English Language Proficiency Gains in an Integrated, Self-Access Program Class of 2005

Part 2: The ETS Test Battery

Thomas M. Pendergast

Part 1 of this two-part consideration of proficiency gains in a two-year（短期大学英語科）, centrally-managed English language program reported the results of four measures of language change, namely the scores of the Geneva-based CEEL（Center for the Experimentation and Evaluation of Language Learning Techniques）test battery, which looks at Listening Comprehension（Channel Capacity）, spoken Fluency, Correctness, and Expression.

Part 2 presents data from six measures of the ETS battery, namely the three components（Listening, Reading, and Total）, respectively, of the TOEIC® and SLEP® examinations.

These data are presented in the hope that comprehensive data from other programs will be forthcoming to provide benchmarks for investigation of what "works" in language learning, especially in cases where the nature of the program has more to do with the outcome than the self-fulfilling induction of already success-oriented students.

Keywords: language learning, TOEIC®, SLEP®, proficiency testing, t-test, Effect Size, SAPL, self-access pair learning, Extensive Reading, StoryTelling, Reliability

Introduction

Theoretically, language education faculty are all in search of "the best"（*effective* at least, *efficient* at best）way to help their students get over the hump in language learning. "Getting over the hump" is one way of saying that a student has become motile in the target language, able and willing to use it for communication, even minimally. At this point experience with the language "as she is spoke" is a key to further improvement. Few in Japan get there.

In the two parts of the article, data are presented from 10 angles: three of listening comprehension, two of reading comprehension, two of total scores reflecting the sum of these two, and one each of spoken fluency, correctness, and expression（communication）. All of these tests are unprepared proficiency measures（NRT[1)]）and none are achievement tests（CRT[2)]）, which indicate the result of only a limited, pre-prepared segment of language gain.

In Part 1, we expressed the hope that these data might serve as a benchmark by which program effectiveness could be compared. This remains our aim, but reflection makes us wonder if a comparison is even possible. Here are two examples of the special circumstances attending our data.

1. For 26 years[3], we had excellent cooperation from our administration. This in itself is remarkable. The curriculum (the Core Language Program --- six classes a week of self-access pair learning, or *SAPL*, and one class a week of Extensive/Intensive Reading and/or StoryTelling --- was maintained intact throughout the program). Pre- and post- proficiency testing (both the CEEL and ETS batteries) was continued during the entire period. Whether or not the curriculum was the best possible in the circumstances is not as important to our overview as the fact that it was an unvarying one with a high level of standardization. For this reason, we have a broad fix on what treatment produced what results and could replicate that treatment. The major imponderable was the quality of the students, which in fact varied greatly over a period of demographic upheaval in Japan.
2. The mean beginning level of our students, as evaluated by the TOEIC® and N73 Channel Capacity in the first semester of each year varied somewhere under the proficiency horizon, but in no year reached an average functional level. More specifically, the mean TOEIC® Total Score at entrance to the program never rose above 300/990, nor the N73 Listening Comprehension above 50/1000 (or even above 35). Further (for the six years that it was administered in addition to the TOEIC®), the average TOEFL® Total Score never rose above 400/677 (or even 370). The ETS documentation for the TOEFL® states that a total score of < 400 is "meaningless." A true "comparative effectiveness" study for these data would require a similar cohort of non-functional beginners.

Proficiency Testing in General

University administrators in Japan often seem to be comfortable with the idea of administering English proficiency tests to all matriculating students, whether the test be the Japan-based *EiKen* (STEP), the TOEIC®, or the TOEFL®. This seems a reasonable way to find out what you are dealing with and a help, over time, in arranging your curriculum to suit. In most cases, however, subsequent proficiency testing is left up to the inclination of individual students. Summary post-testing data to verify that the program has done its job seem scarce --- hence, the current report, as a start.

Test Periods and Exceptions

In the program under discussion, the TOEIC® pre-test was administered during the orientation week of the first semester and the post-test two weeks before the end of the fourth semester. Since the test is administered to the general public six times a year, it happens that some students sit for these tests at their own discretion, especially in those cases where they want an official score for their resumes. In some cases, these intermediate scores are higher than those obtained at the official university post-sitting.

In all cases, we have considered as the post-test the highest score obtained by the student after the first semester and before graduation.

Contact Hours

The core program was four semesters of seven classes a week. Six of these were self-access pair learning (Ferguson, 1980), or *SAPL*, classes in which students directly accessed their learning material, working in pairs while instructors (or "coordinators") assisted them in their activities. One other class a week was devoted to StoryTelling in the first year, with a relatively heavy load of Extensive Reading of graded readers as homework. In the second year, students had the option of continuing with the StoryTelling class or, alternatively, choosing an Intensive Reading class.

Altogether, the Core Program provided about 450 hours of class over two years. Students also enrolled in electives, and the total number of Contact Hours has been rounded to about 500 hours for the four semesters. In the case of Tables 4-6, the SLEP® tests were given at the end of the first and fourth semesters (allowing a total of three semesters worth of contact hours between tests). Hence, the total number of contact hours used in the SLEP® calculations has been reduced by one-fourth to 375 hours.

Use of the TOEIC® at this Level

Most of the students who sat the pre-test had never "spoken" English, had never met a native speaker of English (this is changing now with the introduction of Assistant Language Teachers (ALTs) into the junior high schools), had never taken a proficiency test with directions in English, and had never listened to 45 minutes (the length of TOEIC® Part 1) of English at a single sitting. For this reason, as a proctor I often felt not only that there was some question as to the usefulness of this instrument as a pre-test for these particular students but also that the experience itself might be so daunting for some of them that they would become seriously discouraged (英語嫌いになる). It even occurred to me that some students might question the wisdom of an institution that would ask them to do something so obviously impossible[5]. This feeling was exacerbated during the six years when the university administration imposed the TOEFL® on entering freshmen at the beginning of the first semester.

Listening

The students in the program had a great deal of focused listening study in their two years. On average, they finished nine units in the *SAPL* (Ferguson, Ferguson, & O'Reilly, 1980) program, which translates into relatively intensive, self-access (in small groups, or modules, of 3-5 students) study of at least 19 cassette tapes. In addition, they listened to and studied at least 24 stories in their Storytelling class in the first year, and more in the second year if they chose this option. All students completed the Valerian Postofsky[6]-inspired *Beginning Listening Cycles* series (Boyd & Boyd, 1986) of four tapes and many continued with the two-tape *Listening Cycles* (Boyd & Boyd, 1985) series in their second year.

The two-tailed t-test gain of $p < 0.0001$ on their post-TOEIC® Part 1 Listening Comprehension test is

rated as "extremely significant" (*GraphPad Instat, v. 3.0a for Macintosh*). 54 out of 57 students posted gains, one obtained the same score, and two declined. The K-R21 reliability rating was 0.945. Brown (1996, p. 197) points out that the K-R21 is the "easiest internal-consistency estimate to calculate," one of the two reasons it was chosen for this study (but see disclaimer for the SLEP® data). The other reason for choosing this measure is that, as Brown emphasizes on p. 198, the K-R21 is a particularly "conservative estimate of the reliability of a test."

Effect Size[7)]

For those not familiar with this measure, please refer to Note #7.

Some Comments on the Listening Training

Self-Access Pair Learning

In self-access pair learning (*SAPL*) students form "modules" of 3-5 students which divide into "classes" of 2-3 students. Each module is equipped with a set of tapes or CDs, workbooks, a playback machine, and two speakers attached to screens (partitions) which visually set off each module of students from other modules in the room. In this way, for example, a room of 28 students could have seven modules of four students each, each module proceeding at its own pace with the help of an instructor/coordinator who circulates among them. Each module is independent of the other modules and can study a different level of the same language that the other modules study, or even a different language if the coordinator is linguistically equipped to assist them. A module of four normally divides into two "classes" of two pairs each. A module of three works as a tryad. A module of five divides into two "classes" - one dyad and one tryad. Initially, students are given written instructions in Japanese which explain how to proceed. Following mathetics principles these instructions soon give way to English ones, and the students quickly get used to the process.

All study material is eventually in the target language (English or other). The language is introduced aurally, with reference to the printed page only following extensive listening activity. In some cases, there is no reference to a printed version which, in fact, does not exist. Students work out the language among themselves, performing role-plays and asking and answering questions to provide mutual feedback as to what they think they understand from the recordings. In this way, they learn to listen carefully and to communicate with gestures and facial expression. They assist each other in discovering --- in the case of the English language and Japanese students --- one of the most difficult parts of the language, namely the unstressed elements or function words which make up the skeleton of the "grammar." There is, however, little or no overt discussion of grammar.

StoryTelling

This once-a-week class focused primarily on the instructor telling stories in English during class, with emphasis on vocabulary acquisition, as well as encouragement to read roughly 1000 pages of graded readers a semester. The StoryTelling program was available to about half of the students in their first two semesters.

Table 1

TOEIC® Part 1 Listening Comprehension		
	Pre-Test	Post-Test
Dates	Apr., 2003	Dec., 2004
Contact hours		Ca. 500
N	57	57
Maximum possible（scaled）score	495	495
Test time	45mins.	45mins.
Mean score	172.190	246.840
Mean gain		74.650
Number of students who gained		54
No change		1
Declined		2
Standard deviation（StD）	46.990	46.528
Kuder-Richardson21（K-R21）	0.951	0.945
Standard error of measurement（SEM）	6.224	6.163
Median score	175	245
Median gain		70
High score	260	380
Low score	50	150
（Two-tailed paired t-test）P value		< 0.0001
Effect size		1.60
Bias corrected（Hedges）		1.59

Listening and Reading Scores（pre-test）

As is immediately apparent from a comparison of the pre-test Listening and Reading scores, the Listening score is low（172/495） … but the Reading score is so low（94/495）as to be meaningless. Part of the reason may lie in the fact that the Listening section of the test better motivates students, who are willy-nilly paced by the tape, whereas many either give up on the Reading section before finishing or fail to finish in the time allotted. Students are more likely to answer all of the questions on Part 1, and any answer is better than none.

Table 2

TOEIC® Part 2 Reading Comprehension		
	Pre-Test	Post-Test
Dates	Apr., 2003	Dec., 2004
Contact hours		ca. 500
N	57	57
Maximum possible（scaled）score	495	495
Test time	75mins.	75mins.
Mean score	94.474	153.600
Mean gain		59.126
Number of students who gained		52
No change		0
Declined		5
StD	41.659	47.290
K-R21	0.958	0.955
SEM	5.518	6.264
Median score	90	145
Median gain		55
High score	200	270
Low score	10	40
（Two-tailed paired t-test）P value		< 0.0001
Effect size		1.33
Bias corrected（Hedges）		1.32

Program Goal for the TOEIC® Total Score

With a pre-test total score of 265（< 300 is generally considered "non-functional"）, the originally-envisioned goal for the post-test of 400 points was challenging for this group but not totally unrealistic, requiring a mean gain of 135 points. Ten years earlier, the Class of 1995 had gained a program record 149 points and most of the classes in the interim had gained an average of more than 135 points.

Statistically, the final mean score of 399.910 points could scarcely have been closer to the goal. All of the 57 students showed gains and the mean gain was 135 points, in spite of the notably inauspicious circumstances referred to in Part 1 of this report.

Table 3

TOEIC® Total Score		
	Pre-Test	Post-Test
Dates	Apr., 2003	Dec., 2004
Contact Hours		ca. 500
N	57	57
Maximum possible（scaled）score	990	990
Test time	120mins.	120mins.
Mean score	264.910	399.910
Mean gain		135.000
Number of students who gained		57
No change		0
Declined		0
StD	70.897	76.587
K-R21	0.962	0.960
SEM	9.390	10.144
Median score	265	395
Median gain		130
High score	400	640
Low score	120	210
（Two-tailed paired t-test）P value		< 0.0001
Effect size		1.83
Bias corrected（Hedges）		1.82

The Secondary Level English Proficiency（SLEP®）Test

The SLEP® test is available for purchase from Educational Testing Service（ETS）. It is an NRT purporting to test general proficiency and therefore suited to the aim of providing a number of proficiency measures for a two-year college English program. It comes in six forms to facilitate re-testing. Each form is provided with scaling to convert raw scores to equivalencies, no matter which form is used. In the current study, Form 1 was used for both the pre- and post-tests, with an interval of a year

and a half between sittings. Like the TOEIC® the test is in two parts (Listening and Reading), providing Listening and Reading part scores and a Total Score. The test was "designed to be an easier test than the TOEFL®" (*SLEP Test Manual*, 1991). The *Manual* provides copious data supporting the test, and claims K-R20 internal consistency reliability scores of 0.94 (Listening), 0.93 (Reading), and 0.96 (Total Score) in two international administrations to 326 students at more than 30 test centers worldwide (*SLEP Test Manual*, 1991).

Test Administration for the Current Data

Unfortunately, the Reliability of this test for the Class of 2005, as can be seen from each cohort of data, was not established for our students. The first administration produced the results below (Tables 1-3), which even prior to data analysis seemed skewed. This could have been partly due to the fact that all instructions are given in English on this test, which might have confused students on some of the eight sections, since the standard in Japan is Japanese instructions on language tests. It must be admitted that some of the instructions were complicated.

An algorithm was developed to identify students who did not seem to perform at their level of competence. Brown (1996, p. 29) discusses "the Competence/Performance Issue" and observes that competence is hard to get at. A re-test was administered to a certain number of students who did not seem to have performed as well as they could have. The re-test was given four weeks after the first test --- in Brown's words, p. 193, "{(long enough) that students are not likely to remember the items on the test, but (short enough) that the students have not … (learned more language)}, in the case of both the pre- and post- tests. There were no classes in the interval. The results are noted in brackets. Both sets of data were analyzed for internal consistency reliability, with the results shown. Considering the high reliability of the TOEIC® scores for the same students, the conclusion is that the SLEP® calculations may be flawed. Nevertheless, the available data are noted.

SLEP® N = 56

Of the 57 students who participated in the other areas of this study, one became seriously ill just before sitting the SLEP® series, leaving 56 scores.

Difference in Contact Hours and Comparison of SLEP® Scores with TOEIC® Scores

The SLEP® pre-tests were administered at the end of the first of four semesters, while the TOEIC® pre-tests were taken at the beginning of the first semester. This accounts for the 25% fewer hours of class time (contact hours) for the SLEP® figure in comparison with the TOEIC®.

As a trial, the Effect Sizes of the SLEP® scores were compared with the Effect Sizes of adjusted TOEIC® scores, i.e., reduced by the equivalent of one semester, or 25%. The result for Effect Size in Listening Comprehension was: SLEP® 1.18 compared to 25%-reduced TOEIC® 1.19; Reading Comprehension: SLEP® 1.09 compared to TOEIC® 0.99; Total Score: SLEP® 1.36 compared to TOEIC®

1.37.

Data in Brackets

Of the 56 students taking Part 1 Listening Comprehension, some were selected for a re-test. As a trial, an algorithm was applied to their original scores to give an indication of which scores might not be reliable. Four students re-took the Part 1 pre-test, and 22 the post-test. The higher of the two results is shown in brackets. The reliability of the first test was low and that of the new results lower. The same procedure was followed with Part 2 Reading Comprehension. For Part 2, the number of re-takers was eight for the pre-test and 26 for the post-test (out of 56 total test takers for each of the two tests).

Effect Size and Internal Consistency Reliability (K-R21)

What influences Effect Size and the Reliability is the same thing, namely, the Mean and the Standard Deviation, but in inverse correlation. Given a static Mean, the smaller the StD, the larger the Effect Size and the lower the Reliability. The data from these tests, due to scaling, appear to result in extraordinarily compressed StD and SEM figures. It should be mentioned that the tests were machine-scored, which raises the possibility that the original scoring data were incorrectly input. This would however not account for the considerable score increase for those who re-took the test. The favorable Reliability reports from other institutions in Japan (e.g., Suzuki 2005) strongly suggest that further and consultative investigation in regard to calculating Reliability is needed.

Test-ReTest and Pearson's *r*

Part 1 and Part 2 were each administered to all 56 students at the end of the first and fourth semesters. Since N=56, the result was 224 scores {(56 students * two tests (Listening and Reading * two initial sittings for each test (pre- and post-)}. Of these 224 scores, 68 were identified as low and a re-test for either Part 1 or 2 was administered a month after the first sitting. The Pearson Linear Correlation of these test/re-test 68 pairs was r = 0.73.

Table 4

SLEP® Part 1 Listening Comprehension		
Pre-Test	Post-Test	
Dates	Jul., 2003	Jan.., 2006
Contact Hours		ca. 375
N	56	56
Maximum possible（scaled）score	32	32
Test time	40mins.	40mins.
Mean	17.018（17.268）	19.339（20.464）
Mean gain		2.321（3.196）
Number of students who gained		41（49）
No change		5（5）
Declined		10（2）
StD	3.024（2.687）	2.974（2.207）
K-R21	0.13（-0.10）	0.14（-0.53）
SEM	0.404（0.359）	0.398（0.295）
Median score	17.00（17.50）	19.50（20.50）
Median gain		2.50（3.00）
High score	23（23）	25（25）
Low score	10（10）	11（14）
（Two-tailed paired t-test）P value		< 0.0001
Effect size		0.77（1.19）
Bias corrected（Hedges）		0.77（1.18）

Classes Specifically Related to Reading

Semesters 1 & 2: StoryTelling/Extensive Reading or Extensive Reading Only

StoryTelling with emphasis on listening and vocabulary was the class format for half of the first-year students, with extensive reading assigned for homework. The remaining half of the first-year students did extensive reading inside and outside of class, with a goal of 1000 pages of graded readers per semester. Surveys indicated that the actual amount read was between 700-800 pages a semester.

Semesters 3 & 4: StoryTelling/Extensive Reading, Extensive Reading Only, or Intensive Reading Only

In their third and fourth semesters, students had a choice（選択必須）, as per the underlined rubric immediately above.

Suzuki（2005）provides a detailed analysis of his use of the SLEP® in an attempt to see if it could be correlated with the Japan-based STEP（or *EiKen* 英検）test. His data are very refined but do not emphasize gains over the period of treatment. His Mean Total Score of 38.97 seems to represent the average level of attainment after three years at a national vocational school. While these teenagers were

Table 5

SLEP® Part 2 Reading Comprehension		
	Pre-Test	Post-Test
Dates	Jul., 2003	Jan., 2006
Contact Hours		Ca. 375
N	56	56
Maximum possible（scaled）score	35	35
Test time	45mins.	45mins.
Mean	19.446（20.018）	21.446（22.500）
Mean gain		2.000（2.482）
Number of students who gained		36（47）
No change		08（06）
Declined		12（03）
StD	2.676（2.195）	2.551（2.328）
K-R21	‘0.21（-0.80）	‘-0.28（-0.50）
SEM	0.358（0.293）	0.341（0.311）
Median score	19.50（20.00）	20.50（22.00）
Median gain		1.00（2.00）
High score	24（26）	27（28）
Low score	11（15）	17（17）
（Two-tailed paired t-test）P value		< 0.0001
Effect size		0.77（1.10）
Bias corrected（Hedges）		0.76（1.09）

technology students, their status as students at a “national” **国立** school suggests potential above the norm. Two of his subjects（N=48）passed the STEP Level Two test and also received a Total Scaled Score of 48+ on the SLEP® test. Although in the present study these two tests（the *EiKen* **英検** and the SLEP®）were not compared, Suzuki’s data seem to indicate a reasonable correlation. In our case, seven out of 56 students obtained a SLEP 48+, and the Mean for the group was 42.964.

Table 6

SLEP® Total Score		
	Pre-Test	Post-Test
Dates	Jul., 2003	Jan., 2005
Contact Hours		Ca. 375
N	56	56
Maximum possible (scaled) score	67	67
Test time	85mins.	85mins.
Mean	36.482 (37.411)	40.750 (42.964)
Mean gain		4.268 (5.553)
Number of students who gained		47 (54)
No change		2 (0)
Declined		7 (2)
StD	5.045 (4.203)	4.408 (3.917)
K-R21	0.34 (0.07)	0.21 (0.13)
SEM	0.674 (0.562)	0.589 (0.524)
Median score	36.50 (37.00)	40.50 (43.00)
Median gain		4.00 (6.00)
High score	46 (46)	50 (52)
Low score	21 (27)	32 (33)
(Two-tailed paired t-test) P value		< 0.0001
Effect size		0.95 (1.37)
Bias corrected (Hedges)		0.94 (1.36)

SLEP® and TOEFL® Score Equivalents

The bracketed Total Score Mean of 42.964 in Table 6 is interpreted in the *SLEP Test Manual* (1991) as roughly the equivalent of a 400 on the old, paper-based TOEFL® which according to TOEFL® descriptive literature is the approximate fault line between functional potential and non-functionality in English. My own (unpublished) equivalency formula rates a TOEIC® score of 400 as roughly equivalent to a TOEFL® score of 410.

Summary

The paper presents in two parts the English language proficiency gains of an integrated, college (two-year) program in Japan. Ten measures were looked at. The most widely-known of these are the three scores of the TOEIC®. Our goal was a Total Score of 400 or better, as recommended for two-year-college English majors by personnel managers at major Japanese corporations in a survey. The actual Total Score for the class of 2005 was 399.91, giving a bias-corrected Effect Size of 1.82. The Total Score Gain for the program was targeted at 100-150 and realized in this case at 135, for a cohort of 57 students. The Core Program which accounted for these results is described in the paper.

Acknowledgements

In 1980 at the TESOL convention in San Francisco I met the man who was to influence this program more than any other – Nicolas Ferguson of the CEEL in Geneva, who developed the methodology of self-access pair learning and the course *SAPL*. I first tried *SAPL* in 1982. By 1983, I was ready to commit to its use in our integrated program, which continued for 26 years, with peak results in 1995. Nicolas provided mentor-like support throughout, for which I am most grateful.

During this time, my greatest support was my wife Sakiko Okazaki, who did everything and more, even to the extent of helping to "coordinate" the students in our Class of 2005.

Notes

1) NRT = Norm-Referenced Tests. Brown (1976) explains that proficiency tests are one kind of NRT which "are specifically designed (to assess) … general knowledge or skills … in comparison to other students." The CEEL battery presented in Part 1 of this paper and the ETS battery presented in Part 2 are both NRT. The main point of the paper is to provide a baseline for comparison with other programs at about the same starting level to see if program differences make a difference, and if so of what kind and how much.

2) CRT = Criterion-Referenced Tests. On the other hand, "achievement tests" are the best example of a CRT, in which what is measured is how much students have learned of what has been presented in a specific course.

3) April, 1983 – March, 2009

4) I once met an instructor from another program who explained that the TOEIC® was mandatory in his program as well, but that they administered the test at the end of the second semester. The reason: experience had taught them that mean scores tended to decline the longer students were in the program after that time!

5) It is for this reason that ETS has introduced the TOEIC Bridge® test, to deal with pre-TOEIC® levels of learners. There are several problems with the situation:

   a. The Bridge® seems to be more appropriate to determine accurate levels at the beginning for low-level students.

   b. If however the program is relatively effective, the regular TOEIC® can be used at the end of the program.

   c. There is pressure on the students to provide TOEIC® scores on their resumes. This is of course a priority for the Career Planning Office at the university. At this point, anything which contributes to employment goals trumps academic concerns.

   d. The result is that there is a disconnect in choosing either to compare scores on a TOEIC® pre- and post- test or to choose the Bridge® for more sensitive analysis at the beginning and the TOEIC® as the post-test for its prestige in helping the student to find employment.

6) See, for example, (Winitz, 1981) for an article by Valerian Postofsky.

7) Effect Size (see "References" for *Effect Size Calculator*):

   0.0 indicates no (gain) effect from instruction (or anything)

   0.2 indicates a small gain

   0.4 indicates a medium gain: This is also the average gain (Hattie, 1999)

   0.8 indicates a large gain

   1.0 indicates a "very good gain" (Hattie, 1999): "An effect size of 1.0 indicates an increase of one standard deviation … improving the rate of learning by 50% … an effect size of 1.0 would mean

that approximately 95% of outcomes positively enhance achievement, or (that) average students receiving that treatment would exceed (excel) 84% of students not receiving that treatment."

Bias correction: the effect-size estimate is slightly biased and is therefore corrected using a factor provided by Hedges and Olkin (1985).

References

Alderson, J., Krahnke, K., & Stansfield C. (1987). *Reviews of English Language Proficiency Tests*. Washington, D.C.: TESOL.

Atherton, J.S. (2005). *Teaching and Learning: What works and what doesn't* [On-line] UK: Available: http://www.learningandteaching.info/teacing/what works.htm

Boyd, J.R. & Boyd, M.A. (1986). *Beginning Listening Cycles*. Normal, IL: ABACA Books, Inc.

Boyd, J.R. & Boyd, M.A. (1985). *Listening Cycles*. Normal, IL: ABACA Books, Inc.

Brown, J.D. (1996). *Testing in Language Programs*. Upper Saddle River, NJ: Prentice-Hall Regents.

Culligan, B., & Gorsuch, G. (1998). "Using the Secondary Level English Proficiency (SLEP®) Test in a One-Year Core EFL Program". Presentation at JALT '98.

Educational Testing Service (1991). *SLEP Test Manual*. Princeton.

Effect Size Calculator: Available on-line at <http://www.cemcentre.org/RenderPagePrint.asp?lin…>

Ferguson, N. (1973). *Listening Comprehension Test N73*. Geneva: CEEL.

Ferguson, N. (1980). *The Gordian Knot*. Geneva: CEEL.

Ferguson, N. (1998). *OLAF N73*. Geneva: CEEL.

Ferguson, N. (1999). *Language Teaching Theory: A Handbook for Professionals*. Geneva: CEEL.

Ferguson, N., Ferguson C, & Maire O'Reilly (1980). Geneva: *SAPL*. Castle Publications SA

GraphPad InStat version 3.0a for Macintosh, GraphPad Software, San Diego California USA, www.graphpad.com".

Hattie, J. (1992). "What Works in Special Education". Presentation to the Special Education Conference, May 1992 [NZ: On-Line, Acrobat File]: Available: http://www.arts.auschland.ac.nz/FileGet.cfm?ID=C302783E-1243-4B65-AC54-B7Fd4A5B7EF7

Heffernan, N. (2003). "Building a Successful TOEFL® Program: A Case Study." *The Language Teacher* (*JALT*), 27.8. Available online at: http://www.jalt-publications.org/tlt/articles/2003/8/heffernan

Rimer, S. (2008). "SAT Changes Policy, Opening Rift With Colleges." The New York Times. Available online at http://www.nytimes.com/2008/12/31/education

Scheibner-Herzig, G., Sauerbrey, H., & Kokoschka, S. (1991). "Repetition --- A Means to Predict Foreign Language Oral Proficiency." IRAL XXIX/3, August, pp. 230-239.

Suzuki, T. "Threshold Measurability on the Secondary Level English Proficiency Test: An Analysis from a View Point of the Criterion Validity in Reference to the Pre-2nd Grade SLEP Test Score." Asahikawa National College of Technology Research Reports, pp. 41-48 (online), 2005.

TOEIC® NEWS INTERNATIONAL, The Reporter (1991). "TOEIC® Scores Help Students Get Jobs," p. 4. No. 6, Winter. Princeton: Educational Testing Service.

Winitz, H. (Ed). (1981). *The Comprehension Approach to Foreign Language Teaching*. Rowley, MA: Newbury House Publishers.

平泉渉・渡部昇一 (1975)『英語教育大論争』東京：文藝春秋　238pp.

Johnson, J., アレン玉井光江・加須屋裕子 (1999) SLEP®テストによる英語能力測定：文京女子大学１年生の分析　文京女子大学研究紀要　第１巻第１号　pp. 141-162

Available online at: http://library1.ba.u-bunkyo.ac.jp/kiyo/1999/kyukiyo/Jeff_377.pdf

三枝幸夫 「実証されたTOEIC®受験車層の拡大増加と新入社員の実力」。*TOEIC® Newsletter*, No. 32, pp. 31-33.

黛 道子（2008）「実践報告:レベル差に応じた対応をめざして --- 2006年度多読授業の分析と考察」日本多読学会　JERA Bulletin 2008　第 2 巻第 1 号

Available online at: http://www.seg.co.jp/era/bulletins/2008-03-bulletin.pdf