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

Part 1

Thomas M. PENDERGAST

Not many English major programs in Japanese universities are centrally managed.[1a] Vague course titles[2] are determined by administrators (教務課)、teachers assigned(教務課)、and the choice of text and course content left up to the individual instructor, who rarely coordinates with other instructors in the department or even knows what they are doing.[1a] The results are well known but little documented.[1b&c]

This paper considers a two-year college English Department in which the core language program consisted of six classes a week managed by a team of instructors who coordinated small groups (3-5 students each) who worked in pairs at their own pace, revising material as necessary, and one class a week of extensive reading and/or storytelling audition. The students were extensively pre- and post- tested with standard proficiency testing instruments. Test results are presented in detail.

Key Words: language learning, self-access, SAPL, extensive reading, proficiency testing, t-test, effect size

Introduction (How do we know what "works?")

Recently, in an article in the New York Times, columnist Gina Kolata raised some questions in regard to conflicts of interest between "physical fitness" (gyms) and marketing claims (Kolata, 2009).

More recently, supplements of vitamin C, vitamin E, beta-carotene, and selenin, long thought to be panaceas for guarding against cancer and other illnesses, now seem to be considered largely ineffective, if not actually harmful (Parker-Pope, 2009).

U.S. Secretary of State Clinton is said to be a champion of "comparative effectiveness

research" (Pear, 2008), which seeks solid evidence of the value of (medical) treatment.

What of the language learning field, in which aids/approaches to language learning come and go – fashions enthusiastically touted before being forgotten or dismissed? The answer is of course *evidence* - lots of evidence - to satisfy the need for "evidence-based" research. But does evidence really trump the conviction/ideology[3] which has driven many approaches over the years (Sweet, W., et al., 1966)?

This paper (in two parts) provides ten (considering part and total scores independently) measures of proficiency in profiling the record of a centrally-managed, self-access, two-year college, English-major program in Osaka, Japan. The program was integrated,[4] the students were homogeneously non-functional[5] at the beginning, evaluation was confined largely to standardized, proficiency tests, and no attempt was made to improve student scores by "prepping" them for standardized tests.[6]

For those who use similar testing instruments, the data can serve as benchmarks for comparison. For those who do not, they may provide a paradigm which will help to bring us closer to the day when language programs will be expected to implement "comparative effectiveness research" using yardsticks common to many.

The Program

In 1983, a new English Department (英語科) was established on the campus of the then International Buddhist University (IBU, or 四天王寺国際仏教大学), formerly Shitennoji Women's College (四天王寺女子大学)、and currently Shitennoji University (四天王寺大学) in Habikino City, Osaka Prefecture.

In the preceding year, the Ministry of Education had authorized this program to accept 100+ students per year; it began in fact with 106. A surge in applications led to the authorization being doubled to 200 three years later and actual intake peaked at 279 students in 1991. The program closed this year (2009) with a final class of 46.

The chief administrator (常務理事)of the university mandated that the curriculum focus on "communication" rather than on the typical "finishing school acquaintance with English" (教養) of the time --- i.e., that it function as a true English Department (英語科) rather than as an English Literature Department (英文学科).

The main reason for this emphasis was that the administration was interested in

students' post-graduation placement (就職率), and English Literature graduates had the notorious reputation of being deficient in useful English --- i.e., "useful" in the real world of business.

Goals for the Program

Soon after the program began, a survey by the Association of Junior College English Departments in Japan (短期大学英語科連盟) of Corporate Personnel Managers (人事課長) showed that the average Manager expected (「望ましい」) a total score on the newly-established TOEIC® test of at least 400 for two-year English Department graduates. This was apparently considered the threshold of potential future usefulness. Further information from ETS (Educational Testing Services), creators of the SAT®, TOEFL®, TOEIC®, and SLEP® suggested that 500 on the TOEIC® was a minimum for English-required job assignments domestically, that 600 was a minimum for working abroad, and that 730 was the minimum score that management could feel reasonably comfortable with for those employees sent abroad.

It was agreed with the IBU administration that English Department progress would be measured by various proficiency instruments (tests), which would be administered regularly as control tests. These included the CEEL battery included in this paper, as well as the ETS battery (TOEIC® Parts 1, 2, and Total, as well as SLEP® Parts 1, 2, and Total) included in Part II of this report --- for a total of ten standardized, proficiency measures. Several years of grace were taken to establish a base for realistic goal projection. The grace period was from 1983-1988 (昭和 58～64)。

Determining Goals

The personnel managers were deferred to in settling on an average Total Score goal of 400+ on the TOEIC® as the average exit goal. The realistic fact that matriculating students' average TOEIC® Total Score had not approached even 300 in the first years (and in fact never has since) helped in deciding on an average gain of 100-150 as the aim in this area.

Homogeneous/Comparable Grouping

The class featured in this report --- the Graduating Class of 2005 --- averaged well under 300 (265) on the TOEIC® at entrance in April, 2003. For reference, a Total Score of less than 300 on the TOEIC® suggests little, if any, practical proficiency in English. The other

test given at matriculation, the N73 Listening Comprehension, averaged out at 032, where 050 is the minimum for what is termed Tourist Survival (see below). At this point, it was clear that in order to achieve the TOEIC® exit goal of 400 an average gain of 135 was necessary, and a gain of around 100 on the N73 for Social Survival level.

Inclusive Reporting

All tests were mandatory, both pre-tests and post-tests. Every student who began and finished the program sat the tests. Only those who either did not finish the program or take the final tests are excluded from the data.

Scores on standardized tests do not necessarily improve. All scores --- up, down, and sideways --- are reported in this paper --- and noted in the data tables.

Results from What?

Japan presents a formidable challenge for English language researchers. Although English is not technically a required subject in the high school curriculum (any foreign language offered is acceptable), almost all students elect and study English for six years in junior and senior high school. Many private grade schools have been offering English in at least the fifth and sixth grades for years. Parents eager to give their children an early advantage regularly send them to private English classes and kindergartens. The result is that almost everyone of college age has experienced at least 700 contact hours of school English and often multiples of that figure.

18-year-olds show various results, from under 200 on the TOEIC® to over 500 on the TOEFL®. If you want to show the result of a given approach to English learning with an adult Japanese, you have not only to isolate his/her results from contamination from other exposure during training, but also to compensate for the "activation" of any latent effect of previous exposure to the language. The data in this study simply purport to show to what extent relatively low scores can be raised in an integrated (managed), synergistic program with the emphasis on self-access learning rather than on teaching.

The CEEL Test Battery

The CEEL, located in Geneva, is the Center for the Experimentation and Evaluation of Language Learning Techniques. There, in 1973, as a control in developing a language learning program, four tests were developed to evaluate learning gains. The program was later called Self-Access Pair Learning (SAPL), which was chosen as the core of the IBU

learning program. The test content is not linked to the learning material.

Table 1

N73 Listening Comprehension		
	Pre-Test	Post-Test
Dates	April, 2004	Dec., 2005
Contact Hours		Ca. 500
Number of students	57	57
Number of sentences given	15	15
Number of items (= max. possible)	141	141
Test time	3.5 mins.	3.5 mins.
Mean	038 (032)	087 (112)
Number of students who gained	57	57
No change		0
Declined		0
Standard deviation	32.000	37.000
Kuder-Richardson21[7]	0.9798	0.9826
Median score	040 (038)	086 (107)
High	086 (107)	107 (193)
Low	012 (000)	070 (066)
(2-tailed paired t-test) P value[8]		< 0.0001
Effect size[9]		1.54
Bias corrected (Hedges)		1.53

Test #1: N73 Listening Comprehension (cf. Table 1): Cf. NOTES for details on figures)

N73 L.C. (Ferguson, 1973) measures the ratio of linguistic output to input through eliciting repetition of simple, unrelated sentences and counting the number of words successfully repeated, in any order. The sentences are on CD and were administered to multiple students at the same time by taping the responses and evaluating the taped

samples. With training, evaluators' inter-/intra- reliability is in the high 0.90s (report on tester reliability forthcoming). There are 30 sentences altogether, sufficient to discriminate native-speaker competence, but for efficiency there are cut-off points at 10, 15, and 25 sentences for lower levels. If a certain score has not been attained by the cut-off point, the test is terminated at that point.

In the current study students were given 15 sentences in both the pre- and post-tests, but the possibility is available of re-testing outlier subjects with 25 sentences when this becomes necessary, as occasionally happens with outstanding students on the post-test. With the full test of 30 sentences, (483 words/items) raw scores scale to a native-speaker/equivalent score of 1000+. The version used was American English.

N73 Listening Comprehension Scaled Score Interpretations

20	Minimum scaled score	No functional proficiency
50+	Tourist survival	Basic tourist language
150+	Social survival	Converse w/caretaker assistance
250+	Social autonomy	Converse in everyday situations
350+	Professional/Academic	Survival at school or workplace
450+	Professional/Academic	Autonomy at school/work
600+	Simultaneous interpreter	Adequate to begin training

Scaled scores are calculated up to 1000, difficult but usually achievable for a native speaker. Sentence #30, for example, consists of 33 words. It is not necessary to obtain a perfect score to qualify as a native speaker or equivalent.

For reference, the 57 students in the current study scored on average, as noted in the brackets, a scaled score of 032 (no functional proficiency) at pre-testing, and 112 (low Social Survival) at post-testing.

N73 Spoken Expression (Ferguson, 1999)

Tests #2-4 were scored using a dedicated, hand-held, digital computer, referred to by users as OLAF ("Oral Language Analysis and Feedback").

Taped samples of student speech were individually analyzed and scored. Students were given a previously-unseen, 16-frame, cartoon story sequence, allowed 2-3 minutes to put their thoughts in order, and then asked to describe what was depicted, in a 90-second

recording. They were allowed to continue to look at the pictures in order to mitigate stress due to lapses in memory or lack of imagination. The pictures portrayed everyday situations in the life of a family and required no extraordinary vocabulary to narrate. The same cartoon sequence was used for both the pre- and post- tests, separated by 18 months.

Table 2

N73 Speaking (OLAF) Fluency		
	Pre-Test	Post-Test
Dates	July, 2003	Dec., 2005
Contact Hours		ca. 380
N	57	57
Maximum possible score	N.A.	N.A.
Test time	90 secs.	90 secs.
Mean	7.702	10.965
Number of students who gained[10]		47
No change		5
Declined		5
Standard deviation	2.686	2.584
K-R21[7]	N.A.	N.A.
Median score (tone groups/minute)	8	11
High	18	16
Low	3	6
(2-tailed paired t-test) P value[8]		< 0.0001
Effect size[9]		1.24
Bias corrected (Hedges)		1.23

The samples were analyzed and scored on three linguistic criteria (*fluency*, syntactic *correctness*, and level of *expression* or communication), considered as three separate tests in this paper.

Test #2: Fluency (cf. Table 2)

Contact hours are fewer than in Test #1 because N73 Listening Comprehension was

administered at the beginning and end of the program. Spoken Expression pre-tests #2-4 were given at the end of the first semester and post-tests at the end of the program.

The *Mean* is essentially the number of tone groups (utterances) per minute. A timing mechanism in the OLAF computer computes this figure per minute even though the duration of this particular test was exactly 90 seconds. For reference, a native speaker or equivalent might produce 18-24 tone groups per minute on similar content.

The figures for the rubrics of *No change* or *Declined* in Tests #2-3 do not take into account the fact that the subjects' *Balance of Expression* (the ratio between Fluency and Correctness) often improved. This is a subject for another paper, but in fact nine out of the ten Fluency cases and six out of the eight Correctness cases improved the Balance of their Expression. As can be seen in Table 2, the average mean of the group as a whole moved slightly more than one standard deviation from Low Correctness to Balanced.

Test #3: Correctness (cf. Table 3)

The *Mean* is a percentage score obtained from a computerized analysis of the test sample by a trained evaluator who rates each tone group at one of four syntactic (S) levels:

a. S_c syntactically conventional
b. S_3 *non sequitur*
c. S_2 corrupt tone group
d. S_1 corrupt stress group

The analysis is explained extensively in Ferguson (1998) and to some extent in Pendergast (1985). A day of training with the computer and its manual will in most cases enable reasonably accurate evaluation of recorded samples.

Although any stimulus may be used for the language sample elicitation, for consistency and simplicity we used a 16-frame cartoon segment showing a day in the life of a family. The students were given exactly 2.5 minutes to look over the pictures and get their thoughts in order. In principle, no questions were taken in regard to the picture sequence and each student interpreted them as he/she saw them. Students were encouraged to say at least one thing about each picture but to avoid periods of silence by skipping a picture if necessary. From experience, we found that almost all students would still be speaking

after 90 seconds, which we set as the cut-off point. As with a urine test, it is best to arrange for the cut-off to occur in mid-stream.

Table 3

N73 Speaking (OLAF) Correctness		
	Pre-Test	Post-Test
Dates	July, 2003	Dec., 2005
Contact Hours		Ca. 380
N	57	57
Maximum possible score	100(%)	100(%)
Test time	90 secs.	90 secs.
Mean	41.895	58.351
Number of students who gained		49
No change		0
Declined		8
Standard deviation	13.045	15.603
K-R21	0.8656	0.9009
Median score (%)	41	60
High	63	86
Low	17	14
(2-tailed paired t-test) P value		< 0.0001
Effect size		1.14
Bias corrected (Hedges)		1.14

Again, *Balance* is an important factor. Very low fluency puts the auditor to sleep (Six tone groups a minute is only one utterance every ten seconds … !).

Hesitation or faltering and repetitive speech, however *Correct*, may be fatally distracting, as the listener's attention wanders and fails to note the following tone group.

On the other hand, *Correctness* under about 35% is gibberish. Gibberish combined with high *Fluency* is worse gibberish. Etc.

One notable aspect of the program: there was only one formal grammar class (an elective --- i.e., not part of the core program of required courses), but both *Correctness* and *Balance*

improved markedly over the period of 18 months (see Table 3 for *Correctness*).

Test #4: Expression (cf. Table 4)

Expression is derived by a formula built into the OLAF computer using the *Fluency* and *Correctness* data and can be understood as the amount of information successfully transmitted in a given amount of time. The maximum score is 1000, which represents a number achievable by any native or equivalent speaker given a simple task of narration.

Lower levels are rated on the same scale as the N73 Listening Comprehension test and present a reasonable approximation of levels required in various professional and academic environments …

Tourist (観光客)	050
Sales Clerk, waiter, waitress (店員、ウエーター)	150
Airline cabin attendant, receptionist, Hotel front desk (乗務員、受付、ホテルのフロント)	250
Lower level language teaching (幼児、小・中学校の英語教育)	300
Tourist guide University studies in U.S., U.K., N.Z., etc. (ツーリストガイド・正規留学)	350-450

Table 4

N73 Speaking (OLAF) Expression		
	Pre-Test	Post-Test
Dates	July, 2004	Dec., 2005
Contact Hours		Ca. 380
N	57	57
Test time	90 secs.	90 secs.
Maximum possible score	1000	1000
Mean score	92.140	180.820
Mean gain		88.680
Median gain		86
Number of students who gained		55
No change		0
Declined		2
Standard deviation	42.593	74.696
K-R21	0.9556	0.9744
Median score	80	166
High	254	400
Low	35	48
(2-tailed paired t-test) P value		< 0.0001
Effect size		1.46
Bias corrected (Hedges)		1.45

Conclusion

The data from the graduating Class of 2005, which was trained in a centrally-managed, integrated, self-access program, reveal average beginning levels of Non-Functionality in Listening and Minimum Tourist Survival in Spoken Expression. All four skills tested (Listening, Fluency, Correctness, Expression) show statistically impressive gains (t-test P values = < 0.0001 and Effect Size greater than 1.0 in all four cases) by the end of the program. Two problems remain:

1. This is a benchmark effort to show how well a given group did in a given program. In fact, this was not one of the most successful classes in the

program in terms of final scores, but was interesting for the fact that it was our first class to accept male students, resulting in a number of social and behavioral difficulties. The turn-around in results was striking. In addition, the ratio of students to instructor was the greatest in our recent experience, at 47:1. The question is: how would these data compare with those of other programs of a similar beginning level, using the same instruments of "comparative effectiveness research." What other instruments could be used and what variables noted? Is there a difference? Enough of a difference to make a difference?

2. Despite their statistical gains, the average final levels of the students in this study are still at the threshold of useful language. The program has at least brought many of the students to their "Rubicon" of choice: either to be satisfied with minimum functionality for tourist or social purposes or to develop their skills to professional-level competence after graduation.

Notes

1. Personal communications:
   a. Part-time Instructor trying to discover from department head what book to use, what to teach, etc.: "He says it's up to me."
   b. Tenured instructor in charge of TOEIC® data: "We only report the gains. More than two-thirds of the scores decline after two years."
   c. Tenured instructor in charge of TOEFL® program: "I only report on students who study. Most don't."
2. E.g.,: 英語、英語演習、英語学、コミュニケーション、リスニング，など
3. Shiono Nanami, in her volume on Caesar Augustus, says in the book's introduction that her favorite saying of Julius Caesar's was that people seldom see any reality but the reality that they wish to see (Shiono, 1997)。
「人間ならば誰にでも，現実すべて見えるわけではない．多くの人は，見たいと欲する現実しか見ない」
4. Integration: the program was managed in the following ways:
   a. The core program, in which all students participated, consisted of seven classes per week, of 90 minutes each. The Class of 2005, considered in this paper, had approximately 360 such classes over two years, or about 540 possible contact hours. Attendance was approximately 96% in the first

year and 90% in the second year (2nd -year students need to search for post-graduation employment and consequently miss some classes.).

b. Six of the classes per week were conducted using a method (SAPL) which provided students with self-access material (books and tapes) and training in how to help each other to learn (Ferguson, 1980). The seventh class was nominally an extensive reading class. It goes without saying that reading is a self-access activity as well, although in many cases students were asked to do their reading outside of class and in fact listened to stories in class.

c. Instructors in the non-reading, self-access classes received an intensive five-day introduction to the methodology and method, coordinated their work through a logbook and daily discussion, and in some cases opted for further in-service training. Keeping students on task and involved was simplified by the fact that many students took major responsibility for their learning, allowing instructors to function primarily as coordinators rather than as teachers.

d. Program testing was for the most part standardized, proficiency testing as reported in this paper.

5. Non-functional: For our students, 450-500 contact hours seems like an inordinate amount of time. To put this into perspective, consider that (unpublished) annual surveys of students in this program reveal that the average student had invested roughly 1150-1300 hours (grade school, junior and senior high school, cram school, miscellaneous) in English study prior to enrolment.

In our program, the average TOEIC® Total Scores at matriculation never exceeded 300 --- in 25 years! The group studied in this paper averaged 265.

By coincidence, over a period of five years in the late 1990s, entering students were also administered the TOEFL® at the behest of the school administration. The average Total Score was 363. The ETS manual states that < 400 is “meaningless.”

The N73 Listening Comprehension scaled score at matriculation remained, from 1983-2007, in the range of 27-32, far below the lowest meaningful level of < 50, suggesting no functional proficiency.

These are the results (for these students) of an average of 1150-1300 secondary-education English contact hours prior to matriculation in our program.

After a further 500 hours, the group studied reached an average of roughly 400 on the TOEIC®「望ましい」, or “a suitable level company entry level for English Department

graduates"), roughly 112 on the N73 Listening Comprehension (minimum Social Survival), and 188 on OLAF Expression (Social Survival). These scores are on the threshold of functionality.

Many of these students then graduated and did nothing further linguistically.

Others went on to achieve scores of 800-900+ on the TOEIC® and 300-400+ on the N73 and OLAF Expression, scores which qualify them for further schooling (some transfer to and graduate from foreign universities) and the public sector (some find employment abroad or domestically with international corporations).

For one perspective on the time it takes to learn a language, consider Table 11, pp. 250-251, of Cleveland, at al., (1966). These two pages present schemata purportedly drawn from a U.S. Foreign Service Institute study showing the relative difficulty for foreign service officers, rated in study hours, of learning foreign languages.

Notable are the following:

a. "The estimates … are based on the assumption that students possess no less than average aptitude and positive motivation."

b. The students in this paradigm are of course not our 2-year college students but officers of the select U.S. Foreign Service.

c. The minimum goal is "sufficient proficiency in speaking a foreign language to satisfy routine travel requirements."

d. This minimum goal for Americans learning Japanese requires on average and for (foreign service officer) students with "average aptitude" …

4-6 class hours per day
plus 4-6 hours of drill and study a day
for 4 months (roughly 90 days at 10-12 hours per day)

e. The intermediate goal is "basic familiarity with the structure of the language with sufficient proficiency in speaking to conduct routine business within a particular field."

f. This goal requires FOUR TIMES the investment of the minimum goal, or roughly 4,000 hours. It is fair to consider that this level is what the average corporation in Japan would consider 即戦力, or "an immediately useful level of English." For most of the students in our study, the program was too little, too late. Language learning, like the "fitness" (Kolata, 2009) mentioned in the first paragraph, takes time and effort.

6. Test prepping: there are reports in the literature (Heffernan, 2003) of notable results

obtained from prepping students for exams. This practice is of course endemic in Japan where the 予備校 *yobikoh* cram schools are known for preparing students not only in general, but even and especially for entrance examinations for specific universities.

In the program studied, it was decided not to do any specific exam preparation.

7. K-R21 (Kuder-Richardson formula 21):

The K-R21 Internal Consistency Formula is applied to estimate the consistency or stability of test results, i.e., whether or not the results would be similar with a second testing. The degree of test reliability is indicated by a reliability coefficient, which can go as high as +1.0 for a perfectly reliable test --- the higher the better (Brown, 1996).

8. P Value: "P" stands for "probability." The result of a "two-tailed t-test" (as well as other tests) is often presented as the "p-value" (e.g., $p < 0.05$ or $p < 0.001$, etc.). The "<" symbol means "less than." The t-test is a calculation applied against the means of the pre-trained mean (or "average") and the post-trained mean and compared to see if the difference *makes a difference* statistically, and *how much of a difference*. It is as if you made a bet that any change between the control (pre-test) group and the treatment (post-test) group were due entirely to chance. If you bet that this were in fact the case, you would be accepting what is called the "null hypothesis" (no significant difference due to training). If you show statistically that the treatment/training had some effect, you reject the null hypothesis. Conventionally, values less than 0.05 are taken to cast doubt on the null hypothesis. The smaller the P Value, the more robustly you reject it.

In this study, a "two-tailed t-test" was performed on the four measures of the CEEL test battery, with software created for the purpose. In all cases considered in this paper, the P Value was < 0.0001. The InStat[12] verbal description for this level of rejection of chance is "extremely significant". For this software, there is no higher level of rejection of the null hypothesis, or, in other words, of affirmation that the training was "extremely significant" statistically.

9. 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).

10. Students who gained, etc.: under the Speaking Fluency, Correctness, and Expression rubrics in the data tables, it can be noted that some students either did not improve or actually declined. There is a further aspect of this calculation which is too complicated to consider adequately in this overview of the situation. The problem is one of "Balance". Some scores show either high fluency with low correctness, or the reverse. Bringing these into *balance* is tantamount to a legitimate proficiency gain, and requires further calculation and explanation. In the present study, nine students out of the ten whose Fluency either did not change --- or even declined --- compensated by improving their Correctness. Of the eight students whose Correctness declined, six improved their Fluency and therefore their Balance. This is partly reflected in the (Spoken) Expression data.

11. Statistical analyses are most useful when observed differences are small ... If you only care about large differences, heed these aphorisms (from the Instat GraphPad manual):

> *If you need statistics to analyze your experiment, then you've done the wrong experiment.
>
> *If your data speak for themselves, don't interrupt!

But in many fields, scientists care about small differences and are faced with large amounts of variability. Statistical methods are necessary to draw valid conclusions from these data.

12. Two-tailed, paired t-tests were performed using GraphPad InStat version 3.0a for Macintosh, GraphPad Software, San Diego California USA, www.graphpad.com.

Acknowledgements

The administrators who made this program happen over 25 years: Deguchi Juntoku (former Chief Abbot of Shitennoji Temple, Osaka) and Okuda Kiyoaki (former President of IBU, current Chief Abbot of Shitennoji Temple, Osaka). The coordinators and evaluators who were in the front lines: Nadia Helwig, Sep Overlaet, Derek Phillips, Ted Sanders (R.I.P.), and many others. Special thanks to Extensive Reading and Storytelling instructor Dr. Beniko Mason, who alerted me to the importance of statistically supporting what to me seemed so obvious.[11]

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/teaching/what works.htm

Brown, J.D. (1996). *Testing in Language Programs*, pp. 197-199. Upper Saddle River, NJ: Prentice-Hall Regents.

Cleveland, H., Mangone, G., & Adams, J.C. (1960). *The Overseas Americans*: *A Report on Americans Abroad*. New York: McGraw-Hill.

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

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.

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

Kolata, G. (2009). "Fitness Isn't an Overnight Sensation." NYT Online (January 21, 2009).

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

Parker-Pope, T. (2009). "Vitamin Pills: A False Hope?" NYT Online (February 17, 2009).

Pear, R. (2009). "U.S. to Compare Medical Treatments." NYT Online (February 16, 2009).

Pendergast, T.M. (1985). "OLAF N73: A Computerized Oral Analyzer and Feedback System." In *New Directions in Language Testing*. Lee, Y.P., Fok A., Lord R., & Low, G. (Eds.). Oxford: Pergamon Press.

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.

Sweet, W.E., Swan Craig, R., & Seligson, G.M. (1966). *Latin: A Structural Approach*. Ann Arbor: The University of Michigan Press.

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

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®受験者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