慶應義塾大学学術情報リポジトリ
Keio Associated Repository of Academic resouces

Title	Constructing a large-scale placement test for measuring students' English proficiency
Sub Title	英語能力測定を目的としたプレイスメントテスト作成法
Author	中村, 優治(Nakamura, Yuji)
Publisher	慶應義塾大学日吉紀要刊行委員会
Publication year	2010
Jtitle	慶應義塾大学日吉紀要. 言語・文化・コミュニケーション (Language, culture and communication). No.42 (2010.) ,p.1- 19
JaLC DOI	
Abstract	
Notes	
Genre	Departmental Bulletin Paper
URL	https://koara.lib.keio.ac.jp/xoonips/modules/xoonips/detail.php?koara_id=AN10032394-20101231-0001

慶應義塾大学学術情報リポジトリ(KOARA)に掲載されているコンテンツの著作権は、それぞれの著作者、学会または出版社/発行者に帰属し、その権利は著作権法によって保護されています。引用にあたっては、著作権法を遵守してご利用ください。

The copyrights of content available on the KeiO Associated Repository of Academic resources (KOARA) belong to the respective authors, academic societies, or publishers/issuers, and these rights are protected by the Japanese Copyright Act. When quoting the content, please follow the Japanese copyright act.

Constructing a large-scale placement test for measuring students' English proficiency

Yuji Nakamura

I. Rationale of the Research

Keio University (faculty of letters) has administered an in-house placement test to incoming freshman students and new sophomore students since the spring of 2006 (placement tests (hereafter PT) are administered twice a year, once at the beginning of the academic year and a confirmation test (hereafter CT) at the end of the academic year. These placement tests measure students' reading ability and overall proficiency in English to provide streamed instruction appropriate to their proficiency levels in order to optimize their learning experience, and to provide multi-faceted English communicative skills.

The goals of this project are as follows:

1) To offer EFL four levels of classes for students according to their English reading ability as ascertained by the method below.
2) To offer classes for those who, according to the method below, need remedial instruction.
3) To offer classes for those who have already reached the required level and desire further study.

Although commercialized tests such as the TOEFL-ITP, TOEIC-IP, G-TELP, Step-EIKEN and CASEC exist, it was agreed among members of the faculty of letters at Keio University that the content, level and purpose of those tests were not appropriate for the accurate academic placement of students. Furthermore, it was agreed that the admissions test could not be used for any other purpose than the entrance examination selection. Admission tests are basically used for screening purposes only. Although there are a variety of admission tests conducted

(e.g. admissions office tests, interview tests, the center-test, the high-school recommendation test), these tests generally are not a dependable predictor for placement purposes. In addition, because people are so concerned about privacy security issues even on scores of the admission tests, it seems difficult to use the admission test results for streaming instruction purposes.

Tests can be valid or not depending on whether they agree with the purpose of the test users. The purpose of the aforementioned tests does not seem to suit the needs of the faculty of letters at Keio University. For example, we are not solely intent on measuring students who will study overseas, or assessing the skills of students who will start business communication after graduation. Our purpose for this project is to encourage students to develop their English reading ability, which is indispensable for their major area studies. Almost all the students in the faculty of letters are required to read materials in English whether their major is English or not. For these reasons, we have decided to develop our own placement test.

In order to construct the whole component of the test, we take into consideration the Michigan English Language Assessment Battery (MELAB), which has been widely recognized as a valid proficiency test and is similar to what we intend to establish.

Reading ability is thought to consist of grammar knowledge, vocabulary knowledge, long passage reading comprehension with full context (e.g. the text material does not have any deleted words or blanks intended for other questions), and passage understanding without sufficient context or information. In other words, a test of reading ability should be composed of grammar, vocabulary, reading comprehension and gap-filling items (cloze).

Along with the PT which is conducted at the beginning of each academic year, a similar format confirmation test (CT) has been administered to examine the effectiveness of the reading-centered curriculum at the end of the academic year.

II. Purpose of the Study

The purpose of the present study is three-fold: 1. to examine the validity and reliability of the test using the Rasch-based statistical program; 2. to report the consecutive four-year investigation of students' reading ability; and 3. to suggest the implementation of an item bank which will eventually have 400 linked items by means of equation.

The validity can be examined whether the results fit the model or not in the Rasch measurement analysis.

Another way of assessing the construct validity of a test is to correlate the different test components with each other. Since the reason for having different test components is that they all measure something different and therefore contribute to the overall picture of language ability attempted by the test, we should expect these correlations to be fairly low–possibly in the order of .3–.5. The correlations between each subtest and the whole test might be expected to be higher-possibly around .7 or more since the overall score is taken to be a more general measure of language ability than each individual component score. And it is common in internal correlation studies to correlate the test components with the test total minus the component in question (Alderson et.al 1995).

Also, the content validity is discussed by using the questionnaire analysis. A test is said to have content validity if the questions reflect the course content or syllabus. The researcher conducted a questionnaire analysis.

The reliability is investigated by the Cronbach alpha index. The benchmark for an acceptable boundary is over 0.75.

The consecutive four-year investigation is examined through the comparison of students' test results (descriptive statistics and latent traits).

The 50 items in each test are linked by 12 anchor items for the whole item bank. Once all the items are calibrated and the difficulty of each item is determined, each item can be put on the continuum of the scale according to their logit scores (difficulty level). These items along with a task can be stored as items in a bank.

III. Research Design and Test Method

a. Materials (Test Design and Content)

The test material, which was used for PT and CT, was based on the MELAB format. The test contained 15 grammar MC questions, 10 vocabulary MC questions, 10 gap-filling MC questions, and 3 long reading passages with 5 MC questions each. Applicants had 60 minutes to complete this test, which was scored by optical readers. The reading section consisted of one beginning level, one intermediate level, and one advanced level passage about 400–500 words in length. Difficulty was rated impressionistically by teachers in terms of content, topic, and vocabulary level.

The test format (for PT and CT) in Table 1 and the test content in Table 2 are as follows:

Table 1 Test Format

Category	Grammar	Vocabulary	Gap-filling	Reading Comprehension
No. of Items	15	10	10	15
Test format	Discrete point	Discrete point	One passage with 10 blanks	3 passages with five questions each
Anchor items	4 items	3 items	None	5 items (one passage with five items)

Table 2 Test Content

Grammar	15 discrete point items; multiple choice questions (MCQ)
Vocabulary	10 discrete point items; MCQ
Gap-filling	One passage with 10 blanks; Gap-filling questions (MCQ)
Reading Comprehension	3 passages with 5 comprehension questions each; MCQ

Except PT1 (2006), each test is anchored by 12 items so that the four-consecutive year information can be obtained. However, for the gap-filling section, because of its uniqueness which requires the total integrative ability, it was difficult to utilize anchor items. Therefore, this section was not included in the three-consecutive- year study. Also, the 2008 CT which was not used for test equating was excluded from the longitudinal analysis. Eventually, three pairs of test takers- (one: PT 1 and CT1, another: PT2 and CT2, and the other: PT4 and CT4)- were compared for the four-consecutive-year analysis.

b. Procedures (Test construction, Administration, Timing)

Test Construction

Nakamura (1998, p.260) proposed four points to consider in assessing reading ability: 1) the nature of reading, 2) the theoretical or linguistic underpinnings of reading, 3) the test format of reading, 4) classroom teachers' ideas based on their teaching experiences. The construct of "reading ability" for this test was established mainly from these points plus the specific aspects of the faculty of letters as follows:

1) the teachers' teaching experience
2) the reading sections of other existing tests
3) linguistic theories (Alderson, 2000; Grabe, 2000; Hughes, 2003)
4) the needs of the Mita campus where students are required to read the major books and references for their study areas. In other words, the

required reading ability at the Mita campus.

5) the text books that are actually used in students' study areas.

The materials were searched and selected in the following way.

1) The grammar items were chosen by taking into consideration almost all of the grammar items that were supposed to have been mastered at the high school level.

To obtain this high school level information, there are textbooks authorized by the Ministry of Education that are available at bookstores. Since we did not pretest items in order to determine their difficulty empirically, we relied on theory to create items and sections at different ability levels. For example, the vocabulary items were based on word frequency counts using the benchmark of English-Japanese dictionaries available at bookstores, the grammar items were based on developmental sequences and on the written structures on textbook analysis. The textbooks authorized by the Ministry of Education, Sports and Science are available at bookstores.

2) The reading passages were selected from three disciplines (humanities, social sciences and natural sciences), and appropriate vocabulary levels were taken into consideration. The text passages were analyzed using L1 Fresch Reading Ease (Readability Formula) together with the judgments of experienced teachers.

c. Subjects (Test takers)

The subjects were the entering students (2006 to 2010). Table 3 show the information about the test takers and the corresponding test form.

Table 3 Information about the test takers and the corresponding test form

Test taker	Test form	Test Date	N
2006 entering students	PT1 CT1	2006 April 2007 February	853 790
2007 entering students	PT2 CT2	2007 April 2008 February	856 830
2008 entering students	PT3 CT3*	2008 April 2009 February	841 794
2009 entering students	PT4 CT4	2009 April 2010 February	830 768
2010 entering students	PT5	2010 April	816

*This test data was not included in the test equation design because no anchor items were provided.

According to Table 3, for example, the test taker group of 2006 took both PT1 and CT1. In this study, in order to examine the change of the students' reading ability between PT and CT, each group who took different tests is regarded as a different test taker group. And different test population was provided accordingly.

The 2006 PT1 test taker group was operationally defined as the norm group in this study to investigate the students' ability change across four consecutive years.

d. Analyses

Test Analysis

The test data was analyzed using the Winsteps statistical program, the Xcalibre statistical program and the Bilog MG calibration program. The fit-misfit information was investigated to determine if the test results fit the model or not in the Rasch measurement analysis. The information about item difficulty and item discrimination was obtained to check each item in terms of the classical test theory. The benchmark for the Cronbach alpha index of the test reliability was set at 0.75 or over. Also, the content validity is discussed by using the questionnaire analysis. The Bilog MG was used to confirm that each item was functioning properly to obtain item information as well as test information.

IV. Results and Discussion

1. Descriptive statistics of each test

Table 4 Descriptive Statistics of Each Test

Test Form	N	Mean/50	SD	Max	Mini	Grammar/15	Vocabulary/10	Gap-filling/10	Reading/15
PT1	853	32.42	6.94	49	6	10.82	5.24	7.58	8.78
CT1	790	31.39	6.42	48	4	9.53	6.08	5.86	9.93
PT2	856	29.89	6.43	48	9	9.98	6.03	5.81	8.06
CT2	830	27.81	6.29	45	1	9.72	4.85	4.53	8.71
PT3	841	31.28	5.97	49	8	10.47	5.56	5.62	9.62
CT3*	794	31.72	6.14	47	11	11.07	5.61	6.22	8.81
PT4	830	32.19	6.82	47	9	10.24	6.61	6.03	9.31
CT4	768	28.71	6.57	47	10	10.29	5.13	4.61	8.67
PT5	816	33.20	6.84	49	9	11.62	7.25	6.41	7.90

2. Validity examination

2.1 Misfit information

In order to check if the response fit the model, the fit-misfit information was investigated. The acceptable range of the Mean Square Value is usually from .7 to 1.3. The figure below .7 means overdiscriminating (overfitting), while the figure above 1.3 means underdiscriminating (underfitting). The number of items either below .7 or above 1.3 is in Table 5 as follows:

Table 5 Information about misfitting items in each test

Test Form	PT1	CT1	PT2	CT2	PT3	CT	PT4	CT4	PT5
Underfitting items	5	1	1	0	0	2	1	1	1
Overfitting items	0	0	0	0	0	2	3	0	6
Misfitting items (total)	5	1	1	0	0	4	4	1	7
Total items in each test	50	50	50	50	50	50	50	50	50
%	10	2	2	0	0	8	8	2	14

N.B.

PT1: 5 underfitting items (reading 3, grammar 1, vocabulary 1)

CT1: 1 underfitting (vocabulary)

PT2: 1 underfitting (grammar)

CT: 4 items (two underfiting items: reading 1, vocabulary 1; two overfitting items: reading 1, vocabulary 1)

PT4: 4 items (one underfitting item: gap-filling; three overfitting items: grammar 2, vocabulary 1)

CT4:1 underfitting item (vocabulary)

PT5: 7 items (one underfitting item: vocabulary; six overfitting items: grammar 5, vocabulary 1)

The discovery of seven problematic items (the biggest number in this data) in PT5 was not a problem from the viewpoint of the whole test. In other words, 86% of the test items fit the model, which technically verifies the construct validity of the test. The same is true with CT1,PT2,CT,PT4 and PT1 in which there were fewer problematic items. In other tests, there were no misfitting items, which verifies the construct validity of each test.

2.2 Content validity examination

In order to examine the content validity, 34 part-time instructors (who were

not involved in the test construction or editing) were asked if the PT (PT1) was appropriately measuring the students' reading ability. Table 6 shows the results. The questions were: 1: It is possible to place the students into their appropriate levels according to the test result. 2: I think we can conduct more effective reading-centered class based on the placement test results.

Table 6

Teacher

Question 1

The students have been placed into their appropriate levels according to their reading ability (including grammar and vocabulary knowledge) following the placement test.

A4 : Strongly Agree **A3 : Agree**
A2 : Disagree **A1 : Strongly Disagree**

Question 2
I can teach more effectively now that the students are classified on the basis of their placement test results.

A4 : Strongly Agree **A3 : Agree**
A2 : Disagree **A1 : Strongly Disagree**

Level 2 **N=23**

	Question 1				Question 2			
	A4	A3	A2	A1	A4	A3	A2	A1
TOTAL	0	23	0	0	3	16	4	0
					13%	70%	17%	0%

Level 3 **N=11**

	Question 1				Question 2			
	A4	A3	A2	A1	A4	A3	A2	A1
TOTAL	1	10	0	0	3	7	1	0
	9%	91%	0%	0%	27%	64%	9%	0%

A test is said to have content validity if the questions reflect the course content or syllabus. The result was that in question 1 (all said yes), in question 2 (29 out of 34 also gave affirmative answers. Therefore, it can be that the test had reliable content validity.

2. 3 Correlations among different test components (sub-sections)

Correlations of the sub-sections can be one way of assessing the construct validity of a test. Since the reason for having different test components is that they all measure something different and therefore contribute to the overall picture of language ability attempted by the test, we should expect these correlations to be fairly low-possibly in the order of .3–.5. The correlations between each subtest and the whole test might be expected to be higher-possibly around .7 or more – since the overall score is taken to be a more general measure of language ability than each individual component score. And it is common in internal correlation studies to correlate the test components with the test total minus the component in question (Alderson et.al 1995).

Table 7 Correlations of sub-sections

Table 7: Correlations among three latent ability variables (theta)			
N=6510	GRAMMAR	VOCABULARY	READING
GRAMMAR	1	0.493	0.433
VOCABULARY		1	0.441
READING			1

Table 7 shows that the correlations among each sub-section (grammar, vocabulary, reading) is between .3–.5 which is set up as a benchmark. This is calculated and presented in the theta data using logit scores of 6510 students. This table indicates that each subsection is sharing about 25% overlapping parts with each other, and that the rest 75 are individually unique to each other.

The same is true with each test (PT1-CT4) in Table 7–2, except one or two above or below the benchmark (either below .3 or above .5). This is calculated using raw scores of students in each individual test (four sections: grammar, vocabulary, cloze, reading.) This table further supports the fact each subsection in each placement test is sharing about 25% overlapping parts with each other, and that the rest 75 are individually unique to each other.

Moreover, this table indicates the correlations between each sub-section and the whole test. Although the results are not as high as anticipated (the expected correlation coefficients would be .7 or more), individual test component (sub-section) still contributes to the overall picture of language ability attempted by the test.

Table 7-2 Correlation matrices among three raw scores by PTs and CTs

(1) 2006-04 PLACEMENT TEST

N=853	GRAMMAR	VOCABULARY	CLOZE	READING	Whole
GRAMMAR	1	0.428	0.519	0.470	0.599
VOCABULARY		1	0.416	0.354	0.489
CLOZE			1	0.528	0.628
READING				1	0.569

(1) 2007-02 CONFIRMATION TEST

N=790	GRAMMAR	VOCABULARY	CLOZE	READING	Whole
GRAMMAR	1	0.486	0.370	0.454	0.566
VOCABULARY		1	0.382	0.466	0.582
CLOZE			1	0.333	0.447
READING				1	0.538

(2) 2007-04 PLACEMENT TEST

N=856	GRAMMAR	VOCABULARY	CLOZE	READING	Whole
GRAMMAR	1	0.441	0.455	0.407	0.555
VOCABULARY		1	0.415	0.390	0.528
CLOZE			1	0.402	0.542
READING				1	0.505

(2) 2008-02 CONFIRMATION TEST

N=830	GRAMMAR	VOCABULARY	CLOZE	READING	Whole
GRAMMAR	1	0.397	0.329	0.398	0.487
VOCABULARY		1	0.320	0.431	0.507
CLOZE			1	0.407	0.458
READING				1	0.544

(3) 2008-04 PLACEMENT TEST

N=780	GRAMMAR	VOCABULARY	CLOZE	READING	Whole
GRAMMAR	1	0.370	0.313	0.356	0.468
VOCABULARY		1	0.258	0.298	0.415
CLOZE			1	0.341	0.412
READING				1	0.450

(4) 2009-04 PLACEMENT TEST

N=830	GRAMMAR	VOCABULARY	CLOZE	READING	Whole
GRAMMAR	1	0.509	0.412	0.408	0.550
VOCABULARY		1	0.432	0.446	0.592
CLOZE			1	0.437	0.535
READING				1	0.536

(4) 2010-02 CONFIRMATION TEST

N=758	GRAMMAR	VOCABULARY	CLOZE	READING	Whole
GRAMMAR	1	0.414	0.342	0.399	0.488
VOCABULARY		1	0.436	0.371	0.520
CLOZE			1	0.463	0.536
READING				1	0.534

(5) 2010-04 PLACEMENT TEST

N=816	GRAMMAR	VOCABULARY	CLOZE	READING	Whole
GRAMMAR	1	0.595	0.470	0.411	0.604
VOCABULARY		1	0.488	0.417	0.627
CLOZE			1	0.449	0.580
READING				1	0.515

Table 7-3 Correlations between each sub-section and the whole test

		GRAMMAR	VOCABULARY	CLOZE	READING
(1) 2006-04	PLACEMENT	0.599	0.489	0.628	0.569
(1) 2007-02	CONFIRMATION	0.566	0.582	0.447	0.538
(2) 2007-04	PLACEMENT	0.555	0.528	0.542	0.505
(2) 2008-02	CONFIRMATION	0.487	0.507	0.458	0.544
(3) 2008-04	PLACEMENT	0.468	0.415	0.412	0.450
(4) 2009-04	PLACEMENT	0.550	0.592	0.535	0.536
(4) 2010-02	CONFIRMATION	0.488	0.520	0.536	0.534
(5) 2010-04	PLACEMENT	0.604	0.627	0.580	0.515

As mentioned above, the results are not as high as anticipated (the expected score was .7 or more) in Table 7-3; however, each subsection seems to make a case for their representative value to the test. Overall the four subsections positively contribute to the whole test.

3. Examination of reliability

The reliability was investigated in relation to the Cronbach Alpha index. The benchmark for the acceptable boundary is over 0.75. The reliability of placement tests had scores of over 0.75 in terms of the Conbach Alpha index. This suggests the items in each test were internally consistent. Table 8 shows the Cronbach Alpha index in each test form.

Table 8 The Cronbach Alpha index in each test form

Test form	PT1	CT1	PT2	CT2	PT3	PT4	CT4	PT5
Cronbach α	0.83	0.78	0.77	0.75	0.76	0.81	0.78	0.82

4. Possibilities of Item Bank using the Rasch model

The 50 items in each test are linked by 12 anchor items for the whole item bank. Once all the items are calibrated and the difficulty of each item is determined, each item can be put on the continuum of the scale according to their logit scores (difficulty level). These items along with a task can be stored as items in a bank.

Up to now we have been able to store 314 items linked by anchor items. Figure A shows the relative position of the students' abilities and the item difficulty on the same scale.

An examination of the whole item pool or item bank as a single test in terms

of fit/misfit aspect will now be offered. Table 9 shows that there are only 15 misfit items (7 underfit and 8 overfit items) out of 314, which is less than five percent of the whole data set. It can be said that this whole item bank as a test could work properly to provide appropriate items to subtests as parallel tests. In other words, there is a possibility now that we can construct several versions of parallel tests in which test difficulty should becomparable by examining the logit scores (difficulty level).

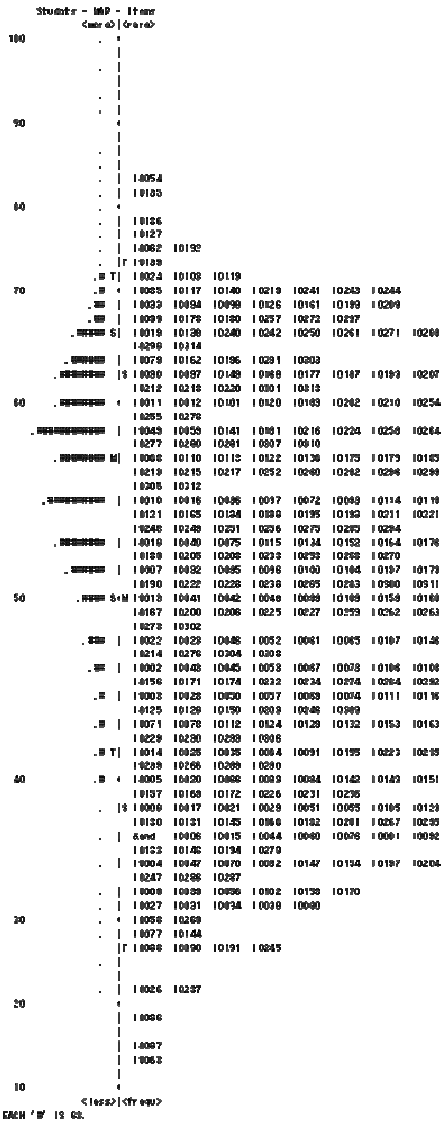

Figure A Relative positions between test takers' ability and item difficulty

Table 9 Information about mis-fitting items in the whole data set of 314items

Test Form	Whole data set
Underfitting items	7
Overfitting items	8
Misfitting items (total)	15
Total items in each test	314
%	4.7

	Underfitting items	Overfitting items
Grammar	1	6
Vocabulary	2	1
Gap-filling	1	1
Reading	3	0
Total	7	8

As Table 10 shows, it is possible to make four tests. One test consists of four sub-sections (grammar, vocabulary, gap-filling, reading). In one test, grammar has fifteen items, vocabulary (10 items), gap-filling (10 items (one passage x 10 items)), and reading (15 items (3 passages x 5 items each)). Four parallel tests can be used to examine students' learning practice or teaching effect within an academic year. The four parallel tests can also be used for longitudinal purposes by using raw scores because all the items were linked and equated by the anchor items.

Table 10 Information about misfitting items in each sub-section (i.e. grammar, vocabulary, reading)

Test Form	grammar	vocabulary	reading	Reading (passages)
Underfitting items	11	7	3	2
Overfitting items	11	2	0	0
Misfitting items (sub-total)	22	9	3	2
Total items of each section	90	59	85	17
Candidates (total-misfit)	68	50	82	15

N.B. The gap-filling section was not anchored in its specific section, but was thought to be linked in the whole 50-item test. Therefore, 90 items (10 items x 9 passages) can be used in the new parallel test to be part of the component of the test.

Possible tests (4)

Necessary items or passages
Grammar: 15 items x 4 (testlets)=60 items/68
Vocabuary: 10 items x4 (testlets)= 40 items/50
Reading: 3 passages x 4 (testlets)= 12 passages/15

5. Comparison of each subtest (consecutive four year analysis)

Note: When we compare the test results on a year basis, the benchmark basis is always the results of 2006 placement test results.

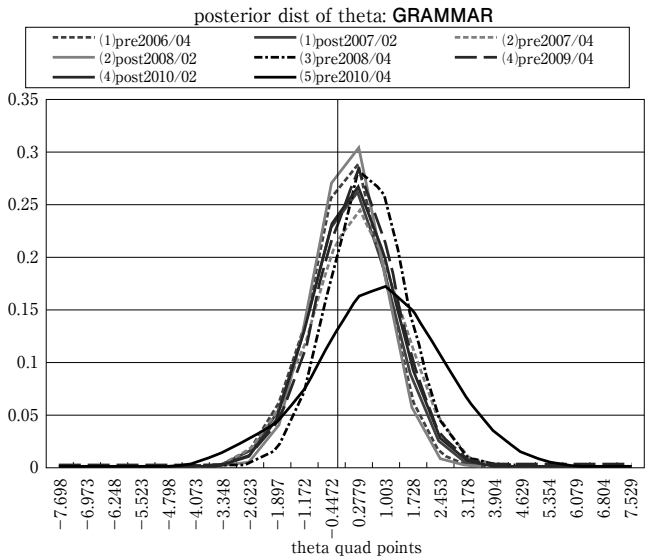

Figure 1 Grammar comparison

Figure 1 shows two things: 1. the difference between the placement test of grammar and the confirmation test of grammar in 2006 academic year, in 2007 academic year, and 2009 academic year; and 2. the change of the four-consecutive-year comparison (2006–2010). This comparison is based on the scores from the placement tests from each year. The dotted line shows the placement test and the solid line indicates the confirmation test.

Figure1 shows that there is little change in the students' grammar ability between the PT and the CT in each academic year. Unlike high school education, there is no English grammar course or class in university education and as a result it is not surprising that students' grammatical abilities would not improve significantly. It should be noted that freshman students' grammatical abilities, in its

separate unit peak, during university entrance examination season. The university education makes a contribution to sustain basic grammatical abilities in students' courses.

Although it is difficult to prove that the placement test results for the last four years show an increase of entering students' grammatical abilities, the pre 2010 test results indicate a greater standard deviation of the students' grammatical abilities. In other words, students' grammatical abilities increase to a greater extent than the previous years.

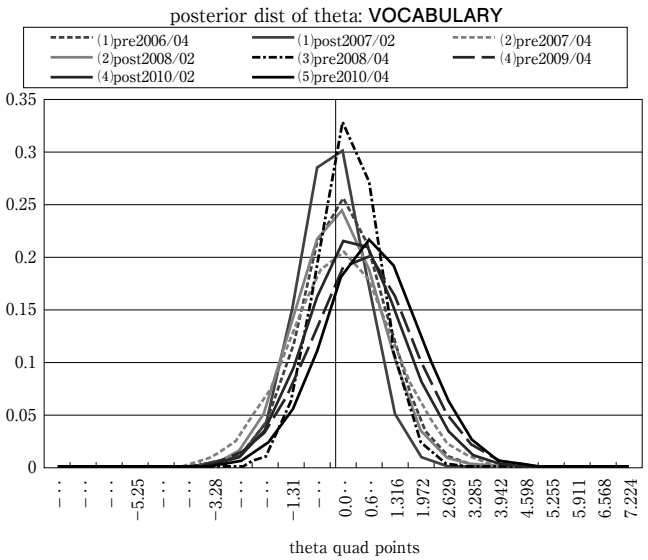

Figure 2 Vocabulary section

Figure 2 suggests that there was no noticeable impact in the vocabulary section when we compare the dotted and solid lines. Moreover, students' vocabulary abilities actually exhibited annual declines for each individual academic year. This lead the researchers to believe that because there were no specific vocabulary building courses in university, students' crammed vocabulary knowledge peaked for the purposes of taking the entrance examination, but after entering university the retention of the learned vocabulary gradually faded. Instead, student vocabulary that focused on textbook reading improved even if the width of the vocabulary did not show a significant increase.

The solid lines suggest that entering students' vocabulary knowledge has been improving for the last four years.

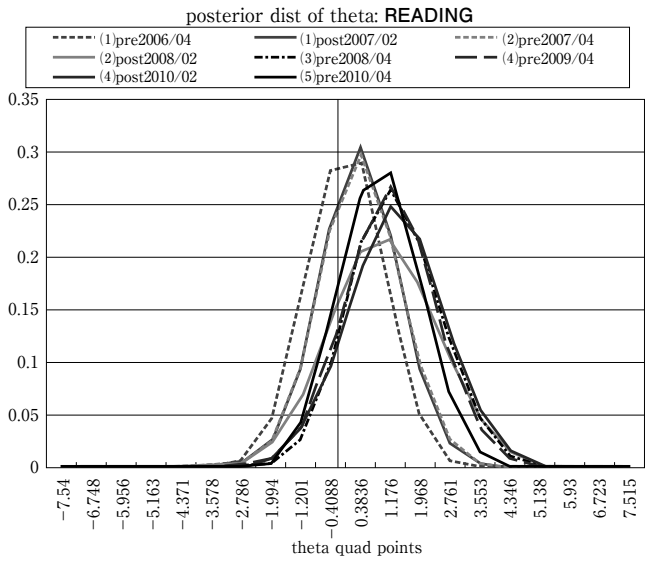

Figure 3 Reading section

Figure 3 shows that there has been a visible change not only in the teaching or learning effect of reading, but also in the entering students' reading ability. In 2006 there was an improvement of students' reading ability between their placement test (pink dotted line) and confirmation test (pink solid line) results. Also, in 2007 there was an increase between their placement test (yellow dotted line) and confirmation test (solid line). Furthermore, in 2009 there was another improvement in students' reading ability between their placement test (blue dotted line) and confirmation test (blue solid line). It can be said that the curriculum change which focused primarily on enhancing students' reading ability has been successful in this regard. One hypothesis is that freshman students, after taking general education courses, increased their background knowledge (schemata) of subject matters in each course. As a result of students' newfound knowledge the subject matter (in other words test topics) in the reading tests began to resemble those in the general education courses. In short, the increase of the background knowledge in the education courses can affect the students' improved reading ability.

Still another possible explanation is that unlike high school education, where test taking strategies are likely utilized, university English education stresses a more thorough understanding of the text, which helps students foster their analytical reading ability.

Figure 3 also indicates that entering freshman students' reading ability

has been improving. One possible explanation for this is that placement test information, which students have access to, can have a positive impact for those who wish to enter Keio University (Faculty of Letters).

6. Comparison of each subtest (pt and ct)

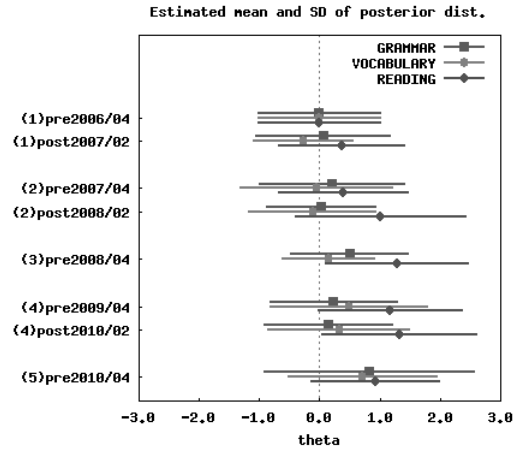

Figure 4 : Three sections (Grammar, Vocabulary and Reading) in comparison at a glance

Figure 4 demonstrates the characteristics of three sections (i.e. grammar, vocabulary, and reading) over four consecutive years. As previously suggested, the PT1 2006 is the benchmark for all the following yearly analyses. One interesting finding was that when grammar and vocabulary knowledge did not improve within the new curriculum, teaching did improve students' reading ability. There was a continual improvement between PT and CT. Furthermore, the improvements between PT2 and CT2 should also be noted. The current findings may also suggest that there has been a teaching effect between pre-teaching (at the placement test time) and the post-teaching (at the confirmation test time).

V. Conclusions and Implications

Considering McNamara's (2000, p.83) statement "The right balance of three basic critical dimensions of tests—validity, reliability and practicality—will depend on the test context and test purpose," the present placement test should be regarded as acceptable, judging from the statistical analyses and the test context as well as

the test purpose. For future improvement, predictive and concurrent validity should be measured. Further, multi-trait multi-method (MTMM) analysis could also offer valuable insights into the potential merits of test validation. Future studies should also explore the issue of face validity and practicality more systematically.

Future research should consider the possibility of a common person test equation as well as a common item test equation (by using another model such as 2PL (2-parameter logistic model) which would offer more suitable illustration item characteristics when analyzing linguistic tests.

One immediate future project will be that using 314 anchored items, at least 4 parallel test forms should be made, so that the change in students' reading ability and teaching effect can be compared on a raw-score basis.

Acknowledgement

I would like to thank all the members of the Keio Placement Test Research Group for their cooperative work editing, as well as administering the tests. The members are as follows: Andrew Armour, Yuji Nakamura (head author of this paper), William Snell, Yoshiko Uzawa, Kenji Adachi, Hikaru Sakamoto, Yoko Hemmi, Nobuya Takahashi, Kyoko Yoshida, Satoko Tokunaga, and Yuichi Akae. I am also grateful to Mr. Haruhiko Mitsunaga for his statistical help.

References

AERA, APA and NCME. (1999). *Standards for educational and psychological testing.* American Educational Research Association.

Alderson, J. C. (2000). *Assessing Reading.* Cambridge: Cambridge University Press.

Alderson,J.C., Clapham, C. and Wall,D.(1995). *Language test construction and evaluation.* Cambridge: Cambridge University Press.

Bachman, L.F. (1990). *Fundamental considerations in language testing.* Oxford: Oxford University Press.

Bachman, L.F. (2004). *Statistical Analyses for Language Assessment.* Cambridge: Cambridge University Press.

Bachman, L.F. and Palmer, A. (2010). *Language Assessment in Practice.* Oxford: Oxford University Press.

Brown, J. D. (2005). *Testing in Language Programs: A Comprehensive Guide to English Language Assessment.* New Edition. New York: McGraw-Hill.

Brown,J.D. and Hudson,T. (2002). *Criterion-referenced language testing.* Cambridge: Cambridge University Press.

Brown, H.D. (2010). *Language Assessment: Principles and Classroom Practices.* Second Edition. New York: Pearson Education.

Brown, H.D. (2007). *Teaching by Principles: An Interactive Approach to Language Pedagogy.* Third Edition. New York: Pearson Education.

Clapham,C.and Corson,D. (Eds.). (1997). Encyclopedia of language and education, vol.7, *Language testing and assessment.* Kluwer Academic Press.

Coombe, C., Folse, K. and Hubley, N. (2007). *A Practical Guide to Assessing English Language Learners.* Ann Arbor: The University of Michigan Press.

Davies, A., Brown, A., et al. (Eds.). (1999). *Dictionary of language testing.* Cambridge: Cambridge University Press.

Douglas, D. (2010). *Understanding Language Testing.* UK: Hodder Education.

Downing, S. M. and Haladyna, T. M. (Eds.). (2006). *Handbook of Test Development.* New Jersey: Lawrence Erlbaum Associates.

Fulcher, G. (1997). An English language placement test: issues in reliability and validity. *Language Testing* 14, 2, 113–138.

Fulcher, G. (2010). *Practical Language Testing.* UK: Hodder Education.

Genesee, F. and Upshur, J.A. (1996). *Classroom-based evaluation in second language education.* Cambridge: Cambridge University Press.

Grabe, W. (2000). *Reading research and its implications for reading assessment.* In A. Kunnan (Ed.), Fairness and validation in language assessment (pp.226–62). Cambridge: Cambridge University Press.

Heaton, J. B. (1997). *Writing English language tests.* New Edition. English Impression. New York: Longman.

Henning, G. (1987). *A guide to language testing.* New York: Newbury House Publishers.

Hughes, A. (2003). *Testing for language teachers.* Second Edition. Cambridge: Cambridge University Press.

Ikeda, H. (2007). (*Tesuto no kagaku*) (Science of Testing). Kyoiku sokutei kenkyujo.

Linacre, M. (2009). *WINSTEPS Rasch Measurement computer program* (Version 3.64). Chicago, Winsteps. com.

McNamara, T.F. (1996). *Measuring second language performance.* London: Longman.

McNamara, T.F. (2002). *Language testing.* Oxford: Oxford University Press.

Messick, S. (1996). "Validity and washback in language testing". *Language Testing* 13,3. Edward Arnold.

Nakamura, Y. (1998). Components of Reading Ability. *Educational Studies,* 40. pp.259–281. International Christian University.

Ozaki, S. (2008). (*Gengo tesutogaku nyuumon*) (Introduction to Language Testing). Daigaku kyoiku shuppan.

Stoynoff, S. and Chapelle, C.A. (2005). ESOL Tests and Testing: A Resource for Teachers and Administrators. Teachers of English to Speakers of Other Languages.

Westrick, P. (2005). Score Reliability and Placement Testing. *JALT Journal,* 27, 1, 71–92.