θ s」とし、リスニング能力を筆記テストで測った能力値を「θ o」と呼ぶことにした。プログラム評価においては、異なる角度のエビデンスを求める必要があるため、θ s とθ o の両方を測るべきであると考えた。英語の筆記テストで測ろうとする能力は、非常に大きな領域の中の僅かなサンプル抽出したものでしかないようにも思える。また、実際の状況の中でコミュニケーションとして何ができるか記述された SCL は、筆記テストよりも大きな領域を包括するかもしれないが、答え方に個人差のある点で信頼性の高い評価とは言えない。この2つの尺度をの関連性を調査することを今後の課題とする。

また、筆記テストに 1 問ずつに回答した直後にその問題について、正解を得られる確信度を 0~100％で答える確信度テスト（張、2007）が θ o と θ s の関係を示すという考え方もあるが、SCL で問う自己評価の内容と確信度テストで問う自己評価の内容を同じ θ s として認識することができるかどうかも今後の研究課題としたい。

スピーキングテストの研究

リスニング＆スピーキングコースのプログラム評価では、スピーキング能力を測定するべきである。しかし、6000 人のスピーキング能力を事前事後テストとして公平に評価することは、リスニング事前事後テストに比べるとはるかに難しい。今後の展望としてコンピュータを使った音声認識システムによる一部自動採点による 3 人グループのオーラルテストの研究に着手する予定である。

EEP-J モデルの普及

第 6 章で実施した英語教育プログラムアンケートでは、EEP-J モデルが他大学でも利用される可能性に対する手ごたえを得ることができた。しかし、アンケートの中には、「学長に呼び出され学生の TOEFL 平均点が下がったことを咎められた」など、その英語教育プログラムで指導していない内容の筆記テストのスコアのみでプログラムの成果を測る傾向がまだ根強く残っているようである。今後は、積極的に普及のためのアクションを起こす予定である。英語教育プログラムアンケートに回答して下さった先生方は、日本言語テスト学会の会員が中心である。まず、当学会で EEP-J モデルについての研究発表や論文投稿を行い、着実に普及に努めたい。

7．3．　本論文の要約

日本の大学英語教育プログラムにおいてプログラム評価があまり実施されてこなかったのは、教員たちが、多角的な客観的エビデンスを得て的確に体系的に自らのプログラムの改善をめざすためのモデルがなかったからではないかと推測する。そこで、本論において「大学英語教育プログラムにおける EEP−J モデル」を提案し、学生の自己評価と事前事後テストという2つの異なった角度（学生と教員）からの客観的評価手法を中心としたプログラム評価モデルの開発を目指す。

EEP-J モデルは、①実施内容の確認⇒②アウトカム評価⇒③改善計画の 3 段階からなっている。①実施内容の確認では、プログラムの設定したゴール・目標と実施内容に矛盾がなく合

致しているか確認する。②アウトカム評価は、学生の自己評価と事前事後テストにより客観的なエビデンスを提示する。③改善計画では、アウトカム評価の結果をうけて改善のための行動を起こす。 Fettermanら (1996, 2001) が提唱したエンパワメント評価の特徴は、プログラムの参加者である教員が、自らのプログラムを評価して問題を発見し、方策を考えて改善を目指すことである。EEP-J モデルの「改善計画」は、エンパワメント評価の真髄である「改善のためのプログラム評価」を取り入れている。本モデルの中心であるアウトカムには、(1)学生の自己評価と(2)事前事後テストを据え、項目反応理論(IRT)を用い、できる限り信頼性の高い客観的エビデンスとして提示する。これは教員が学生に自己修正能力を意識させながらプログラムの改善をめざす為である。

第 2 章は、本論の背景である最近の日本の大学英語教育の傾向について、Can-Do statement (CDS)の授業での利用について考える。CDS はテストスコアのような数値ではなく、現実の状況のなかでコミュニケーションとして何ができるかを自然語で記述した文のことである。この CDS の中でも、多くの研究者たちが 20 年以上の年月をかけて作成し世界で最も普及しているのが Common European Framework (CEFR)である。ここでは CEFR を日本の大学英語教育に到達目標、評価基準、習熟度別クラスの基準などとして導入する取組と、これを日本人学習者に適用させるための先行研究についてまとめた。

第 3 章では、Can-Do 自己チェックリスト(SCL)を項目反応理論(IRT)を用いて分析し、そのプログラムの学習者に適応したものにカスタマイズする過程を、実データを使って研究した。T 大学のリスニングコースを担当する教員 10 名が作成したリスニング SCL を、約 330 人の学生に対して実施した。この結果に IRT を用いて項目困難度を推定し、教員が想定した難易度と比較してより学生の実態を反映したものに改訂した。このように IRT を用いて SCL の難易度を分析する先行研究は、既存の CDS の妥当性検証として実行したものはあるが、教育プログラムで利用するものはなく、これは本研究の新規性の 1 つである。

第 4 章では、SCL の question order effects(質問順序効果)に注目し、質問項目の順序が結果に与える影響を明らかにするための調査を実施した。その結果、回答者に適した難易度の差のない項目を質問することで質問順序効果を小さくする可能性が高いという示唆が得られた。IRT を用いて全てのレベルの SCL を同一尺度化することで、回答者は自分のレベルの SCL にしか答えなくても他のレベルの SCL との差は分かる。これにより多くの項目の SCL に回答することに比べ、回答者から得られる情報量は減るが、レベルに応じた短い SCL に回答することで以下のような利点が挙げられる。(1)学生の負担減、(2)レベルに合わない SCL に回答する必要

がない、(3)授業の到達目標として記憶・認識しやすい、(4)授業中に複数回実施しても支障が少ない、(5)質問順序効果の影響が少なくなる。

第 5 章では、ある大学英語教育プログラムの 1 学年約 6000 人が事前リスニングテストを受験し、2 年後にそのうちの約 700 人が事後テストを受験した。この事前事後テストのデータを用いて、10 通りの中から最適な等化の方法を判定する研究を実施した。本研究では(1)等化する2つのテストの人数に大きな違いがある、(2)等化法を比較する基準をパラメタではなく θ をにしている、(3)シミュレーション研究ではなく実データを使用、これらの3つの条件下で等化法を比較したことが本研究の新規性である。テスト等化法の先端的研究は専らテスト開発に関わる研究者が行うことが多く、教育現場で授業の一環として実施された実際のテストデータを使うことは希である。ところが、実際の教育現場では事前事後テストの受験者が大幅に違うことも容易に想定される。また、等化法を比較する基準にパラメタでなく θ を使っているのは、実際にその教育プログラムの学生の能力値である θ を基準にして等化法を比較するほうが、実態に合った結果となると考えられる為である。

本論で提案する EEP-J モデルの中心は、2つの異なる角度(学生と教員)からの客観的エビデンスを提示するところである。学生は SCL に回答することによって、実際の状況の中でコミュニケーションとして何ができるか考え自己修正能力を高める意識を持ち、教員は学生の自己評価や事前事後テストのデータを IRT によってより客観性の高いプログラム評価エビデンスとする。このように、EEP-J は教員がプログラムを多角的かつ体系的に評価して改善をめざすためのモデルである。そして、第 6 章で実施した英語教育プログラム評価アンケートの結果と EEP-J モデルの整合性は低くなく、EEP-J モデルが他の大学でも利用される可能性は高い。

謝辞

2006 年に入学して以来、変わらず親切にご指導いただいた前川眞一教授に心から感謝します。先生のご指導を受けていなければ、本研究を完成させることは決して叶わなかったと思います。水曜日しか大学に来られない私のために、毎週ゼミを水曜日にして下さったおかげでここ数年、ほとんど毎回ゼミに出席することができました。そして、何と言っても前川先生が、時には学生室の私のデスクパソコンのスクリーンを何時間も一緒に見て下さり、気長に、忍耐強くご指導して下さったおかげです。

2011 年に中間発表をしましたが、その時に審査をして下さった先生方のうち大学入試センターの石塚智一教授は非常に残念なことにお亡くなりになり、牟田博光教授は退官され、4年の月日がとても長く感じられます。中間発表の時に、大変貴重なアドバイスをいただいた両先生に心から感謝いたします。また、牟田先生には、お忙しい中何度か時間を作っていただきアドバイスをしていただきました。研究についてはもちろんのこと、私がサバティカル先を迷って相談した時、「ケンブリッジにしなさい。有名な大学に行くと、あなたがそこに居るだけで世界中の立派な研究者たちが集まってきて、知り合いになれる。」とおっしゃって下さり、そのお言葉のとおり、今でも私はケンブリッジで知り合った友人たちに助けられています。中間発表だけでなく、中川正宣教授と室田真男教授には最終審査までお付き合いいただき、本当にありがとうございます。また、中山実教授と山元啓史准教授には、お忙しいなか拙論の審査に関わっていただき、深く感謝いたします。

前川研究室にかつていらっしゃった村山航先生、岡田謙介先生、宇佐美慧先生には、研究に関して度々ご助言をいただきました。また、栗山直子先生にも博士論文のまとめ方について教えていただきました。荒井清佳さん、野上康子さん、光永悠彦さん、沖嘉訓さんにも研究面で教えていただくことが数多く、皆さま、大変お世話になりました。

それから、本論の発想の源ともなった日本英語検定協会の委託研究プロジェクト「言語テストの規準設定」のメンバーに私を入れて下さり、研究について様々なアドバイスを下さった大友賢二先生、そして 3 年間プロジェクトを進める中で多くのことを教えて下さり、研究に関して刺激を下さった渡部良典先生、伊東祐郎先生、法月健先生に心から感謝いたします。

最後になりましたが、お忙しい中、英語教育プログラム評価アンケート調査に協力して下さった多くの先生方、またそのアンケートのために多くの先生を紹介して下さった松本佳穂子先生、そして、データ収集に協力いただいた東海大学外国語教育センターの先生方、先生方のご協

力があったからこそ、本論を完成させることができました。深く感謝します。

そして、2つめの博士号に挑戦するという私の暴挙にもあきれず、8年間励まし続けてくれた母に感謝します。

参考文献

Alderson, C., and Beretta, A. (1992). *Evaluating second language education.* Cambridge, England: Cambridge University Press.

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

Angoff, W.H.(1984). *Scales, norms, and equivalent scores.* Princeton, NJ: Educational Testing Service.

Arai, S., and Mayekawa, S. (2011). A comparison of equating methods and linking designs for developing an item pool under item response theory. *Behaviormetrika, 38* (1)1-16 .

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

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

Bachman, L. F. (1990). *Fundamental considerations in language testing.* Oxford: Oxford University Press.

BILOG-MG version 3. (2003). Lincolnwood: Scientific Software International.

Boldt, R., and Ross, S. (2005). Language Proficiency Gain on the Test of English for International Communication: Meta-Analyses of Japanese and Korean Corporate Language Programs, from https://www.toeic.or.jp/toeic_en/pdf/newsletter/boldt_ross2005.pdf

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

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

Brown, J. D. (1996). *Testing in Language Program.* New Jersey: Prentice Hall Regents.

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

Busch, M., Elsea, M., Gruba, P., and Johnson, F. (1994). A study of the needs, preferences and attitudes concerning the learning and teaching English proficiency as expressed by students and teachers at Kanda University. *The Journal of Kanda University of International Studies,*6, 1-62.

Campbell, D.T. (1969). Reforms as Experiments. *American Psychologist, 24,* 409-429.

Chen, H. T. (2005). *Practical Program Evaluation: Asessing and improving planning implementation, and effectiveness.* Newbury Park, CA: Sage.

Choi, I. C., and 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., and Gorsuch, G. (2000). Using item response theory to refine placement decisions. *JALT Journal, 22*(2), 315-325.

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

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

Cronbach, L. J. (1984). *Essentials of psychological testing* (4th ed). New York: Harper & Row.

Cronbach, L. J. (1989). Construct validation after thirty years. In R. L. Linn (Ed.). *Intelligence: measurement theory and public policy* (pp. 147-171) Urbana, IL: University of Illinois Press.

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

Fetterman, D. M. (2001). *Foundations of Empowerment Evaluation.* Thousand Oaks, CA: Sage.

Fetterman, D. M., Kaftarian, S. J., and Wandersman, A. (1996). *Empowerment Evaluation*: *Knowledge and tools for self-assessment and accountability.* Thousand Oaks, CA: Sage.

Fetterman, D.M. and Wandersman, A. (2005). *Empowerment Evaluation Principles in Practice.* Guildford Press. NY.

Fujita, T. (2001). Peer, self-, and instructor assessment in an EFL speech class. *The Journal of Rikkyo University Language Center, 3,* 197-207.

Fujita, T. (2002). The effectiveness of peer review in a second language writing class in Japan. *The Bulletin of Foureign Language Center, 23,* 1-15.

Fujita, T. (2005). *Validation of a Japanese university English language placement test.* Ann Arbor, MI: UMI. ProQuest Information and Learning.

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

Geoff Brindley (2001). Outcomes-based assessment in practice: some examples and emerging insights. *Language Testing 18* (4) 393-407.

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

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research, 22*(3), 144-149.

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

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

Horigtuchi, S., Harada, Y. Imoto, Y., and 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 and K. Sakai (Eds.), *Can-do statements in language education in Japan and beyond - Applications of the CEFR* (pp. 138-154). Tokyo: Asahi Press.

Jodoin, M., Keller, L., and Swaminathan, H. (2003). A comparison of linear, fixed common item, and concurrent parameter estimation equating procedures in capturing academic growth. *The Journal of Experimental Education, 71,* 229-250.

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

Keller, L., Keller, R., and Parker, P. (2011). The examination of the classification of students into performance categories by two different equating methods. *The Journal of Experimental Education, 79*, 30-52.

Kikuchi, K. (1995). Student and teacher perceptions of learning needs: A cross analysis. *JALT Testing and Evaluation SIC Newsletter,* 6(2), 8-20.

Kim, D., Choi, S., Lee, G., and Um, K. (2008). A comparison of the common-item and random-groups equating designs using empirical data. *International Journal of Selection and Assessment, 16*(2), 83-92.

Kiely, R., and Rea-Dickins, P. (2005). *Program evaluation in language education.* London: Palgrave MacMillan.

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

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

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.

Lenz, P., and Schneider, G. (2004). *A bank of descriptors for self-assessment in European language portfolios.* Strasbourg: Council of Europe.

Lee, W., and Ban, C. (2010). A comparison of IRT linking procedures. *Applied measurement in education, 23,* 23-48.

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.

Lynch, B. K. (1996). *Language program evaluation.* Cambridge, England: Cambridge University Press.

Lenz, P., and Schneider, G. (2004). *A bank of descriptors for self-assessment in European language portfolios.* Strasbourg: Council of Europe.

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

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

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement, 14,* 139-160.

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.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 35,* 1012-1027.

Messick, S. (1988). The once and future issues of validity: Asessing the meaning and consequences of measurement. In H. Wainer & Braun (Eds.), *Test validity* (pp. 33-45). Hillsdale, NJ: Lawrence Erlbaum.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23,* 13-23.

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

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.

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

Naganuma, N., and Miyajima, M. (2006). The development of Seisen academic Can-Do framework. *Bulletin of 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.

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

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

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

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

Patton, M. Q. (1997). *Utilization-Focused Evaluation* (3rd ed.). Thousand Oaks, CA: Sage.

Plake., and 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.

Popham, W. J. (1978). *Criterion-referenced measurement.* Englewood Cliffs, NJ: Prentice-Hall.

Popham, W. J. (1981). *Modern educational measurement.* Englewood Cliffs, NJ: Prentice-Hall.

Popham, W. J. (1997). Consequential validity: Right concern-wrong concept. *Educational measurement: Issues and practice, 16,* 9-13.

Purcell, S. (2007). *PLINK* (version 2). http://pngu.mgh.harvard.edu/purcell/plink/

Rea-Dickens, P., and Germaine, K.P. (1998). The price of everything and value of nothing: Trends in language program evaluation. In P. Rea-Dickens andK.P. Germaine (Eds.), *Managing evaluation and innovation in language teaching: Building bridges*. (pp.3-20) London: Longman.

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

Ross, S. (2003). A diachronic coherence model for language program evaluation. *Language Learning 53*, (1), 1-33.

Ross, S. (2009). Program evaluation. In Michael H. Long and Catherine J. Doughty (Eds.), *The Handbook of Language Teaching.* (pp. 756-778) Oxford, Wiley-Blackwell.

Rossi, P. H., Freeman, H. E., and Lipsey, M.W. (1999). *Evaluation: A systematic approach*.6th ed., Thousand Oaks, Calif.: Sage Publications.

Saida, C., & Hattori, T. (2008). Post-hoc IRT equating of previously administered English tests for comparison of test scores. *Language Testing, 25*(2), 187-210.

Saito, H., and Fujita, T. (2004). Characteristics and user acceptance of peer rating in EFL writing classrooms. *Language Teaching Research 8,* 1. 30-53.

Saito, H., and Fujita, T. (2009). Peer-assessing peers' contribution to EFL group presentations. *RELC Journal, 40*(2). 149-171.

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

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

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

Stocking, M. L., and Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*(2), 201-210.

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.

Wandersman, A. (2003). Community science: Bruidging the gap between science and practice with community-centered models". *American Journal of Community Psychology,* 31, 227-242.

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

Weiss, C.H. (1998). *Evaluation: Methods for studying programs and policies* (2nd ed.), Upper Saddle River, NJ: Prentice Hall.

Wholey, J. S. (2004). Evaluability Assessment. In J. S. Wholey, H. P. Hartry, & K. E. Newcomer (Eds.), *Handbook of Practicl Program Evaluation* (pp. 33-62) Jossey-bass A Wiley Imprint. .

Yu, J., and Murphy, K. (1993). Modesty bias in self ratings of performance: a test of the cultural relativity hypothesis. *Personnel Psychology 46*, 357-63.

Zu, J., and Liu, J. (2010). Observed score equating using discrete and passage-based anchor items. *Journal of Educational Measurement*. *47*(4), 395-412.

伊東祐郎（2008）『日本語教師のためのテスト作成マニュアル』東京:アルク.

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.

植山剛行（2004) コンテックスト手法によるプログラム評価:英語教育への応用『教育制度研究紀要』35, 15-31. 日本大学教育制度研究所

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

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

大友賢二 (1996)『項目応答理論入門』. 東京: 大修館.

鎌田倫子・中河和子・峯正志・後藤寛樹 （2010） エンパワメント評価の可能性と限界:原理と特徴より 『研究紀要:富山大学杉谷キャンパス一般教育』(38) 55-70. http://hdl.handle.net/10110/11255

鎌田倫子・中河和子・後藤寛樹 （2012） 理科系キャンパスの小規模日本語プログラムにおけるエンパワメント評価の実践 『研究紀要:富山大学杉谷キャンパス一般教育』(40) 45-62. http://hdl.handle.net/10110/11231

金谷憲他編（2003）『英語教育評価論:英語教育における評価行動を科学する』河源社

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

串本 剛(2006) 大学教育におけるプログラム評価の現状と課題『広島大学 高等教育研究開発センター 大学論集』第 37 集 263-276.

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, 127-140.

斉田智里, 小林邦彦., and 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.

芝祐順(編) (1991)『項目応答理論－基礎と応用』東京大学出版会.

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

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」 財団法人日本英語検定協会『STEP BULLETIN』 vol. 20.

張一平 (2007) 『確信度テスト法と項目反応理論』 東京:東京大学出版.

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

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

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

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

中村和彦 (2008) アクションリサーチとは？『人間関係研究』南山大学 人間関係研究センター 第 7 号 1 - 25.

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 自己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.

根岸雅史 (2005)「日本における英語能力記述の枠組みの開発」『ARELE: annual review of English language education in Japan』全国英語教育学会, 16, pp. 191-200.

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

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

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

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

前川眞一 (1991) パラメタの推定.　芝祐順編『項目応答理論』(第 4 章, 87-129). 東大出版会.

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. 38－48. くろしお出版.

光永悠彦・前川 眞一 (2012) 項目反応理論に基づくテストにおける項目バンク構築時の等化方法の比較『日本テスト学会誌』8, 31-48.

三好皓一・田中弥生（2001）参加型評価の将来性: 参加型評価の概念と実践についての一考察『日本評価研究』日本評価学会　第 1 巻　第 1 号　65-79.

安田節之(2010) プログラム評価の意義と展望:方法論の視点から『人事試験研究』214, 2-15.

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

山中司（2007） 大学英語教育における「プログラム評価」導入と実施の提言『政策情報学会』2 (1) 21-36.

山中司・鈴木祐治（2006） 大学英語教育評価論の新たなパラダイムを見据えて『政策情報学会』1 (1) 53-85.

安田節之・渡部直登（2011） プログラム評価研究の方法. 下山晴彦（シリーズ編著）. 臨床心理学研究法第 7 巻. 新曜社.

安田節之（2010） プログラム評価の意義と展望:方法論の視点から『人事試験研究』. No. 214. 2-15.

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.

吉島茂・大橋理枝（訳編）（2008）『外国語教育 II- 外国語の学習、教授、評価のためのヨーロッパ共通参照枠』東京:朝日出版社.

渡辺直登 (2000) プログラム評価研究 下山晴彦（編著）, 臨床心理学研究の技法 147-156 福村出版

Appendix　A：Can-Do 自己チェックリスト項目難易度（第 3 章）

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

Appendix B: Can-Do 自己チェックリスト　難易度別フォーム　（第 4 章）

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

Form L

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

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

最後に、このアンケートの答えやすさを教えて下さい

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

Appendix C: プログラムニーズアンケート(第５章)　　(Kikuchi (2005) から一部引用)

これは「英語の授業を改善する」ためのアンケートです。以下の質問に合意するか４段階基準（１．そう思わない、　２、あまりそう思わない、　３、まあそう思う　４、そう思う）をマークしてください。

	１．そう思わない、　２、あまりそう思わない、　３、まあそう思う　４、そう思う	1	2	3	4
1	他の学習者とペアを組んだり、3−4 人のグループで勉強するのが好きだ。	○	○	○	○
2	ひとりで（ペアやグループとはなく）学習するほうがはかどる。	○	○	○	○
3	授業が教科書にきちんと沿っているとよく学習できる。	○	○	○	○
4	教室が和気あいあいとしているときが一番勉強できる。	○	○	○	○
5	ビデオやテレビなどを使って勉強するのが好きだ。	○	○	○	○
6	自分のライティング（筆記の）課題をお互いに直しあうと勉強になる。	○	○	○	○
7	LL 教室などでコンピュータ/CD/カセットなどを使って勉強するのが好きだ。	○	○	○	○
8	ひとつのスキルだけ（「リーディング」ならそれだけ学ぶ）授業がよい。	○	○	○	○
9	聞く話す読む書くあわせた総合英語の授業がよい。	○	○	○	○
10	英語の学習に学習支援センターや図書館を利用する。	○	○	○	○
11	日本人教師から英語を学ぶのはためになる。	○	○	○	○
12	先生が教室を歩き回り、一人一人生徒に指導してくれるのがよい。	○	○	○	○
13	先生が授業を楽しくしてくれるとよく勉強できる。	○	○	○	○
14	先生が厳しく授業をしてくれるとよい。	○	○	○	○
15	先生がテストをしてくれたり、宿題をだしてくれるとよく学習できる。	○	○	○	○
16	先生が日本語で説明をしてくれるとよく学習できる。	○	○	○	○
17	外国人の先生から英語を学ぶのが好きだ。	○	○	○	○
18	小さなテストを何回か受ける方が、1 回の大きなテストよりも好ましい。	○	○	○	○
19	筆記テストのほうが、会話のテストよりも得意である。	○	○	○	○
20	期末テスト対策の方法（ストラテジー：方略）をもっと知りたい。	○	○	○	○
21	可能であれば、コンピュータを使ってテストを受けたい。	○	○	○	○
22	選択肢から答えるテストよりもエッセイ・ライティングによるテストのほうが好ましい。	○	○	○	○
23	テストがあると思うと良く勉強する。	○	○	○	○
24	可能な限り、よい成績をとることは重要だ。	○	○	○	○

25.　今まで TOEIC, TOEFL、英検のどれかをうけたことがありますか？	ない ○	ある ○

Appendix D: 英語教育プログラム評価を実施された方へのアンケート

（お名前：　　　　　　　　　　　　　本務校：　　　　　　　　　　　　）

1. プログラム評価を実施された目的は何ですか？

・・

2. 貴言語教育プログラムの**ゴールや教育目標**は設定していらっしゃいましたか？　いる / いないを（　　）にお願いします。　　答え：**設定して（　　　）。**

設定している場合、それはどのようなものですか？

・・

3. プログラム評価の実施を主に実行されたのは、どのような立場の方ですか？（例：教員、外部の専門家に委託など）

・・

4. プログラム評価の結果は誰(または何）に対して報告されましたか？

・・

5. 以下のうち、どれに重点を置いてプログラム評価をされましたか？A~Fでお答え下さい（複数回答可)。またその理由は何でしょうか？

（A）プログラム理念（ゴール/目標）　　（B）カリキュラムの内容例：教員、授業内、教材
（C）プログラムによる効果　　　　　　（D）学生の反応
（E）教員の反応　　　　　　　　　　　（F）その他（　　　　　　　　　　　　　　　）

重点１（　　　）　重点２（　　　）　重点３（　　　）
理由：

・・

6. プログラム評価を実施したことによって何がもたらされましたか？

・・

7. 統計を含むテスト科学は、プログラム評価にどれくらい役立ちましたか？　１~100％でお答え下さい。

（　　　　　％）

8.　以下の８項目について、**プログラム評価のためのデータとして**①**重要**だと思われる項目の（　）に**ｉ** を、次に、②実際に**実施した**項目の（　）には **a** を記入し（複数回答可)、最後にその理由を書いて下さい。

	①重要性	②実施の有無
＊学生の市販テストのスコア	（　）	（　）
＊学生の事前事後（入口出口）テストのスコア	（　）	（　）
＊学生の授業内容に対する到達度を測る中間テストや期末テストのスコア	（　）	（　）
＊学生の自己評価　例、Can-Do 自己チェックリスト	（　）	（　）
＊学生の授業に対する評価　例、学生による授業アンケート	（　）	（　）
＊教員の学生に対する評価　例、最終総合成績、授業中の学生観察	（　）	（　）
＊教員たちのカリキュラムに対する評価　例、アンケート	（　）	（　）
＊その他　（　　　　　　　　　　　）	（　）	（　）

理由：

・・

9. プログラム評価を実施されて、どのような問題点があったか具体的に書いていただけないでしょうか？

。。。

10.　**貴言語教育プログラムの成果**（実際にそのプログラムでゴール/教育目標としていることが、どのくらい達成できたか）を確認するための事前事後（入口出口）テストとしては、**市販テスト**と、そのプログラムの**教員作成テスト**のどちらが適切だと思われますか？

答え：(　　　　　　テスト)

理由：

ご協力いただき、本当にありがとうございました。

Appendix E: 言語教育プログラム評価を実施していない方へのアンケート

（お名前：　　　　　　　　　　　　　　本務校：　　　　　　　　　　　　）

1．プログラム評価を実施しない理由は何ですか？

・・

2．貴言語教育プログラムの**ゴールや教育目標**は設定していらっしゃいますか？　いる / いないを（　　）にお願いします。　　答え：**設定して（　　　）。**

設定している場合、それはどのようなものですか？

・・

3.　もしプログラム評価を実施するとしたら、主に実行されるのは、どのような立場の方ですか？（例：教員、外部の専門家に委託など）

・・

4.　もしプログラム評価を実施するとしたら、その報告をするのは誰(または何）に対してですか？

・・

5. もしプログラム評価を実施するとしたら、以下のうちでどれに重点を置いて評価をされますか？A~Fでお答え下さい（複数回答可)。またその理由は何でしょうか？

（A）プログラム理念（ゴール/目標）　　　（B）カリキュラムの内容、（例：教員、授業内容、教材）
（C）プログラムによる効果　　　　　　　（D）学生の反応
（E）教員の反応　　　　　　　　　　　　（F）その他（　　　　　　　　　　　　　　　　　）

重点１（　　　）　重点２（　　　）　重点３（　　　）

理由：

・・

6. もしプログラム評価を実施したとしたら、何がもたらされると思いますか？

・・

7. 統計を含むテスト科学は、プログラム評価にどれくらい役立つと思われますか？　1~100%でお答え下さい。

（　　　　　%）

8.　以下の８項目について、**プログラム評価のためのデータとして**①重要だと思われる項目の（　）に **i** を、次に、②実際に実施したいと思われる項目の（　）には **w** を記入し（複数回答可)、最後にその理由を書いて下さい。　　①重要性　②実施したい

	①重要性	②実施したい
＊学生の市販テストのスコア	（　）	（　）
＊学生の事前事後（入口出口）テストのスコア	（　）	（　）
＊学生の授業内容に対する到達度を測る中間テストや期末テストのスコア	（　）	（　）
＊学生の自己評価　例、Can-Do 自己チェックリスト	（　）	（　）
＊学生の授業に対する評価　例、学生による授業アンケート	（　）	（　）
＊教員の学生に対する評価　例、最終総合成績、授業中の学生観察	（　）	（　）
＊教員たちのカリキュラムに対する評価　例、アンケート	（　）	（　）
＊その他　（　　　　　　）	（　）	（　）

理由：

・・・

9.　もしプログラム評価を実施するとしたら、どのような問題点が予想されるか具体的に書いていただけないでしょうか？

10.　**貴言語教育プログラムの成果**（実際にそのプログラムでゴール/教育目標としていることが、どのくらい達成できたか）を確認するための事前事後（入口出口）テストとしては、**市販テスト**と、そのプログラムの**教員作成テスト**のどちらが適切だと思われますか？

答え：(　　　　　　テスト)

理由：

ご協力いただき、本当にありがとうございました。