英語教育カリキュラムの中の評価、その方法と問題点

竹蓋　幸生・与那覇　信恵

［**要旨**］文献調査、大量のテストスコアの分析的観察、出張視察、等をとおして、効果の期待できる英語教育の評価はいかにあるべきかについて考察する実験研究を行った。全国的に採用されている外部テストにも種々の問題点が存在することが明らかにされたが、テストスコアの変動に影響を与える多くの因子の存在を正しく認識して科学的な観察の姿勢を崩さずに使用すれば、外部テストは英語教育の改善に有効に活用できる可能性のあることが判明した。

1．はじめに

本論文は、文京学院大学総合研究所の平成 17-18 年度共同研究の経費助成に採択されて行われた研究、「仕事で使える英語力の養成を目指すカリキュラムの編成に関する研究」のうち、中間報告（竹蓋、与那覇 2006a）と最終報告（竹蓋、与那覇 2007 予定）に含めることができなかった、英語教育カリキュラムの中の「評価」に関する研究のまとめである。

2．研究の目的

多くの大学現場教員の発言に TOEFL や TOEIC は英語コミュニケーションの総合力（とくに発信力）をテストしていない、TOEFL や TOEIC の点が高くても英語があまり使えない者が少なくない、などの批判がある。Childs（1995, 2002）や Gilfert（1995）もそのスコアの安定性やスコアを使った比較に問題があり得る等の指摘をしている。そのような中、我々は、自分たちの身近にある外部テストのデータを種々の方向から分析的に観察することにより指摘された問題点の「実態」を観察することにした。ただ、我々は妥当性、信頼性、実用性がすべて十分に高い発信力の評価法が存在するという確証はまだないと考えている。したがって、そのような面からの妥当性の検証は別の機会にゆずることとして、テストスコアの「揺れの実態」、「スコアの意味」を客観的データで観察し、外部テストの使用にあたって留意すべき事項の一部を明らかにすることとした。

3．研究の方法

研究の方法としては、まず、評価に関する文献、資料をできるだけ多く収集し、他の研究者が外国語教育関連の評価に関してどのような指摘をしているかを概観する。その後、自分たちの身近にあるデータ、または学会誌等に公刊されたデータ等の分析的観察、相対的比較等により、指摘されている内容の実態を観察する。

4．文献の調査

まず、Richards らは、評価について、*Longman Dictionary of Applied Linguistics* の中で次のように定義している。

> EVALUATION:　… In language planning, evaluation frequently involves gathering information on patterns of language use, language ability, and attitudes towards language. In language teaching programmes, evaluation is related to decisions to be made about the quality of the programme itself, and decisions about individuals in the programmes. The evaluation of programmes may involve the study of curriculum, objectives, materials, and tests or grading systems. The evaluation of individuals involves decisions about entrance to programmes, placement, progress, and achievement. In evaluating both programmes and individuals, tests and other measures are frequently used. （Richards, *et al*.1985 ： 98）

また、文部科学省（2003）がその期待を表明したものとしては、とくに大学に求められている部分として、以下のような指摘がある：

> 日本人全体として、英検、TOEFL、TOEIC 等客観的指標に基づいて世界平均水準の英語力を目指すことが重要である。

英語学習の目標として、客観的指標に基づいて世界平均水準の英語力（たとえば TOEFL-PBT 536、竹蓋、与那覇 2006b ： 33）を目指すべきだということは、教育効果の評価においてもそのことを無視すべきでないということであろう。しかしながら、一口に評価といってもその目的はひとつではない。

テストの目的： Richards らや文部科学省などの指摘とは別に、竹蓋（1984 ： 158-162）にはテストに少なくとも 3 種の目的があると述べられている。それは、1）基本的な能力の有無を判定する「選別テスト」による評価、2）指導したことがどのくらい学習されたかの判定をする「学力テスト」、それに、3）どこが悪いのかを分析的に明らかにする「診断テスト」、などと呼ばれるテストのそれぞれで行う評価である。カリキュラムの中での評価の位置づけ、その役割を考えるにはこれらのすべてを適切に活用して指導を最適なものにする必要があるということである。

次に、テストをその形式的内容から分類すると、まず、リスニング、スピーキング、リーディング、ライティングのような 4 技能の力を分析的に評価するもの、さらに細かく、音声識別力テスト、語彙力テスト、文法力テストもあるが、総合力テストと称するものもある。わが国でも多くの受験者のある TOEIC は、最近まで listening section と reading section の 2 セクションのスコア、また、TOEFL は listening section と reading section のスコアに structure and written expression の 3 セクションのスコアを加えて total score をだし、総合力の評価できるテストであると称してきた。

一方、Caulfield 他は竹蓋（1984: 240-241）に引用されているような分析的、または総合的に行われたテストのスコア間の相関関係を示しているが、表現力のテストを直接は行わずに総合力が評価できると TOEIC、TOEFL が称して来たのはこのようなデータがあったからかもしれない。

テストの方法：テストの技術的な面からみると、理解力のテストにはミニマルペア（音素）の識別力テスト、語彙力テスト、文法力テスト、ディクテーション、内容理解テスト（自由筆記、多肢選択式）等がある。また、表現力テストには発音のテスト、スピーキングテスト、インタビューテスト、作文テスト（主に文単位）、パラグラフ・ライテイング、自由英作文などがあるが、Caulfield 他（竹蓋 1984 ： 240-241）は、表面的には形式のテストに見えるクローズテストやノイズテストでもある程度まで総合力を評価できると述べている。

テストの評価：人事の査定や研究論文の審査等でもしばしば問題になることに、評価や審査が適切かどうかがある。教育効果の評価においてもそれは第一に問題にされなくてはならない。したがって、テスト自体も少なくとも以下の 3 方向の観点から評価されることが多い。それは、妥当性、信頼性、実用性である（竹蓋 1984 ： 162-168）。妥当性とはそのテストが評価すべき内容を正しく評価しているかどうかの指標であり、信頼性とはそのテストの評価が安定しているかどうかの指標、そして、実用性とは経済性や必要な時間数、実施の困難さ、等の使い易さの指標である。これらの指標のすべてが完璧なテストはもちろんまだ存在しないが、できるだけ指標の高いテストを選ばないと学習者のやる気を削ぐことになるので注意が肝要である。

テストの制作：テストは、また、個々の教員が自作するものと外部機関が制作するものにも分けられる。前者は一般に手軽に、また、目的を絞って作れるので実用性に富んでいる場合が多いが、妥当性、信頼性が低い場合があると言われる。後者は妥当性、信頼性は比較的高いが、実施に時間や多額の費用がかかり、実用性が低いと言われることがある。外部テストの TOEIC や TOEFL が最近まで発信力のテストを含めないで総合力のテストであると称して来たことの背景の少なくとも一部は発信力のテストが、行われたとしても、信頼性、実用性が低いと見られてきたからだと思われる（竹蓋 1984 ： 240）。

従来型の受信力のみのテストでは、スコアが高くても英語がほとんど出来ない者がいるとの指摘が高まり、最近は TOEFL も TOEIC も発信力のテストを含めるようになったようであるが、その部分のテストの信頼性、実用性が高くなったかどうかは明らかでない。むしろ筆者

（竹蓋）は、真の英語力養成の学習をして来ず、入試の受験対策同様、TOEIC や TOEFL のテスト対策の学習に焦点をあわせた勉強しかしてこなかった者が「スコアは高くても英語力が高くない」という実態を生み出しているのだと考える。しかしながら、聴解力や読解力（共に受信力と言われることが多い）はコミュニケーション能力の基礎であると考えれば、それを、テスト対策としてではなく、広範囲にわたり、適切な教材を使って十分に学び、その結果として TOEFL または TOEIC スコアが高く出た者は、それがリスニングとリーディングセクションのみのものであっても、少なくとも基礎力は高いと言えるはずである。なお、Woodford（1982 ： 16）には以下のような説明がある。

> The Listening Comprehension part score of TOEIC correlates very highly with other measures of both listening and speaking. The TOEIC Reading part score correlates highly with other measures of candidates' abilities in both reading and writing.

スコアの解釈：一方、テストを受験して得られるスコアをどう解釈すべきかについて分析的に見た興味深い論文がある。それは、Childs（1995 ： 71-72）のものであるが、彼は TOEIC スコアの上昇量の中身（Causes of Score Gains）を分析して、① The formal teaching of English, ② Differences in motivation levels, ③ Heeding English in the environment, ④ Practice effect, ⑤ Temporary vs. long-term gains のような要素があると指摘している。テスト対策でスコアの上がった人で実力のない人とは、主に④の要因でスコアの数値が上がっており、①の要因で上がったのではない人ということなので、スコアが高くても、実力がないということになる。

テストの信頼性： TOEFL、TOEIC は、ともに、米国の Educational Testing Service（ETS）で作成されたテストである。テストの信頼性の指標のひとつに Standard Error of Measurement（SEM）があるが、たとえば、TOEIC の SEM は 34.93 であるという（Woodford1982 ： 8）。SEM とは統計的指標であり、それが 34.93 であるということは、テストで受験者の得た得点が「真の得点から 34.93 点以内にある確立が 68 %であること」を意味すると言われる。しかしながら、Childs （1995 ： 70）は自分たちでもこのテストの信頼性を観察し、SEM は 43 であったとしている。

この説明だけを読むと外部テストの TOEIC もあまり軽々しくそのスコアを信用はできないテストであるようにも見える。しかし、一方で、文部科学省は「英語が使える日本人育成のための戦略構想」のなかで「英検、TOEFL、TOEIC 等客観的指標に基づいて」と言い、多くの大学や大学院でも最近その英語力の評価に TOEFL または TOEIC のスコアを使い始めているようである。そして、たとえば以下のような説明が多くの大学でされている。

外部テスト採用の理由：

> TOEFL の受験は，主にアメリカ、カナダの大学や大学院に正式に留学するための必須条件となっていますので、TOEFL-ITP 受験は、皆さんが留学をめざす場合の自分の英語力を試すチャンスとなるでしょう。また、留学を考えていない人にとっても、アメリカでの日常生活や大学での学園生活が問題なく過ごせるぐらいの英語運用能力

は、社会に出てからの強い味方になってくれるでしょう！

（大阪大学言語文化研究科英語教育部会 2006）

広島大学では、学生の英語学力を客観的に把握するために TOEIC（R） IP テストを用いています。この試験を継続して行うことにより学生の習熟度を把握して、短期的には習熟度別クラス編成や習熟度に応じた成績評価を、長期的にはカリキュラムや教育方法の改善を行います。また、各自の英語学力を社会的・国際的に通用するスコアで確認することで、自主学習に役立てることもできます。また、スコアは成績表にも記載します。そのスコアは、学生にとって次のように役立ちます。

＊自分の力を、社会的、国際的に通用するスコアで知ることができる
＊社会的に認められたテスト結果で、就職などの自己 PR に使用できる
＊高スコアを得れば、教養教育の英語科目の単位として認定される

（広島大学 2006）

なぜ、TOEIC テストなのか？

目標や達成が、だれにも見える
コミュニケーション指向・必要性の自覚
授業改善
公平感
品質保証

（山口大学、宮崎 2006）より

外部テストの採用校の数：全国の大学で外部テストがどのくらいすでに採用されているかについては以下のようなデータが公表されている：

2005 年度 TOEIC テスト採用校　大学： 58 %（726 校中 422 校）

（TOEIC 運営委員会 2006、文部科学省 2005 より）

2004 年度 TOEFL テスト採用校　大学： 55 %（回答 570 校中）

（国際教育交換協議会 2004 より）

科学的評価：大変な労力をかけてデータを収集してもそこから必要な結論を正確に抽出できなくては無駄な骨折りになってしまう。もっとひどい場合には不正確な結論を信じたために大きな損害を蒙ることもある。そのような事態を避けるために科学的手法を採用する研究者が多いのであるが、収集したデータの科学的評価には以下のような条件を満たす必要があると Anderson（竹蓋 1982 ： 21）は指摘している： operational definition, generality, controlled observation, repeated observation, confirmation, and consistency.

文献調査のまとめ：一言で評価といっても目的、形式がいろいろあり、完璧な評価法も存在するとは言えない可能性が見えてきた。それにもかかわらず、社会的な要請として客観的評価

が求められており、その適切な活用が効果的な英語学習を助ける場合があることも事実のようである。そのような中で忘れてならないのが教員の良心と研究者の科学の目による真実の評価ではなかろうかというのが文献調査から得た我々の結論である。

5. 収集されたデータ・資料の分析

文京女子短期大学では 10 年以上にわたって継続的に TOEIC-IP を使用してきた実績があり、また 2001 年度に文京女子大学外国語学部が設置された後もその伝統が受け継がれて学生の全員が TOEIC-IP を受験してきた。そのデータを基に幾つかの貴重な提言ができる。

5.1　テストの難易度

難易度の揺れ：文京女子短期大学には 1994 年度から 2003 年度までの平均で毎年約 560 名の入学生があった。その学生が在学中にほぼ全員 2 回ずつ TOEIC-IP を受験していたのであるが、1993 年度から 2004 年度までの各年度の 1 回目のテストの平均点を年度毎に図示（折れ線グラフ）したものが（竹蓋他 2006 ： 33 ：図 1-2-7）に示してある。そのグラフからは右肩下がりになっている全体的傾向に加えて、局部的に必ずしも小さくない上下動が読み取れ、そのことから TOEIC-IP には難易度の揺れがあることが推定された。そのように考えた理由は、最近の大学や短大は偏差値で輪切りにされ、入学生のレベルがほぼ固定されており、大学や学部単位の学生の平均的な成績は年度毎に大きく変動することは少ないと言われているからである。

我々は、このようにして推定された TOEIC-IP の難易度の揺れがどの程度あるのかを調べるために上記の短大生の 12 年分のデータの直線回帰分析を行い、その結果から観察できる、「各年度の推定平均値と各年度の素点の平均値との差」を求めた。このような作業を行ったのは、推定平均値より素点の平均値が高い場合はその年度の TOEIC-IP は難易度が低かった、推定平均値より素点の平均値が低い場合はその年度の TOEIC-IP は難易度が高かったと推定でき、かつその回帰直線上のスコアとの差を見ることによりどのくらい難易度が高かったか、低かったかが推定できるからである。この短大のデータで、たとえば、2000 年度のテストは 24.2 点難易度が高かった、2003 年度のテストは 20.9 点難易度が低かったということが推定された。

一方、2001 年度に開設した文京女子大学外国語学部の入学生は 160 名であったが、この学生たちは、ほぼ全員が入学時から 2 年間で 4 回 TOEIC-IP を受験した。そこで、この学生たちの入学時の平均点の変動も折れ線グラフにしてみた（竹蓋他 2006 ： 33-34 ：図 1-2-8）が、やはり大きな揺れが見られた。そして、こちらのデータから、2002 年度前期のテストが 25.4 点易しかったことが推定された。TOEIC-IP に難易度の揺れがあると推定されるデータは（竹蓋他 2004 ： 11 ：図-10）にも報告されている。一ヶ月くらいしか間を置かずに受験した複数（6 名、16 名）の学生の 2 回の TOEIC-IP のスコアに方向性の一致した大きな差が見られたのである。

このような、TOEIC-IP の難易度に小さくない揺れがあるという推定の妥当性は、これまで見てきたような観察の再現性によってもある程度検証できるが、これに加えて、推定された「プリテストとポストテストの難易度の差」と「学習によるスコアの上昇量」との関係を繰り返し観察することにより別の面からの検証も可能になる。竹蓋他（2004 ： 9-12）と竹蓋他（2006 ： 35 ：表 1-2-6）にはそのことを示すデータも報告されており、我々は TOEIC-IP には難易度の少なくない揺れがあると結論した。

そのトータルスコアの変動を見て、TOEIC-IP には難易度に揺れがあることが推定されたのであるが、揺れはトータルスコアに留まらず、上位群の難易度だけが高かったり、下位群の難易度だけが高かったりすることもあること、さらには reading section のみ、または listening section のみの難易度が高かったり低かったりすることもあると推定されるデータがある（竹蓋、与那覇 2006b ： 90-93）。

5.2　テストスコアの変動要因

外部テストを受験して受け取るスコア票を見て、学習者は一喜一憂する。それが自分の学習効果、少なくとも努力、を正確に反映したスコアが得られると期待するからである。しかしながら、その期待は裏切られることが少なくない。そのひとつの理由は、上にも述べたように、テストの難易度が必ずしも安定していないからである。しかし、多くのスコアデータを観察すると、スコアの上昇、下降は、他にも想像以上に多くの要因に影響を受けることが判明する。

学習者の習熟度レベル：　Childs（1995 ： 71-72、2002 ： 18）が繰り返し指摘しており、竹蓋（2000 ： 31）、玉井（2005 ： 44）、茅野（2006 ： 102）土肥（2006 ： 25）らも異なる指導法、異なるテストで観察されたデータを数多く報告しているように、まずは、学習者の習熟度が上がれば上がるほど外部テストのスコアは上がりにくくなるという事実がある。このことは、普遍的事実と言ってよいようで、竹蓋他（2006 ： 35-39）にも、田村（2005）、杉田（2005）などのデータとともに、文京女子短期大学でのスコアの上昇、下降の実態が報告されている。竹蓋他（2006）は、4,392 名の受験者群を下位群、中位群、上位群に 3 分割して各群の平均得点上昇量を観察したのであるが、それぞれに 32.7、3.9、－ 17.7（竹蓋他 2006 ： 12 ：表 1-2-2）であったと報告している。外国語学部でも傾向は同様で、2002 年度から 2004 年度の統制群（非 3R）のみの上昇量を 3 群に分けてみると、下位群、中位群、上位群の半年平均上昇量はそれぞれ、33、20、－ 7（竹蓋他 2006 ： 37 ：図 1-2-11）であった。

このような理由によるデータの不正確な比較を避けるためには、あらかじめ実験群と統制群に等質の被験者を配置する、とくに同レベルの学生を配置する、という処置が取られる（controlled observation）ことが多いが、我々はそのようなことをするのは学生をモルモット扱いするものであるとの考えから習熟度の異なる学生群を実験群、統制群としても素点の補正によりある程度正確な比較を可能にする手法を開発している（竹蓋他 2006 ： 40-41）。

学習態度・学習動機：一方、同じ教師が、同じクラスで、同じ教材を使い、同じ指導法で指

導しても学習者の TOEFL スコア（公開テスト）の上昇量は、＋ 87 から－ 36 まで大きくばらついたという事実がある。こんなことは常識だと考える教員は多いだろう。学生の能力が大きくばらついている以上同じ指導をしても効果がばらつくのは当然で、だから少人数クラスの習熟度別指導が不可欠なのだ、と主張する教員も多い。我々も習熟度別指導は必要であると考える。しかしながら、テストスコアの変動の主因は学生の先天的能力の差がすべてではないということを示唆するデータも繰り返し観察されている。

我々は、まず、＋ 87 から－ 36 までスコアの上昇量が大きくばらついたクラスの学生全員に、この 3R の CALL を使って指導したクラスで、「真面目に学習した学生」、「形式的に学習した学生」、それに、「ほとんど学習をしなかった学生」の割合はそれぞれどのくらいだと思うかを報告させた。その結果、それぞれに 59 %、26 %、15 %という推定値（平均）が得られたのであるが、興味深いことに、約 4 ヶ月の学習で、「スコアの上昇量がマイナスとなった学生」の割合と「ほとんど学習しない学生」の割合として学生たちが推定した数値が一致したのである。このことから、我々は、学習効果は学生の能力だけでなく、彼らの学習態度（含学習時間）も大きく影響するらしいとの仮説を立てた。

その後、文京学院大学の年間を通した指導でも同様にスコアの上昇量が大きくばらつく傾向が見られた。そこで、次は学習前の習熟度レベルでなく、前期終了後に観察された「前期の上昇量の大きさ」で 3 分割して同じ学生群の「後期の上昇量」を観察した。すると、竹蓋・水光（2005 ： 178）の、図 4.3 の右欄にまとめて示したような事実が明らかになった。つまり、学習者のスコア向上量は学生の能力の高低だけでなく、学習者のやる気の高低（危機感、慢心など）に影響される部分がかなり大きいことが判明したのである。

テスト慣れ：竹蓋（2000 ： 34）には、TOEFL を 1 ～ 6 回受験した複数の学習者の 5 回のスコア上昇量の変遷を観察した結果が示されているが、その特徴は、初回は大きく伸びるが 2 回目以降は上昇量が緩やかに下がりながら安定していく傾向が観察されている。この 1 回目の上昇のみが大きいのは Childs（1995 ： 71-72）の指摘した practice effect の影響と考えられる。

また、多くの大学では長い間、1、2 年次が教養課程、3、4 年次が専門課程といった分け方が大勢を占め、英語を含む外国語は 1、2 年で指導されることが多かった。それは、2 年間で英語の力をつければ、それが 3 年次以降の専門の授業で使えるとの暗黙の期待があったものと考えられるが、実態としては、最初の 2 年間で実用になるレベルまで届かないだけでなく、学習を止めると、半年ぐらいで外国語力は大きく下がりはじめるという実態も観察され、竹蓋、与那覇（2006a ： 9）に報告されている。1 年生の前期から 4 年生の前期まで 1 学期ごとの統制群（非 3R）のみの素点で見た平均上昇量は 24、13、11、19、9、－ 7、－ 3 であり、3 年次の前期には多少の上昇傾向が残るもののそれ以降はマイナスとなってしまうことが明らかにされている。

学習時間：　TOEIC スコアの上昇量が学習時間によって大きく影響を受けると言う事実は想像に難くないが、竹蓋他（2006 ： 44 ：図 1-2-15）にも分析的なデータが示されている。文

京学院大学で 2004 年度に 3R の CALL 教材を自習で使用することを含めて学習した学生の TOEIC スコアの上昇量をその 1 学期あたりの学習時間で、0、5 時間未満、5 時間以上 10 時間未満、10 時間以上 15 時間未満、15 時間以上の 5 群に分けて観察したところ、それぞれに、23、45、50、83、89 であった。これは時間を増やせば無限に効果が上がるということを示しているわけではないが、少なくとも 1 学期に 20 時間前後は使用しないと効果を十分に得られないということを明らかにしている。なお、CALL 教材での学習時間がゼロでも 23 点上昇しているのは、これが通常の教員による対面授業 5 コマ分の効果と考えられる。

一方、同じ 3R の実験群でありながら 2002 年度の実験群と 2004 年度の実験群ではスコアの年間上昇量に無視できない差がでた。理由は、少なくとも以下の 3 点が推定される：1）2004 年度入学生のトータルの自習時間が少なかった、2）2004 年度後期のポストテストの難易度が高かった、3）2004 年度入学生の場合、前期にスコアが大きく上昇したため、後期に慢心の影響が出た。逆に、2002 年度入学生の場合、前期の上昇量が少なかったため危機感が働き頑張った。このことが両群の後期の学習時間の差にも現れている。したがって、一概に学習時間だけが原因とは言えない。しかし、年間上昇量も学期ごとの上昇量も学習時間の多寡と完全に正比例しているという事実は無視できない。

表 1　2002 年度実験群と 2004 年度実験群の学習時間と TOEIC スコア上昇量の比較

	CALL 使用時間			TOEIC スコア上昇量
	前期	後期	計	
2002 年度（15 名）	27 時間　6 分	30 時間 20 分	57 時間 26 分	100.4
2004 年度（15 名）	29 時間 56 分	17 時間 12 分	47 時間　8 分	77.7

学習前後でテスト種が異なる場合の比較： 2005 年度の外国語学部 1 年生と 2006 年度の短期大学 1 年生の場合、プリテストとポストテストで、TOEIC Bridge と TOEIC という異なるテストを使用したため、当初、学習効果の確認は出来ないだろうと考えられた。しかしながら、プリテストとポストテストそれぞれでの順位を個別に観察したところ、大きく上昇したものと下降した者があることが明らかになった。この順位変動の理由を発見するために学部生、短大生それぞれを別々に上昇群と下降群に 2 分したところ、当然ながら、両群にはスコアの上昇量に大きな影響を与える学習前のスコアに無視できない差があることが判明した。そこで上位群、下位群をどちらもプリテストの成績順に並べ、学習前のスコアの平均値が同一になるところまで機械的に両群の下位の学生または上位の学生を削除していったところ、学部生は 22 名ずつ、

表 2　2005 年度外国語学部 1 年生の学習履歴

	CALL 使用（年間計）		同学年内の順位の変化	TOEIC Bridge スコア
	時間	日数		
順位上昇群 22 名	29 時間 48 分	47 日	42.5 位上昇	152.7
順位下降群 22 名	22 時間 50 分	33 日	41.7 位下降	128.6
差	6 時間 58 分	14 日	84.2 位	24.1

表 3　2006 年度短期大学 1 年生の学習履歴

	CALL使用（1学期）		同学年内の順位の変化	TOEICスコア
	時間	日数		
順位上昇群20名	16時間30分	19日	29.2位上昇	421.5
順位下降群20名	11時間59分	15日	29.5位下降	287.5
差	4時間31分	4日	58.7位	134.0

短大生は 20 名ずつを残して、学習前の成績がほぼ同点となった。このようにして得られた 2 群それぞれの CALL 使用時間、順位の変化、TOEIC スコアの変化をまとめたものが表 2 と表 3 である。両表から 3R の CALL での学習時間の長いものは短いものに比べて順位に大きな上昇があり、スコアも大きく上昇することが明らかになった。

指導者の影響： 3R の CALL 教材は使用効果の高いことが繰り返し検証されていることもあり、多くの大学や高校から「借用依頼」がくる。平成 19 年 9 月の調査時点で 94 大学、47 高等学校からの依頼があり、教材の使用後にアンケートを送ってくれた大学だけでも 21 大学にのぼるが、そのアンケートから成就感、満足感、継続学習意欲を表明していると考えられる学生の割合を観察したところ、それぞれに 61 %、57 %、55 %であることが判明した。高校や大学レベルでの英語学習に対する印象を調査した結果の多くが 30 %くらいしか肯定的に回答していない現状からすれば決して悪い方ではない。

しかし、興味深いことは「教材作成者」が 3R の CALL 教材を使用して指導した場合は成就感、満足感、継続学習意欲が、79 %、78 %、84 %と、いずれの項目への回答でも 20 %前後高くなることである。以下にも述べるように、学生の肯定的態度が学習効果に影響することは多くの場面で確認されているので、教材の質だけでなく、教員がどれだけ教材の中身を知っているか（教材作成者＞教材借用者）、教員が情熱を持って指導しているか（教材作成者＞教材借用者）がその指導効果に大きな差をもたらすことも示唆された。

文京学院大学には、3R の構想の開発者、3R に基づいた教材の制作者、非 3R の教材採用者がおり、2002 年から 2004 年にかけていずれも複数クラスの英語授業を担当したので、その学生群の受験した TOEIC-IP のスコアを表 4 にまとめてみた。そこから、英語指導の効果は、3R 構想の開発者教員＞ 3R 教材の制作者教員＞非 3R 教材採用教員の関係にあることが客観的にも明らかになった（補正値については竹蓋他 2006 ： 40-41 参照）。

表 4　二学期（年間）連続指導の平均的効果

教員	学生延べ人数	ﾌﾟﾘﾃｽﾄｽｺｱ（TOEIC-IP）	ﾎﾟｽﾄﾃｽﾄｽｺｱ（TOEIC-IP）	上昇量	補正値上昇量
3R構想開発	42名	526.2	594.4	68.2	**201.1**
3R教材制作	43名	458.9	493.5	34.6	**120.2**
非3R教材採用	246名	347.6	379.1	31.5	**31.8**

丸めの誤差あり

教材の内容：実施されたアンケートの項目のひとつに素材の内容に興味が持てたかを尋ねたものがあったが、その項目に肯定的に回答した学生と否定的に回答した学生の差は TOEIC スコアの上昇量で 34 点（＝ 77 － 43）であることも観察されている（竹蓋他 2006 ： 54 ：図 1-2-26）。大学入試対策の勉強は、あまり興味がもてるとは考えられないが、塾に入ってまで懸命にする生徒が多いことを考えると、ニーズを感じられる教材での学習は興味の持てる教材を使う効果よりさらに高い効果を生むであろうことも容易に推測できる。

語彙力：外部テストスコアの高低に直接的に大きな影響を与える要因のひとつは語彙力であることを示すデータもある（竹蓋他 2006 ： 50 ：図 1-2-22）。語彙力は言語力のビルデイング・ブロックであるなどと言われることもあるくらいで、たしかにその重要性は無視できない。使える英語力の養成・保持には最低 7,000 語から 8,000 語が必要と言われるなかで、2,000 語前後の語彙力の学生をどう指導するかは今後の英語教育の大きな課題である。

教材の難易度：教材が難しすぎる、易しすぎるという批判を教員からも学生からもよく聞くので、それが教育効果に影響するらしいことは容易に推測できる。しかし、難易度の問題は教材（素材）だけできまるのではなく、それをどう教えるか、どう学ぶかによっても、また学生の態度によっても変わってくるので、実は、簡単に言えることではない。重要なことは、もっとも効果の上がる教材（コースウエア）は難易度がどの程度の教材なのかを学生のレベルとの相対的な見地から明らかにすることである。そのことを念頭に、我々は 3R の教材の難易度と学習効果の関係について観察した。その結果、竹蓋、水光（2005 ： 185 ：図 4.7）に見られるような図が得られたのであるが、教材の難易度と学習者の習熟度レベルが最適な関係より TOEIC で 100 点ずれると効果がほぼ半減することが明らかにされた。

あまり表にでないことではあるが、TOEIC、TOEFL のスコアが伸びない原因が、文法力や語彙力だけでなく、常識や専門的知識の欠如にあることも少なくない。たとえば、bookkeeping や investment という英語表現が和訳すれば「簿記」と「投資」であることがわかってもそれが内容的にどのようなものであるのか、どのような場面でどのような機能で使われるかを知らなければやはりコミュニケーションのための言語力とはならないからである。TOEIC より TOEFL のほうが難しいなどと言われることがあるが、テストの素材で扱っているトピックの幅がほぼ日常生活とビジネスに限られる TOEIC に比べ、いわゆる大学教養課程の話題を広く網羅している TOEFL のほうが必然的に難易度が高くなる。

5.3　テストによる指導効果の比較

外部テストを使用してもその得点や上昇量（素点）の値が多くの要因に影響されて、必ずしも個々のスコアから学習の効果がストレートに読み取れるものではないことが多くの事例から示唆された。しかし、それでは外部テストを使用して指導法や教材、それに指導教員の指導力等を比較することはできないのかというとそこまで悲観的になる必要はないことも示唆できる。たとえば、高校での指導の場合であるが、玉井（2005 ： 41）と茅野（2006 ： 102）は、それ

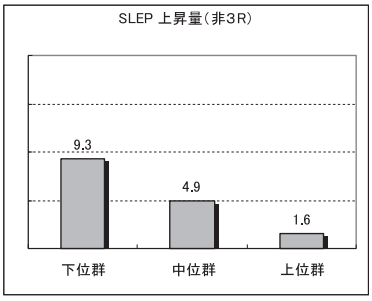

図 1　玉井（2005）より

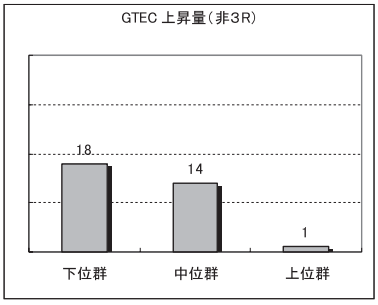

図 2　茅野（2006）より

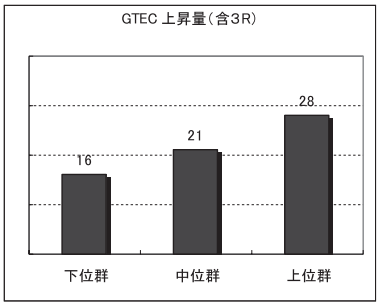

図 3　文京学院女子高等学校

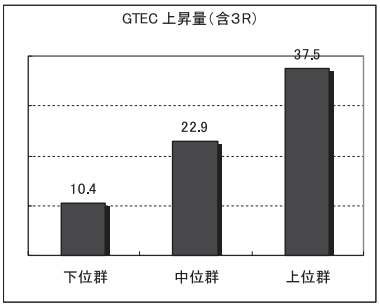

図 4　千葉市立稲毛高等学校

ぞれに SLEP と GTEC を使用して自分たちの指導効果を別々に観察（図 1,2）し、シャドーイングと呼ばれる指導法が高校生の上位群には効果が得られなかったと結論している。

一方、3R の CALL 教材の使用を含めて指導した文京学院大学女子高校や千葉市立稲毛高校では、図 3、4 に見られるように、ほぼ同レベルの高校生への指導で、平均点は言うに及ばず上位群にも大きな上昇が観察され、教材・指導法の差は一瞥して明らかである。

さらに、竹蓋、与那覇（2006a ： 24 ：表 2）には、複数の高校での指導で、3R の CALL による指導を含めた SELHi 校がいずれも「SELHi 一期校の平均スコア」より高いスコアを記録したことが示されており、3R による指導を含めた指導の優位が確認されている。

また、大学での指導の比較では、3R の CALL を含めて指導した 2 クラス（3R ：実験群）の場合とそれを含めないで指導した 2 クラス（非 3R ：統制群）での指導が比較された。このケースでは、実験群も統制群も、一般的にスコアが上がらないと言われる、それぞれの年度で上位群を構成していた学生 2 クラスずつであった。しかしながら、プリテストとポストテストに、実験群は TOEIC、そして統制群は TOEIC Bridge と、異なるテストを採用したので、当初、指導法の優劣の評価をすることは困難と考えられていた。そこで、我々は実験群、統制群の両群に同年度の他のクラスのデータを加え、それぞれを上位群（2 クラス）、中位群（8 クラス）、下位群（2 クラス）とに再構成した（図 5、6）。その後、実験群と統制群（いずれも上位群）の成績を同年度の同じテストを受験した中位群の成績と比較することによって相対的な学習効

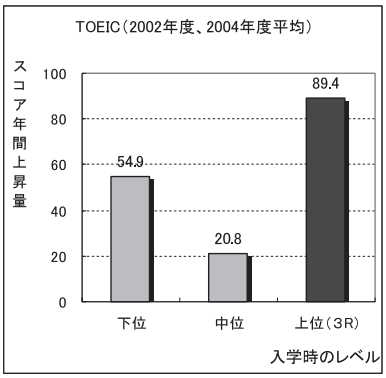

図5　上位群のみ 3R による指導

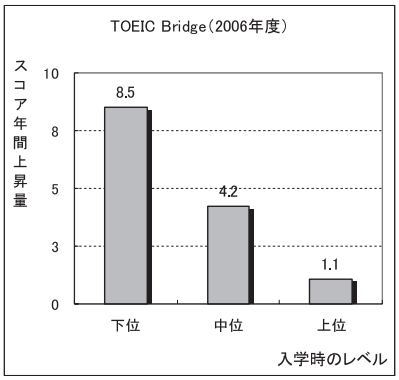

図6　全群　非 3R による指導

果を観察した。その結果、それぞれの比率が 89.4 対 20.8 ＝ 4.3（A ：実験群）と 1.1 対 4.2 ＝ 0.26（B ：統制群）であることが判明した。続いて、A/B ＝ C（＝ 16.5）を求めることによって、指導法の効果の相対的な差が観察でき、3R を含めた上位群への指導の相対的効果は非 3R による上位群への指導の 16.5 倍であることが推測された。この相対的比較の結果にはあまりにも大きな差があるように見えるかもしれないが、竹蓋他（2004 ： 5 ：表-2）で実験群と統制群の TOEIC スコアを比較した際に、学習前に 450 以上であった学生（上位群）に絞った比較の結果得られた上昇量の割合の差とほぼ同じである。このことからこの結論には再現性があり、結論の妥当性が検証されたと結論した。

教材・指導法間の優劣の比較が不安定に見える外部テストのスコアを使用しても不可能ではないと我々が結論する根拠は、複数存在する。ひとつは、「科学的論拠に基づいて制作された教材（NUWHIKS1997 ： 79-80、水光 2000 ： 4、水町 2006 ： 197）を使用し、そこに体系的な指導法が結び付けられた指導を行った場合、上位群への指導効果も向上する」とする仮説が、図 1、2 対図 3、4 の比較で、「再現性を含めて」検証されていることである。さらに図 5 と図 6 のデータも、いずれも 2 クラスずつのデータの比較であり、こちらでも再現性を含めて検証されている。加えて、図 1、2 と図 3、4 が高校での比較、図 5 と図 6 の比較が大学での比較と「異なる学習環境」でも同様の結果が得られているということである。

5.4　スコアのデータ処理

科学的な論文ではデータの客観的処理に統計的処理手法が使われることは少なくないが、それは適切に使われて初めて効果が期待できるのであり、乱用されると使わない方が良いこともあるという例を二つ挙げる。そのひとつは、竹蓋他（2006 ： 11 ：表 1-2-1）に採録されているものであるが、1994 年から 2003 年までの本学の短大生の TOEIC スコアを全部まとめてプリテストとポストテストの平均スコアを求めたところ、それぞれ 341.6 と 347.9 であった。したがって平均では 6.3 点の上昇があったということになるのであるが、この差を t-test により

検定したところ、t = 6.73 で、統計的には、有意な差であることがわかったというものである。

このデータを採録した竹蓋他（2006 ： 11）でも、「（10 点から 990 点という大きな幅のあるテストで）342 点から 348 点への 6 点の上昇はコミュニケーション能力の変化としてはほとんど意味が無い」と指摘している（竹蓋他 2006 ： 45）が、この例は、ただ形式的に統計的手法を使ったから科学的推論であり、有意差が認められたからというだけで「差が十分に大きかった」という結論にはならないと言えるよい例である。むしろ乱用を避けるべきだということを示している。この例は、サンプル数が大きくなると実質的にはほとんど意味の無い差でも統計的には有意な差であるという結論になることが多いという例であるが、同様な指摘は芝、南風原も相関係数の評価のケースを例にとり、以下のように警告している。

> … とくに大きな標本の場合、相関係数が統計的に有意であるというだけからは、ほとんど何も情報が得られないことが分かる。したがって、統計的有意性をもって“相関あり”とし、さらにあたかも同順率の高い状態であるかのように解釈することは明らかに誤りである。（芝、南風原 1992 ： 127-129）

統計手法の誤用の例は、他にも、研究会での発表や審査の無い研究論文集等にしばしば見られるので注意が肝要である。

実験データ観察のまとめ：実際に得られた大量の外部テストのデータを種々の報告からとり、分析的に比較観察すると、外部テストのスコアだからと言ってとくにその素点を安易に受け入れることはすべきでないとの結論になる。また、学習前後のテストの差、上昇量、にしても多くの要因に影響され、Ostle（1963 ： 246-248）や Hoel & Jessen（1971 ： 336）らに confounding と呼ばれる、望ましくない状況が生まれる。その中から学習効果と呼べる部分を取り出すことの困難さも判明した。しかしながら、教員が良心と熱意を持ち、科学の目（Anderson ；竹蓋 1982 ： 21）をとおして観察すれば、必要な結論を抽出することが不可能ではないことも見えてきた。

6. まとめ

仕事で使える英語力の養成を目指すカリキュラムの中で評価をどのように考えるべきか、その実践にあたって遭遇する問題点は何なのか等について、文献の調査と外部テストのデータの分析的観察による実験研究を行った。その結果、社会的にも客観的データによる評価が求められ、多くの大学ですでに TOEFL、TOEIC 等の外部テストによる評価が導入されていることが判明した。しかしながら多くの大学で採用されている外部テストも完璧なものではなく、スコアに揺れがある、スコアの変動をもたらす可能性のある要因が数多くあり、個々のテストスコアのみではその特定が容易ではないなど、現在でも幾つか難問を抱えていることも判明した。そのようなことを知らずにテストスコアの素点の高低、上昇量の多寡に一喜一憂していれば、学習者、教員両者の学習、教育意欲を削ぐことにもなり得る。にもかかわらず、テストスコア

の諸性質を知った上でうまく活用すれば、TOEFL、TOEIC、TOEIC Bridge、GTEC、SLEP 等の外部テストが指導法や教材、それに教員の指導力等の評価に使うことができ、学習者の目標設定や動機付けの高揚等にも有効に使えそうだということも見えてきた。許された誌面に限りがあり、ここでは十分な扱いが出来なかったが、我々は、教育における評価の中では診断テストが最も重要であるという立場は変えたくないし、正しい評価は指導への道標であり、牽引車であるという事実も忘れたくない。

参考文献

Childs, Marshall R.（1995）“Chapter 8: Good and Bad Uses of TOEIC by Japanese Companies,” *Language Testing in Japan*（Jamae Dean Brown and Sayako Okada Yamashita, eds.）, A Special Supplement to The Language Teacher, The Japan Association for Language Teaching, pp. 66-75

Childs, Marshall R.（2002）“What You Can Expect from TOEIC Preparation, ” *The Daily Yomiuri,* October 18, p.18

Gilfert, Susan（1995）“Chapter 9: A Comparison of TOEFL and TOEIC,” *Language Testing in Japan*（James Dean Brown and Sayako Okada Yamashita, eds.）, A Special Supplement to The Language Teacher, The Japan Association for Language Teaching pp. 76-85

Hoel, Paul G. and Raymond J. Jessen（1971）*Basic Statistics for Business and Economics*, John Wiley & Sons, Inc.

Ostle, Bernard（1963）*Statistics in Research,* The Iowa State University Press

Richards, Jack, John Platt, Heidi Weber（1985）*Longman Dictionary of Applied Linguistics,* Longman, Essex, England

Woodford, Protase E.（1982）*An Introduction to TOEIC: The Initial Validity Study*, Educational Testing Service

大阪大学言語文化研究科英語教育部会（2006）「TOEFLR-ITP Q&A（平成 18 年度版）」http://www.lang.osaka-u.ac.jp/hp/index.cgi?page=TOEFL%26reg%3B-ITP, accessed 2006/12/04

国際教育交換協議会（CIEE）日本代表部 TOEFL 事業部（2004）『2004 年 TOEFL?テストスコア利用実態調査報告書』

芝祐順、南風原朝和（1992）『行動科学における統計解析法』東京大学出版会　東京

水光雅則（2000）「英語自習用 CD-ROM を使用して英語教育に関する諸問題を解決することに向けて」『MM News』No.3 pp.1-8 京都大学総合人間学部マルチメディア教育運営委員会

杉田由仁（2005）「区切り聞きによるリスニングの指導－実践とその効果」『関東甲信越英語教育学会紀要』第 19 号

竹蓋順子(2000)「大学英語教育における複合システムの実践的研究」『言語行動の研究』第 7 号増刊号　千葉大学　pp.1-54

竹蓋順子，竹蓋幸生（1996）「文献に見る語彙指導の諸相－背景，理論，方法，課題」『千葉大学教育学部研究紀要』44 巻　第 2 部　pp.27-38

竹蓋幸生（1982）『日本人英語の科学』研究社出版

竹蓋幸生（1984）『ヒアリングの行動科学』研究社出版

竹蓋幸生（1988）『言語行動の科学』1 号

竹蓋幸生（1989）「第 7 章　指導に生きる評価」『ヒアリングの指導システム』pp.66-83
竹蓋幸生，草ヶ谷順子，与那覇信恵（2004）「外国語学部における英語教育改善の歩み(2)」『文京学院大学外国語学部・文京学院短期大学紀要』第 3 号　pp.1-15
竹蓋幸生，水光雅則編（2005）『これからの大学英語教育』岩波書店
竹蓋幸生，高橋秀夫（1991）「英語ヒアリング力診断用標準テストバッテリーおよび運用力スケールの開発」『平成 2 年度研究助成事業助成研究成果報告書』サウンド技術振興財団
竹蓋幸生，与那覇信恵（2006a）「仕事で使える英語力の養成を目指すカリキュラムの編成に関する研究（中間報告）」『文京学院大学総合研究所紀要』第 7 号　pp.3-29
竹蓋幸生，与那覇信恵（2006b）『仕事で使える英語力の養成を目指すカリキュラムの編成に関する研究　資料集』文京学院大学総合研究所
竹蓋幸生，与那覇信恵(2007 予定)「仕事で使える英語力の養成を目指すカリキュラムの編成に関する研究」『文京学院大学総合研究所紀要』第 8 号
竹蓋幸生，与那覇信恵，竹蓋順子（2006）『文京語学教育研究センター活動報告（2001 ～ 2004 年度）』文京語学教育研究センター
玉井健（2005）『リスニング指導法としてのシャドーイングの効果に関する研究』風間書房
田村哲夫（2005）「マルチメディアを活用した内容中心教授法による高校英語学習プログラムの開発－自ら英語で，学び・考え・表現する生徒の育成を目指して－」『スーパーイングリシュランゲージハイスクール研究開発実施報告書』（平成 14 年度～平成 16 年度）渋谷教育学園幕張高等学校
茅野潤一郎（2006）「ディクテーションとシャドーイングによる指導法が聴解力に与える効果」『Language Education and Technology』第 43 号　外国語教育メディア学会　pp.95-109
土肥充（2006）「TOEIC-IP による千葉大生の英語力の現状分析」『人文と教育』第 2 号　千葉大学国際教育開発センター　pp.15-29
広島大学（2006）「TOEIC（R）IP 全学一斉実施」http://home.hiroshima-u.ac.jp/flare/toeicip/, accessed 2006/12/04
水町伊佐男（2006）『コンピュータが支援する日本語の学習と教育―日本語 CALL 教材・システムの開発と利用』溪水社
宮崎充保（2006）「メモ：統一評価テストと習熟度別クラスの効果」『第 5 回愛媛大学英語教育改革セミナー　新時代の英語教育のあり方　報告書』愛媛大学英語教育センター　pp.31-32
文部科学省（2003）「『英語が使える日本人』の育成のための行動計画」http://www.mext.go.jp/b_menu/houdou/15/03/03033101.htm.
文部科学省（2005）「学校基本調査 －平成 17 年度－ 高等教育機関 統計表一覧」http://www.mext.go.jp/b_menu/toukei/001/05122201/004.htm.
ILC 国際語学センター（2006）「新しい TOEFL テストとは？」『週刊 ST』4 月 28 日
NUWHIKS（1997）「英語教育　研究と実践」『英語教育』第 46 巻 8 号　pp.78-80
TOEIC 運営委員会（2006）『TOEIC® テスト採用学校一覧』