神戸大学　大学教育推進機構 『大学教育研究』 第 25 号　2017 年 3 月：63－81

大学英語教育における形成的評価：
統計手法を用いた形成的評価データの特性の解明

Formative Assessment in Tertiary-Level English Education:
A Statistical Analysis of Formative Assessment Data

石川　慎一郎（神戸大学　大学教育推進機構　教授）

要旨

大学英語教育においては，かつては，総括的なテストによって成績をつけることが少なくなかったが，英語力，とくに，コミュニケーション英語力の多元性への理解が深まる中で，近年，指導と評価のプロセスを一体化させる形成的評価が次第に普及しつつある。もっとも，指導の目的や過程が明確に定まっている中学校や高等学校に比べると，大学英語教育における形成的評価の導入には課題も多い。本論文では，実際の指導過程で得られた 7 種の形成的評価指標を用い，各指標の特性・相互関係・分類，各指標値と外部試験成績の関係，主成分分析を用いた指標の合成，異なる指標合成手法の選択が最終成績に及ぼす影響について実証的検討を行った。その結果，複数の形成的評価指標データを活用することで，学習者のコミュニケーション英語能力をより多面的に評価できる可能性が示唆されたが，一方で，指標の合成手法の選択が最終成績に大きな影響を及ぼすことも明らかになった。

1.　はじめに

大学英語教育の質的改革が求められる中で，教材や教授過程だけではなく，評価についても関心が高まっている。英語力の評価に関しては，中学校・高等学校のレベルでは，すでに 2000 年ごろから大規模な変革が進んでおり，学習者の英語力を複数の観点でとらえ，日々の授業過程の中で形成的に評価するスタイルが定着しているところであるが，大学英語教育はこれまでこうした動きとほぼ無縁であった。

もっとも，指導要領や検定教科書によって，単元ごと，学期ごと，学年ごと，学校段階ごとの達成目標が明確に示され，それに基づき評価が行われている中等学校と異なり，大学英語教育においては，どのような評価データを収集するか，収集した評価データをどのように組み合わせるか，授業外の能力試験等で測定された英語力を加味すべきかどうか，といった点について，必ずしも共通理解が醸成されていない。そこで，本稿では，国立大学における初年次の英語教育の実践をふまえ，これらの点について探索的な検討を試みることとする。

2. 総括的評価と形成的評価：英語力の多元性をめぐって

2.1 英語力のとらえ方

近年の応用言語学の潮流をふまえ，大学で涵養するべき英語力を「コミュニケーション言語能力」と規定する場合，そこには多様な要素が包含される。Lyle Bachman は，コミュニケーション言語能力の内部構成に関して以下のようなモデルを提案している（Bachman, 1990）。

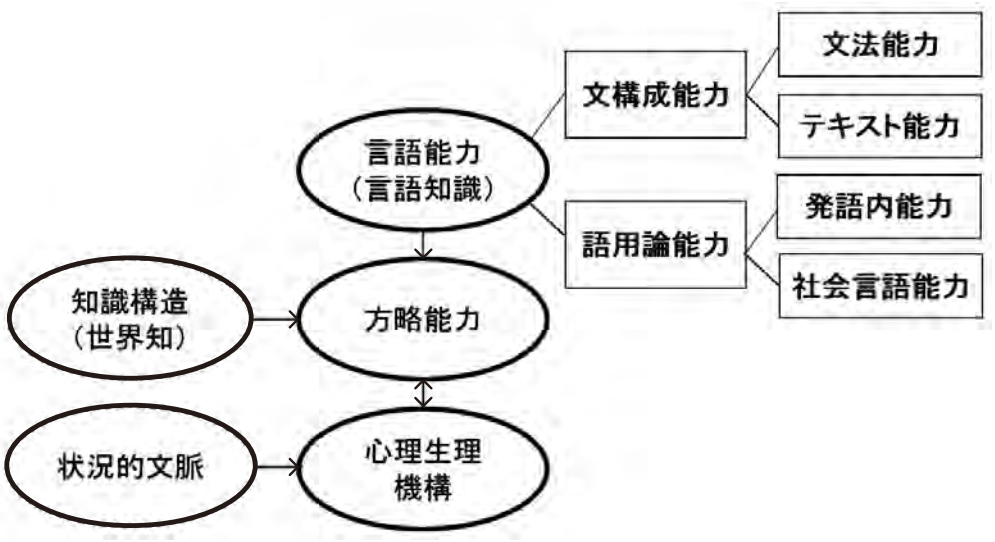

図 1 Bachman のコミュニケーション言語能力モデル

（Bachman，1990，p.85 の Fig.4.1 と p.87 の Fig.4.2 を改変して筆者が作成）

Bachman のモデルでは，コミュニケーション言語能力は，言語に関する知識や運用力に関わる言語能力（language competence）だけでなく，言語使用の計画・遂行・調整といった認知的処理に関わる方略能力（strategic competence）や，所与の環境下での実際的な言語使用の遂行に関わる心理生理機構（psychophysiological competence）からなるとされる。このうち，言語能力には，語彙・形態論・統語論等に関わる文法能力（grammatical competence）や，結束性を保って文を構成するテキスト能力（textual competence）に基づき，文法的に正確な文を構成する文構成能力（organizational competence）と，相手の発話意図の理解に関わる発語内能力（illocutionary competence）や，方言・言語使用域（register）・文化的意味・比喩の理解に関わる社会言語能力（sociolinguistic competence）に基づき，対人的・社会的関係性の中で文を適切に使用する語用論能力（pragmatic competence）の両面が含まれる。Bachman のモデルに従えば，いわゆる英語力は狭義の言語能力に限定されるものではなく，方略能力を介して幅広い要素を巻き込み，一般常識や文脈の理解力を前提として，他者との関係性の中で言語を正確かつ適切に使用する能力の総体であることがわかる。

英語力を多元的にとらえる姿勢は，我が国の小学校・中学校・高等学校における指導の目標や指針を定めた指導要領にも見出すことができる。高校を例にして言えば，外国語教育で涵養すべき資質･能力は,外国語を通じた「言語や文化に対する理解」,「積極的にコミュニケーションを図ろうとする態度」,「情報や考えなどを的確に理解したり適切に伝えたりするコミュニケーション能力」の 3 点で定義される（「高等学校指導要領」第 8 節･第 1 款)。また,「コミュニケーション英語Ⅰ～Ⅲ」においては，最終的に「社会生活において活用できるようにする」ことが目指されている（第 8 節・第 2 款・第 2 ～ 4)。こうした一連の目標記述から浮かび上がってくるのは，態度や文化理解も含め，英語力を，実社会においてコミュニケーション活動を遂行するための幅広い能力ととらえる立場である。

こうした英語力観は，観点別評価の枠組みの中でさらに具体的に規定されている。そもそも，学校教育における評価観点については，1977 年（昭和 52 年）の指導要領改訂に伴う指導要録の見直しにおいて全教科に共通する観点として「関心・態度」が示され，その後，1989 年（平成元年）の指導要領改訂に伴う指導要録の見直しで「関心・意欲・態度」,「思考・判断」,「技能・表現（表現)」,「知識・理解」の 4 観点が明示された。これにより，すべての教科の学力が，主体的に学習に取り組む姿勢（関心・意欲・態度）を持ち，基礎的･基本的な知識や技能を獲得し（知識･理解／技能)，必要な認知的判断を行って（思考･判断・表現)，課題を解決する力と定義されたことになる。2002 年（平成 14 年）には小学校･中学校向けに，また，2004 年（平成 16 年）には高等学校向けに,「評価規準の作成,評価方法の工夫改善のための参考資料」が提示され, 2000 年代後半には, 小中高において,こうした観点に基づく評価規準の作成や評価が実施されるようになった。

なお，外国語については，その特性をふまえ,「コミュニケーションへの関心・意欲・態度」,「外国語表現の能力」,「外国語理解の能力」,「言語や文化についての知識・理解」という 4 観点が示されている。もっとも,「外国語表現」というのは，単なる作文力・発話力ではなく，思考や判断と，その結果としての表現を一体的にとらえた概念であり,「外国語理解」も，単なる読解力・聴解力ではなく，各種の資料を理解し，わかったことをまとめて発表するまでの過程を一体的にとらえた概念である。この意味で，他の教科と変わるものでなはい。

このように，応用言語学においても，また，小中高の英語教育においても，英語力の多面性についての理解が進み，これに対応した評価のシステムが整備されてきているところであるが，大学英語教育について言えば，この面での対応はいまだ必ずしも十分とは言えない。

2.2 総括的評価から形成的評価へ

前節で述べたように,近年,英語力を多元的にとらえる立場が一般的になっている。では,そうした能力はどのような形で評価されるのであろうか。大学英語教育では，従来，定期

試験のスコアだけで機械的に評価を行うことが少なくなかった。こうしたアプローチを一般に総括的評価（summative assessment）と呼ぶ。総括的評価の場合，学習内容の全体を総括するテストで一定のスコアを得ていることをもって，所与の教育目標が全体として達成されたと判断する。

しかし，いかに定期試験の内容を工夫したとしても，多様な英語力の諸相を 1 回の試験で網羅的に評価することには無理がある。たとえば，対人関係や社会的文脈をふまえた語用論的能力，認知的な方略能力，さらには，コミュニケーションへの関心・意欲・態度，また，異文化の知識や理解などを信頼できる形で問うことは容易ではない。

この点をふまえると，総括的評価に代え，授業過程と評価過程を一体化させ，日々の指導の場面で多様な評価データを体系的に収集していくという方向が考えられる。こうした評価の在り方を一般に形成的評価（formative assessment）と呼ぶ（石川，2014b ; 石川，2015）。

経済協力開発機構（OECD）の報告書は，総括的評価と形成的評価の関係について，次のように述べている。

> 評価は教育過程に欠かせないものである。最もわかり易い評価は，テストなどによって生徒が何を学習したかを測定するもので，言い換えれば，学校側に生徒の達成度について説明責任を負わせる総括的評価である。しかし，「形成的な」評価もあり得る。形成的評価とは，生徒の学力向上や理解度を頻繁かつ双方向的に評価することである。そうすれば，教師は，明らかになった学習ニーズに合わせて授業のやり方をよりうまく調整することができる。形成的評価が総括的評価と異なるのは，形成的プロセスで収集した情報を達成度の要約としてではなく，学力の向上に利用するという点にある。学校レベルや政策レベルで形成的評価の原則を活用すれば，学力を向上させるべき分野を特定したり，教育システムの中で評価に対する建設的な雰囲気を形成したりすることができる。各種調査によれば，形成的評価は生徒の達成度を向上させる最も効果的な戦略の 1 つである。また，形成的評価は生徒間の達成度の格差を少なくしたり，生徒が「学び方を学ぶ」スキルを開発したりするためにも重要である…（OECD, 2005b，一部改変）

「頻繁かつ双方向的」な評価を行う際には，「学習到達目標に対応した学習活動の特質等に応じて，多肢選択形式等の筆記テストのみならず，面接，エッセー，スピーチ等のパフォーマンス評価，活動の観察等，様々な評価方法の中からその場面における生徒の学習状況を的確に評価できる方法を選択することが重要」となる（文部科学省，2013）。形成的評価は，正しく実践されれば，近年注目されているアクティブラーニングなど，教え方や学び方の根本的な改革にもつながるものと言える（石川，2016）。

ただし，形成的評価の効果的な実践方法については必ずしも共通理解が存在せず，現場

での模索が進められている段階である。前出の OECD のレポートの後半には以下のような言葉が続く。

> …しかし，形成的評価は体系的に実践されていない。特に…イノベーションや変革への障害を克服するのに困難を伴うことが多い中等学校ではそうである。こうした障害としては，教室をベースとした形成的評価と学校の説明責任が一目瞭然となる総括的テスト（教師はテストのための授業をする傾向にある）との間に緊張関係が存在することや，評価と評定への体系的なアプローチ，学校側のアプローチ，教室におけるアプローチの間に繋がりが欠けていることなどが挙げられる。（OECD, 2005b, 一部改変）

前述の問題は，中等学校に限られるものではない。大学英語教育における英語力の評価にもまさに同様の問題が残るのである。

3. リサーチデザイン

3.1 ねらいと RQ

日々の授業において学習者の英語力を多面的・継続的に測定することで，教授と評価を一体化させ，評価を学習者の支援や授業・カリキュラムの改善にも活用する形成的評価は，大学英語教育においてもメリットの大きいものであるが，一方で，評価の妥当性の担保は必ずしも容易ではない。たとえば，どのような観点で評価するのか，様々な方法で得られた評価データをどのように合成して最終評価に落とし込むのか，近年，大学英語教育で強調されている TOEIC や TOEFL 等の外部試験を受験させた場合，その結果をどのように位置付けるべきなのか，といった点について，明確な立場は存在しない。

本稿は，筆者の授業実践で得られた評価データを資料として，これらの点を探索的に検討することを目標とする。この目的に沿い，以下の 4 つのリサーチクエスチョン（RQ）を設定した。

RQ1 授業内で収集される 7 種類の形成的評価指標のうち，学習者の英語力を弁別しやすいものは何か。また，各々の指標はどのように相関しており，どのように分類されるか。

RQ2 授業内で収集される 7 種類の形成的評価指標のうち，外部試験指標と相関が高いものは何か。

RQ3 授業内で収集される 7 種類の形成的評価指標を主成分分析法によって合成する場合，全体の中で重みづけが高くなるものは何か。

RQ4 授業内で収集される 7 種類の形成的評価指標を異なる方法で合成した場合，最終的な評価結果にどのような差異が生じるか。また，学生の専攻タイプによる影響

はあるか。

3.2 形成的評価データの収集

すでに述べたように, 形成的評価では評価と指導が一体化される。本研究で分析するデータは，筆者が関西圏の研究型国立大学において，2016 年度の第 1 クォーター（8 回）に担当した「English Communication」の授業で取得したものである。

この授業は「口頭英語を中心とした英語能力の向上」をテーマとするもので,「英語の発音や表現パターンの理解，また，英語による口頭での意見交換などを通し，聞く力と話す力を中心とした英語能力の総合的開発を目指す」ことが科目の共通目標となっている。筆者の授業では，この方針に沿いながら，(1) English as a Lingua Franca（ELF）および Lingua Franca Core（LFC）という理念の理解，(2) 現代世界の諸問題に対する批判的思考力の育成（石川，2014a），(3) 世界の多様な英語に対する聴解力の涵養，(4) Lingua Franca Core に基づく音素の発音に関する知識および技能の拡大, (5) 英語による発信力（ライティング，スピーキング）の涵養，(6) 対人関係力の向上，(7) 他者との協働作業力の向上，(8) 自律的学習態度の育成，の 8 点を具体的な目標としている。

授業は，モジュール制を取っており，Ⅰ：事前指定課題に基づくニュース英文のディクテーション，Ⅱ-a：課題スピーチ，Ⅱ-b 即興会話，Ⅲ：主要音素の発音に関わる講義と演習，Ⅳ：グループ単位での楽曲のディクテーション，Ⅴ：グループ単位での楽曲の背景に関する英文の読解というフローで全体が構成されている。

Ⅰでは，たとえば,「中東圏における女性の抑圧」「中国における割り箸使用禁止」「Wikipedia の普及と問題点」「カカオ価格の高騰と先物取引の問題点」など，世界の時事的な問題を報じるニュース素材（1 分程度）が使用される。使用する音声ファイルは事前に学習者に配布され, 学習者は自宅で何度も聞き取りを行い, それをふまえて授業内でディクテーションテストに回答する（各回 10 点満点）。解説においては，言語面よりも，むしろ, 時事問題の背景について十分な時間をかけて指導を行うよう留意している。たとえば,「中東圏における女性の抑圧」であれば, 世界における女性解放の歴史, 女性参政権の問題, イスラム圏における宗教法（シャリア）の理念，ウーマンリブ運動の勃興と展開などについて，体系的な解説が加えられる。Ⅰで実施されるディクテーションテストは，狭義には聴解力を問うものであるが，時事問題についての一定の知識が必要であり，事前の準備やクラスメートとの相談も可能であることから，現代社会への問題意識，対人関係力，自律的学習態度なども同時に測っていると推定される。また，前後関係をヒントとして活用できるかどうか，といった点で方略的能力とも関係している。

Ⅱ-a では，事前に与えられた課題について，受講生があらかじめ原稿を執筆し，それを使ってパートナーにスピーチを行うことが求められる。スピーチ課題はⅠと連動した身近なテーマを事前に指示しており，たとえば,「中東圏における女性の抑圧」については「JR

の女性専用車は是か非か」,「中国における割り箸使用禁止」については「マイ割り箸運動は是か非か」,「Wikipedia の普及と問題点」であれば「大学生と Wikipedia」など, ディクテーションテストで扱った問題を自分の視点で見つめ直すものとなっている。Ⅱ -a に関しては, 毎回，各自が作成した原稿の提出が求められる。原稿は外形的観点（課題を正しく扱っているか，規定の分量を満たしているか）に基づいて 10 点満点で評価される。あくまでもスピーチの下準備がきちんとできているかどうかの確認を目的としているので，内容や構成の良し悪しは評価観点に入っていない。このタスクも，直接的には（スピーキングの準備としての）ライティング力を見ているものであるが, 同時に, 現代社会への問題意識や, 自律的学習態度などにもかかわるものである。

Ⅱ -b では，学習者は立ち上がって教室を自由に移動し，教授者がその場で示す 3 つのトピック（たとえば, Ⅱ -a で「JR の女性専用車の是か非か」を扱った場合は,「列車」「飛行機」「通学」など）についてそれぞれ 1 分間の即興会話を行うことが求められる。会話相手はその都度変えることがルールとなっている。ここでは，周囲のクラスメートの中から毎回違う相手を探し，自分で話しかけ，会話の糸口をつかまなければならない（時間が制限されているので，早く相手を見つけなければほとんど話せないうちに持ち時間が終わってしまう)。授業者としてこのタスクに求めているのは, スピーキング力や対人関係力のトレーニングである。なお，Ⅱ -b について評価タスクは設定していない。

Ⅲでは，Seidlhofer（2011）の言う English as a Lingua Franca モデル（非母語話者同士のコミュニケーション手段として英語を位置付ける主張）や，Jenkins（2007）らが言う Lingua Franca Core モデル（国際的な英語使用場面において通用性（intelligibility）に大きくかかわる発音上の問題の大半は子音に限定されるという主張）に基づき，毎回，ターゲット子音を設定し，音声学的に解説を加えた上で，発音テストを行う。発音テストの素材は，当該子音を含む語・句・短文である。たとえば，破裂音として [p] を扱った場合は “put” という語を発音させ，十分な破裂が起こっているかどうか確認する。摩擦音として [s] や [ʃ] を扱った場合は “seat/ sheet” を連続して発話させ，2 つの子音を区別して発音できているかどうかを確認する。また，破擦音 [tʃ] を扱った場合は，“Check it out” などの短文を読ませ，破擦音の発音が正しくできているかどうかを確認する。発音テストは授業内で受講生全員にその場で行わせ，授業者が 5 点満点で採点する。発音テストは，音素の発音に関する学習者の知識・技能を測定するものであると同時に，教室環境の中で全員に聞こえるように適切な音量で発音を行わなければならないという点で，学習意欲や対人配慮態度なども測っていると考えられる。

ⅣおよびⅤでは，英語圏文化の一側面として音楽を取り上げ，人口に膾炙したポップス音楽を素材として，4 人グループで，歌詞の聞き取りに基づく空欄補充（10 点）と，楽曲に関係する英文の読み取りに基づく空欄補充（5 点）を行う。学習者はグループで合議して解答をまとめて提出する。従ってグループワークテストの得点はグループ構成員全員が

同一である。歌詞の聞き取りは聴解力を，英文の空所補充は読解力や文法力を問うものであるが，これらはグループディスカッションによる協働作業となるため，学習者の対人関係構築力や協働力も同時に測定されていると考えられる。実際，すぐれたグループでは誰かが自然にリーダーシップを取って，構成員全員に聞き取った内容を述べさせてディスカッションを主導し，意見が異なる場合は様々な観点を示して合議によって答案内容をまとめていくような姿が認められる。また，文脈的ヒントを活用するという点で方略能力も関係している。なお, 現段階では, 歌詞の聞き取りと読解を合わせて 15 点で評価しており，両者は区別していない。

上記のⅠ～Ⅴのモジュールからなる通常授業を行う回のほか，全学実施される外部試験に先立って TOEIC や TOEFL の模擬テストを行う回，また，期末にはリスニング総括実力テスト（授業内容に直結しない多様なリスニング問題によって標準的な英語マテリアルに対するリスニング力を総合的に問う）を行う回や，授業内容総括テスト（授業で扱った各種の英文や，その文化的・社会的背景等について，事前に範囲を示して準備させた上で，当該知識の定着を問う）を行う回がある。前者の総括テストには聴解力や方略能力が，後者の総括テストには時事問題に対する関心や，既習内容を理解・記憶する能力，また，こうした活動を継続的に行う自律的学習態度などが関係する。

以上の授業過程で得られた 7 種の形成的評価指標をまとめたのが表 1 である。

表 1　形成的評価指標

指標	作業	評点	回数	総点	関連する資質・能力
DT：ディクテーションテスト	個人	10	5	50	聴解力，自律的学習態度，（対人関係構築力），方略的能力など
SS：課題スピーチシート	個人	10	5	50	作文力，自律的学習態度，方略的能力など
PR：子音発音テスト	個人	5	4	20	発話力，自律的学習態度など
GW：グループワークテスト	集団	15	4	60	対人関係構築力，協働力，聴解力，読解力，語彙力など
MT：模擬テスト	個人	50	1	50	聴解力，方略的能力など
LT：リスニング総括実力テスト	個人	100	1	100	聴解力，方略的能力など
ST：授業内容総括テスト	個人	100	1	100	自律的学習態度，社会意識，批判的思考力，記憶力など

3.3 対象者

対象学生はすべて 1 年生で，専攻は，社会系，国際系，人文融合系，医療系，工学系の 5 種（5 クラス）である。また，授業内で収集した形成的評価指標に加え，これらの学生がクオータ期間内に受験した TOEIC または TOEFL-PBT のスコアデータをあわせて分析に利用する。分析にあたっては，通常授業において欠席のある学習者のデータは分析対象から除外している。

3.4 手法

まず，RQ1（指標の弁別性・関係性）については，標準偏差を平均で割った変動係数を手掛かりに各指標による学習者の弁別力を調べ，ついで，ピアソンの積率相関係数によって指標間相関を見る。その後，変数クラスター分析（1 次距離は（2-2r）の平方根で定義，クラスター間距離は Ward 法で計算）を実施し，7 指標がどのように分類されるか確認する。

次に，RQ2（外部指標との相関）については，7 指標それぞれと外部試験指標（TOEIC のリスニングスコア，同リーディングスコア，TOEFL のリスニングスコア，同文法スコア，同リーディングスコア）との相関を調べる。

RQ3（指標合成）では，7 種の指標値を主成分分析で合成し，第 1 主成分における指標ごとの負荷量を比較する。

RQ4（異なる合成指標による最終評価の変化）では，(i) 7 指標の配点を保持したまま合計点を求め，配点総点で割って百分率化する方法（単純合算法），(ii) 7 つの指標をそれぞれ得点率に換算した上で単純平均する方法（得点率平均法），(iii) 7 指標を合成した第 1 主成分得点を総合得点とみなす方法（主成分法），(iv) 7 種の指標のうち，リスニング総括実力テスト・授業内容総括テストを除く 5 種の指標値について配点を保持したまま合計点を求め，配点総点で割って百分率化したものに，2 種の総括テストの良いほうのスコアを加算する方法（単純合算＋総括テストスコア法），(v) 5 種の指標を得点率に換算した上で単純平均したものに，2 種の総括テストの良いほうのスコアを加算する方法（得点率平均＋総括テストスコア法）という 5 種を比較し，個々の学生ごとに，すべての指標に欠損がない 113 名中での順位がどう変化するかシミュレーションを行う。比較に当たっては，5 つの合成指標による順位値の最大値から最小値を引いた値を計算し，これを変動幅と定義する。

4. 結果と考察

4.1 RQ1 指標の弁別性・関係性

7 種類の形成的評価指標の概要は表 2 の通りである。N は人数，Av は平均，% は得点率，SD は標準偏差，CV は変動係数を示す。

表 2 指標ごとの基礎統計量

	DT	SS	PR	GW	MT	LT	ST
N	160	160	160	160	160	158	115
Av	31.63	49.22	15.08	49.13	30.59	63.09	75.43
%	63.26	98.44	75.41	81.88	61.18	63.09	75.43
SD	7.83	3.00	2.07	4.58	5.92	12.49	14.63
CV	0.25	0.06	0.14	0.09	0.19	0.20	0.19

変動係数に注目すると，7 種の形成的評価指標のうち，学習者間のばらつきが大きいものはディクテーションテスト (0.25)，リスニング総括実力テスト（0.20），模擬テストおよび授業内容総括テスト（0.19）などであった。一方, ばらつきが小さいものは課題スピーチシート（0.06），グループワーク（0.09）などであった。学習者の能力を弁別して評価するという目的に限って言えば，後者の寄与は限定的であることが示された。ただし，テストを通して学習者に必要な学習を行わせるという点では後者にも教育的意義が認められる。

次に，7 種のデータ間の相関を調べたところ，表 3 の結果を得た。なお，欠損データについてはペアワイズで除去を行っている。分析により，模擬テストとリスニング総括実力テスト（0.709），ディクテーションテストと模擬テスト（0.642），ディクテーションテストとリスニング総括実力テスト（0.619）の相関が相対的に強く，このほか，グループワークテストと模擬テストの間にも中程度の相関があること（0.46），一方，発音テストとその他の指標（-0.031 ～ 0.222），スピーチシートとその他の指標（-0.065 ～ 0.190）の間はほぼ無相関となることがわかった。なお，ここで興味深いのは，個人課題であった模擬テストとグループワークテストの間に中程度の相関が出ていることである。このことは, グループワークテストが協働力や対人関係構築力だけでなく，構成員個々の聴解力にも影響されている可能性を示唆する。

表 3 指標間の相関

	DT	SS	PR	GW	MT	LT	ST
DT	1.000						
SS	0.182	1.000					
PR	0.132	0.081	1.000				
GW	0.423	0.190	0.222	1.000			
MT	0.642	0.122	0.133	0.486	1.000		
LT	0.619	0.125	0.119	0.481	0.709	1.000	
ST	0.219	-0.065	-0.031	0.002	0.011	0.001	1.000

次に，7 指標についてクラスター分析を実施したところ，図 2 の結果を得た。

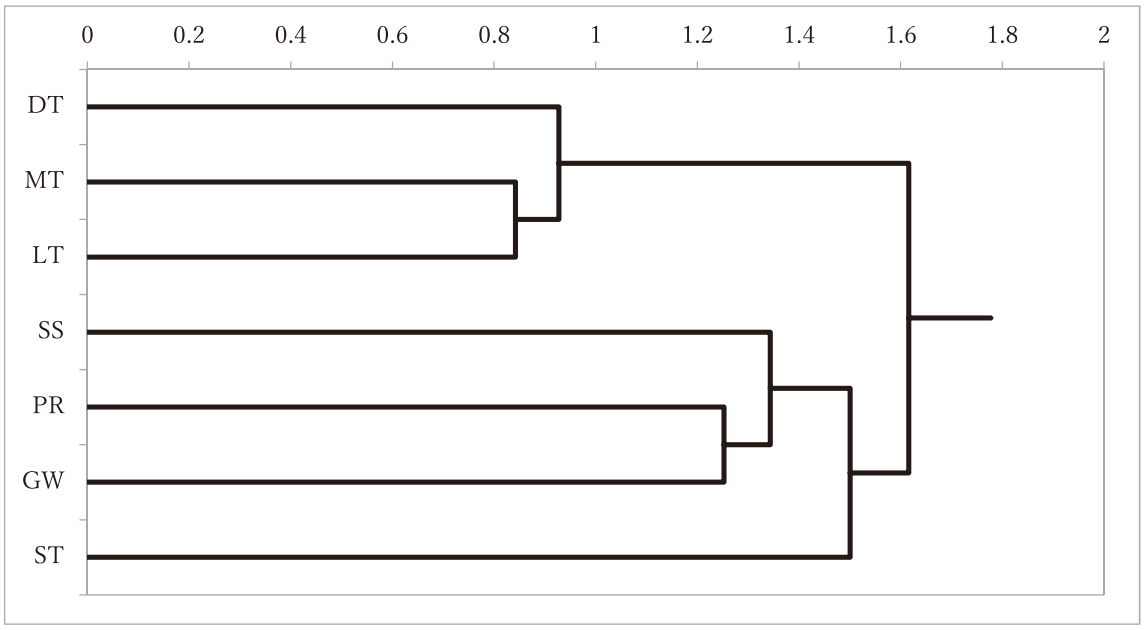

図 2 指標の分類（クラスター分析に基づく樹形図）

距離 1.4 あたりにカッティングポイントを設定すると，授業過程で継続的に収集された 7 種の形成的評価指標は，第 1 クラスター（ディクテーションテスト，模擬テスト，リスニング総括実力テスト），第 2 クラスター（スピーチシート，発音テスト，グループワークテスト），第 3 クラスター（授業内容総括テスト）の 3 群に分かれることが示された。第 1 クラスターは学習者個人の聴解力を，第 2 クラスターは学習を管理する自律的学習態度，積極的にコミュニケーションを行おうとする態度，他者と積極的に関わり，協働しようとする姿勢といったコミュニケーション言語能力の非言語的側面を，第 3 クラスターは学習に真摯に取り組み，必要な事項を記憶する能力を表していると考えられる。ここで注目すべきは，多様な英語力がおよそ 3 つに区分されることと，従来の大学英語教育で多用されてきた授業内容総括テストで測られる力が他の能力と必ずしも一致していないということである。

4.2 RQ2 外部指標との相関

すでに述べたように，近年の大学教育では，正課の英語授業の一部として，あるいは，それとは独立した形で，TOEIC や TOEFL といった何らかの外部試験を受験させることが増えている。こうした動きの背景には，大学での英語力評価に対してアカウンタビリティが求められるようになっているという現状がある。この場合，英語授業で形成的に評価された英語力と外部試験によって測定される英語力には一定の相関が存在することが期待されるが，一方で，あまりに高い相関が出ていると，大学英語教育としての独自の目標設定

がうまくなされていないことになる。

今回，分析対象とした学習者について言うと，学部ことに受験すべきテスト種別が決定されており，全体の 6 割弱が TOEIC を，4 割程度が TOEFL を受験している。それぞれの外部試験の成績概況は表 4 の通りである。

表 4 外部試験の成績データ

Sub	TOEIC_L	TOEIC_R	TOEIC	TOEFL_L	TOEFL_G	TOEFL_R	TOEFL
N	80	80	80	60	60	60	60
Av	281.06	266.25	547.31	49.25	51.82	50.75	510.73
%	56.78	53.79	55.28	72.43	76.20	75.75	75.44
SD	74.65	75.53	142.20	4.11	6.04	5.57	42.81
CV	0.27	0.28	0.26	0.08	0.12	0.11	0.08

変動係数に注目すると，TOEIC が学習者の能力差を相対的に細かく弁別しているのに対し（0.26 ～ 0.28），TOEFL の弁別力は低い（0.08 ～ 0.12）。このデータを用い，形成的評価指標との相関係数を調べたところ，表 5 の結果を得た。指標ごとの相関平均値に注目すると，外部試験の各スコアは，模擬テスト（0.567），リスニング総括実力テスト（0.515），ディクテーションテスト（0.493）と中程度の相関を示し，グループワークテスト（0.347）および授業内容総括テスト（0.212）と弱い相関を示すのに対し，スピーチシート（0.117）および発音テスト（0.114）とはほぼ無相関であった。以上の平均は 0.338 となり，相関の平方で求められる説明力は 13.22%となる。

表 5 指標と外部試験の相関

	DT	SS	PR	GW	MT	LT	ST
TOEIC_L	0.536	0.113	0.209	0.356	0.666	0.664	0.186
TOEIC_R	0.507	0.127	0.264	0.386	0.727	0.633	0.087
TOEIC	0.551	0.127	0.250	0.392	0.736	0.685	0.145
TOEFL_L	0.480	0.140	0.002	0.397	0.547	0.614	0.208
TOEFL_G	0.432	0.053	-0.021	0.202	0.341	0.178	0.274
TOEFL_R	0.282	0.119	0.077	0.200	0.264	0.209	0.278
TOEFL	0.663	0.141	0.014	0.495	0.686	0.625	0.308
Av	0.493	0.117	0.114	0.347	0.567	0.515	0.212

つまり，英語力の多面性を意識し，異なる側面に焦点を当てて収集した形成的評価データは，外部試験で測定される英語力と一定の相関を持ちつつも，それだけに限定されず，授業目標に即した授業独自の要素が加味された評価となっていることを示す。

なお，大学英語教育では外部試験の成績を一定程度（たとえば 50%）組み込んで成績評価を行わせている例も存在するが，今回のデータに限って言えば，両者の相関性は限定的で，安易な合成には問題が多いと言うべきであろう。

4.3 RQ3 指標合成

複数の形成的評価指標を合成する場合，一般的には平均化が行われるが，すでに見たように，個々の指標値は異なる性質を持っている。これらの全体をよりよく代表するよう指標の合成を行おうとする場合，主成分分析の手法を用いることができる。

主成分分析の結果，固有値 1 を超えるものとして，第 1 主成分（寄与率 35.61%）と第 2 主成分（17.20%）が得られた。このうち，第 1 主成分はすべての指標の負荷量がプラスとなっており，全体の総合得点となっていることがわかる。

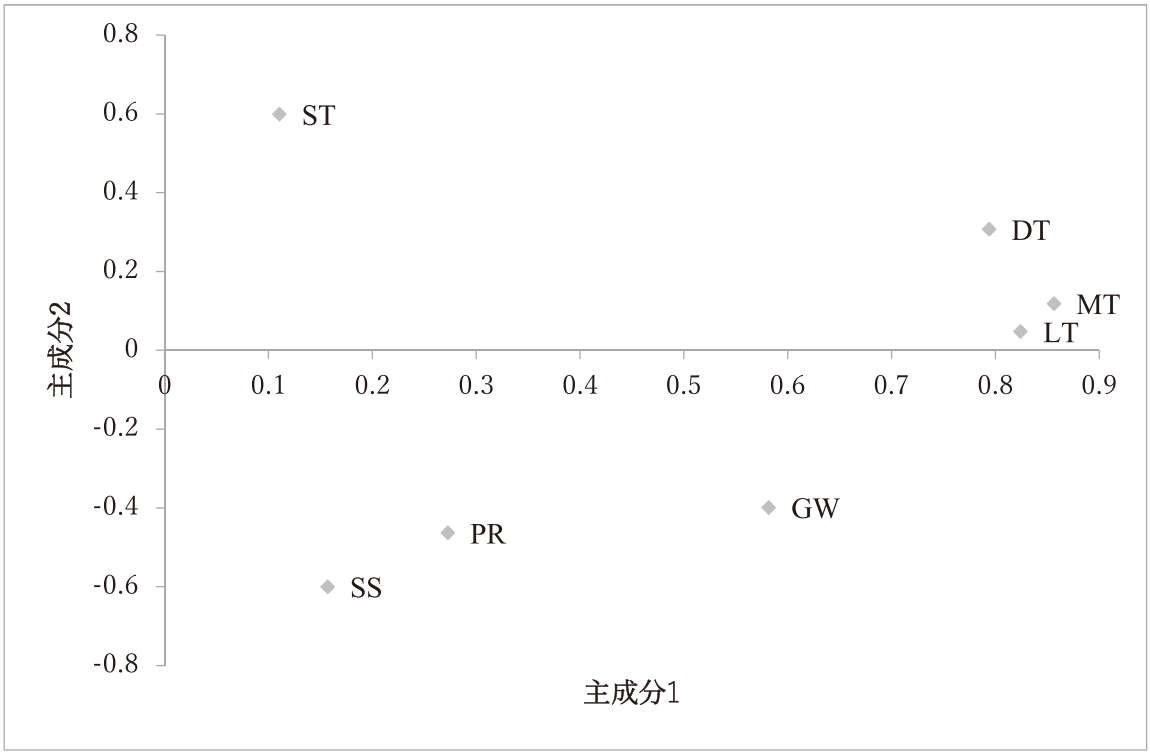

図 3 合成指標と各指標の関係（第 1 ・第 2 主成分に基づく散布図）

第 1 主成分負荷量の降順で 7 種の指標を整理したところ，表 6 のようになった。

表 6 主成分負荷量（降順）

変　数	第 1 主成分負荷量
MT	0.856
LT	0.824
DT	0.794
GW	0.582
PR	0.273
SS	0.157
ST	0.110

7 指標の中では，模擬テスト（0.856），リスニング総括実力テスト（0.824），ディクテーションテスト（0.794）の負荷量が高く，グループワークテストがそれらに次ぎ（0.582），発音テスト（0.273），スピーチシート（0.157），授業内容総括テスト（0.110）の負荷量が低いことがわかった。ここで興味を引くのは，授業内容総括テストの負荷量の低さである。既習内容の中から範囲を定め，事前準備の上で受験させる総括的試験は大学において古くから行われてきたわけだが，そうしたテストで測られる能力は，記憶力のような一般的認知能力，ないしは，まじめさといった態度要因であって，前述のように，必ずしも一般的な意味での英語力が測れているとは限らない。

4.4 RQ4 異なる合成指標による最終評価の変化

この授業実践では，7 種の形成的評価指標に基づいてデータを収集したわけであるが，これらを均等に扱って合成する方法として，(i) 単純合算法，(ii) 得点率平均法，(iii) 主成分法の 3 つを，また，通常の授業の中で取得された 5 種の指標（ディクテーション，スピーチシート，発音テスト，グループワークテスト，模擬テスト）と授業期間の最後に行う総括テスト（リスニング総括実力テスト，授業内容総括テスト）指標を別扱いにして合成する方法として，(iv) 単純合算＋総括テストスコア法，(v) 5 得点率平均法＋総括テストスコア法の 2 つを比較検討した。なお，(iv) と (v) に関して，総括テスト成績の加算にはいくつかの方法が考えられるが，ここでは，実際の授業で採用した方法（2 つの総括試験を受験させて良いほうの成績のみを評価に使用する）に準じる。

以上，5 種類の合成指標を用いて最終成績のシミュレーションを行った結果は，表 7 の通りである。なお，順位はすべての指標が揃っている 113 名中での位置を示し，変動は 5 つの合成指標による順位値の最大値から最小値を引いた値を表す。表 7 はシミュレーション結果の一部として，10 名分の学生データを示したものである。

表 7 異なる合成指標を採用した場合の全体順位の変動（10 名分）

	スコア					順位（113 名中）					
	i	ii	iii	iv	v	i	ii	iii	iv	v	変動
X_01	73.3	74.2	0.2	159.0	158.3	38	38	43	37	38	6
X_02	68.6	69.7	-0.2	136.6	135.9	71	75	57	96	97	40
X_03	77.4	76.2	0.7	172.7	170.3	21	26	33	6	7	27
X_04	68.4	69.7	-1.3	149.6	150.7	77	75	88	71	65	23
X_06	84.0	82.4	3.4	171.7	168.5	6	9	3	7	10	7
X_07	72.6	75.1	0.5	150.5	151.9	42	32	36	65	58	33
X_10	78.1	80.6	2.1	163.5	164.1	16	11	11	27	22	16
X_11	64.4	68.0	-1.6	154.3	154.1	100	88	100	52	52	48
X_12	70.7	70.2	-0.6	157.0	155.1	56	71	71	44	50	27
X_13	69.8	71.2	-0.3	148.8	148.1	62	62	60	73	79	19

サンプル学生について言うと，最小で 6 位分，最大で 48 位分の順位変動が生じている。このうち, X_03 や X_11 は（iv）と（v）の合成指標において順位が大幅に上昇しているが，これは範囲を予告して行った授業内容総括テストにおいてきわめて高い成績を収めたためである。また，113 名全体について言うと，変動幅（順位数）の平均値は 27.15（SD=17.04）であった。同じ形成的評価データを使っていても，最終的な合成指標の選択によって，想像以上に大きな順位変動が生じることが確認された。

図 4 は，5 種の合成指標によるクラスごとの順位平均値を示したものである。縦軸は順位値を示し, 値が小さいほうが全体の中での上位となる。図に示される通り, 自然科学系（医療系・工学系）のクラスでは，（i）～（iii），つまり，単純合算法，得点率平均法，主成分法の順で平均順位が上がるのに対し，人文系（国際系・人文融合系）のクラスでは逆に平均順位が下がることが確認された。主成分において，模擬テスト・リスニング総括実力テスト・ディクテーションテストの負荷量が高く，発音テストやグループワークテストの負荷量が低くなっていたことをふまえれば，自然科学系の学習者の英語力は個人レベルでの正確な聴解に強みを持ち，一方，人文系の学習者の英語力はコミュニケーションや協働作業に強みを持つと言える。また，総括テストを別枠として処理した場合は，人文融合系で順位が大きく上昇するが，自然科学系では順位が総じて下がっている。授業では，最初にリスニング総括実力テストを行って結果をその場で告知し，その後，範囲を指定して準備させ，授業内容総括テストを行ったわけだが，人文融合系の学習者の多くがより良い評価を求め，十分な準備の上でテストに臨んだのに対し，自然科学系の学習者は，最初のテストで最低合格基準に達すると，それ以上の評価を望まず，必ずしも十分な準備を行わずに

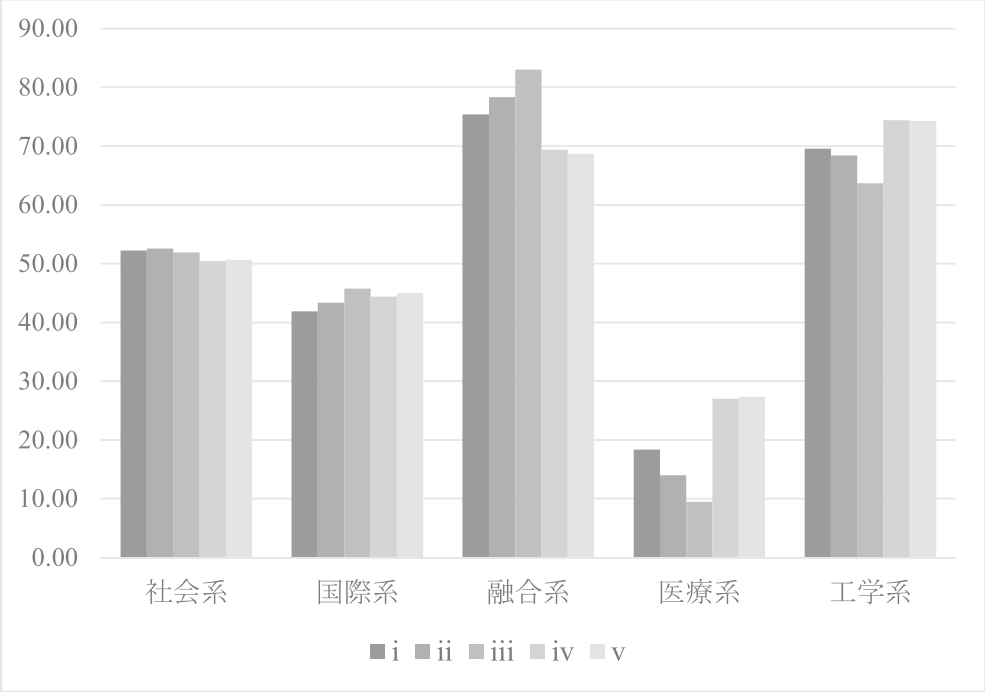

図 4 異なる合成指標を選択した場合のクラス別の平均順位

テストを受験したものと考えられる。

このように，合成指標の選択は学習者の評価順位に大きな影響を及ぼすわけであるが，一方，社会系の学生については，どの手法をとっても，平均順位はほぼ同等で変わることがなかった。合成指標の選択が，英語コミュニケーション能力の多面性の中でどの部分を優先的に評価するかにつながっているとすれば，社会系の学習者の英語力は，他と比べて，安定性の高い（どの側面でも優劣の差が小さい）ものである。これらのシミュレーション結果は，大学英語教育における学習者の英語力の評価方法の検討において，学習者のタイプ要因の考慮の必要性を示唆していると言えるだろう。

5 まとめ

以上，本研究では，英語力の多面性について概観した後，大学英語教育での形成的評価導入の可能性について検討してきた。実際の授業実践で得られた 7 種の形成的評価指標を基に分析を行った結果，4 つのリサーチクエスチョンについて以下の知見を得た。

まず, RQ1（指標の弁別性･関係性）については, 7 種の形成的評価指標のうち, ディクテーションテストやリスニング総括実力テストにおいて学習者の能力が細かく弁別され，課題スピーチシートやグループワークでは弁別性が低いことが確認された。また，指標間の相関については，模擬テスト，リスニング総括実力テスト，ディクテーションテスト間の相関が高い一方，発音テストやスピーチシートはその他の指標とほぼ無相関であり，異なる

能力を評価している可能性が示された。さらに，クラスター分析により，7 種の指標は，個人単位での聴解力に関係する第 1 クラスター（ディクテーションテスト，模擬テスト，リスニング総括実力テスト），コミュニケーション志向や学習態度に関わる第 2 クラスター（スピーチシート，発音テスト，グループワークテスト），勤勉さや記憶力に関わる第 3 クラスター（授業内容総括テスト）の 3 群に分かれることが確認され，伝統的に大学で行われてきた授業内容総括テストは一般的な英語力とは異質な能力を測っている可能性が示唆された。

RQ2（外部指標との相関）については，模擬テスト・リスニング総括実力テスト・ディクテーションテストは外部試験と中程度の相関を示した。一方，スピーチシートや発音テストは外部試験とほぼ無相関であった。7 種の指標値の外部試験との相関の平均は 0.338 となり，授業実践で得られた 7 種の評価データが持つ情報の分散のうち，約 13%が外部試験で測定される一般的な英語力要素で説明されることがわかった。

RQ3（指標合成）については，主成分分析で得られる第 1 主成分に 7 指標すべてを合成することができた。このとき，主成分負荷量に注目すると，7 指標中，模擬テスト，リスニング総括実力テスト，ディクテーションテストの重みが高く，スピーチシート，授業内容総括テストの重みが低くなることがわかった。

最後に，RQ4（異なる合成指標による最終評価の変化）に関しては，5 種類の合成指標に基づく全体順位を比較した結果，異なる合成指標を選択することで，113 名中での順位値が平均で 27 位ほど変動することが確認された。また，人文系学習者と自然科学系学習者では，コミュニケーション英語力に包含される多面的要素の中で得意とする分野に一定の違いがあり，合成指標の選択によって，有利・不利の差が生じる可能性が示唆された。

本研究の結果は，(1) 試験範囲を明示して事前準備の上で受験させる伝統的な総括評価（期末テスト）によって測定される英語力が「コミュニケーション英語力」としてかなり異質なものであること，(2) 評価と指導のプロセスを融合し，各種の評価データを形成的に収集することで，学習者の「コミュニケーション英語力」を多面的に測定しうること，(3) 一方で多様な形成的評価データを統合して最終評価を行う場合は，指標を合成する手法の選択が最終評価に影響を及ぼしうること，などを実証的に明らかにした点で，今後の大学英語教育における評価の問題を考える際に，一定の意義を持つものと言える。

もっとも，今回の実践で取り入れた 7 つの指標だけで英語力の多面性のすべてがカバーされているわけではない。今後は，スピーチやライティングといったパフォーマンスの直接評価を加えるなど，新たな形成的評価指標の開発を行い，評価の妥当性とその教育的意義を高める工夫を継続的に行っていきたい。

参考文献

Bachman, L. F. (1990) *Fundamental considerations in language testing*. New York, NY: Oxford

University Press.

Bachman, L. F., & Palmer, A. S. (1996) *Language testing in practice*. New York, NY: Oxford University Press.

Black, P., & Wiliam, D. (1990) *Inside the black box: Raising standards through classroom assessment*. London, UK: GL Assessment Limited.

Black, P., & Wiliam, D. (1998) Assessment and classroom learning. *Assessment in Education: Principles, Policy and Practice*, *5*(1), 7-74.

石川慎一郎 (2014a)「英語教育における異文化理解教育の課題と展望：文化の定義の再考と異文化理解教育の発達段階別多層モデルの提案」『東京外国語大学 世界言語社会教育センター国際シンポジム 2013 報告書』105-119.

石川慎一郎 (2014b)「観点別評価から始まる目標志向型教育」神戸市高等学校教育課程研究協議会（2014 年 8 月 1 日）講演資料 .

石川慎一郎 (2015)「観点別評価と can-do 評価：指導と評価の科学」神戸市立葺合高等学校 SGH 教員研修会（2015 年 3 月 19 日）講演資料 .

石川慎一郎 (2016)「アクティブラーニングの二重性：英語教育への示唆」『チャートネットワーク』（数研出版）*79*, 1-4.

Jenkins, J. (2000) *The phonology of English as an international language*. Oxford, UK: Oxford University Press.

Jenkins, J. (2007) *English as a lingua franca: Attitude and identity*. Oxford, UK: Oxford University Press.

文部科学省 (2009)『高等学校学習指導要領』東山書房 .

文部科学省 (2010)「児童生徒の学習評価の在り方について（報告）」Retrieved from http://www.mext.go.jp/b_menu/shingi/chukyo/chukyo3/004/gaiyou/attach/1292216.htm

文部科学省 (2013)「各中・高等学校の外国語教育における『CAN-DO リスト』の形での学習到達目標設定のための手引き」Retrieved from http://www.mext.go.jp/a_menu/kokusai/gaikokugo/__icsFiles/afieldfile/2013/05/08/1332306_4.pdf

OECD. (2001) *Knowledge and skills for life: First results from PISA 2000*. Paris, France: OECD.

OECD. (2004) *Learning for tomorrow's world : First results from PISA 2003*. Paris, France: OECD.

OECD. (2005a) *Policy Brief − Formative assessment: Improving learning in secondary classrooms*. Paris, France: OECD.

OECD. (2005b) *Formative assessment: Improving learning in secondary classrooms*. Summary in Japanese. Paris, France: OECD.

Seidlhofer, B. (2011) *Understanding English as a lingua franca*. Oxford, UK: Oxford University Press.

Stiggins, R. (2007a) Assessment through the student's eye. *Educational Leadership*, *64*(8), 22-26.

Stiggins, R. (2007b) Five assessment myths and their consequences. *Education Week*, *27*(8), 28-29.

Walker, R. (2010) *Teaching the pronunciation of English as a lingua franca*. Oxford, UK: Oxford University Press.