Proficiency Testing and Gain Scores: A Research Overview and Use of the TOEFL at One Japanese College

David Schneider[※]

Introduction

In a recent publicaton, the Educational Testing Service (ETS) compares and contrasts five widely administered English language proficiency tests. Among these tests is the Test of English as a Foreign language (TOEFL). A list of five uses for the TOEFL is provided, including placement in language programs and as an exit test. In fact, many programs do use the test for just these purposes. In some cases, placement or entry scores are compared with exit scores and a raw gain score is calculated. This number may have varying degrees of influence on educational decision-making. Teachers in English language programs are likely to view gain scores as one factor among several in assessing student progress. They might, for example, also look at scores from tests related directly to class-work, student motivation, attendance, participation, or overall learner improvement. On the other hand, it may be tempting for administrators to look at gain scores apart from any real context. Reasons for this could include lack of time, more interest in items such as pass/fail rates, or general non-understanding of the use and meaning of testing in language programs. In addition, a single number metric has appeal as something "concrete." When thought of in these terms, this "concreteness" may be wrongly used to make judgements about degree of student improvement, competence of teachers as a group, or even overall program effectiveness. As opposed to a single test result, these kinds of assessments should surely be made on the basis of measurable outcomes clearly linked to curricular goals.

Under the circumstances described above, it becomes critical for all concerned to have a shared frame of reference for understanding proficiency test gain scores. Earlier research has focused on such questions as the following:

1) What is the influence of coaching, or teaching for the test, on TOEFL scores?

2) What factors explain student gains or losses on the TOEFL?

3) When gains or losses occur, what do they really mean?

※ Aomori Public College

4) Is the TOEFL a good measure of short-term language proficiency gain?

Keeping these questions in mind, the next section will review earlier work devoted to the topic of proficiency test gain scores. This overview will be followed by a report on gain scores and the TOEFL at one Japanese college.

Background – Gain Score Related Studies

Before referring to earlier studies, several points are worth noting. English language proficiency testing in general (TOEFL gain scores in particular) is a rather messy business; locating information, and sifting through and assessing it is not easy. It appears there has been very little previous research done on the specific topic of TOEFL gain scores. However, there is some supplemental gain score related material from other tests, such as the Test of English for International Communication (TOEIC). Because this non-TOEFL material is applicable to the TOEFL itself, it is included in the discussion. Second, latebreaking changes in TOEFL format, namely the switch to a computer based system, may have impact on the relevance and accuracy of some of the points that follow. For purposes here, it should be assumed the focus is on the paper and pencil Instutional Testing Program (ITP) TOEFL.

Influence of Washback

The literature from gain score connected sources addresses several of the issues surrounding the four focus questions in the introduction. First, there is the business of the influence of testing on what is taught in the classroom. This "teaching for the test" and how it affects learning, called "washback", is a subject that serves as a good starting point for gain score discussion. Gates (1995) explains the two types of washback, positive and negative. Course goals in tune with what is to be tested foster positive washback; when goals and tests don't match the washback will be negative. There is also commentary on possible adverse effects of making a course/curriculum change, but not corresponding test changes; the reverse is also to be avoided. Gates provides a useful list of factors that make standardized tests like the TOEFL strong in washback. For Japanese students, test prestige and format (multiple choice) are important sources of positive washback. He concludes that because standardized testing is hardly about to disappear, it is best to try and offer students a mix of test prep and other activities that meet more general language needs.

Wadden and Hilke (1999) and Hamp-Lyons (1996) have engaged in a spirited debate on several washback related issues. Hamp-Lyons questions the ethics of even conducting test prep courses, claiming such classes distort the normal learning process. She is not convinced, either, that test coaching meaningfullly raises TOEFL scores. In support of her stance, she says she is not aware of any research that proves the point. Going further, she seems to suggest even if scores can be raised, the increases are suspect in terms of true learner progress:

"English language programs .... treat TOEFL scores from different countries differently. Most common among these differences is the tendency to take TOEFL scores from Japan .... with a vey large grain of salt. The folk wisdom holds that TOEFL scores from Japan tend to be 20-30 points above the test taker's actual ability. This seems to be about the amount that practice and preparation materials can raise scores without boosting mastery of the language." (page 273)

Wadden and Hilke counter that data in support of test prep materials and classes raising TOEFL scores is available at schools and from test practice companies all around the world. The problem, they say, is that to this point, nobody has made the effort to crunch the numbers. They also point to an ETS study, in which Wilson (1987) found "inconclusively" that test scores often went up for test takers who simply retook the test. Lastly, the two researchers refer to an ongoing study of their own. They state a 20 hour TOEFL prep class targeting specific skills and vocabulary, plus "language activation," resulted in 200 students gaining an average of 65 points. The data from this study is apparently yet to appear.

A final research study concerning washback involves the Test of English for International Communication (TOEIC). While the TOEIC and TOEFL differ in several aspects, it is reasonable to assume there are enough similarities between the two proficiency tests to make research into one applicable to the other (see Gilfert, 1995). Robb and Ercanbrack (1999) address the specific question of whether or not Japanese university students who use materials designed to produce TOEIC test gains actually improve their scores. In their review of past studies, the authors echo Hamp-Lyons by stating there is little or no research into gain scores on language proficiency tests. Some commercial testing companies have touted their coaching systems by offering examples of big increases in student scores on other types of standardized tests. The authors, however, cite research that shows large variations, both up and down, are common for students who take the tests several times. Further, these fluctuations may well have more to do with statistical features such as regression to the mean and standard error of measurement rather than teaching to the test. The authors make the point that research has important implications for the validity of language proficiency tests; if it can be demonstrated that preparation materials and techniques help raise test scores significanly, then the tests become more an exercise in "test-wiseness" than a measure of overall language proficiency.

Robb and Ercanbrack note that there are several research studies into one particular standarized test in the U.S., the Scholastic Aptiude Test (SAT). The test is taken by many high school students, usually seniors, and the scores are often one important component universities and colleges use for admission purposes. The data from SAT research suggests that test coaching has very little effect on gain scores. In looking for an area of test coaching that might explain modest increases in SAT scores, the authors cite one study (Johnson, et al., 1985) that indicated test coaching helped examinees to work faster and increase the number of items they were able to complete. This ability would be helpful on an SAT style test where scores are tallied on the basis of the number of correct items and wrong guesses don't matter.

In their actual study, Robb and Ercanbrack set up two groups of freshmen students from a Japanese university. One group consisted of English majors who had seven hours of 90 minute English classes each week. A second group of non-majors were in English classes for two 90 minute classes per week. These first two groups were divided once more into 8 treatment sub-groups each. The classes for all students had reading and listening components, but the way of teaching the material varied. Some groups, both majors and non-majors, had an exclusively TOEIC treatment, some a business treatment, others a general treatment. Regardless of treatment, all students were told the class objective was to increase TOEIC scores. Within this framework, the two researchers hypothesized that class treatment would have no impact: gain scores would be equal for all 16 student sub-groups.

Data analysis, a major part of which compared reading and listening TOEIC sub-set gains, confirmed the Robb/Ercanbrak hypothesis, with one exception: The non-major students given the TOEIC treatment had significant gains on the test's reading section, while non-majors exposed to the business and general treatments did not. The authors conclude that there may be some value in using TOEIC materials for raising TOEIC reading sub-section scores of non-majors of English, at least in one Japanese university setting. The conclusion is a very tentative one. The writers point out, for example, the fact that one group of non-majors increased reading scores dramatically and no major group did, might be a result of the non-majors' much lower initial sub-section scores. Finally, students studying English more intensively appear not to have benefitted at all from TOEIC prep materials.

To this point, then, it can be stated that researchers agree that washback is a very real phenomenon that can have a positive or negative effect on classroom instruction. The influence of washback on gain scores is much less clear. Research into standardized tests other than the TOEFL indicates slight gains may be attributed to test coaching, but the connection is not definite.

Ups and Downs – What do they mean?

This section offers discussion on the question of interpretation of proficiency test score increases and losses. To begin, De Avila (1997) presents some basic, but key concepts in relation to expected gains on standardized tests. His work actually involves bilingualism and English language instruction for non-native school children. Nonetheless, his points are relevent to the topic at hand and include the following: To be a fair measure of student progress, expected gains on standardized English tests depend on the tests being consistent and matching what they measure. Second, progress benchmarks should be reasonable. This means score growth is linked directly to where students enter a language program, i.e., their initial test score. Students with lower entering tests will show larger increases over the short term than students on the upper end – there are diminishing returns on the overall learning curve and a meaningful metric needs to account for incremental progress at differing proficiency levels. Lastly, the writer cites research indicating the necessity of using expected gain as a measure of group performance, not individual assessment (except in terms of probability value).

With the De Avila material as a backdrop, studies by Childs (1995) and DesBrisay and Ready (1991), are well worth noting. The Des Brisay/Ready research is apparently one of the only studies that deals specifically with TOEFL gain scores. The writers introduce their work by stating that test score interpretation is very much an inexact, complex business and needs to be looked at with extreme care. The subjects in this study were Indonesians whose basic goal was to achieve a 550 TOEFL, which would enable them to to study in mainstream North American university classes. 129 students were divided into 8 classes; the classes were partly designed to raise TOEFL scores, so gains on the test were an essential class goal. The researchers wanted information on the questions of what kinds of gains were reasonable, and who and how many of the examinees would be successful in reaching the 550 benchmark.

The 8 classes were put into 3 separate study centers. Although student assignment to the classes was not random (the formation of classes was based on groups that had been studying together in different ESL or EAP [English for Academic Purposes] classes), entry TOEFL scores showed similar range. The entry TOEFL was given in July and, following 12 weeks (15 hours a week) of course work, again in October and January. The centers were allowed leeway to spend differing amounts of class time on TOEFL preparation, which they did. The mean gain for all groups together was 34 points. Standard deviations for the October and January tests were 32.6 and 27.0. In an effort to account for such individual variation, differences in gain scores were measured against length of study in the program (students had entered into the program at different times) and entrance proficiency (entry TOEFL score). As might be expected, students who had entered the program the earliest as a group, and with the lowest mean scores, had the biggest increases between the October and January TOEFLs. Similarly, the lowest starting group based on entry score also had the largest increases, between July and January. And again, it must be remembered that the gain score diifferences in group levels point up what has been noted earliier: incremental improvement cannot be measured in equal terms on the learning curve.

The researchers also look at the variable of teaching method/class content and conclude that differences are not statistically significant. Variability of individual scores makes apparent group score differences largely insignificant. The writers make the point that better baseline data (the students in the study took different institutional TOEFLs at different times to establish the "July" entry score.) would have been useful. They conclude that the statistics in their study might be interpreted differently, as there were differences in center test scores and, overall, mean scores did go up 34 points. But they also are clear on the fact that deeper analysis of data shows none of the center and/or class gains was found to be significant. Thus, what to make of the results may be determined by how much the reader knows about statistics.

Finally, the study results are looked at in terms of impact on language programs. Key concepts are put forward:

1) "....... general proficiency tests are not appropriate for measuring gains over short periods of time ...." (page 22)

2) "The poor performance of these 129 subjects on the January TOEFL offers a compelling argument against the use of a norm-referenced general proficiency test to measure achievement in an academic skills program. You recall that there was a group gain [from October] of only 6.2 points and perhaps the most striking finding in the study is the large number of students (48 out of 129) who had lower TOEFL scores in January than .... in October, something that cannot be explained satisfactorily by referring to standard error and regression toward the mean or standard error." (page 22)

3) "Even in studies where differences in gains .... can be shown to be statistically significant it is difficult to show causal relationships." (page 22)

The previously mentioned work of Childs (1995) deals once more with the TOEIC. In relation to the test, the writer investigates several questions: Does the test adequately measure group mean gains and are any differences significant? How is progress measured when test scores, particularly on the final test, go down after students have just finished a course that should help on the final? Can we measure the SEM (standard error of measurement) and how are we to interpret it? How should we assess extreme individual learner increases and decreases? Does putting students in ability groups make a difference in score fluctuation?

Childs begins by drawing a distinction between how the TOEIC is meant to be used and how it is actually used in Japan. He notes that the TOEIC is a norm-referenced test and not designed to measure how well examinees have learned material in particular courses or classes. Still, he states the test is often used inappropriately as a criterion-referenced language gain measurement in Japanese businesses and educational institutions. Teachers have to simply deal with reality and accept that adminisrators in particular are often going to look at TOEIC scores in this fashion.

This 1992/93 study, conducted over a period of 6 months, involved 113 new company employees who tested at a beginning mean of 269, a score Childs says is 100 points below the average for "all recent college graduates" in Japan. Overall, the students had 53 hours of general English instruction. 37 hours of classes were given over a period of 5 days. The remaining 16 hours came in four half day 4 hour classes. Four TOEIC tests were given: one at hiring, one after the five day class, and one after each of the 4 half day classes. The mean scores on the tests were 269 – 341 – 322 – 326. The raw gains on the test were decent, based on improvement over the initial score.

After subjecting the data to rigorous statistical evaluation, Childs attempts to answer the questions he posed in the beginning of his paper. First, he concludes the TOEIC is not reliable for this group of learners. This finding is based on a low (.57) internal consistency reliability estimate, a 43 point standard error of measurement, and differences in actual and proportional scores. Concerning measurement of group mean gains, the differences on the whole were judged to be significant, in so far as they stood outside the SEM. Once again, there is no exact way of accounting for the gains. Possibilities include teacher instruction, student motivation, degree of opportunity to use English, and the "polish" effect . (Note that the highest group mean score was on test #2, given just after students had finished their 37 hour 5 day intensive classes. Thus, their English skills, in the short-term, can be said to be at a relatively high, or “polished” level.) The 56 point gain between the initial and final TOEIC scores could be viewed in terms of one hour of study equals a one point gain. But it must be remembered that sound gain analysis depends on knowing where the learner starts on the score continuum. As we have also seen, improving at the high end is much more difficult than at the lower end. The rule of thumb appears to be that raising a TOEIC score one point at the 700 level requires 2.5 – 3.0 class hours. In any case, Childs concludes the overall mean increase is a combination of the four possibilities mentioned above; the exact reason(s) for the gain cannot be pinpointed.

Concerning big jumps in individual scores, up or down, Childs feels they are one good argument for the TOEIC being a poor gauge of single student progress. He notes that for this study the SEM was in range of proficiency gains. It is therefore problematic to judge individual learner improvement. He finishes with some thoughts aimed at administrators. Assessing student progress on a raw gain basis makes for poor use of the TOEIC. It is far better to have a criterion-reference based test, or tests, in accord with program goals and methods. The TOEIC should be used in the context for which it was designed: as a general look at how learners compare in a global sense.

Two other projects, those of Swinton (1983) and Gradman and Hanania (1995) are also worthy of mention. Gradman and Hanania worked on a study directly linked to variables that might raise TOEFL scores. They investigated the language backgrounds of some 100 students enrolled in an Intensive English Program at Indiana University. The students came from a variety of differnt language groups. They were first given an oral questionnaire to obtain language background information in four broad categories. This enabled the researchers to eventually generate a collection of 44 background variables that might affect scores on the TOEFL. The original variables were subjected to further statistical analysis: 1) pair correlation for TOEFL scores and all variables 2) multiple regression analysis, which helped trace the most significant variables 3) a procedure called "path analysis," designed to distinguish direct from indirect effects. The results indicated two items as having a significant effect on increasing TOEFL scores: extensive outside reading and teachers with a clearly strong understanding of English (in most cases, this would probably be a native speaker, though the study did not include this exact variable). Lastly, TOEFL scores can be helped indirectly by use of oral English for communication in and out of the classroom: such a use has a positive impact on extensive reading. It is perhaps useful to add that the researchers do not present any TOEFL data to support their conclusions.

The final material to be discussed comes in the form of a "manual" rather than a research study. It is worth going over because it offers a somewhat different gain score approach. Swinton has written a guide to interpreting student gains on tests in an ESL context. Though he is thinking in terms of the overall ESL testing picture, he uses the TOEFL as an example to make his points. He first reviews some basic statistical concepts. Then he goes on to offer a method of taking raw gain score data and using it to make some predictions about expected student progress. Specifically, he deals with raw gains in relation to "regression to the mean." The idea behind the whole project was to examine the regression effect using TOEFL data to separate regression from real language growth. To see the distinction, Swinton's approach is to use a reliability estimate "to predict the expected final score for each pretest score under the assumption that no real change has taken place, and call only observed discrepancies from this predicted score "change." (page 15) Thus, in order to account for measurement error and practice effect, Swinton suggests use of this baseline data as a more accurate measure of change than with raw scores.

In actual application of his system, Swinton notes that three tests must be administered, including a pretest and reliability test within a one week period in order to establish the no-change baseline. The baseline results yield a number that can then be compared to the posttest score.

Research Summary

What conclusions can be drawn from the body of research encountered here? The question might be dealt with by listing issues about which the researchers agree and disagree.

Areas of general agreement

–Score growth on TOEFL depends on where students enter a language program: the initial score. Lower initial scorers as a group will register mean gains greater than higher initial scorers.

–Because of the above, in terms of assessment, the same metric cannot be used to measure the proficiency test gains of students at different levels.

–Group tenden cies in test scores cannot be applied to individual scores, except in terms of statistical probabilities. Similarly, group gains can be projected at times, individual gains cannot. (Swinton gets around this difficulty by separating raw gain into regression and true gain. In accounting for the influences of test unreliability and practice,
his formula enables a degree of individual score projection).

–Tests need to be consistent with specific program goals; norm-referenced proficiency tests like the TOEFL/TOEIC should not be used to measure progress in the same way a final exam for a particular class measures mastery of formal instruction. Measure of class specific knowledge should be done using criterion-referenced tests.

–TOEFL and TOEIC measure receptive skills, not productive ones.

– Using the results of any single test for high-stakes eductional decisions is risky.

–Administrators and teachers should interpret research data with extreme caution -- as with single tests, no single research study can completely answer all questions related to a particular topic.

–More research is needed into all aspects of language proficiency testing.

Areas of disagreement/uncertainty

–What, precisely, should be done to raise short-term TOEFL scores?

–The whole question of washback and how (or to what extent) proficiency testing makes it positive or negative.

–Can the specific factors for proficiency test mean scores going up or down be isolated? (Range of possibilities: test preparation, student motivation, "test-wiseness," formal instruction, chance/luck, statistical factors -- regression to the mean, for example).

–Data interpretation -- What do score gains on the TOEFL really mean? (Role of statistical sophistication).

With the above summary material as a backdrop, we are now ready to turn to a research study of TOEFL test score use within the context of one Japanese college.

Context for Study

For the past several years, the English language curriculum at Aomori Public College, in northern Tohoku, Japan, has been undergoing a series of reforms. Amid a sea of change, administrators have been unwavering in one point: the use of TOEFL gain scores as a key element in program goal setting and evaluation. Accordingly, taking us back to several points in the introduction, the purpose of this study is to examine some questions surrounding the TOEFL in one Japanese setting: Is the TOEFL a good measure of short-term language proficiency gains for this particular group of learners? Is the TOEFL a good measure of short-term language proficiency gains for individual learners in this setting? How are we to understand the role of certain statistical features, such as the standard error of measurement, in relation to TOEFL gain scores?

Methodology

The subjects in this study were three separate groups of Japanese college students who entered school over a three year period, from 1997--1999. The total population numbered 742; 270 from 1997, 282 in 1998, and a 1999 group of 290. All three groups were required to take a series of ITP TOEFL tests. The '97 and '98 groups took two TOEFL tests each, while the '99 class sat for three.

Students were all enrolled in required EFL classes which met for 60 minutes, four times a week. With one exception (an April '99 TOEFL) the tests were taken toward the end of a school term, after some 40 hours of instruction. Average class size was around 30.

In general, the curriculum during the three year period at issue was in a constant state of change. The key elements involved made up a somewhat uneasy mix of "critical and analytical reading," speaking, and TOEFL preparation. The five full-time faculty had a fair amount of autonomy in the classroom. TOEFL prep was seen as highly important by some, less so by others. Those who were TOEFL supporters took more class time for TOEFL specific work. Throughout this time, it can be said there was pressure – sometimes subtle, sometimes not – from administrators to get scores moving up.

The mean TOEFL scores for first year students, 1997 – 1999, are given in Table One below.

AN EXAMINATION OF REPORTED TOEFL SCORES 1997 -- 1999

TABLE ONE: Overall Results for First Year Students on TOEFL Total and Part Scores (n = 842)

1997 [n=270 / 270]

	Listening	Structure/Grammar	Reading	Total Score
JULY	39.05	41.73	39.04	399.38
	(4.12)	(5.48)	(5.19)	(35.62)
DEC.	39.94	42.14	42.55	413.94
	(3.74)	(4.99)	(5.54)	(33.27)

1998 [n=282 / 282]

	Listening	Structure/Grammar	Reading	Total Score
JUNE	38.79	39.31	39.57	392.62
	(4.43)	(5.34)	(5.57)	(34.63)
NOV.	39.70	42.64	43.65	419.94
	(3.75)	(4.70)	(5.03)	(33.89)

1999 [n=313 / 290 / 290]

	Listening	Structure/Grammar	Reading	Total Score
APRIL	38.48	40.17	36.72	384.55
	(4.18)	(5.25)	(5.69)	(35.71)
JULY	40.31	39.95	40.83	403.73
	(3.73)	(5.39)	(5.08)	(37.01)
DEC.	39.28	42.10	40.53	406.84
	(3.84)	(5.93)	(5.60)	(39.30)

(—) = sd (standard deviation)

A Look at Group Gains

Looking at mean test scores of dependent groups (the pretest/postest design had no control groups) has its limitations. Still, some useful information can be found by taking a closer look at Table One. First, it's clear that of 28 possible mean comparisons, in 26 instances, average point totals have increased. The increases range from modest to fairly impressive (the 1998 total score mean average, for example, was up over 27 points). Let's walk through the data in more detail and consider some issues connected to significance and meaningfulness.

From Table One, two of the largest mean increases are for total scores in 1997 and 1998. 1997 mean scores went up an average of 14.56 points and 1998 scores rose by 27.32 points. The increases can be examined by doing some hypothesis testing. Let's take the 1997 increase first. To test for significance, in this instance, a paired t test can be used. A paired t test is used to test if two dependent sets of numbers are different, on average, when there is a natural pairing between the sets. Such a test fits the situation here, where there are "before/after" test scores for the same group of students. It is important that the data be paired; if not, there is no clear way to identify differences between pairs. In order to get useful information out of the pairings, it is necessary to work with the average and standard deviations of the differences. How this whole process works is summarized in Tables Two and Three below. (The data was checked for normal distribution by histogram; a normal distribution is a required assumption for paired t test validity).

TABLE TWO / 1997 Total Mean TOEFL Scores

	Test 1 (JULY)	Test 2 (DEC.)
Student 1	410	423
Student 2	363	407
Student 3	363	410
....	....	....
....	....	....
....	....	....
....	....	....
....	....	....
Student 270	377	443
Sample size =	270	270
Average =	399.38	413.94
Standard deviation =	35.62	33.27

TABLE THREE / 1997 TOEFL Score Change

	DEC. / JULY
Student 1	13
Student 2	44
Student 3	47
....	....
....	....
....	....
....	....
....	....
Student 270	66
Sample size =	270
Average =	14.56
Standard deviation =	48.89
Standard error =	2.98

The tables illustrate the fact that student 1, for example, went from 410 to 423, a mean gain of 13 points. (The convention for computing the change is to take "after" minus "before," or, here, the December test 2 score minus the July test 1 score . This is done so that increases end up as positives and decreases as negatives). As is shown in Table Three, we end up with what is basically a one sample problem.

The underlying hypotheses for the 1997 total mean score change are as follows:

H_0: There is no significant systematic relationship between the mean scores on the July and December TOEFL tests.

H_1: There is a significant mean change in total TOEFL scores from July to December.

In short form, the statistical details come down to the following (adapted from Brown, 1988):

1) Look at the hypotheses (as above): H_0: $\mu = 0$, H1: $\mu \neq 0$

2) Look at the alpha level: alpha $< .05$, two-sided test

3) The observed sample statistics are
$\bar{D}$ (average mean score change) = 14.56 $S_D = 48.89$

4) For paired data at 95% confidence level use the general formula

$$\bar{D} \pm t_{crit} \times \frac{S_D}{\sqrt{n}} = 14.56 \pm 1.960 \times \frac{48.89}{\sqrt{270}} \quad \text{OR} \quad [20.40,\ 8.72]$$

5) The hypothesis test simply involves seeing whether $\mu = 0$ is in the confidence interval or not. It isn't, so the result is statistically significant. For t tests, the confidence interval represents a difference between mean values. So for the data in this case, 95% of the mean differences should fall between 8.72 and 20.40 points.

The 1998 mean total score gain of 27.21 (Table One) can be analysed using the same treatment applied to the 1997 data. At the end of the day, the same formula (#4 above) yields the following numbers:

$\bar{D} = 27.21$ $t_{crit} = 1.960$ SD: 50.65 $n = 282$

$$27.21 \quad + \quad 1.960 \times \frac{50.65}{\sqrt{282}} \qquad \text{OR} \qquad [33.14,\ 21.27]$$

Again, the interval does not cover 0 so we can conclude the mean total score gain is significant. The numbers say 95% of mean differences should fall between 33.14 and 21.27 points.

While the total mean increases are significant, the job at hand is not quite done. Brown (1988) points out that significance needs to be distinguished from meaningfulness. Statistics can be significant without being meaningful. There is no "road map" for deciding meaningfulness and conclusions about it are often tentative. Concerning the meaingfulness of the data at issue, several points can be raised. For example, the '97 paired sample correlation coefficient is equal to minus .006. This number suggests a weak negative correlation; in other words, low scorers on the initial TOEFL have tended to go up in score and higher scorers have tended to go down. This finding, based on what was covered in the earlier research review, is perhaps an expected one.

The point here is simply that it is easy to overemphasize or overinterpret on the basis of a single statistic, such as a mean gain score. Things just aren't that simple.

The next section deals with reliability and it would make sense to continue working with the data from 1997 and 1998. However, the 1999 statistics are more suitable for a reliability discussion because three tests, rather than two, give a better view of the consequences of standard measurement error.

Reliability

The procedure followed in this section parallels the previously mentioned work of Childs (1995). The present study is different, however, in its use of the TOEFL (versus the TOEIC) and dependent (versus independent) student groups.

To examine some TOEFL test reliability features with a particular student group, the 1999 total mean score gains were used. Two measures were involved: an internal reliability estimate, plus a standard error of measurement instrument. The Kuder-Richardson formula 21 (K-R21) was used for the internal reliability estimate. The formula is not hard to employ and was applied to all three 1999 TOEFLS. All one needs to calculate are the number of test items, the mean of the test items and the standard deviation of the test items. The K-R21 formula is

$$\text{K-R } 21 = \frac{k}{k-1} \quad \frac{(1 - M(k - M))}{ks^2}$$

Where k = number of items
M = mean of test items
s = standard deviation of test items

The K-R21 computations for the three 1999 TOEFLs are below in Table Four:

TABLE FOUR / K-R21 Internal Reliability Estimates for 1999 TOEFL SCORES

	M (test mean)	k (# of test items)	s (standard deviation)	K-R21
Test #1 (April) (n = 313)	384.55	140	35.71	.8289
Test #2 (July) (n = 290)	403.73	140	37.01	.8736
Test #3 (Dec.) (n = 290)	406.84	140	39.30	.8936
			Average =	.8654

The K-R21 formula is said to be "conservative" in that it tends to underestimate internal reliability. Nonetheless, going to Table Four, it's evident that all the numbers are in the .8200 to .9000 range, including the average. These are relatively high reliability rates.

K-R21 is a reliability estimate based on group performance and offers information on test consistency for groups of individuals. So it doesn't tell us anything about the variance of individual scores. For information on single test takers we must look elsewhere.

Bachman (1997) offers a lucid explanation of an important statistic for use in the current context, namely, the standard error of measurement (SEM). He starts by saying if tests were 100% reliable, it follows that a test taker's obtained score, assuming no extra practice or improvement over time, should be exactly the same as his/her true score. Of course, "real world" test results don't quite work out in this way. Tests are never perfectly reliable and the resulting measurement errors cause observed scores to vary from true scores. These errors in measurement will make obtained scores higher than true scores in some cases and lower in others. To track individual scores in relation to these issues, the SEM is a very useful tool. The SEM can be thought of as a measure of the distance between a learner's observed and true scores. It sets a kind of range or band around where an individual test taker's score would fall if he/she took the same test many times.

Once the internal reliability of a test has been figured, we can use it to find the SEM:

$$\text{SEM} = \text{s (standard deviation)} \ \text{X} \ \sqrt{1 - \text{R}} \ \text{(reliability rating, from Table Four)}$$

Using this formula, the three SEMs for the tests are 14.62 for April, 13.03 for July, and 12.69 for December. (*See Appendix A*). Since the SEM is based on the assumption of a normal test score distribution, we might say, for example, that a student who had 420 in April would be within + 14.62 points of that same score if he/she repeatedly took the test, assuming no improvement or practice between test administrations. Note that the bands for the SEMs cover a total 29.24 points, 26.06 points, and 25.38 points. The concept of score bands is important and will be discussed in more detail shortly.

Gains and Losses

The SEM numbers and K-R21 values to a degree, point to a lot of ups and downs in individual scores. This idea is supported by simply taking the three tests and noting the pattern for losses and gains. Table Five shows the gain/loss figures between 1999 tests one and two, and between tests two and three.

TABLE FIVE / Gain and Loss Patterns for 1999 TOEFL Tests

Test 1	Test 2	Test 3	n (total = 260)
------	Gain	Loss	92 = 36.53%
------	Loss	Gain	72 = 27.70%
------	Gain	Gain	67 = 25.77%
------	Loss	Loss	14 = 5.4%
------	no change / Loss/ Gain combination		12 = 4.6%

One thing worth noting about the table is the fact over a 3 test period that saw mean scores go up over 22 points only about 26% of the test takers were able to register gains on two consecutive tests. The implications of this point are discussed in the next section.

Discussion

This walk through tables is admittedly quite selective and far from exhaustive in scope. Still, we are perhaps in position to return to the research questions we began with. Questions one and two asked if the TOEFL is a good match for measuring group and individual short term proficiency gains in the present setting. First, in my view, the TOEFL can be useful for looking at group gain scores provided its importance is kept in perspective, and with the realization that in many cases results will be statistically ambiguous. For example, the group gains looked at for 1997 and 1998 were found to be significant. But one of the reasons the gains are good is probably the fact that our students enter with such low scores. Recall a main point from the research overview: It is easier for a learners at the lower end of the TOEFL scale to make gains than those at the higher end. A jump from 350 to 400 should not be equated with an increase of, for example, 450 to 500. Note that in Table Five (1999) the most common pattern, for 36.53% of the test takers, is gain on test 2, then loss on test 3. The gain from test 1 to test 2 is over 20 points, which, since there are so many low scorers, again fits the idea of low scorers making substantial gains. But between tests 2 and 3 the increase was a scant 3 points. The 1999 data would seem to indicate that initially low group means will increase decently on a second test. At the same time, expecting similar score gains on subsequent tests, for both groups and individuals, may not be realistic; many of the low entry scorers who show gains on test two will fall back to a score more accurately reflecting their true scores on test 3.

In answer to research question 2, the TOEFL does not appear to be a particularly good measure of individual short term progress in this setting. Though the 1999 TOEFLS together showed an average internal reliability estimate of approximately 86%, an informed observer would do well to be aware of some other issues. First, as mentioned earlier, only 25.77% of 1999 students managed two consecutive score gains during a period in which average group mean scores rose over 20 points. Research question 3, concerning the standard error of measurement (SEM), also connects to question 2. We return to the idea of the SEM showing a kind of band into which an individual learner's score is likely to fall. For the 1999 data, as we have seen, these bands covered, 29.24, 26.06, and 25.38 points. Let's look at an example applying these bands. Suppose there is a test on which one learner scores 400 and another 417. On the surface, there is a fairly "concrete" difference between the two total scores. If we use the 29.24 band as our SEM for this test, the 400 student's confidence interval would be + 14.62 points. This means, with the SEM based on a normal distribution, that there is a 68% chance the true score falls between 385.38 and 414.62 points. Using the same logic, there is a 68% chance the 417 student's true score is somewhere between 402.38 and 431.62 points. The 14.62 SEM and the fact that the bands overlap mean that there is a 68% chance that the observed score difference of 17 points – between 400 and 417 – is the result of measurement error. Under these circumstances, the TOEFL should not be used as a measure of individual student progress over the short term.

The SEM is also an interesting statistic to look at in a level placement or minimum qualification score context. For example, let's say a student scores 487 on a test for which the program entry standard is 500. Again using the 14.62 SEM, while short of the goal, the obtained score of 487 is clearly within 1 SEM of 500. This student's test score confidence interval indicates she could well make the cut-off number by simply retaking the test. Should this student be accepted into the program? This is perhaps a debatable issue, but more importantly, it should be clear that considering the answer solely on the basis of THE SCORE does not make for good use of the TOEFL. To use an analogy, the TOEFL might be considered akin to a hammer in a tool box; it's an important tool, but it would be difficult to build a house using nothing else.

Further Discussion

One of the real weaknesses of what has been presented in the research section of this paper is that there is little said concerning causes of gain scores. Possibilities for investigation include such things as influence of classroom teaching (back to "washback"), number of classroom hours, the "polish effect" (mentioned in the research section), and student motivation. For those with views that differ from the conclusions drawn here, there is no shortage of angles from which to address the material.

Received:August 27, 2001

References

Alderson, J.C., and Hamp-Lyons, L. (1996). TOEFL preparation courses: A study of washback.*Language Testing* , 13, 280-297.

Bachman, L.F., (1997). Fundamental considerations in language testing. Oxford: Oxford University Press.

Brown, J.D., (1996). *Testing in language programs.* New York: Prentice Hall

Brown, J.D., (1998). *Understanding research in second language learning.*2nd ed. Cambridge: Cambridge University Press.

Brown, J.D., and Yamashita, S.O. (Eds.) (1995). *Language testing in Japan.* Osaka: Koshinsha.

Childs, M. (1995). Good and bad uses of TOEIC by Japanese companies. In J.D. Brown and S.O. Yamshita (Eds). *Language testing in Japan.* Osaka: Koshinsha.

De Avila, E. (1997). *Setting expected gains for non and limited English proficient students.* Washingto D.C.: National Clearinghouse for Bilingual Education.

Des Brisay, M., and Ready, D. (1995). Defining an appropriate role for language tests in intensive english programs. In A. Srainee (Ed.). *Issues in language program evaluation in the 1990s. Anthology series 27.* Singapore: Southeast Asian Regional Language Center.

Ercanbrak, J. and Robb, T. (1995). A study of the effect of direct test preparation on the TOEIC scores of Japanese university students. *TESL-EJ. 3:4*

ETS (Educational Testing Service). (1999). *Bulletin of the TOEFL, international edition.* Princeton, N.J.: Educational Testing Service.

ETS (Educational Testing Service). (1998). *TOEFL test and score manual.* Princeton, N.J.: Educational Testing Service

ETS (Educational Testing Service). (1999). *English proficiency tests: a comparative study.* Princeton,N.J.: Educational Testing Service.

Gates, S. (1995). Exploiting washback from standardized tests. In J.D. Brown and S.O. Yamashita (Eds.). *Language testing in Japan.* Osaka: Koshinsha.

George, D., and Mallery, P. (1999). *SPSS for windows: a step-by-step guide.* Boston: Allyn and Bacon.

Gilfert, S. (1995). A comparison of TOEFL and TOEIC. In J.D. Brown and S.O. Yamashita (Eds.). *Language testing in Japan.* Osaka: Koshinsha.

Gradman, H.L. and Hanania, E. (1995). Program evaluation in light of language learning background, student assessments and TOEFL performance. In A. Sarinee (Ed.). *Issues in language program evaluation in the 1990s. Anthology series 27.* Singapore: Southeast Asian Regional Language Center.

Hamp-Lyons, L. (1996). *Ethical test preparation practice: the case of the TOEFL.* Paper presented at the 18th Annual Language Testing Research Colloquium , Tampere, Finland.

Hilke, R., and Walden, P. (1997). The TOEFL and its imitators: Analyzing the TOEFL and evaluating TOEFL-prep texts. *RELC Journal, 23,* 28-53.

Swinton, S. (1983). A manual for assessing language growth in instructional settings. *TOEFL research report #4.* Princeton,N.J., : Educational Testing Service.

Wilson, K. (1987). Patterns of test taking and score change for examinees who repeat the Test of English as a Foreign Language. *TOEFL research report #22.* Princeton, N.J., : Educational Testing Service.

Appendix

Appendix A: : Calculations for K-R21 and Standard Measurement Error

Paralleling Childs (1995), a score reversal procedure was used for K-R21 and SEM calculations. To justify his method Childs offers a 1980 ETS citation, which I was not able to track down. In any case, in the Childs study the TOEIC was the test at issue. He used the following approximation for relating standard and converted scores:

Standard score = 6.25 X raw score

In this case, for the TOEFL, a different approximation was used:

Standard score = 3.33 X raw score

The approximation is based on the usual TOEFL raw score conversion procedure, as in the following example:

46 (List.) + 54 (Gram.) + 50 (Read.) = 150

150 X 10 / 3 = 500 (total)

For this study, to use K-R 21, standardized scores for total means and standard deviations were turned into raw values by reversing the conversion formula, as below:

April 1999 test mean = 384.55	384.55 X .3 (3/10) = 115.365
April 1999 test standard deviation = 35.71	35.71 X .3 = 10.713

The raw values were then plugged into the K-R21 formula (with k = 140) to obtain the internal reliability estimates. (.8289 for April 99)

To next calculate the SEM

SEM = standard deviation X $\sqrt{1 - R}$ = 10.713 X $\sqrt{1 - .8289}$ - = 4.431

4.431 = raw SEM To obtain converted SEM multiply by 3.33 (10 / 3)

4.431 X 3.33 = 14.62