Learning Activities Associated with English Proficiency Test Score Improvements

Yuka MASDA

(Received on November 2, 2009)

要　　　約

コンピューターを使用した英語の授業において，聴解・文法課題や語彙学習といった学習活動や，出席状況等のうち，何が習熟度テスト得点の伸びにつながっているかを調査した。結果は，習熟度テスト得点の伸びは，各種の達成度テストの得点，具体的には，聴解，文法，少量づつあるいは多量の語彙学習のテスト得点と関連づけられ，これらの学習が得点の伸びに貢献することを示唆した。補足的に，文法や語彙のような，以前経験した種類の学習のテスト点は，学生の習熟度によって上下するため，これによって聴解テストが，得点の伸びを示す信頼性の高い指標となることがわかった。

Summary

This study explored which of the various learning activities provided by a computer-assisted language learning course are associated with the subsequent proficiency test score improvements. The results showed the proficiency score improvements were associated with the scores of various types of achievement tests, namely those testing progress in learning of listening, grammar, and vocabulary in both small amounts and in bulks, suggesting the effects of such learning over score improvements. Additionally, test scores on previously experienced types of learning, such as that of grammar and vocabulary knowledge, tend to rely on students' previously acquired proficiencies, making listening skills tests a particularly reliable indicator of proficiency score improvements.

Objectives

One of the general trends currently observed in higher education in Japan is the use of proficiency tests whose scores are standardized and costs relatively reasonable. Standardized scores, in principle, are those statistically processed so that a certain score indicates approximately the same proficiency level regardless to which batch of the test the score holder took. Together with the low costs, such tests appear to be practical in aspects of higher education, for instance as a good indicator of students' abilities when seeking employment.

TOEIC is often known as one such test, but there is another called TOEIC Bridge, which targets learners who are not yet accustomed to the testing format of TOEIC. Unlike a typical test given in Japan — a bilingual test whose instructions are written in Japanese, both TOEIC and TOEIC Bridge are conducted entirely in English, i.e., all the instructions as well as questions are provided in English. Other features common between these two tests are that all questions offer multiple-choice answers, and that a test is divided in two parts, the first part testing listening skills by use of recorded sound, and the second testing grammar knowledge, vocabulary use, and reading skills. The difference on the other hand lies in the time limits and the levels and numbers of questions: TOEIC consists of 45 minutes of 100 listening questions and 75 minutes of 100 reading questions, all together making up two hours of 200 questions, while TOEIC Bridge asks 100 questions in an hour, divided in 25 minutes of 50 listening questions and 35 minutes of 50 reading questions.

Another noticeable trend in higher education is computer-assisted language learning. In a study, 75–76% of the institutions in higher education in Japan had introduced ICT (Information and Communication Technology) and more and more are conducting e-learning classes, where students make use of ICT (e-learning White Paper 2008/2009, 2008). These two trends are sometimes combined, so that a certain level of proficiency test scores is at least part of the requirements and goals of such a computer-assisted course. The question pertinent to the fundamental course design then would be what aspects and activities of such a language class are involved in proficiency test score improvements. This study examines it by comparing students' score improvements in one such a proficiency test, TOEIC Bridge at the beginning and end of a semester to their achievement levels in various learning activities provided in a computer-assisted language course.

Procedures

Targeted learners

The data were collected from 136 first year university students studying on a compulsory computer-assisted English course. They belonged to four different classes varying in size from seven to 52 students and to similar proficiency levels. During this semester period, all the students took only one other English course, namely general English, besides this computer-assisted one, as a compulsory component of the curricula of their respective departments, which spanned from that of business to law. As a result the students practically faced no necessity or opportunities to study English in a formal setting outside these two English classes.

Most classes on this computer-assisted English course were organized in three proficiency levels according to the scores of the proficiency test the students took at the beginning of the semester. Majority of the non-English students on this course including all the students in the current study belonged to the beginner level. Especially at the beginning of but also throughout the course, many of these students expressed on various occasions that they had not been good at or fond of English learning.

Learning activities and proficiency tests

On this computer assisted language course, the students were assessed in several categories including class attendance rates, task completion rates, various types of achievement tests targeting different areas of language knowledge and skills and set at different intervals, and the proficiency tests at the beginning and end of the term. Out of these assessment records, 8 categories[1] were used for the current analyses and are listed below.

Categories of scores analysed in this study

1. Proficiency test scores at the beginning of the semester (Proficiency 1)
2. Proficiency test scores at the end of the semester (Proficiency 2)
3. Score differences between the two proficiency tests (Proficiency 2 – Proficiency 1)
4. Attendance rates
5. Weekly task completion rates
6. Weekly listening test scores
7. Weekly vocabulary and grammar test scores
8. Periodical vocabulary test scores[2]

Because the purpose of the study is to determine factors associated to TOEIC Bridge score improvements, scores from the both beginning and end of the semester tests (1. and 2.) as well as the differences between the two test scores (3.) were analysed. TOEIC Bridge gives scores in even numbers within the range of 20 to 180, e.g., 20, 22, 24 and so on. These scores are not the proportion of the correct answers but are scaled, computed from the original raw scores using statistical process called equating in order to stabilise the scores across tests. For this reason, the ‘raw’ TOEIC Bridge scores were used in these analyses.

[1] Not all these records were used for course grading, however. These were Proficiency 1 and Vocabulary 1–3. Students were informed of the assessment scheme.

[2] This has 4 sets of data: Periodical vocabulary test 1, 2, 3, and 4.

For Category 4–8, which are records of learning activities in and outside class, each score was converted into the proportion to its maximum score before performing the analyses. For example, there were 15 class hours in total for the course, so if a student misses a class, the number of classes attended out of 15 (an attendance record of 14) will be treated as 14/15, which may be expressed as .9333333 if rounded at the seventh digits after the decimal point, or simply 93%. As a rule, six digits after the decimal point were used for Category 4-8 in the statistical analyses for this study.

The weekly task completion in Category 5 recorded the number of weeks a student completed the on-line listening questions given in the TOEIC or TOEIC Bridge format. These were to be completed by the students outside class hours at their own pace to meet the weekly deadlines. The tasks were given 10 times so the maximum score was 10 in most cases[3].

The two kinds of weekly tests as in Category 6 and 7 in the list tested the language items practised the previous week. These items were listening skills (6. listening test), grammatical items such as verb inflections, lexical items such as individual words, phrases, and whole sentences, either independently as Japanese-English translations, in gap-fill or word reordering questions (7. vocabulary and grammar test). Altogether, 10 tests of 10 marks each were given, with the maximum scores of 100 for each of the two kinds.

The vocabulary tests (8. periodical vocabulary tests) were given four times in total during the semester and targeted a list of 200–400 items of vocabulary selected as being essential for TOEIC. The amount of target vocabulary was adjusted according to the class proficiency level, resulting in a range of maximum scores from 248–400. This should hold no obstacles, because the current study focuses on the relationship between proficiency levels and the levels of achievements relative to the expectation in various learning practices. However, it is conceivable that the difference in the amount of tested language items may have distorted the difficulty levels across classes by making the tests with more items more difficult, so a separate set of analyses was performed in order to explore the possibility. This will be reported in the results.

Although the majority of assessment records were analysed for this study, some were not included in this report for various reasons. For example, the rate of final task completion was not included because the data offered only redundant information.

[3] One of the classes had the maximum score of 11 because of an extra mark given for an administrative reason.

Results

Preliminary analyses

Before we move on to the main analyses, we need to examine the possible effects of target language amount differences in vocabulary tests between classes as discussed in the last section.

All the tests for Class 1, 2, and 4 had the same number of target lexical items (N_1 = 50, N_2 = 100, N_3 = 100, N_4 = 150, for each of the four tests), while there were considerably fewer for Class 3 (N_1 = 50, N_2 = 50, N_3 = 50, N_4 = 100, likewise). If these differences influenced the results as hypothesised, the scores are expected to be consistently and/or significantly higher for Class 3 than for the rest of the classes. However, this was not the case. Class 3 came third twice and fourth twice, with various score differences between 2% and 30% from the top class. there were no obvious score patterns as a whole except for that Class 1 and 2 constantly scored higher than the other two, as their general proficiency score levels predicted. When the data from the three classes sharing the testing conditions were pooled together and compared against those of Class 3, score bias was observed but only in the direction opposite to that predicted by the condition differences, thus clearly discounting the doubt regarding the testing condition bias.

With these results, all the vocabulary test data were incorporated into the following analyses, after averaging out the proportional scores of the four tests, not the raw scores, for each student.

The main analyses

Table 1 shows the means, standard deviations, and numbers of students for each assessment category for each class. As the note to the table explains, the improvement values were calculated as the difference between Proficiency 1 and Proficiency 2 for each student and not as that for the entire group. The number of students (*N*) differs depending on the data availability in each category. In the following analyses these missing values were excluded pair-wise.

All the class assessment categories had a high level of means, as their purposes were to motivate students, to serve as reviewing occasions, and to assess progress not objective proficiency level, and therefore, were designed to score approximately slightly below or above 80%, as was considered to be most effective for these purposes. Overall, the proficiency scores improved by 14.5 points during the semester.

	Means	*SD*	*N*
Attendance	.928	.129	136
Task completion	.813	.249	136
Listening tests	.756	.154	135
G&V tests	.797	.105	135
V tests	.865	.124	135
Proficiency 1	109.1	16.5	107
Proficiency 2	124.7	14.1	129
Improvement	14.5	11.5	103

Table 1　The means and standard deviations of the assessment categories and proficiency test scores
Figures are rounded at the fourth digit after the decimal point for the first five categories, and at the second for the last three. G&V tests indicates the weekly tests on grammar and vocabulary, V tests the average of the four periodical vocabulary tests. The improvement values are calculated as the difference between Proficiency 1 and Proficiency 2 for each student and not as that for the entire group. The number of *N* differs depending on the data availability. In the later analyses these missing values were excluded pair-wise.

Simple correlation analyses were performed on the data summarised above. Table 2 shows the correlation coefficients between all the 8 assessment categories. The target of this study, the proficiency score improvements, were significantly correlated with attendance, $r = .219$, $p < .05$, task completion, $r = .284$, $p < .01$, listening skills tests, $r = .283$, $p < .01$, weekly grammar & vocabulary tests, $r = .236$, $p < .01$, with small to medium size effects (Cohen, 1992)[4], but not the periodical vocabulary test results. However, the vocabulary tests were strongly correlated to Proficiency 1, $r = .533$, $p < .01$ as well as Proficiency 2, $r = .633$, $p < .01$. The improvements were also negatively correlated with Proficiency test 1, $r = –.513$, $p < .01$, which indicates that it was clearly more advantageous for the relatively low score receivers to achieve bigger score improvements, at least after 4 months of computer-assisted language learning. Improvements were mildly correlated with Proficiency test 2, $r =.220$, $p < .5$, although this result is tautological, as the higher these scores are the more chance there is of bigger improvements regardless of the Proficiency 1 scores.

The attendance highly correlated with task completion, $r = 567$, $p < .01$, and the two weekly tests, $r = 665$, $p < .01$, $r = 697$, $p < .01$ respectively, also relatively strongly with the vocabulary tests, $r = 408$, $p < .01$, but not with either of the proficiency tests. The task

[4] The proposed index of effect size is that $r = .10$ is small, $r = .30$ is medium, $r = .50$ is large (Cohen, 1992). Note that these effect sizes are not necessarily a direct indication of how effective these activities are relatively to each other, especially when you compare two tests of different sensitivities.

	attn	task	L	G&V	V	P1	P2
task	.567** (136)						
L	.665** (135)	.649** (135)					
G&V	.697** (135)	.431** (135)	.647** (135)				
V	.408** (135)	.272** (135)	.397** (135)	.552** (135)			
P1	–.045 (107)	–.191* (107)	–.052 (107)	.182* (107)	.533** (107)		
P2	.024 (129)	–.065 (129)	.167* (129)	.389** (129)	.633** (129)	.724** (103)	
Impt	.219* (103)	.284** (103)	.283** (103)	.236** (103)	–.005 (135)	–.513** (103)	.220* (103)
	attn.	task	L.test	G&V	Vavg	P1	P2

Table 2 Pearson's correlation coefficients (*r*) for the 11 assessment categories
The figures marked with ** are significant at 1% and those marked with * at 5%, both one tailed. The numbers in brackets indicate the numbers of participants (*N*). 'attn.' means attendance, 'L' weekly listening skills tests, 'G&V' weekly grammar & vocabulary tests, 'V' the mean scores of periodical vocabulary tests 1–4, 'P1' and 'P2' proficiency test 1 and 2 respectively, and 'Impt' the improvements from Proficiency 1 to Proficiency 2.

completion rate also showed a strong correlation with listening skills tests, $r = 649$, $p < .01$, a medium to strong correlation with grammar & vocabulary, $r = 431$, $p < .01$, and a medium correlation with the vocabulary tests, $r = 272$, $p < .01$, but notably, a negative correlation with Proficiency 1 if not to Proficiency 2. This means high scorers in Proficiency 1 tended to fail to demonstrate the diligence to complete weekly tasks.

The two weekly tests were correlated not only with attendance and task completion, but also with each other $r = 647$, $p < .01$ and the periodical vocabulary tests, with stronger correlation between grammar & vocabulary tests and vocabulary, $r = 552$, $p < .01$, than between listening tests and vocabulary, $r = 397$, $p < .01$, despite the fact that these two tests (weekly grammar & vocabulary and periodical vocabulary) tested different sets of lexical items. Grammar & vocabulary tests were also more highly correlated to Proficiency 2, $r = 389$, $p < .01$, than Proficiency 1, $r = 182$, $p < .05$, again more so than listening tests were with either proficiency test. These results regarding the weekly tests may be taken to indicate that the grammar & vocabulary tests were a better indicator of the students' overall proficiencies, while both test results were as important to proficiency improvements.

The periodical vocabulary tests demonstrated high correlation with the both proficiency tests, and the correlation was stronger with Proficiency 2. This was true even when each vocabulary test was compared separately with the proficiency tests. However, as reported above, these periodical tests were not shown to be related with Improvements themselves. This seeming contradiction will be discussed later.

Discussion

Assessment categories may be divided in three different groups for two reasons: first, the nature of activities involved were clearly different between these tests, and second, the general pattern of correlation were more or less similar within each group. The three groups are listed below.

Grouping of assessment categories

1. Attendance rates; task completion rates (motivation and self-management)
2. Weekly listening tests; weekly grammar & vocabulary tests (progress on class work and assignments)
3. Periodical vocabulary tests (progress on assignments and self-management)

Attendance rates and task completion rates are not directly concerned with students' language skills, but more with their motivational and self-managerial aspects. The two weekly tests assess their progress in listening tasks and grammar & vocabulary tasks. The periodical vocabulary tests also assess progress in learning a vocabulary list, although, unlike the weekly tests, the lexical items tested on these tests were not covered in class so individual students should study for the tests outside the class hours, and also the timing and frequency of the periodical tests differed from the weekly tests. For these reasons, I will discuss the results in each of these three groups in this section.

Group 1: measurements of motivation and self-management

If students attend more classes, it is assumed that they would be more motivated and therefore would more regularly work on the weekly tasks, and also that by default they would miss fewer weekly tests, collecting more marks in these tests as a result. These assumptions were supported by the high correlations between these categories.

Both measures were significantly correlated with the proficiency score improvements, although the attendance rates were not with either of the proficiency tests, and in regards to the task completion rates, negatively with Proficiency 1 and not at all with Proficiency 2. These results suggest that for one thing, as mentioned in one of the previous sections, early proficiency high scores contradictorily often lead to the lax attitude in weekly task completion, that regularly attending the class and completing given tasks are not strong factors in achieving high proficiency score levels at least by themselves. However, it should not be forgotten that both these learning attitude measurements were associated with score improvements, probably because they are both vital requirements in providing learning environments.

Group 2: measurements of progress on class work and assignments

To summarise the results, the listening skills tests and grammar & vocabulary tests were both shown to be moderately related with proficiency score improvements, strongly with each other, with the periodical vocabulary tests though with stronger correlation between grammar & vocabulary tests and vocabulary, and also between grammar & vocabulary and each proficiency test. To locate the small but noticeable difference, further analyses were performed.

Previously to this course, the students had not had a large amount of input in listening compared to reading and writing, were not used to listening skills tests, while they had been subjected to the kind of grammar and/or vocabulary tests as those discussed in this study. This internal difference may have influenced the students' learning experience. For example, it is plausible that, similarly with the vocabulary tests as will be reported in the next section, the grammar & vocabulary tests measured levels of certain qualities, which built upon and depended on the skills and knowledge already existing in students through the past learning experience, while the listening tests did not. In order to test this question, the data were reassessed using the following approach and statistically analysed.

There were ten tests with the maximum score of ten in either of the weekly tests, and the total marks were divided by the total maximum of 100. Not all students sat all these ten tests; if they missed a class in which any of these tests were given, they will automatically lose the marks they would have received otherwise. The data gained thus are unlike the data acquired by averaging out the scores of all the tests taken, because the current data are more likely to be an indicator of how much input of the whole course students absorbed by investing time and efforts accordingly, while the simple average scores may be a purer indicator of how well they could do in a given test on a limited amount of input.

A new set of data was collected by dividing each student's total marks by the maximum possible scores of all the tests taken by that student. Simple correlation analyses were performed on this. The correlation coefficient between the two weekly tests was much lower, $r = .336$, $p < .05$, than that in the former analyses, $r = .647$, $p < .01$. Listening tests showed no correlation with Proficiency 1, moderate correlation with Proficiency 2, $r = .216$, $p < .01$, and Improvements, $r = .254$, $p < .01$, while grammar & vocabulary more highly correlated to Proficiency 1, $r = .380$, $p < .01$, and Proficiency 2, $r = .506$, $p < .01$, but not to Improvements.

These results support the hypothesis that listening tests and grammar & vocabulary tests tested qualities to an extent different from each other (pattern of correlation between the two types of tests), and that the qualities tested in grammar & vocabulary were dependent on the already existing knowledge and skills (pattern of correlation with the proficiency tests). If this was the case, the previously acquired proficiency is expected to interfere with the actual effects of such learning and to confound the results. To further examine the possible contribution of grammar and vocabulary items, post-hoc analyses were performed to determine the relationship between the all types of measurements of listening and grammar & vocabulary tests and the proficiency score improvements (or Proficiency 2 scores[5)]) while controlling the proficiency levels. Significant correlation was found between the improvements and all the achievement test scores, semester total listening scores, $r = .301$, semester total grammar & vocabulary, $r = .372$, average listening, $r = .318$, average grammar & vocabulary, $r = .367$, all at $p < .01$, suggesting that both types of learning, i.e., listening and grammar & vocabulary improved the subsequent proficiency scores.

Additionally, it further indicates that, in the case of grammar and vocabulary tests, the overall amount and achievement levels through the learning may be a handy and more reliable predictor of subsequent proficiency score improvements, but not how well students did on average, which tends to be crowded over by their previous proficiency levels, while both scores predict score improvements as well as each other in case of listening skills.

Group 3: measurements of progress on assignments and self-management

The periodical vocabulary test results were not shown to be associated with the proficiency score improvements, while they were strongly correlated to the both proficiency tests, more so with Proficiency 2. The general pattern of the results seems to indicate that these vocabulary test scores largely and effectively measured students' general proficiency levels existing

[5)] When Proficiency 1 is controlled for, Proficiency 2 and Improvements have the same mathematical values.

previously to the learning, probably because of the former's dependency on the latter as speculated regarding the grammar & vocabulary tests in the last section, making the contribution of vocabulary learning during the current timeframe hard to observe. Post-hoc analyses similar to those conducted earlier, where Proficiency 1 is controlled for, indicated significant correlation, $r = .437$, $p > .01$, supporting this.

Conclusion

It is an often noted valuable caution that correlations do not directly support causation by default. Likewise, this study does not claim to have found any such causation. However, it appears as if it was not mentioned often enough when the causation is more likely in a correlational analysis, and it is when one variable may be safely assumed to have affected the other, rather than the other way. This study approached this problem by incorporating the two proficiency tests before and after the various learning activities, which approximates an experimental procedure of pre-test, treatment, and post-test in a real-life learning environment. It is not as likely that the learning activities studied here affected the proficiency test scores previous to the activities, nor the score improvement had factors to contribute to such activities.

The overall results indicate that all these learning activities were either directly or indirectly related to the proficiency score improvements, although the effects may have been exerted through various courses. Among these activities, directly involved were learning attitude related scores such as class attendance and task completion rates, and overall progress in class input learning measured by weekly achievement tests. In the case of the weekly achievement tests, the mean scores of listening skills, not only the overall achievement, were a good indicator of proficiency score improvements, while, at least in this study, it did not matter how well a student could score in grammar & vocabulary tests on average. Bulk vocabulary learning measured by the periodical vocabulary tests did not demonstrate direct influence over the score improvements, but they may have influenced the later proficiency test results.

It is noted that the types of learning similar in style to students' previous experience arguably depend on the skills and knowledge acquired through such past learning, and therefore are more difficult to observe their influence over proficiency test score improvements. This point needs further investigation to clarify.

References

Cohen, J.（1992）. A power primer. *Psychological Bulletin*. 112（1）, 155–159.

公式データ・資料. http://www.toeic.or.jp/bridge/data/data.html 2009年9月14日

特定非営利活動法人日本イーラーニングコンソシアム（編）（2008）. E ラーニング白書2008／2009年版　東京　東京電機大学出版局