The use of TOEFL to measure a change in English proficiency

Gary J. Ockey
International University of Japan

Abstract

This study investigates the use of TOEFL as an instrument for measuring change in English language proficiency of graduate students at the International University of Japan. The data, which was analyzed, included entry and exit TOEFL scores of 181 students who participated in the nine-week Intensive English Program during one of the summers between 1994 and 1997. The results suggest that TOEFL scores can be used to show a change in English language proficiency for the students as a group, but should not be used to report to individual students a change in proficiency. It is argued that the results can be explained by the error of measurement associated with TOEFL and a design which measures change in proficiency by comparing pre-test and post-test scores.

Key words: assessment, TOEFL, language proficiency, SEM

INTRODUCTION

It is no secret that tests are often used for purposes for which they are not designed. This is not surprising since it takes a long time for a test to gain the respect necessary for students, instructors, and administrators to give credence to its results. One such well-known test that has been used for various purposes is the Test of English as a Foreign Language (TOEFL). Because of its design, one would not expect that TOEFL could be used to measure a change in proficiency over the course of a language program. However, research suggests that TOEFL has been used for this purpose (Hilke, 1999, Swinton, 1983). Although it is believed that a test designed specifically to measure a change in proficiency of the International University of Japan (IUJ) student population would be a more appropriate test for this purpose, TOEFL has been considered because it is well-known, respected, and accessible. The aim of this study therefore was to investigate the appropriateness of TOEFL as an instrument for measuring a change in English proficiency of the students at IUJ.

TOEFL

The Test of English as a Foreign Language (TOEFL) is a well-known test which is respected in the English language teaching field. TOEFL was first developed in 1963-64, "... to measure the English proficiency of international students wishing to study at colleges and

universities in the United States and Canada, and this continues to be its primary function" (Educational Testing Service, 1997, p. 7). From this statement, it is clear that TOEFL was not designed to measure a change in English proficiency.

TOEFL consists of three sections: listening comprehension, structure and written expression, and reading comprehension. The listening comprehension section contains three components. The first part consists of short conversational exchanges between two speakers. The second and third parts contain short lectures and dialogs (up to two minutes in length). The structure and written expression section has two sections. The first part requires examinees to select information which will correctly complete a given sentence. The second part asks examinees to recognize errors in written English. The final section is reading comprehension. It contains a number of short passages that are based on factual information followed by comprehension questions which are based on the information in the passages. All questions are multiple choice formats with only one correct answer and three distractors. The test takes about two and a half hours to administer (Educational Testing Service, 1997, p. 11-12).

The Intensive English Program

Students who enter the Intensive English Program (IEP) at IUJ have been accepted as post-graduate students to study at the university in the fields of either international management or international relations. Since all content courses in both of these fields of study are taught in English, the IEP is designed to prepare students to function in English in one of these programs. The IEP includes 35 days of English instruction for five to six hours a day in a nine-week program. Each class day includes three hours of instruction for speaking and listening and two hours of instruction for reading and writing. The program also provides 5 hours of individual tutorial instruction and 16 hours of computer instruction taught in English and designed to facilitate the learning of English. There is no formal preparation for TOEFL taught in the classes. The class sizes range from 9 to 13 students. Various teaching styles are employed in the courses, including lectures, discussions, roleplays, simulations, case discussions, and student-led activities. Students have a substantial amount of homework which includes writing essays, preparing for oral presentations, researching current issues, and reading assigned materials. Students are also encouraged to use only English at all times during the nine-week program; they are provided with many opportunities to use English

outside of the classroom including planning and participating in various informal activities which are conducted in English.

THE STUDY

Since it is clear that what TOEFL was designed to measure and what is taught in the IEP are quite different, it would be unreasonable to believe that TOEFL could be used to accurately measure the change in proficiency of students during the nine-week IEP. On the other hand, it is likely that there is some measure of general proficiency overlap between what is learned in the IEP and what is tested by TOEFL. With this in mind, the general research question this study aimed to answer was whether or not TOEFL could be used to detect any change in students' English proficiencies, and if so, if it could be used to report to students this change in proficiency.

In order to examine this basic research question, three specific questions were considered. The first question that was dealt with was whether or not TOEFL could be used to demonstrate that there is a change in students' English proficiency during the IEP. It was predicted that TOEFL could be used to show a change in students' English proficiencies as a group. It was hypothesized that there would be some general proficiency overlap between what TOEFL measures and what is learned in the IEP which could be measured with a large group of students, such as the 181 in the study.

The second question to be answered in the study was whether or not TOEFL could be used to report to students a gain in English proficiency. It was predicted that TOEFL could not be used to show a change in a student's English proficiency over the nine-week period because TOEFL was not designed to detect such small changes accurately. Furthermore, research indicates that students at the middle and high end of the TOEFL scale show less change with the same amount of instruction as students at the low end of the TOEFL scale (Swinton, 1983). Therefore, since the students in the study mostly had scores at the high end of the TOEFL curve (see figure 1), it was believed that it would be very unlikely that TOEFL could be used to report to an individual student a change in proficiency.

The final question to be answered in the study was whether or not TOEFL could be used to report to a student a change in proficiency over the two-year period a student studies at IUJ. It was hypothesized that TOEFL could be used to show change in proficiency over a two-year period because there would be enough change in two years for TOEFL to detect it.

Data collection

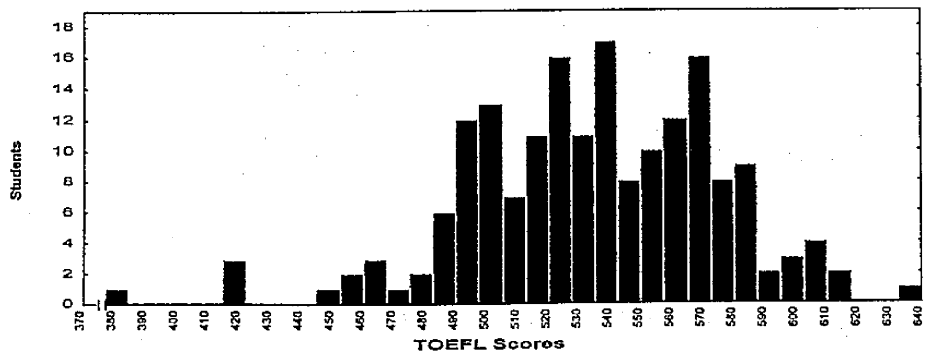
Institutional TOEFL scores from students who had taken one form of the institutional TOEFL paper and pencil test at the beginning of the IEP (the pre-test) and another form of the test at the end of the IEP (the post-test) were utilized in the study.[i] Ten of these students had also taken Institutional TOEFL upon the completion of their two years of study at IUJ (the graduation test); this data was included in the study as well.

Institutional TOEFL tests are ones that have previously been used as regular TOEFL exams. There are no differences between TOEFL and Institutional TOEFL test designs. The tests were computer scored by the Educational Testing Service.

Students

Data from 181 students was analyzed in the study.[ii] All of the students were in the IEP during one of the summers of 1994, 1995, 1996, or 1997. The students in the IEP had the following characteristics during that period. One hundred twenty-nine students were Japanese, 116 male and 13 female; 63 were Indonesian, 50 male and 13 female; 6 were Thai, 4 male and 2 female; 2 were Chinese, both female; 1 was a Korean male; and 1 was a male from Papua New Guinea. All students were graduate students. The oldest student was 40 and the youngest was 23; the mean age for all students was 29.8. TOEFL scores at the beginning of the IEP ranged from 640 to 377 (with the majority of students in the 490 to 590 range) with a mean of 533.6. A profile of the student populations' TOEFL scores can be seen in figure 1.

Figure 1



TOEFL scores of incoming 1994-1997 IEP students

RESULTS

The results are summarized in tables 1 and 2. In table 1, it can be seen that for the 181 students who took both the IEP pre-test and the IEP post-test, the average score on the pre-test was 533.5 and the average score on the post-test was 547.6, a difference of 14.1 points. The standard deviation on the pre-test was 40.6, and the standard deviation on the post-test was 37.7. A paired t-test showed that the difference between the pre-test and post-test scores was significant at the .05 level.

Table 1

	Pre-test	Post-test
Mean	533.5	547.6
Standard deviation	40.6	37.7
Observations	181	181
df	180	

P< .05

IEP students' pre-test and post-test scores on TOEFL

It can be seen in table 2 that of the 10 students who took the IEP pre-test and the graduation test, the average score on the IEP pre-test was 510.9 and the average score on the graduation test was 533.4. There was a difference of 22.5 points on the two tests. The standard deviation on the pre-test was 11.7, and the standard deviation on the post-test was 25.9. A paired t-test showed that the difference between the pre-test and post-test scores was significant at the .05 level.

Table 2

	Pre-test	Graduation test
Mean	510.9	533.4
Standard deviation	11.7	25.9
Observations	10	10
df	9	

P< .05

IEP students' pre-test and graduation test scores on TOEFL

DISCUSSION

Group change in proficiency during the IEP

The data give support to the first hypothesis that TOEFL could be used to argue that there is a change in students' English proficiency during the IEP. The change in students' scores on the two tests was positive (14.1 points), and this difference proved to be significant at the .05 level.

It might be argued, however, that factors other than the IEP experience contributed to this increase in TOEFL scores. For instance, in studies with a pre-test/post-test design, it could be argued that when the students take the post-test they are taking the test for the second time. The improvement might therefore be attributed (in part at least) to test wiseness, the concept that taking a test might result in better performance on another form of the test because students become familiar with the format, and procedures of the test (Bachman & Palmer, 1996, p. 31). Test-wiseness is unlikely to have had a significant effect in this study, however, since all students in the study were required to take TOEFL to enter IUJ only a few months before the pre-test. Generally speaking, other sources of test error[iii] would not have been significant because the tests were administered with strict compliance to TOEFL testing procedures.

It can thus be concluded that this study supports the claim that English proficiency, as measured by TOEFL, is increased during the course of the nine-week IEP. This means that pre-test and post-test TOEFL scores could be used to argue that students show increased English proficiency over the course of the nine-week IEP when the group is considered as a whole.

Individual change in proficiency during the IEP

The second purpose of the study was to determine whether or not TOEFL could be used to report a change in proficiency to individual students over the nine-week IEP. In order to answer this question, it became necessary to look at the students' scores on the tests individually. Students' gain scores based on the difference between their TOEFL pre-test and post-test scores, as can be seen in table 3, ranged from a high of 80 points to a low of –43 points--a 123 point range.

Table 3

Student	pre-test	post-test	change
1	377	457	80
2	503	580	77
3	500	567	67
4	467	530	63
5	420	480	60
6	520	577	57
7	513	567	54
8	510	563	53
9	550	603	53
10	453	503	50
11	457	507	50
12	493	543	50
13	573	620	47
14	567	613	46
15	583	627	44
16	533	577	44
17	527	570	43
18	567	610	43
19	520	560	40
20	493	533	40
21	527	567	40
22	547	587	40
23	517	557	40
24	533	573	40
25	533	573	40
26	497	537	40
27	513	550	37
28	500	537	37
29	570	607	37
30	510	547	37
31	503	540	37
32	524	567	36
33	487	523	36
34	587	623	36
35	503	537	34
36	493	527	34
37	603	637	34
38	513	547	34
39	513	547	34
40	520	553	33
41	540	573	33
42	507	537	30
43	417	447	30
44	513	543	30
45	570	597	27
46	563	590	27
47	563	590	27
48	493	520	27
49	583	610	27
50	487	513	26
51	507	533	26
52	493	517	24
53	563	587	24
54	533	557	24
55	540	563	23
56	460	483	23
57	450	473	23
58	497	520	23
59	507	530	23
60	517	540	23
61	477	500	23
62	520	543	23
63	527	547	20
64	470	490	20
65	503	523	20
66	580	600	20
67	583	603	20
68	527	547	20
69	560	577	17
70	583	600	17
71	540	557	17
72	570	587	17
73	487	503	16
74	577	593	16
75	493	507	14
76	493	507	14
77	530	543	13
78	577	590	13
79	520	533	13
80	507	520	13
81	530	543	13
82	577	590	13
83	570	583	13
84	560	573	13
85	530	543	13
86	517	530	13
87	567	580	13
88	513	523	10
89	557	567	10
90	503	513	10
91	537	547	10
92	537	547	10
93	500	510	10
94	537	547	10
95	563	573	10
96	537	547	10
97	503	513	10
98	500	510	10
99	533	543	10
100	557	567	10
101	543	551	8
102	483	490	7
103	480	487	7
104	563	570	7
105	490	497	7
106	530	537	7
107	543	550	7
108	573	580	7
109	580	587	7
110	540	547	7
111	537	543	6
112	467	473	6
113	497	503	6
114	567	573	6
115	503	507	4
116	563	567	4
117	523	527	4
118	543	547	4
119	420	423	3
120	617	620	3
121	567	570	3
122	600	603	3
123	610	613	3
124	537	540	3
125	530	533	3
126	537	540	3
127	567	570	3
128	577	580	3
129	577	580	3
130	520	520	0
131	523	523	0
132	537	537	0
133	590	590	0
134	550	550	0
135	540	540	0
136	487	487	0
137	513	513	0
138	497	497	0
139	503	503	0
140	523	523	0
141	537	537	0
142	607	607	0
143	560	557	-3
144	500	497	-3
145	570	567	-3
146	600	597	-3
147	537	533	-4
148	567	563	-4
149	543	537	-6
150	603	597	-6
151	543	537	-6
152	550	543	-7
153	570	563	-7
154	580	573	-7
155	550	543	-7
156	567	560	-7
157	577	570	-7
158	560	553	-7
159	530	523	-7
160	560	550	-10
161	567	557	-10
162	560	550	-10
163	580	567	-13
164	543	530	-13
165	543	530	-13
166	600	587	-13
167	550	537	-13
168	537	523	-14
169	517	503	-14
170	593	577	-16
171	640	623	-17
172	557	540	-17
173	520	503	-17
174	487	467	-20
175	507	487	-20
176	527	507	-20
177	567	547	-20
178	550	530	-20
179	603	580	-23
180	540	510	-30
181	553	510	-43

Pre-test, post-test, and gain scores of IEP students on TOEFL between 1994 and 1997

While at first this might appear to be a surprisingly large range of scores, in fact, it is quite a predictable range considering the possible measurement error. One way to predict how closely an acquired score will be to a student's true score (his actual proficiency) is to look at the Standard Error of Measurement (SEM) on the test (Bachman, 1997, p. 199). The SEM is a confidence band which can be used to predict how far a student's true score might be from his measured score. Assuming that a student's true gain score was the average gain score of 14.1 points on the test, one can predict how far a student's score will vary from that point, based on

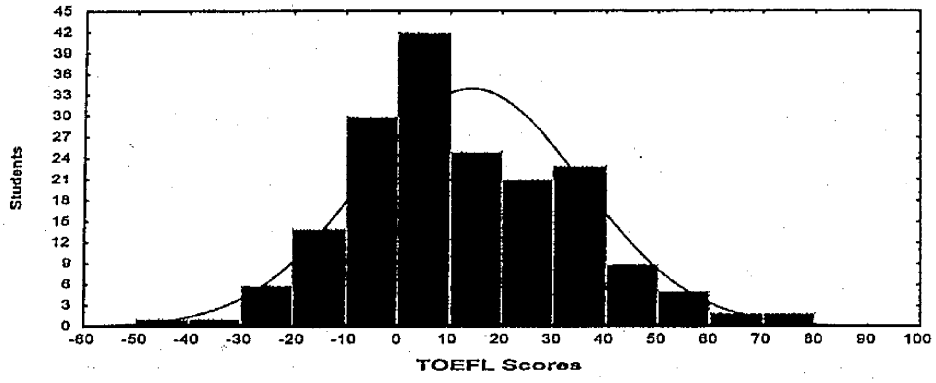
the SEM. The SEM is reported based on one standard deviation, 68% probability (Hatch & Lazarton, 1991, p. 478-479). An SEM score predicts that 68% (one standard deviation) of scores will be within the number of points reported. For example, if a test had an SEM of 5, it could be predicted that 68% of the students that took the test would achieve scores which would be within 5 points of their true scores on the test. For TOEFL, during the years of the study, an SEM of 13.9 points was reported (Educational Testing Service, 1997, p. 30).[iv, v] This means that approximately 68% of the students who took TOEFL during that period of time would achieve scores within 13.9 points of their true scores.

When gain scores are compared, as is the case in this study, it is necessary to consider the SEM of both tests. For example, if a student scored one standard deviation (the SEM value) below his true score on the first test and one standard deviation (the SEM value again) above his true test score on the second test, his reported gain score would be two standard deviations away from his true gain score. With this in mind, it is necessary to double the reported SEM to be able to predict the error in a gain score.[vi] This means that the expected SEM for the gain score of TOEFL would be 27.8 (the 13.9 reported SEM doubled to take into account the SEM of both the pre-test and the post-test).

This error of measurement explains why gain scores on the test appear to be somewhat random when viewed in isolation. A 14.1 point gain in proficiency, the average gain for students in this study, cannot be detected when looking at students individually—there is too much potential for random error. As was shown above, one could predict that 68% of the students in this study would have gain scores within 27.8 points of the mean gain score. On the other hand, 32% of the students would be expected to have gain scores more than 27.8 points from the mean gain score. In fact, 32 of the 181 (17.7%) students have gain scores further than 27.8 points from the mean gain score (see table 3). It could also be hypothesized that 95% (2 standard deviations from the mean) of the students would have gain scores within the range of 55.6 points of the mean gain score while 5% of the students would have gain scores further away from the mean gain score. In fact 3 of the 181 students (1.7%) have gain scores more than 55.6 points from the mean gain score (see table 3).[vii] As predicted, a number of students have scores which are far from the mean gain score.

The SEM for the gain scores of the students is shown in figure 2. It shows that the largest cluster of students is near the mean, with fewer scores appearing the further one moves from this point. The line in figure 2 indicates a normal curve which would be expected if the students gained on average 14.1 points with SEM explaining the deviance from the mean.

Figure 2



TOEFL gain scores on IEP 1994-1997 (based on the IEP pre-test and post-test)

Assuming all students did make the average gain, the fact that the average gain score is only slightly higher than the SEM of both the pre-test and the post-test suggests that many students would not get a higher post-test score than pre-test score. This in fact was the case in this study. In table 3, it can be seen that of the 181 students in the study, 41 students (student numbers 143-181) had lower scores on the post-test than on the pre-test, and eleven students (student numbers 130-142) showed no change on the test. Therefore, if TOEFL were used to show students their change in proficiency over the course of the IEP, fifty-two students (28.7%) would be given the message that they had made no improvement in English proficiency over the course of the IEP.[viii] With this in mind, it is clear that because of the error of measurement on the test, it is problematic to use TOEFL to show individual students that they have made a gain in proficiency over the course of the nine-week IEP.

Individual change in proficiency over a two-year period

The final aim of this study was to see if TOEFL could be used to report to a student a change in proficiency over his two-year period of study at IUJ. The logic was that if there was enough of a change in proficiency that could be measured by TOEFL, the SEM would have less effect on the gain score. Unfortunately, data was only available for 10 students, so the results must be treated with extreme caution. As a group, the students showed a significant gain (an average of 22.5 points) on the test after completion of the two-year course (see table 2),[ix] but when the students are looked at individually, the problem of SEM on TOEFL again

arises. The results show that 9 out of the 10 students had a higher score after they graduated than on the pre-IEP test; one student, however, had a negative gain score—number 10 got three points lower on the graduation test than on the pre-IEP test (see table 4).

With only 10 students in the study it is unreasonable to come to any clear conclusions. However, since 1 of the 10 students had a lower score on the post-test than on the pre-test, TOEFL is probably not a useful instrument for reporting a change in proficiency to a student over his two-year period of study at IUJ. On the other hand, 9 out of the 10 students did show positive gain scores, suggesting that TOEFL might be used to report to a student a positive gain in proficiency. Thus, as to whether TOEFL can be used to report to an individual student a change in English proficiency over a two-year period remains unanswered in this study.

Table 4

Student	Pre-IEP score	Graduation score	Change
1	500	597	97
2	513	540	27
3	513	540	27
4	500	520	20
5	517	537	20
6	503	520	17
7	503	517	14
8	520	523	3
9	537	540	3
10	503	500	-3

Pre-IEP TOEFL scores compared to scores at graduation

CONCLUSION

This study suggests that TOEFL could be used to show a gain in the English proficiency of students over the course of the IEP. Furthermore, the results indicate that the IEP experience contributes to a student's improvement in English proficiency. Thus, because a significant group gain can be detected by TOEFL, and it is a well-known test which is highly respected by most faculty and administrators, TOEFL is arguably a useful instrument to use in order to indicate the success of the IEP. On the other hand, this study suggests that it is unreasonable to use TOEFL to report to a student his change in proficiency over a short

period of time. Because the average gain on the test is almost the same as the SEM on the test, there is a high probability that many students will attain negative gain scores. One likely effect of reporting gain scores to students would be that the students who did not achieve positive TOEFL gain scores (28.7% of the students in this study) would feel the IEP was not effective in helping them to improve their English. This might result in negative feelings of students toward the IEP, instructors, and administrators—clearly an undesirable outcome of assessment. Further research is needed to determine whether or not TOEFL could be used to report to students a gain in proficiency over the two-year period that they study at IUJ.

If TOEFL cannot be used to report to students their gains in proficiencies over the course of the IEP or their two years of study at IUJ, what can be done to show students how much their English has improved while studying at the university? One solution might be developing an in-house test, or as Brown and Hudson (1998) recommend developing a variety of instruments for assessing proficiency (p. 670), designed for the population of students at IUJ. While it would be extremely difficult to develop a reliable battery of tests for measuring English proficiency, if IUJ students are to be given the kind of information they need to recognize their English improvement, this is the task to be accomplished.

---

[i] See TOEFL 1997 for TOEFL procedures.

[ii] There were a total of 202 students in the IEP during this time period, but only 181 were included in the study because some students lacked either an entrance TOEFL score or an exit TOEFL score on account of being absent on one of the test days.

[iii] See Brown 1996, page 189 for a further list of sources of error.

[iv] An SEM of 13.9 was reported for 1996-1997 while an SEM of 14.1 was reported for 1995-1996. Although there is a slight difference here, for the purposes of this study, it is not a large enough difference to have an important impact on the results.

[v] The TOEFL scores of IUJ students are bunched up in the 490-590 range, so they will probably not show as much SEM as would be expected of the general TOEFL population. For the purposes of this study, however, the reported SEM should be adequate.

[vi] Doubling the SEM may not give a precise estimate of the expected error, but for the purposes of this study it should be an adequate procedure.

[vii] The scores might not show as much error as predicted because most students in the study are close to the mean of TOEFL, and there might not be as much error associated with these scores as with scores which are far away from the mean.

[viii] Other students who only gained a few points would no doubt also feel they had not made reasonable improvement during the course.

[ix] The large gain by student number 1 probably had an unreasonable effect on the group average gain.

Acknowledgements

I wish to thank Professors James Ockey, University of Canterbury; Mohammed Ahmed and Donna Fujimoto, International University of Japan; and William Bonk, Kanda University of International Studies who provided valuable feedback on drafts of this paper. I am also indebted to Ms. Mitsuko Nakajima who assisted in collecting and compiling much of the data in this study.

REFERENCES

Bachman, L. F. (1997). Fundamental considerations in language testing. New York: Oxford.

Bachman, L. F. & Palmer, A. S. (1996). *Language testing in practice: designing and developing useful language tests.* New York: Oxford.

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

Brown, J. D. & Hudson, T. (1998). The alternatives in language assessment. *TESOL Quarterly, 32 (4),* 653-675.

Educational Testing Service, (1995). *TOEFL test & score manual.* Princeton: educational testing service.

Educational Testing Service, (1997). *TOEFL test & score manual.* Princeton: educational testing service.

Hatch, E. & Lazaraton, A. (1991*). The research manual design and statistics for applied Linguistics.* New York: Newbury House.

Hilke, R. (1999, January 19). Personal communication.

Swinton, S. (1983*). A manual for assessing language growth in instructional settings (TOEFL Research Report 14).* Princeton, NJ: Educational Testing Service.