Discriminant analysis on university students' reading skills:
Experiments using class learning outcomes and certificate exam scores

Toshihiko Shimauchi (Advanced Course, Komatsu College, shimauchi@komatsu-c.ac.jp)
Hidetaka Nambo (Graduate school of Natural Science & Technology, Kanazawa University, nambo@blitz.ec.t.kanazawa-u.ac.jp)
Haruhiko Kimura (Faculty of Production Systems Engineering and Sciences, Komatsu University, haruhiko.kimura@komatsu-u.ac.jp)

Abstract
Scores of English certificate exams have been drawing increasing attention as one of the indexes for English learning activities. The certificate exams can evaluate proficiency of test takers objectively. However, the exams are usually expensive to take, making it difficult for students to take ones frequently. If students' proficiency based on certificate exams' result can be estimated by using learning outcomes of English course, their objective proficiency can be assessed without additional cost. This study used learning outcomes of 76 students in a public university in Ishikawa. Their learning outcomes (weighted average scores of mid-term and final exams, assignment average scores, assignment average study hours, times of assignment submission) in an English reading course in 2018 fall semester and reading scores of TOEIC Bridge in February of 2019 were used for discriminant analysis. A model using weighted average sores of mid-term and final exams and assignments average scores yielded 89.04% accuracy to classify the students into upper and lower groups. The threshold of the two groups are set on 80 % correct score of TOEIC Bridge reading scores. The results also showed a necessity to give students appropriately calibrated assignments on regular basis.

Key words
EFL, TEOEIC Bridge, out-of-class-study, discriminant analysis, readability

1. Introduction

The Courses of Study determined by the Ministry of Education, Culture, Sports, Science and Technology (MEXT) of Japan are broad standards for all schools, from kindergarten through high schools, to organize their programs in order to ensure a fixed standard of education throughout the country (Ministry of Education, Culture, Sports, Science and Technology, n.d.). Since the end of World War II, the Course of Study have been revised 7 times, approximately at 10-year intervals. With the revisions, English education in Japan has gone through several reforms. The latest Courses of Study aim to nature communication skills in a foreign language (Koreto, 2017; Kumazawa, 2018; Tada, 2016).

The Course of Study does not apply to universities. However, a reform for university entrance examination has been implemented. In 2020, the current National Center Test will be replaced with the Common Test for University Admissions. In the new test system, applicants' English proficiency will be assessed in four skills (reading, listening, speaking and writing). In order to assess speaking and writing skills, a certificate exam will be used.

One of the certificate exams to be used is TOEIC. The test has been drawing attentions from business world, owing to the widening and acceleration of the economic globalization. A study conducted in 2013 to listed companies in Japan showed that 69 % of respondents used TOEIC score in evaluating job applicants' qualification. The test was also used widely in deciding who to be promoted or transferred to overseas offices (IIBC, 2013). As a result, 2.5 million people took the test in 2017, of which 1.1 million were students (IIBC, 2018).

In many private universities, the tests are used widely for entrance exam or credit certification. A survey conducted in December 2016 by the Institute for International Business Communication (IIBC), an administrator of the TOEIC in Japan, showed out of 751 universities surveyed, 427 used the test for entrance exam and 378 used the test for credit certification (IIBC, n.d.).

Although the test has been attracting as an objective certificate to assess examinee's proficiency, the cost is rather high. Depending on the test type, minimum of 5,000 JPY per test is necessary, making it harder to take one frequently. In an English course offered in a school, a workbook can be used for trial exam. However, duplicating the workbook is prohibited by the copyright law. If students have to buy the workbook, preprocessing is necessary to hide answer keys included in the workbook.

When a student can estimate the TOEIC score from own learning outcomes related to course activities, it is possible for him or her to decide when to take an official test based on the estimated scores. The estimation will also allow a teacher to give necessary instruction and advice to each student according to his or her proficiency.

This study is a first step for the estimation of the score from the learning outcomes. The discriminant analysis has been applied to classify students into upper and lower groups

based on TOEIC Bridge reading scores by using learning outcomes in an English course conducted in a public school in Ishikawa.

2. Existing studies

2.1 TOEIC Program

TOEIC is the abbreviation for Test Of English as International Communication and is developed by ETS, a non-profit educational foundation in the U.S. At the beginning, there was only one test type called TOEIC Listening and Reading (L&R). Later, a simplified test called TOEIC Bridge and a test to evaluate output performance called TOEIC Speaking and Writing (S&W) were developed. Table 1 shows a score comparison between TOEIC Bridge and TOEIC L&R.

Table 1: Score comparison: TOEIC Bridge and TOEIC L&R

Bridge	90	100	110	120	130	140	150	160
L&R	230	260	280	310	345	395	470	570

Source: IIBC.

2.2 Estimation of TOEIC score

Taguchi (2018) reports a correlation between TOEIC scores and self-study hours not related to the course based on data from students in Aichi University of Education. The study hours were self-reported in 6-point scale, so the reliability of the data is not sufficient.

There are many existing studies to estimate TOEIC scores from other certificate scores. Maruyama et al (2018) have conducted an experiment to estimate TOEIC scores from G-TELP scores using a generalized linear model. Dohi and Cheung (2014) estimate TOEC scores from TOEFL scores based on data of students in Chiba University by using a simple linear regression model. Eguchi (2011) calculates a correlation between TOEIC scores and TOEFL scores based on data of students in Hokusei University. These existing studies reports interesting findings. However, each estimation requires scores from another certificate test, which requires additional cost.

This study aims to overcome the cost problem using learning outcomes from course activities which do not require additional cost.

3. Experiment

3.1 Participants

First year students in the Faculty of Intercultural Communication of Komatsu University participated in this project. In fall semester of 2018, they were required to take "English III" course. Due to the physical size of the classroom, the course was divided into three classes held on Tuesdays, Wednesdays and Fridays. The students freely decided which class to register (Table 2).

76 students took TOEIC Bridge conducted in February of 2019.

Table 2: Course name, days and students registered

Class	Date	Male	Female	Total
English IIIa	Tue. 13:00-	3	20	23
English IIIb	Wed. 13:00-	12	12	24
English IIIc	Fri. 13:00-	2	33	35
Total		17	65	82

Three of them did not submit assignments described in 3.2.1.

Data from 73 students with both TOEIC Bridge and course assignments were used for discriminant analysis.

3.2 Textbook

"Skills for Better Reading, revised edition" published by Nanun-do was used for the course. According to the publisher, intended students for the text are expected to have TOEIC scores from 400 to 600. From Table 1, these scores can be interpreted as between 140 to 160 in TOEIC Bridge.

TOEIC Bridge is a simplified version of TOEIC. It consists of listening part (50 questions for 25 minutes) and reading part (50 questions for 35 minutes). The scores are 90 each (180 in total). Table 3 shows the test result conducted in February 2019. Compared to national average scores of 1st year university students (Total 124.0, Listening 61.1, Reading 62.9), the participating students have higher proficiency in reading and listening on average (IIBC, 2018).

Since one of the course objectives was to improve reading skills, the reading scores were analyzed in this study.

Table 3: Descriptive statistics of TOEIC Bridge scores

	Total	Listening	Reading
Mean	149.29	72.30	76.99
SD	13.22	6.38	8.39
Min	108	52	52
Max	174	90	90
Median	152	72	80
Skewness	–1.01	–0.45	–1.37

3.3 Learning outcomes

3.3.1 Assignments

The course was conducted over 16 weeks. For 13 weeks, reading assignments were given to the students. The assignments were adopted from written tests of EIKEN Grade P1 conducted and published by Eiken Foundation of Japan. A written test is composed of multiple-choice section (41 questions) and composition section (1 question). Out of 41 questions, 10 questions in reading sub-section were printed and distributed to the students as an assignment at the end of class each week. The sub-section is divided into three ques-

Table 4: Average words size for each question set

	QS1	QS2	QS3	Total
Number of questions	3	3	4	10
Words in texts	312	412	494	1,218
Words in questions	241	239	343	823
Words in total	553	651	837	2,041

Note: QS: question set.

tion sets, as shown in Table 4.

According to Eiken Foundation's evaluation criteria, the intended examinees of the Grade P1 are second to third year university students. Successful examinees are expected to fully understand and use English required in actual social life situations. The level of the test is difficult for the participating students. For the copyrights of the past test, section1, article 35 of the copy rights law is applied for this project so that the tests can be reproduced without a permission from the copyright holder.

3.3.2 Mid-term and final exams

The mid-term exam was conducted in November 2018. The final exam was held in February 2019. The questions were made based on the textbook used in the course. The mid-term exam covered 6 units in the textbook and was conducted separately for each class. 50 % of the questions were same among three classes and the remaining 50 % were different among each class. The final exam covered 4 units of the textbook and was conducted all together.

For the experiments, a weighted average score of mid-term and final exams was calculated for each student. Table 5 shows the statistical information and readability scores for each exam.

There are many readability scores (Klare, 1974; Zakaluk and Samuels, 1988). Flesch-Kincaid Grade Level (FKGL) is most widely used for assessing the appropriateness of a given textbook (Coleman and Liau, 1975). FKGL was developed by Flesch and Kincaid in 1975 and the number indicates an appropriate grade level in the U.S. school system for a document being analyzed. Ishioka et al (2010) report English entrance tests in the National Center Test are at 4 to 8 in FKGL. Table 5 shows the exams for the course were more difficult than the entrance examination commonly used in Japan.

Table 5: Statistical information and readability scores for each exam

Exam	Class	Sent	Words	Syll	Lett	FKGL	ARI
	A	65	860	1450	4432	9.5	9.5
Mid	B	65	917	1529	4687	9.6	9.7
	C	70	961	1600	4862	9.4	9.3
Final	all	62	963	1558	4523	9.6	8.5

Note: Sent; sentences, Syll; syllables, Lett: letters.

The formula of FKGL is as follows.

$$FKGL = 0.39 \times WPS + 11.8 \times SyPW - 15.59$$

where

WPS = Words Per Sentence
SyPW = Syllables Per Word

As the formula shows, information on syllables are required for each word to calculate FKGL. When there is a word in a text to be analyzed which does not have syllable information, FKGL cannot be calculated.

Automated Readability Index (ARI) was developed by U.S. Airforce and uses the number of letters to calculate the readability (Smith and Senter, 1967). Its result shows the appropriate grade level, same as FKGL. The formula is as follows:

$$ARI = 4.71 \times LPW + 0.5 \times WPS - 21.43$$

where

LPW = Letters Per Word
WPS = Words Per Sentence

As shown in Table 5, except for the final exam's ARI, the readability scores did not show significant fluctuations. The result allowed to suppose all the students took same level of examinations.

Abbreviations in Table 6 are used in tables and figures for

Table 6: Legend

Abbreviation	Meaning
WAS	Weighted average scores of mid and final exams
AAS	Assignment average scores
AHRS	Assignment study hours
ATMS	The number of submissions of assignment

Table 7: Descriptive statistics of learning outcome

	WAS	Assignment		
		AAS	AHRS	ATMS
Mean	71.17	5.15	1:11:15	10.77
SD	13.52	1.71	0:27:54	2.94
Min	19.42	1.67	0:30:00	2
Max	97.40	8.58	2:40:00	13
Median	73.15	5.18	1:05:50	12
Skewness	−0.97	−0.08	1.00	−1.75

the remainder of the paper. Table 7 shows the learning outcome data.

3.3 Software

Google Apps were used for collecting and scoring the assignments. The students were required to register study hours in addition to the answers to the assignment. Bell Curve for Excel (version 2.15) from Social Survey Research Information was used for discriminant analysis. Mint Reading Grade Level Formulas from Mint Phonetics Education Institute was used to calculate readability scores for the mid-term and final exams.

4. Results

4.1 Correlation

Data from 73 students were used to conduct discriminant analysis. The students were divided into upper and lower groups based on TOEIC Bridge reading scores. The threshold between the group was set at 80% correct scores. Explanatory variables were four learning outcomes listed in Table 7. The number of combinations of these 4 variables are 15. In order to simplify the analysis, the partial correlation coefficients were calculated to find a variable having a weak correlation with the reading score.

Table 8 and Figure 1 show that the number of submissions has a weak correlation with the reading score. This explanatory variable was excluded and the remaining three learning outcomes were used for discriminant analysis.

Table 8: Partial correlation matrix ($n = 73$)

	1	2	3	4	5
1 Reading Score	–	.317	.332	-.276	-.056
2 WAS	**	–	.462	.001	-.027
3 AAS	**	**	–	-.073	.362
4 AHRS			**	–	.098
5 ATMS					–

Note: ** $p < 0.01$, * $p < 0.05$.

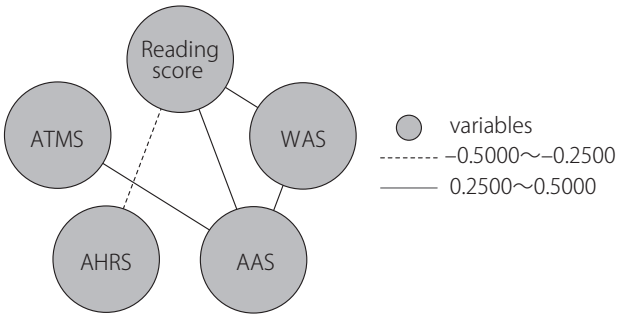

Figure 1: Undirected graph ($p < 0.05$)

4.2 Overview of discriminant analysis

The analysis overview is shown in Table 9.

Table 9: Experiment results;percentage of correct classifications

Model	Variable(s)	% of correct classifications		
		Upper	Lower	Total
1	a	85.00 %	76.92 %	83.56%
2	b	76.67 %	84.62 %	78.08 %
3	c	81.67 %	61.54 %	78.08 %
4	a × b	88.33 %	92.31 %	89.04 %
5	a × c	88.33 %	69.23 %	84.93 %
6	b × c	80.00 %	84.62 %	80.82 %
7	a × b × c	88.33 %	76.92 %	86.30 %

Note: a; WAS, b; AAS, c; AHRS.

The value for upper group corresponds with recall, the lower group value corresponds with specificity, and the total value corresponds with accuracy. Discriminant analysis does not yield precision. The data from the experiment were applied to the formula for precision to yield F-value as shown in Table 10.

Table 10: Experiment results: precision and *F*-value

Model	Variable(s)	Precision		*F*-value	
		Upper	Lower	Upper	Lower
1	a	0.9444	0.5263	0.8947	0.6501
2	b	0.9583	0.4400	0.8519	0.5591
3	c	0.9074	0.4211	0.8596	0.5556
4	a × b	0.9814	0.6316	0.9298	0.7365
5	a × c	0.9298	0.5625	0.9060	0.6873
6	b × c	0.9600	0.4783	0.8727	0.5986
7	a × b × c	0.9464	0.5882	0.9138	0.7062

Note: a; WAS, b; AAS, c; AHRS.

4.3 Detailed results for several models

Tables 11 and 12 respectively show the detailed results for models 4 and 6 which produced higher than 80% correct classification for both upper and lower groups.

Table 11: Experiment result of model 4

		Prediction		% of correct classification
		Upper	Lower	
Observation	Upper	53	7	88.33 %
	Lower	1	12	92.31 %
			Total	89.04 %

Table 12: Experiment result of model 6

		Prediction		% of correct classification
		Upper	Lower	
Observation	Upper	48	12	80.00 %
	Lower	2	11	84.62 %
			Total	80.82 %

Tables 13 and 14 respectively show the results for models 2 and 7, both with average assignment scores as an explanatory variable.

Table 13: Experiment result of model 2

		Prediction		% of correct classification
		Upper	Lower	
Observation	Upper	46	14	76.67 %
	Lower	2	11	84.62 %
			Total	78.08 %

Table 14: Experiment result of model 7

		Prediction		% of correct classification
		Upper	Lower	
Observation	Upper	53	7	88.33 %
	Lower	3	10	76.92 %
			Total	86.30 %

5. Discussion

5.1 Overview

As shown in Table 9, all the models except model 2 exceed 80 % for correct classification for upper group. However, the correct classifications for lower group varies from 60 % to 90 % among different models. As for precision, similar tendency can be observed from Table 10. Although precision for upper group exceed 0.9 in all the models, for lower group, the maximum is 0.6316 in model 4.

Incorrect classifications include actual lower group students being misclassified as upper group and vice versa. The former case will likely to occur when the difficulty levels of assignments and examinations are low. The latter case will likely to occur when these difficulties are high.

5.2 Model 2

This model is the only model with less than 80 % correct classification for upper group classification. Approximately 25 % of the students who actually belong to upper group were predicted to be lower group students. This model employs only one variable (average assignment scores) for the prediction. The assignments were adopted from EIKEN Grade Pre-1 tests, which were at a higher difficulty level. The result suggests that even those who belong to upper group could not have higher score for the assignment.

5.3 Models 4 and 6

Models 4 and 6 have higher correct classifications for lower group than those of upper group. Both models employ average assignment scores for the prediction.

Model 4 uses weighted average scores of exams, too. These two explanatory variables allowed higher prediction not only for lower group but also for upper group.

Model 6 uses assignment study hours, in addition to average assignment scores. Compared to model 2, correct classification improves just by 3 % point for upper group and precision improves just in 0.03 point. As shown in Table 8, partial correlation coefficient between assignment study hours and assignment average scores was negative weak (-0.073), leading to lesser improvement in prediction for additional variable.

5.4 Model 7

The model uses all the explanatory variables. Compared to model 4 which does not use assignment study hours, accuracy and precision for lower group yield lower numbers. As shown in Table 8, the partial correlation between assignment study hours and reading score was statistically insignificant. This partly explains degrade in prediction.

6. Conclusion and future direction

Using certificate exam is effective to objectively evaluate proficiency. TOEIC has been attracting attention from business world. However, the cost to take test is one of the obstacles for students to take one regularly. When prediction can be achieved through learning outcomes of English course, students can obtain necessary guideline for the preparation. Also, teachers can have valuable index for class management.

In this study, discriminant analysis was applied in order to predict upper and lower group of TOEIC Bridge reading scores from learning outcomes. The results showed that a model employing weighted average exam scores and average assignment scores predicted upper and lower group with 89.04 % of accuracy. However, for the lower group, the precision was lower, suggesting assignments with higher difficult level misclassified upper group students as lower group students. In order to predict students' proficiency accurately, assignments whose difficulty levels are calibrated properly according to students' skill are important.

One future direction of the study is to expand the participating students to include different faculties in order to find a better model in predicting students' reading proficiency. One of the authors is scheduled to teach technical English for students in a faculty of science. The new experiments and results

by applying this method to the students will be useful for many professors in similar faculties to review overall English course structures.

Another future direction of the study is a further analysis of study hours. Discussions in 5.3 and 5.4 suggest that although many upper group students spent shorter hours for the assignments, some used longer hours. Analyzing factors influencing the difference in study hours among the upper group students will lead to a new finding for more effective classification criteria.

References

Coleman, M. and Liau, T. L. (1975). A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, Vol. 60, No. 2, 283-284.

Doi, M. and Cheung, C. K. (2014). A descriptive and longitudinal analysis of TOEIC IP and TOEFL ITP scores at Chiba University. *Papers on Languages and Cultures*, No. 8, 15-32. (in Japanese)

Eguchi, H. (2011). An examination of correlations between TOEFL and TOEIC scores among Hokusei Gakuen Students : An attempt to find a score conversion formula. *Hokusei Review, the School of Humanities*, Vol. 48, No. 2, 35-44. (in Japanese)

IIBC (n.d.). TOEIC Tests usage: Entrance exam and credit certification. Retrieved March 22, 2019, from https://www.iibc-global.org/toeic/official_data/lr/search.html. (in Japanese)

IIBC (2013). Survey on English usage in listed companies in Japan. Retrieved March 22, 2019, from https://www.iibc-global.org/toeic/official_data/lr/katsuyo_2013.html. (in Japanese)

IIBC (2018). TOEIC Program DATA & ANALYSIS 2018. Retrieved March 22, 2019, from https://www.iibc-global.org/library/default/toeic/official_data/pdf/DAA.pdf. (in Japanese)

Ishioka, T., Takamitsu, H., and Otsu, T. (2010). Statistical analysis of National Center Test by using Natural Language Processing Techniques. *Journal of University Entrance Study*, No. 20, 145-150. (in Japanese)

Klare, G. R. (1974). Assessing readability. *Reading Research Quarterly*, Vol. 10, No. 1, 62-102.

Koreto, Y. (2017). The history of English language education in postwar Japan: Focusing on the revision of the course of study. *Bulletin of Educational Science*, No. 20, 1-12. (in Japanese)

Kumazawa, M. (2018). Reforms in English language education in Japanese secondary schools: Impacts and issues. *Studies in Language and Culture, J.F. Oberlin University*, No. 9, 17-31. (in Japanese)

Maruyama, M., Ogasawara, S., and Utsunomiya, Y. (2018). Estimating the TOEIC scores from the G-$_{TEL}$P scores by the generalized linear model: From the data obtained from Nagasaki University Students from 2011 to 2016. *Journal of Center for Language Studies, Nagasaki University*, No. 6, 33-51. (in Japanese)

Ministry of Education, Culture, Sports, Science and Technology (n.d.). Improvement of academic abilities (Courses of Study). Retrieved April 2, 2019, from http://www.mext.go.jp/en/policy/education/elsec/title02/detail02/1373859.htm.

Smith, E. A. and Senter, R. J. (1967). *Automated readability index* (No. AMRL-TR-66-22). Wright-Patterson AFB, Ohio: Aerospace Medical Division.

Tada, M. (2016). Recent reform to the English education system in Japan. *21st Century Education Forum*, No. 11, 21-29.

Taguchi, T. (2018). TOEIC, study time, and grit: A case study of Aichi University of Education. *Liberal Arts and Education, Aichi University of Education*, No. 18, 1-9. (in Japanese)

Zakaluk, B. L. and Samuels, S. J. (1988). *Readability: Its past, present, and future*. International Reading Association.

(Received: April 4, 2019; Accepted: April 17, 2019)