Does English Ability Really Matter ?
— Validating the Final Grades of University EFL Courses

Yoshinori Watanabe

Introduction

It is normal that the final course grade includes various factors other than the target ability that is to be developed in the course. University courses of English as a Foreign Language are no exception. The final grade normally includes various extraneous factors, such as efforts, submission of homework assignments, attitudes, attendance, and so forth. The problem with this type of practice is that the information thus 'polluted' with extraneous factors cannot be a valid indicator, in the sense that the students' ability cannot be inferred therefrom. This means that this type of evaluation is not held accountable for the outsiders who wish to refer to it a source of information when they would like to make a decision based on it. It would certainly be ideal if the summative information in the form of a final grade reflects students' true competence in the target language. However, if it is inevitable that the grade includes a variety of extraneous factors, it becomes crucial to indicate what sorts of information have been referred to in giving a grade, and to provide an independent measure indicating the accurate proficiency level of course participants.

The present research was an attempt to examine how much English ability is reflected in the final grade of the EFL courses offered at Akita University. Because of a paucity of hard data in the field to date, the purpose of the research was not to test a specific set of hypotheses, but rather descriptive, or a hypothesis generating exercise at best, the result of which should be tested in more focused research in the future.

Research questions

The questions that the present research addressed were as follows;
1. Would the final grade help understand the English ability of those students who would have completed the course? If it would help, how much would it help?
2. Which of the factors, if any, would best predict the student's final attainment level of and improvement in an English ability?
3. What would be the degrees of relative contribution of various factors included in the final grade for us to understand an English ability of the students?

Participants

A total of 136 students learning English as one of the required foundation course subjects took part in the present study; 42 freshmen were from the Faculty of Education and Human Studies (hereafter, Education), 52 freshmen and 42 sophomores were from the Faculty of Engineering and Resourse Science (hereafter, Engineering 1 and Engineering 2 respectively).

The freshmen course was a required credit course that was offered to a total of

approximately 600 students, running through the whole year consisting of two terms. The purpose of the first term of the course was to help students develop listening skills, whereas the second term was to help them develop writing skills. The students were divided into 18 classes, each of which consisted of approximately 50 students, in addition to several re-takers. The classes were taught by ten teaching staff including part-time teachers, who were supposed to use the same textbook, and measure students' performance by the same test (i.e., program-wide tests). The present research dealt with those students who had taken the courses of the first term. The textbook used for the listening course was *California Dreaming* (Seibido, publishing inc. 2003), which contained various exercises that were intended to develop students' listening comprehension skills by means of a variety of interviews that were video-recorded in authentic situations in the USA. The types of activities students were engaged in involved note-taking, production of short-sentences based on the notes, identifying key words, shadowing, and a gap-filling summary.

Unlike the freshmen component, the sophomore component did not require the teachers to use the same textbook, nor a common test. The textbook that the present writer used in his course was *Reader's Choice* (Michigan University Press 1993). The textbook included a variety of exercises intended to help students develop reading sub-skills, including skimming, scanning, inference, restatement, prediction, and so forth. The textbook also included vocabulary and structure exercises.

Both courses ran though the first term of 2003, consisting of 30 90-minutes sessions.

Data sets

The following sets of data were used for the present investigation. All the data sets were originally gathered to serve as sources of information for providing students with the feedback on their levels of achievement on a regular basis, and for giving a final grade. There were no data that were specifically prepared for the purpose of the research.

The ACE test

The first data set was the test scores derived from the administration of the Assessment of Communicative English or ACE (by Kirihara Publishing Company). This test was intended to measure the English use ability of intermediate and advanced levels of senior high school students. The test was based on Item Response Theory (IRT), "a systematic procedure for considering and quantifying the probability or improbability of individual item and person response patterns given the overall pattern of responses in a set of test data" (Henning, 1989, pp. 108-109). Among three families of IRT, the two-parameter model was employed for the present version of the test, taking account of a scale of person ability and item difficulty, and a continuous estimate of discriminability. The test consisted of three components: Listening (30 items), Reading (20 items), and Grammar/Vocabulary (48 items), but for the present research only relevant components were used for each course: Listening for the two courses offered for the first year students, and Reading and Grammar /Vocabulary for the sophomores. Two parallel versions of the test were administered at the beginning of the term in April (version AL006), and at the end of the term in July (Version AL009). (See http://www.kirihara-kyoiku. net/BACE-ACE/moushikomi.html for details of the ACE test).

Class attendance

The frequency of each student attending 30 sessions of the course was computed in percentages. Attendance was considered to be an

important part of the course, and this had been announced to all the students in a written form at the beginning of the course. Also to be noted is that those students who were late for the class were considered to be absent from the class, which had also been announced to the students at the beginning of the course.

Homework assignments

In the listening courses, two pieces of homework were assigned, whereas 14 pieces were assigned for the reading course. The homework for the listening course involved diagramming the content of the interview script that had been dealt with in the lesson. In the next lesson, students were asked to present and explain the diagram to the member of a group. The homework for the reading course involved not only diagramming the text structure, but also other reading related activities such as drawing up a semantic network of the vocabulary, summarizing a text, and so forth.

In-class tests

In-class tests were administered two times for the Education 1 course and four times for Engineering 1. No in-class tests were administered for Engineering 2. The purpose of the tests was to measure how much students had learned in the previous part of the lesson, thereby giving them a sense of achievement. The test scores also comprised part of the final grade. Thus, the content and method of the tests were closely related to what had been taught in the lesson. For example, listening in-class tests involved note-taking by listening to the passages which had been covered in the class. However, for the test, audio-recording rather than VTR was used, for a practical reason that repeating the video recording three times in the testing situation was extremely difficult.

Program-wide tests

The program-wide test was a type of achievement test that was intended to measure the levels of achievement of the students on a common ground. The test was administered twice, in the middle and at the end of the term. To construct a test, two teams, each consisting of two teachers, were formed, and one team took charge of producing a test paper for the Education course, while the other took charge of the Engineering course. Slightly different policies were employed by the two teams, and the difference seemed to be reflected in the test contents. The tests for the Education course consisted of virtually the same material as the textbook, while the tests for the Engineering course were made up of the content which was slightly different from the course textbook.

There were no program-wide tests for the Engineering 2 students.

Self-evaluation

To examine the degree of students' awareness about their own proficiency level of English, the self-evaluation instrument called DIALANG was administered. The test was a self-evaluation measure developed by the Council of Europe. The translation was made with the cooperation of Satsuki Shirasawa, a senior of Akita University, who was writing her senior thesis using the instrument. For the present research, only the listening component (consisting of 44 items) of the system was used. The listening component, like the other components was, divided into three major levels of proficiency: basic user (A), independent user (B), and proficient user (C). (See Council of Europe, 2001 for details). Students were asked to rate their proficiency according to a five-point scale, ranging from 'very difficult' (1) to 'very easy' (5). The reading component had also been prepared, but it could not have been administered due to the time constraint

of the course schedule.

Final grade

The final grade to be reported was a single letter grade, which consisted of the aggregation of all the above data sets except for the ACE scores. The letter grade was based on the following standards: 'A' for the students whose final amalgamated scores amounted to 80% and above, 'B' for 70% to 79%, 'C' for 60% to 69%, and 'D' (fail) for 59% and below.

Data Analysis

All the data sets were analyzed by SPSS (Statistical Package for Social Sciences, Version 11.0). The analysis dealt with the ACE scores as dependent variables, so it may be possible to examine which of the other independent variables would be the best predictor of the ACE gain scores and the scores obtained by the second administration of the test. The ACE gain scores were computed by subtracting the scores of the second administration

Table 1

Descriptive Statistics

| | | Listening | | Reading |
		Education 1 (N=42)	Engineering 1 (N=52)	Engineering 2 (N=42)
ACE 1	Mean	174.209	160.636	320.881
	SD	20.951	20.300	53.375
	α	.503	.503	.826
ACE 2	Mean	193.674	176.527	330.095
	SD	24.731	29.396	50.995
	α	.537	.646	.817
ACE gain	Mean	19.571	15.230	9.214
	SD	17.242	27.691	38.410
homework	Mean	0.685	48.885	0.62
	SD	0.163	11.066	0.17
attendance	Mean	0.960	0.969	0.945
	SD	0.060	0.037	0.081
inclass fin	Mean	0.820	0.521	0.640
	SD	0.131	0.114	0.155
final grade	Mean	0.771	0.560	0.726
	SD	0.090	0.112	0.083
inclass mid	Mean	0.630	0.712	0.640
	SD	0.170	0.156	0.129
pwt mid	Mean	0.750	0.701	NA
	SD	0.130	0.150	NA
pwt final	Mean	0.793	1.500	NA
	SD	0.120	0.728	NA
DIALANG	Mean	2.370	0.694	NA
	SD	0.380	0.99	NA
	α	.952	.989	NA

Notes: ACE gain = gain scores in the second administration of the ACE tests. ACE1 = scores of the first administration of the ACE test. ACE2 = scores of the second administration of the ACE test. homework = submission of homework assignments. attendance = percentage of attendance. inclass fin = in-class final test. final grade = final grades expressed in percentages. inclass mid = in-class mid-term test. pwt mid = program-wide mid-term test. pwt final = program wide term-end test. DIALANG = self-evaluation. NA = Not applicable.

from those of the first administration.

Results

The results of the analyses are provided in Tables 1 through 16. Below is a summary of two sets of findings that were made. The first set of findings is based on basic statistics, and the second set is based on the results of multiple regression analyses.

First, the reliability of the listening tests was not so high as had been expected: in the case of Education, $\alpha = .503$ for the first administration of the ACE test (ACE1), and $\alpha = .537$ for the second administration (ACE2),

and $\alpha = .503$ (ACE1) and $\alpha = .646$ (ACE2) in Engineering 1 (Table 1). On the other hand, fairly high reliability was established for the reading component of the ACE: $\alpha = .826$ (ACE1) and $\alpha = .817$ (ACE 2). It follows that the present results, especially those based on multiple regression analyses, should be interpreted with caution. Also, to be noted is that the reliability of other measures, including in-class tests, and program-wide tests, was not established, so correction for attenuation could not be carried out, which also necessitated the cautious interpretation of the results.

Second, though reservations are required, perhaps good news is that the scores of the lis-

Table 2

Pearson product moment correlations between variables (Faculty of Education and Human Studies) (N= 42)

	homework	inclassmid	inclassin	pwt mid	pwt final	DIALAN	final grade	ACE1	ACE 2	ACE gain
attendance	.130	.080	.171	.135	.192	-.075	.294*	.232	.131	-.091
homework		.306*	.275*	.292*	.105	.225	.589**	.035	.128	.154
inclass mid			.357*	.659**	.663**	.041	.829**	.430**	.462**	.153
inclass fin				.392	.319*	.097	.639**	.341*	.244	-.058
pwt mid					.572**	-.013	.791**	.569**	.635**	.245
pwt final						-.150	.716**	.384*	.488**	.255
DIALANG							.062	-.028	-.045	-.031
final grade								.477**	.520**	.191
ACE 1									.733**	-.165
ACE 2										549**
ACE gain										-

Notes: ** = $p < .01$ level (2-tailed). * = $p < .05$ (2-tailed).

Table 3

Pearson product moment correlations between variables (Faculty of Engineering and Resource Science 1) (N = 52)

	In-class	Inclass mid	Inclass fin	pwt mid	pwt final	DIALANG	Final grade	ACE 1	ACE 2	ACE gain
attendance	.003	.299*	.201	.140	.194	.116	.394**	-.137	-.072	.024
In-class		.018	.342**	.066	-.095	-.137	.140	.045	.307*	.298*
Inclass mid			.479**	.520**	.477**	-.138	.870**	.384**	.240	-.028
Inclass final				.590**	.365**	-.264	.719**	.241	.092	-.081
PWT mid					.239	-.085	.669**	.307*	.259	.051
PWT final						.024	.608**	.250	.430**	.278*
DIALANG							-.131	-.153	-.080	.029
Final grade								.262	.199	.019
ACE 1									.449**	-.261
ACE 2										.745**
ACE gain										-

Notes: ** = $p < .01$ level (2-tailed). * = $p < .05$ (2-tailed).

Table 4

Pearson product moment correlations between variables (Faculty of Engineering and Resource Science 2) (N=42)

	Mid-term	Final test	Home work	Final Grade	ACE1	ACE2	ACE gain
Mid-term		.028	.508**	.694**	.651**	.573**	-.144
Final test			-.200	.302*	.055	.152	.125
Homework				.822**	.339*	.209	-.195
Final Grade					.474**	.403**	-.123
ACE1						.730**	-.420**
ACE2							.313*

Notes: Final test = final test score. Homework = Submission of homework assignments. Final grade = the final grade. ACE gain = gain scores of the ACE tests, ACE 2 = the score of the first administration of the ACE test. ACE 2 = the score of the second administration of the ACE test. ** = $p < .01$ level (2-tailed). * = $p < .05$ (2-tailed).

tening tests significantly increased from ACE1 to ACE2 both in Education (t = -7.487, df = 42, p <.001) and in Engineering 1 (t = -4.730, df = 54, p <.001). There was also a tendency indicating some increase for Engineering 2, though the difference was not significant (t = -1.555; df = 41, ns).

Third, the final grades moderately correlated with the ACE2 scores in the listening course for Education (ACE1, r = .520, p<.001; ACE2, r=.520, p<.001) (Table 2), and in the reading course for Engineering 2 (ACE 1, r=.474, p<.001; ACE2, r=.403, p<.001) (Table 4). The reason is yet to be established as to why there was very little correlation between these two variables in Engineering 1 (ACE1, r=.262, ns; ACE2, r=.199) (Table 3), but these results could be interpreted to indicate that the final grade may reflect part of language competence the students had developed during the course.

Fourth, the final grades did not correlate with the ACE gains scores in any case: Educa-

tion 1 (r=.191, ns.) (Table 2), Engineering 1 (r=.019, ns.) (Table 3), or Engineering 2 (r= -.123, ns.) (Table 4). These results indicate that the final grades did not provide accurate information regarding the degree of improvement of the students' proficiency.

Next, the results of multiple regression analysis are presented. As noted at the beginning of this paper, the present research was a descriptive study, rather than testing specific hypotheses, because it was not possible to identify and exclude specific variables prior to the analysis. In using multiple regression analysis, then, the "enter" command rather than stepwise or other methods of entering variables was employed.

As Tables 5 through 16 show, the results of multiple regression analyses were quite mixed. First, the variables in the present model predicted very little of the ACE gain scores of Education, where the variables accounted for only .166 (Table 6), and only .052 for Engineering 2 (Table 14).

Table 5

Predictor Variables for ACE gain scores (Faculty of Education and Human Studies) (N = 42)

	Unstandardized Beta	Coefficients Std. Error	Standardized Beta	t	Sig.
(Constant)	27.252	52.420		.520	.607
homework	19.990	18.596	.189	1.075	.290
attendance	-48.377	49.898	-.158	-.970	.339
inclass fin	-27.737	23.347	-.210	-1.188	.243
inclass mid	-18.183	24.883	-.179	-.731	.470
pwt mid	31.214	30.153	.229	1.035	.308
pwt final	46.141	32.911	.319	1.402	.170
DIALANG	-.325	7.534	-.007	-.043	.966

Note: Dependent Variable = ACE gain

Table 6

Model Summary of Multiple Regression Analysis and ANOVA for Predictor Variables and ACE gain scores (Faculty of Education and Human Studies) (N = 42)

Multiple R	.407(a)
R Square	.166
Adjusted R Square	-.006
Std. Error of the Estimate	17.29126

Analysis of Variance

	Sum of Squares	df	Mean Square
Regression	2022.702	7	288.957
Residual	10165.583	34	298.988
Total	12188.286	41	
F = .966		Significance of F = .471	

Notes: Predictors: (Constant), DIALANG, pwt mid, attendance, homework, inclass fin, pwt final, inclass mid. Dependent Variable: ACE gain

Table 7

Predictor Variables for ACE2 (Faculty of Education and Human Studies) (N = 42)

	Unstandardized Beta	Coefficients Std. Error	Standardized Beta	t	Sig.
(Constant)	95.711	62.571	1.530		.135
homework	-14.035	22.198	-.095	-.632	.531
attendance	9.386	59.561	.022	.158	.876
inclass fin	-7.933	27.868	-.043	-.285	.778
inclass mid	2.179	29.702	.015	.073	.942
pwt mid	104.780	35.992	.551	2.911	.006
pwt final	28.941	39.285	.144	.737	.466
DIALANG	.637	8.993	.010	.071	.944

Note: Dependent Variable: ACE 2

Table 8

Model Summary of Multiple Regression Analysis and ANOVA for Predictor Variables and ACE2 (Faculty of Education and Human Studies) (N = 42)

Multiple R	.623
R Square	.388
Adjusted R Square	.262
Std. Error of the Estimate	20.63968

Analysis of Variance

	Sum of Squares	df	Mean Square
Regression	9192.026	7	1313.147
Residual	14483.879	34	425.996
Total	23675.905	41	

F = 3.083   Significance of F = 0.13

Notes: Predictors: (Constant), DIALANG, pwt mid, attendance, homework, inclass fin, pwt final, inclass mid. Dependent Variable: ACE 2

Table 9

Coefficients of Predictor Variables for ACE gain scores (Faculty of Engineering and Resource Science 1) (N = 52)

	Unstandardized Beta	Coefficients Std. Error	Standardized Beta	t	Sig.
(Constant)	-89.073	107.316		-.830	.411
In-class	.732	.342	.293	2.139	.038
attendance	58.409	109.193	.078	.535	.596
PWT mid	4.734	59.419	.019	.080	.937
PWT final	147.626	52.266	.595	2.825	.007
Inclass mid	-4.948	63.858	-.028	-.077	.939
Inclass final	-38.440	50.563	-.209	-.760	.451
Final grade	-51.343	176.257	-.183	-.291	.772
DIALANG	-.252	.720	-.049	-.350	.728

Note: Dependent Variable: ACE gain

Table 10

Model Summary of Multiple Regression Analysis and ANOVA for Predictor Variables and ACE gain (Faculty of Engineering and Resource Science 1)

Multiple R	.536
R Square	.287
Adjusted R Square	.155
Std. Error of the Estimate	25.45815

Analysis of Variance

	Sum of Squares	df	Mean Square
Regression	11236.172	8	1404.521
Residual	27869.059	43	648.118
Total	39105.231	51	

F = 2.167   Significance of F = 0.50

Notes: Predictors: (Constant), grade raw, In-class, DIALANG, attendance, PWT final, PWT mid, Inclass final, Inclass mid. Dependent Variable: ACE gain

Table 11

Coefficients of Predictor Variables for ACE2 (Faculty of Engineering and Resource Science 1) (N=52)

	Unstandardized Beta	Coefficients Std. Error	Standardized Beta	t	Sig.
(Constant)	-75.453	99.794	-.756		.454
In-class	.741	.338	.296	2.193	.034
attendance	54.902	107.684	.073	.510	.613
PWT mid	-3.868	54.135	-.016	-.071	.943
PWT final	143.942	50.792	.581	2.834	.007
Inclass mid	-7.256	62.912	-.041	-.115	.909
Inclass final	-40.839	49.640	-.222	-.823	.415
Final grade	-61.531	172.339	-.219	-.357	.723
DIALANG	-.268	.712	-.052	-.376	.708

Note: Dependent Variable: ACE2

Table 12

Model Summary of Multiple Regression Analysis and ANOVA for Predictor Variables and ACE gain (Faculty of Engineering and Resource Science 1) (N=52)

Multiple R	.673
R Square	.453
Adjusted R Square	.352
Std. Error of the Estimate	24.13004

Analysis of Variance

	Sum of Squares	df	Mean Square
Regression	20771.398	8	2596.425
Residual	25037.122	43	582.259
Total	45808.519	51	
F = 4.459		Significance of F = 0.01	

Notes: Predictors: (Constant), grade raw, In-class, DIALANG, attendance, PWT final, PWT Mid, Inclass final, Inclass mid. Dependent Variable: ACE 2.

Table 13

Coefficients of Predictor Variables for ACE gain scores (Faculty of Engineering and Resource Science 2) (N=42)

	Unstandardized Beta	Coefficients Std. Error	Standardized Beta	t	Sig.
(Constant)	-5.220	107.651		-.048	.962
attendance	101.814	208.298	.215	.489	.628
Mid-term	33.110	119.967	.112	.276	.784
Final test	72.799	75.567	.294	.963	.342
Final grade	-205.372	306.741	-.442	-.670	.507

Note: Dependent Variable: ACE gain (scores in vocabulary, grammar and reading components).

Table 14

Model Summary of Multiple Regression Analysis and ANOVA for Predictor Variables and ACE gain scores (Faculty of Engineering and Resource Science 2) (N=42)

Multiple R	.229
R Square	.052
Adjusted R Square	-.050
Std. Error of the Estimate	39.36099

Analysis of Variance

	Sum of Squares	df	Mean Square
Regression	3165.431	4	791.358
Residual	57323.641	37	1549.288
Total	60489.071	41	
F = .511		Significance of F = .728	

Notes Predictors: (Constant), final grade, final test, mid-term, attendance. Dependent Variable: ACE gain

Table 15

Coefficients of Predictor Variables for ACE2 (Faculty of Engineering and Resource Science 2) (N=42)

	Unstandardized Beta	Coefficients Std. Error	Standardized Beta	t	Sig.
(Constant)	93.103	117.145		.795	.432
Attendance	200.673	226.669	.318	.885	.382
Mid-term	346.285	130.547	.879	2.653	.012
Final test	112.665	82.232	.343	1.370	.179
Final grade	-338.235	333.793	-.548	-1.013	.317

Note : Dependent Variable: ACE2 (vocabulary, grammar, reading)

Table 16

Model Summary of Multiple Regression Analysis and ANOVA for Predictor

Variables and ACE2 (Faculty of Engineering and Resource Science 2) (N=42)

Multiple R	.603
R Square	.363
Adjusted R Square	.295
Std. Error of the Estimate	42.83238

Analysis of Variance

	Sum of Squares	df	Mean Square
Regression	38740.942	4	9685.235
Residual	37880.677	37	1834.613
Total	106621.619	41	
F = 5.279		Significance of F = .002	

Notes: Predictors: (Constant), final grade, final test, mid-term test, attendance. Dependent Variable: ACE2

Second, the variables of the course for Engineering 1 accounted for .287 (F=2.167, p<.005), where the largest contribution was made by the program-wide test final (t = 2.825, p=.007), and the second largest by in-class tests (t = 2.139, p=.038). These variables met the guideline regarding the identification of useful predictors; i.e., t values well below -2 or above +2 (SPSS Inc., 1999, p. 209).

Third, in the case of ACE2, or the final attainment level of students, meaningful predictor variables were identified in all the three cases. In Education 1, the variables accounted for R^2=.388 (F=3.083, p=.013) (Table 8) in Engineering 1, R^2=.453 (F=4.459, p=.001) (Table 12), and in Engineering 2, R^2=.363 (F=5.279, p=.002) (Table 16). Among others, the program wide mid-term exam was the strongest predictor (β=104.780, unstandardized, β=.551 standardized; t=2.911, p=.006) in Education (Tables 7 and 8), in-class tests (β=.741, unstandardized, β=.296, standardized; t=2.193, p=.034) and the program wide final test (β=143.942, unstand-

ardized, β=.581, standardized; t=2.834, p=.007) in Engineering 1 (Table 11), and the mid-term exam (β=346.285, unstandardized, β=.879, standardized; t=2.653, p=.012) in Engineering 2 (Table 15).

Discussion

The present research set out to answer the following questions: (1) Would the final grade help understand the English ability of those students who would have completed the course? If it would help, how much would it help? (2) Which of the factors, if any, would best predict the student's final attainment level of and improvement in an English ability? (3) What would be the degrees of relative contribution of various factors included in the final grade for us to understand an English ability of the students?

The results were too complex to generalize to answer these questions directly. And yet, several tentative answers could be suggested. First, the final grade did not seem to be a predictor of students' increase in proficiency,

nor the level of their final attainment. The grade seemed to be polluted, so it may not be a suitable indicator to certify the level of proficiency that the students have gained (or lost) by the end of the term. Second, there were no cases where self-evaluation was an accurate predictor of gain scores, nor the final score of the ACE test. Nowadays, it is becoming increasingly popular to incorporate self-evaluation in the classroom evaluation(e.g., Benson, 2001). However, the present result indicates that this type of information may not be an accurate measurement for indexing the students' levels of proficiency. Thus, care must be exercised to use this type of instrument for the purpose of measuring students' proficiency. Third, nevertheless, there were several measures that may independently predict proficiency and final attainment levels of students. For example, in-class tests that are carried out on a regular basis may be one, and the tests that are carried out on a university-wide basis may be another. And yet, again, these instruments should be used with caution, if they are used to certify students' levels of proficiency.

Conclusion

The present research was an attempt to validate the final grade of university English courses, on the basis of external proficiency test scores. Though the reliability of the external test was not sufficiently high, the results indicated that the final grade was not 'pure,' but rather contaminated with various extraneous factors, such as efforts, attitudes, and so forth, which in turn would decrease the reliability of the final grade. This does not mean, however, that to include various factors in the final grade is wrong. The result simply suggested that the final grade could not be a valid indicator of students' actual proficiency levels. It follows that to indicate the final attainment of students as well as their improvement throughout the course, some independent measures should be provided.

References

Benson, P. (2001). *Antonomy in Langnage Learning*. London: Longman.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

Henning, G. (1989). *A Guide to Language Testing: Development, Evaluation, Research*. New York: Newbury House.

SPSS Inc. (1999). *SPSS® Base 10.0 Application Guide*. Chicago: SPSS Inc.