学生の英語能力上達の測定

学生の英語能力上達の測定

モロイ・H. P. L.

はじめに

この論文では学生の英語の能力の上達を測るための研究について述べる。英語能力を測ることは一般的だが、授業中に行うことは珍しく、この研究が初めてだと言えるのではないだろうか。

本研究はいくつかの仮定から始まった。１．第２言語の教育の目的はその言語の能力を改良し、向上させるためと思われる。２．学生は将来読むものを理解できるように教えられる方がいい。３．読み方の能力の測定は、無作為に決めた様々な質問によって出来る。４．学生の英語能力は個々で違いがある上、能力があまり高くない学生は授業中に上達出来る。よって、まだ他の上達出来なかった学生より普通の試験のスコアは少ないが、上達がよく見られた人はその分の成績を与えられるはずである。５．勉強の為に使った読み方をそのまま試験のテキストに使用した場合、記憶の能力を測っているのか、読み方の能力を測っているのか、分からない。

これらの仮定で、この研究をする前に考慮しなければならない問題点があり、それは以下の点である。最も大事な問題は、どのように上達を測定した方がよいかということである。毎回、すべての学生に同じ試験をした場合、試験の難易度が前に行った試験と同じレベルかどうか分からない為、上達が出来たかどうか分からない。たとえば、１つの授業の平均のスコアは最初の試験で75点であったが、次の試験では80点になった場合、最初の試験が２番目のものより難しい場合には、上達出来なかったと言え、上達度は判定できない。２番目の試験が最初のものより難しい場合には、上達を出来たとは言えるが、古典的テスト理論では試験の難しさは試験を受けた人から独立しないため、ラッシュモデルを用いて分析することに決めた。ラッシュモデルでは、質問の難しさは試験を受け

る人の能力から独立するからと考えられており、同じ人が同じ時に試験を２つ受ける場合にも独立性が守られる。私達の授業では、試験を４回（９回授業が終了した毎）行った。つまり、１つの試験と次の試験の間、学生は英語の読み方の上では、同じ人ではないと言える。個々の学生の勉強はそれぞれ異なり、授業を欠席することもあるので、上達度は違う。

この研究では、試験４種類を作り、全ての学生が全種類の試験を受けたが、毎回試験４種類全部をやった。つまり、毎回、学生の１/４が試験 A を受け、他の１/４の学生が試験 B を、さらに１/４の学生が試験 C、試験 D といった具合である。全ての学生が無作為に決めた順番で試験を受けた。最初の試験の時、ラッシュモデルを用いて試験の質問の難易度を計算した。この難易度により、学生の上達を測定することが出来た（詳細は Molloy & Weaver, 2009参照）。

この論文における調査質問は以下の４つである。

１．勉強しながら英語の読み方の能力を上達させることが出来たか。

２．別の学生の英語の読み方の能力の上達（あるいは低下）に一貫性があったか。

そして、ラッシュモデルを使って、もう２つの調査質問に答えることが出来る。

３．学生の上達を分析するためには、ラッシュモデルによる分析の方が古典的テスト理論よりも良いかどうか。

４．試験の質問の難易度を測定するためには、ラッシュモデルによる分析の方が古典的テスト理論よりもよいかどうか。

方法論

使用したもの

授業では教科書３冊を使った。教科書は *Essential Reading*（Gough, 2008）と *Think in English （1）*（Stanley *et al.*, 1998）、*Strategic Reading : Level 1*（Richards & Eckstut-Didier, 2002）であった。最初の教科書は１年生の必須科目の授業の為のもので、他の教科書は２年生の選択科目の為のものであった。他の２つの授業では教科書を使わなかった。

試験のテキストはほとんどの部分を本４冊から取った。１つは、専門外の人の為の鮫の説明書（Lineaweaver III & Backus, 1970）（試験 A）だった。試験 B のテキストは、集団遺伝学に近い専門の学生の為の集団遺伝学の教科書（Hartl, 1981）から取った。試験 C のテキストは、プリンストン大学の入門生物学の教科書（Gould & Keeton, 1996）から取り、試験 D のテキストは、一般の読者を対象としている対称の説明書（Gardner, 1982）から取った。これらの本から３つずつ抜粋し、その上、下の４番の質問の種類のように取った抜粋１つずつはカジュアルな英語（Eggins, 1994, の通りに）に書き改めた。

ほとんどの場合、テキストが授業の教科書より困難だった。難しいテキストに決めた理由は、100点をとった学生がいる場合、その学生の能力が分からないからである。説明的な英語を分かるように教えるつもりだったため、テキストは科学書、及び専門的な本から取った。その上、試験のテキストが専門的な事であるなら、学生は試験を受けながら内容を学べると考えたからである。

試験は、６種類の質問から構成された。

１．新しい単語の意味を理解するための質問。将来、新しい単語が作られるはずだし、言語を使う人はそのような単語を理解しなければならないため、このような質問が大切だと言える。(たとえば、30年前は「hypertext」は普通に使われる単語ではなく、60年前は「fission」も、90年前には「antibiotic」も普通に使われる単語ではなかった。)

２．テキストから含蓄出来たものと出来ない質問。

３．指示物の単語を理解するための質問。

４．修辞的技巧を理解するための質問。

５．テキストの形式を理解するための質問と、形式を理解出来ない質問。

６．テキストの次のテーマを予想する質問と前のテーマを推論する質問。

可能な限り、いつも同じ種類の質問は同じ所で出題した。たとえば、全ての試験で最初の質問は、次のテーマは何かを問う質問であったし、最後の質問はいつも最もふさわしいタイトルを決めるものであった。

質問に対する回答方法は２種類あった。１つは、複数の選択肢（選択肢は２つから４つまで）だった。もう１つは、丸バツ式だったが、毎回もう１つの選択肢は「不明」だった。試験の質問は30個ずつあった。質問は、英語でも日本語でも書いてあったが、単語の意味を問う質問の日本語のバージョンでは、単語はそのまま英語で書いてあった。(例　(試験 A から):「１段落目で、“locality”は何と言う意味ですか？」)

和訳は以下の通りに行った。最初に質問を英語で書き、次に、私達が和訳した。次に、英語が分からない人が日本語の質問を読み、相応しくない文法と単語を直した。そして、もう１人のバイリンガルな人が英語と直した日本語のバージョンを比較した。その後、問題があった場合には英語と日本語のバージョンが同じ意味になるように直した。

試験はＡ３紙に印刷された。

被験者

被験者は４つの授業を履修する177人で、被験者の学生は全員同じ専攻であった。３つの授業は

２年生の選択英語のリーディングであり、もう１つは１年生の必修英語のリーディングの授業であった。１年生の授業は TOEIC 試験で能力別編成した。TOEIC 試験と別のレベル・チェック試験の相互関係は高い(Molloy 2008; Molloy & Shimura, 2006)。

試験の問題の難易度の基礎を確立するために、さらに同じ大学の学生と別の大学の学生102人が最初の試験を行った同時期に試験を受けた。(下記の説明の通り、この最初の試験の管理で、次の問題の難易度を確立した。)

試験の管理

最初の試験の時、隣同士の学生が別の試験を受けるように４種類の試験が配られた。つまり、学生279人の１/４が試験 A を受け、１/４の学生が試験 B を受けたといった具合である。このように試験は無作為に管理された。次の試験を受ける時には、全ての学生が24パターンから１つのパターンのとおりに試験を受けた。たとえば、１人が BACD の順番に受けた場合、もう１人は CDAB のとおりに受けた。それぞれの授業は40人位の学生が履修していたため、１つの授業で試験を同じ順番に受けた学生は２人くらいであった。

試験を行った後、すぐに答えをエクセルのスプレッドシートに入力し、Winsteps（Linacre, 2006）というラッシュモデルのプログラムを使って分析した。学生には、ラッシュ能力スコア、誤差、(２番目の試験から)上達度を教えた。

最初と次の３つの試験では Winsteps の使い方が違った。最初の試験は、無作為に受けさせたため、学生の能力は正規分布になると考えた。そこで、Winsteps では学生達の平均の能力を０にセットした。質問の難易度は変化するようにした。次の試験では、最初の試験の分析で分かった難易度を使ったので、学生の上達度が分かった。そして、上達度が分かるように、最初の試験をもう１回分析した。その際、最初の分析の難易度をセットし、能力は変化するようにした。

分析

上記のような分析を行った上、ラッシュ能力スコアで学生の返事の能力スコアと別の授業の学生の能力スコアを計算した。

それから、全ての学生が全ての授業が終わった時には４種類の試験の全ての質問に答えたため、合成の試験を作ることが出来た。そのような合成の試験の質問は120個あった。これは、合成の試験と本物の試験の質問の難易度を比較を行い、質問が本当に能力を測定するならば、合成の試験の質問の難易度は本物のより低いから学生が答え方を授業中に学べるといえる。

調査質問の３番と４番の為、古典的テスト理論のとおりに学生の偏差値(*z*-スコア)を計算した。

また、古典的テスト理論のとおりに質問の難易度が分かるように項目容易度を計算した(Brown, 1996)。そして、古典的テスト理論の結果とラッシュモデルの分析結果とを相関した。

結果

調査質問１：授業の場合英語の読み方の能力の上達を出来たか。

上達はしたが、いつもではなかった。図１は177人の学生の返事の能力スコアの進行を表している。最初の試験は左の点で、右の点が最後の試験である。Y 軸はラッシュ分析の能力スコアである。

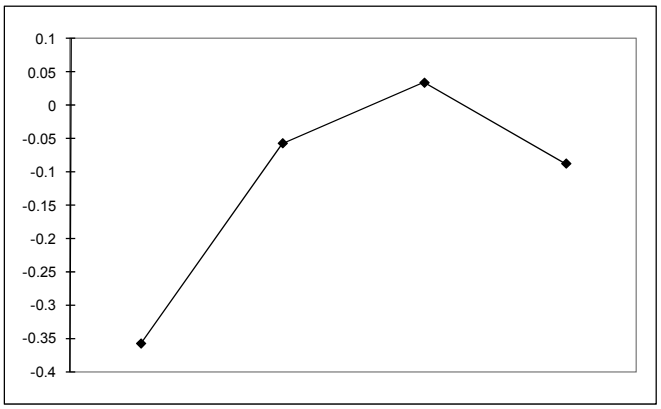

図１．学生全体の返事の能力スコアと試験結果

図２から図５は、授業ごとの学生の返事の能力スコアの進行を表している。最初の試験は左の点で、右の点が最後の試験である。Y 軸はラッシュ分析の能力スコアである。最初の試験から最後の試験まで、読み方は上達を出来たが連続的ではなかった。

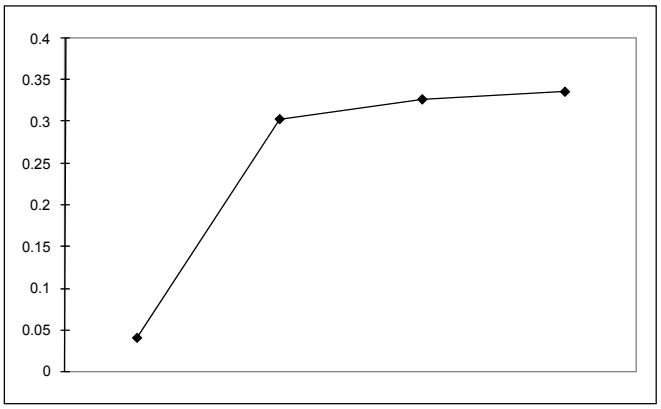

図２．クラス A（１年生）

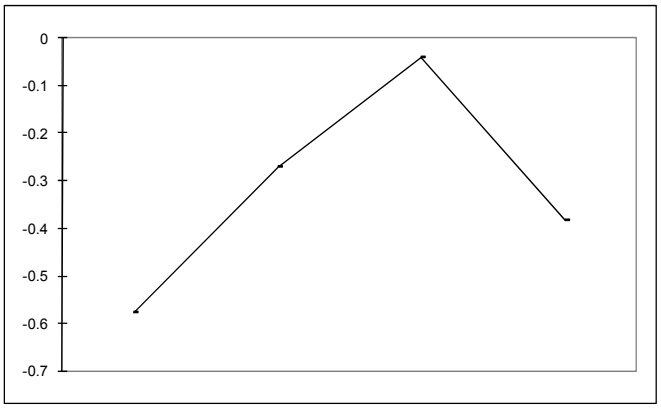

図３．クラスB（２年生）

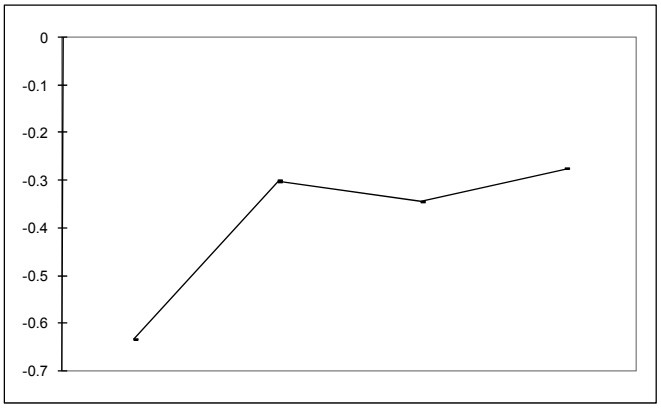

図４．クラスC（２年生）

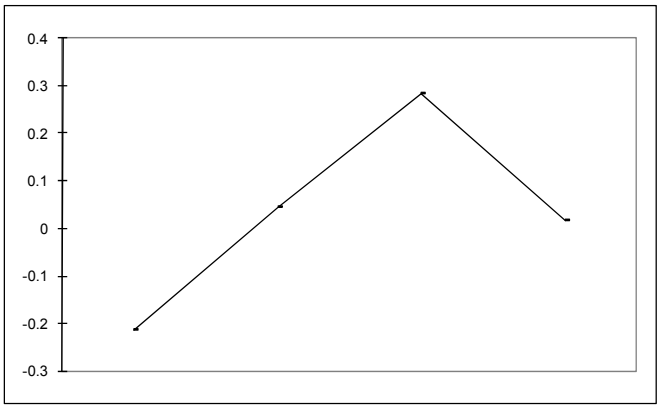

図５．クラスD（２年生）

研究質問２：別の学生の英語の読み方の能力の上達（あるいは低下）に一貫性はあったか。

読み方の能力の上達に一貫性は見られなかった。

図６は、学生177人の４つの試験ごとの能力スコアを示している。Y 軸はラッシュ分析の能力スコアであり、X 軸は試験１から４まで(左から右)である（一部の学生が脱落した為、左の辺の点が右の辺の点より多い)。

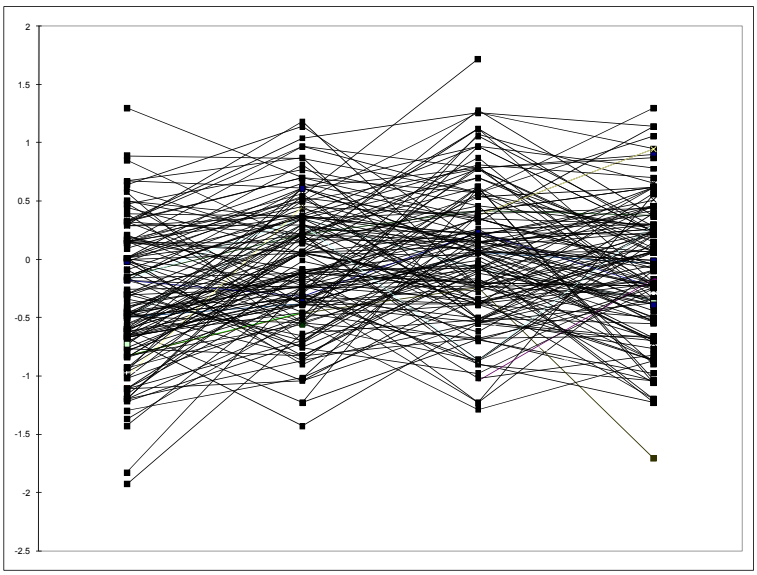

図６．全ての学生の能力スコア（*n* = 177）

分かりやすいように、図７と図８では授業ごとの学生の能力スコアが示してある。Y 軸はラッシュ分析の能力スコアで、X 軸は試験１から４まで(左から右)である（一部の学生が脱落した為、左の辺の点が右辺の点より多い)。図７は、１年生の必修の授業の学生のスコアである。図８は、２年生の選択授業の学生のスコアである。

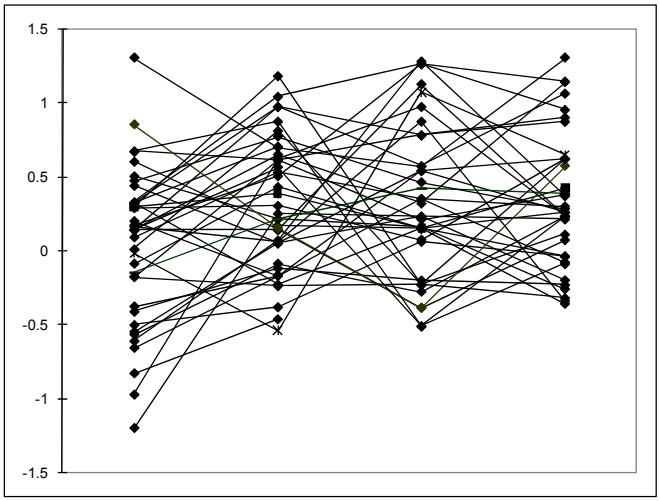

図７．クラス A（１年生)の学生の能力スコア

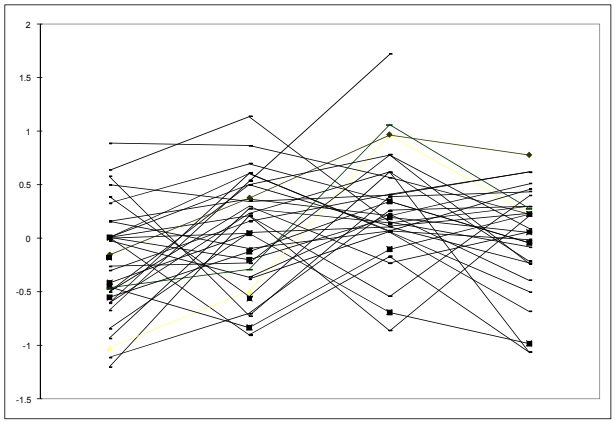

図８．クラス D（２年生)の学生の能力スコア

上記と同じ事が能力スコアの相関係数で分かる。表１は、ピアソンの積率相関係数を示す。もしすべての条件がそろった場合（はっきりと(i)全ての学生が堅実に上達を出来た場合、(ii)試験が英語の能力を正確に測定した場合)、積率相関係数が高くなる。

表１　試験と試験のピアソンの積率相関係数。

$r_{1,4}$は試験１と試験４の積率相関係数を意味する。ラッシュ能力スコアの相関係数 $n = 177$.

$r_{1,2}$	$r_{1,3}$	$r_{1,4}$	$r_{2,3}$	$r_{2,4}$	$r_{3,4}$
0.46	0.28	0.39	0.31	0.43	0.40

研究質問３：学生の上達を測定するためにラッシュの分析は古典的テスト理論よりもよいか。

この研究では、ラッシュモデルによる分析が古典的テスト理論による分析よりも良いという結果は得られなかった。素点も偏差値(*z*-スコア)とラッシュ能力スコアを相互に関係させた。

表２　素点(%)と偏差値(*z*)、
ラッシュ能力スコア(abil)のピアソンの積率相関係数(*r*)

	時間１	時間２	時間３	時間４
$r_{abil,\%}$	0.98	0.93	0.96	0.96
$r_{abil,z}$	0.97	0.96	0.98	0.98

研究質問４：試験の質問の難易度を測定するためには、ラッシュモデルを使った分析は古典的テスト理論を使った分析よりよいか。

この研究では、試験問題の難易度を測定するためには、ラッシュモデルを使った分析が古典的テスト理論よりも良いという結果は得られなかった。ラッシュの難易度のスコアと古典的テスト理論の項目容易度は相互に関係した。表３は、「全体」の縦隊で合成の相関関係が書いてある。相関係数が負数であり、ラッシュ分析では負数は質問が簡単であることを意味し、古典的テスト理論では正数は質問が簡単であるという意味である。

表３　ラッシュの難易度のスコアと古典的テスト理論の項目容易度の相関関係

	全体	時間１	時間２	時間３	時間４
試験 A	−0.74	−0.76	−0.92	−0.89	−0.97
試験 B	−0.96	−0.99	−0.93	−0.95	−0.996
試験 C	−0.95	−0.996	−0.996	−0.996	−0.96
試験 D	−0.96	−0.98	−0.99	−0.98	−0.99

論考

研究質問１：勉強しながら英語の読み方の能力の上達を出来たか。

本研究から、上達が出来ていることが明らかになった。図２から図５までによると、学生の平均能力スコアは上昇した。図１によると全ての学生が上達を出来たことがわかる。３番目の試験から４番目の試験は、平均のスコアが下ったが、なぜかは分からない。最初の試験から２番目試験までは、毎回、上達の幅が最も大きかった。これには３つの理由が考えられる。１．学生が真剣に最初の「レベル・チェック」試験を受けなかった。２．学生はテスト形式に慣れていた。３．学生が最も大事なことのほとんどを２番目の試験の時までに学んだ。

この研究の被験者の学生の専攻は英語ではなかった。４番目の試験の時には、学生は専門の授業のためのレポートを書き、プレゼンの準備をし、専攻の授業の試験のために勉強したと思われる。

研究質問２：別の学生は英語の読み方の能力の上達（あるいは低下）に一貫性があったか。

学生の上達に一貫性はなかった。なぜかは分からない。上記の理由で上達に一貫性がなかった。他にも原因があると思われるが、その原因は明らかではない。

研究質問３：学生の上達を測定するためには、ラッシュの分析の方が古典的テスト理論による分析よりもよいか。

表２から、この研究では、ラッシュによる分析と古典的テスト理論による分析に違いはないことが明らかになった。

研究質問４：　試験の質問の難易度を測定するためには、ラッシュの分析の方が古典的テスト理論による分析よりもよいか。

表３から、ラッシュの分析と古典的テスト理論による分析の結果は同じである。古典的テスト理論を使ったが、研究成果が同じだった。この研究におけるラッシュ分析の最大の利点は、便利さという点である。ラッシュ分析はすぐれた理論を基礎としており、これも重要な点である。

欠点と将来の研究

この研究で使った試験の信頼性は、質問の分析のためにはよいが、学生の能力スコア（Molloy & Weaver, 2009）のためには更にいい試験であれば良かったと考えている。今回使用した試験では、学生を確実に分離することができなかった。これは以下のような理由からだったと考えている。

１．特定の大学に入学した学生はほぼ同じ英語の能力である（Molloy & Shimura, 2006）。２．試験のテキストが難しかった。一部の学生には難しすぎたと思っている。そのような学生は、試験ができず、がっかりしてしまったと思われる。誰かが質問の全部を答え、また、一部の学生は全ての試験に解答し、全ての学生が、さらに分かりやすいテキストでも研究を行いたい。試験の長さ（質問が30個以上の試験）は適切であったと思われる。質問の難易度と順番の相関関係は高くなかったため、疲労が影響したとは考えられず、次回はより長い試験を使えるだろう。

この研究の目的は上達度を測定することであったため、ある程度成功したと言える。この研究は次のリーディング授業において他のテキストで同じような試験を行うつもりである。その上、リスニングの授業にも同じような試験の使い方を行う予定である。

参考文献

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

Eggins, S. (1994). An introduction to systemic functional linguistics. London: Pinter.

Gardner, M. (1982). *The ambidextrous universe: Left, right, and the fall of parity*（２ｎｄ ed.). Harmondsworth, Middlesex, England: Penguin.

Gough, C. (2008). *Essential reading*. Oxford: Macmillan Education.

Gould, J. L., & Keeton, W. T. (1996). *Biological science*（６ｔｈ ed.). New York: W. W. Norton & Co.

Hartl, D. L. (1981). *A primer of population genetics.* Sunderland, MA: Sinauer Associates.

Linacre, J. M. (2006). Winsteps Ver. 3.64.0 (Computer software). Chicago: Author.

Lineaweaver III, T. H., & Backus, R. H. (1970). *The natural history of sharks.* New York: Lyons & Burford.

Molloy, H. P. L. (2007). Comparing the reliability of multiple-choice Ｌ２ cloze tests with ４ and ５ choices. *Journal of Bunka Women's University, 15*, 115-126.

Molloy, H. P. L., & Shimura, M. (2006, June 25). *Comparing approaches to proficiency assessment problems facing faculty*. Paper presented at the JACET Kanto Regional Meeting, Tokyo.

Molloy, H. P. L., & Weaver, C. (2009). *Approaching the linking problem in classroom testing.* Proceedings of the Temple University Applied Linguistics Colloquium. Tokyo: Temple Univeristy.

Richards, J., & Eckstut-Didier, S. (2002). *Strategic Reading: Level 1.* Cambridge: Cambridge University Press.

Stanley, N., Brown, L., Kasprowicz, K., Kagata, T., & Tsumura, S. (1998). *Think in English （1）*. Tokyo: Macmillan Languagehouse.

（2009年９月14日受理）