ASIAN EFL JOURNAL
The EFL Professional's Written Forum

Volume 7. Issue 2.
Article 6

Article Title

A Validation Study on the English language test in a Japanese Nationwide University Entrance Examination

Author

Akihiro Ito

Aichi Gakuin University, Japan

Bio Data

Akihiro Ito, PhD in Applied Linguistics (Hiroshima University, Japan) is currently an Associate Professor of applied linguistics at Aichi Gakuin University, Japan. His research interests include second language acquisition and language testing. In September 2001, he was given the JACET Award for the Most Promising New Scholar (Young Applied Linguist of the Year) for his article titled 'Japanese EFL Learners' Sensitivity to Configurational Distinction in English Realtivization' in ITL Review of Applied Linguistics (Afdeling Toegepaste Linguistiek at Katholieke Universiteit Leuven, Belgie), 131&132, 11-33, 2001.

Abstract

The present study employs validation study on the English language test of the Japanese nationwide university entrance examination- the Joint First Achievement Test (JFSAT). Two studies are presented. The first examines the reliability and concurrent validity of the JFSAT-English test. The reliability was acceptable. Criterion validity was estimated by correlating the JFSAT-English test and English language ability measure (a

carefully constructed cloze test) and was found to be satisfactory. The second study reports on a construct validation study on the test through internal correlation study. The JFSAT-English test was divided into five subtests. Examination of the correlation matrix indicated that the paper-pencil pronunciation test had low validity with almost no significant contribution to the total test score. It is argued that though the JFSAT-English test can work as a reliable and somewhat valid measure of English language ability, the paper-pencil pronunciation test should be eliminated and a listening comprehension test might be included as one of the subtests in the JFSAT-English test. The other subtests, however, showed satisfactory validity.

Introduction.

English language tests are widely used as one of the components of the tests for screening students in Japanese university entrance examinations. However, there has been very little research to indicate just how effective an English language test in university entrance examination settings is in terms of reliability and validity. The test data used in the entrance examinations are almost never disclosed, presumably for security reasons. This has served as the motive for the present research. In this paper the author would like to report on a study investigating reliability and validity of the most widely taken English language test in Japan. The test is the English test in the Joint First Stage Achievement Test' (henceforth, the JFSAT-English test).

Each Japanese university used to construct its own annual entrance examinations independently. There had been a wide variety of types of entrance examinations. As a consequence, the content and difficulty of items in the entrance examinations were never standardized. As a trial to promote the standardization of test content, the Mombusho (henceforth, Japanese Ministry of Education, Science and Culture) made a decision to administer a nation wide test to be called the Kyotsu Ichiji Gakuryoku Shiken (JFSAT) in 1971. The Japanese Ministry of Education, Science and Culture spent 8 years planning and developing the test with multiple-choice format and giving ample notification about the nature of the test to the entire educational system. In so doing, the Daigaku Nyushi Center (henceforth, the National Center for University Entrance Examinations), where the JFSAT is constructed, distributed and scored, was established. In 1979, the first

JFSAT was introduced to Japanese government-sponsored universities as one of the components of the entrance examination. The purpose of the JFSAT was to measure applicants' basic knowledge and ability in various subjects such as Japanese, mathematics, science, social studies and foreign language (about 99 per cent of the JFSAT examinees choose English though a test in German, French, Chinese and Korean can be chosen at present).

In 1989, the university examination system slightly changed and the JFSAT was revised to a different version. However, the content remained virtually the same. The JFSAT is "a first stage exam, somewhat analogous to the College Board SAT, in that many universities subscribe to it" (Brown and Yamashita, 1995, p. 12). As a result, the present Japanese university entrance examination system requires the examinee eager to enter the government-sponsored university to take two examinations: the JFSAT and second stage test at their prospective university. In most government-sponsored university entrance examinations, the total score of the two tests (the JFSAT and the second screening test) are used for screening the applicants, though each university 'has the power to determine the relative value of the JFSAT and the second stage test, so some universities focus on the JFSAT, others on the second stage test' (Ingulsrud, 1994, p. 67). In addition, since 1989, some private universities started to use the results of the JFSAT as one of their tools for screening the applicants, not only the applicants for government-sponsored university but also those for some private universities take the JFSAT. Since then, the number of JFSAT test-takers has been increasing every year. In January 2002, the JFSAT-English test was administered to 552,971 examinees.

Empirical research on the quality of the JFSAT-English tests is extremely limited. The first reason might be related to the difficulty of access to the test data for the people outside of the central location: The National Center for University Entrance Examinations. The National Center for University Entrance Examinations in Tokyo reports annually on the quality of components of the JFSAT through soliciting opinions from a variety of high-school teacher organizations. This source of information is the only one that we can access. Though we may see descriptive statistics of the JFSAT-English test such as mean scores and standard deviations, we cannot see if the test worked well in terms of reliability and validity. The second reason might be related to the

difficulty of defining the nature of the JFSAT-English test. The JFSAT is supposedly an achievement test based on the Japanese high school curriculum centrally determined by the Mombukagakusho (Japanese Ministry of Education, Culture, Sports, Science and Technology, formerly the Japanese Ministry of Education, Science and Culture). An achievement test is a test given to language learners at the end of a program to check if the learners have mastered the target grammatical items or skill. However, some people may doubt whether the JFSAT is really an achievement test. The reason might be related to the fact that high school teachers are not appointed to the test development committee and the JFSAT is basically considered to be a university entrance examination developed by Japanese university professors (see Ingrulsrud, 1994 for further discussion on this matter). In fact, the JFSAT is not a test based on specific English language programs or classes in Japanese high school education. Based on this, the author believes that the JFSAT could be called a proficiency test. However, the author has no intention to completely deny the concept of the JFSAT as an achievement test. This is because, as Brindley (1989) argues, the distinction between proficiency and achievement is not as clear-cut as it would first seem. In this paper, therefore, the author would like to use the term 'ability' as the wider concept including 'achievement' and 'proficiency' when discussing what the JFSAT may be measuring.

Because of the difficulty of access to JFSAT test data and the difficulty of defining the test type of the JFSAT, there is not much research directly investigating the question: Is the JFSAT-English test really measuring the applicant's English ability? The author could only find two studies dealing with the issues of reliability and validity of the JFSAT-English test. Kiyomura (1989) argues that the English questions in the JFSAT are relatively valid for predicting students' performance levels in English in high school. He reported a moderate correlation coefficient (r=.330 to .620) between the examinees' scores in the JFSAT-English test and their high school performance as measured by their transcripts. Yanai, Maekawa and Ikeda (1989), who were members of the test development section in the National Center for University Entrance Examination in Tokyo, report constantly high reliability coefficients of English tests (r=.940 to .956) in the JFSAT.

There is not much specific criticism on English tests in the JFSAT from the

public. Only one serious problem with the construction of JFSAT-English test, however, has been discussed in Japan for about the last 25 years. This concerns the use of paper-pencil pronunciation test for measuring the students' oral English ability (Ishii, 1981; Kira, 1981; Kuniyoshi, 1981; Masukawa, 1981; H. Suzuki, 1981; Ibe, 1983; Kashima, Tanaka, Tanabe and Nakamura, 1983; Ohtomo, 1983; Shiozawa, 1983; S. Suzuki, 1985; Ikeura, 1990; Takanashi, 1990; Wakabayashi and Negishi, 1994). Buck (1989) explains that written tests of pronunciation used in Japanese university entrance examinations are based on the simple idea, which was suggested by the American structural linguist and language testing expert Robert Lado, that the language learners choose appropriate places in stress or pronunciation in written tests if they can speak with appropriate stress and pronunciation in actual language use. Written test of stress and pronunciation are easy to administer to large numbers of students without the use of an actual interview or recorder and may even be more economical in that raters do not have to have good English pronunciation. A several Japanese researchers have recently conducted research on the validity of paper-pencil pronunciation tests. Shirahata (1991) sampled 40 high school students and administered paper and oral tests of primary stress. Each of the tests contained 40 items. The test items were taken from English tests (1990 and 1991 version) in the JFSAT. Results indicated that learners' actual performance in primary stress can be predicted by paper and pencil pronunciation tests because there was a relatively high agreement rate (82.6%) between the learners' actual performance in primary stress and their scores in the written test of pronunciation.

The target of the item type of written pronunciation test shifted in another study by Inoi (1995). His experiment was administered to 60 college freshman students. The participants were given a written test first, then an oral version followed. Each test contained 30 items which were sampled from recent English tests in the JFSAT. Results indicated that a relatively strong correlation was obtained between the scores in the two tests ($r=.810$, $p<.01$). The agreement rate, however, was unexpectedly not so high (66.8%). Inoi (1995) concludes that some phoneme discrimination items on his test were not valid for assessing the participants' actual pronunciation ability of English words. As a consequence, he added that good performance on a written test does not guarantee good performance on an oral test. These two studies have shown a moderate validity of the

paper-pencil pronunciation test for assessing students' actual performance in pronunciation such as primary stress and/or phoneme discrimination (Komazawa and Ito, 1997).

Wakabayashi and Negishi (1994) argue that the simplest way of measuring the student's aural/oral ability in English is to develop a listening comprehension test and administer the test as one of the sections (subtests) in the JFSAT-English test. However, the paper-pencil pronunciation test has been used to indirectly measure student's oral ability, not aural ability (listening comprehension ability). Therefore, if we would like to argue that a listening comprehension test should be administered in the JFSAT-English test, we should first prove that there is no correlation between the pronunciation ability measured in the paper-pencil pronunciation test and the ability measured in the listening comprehension test.

The possible need of introducing a listening comprehension test in the JFSAT-English test is also discussed in relation to the content of the Course of Study by the Japanese Ministry of Education, Science, and Culture. Though Japanese Ministry of Education, Science and Culture (1994) implemented aural/oral guidelines for the Course of Study for Upper Secondary School (high school) and has tried to enhance high school students' listening comprehension skills, the JFSAT-English tests have not changed to reflect the new direction toward emphasizing aural skills. In short, as Brown and Yamashita (1995, p. 28) say, "there is a contradiction between what is tested in the JFSAT-English test and what the Ministry of Education, Science and Culture promotes in its curriculum." The Japanese Ministry of Education, Science and Culture (1999) has issued a new version of the Course of Study for Upper Secondary School (high school) and in its new curriculum has again emphasized the importance of developing high school students' listening comprehension skills as was the case in the old version.

According to the Final Report on the Contents of Future JFSAT of 2008 (National Center for University Entrance Examinations, 2003), which is displayed on the web site (http//www.dnc.ac.jp/center_exam/18kyouka-saishuu.html), the National Center for University Entrance Examinations reports with the date of 8 June, 2003, that the center has made a final decision to administer a listening comprehension test as one of the subtests in the JFSAT-English test to be administered from 2008 on the basis of

examination of 'the Council on the Improvement of Method in Screening University Applicants' in the Ministry of Education, Culture, Sports, Science and Technology (formerly Ministry of Education, Science and Culture).

In summary, an ongoing debate has been discussed regarding the introduction of listening components to the JFSAT-English test for about 25 years. A listening test will indeed be administered from 2008 in the JFSAT-English test. However, as mentioned, there is not much research proving whether there is a lack of correlation between the paper-pencil pronunciation test and the listening comprehension test. In addition, the reliability and validity has not yet been much investigated.

Two studies are reported here. The first was designed to examine the reliability and criterion validity of the JFSAT-English test. Specifically, the criterion validity of the JFSAT-English test and correlation between the paper-pencil pronunciation test and a listening comprehension test will be examined. The second study looks at the construct validity of the JFSAT-English test, employing internal correlation study.

The Study

Purpose

As the purpose of the present study was to investigate the reliability and validity of English questions in the JFSAT, the goals were to determine if the JFSAT-English test is a reliable and valid measure of student's English ability. The present study investigates the criterion validity and construct validity of the JFSAT-English test.

The following research questions are set up.

Study 1

(1) How high are the reliability and criterion validity of the JFSAT-English test? Does the JFSAT-English test measure student's English ability?

(2) Is there a lack of correlation between the paper-pencil pronunciation test and a listening comprehension test?

Study 2

(3) How high is the construct validity of the JFSAT-English test?

The first research question will be addressed by measuring the reliability coefficient of the English test in JFSAT such as Cronbach's alpha and Pearson's product-moment correlation coefficients between the JFSAT-English test and an external criterion test such as a cloze test and between the paper-pencil pronunciation test and a listening comprehension test such as the listening section of the Test of English as a Foreign Language (TOEFL).

The second research question will be addressed by measuring the correlation coefficients among the subtests in the JFSAT-English test and examining the correlation coefficients on the basis of the criteria used in past internal correlation studies (e.g. Alderson, Clapham and Wall, 1995; Ito, 2000b; Imai and Ito, 2001). In order to carry out the internal correlation study, the JFSAT-English test should be divided into subtests supposedly measuring the same ability. According to Hitano's (1983, 1986) theoretical classification of the internal structure of JFSAT-English test, the test can be basically divided into five subtests. The first subtest is constructed to measure the knowledge and ability of pronunciation and stress; the second the knowledge and ability of English grammar; the third the knowledge and ability of spoken English; the forth the knowledge and ability of written English; and the fifth the knowledge and ability of reading comprehension. On the basis of the Hitano's classification, the JFSAT-English test will be categorized into five sections (subtests) and they are named as follows for the conveniece: (1) Pronunciation, (2) Grammar, (3) Spoken English, (4) Written English and (5) Reading Comprehension. Though it cannot be denied the internal structure of the JFSAT-English test may change each year, the author believes that the subtests of the JFSAT-English test to be engaged in the present study generally follow Hitano's classification.

Participants

The participants (N=100) were sampled from first-year students who were enrolled in an undergraduate class in general English at Aichi University of Education in Japan. Most of them were eighteen years old. The average age was 18 years and 2 months. The ratio between male and female was 1 to 1. All of them would have taken more than six years of formal English courses prior to this study. They were majoring in a scientific field. The

sample was thus homogeneous with regard to nationality, language background, educational level and age.

Instruments

The following instruments were used in the present investigation.

1. An open-ended cloze test (70 items) (Appendix 1). The full score was 70 points. The participants were allowed 30 minutes for completion.
2. The TOEFL Listening Comprehension Test (Steinberg, 1987) (50 items). The full score was 50 points. It took 25 minutes to finish the listening test.
3. The JFSAT-English test 1991 version (Kyogakusha, 1994) (58 items). The full score was 200. The description of test content and the item point value in each subtest are as follows:

Subtest 1: Pronunciation (7 items (5 items X 2 points and 2 items X 3 points: Total: 16 points)

In this section, the participants are asked to match one word with another having the same segmental phoneme out of four choices.

Subtest 2: Grammar (14 items X 2 points: 34 points)

In this section, the participants are required to select from the four choices the most appropriate word to be filled in the presented sentence out of four choices.

Subtest 3: Spoken English (6 items X 3 points: 18 points)

In this section, the participants are required to select the most appropriate word from four choices to be filled into the presented sentence.

Subtest 4: Written English (10 items X 4 points: 40 points)

In this section, the participants are asked to place given words into the correct order to produce a meaningful sentence.

Subtest 5: Reading Comprehension (18 items (16 items X 5 points: 80 points and 2 items X 6 points: 12 points. Total: 92 points)

In this section, the participants are asked to read passages and answer multiple choice questions. The participants are required to select the most appropriate answer from four choices.

The full score is 200 points.
The participants were allowed 80 minutes for completion.

4. The paper-pencil pronunciation test 1989 and 1992 versions (Kyogakusha, 1992, 1994) (20 items X 2 points: 40 points). The paper-pencil pronunciation tests in the JFSAT-English test are basically divided into three types. In the present study, the paper-pencil pronunciation test where participants are required to distinguish the segmental phonemes was used. The participants were given 10 minutes for completion.

The Construction of the External Criterion Test: The Cloze Test

In order to examine the JFSAT-English test in question, a carefully constructed cloze test was used as an external criterion test. The cloze test was produced on the basis of recent research on cloze test construction: (1) the selection of appropriate cloze texts for Japanese learners of English (Nishida, 1986, 1987; Mochizuki, 1984, 1994; Takanashi, 1984, 1988; Ito, 2000a); (2) appropriate word-level of the cloze text for Japanese learners of English (Mochizuki, 1992); and (3) the essential number of questions and scoring methods in cloze testing (Alderson, 1979; Nishida, 1985; Siarone and Schoorl, 1989). The cloze passage was adapted from a low intermediate reader (700 word level) for Japanese high school students by Ishiguro and Tucker (1989). The passage selected, "Wang's story," a relatively neutral, narrative topic, contained 457 words. Its readability level was about 8th grade level as measured by the Flesch-Kincaid readability formula by using computer program Grammatik IV (1988). The cloze test was created by deleting every 6th word for a total of 70 blanks. Two sentences were left intact: one was beginning of the passage and one at the end to provide a certain level of context. Siarone and Schoorl (1989) argue that a cloze test of about 75 items should be scored with a

contextually acceptable method in order to maintain a satisfactory reliability (r>.8). Then the cloze test was scored by the author based on the contextually acceptable words method with the help of a native speaker of English who was working as a full time English language professor at Aichi University of Education where the present research was conducted by the author.

In order to examine the concurrent validity of the cloze test itself, the correlation between the TOEFL (Steinberg, 1987) and the cloze test was measured in a pilot study. Klein-Braley and Raatz (1984) propose the criteria to judge the quality of the C-Test (a modified version of cloze test). In their six C-Test construction axioms, they argue that a valid C-Test should correlate with a reliable discrete-point test at .5 or higher. Since the C-Test is a modified version of a cloze test, the author applied Klein-Braley and Raatz's idea for judging the concurrent validity of the cloze test in the present study. The author also decided to apply the cloze test as a criterion test measuring the validity of the JFSAT-English test if the cloze test showed a higher correlation coefficient than .5 with the a reliable discrete-point test (TOEFL).

In the pilot study, the participants were 100 second year students enrolled in general English classes at Aichi University of Education in Japan. The average age was 19 years and 4 months. The ratio between male and female was 1 to 1. All of them would have taken more than seven years of formal English courses prior to this study. They were majoring in Japanese language education and art education. The sample was thus homogeneous with regard to nationality, language background, educational level and age. This group of participants was selected as the second population similar to the one to be used in the main study. This is because the two groups were placed in the same course level and were taught by the same teacher with the same text based on the same syllabus. Unfortunately, however, there was no data available showing the no statistically significant difference between the two groups in English ability level.

Table 1

Table 1 Descriptive statistics (N=100)

Test	Reliability (α)	Mean (M)	Full Score	SD
Cloze Test	0.840	33.450	70.000	7.815
TOEFL	0.781	31.720	100.000	8.128

The results indicated high reliability of the tests and moderate correlation coefficients between the TOEFL and the cloze test (r=.489, p<.01). The correlation coefficient corrected for attenuation by Henning's (1987) method is r=.604. In this regard, the cloze test had a relatively high reliability coefficient (r=.840) and moderate correlation (r=.604) with a reliable discrete-point test such as TOEFL.

The final decision on the cloze test

The author made a final decision to employ the cloze test as an external criterion test for measuring participants' English in the present investigation. This is because the reliability of the cloze test exceeds the critical threshold level of .8 (r=.840) and the test correlated with the reliable discrete-point test, TOEFL, at higher than .5 (r=.604, p<.01). The correlation corrected for attenuation in the cloze test was more than .5.

The TOEFL listening comprehension test

The TOEFL listening comprehension test was used as a listening comprehension test for examining if there was a lack of correlation with the paper-pencil pronunciation test. The author did not conduct a study on the validity of the TOEFL listening comprehension test. The reason for this is that the test has been utilized in real testing session, and we can therefore conclude that the test might be reliable and valid enough for the purpose of the present investigation.

Scoring procedure for the other tests

After all the test except the cloze test were administered, the test papers were exchanged between students and scored under the author's direction. After the test papers were collected, they were reviewed by the author twice before the statistical calculations were performed.

Statistical software and alpha level for significance

All statistical analyses were performed by computer programs in Statistical Package for Social Sciences Windows 7.5 version (SPSS Inc., 1996) in the library room of the Department of English Language Education, Faculty of Education, Hiroshima University, Japan.

Alpha level

The sample (N=100) necessitated a conservative treatment of statistical analysis because such a small amount of data does not show the normal distributions which are necessary for parametric statistical analyses. Therefore, the alpha level for all statistical decisions was set at α=.01.

Results

Study 1

Descriptive statistics

Descriptive statistics for each of four tests are given in Table 2. The low mean in TOEFL listening comprehension test suggests that the test was rather difficult for the population who took them, with the result that there is not as much variance in the tests as would be ideal. The reliability coefficients of the cloze test (r=.853), slightly higher than in the pilot study, and the JFSAT-English test (r=.817) are high. The reliability coefficients of the TOEFL listening comprehension test (r=.398) and the pronunciation test (r=.208) are low.

Table 2 Descriptive statistics (N=100)

Test	Reliability (α)	Mean (M)	Full Score	SD
Cloze Test	.853	32.850	70.000	7.697
JFSAT	.817	119.031	200.000	8.013
TOEFL Listening	.398	13.060	50.000	3.484
Pronunciation Test	.208	16.700	40.000	2.355

Concurrent validation

Table 3 displays the correlation coefficients between the cloze test and the JFSAT-English test. The correlation coefficient was moderate (r=.462, p<.01).

Table 3 Correlation between Cloze Test and JFSAT (N=100)

Test	r	p
Cloze Test and JFSAT	.462	<.01

Table 4 shows that there was no significant correlation between the TOEFL listening comprehension test and the paper-pencil pronunciation test (r=-.078, n.s.).

Table 4 TOEFL Listening Comprehension Test and Paper-Pencil Pronunciation Test (N=100)

Test	r	p
TOEFL Listening and Pronunciation	-.078	n.s.

Study 2

Descriptive statistics

Table 5 shows the descriptive statistics of the five subtests in the JFSAT-English test.

Table 5 Descriptive statistics of the Five Subtests (N=100)

Test	Mean (M)	Full Score	SD
Pronunciation	8.897	16.000	1.275
Grammar	19.860	34.000	2.731
Spoken English	11.568	18.000	1.125
Written English	18.304	40.000	2.186
Reading Comprehension	60.402	92.000	12.110
Total	119.031	200.00	8.013

Internal correlation study

Table 6 shows the full correlation matrix for the five subtests in the JFSAT-English test. Such a matrix is usually used to examine the construct validity of the test by internal correlation study.

Table 6 Correlation Matrix for the Five Subtests (N=100)

Test	1	2	3	4	5	Total(s)*
1. Pronunciation	---	---	---	---	---	.238 C3
2. Grammar	.280 C1	---	---	---	---	.564** C3
3. Spoken English	.136 C1	.437** C1	---	---	---	.493** C3
4. Written English	.159 C1	.522** C1	.395** C1	---	---	.543** C3
5. Reading Comprehension	.153 C1	.351** C1	.358** C1	.364** C1	---	.432** C3
6. Total	.392** C2	.782** C2	.600** C2	.729** C2	.746** C2	---

Notes: *Total score excluding the subtest which it is being correlated correlations **p<.01

Alderson, Clapham and Wall (1995) employed internal correlation study in order to examine the construct validity of the college English placement test. The placement test consisted of four different 'components,' which are equivalent to the 'subtest' in the present study. They go on to explaine the reasoning and the criteria each subtest needs to meet as follows.

> "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 +.0 - +.5. If two components correlate very highly with each other, say +.9, we might wonder whether the two subtests are indeed testing different traits or skills, or whether they are testing essentially same thing. If the latter is the case, we might choose to drop one of the two. The correlations between each subtest and the whole test, on the other hand, might be expected, at least according to classical test theory, 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. Obviously if the individual component score is partly between the test component and itself, which will artificially inflate the correlation. For this reason it is common in internal correlation studies to correlate the test components with the test total minus the component in question." (Alderson, Clapham and Wall, 1995, p. 184).

Reasoning for examining construct validity

This same reasoning used in Alderson, Clapham and Wall (1995) can be used here to examine the construct validity of the JFSAT-English test in the present investigation. In order to examine this, it is convenient to identify different types of correlation coefficients:

(1) The correlation coefficients between each pair of subtests. These correlation coefficients are labeled C1;
(2) The correlation coefficients between each subtest and the whole test. These

correlation coefficients are labeled C2; and

(3) The correlation coefficients between each subtest and the whole test minus the subtest. These correlation coefficients are labeled C3.

Criteria

In order to establish that the subtest under consideration works as a valid subtest and contributes to the total test score, the following criteria would need to be met:

Criterion 1: The correlation coefficients between each pair of subtests should be from .3 to .5. That is $.3 < C1 < .5$.

Criterion 2: The correlation coefficients between each subtest and the whole test should be around .7 or more. That is $C2 \fallingdotseq .7$.

Criterion 3: The correlation coefficients between each subtest and the whole test should be higher than those between each subtest and the whole test minus the subtest. That is $C2 > C3$.

On the basis of the three criteria above, the construct validity of each subtest will be examined.

First, looking at the paper-pencil pronunciation test (Pronunciation) :

Criterion 1: The C1 correlations (.280, .136, .159, .153) are all lower than .3. Therefore, the Criterion 1 was not met.

Criterion 2: The C2 correlation (.392) is rather lower than .7. Therefore, the criterion 2 was not met.

Criterion 3: The C2 correlation (.392) is higher than C3 correlation (.238) indicating that the criterion 3 was met.

Second, looking at the grammar and usage test (Grammar):

Criterion 1: The C1 correlations (.136, .437, .522, 351) are all higher than .3 except the correlation coefficient with the paper-pencil pronunciation test (.136). Therefore, the Criterion 1 was not met.

Criterion 2: The C2 correlation (.782) is higher than .7. Therefore, the criterion 2 was met.

Criterion 3: The C2 correlation (.782) is higher than C3 correlation (.564) indicating that the criterion 3 was met.

Third, looking at the spoken English expression test (Spoken English):

Criterion 1: The C1 correlations (.136, .437, .395, .358) are all higher than .3 except the correlation coefficient with the paper-pencil pronunciation test (.136). Therefore, the Criterion 1 was not met.

Criterion 2: The C2 correlation (.600) is lower than .7. Therefore, the criterion 2 was not met.

Criterion 3: The C2 correlation (.600) is higher than C3 correlation (.493) indicating that the criterion 3 was met.

Fourth, looking at the sequencing test (Written English):

Criterion 1: The C1 correlations (.159, .522, .395, 358) are all higher than .3 except the correlation coefficient with the paper-pencil pronunciation test (.159). Therefore, the Criterion 1 was not met.

Criterion 2: The C2 correlation (.729) is higher than .7. Therefore, the criterion 2 was met.

Criterion 3: The C2 correlation (.729) is higher than C3 correlation (.543) indicating that the criterion 3 was met.

Finally, looking at the reading comprehension test (Reading Comprehension):

Criterion 1: The C1 correlations (.159, .351, .358, .364) are all higher than .3 except the correlation coefficient with the paper-pencil pronunciation test (.159). Therefore, the Criterion 1 was not met.

Criterion 2: The C2 correlation (.746) is higher than .7. Therefore, the criterion 2 was met.

Criterion 3: The C2 correlation (.746) is higher than C3 correlation (.432) indicating that the criterion 3 was met.

Discussion

In this section, the original research questions are addressed. Before we discuss

the results, however, other aspects must be considered:

(1) the effects of sample homogeneity on statistical results concerning reliability coefficients and correlation coefficients; and

(2) the size of the sample (N=100).

(1) How high are the reliability and criterion validity of the JFSAT-English test? Does the JFSAT-English test measure student's English ability?

The reliability coefficient of the cloze test was .853. This value was slightly higher than that in the pre-test (r=.840). The result indicates that the cloze test has shown a consistent relatively high reliability. The reliability coefficient of the JFSAT-English test was also relatively high (r=.817). However, the reliability coefficient of the paper-pencil pronunciation test (r=.208) and TOEFL listening comprehension test (r=.398) were very low. Since the cloze test and the JFSAT-English test have shown relatively high reliability coefficients, we correlated the two tests without correcting for attenuation in the cloze test. The correlation coefficient between the tests was moderate (r=.462, r-square=.213). This gives an idea that shared variance between the tests of 21.3 % and the JFSAT-English test was a somewhat valid test of English ability.

(2) Is there no significant correlation between the paper-pencil pronunciation test and a listening comprehension test?

The lack of the correlation between the paper-pencil pronunciation test and the TOEFL listening comprehension test may be due to the fact that the ability to distinguish among segmental phonemes in the paper-pencil pronunciation test is completely different from the wide range of listening abilities which the TOEFL listening comprehension test tries to examine. In this sense, the answer to the research question (2) seems to be 'Yes.' However, there is a serious drawback with this study. The problem is the low reliability of the TOEFL listening comprehension test, a finding which calls for further research on the matter. Therefore, whether or not there is no significant correlation between the paper-pencil pronunciation test and the TOEFL listening comprehension test remains

unclear.

(3) How high is the construct validity of the JFSAT-English test?

The present internal correlation study suggests that the correlation coefficients between the paper-pencil pronunciation test and other subtests (r=.153, n.s. to .280, n.s.) are too low to meet the three criteria. In addition, the correlation coefficient between the paper-pencil pronunciation test and the whole test minus itself was very low (r=.238, n.s.). These results show that the paper-pencil pronunciation test shows low construct validity. Though there was a lower correlation coefficient between the spoken English expression test and the whole test (r=.600) than criterion 2 (r>.7), the other correlations in the matrix met the criteria. Therefore, it is justifiable to say that only the paper-pronunciation test does not significantly contribute to the total test score, which means indeed that the test has low construct validity. Though the research question (2) whether there is a lack of significant correlation between the paper-pencil pronunciation test and a listening comprehension test has not been well answered because of the low reliability of both of the paper-pencil pronunciation test and the TOEFL listening comprehension test, internal correlation study clarified the low validity of the paper-pencil pronunciation test.

In order to estimate how little the paper-pencil pronunciation test contributes to the total test score of English test in the JFSAT, the author calculated the correlation coefficient between the whole test and the whole test minus the paper-pencil pronunciation test. The correlation coefficient was very high (r=.998, p<.01, r-square=.996), which means that without the paper-pencil pronunciation test, the other 4 subtests (Grammar, Spoken English, Written English and Reading Comprehension) can account 99.6% of the variance of the whole test.

Conclusion

Summary

The present study investigated the reliability and validity of the English test in a Japanese Nationwide Entrance Examination (the JFSAT-English test). Participants

consisted of 100 Japanese university students learning English as a foreign language. The results of the first study revealed that the JFSAT-English test is, to some degree, an appropriate measure of the examinee's English ability in terms of reliability and validity. Second, there was no statistically significant correlation between the TOEFL Listening Comprehension Test and the paper-pencil pronunciation test. However, since both of the tests showed low reliability, this result should be considered with a certain amount of caution. The second study was an internal correlation study investigating the construct validity of the test. The results of the second study revealed that the correlations between the paper-pencil pronunciation test and the other subtests are very low, showing the low construct validity of the test.

Limitations and remaining issues

A few characteristics of the present study limit the generalizability of its results. First, the results can only be generalizable to Japanese learners of English who have studied English in a formal educational setting. However, selecting only one nationality is often one of the great strengths of this kind of empirical study. In many studies conducted by other researchers in the past, various language backgrounds, age and educational background were mixed together. As a result, the findings of those studies are often hard to interpret because they can be only generalized to the single situation in which the data happened to be collected. Second, the present study focused on the internal construct validity of the JFSAT-English test by calculating the correlation coefficients among the subtests. Therefore, though all the tests outside of the paper-pencil pronunciation test show somewhat high validity in the internal correlation study, the results do not guarantee that each subtest is really measuring what it is constructed to measure. For example, some people might question if the spoken English expression test can indirectly measure the examinee's English speaking ability. Therefore, future research should conduct concurrent validation study on each of the subtests in order to know more about the problems with the quality of the questions in the JFSAT-English test. In the future, we should investigate again whether there is a lack of correlation between the paper-pencil pronunciation test and a listening comprehension test of any type. If we find that there is no correlation between the two tests, then we should investigate what

kind of listening comprehension test should be included as a subtest in the JFSAT-English test. Perhaps one of the issues that we should consider might be the number of items in a listening comprehension test in order to make the test work as a subtest significantly contributing to the total test score.

The following two general questions are posed as a summary of this section in the hope that other researchers interested in this area will find these lines interesting enough to pursue in the future:

(1) Is there no significant correlation between the paper-pencil pronunciation test and a listening comprehension test?
(2) How high is the concurrent validity of each subtest of English test in the JFSAT?
(3) How many items should be included in the future listening comprehension test in the JFSAT-English test?

Acknowledgements

This work was made possible by the author's study leave from July to August 2002 in the Department of Linguistics and Modern English Language at Lancaster University, England. I would like to express my gratitude to Professor Charles Alderson, Dr. Caroline Clapham, Dr. Dianne Wall, Dr. Rita Green, Jayanti Banerjee for their constructive comments on my research project in the research sessions in Language Testing at Lancaster 2002, and J.D. Brown, University of Hawaii at Manoa for his special supervising and frequent personal discussion on this topic. A preliminary version of this article was presented at the 21st Language Testing Research Colloquium (International Language Testing Association) held at Tsukuba International Congress Center, Tsukuba, Japan on July 30, 1999.

References

Alderson, C. (1979). The cloze procedure and proficiency in English as a foreign language. *TESOL Quarterly, 13* (2), 219-227.

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

Brindley, G. (1989). *Assessing achievement in the learner-centered curriculum*. Sydney: National Center for English Language Teaching and Research.

Brown, J.D. and Yamashita, S.O. (1995). English language entrance examination at Japanese universities: What do we know about them? *JALT Journal, 17* (1), 7-30.

Buck. G. 1989: Written tests of pronunciation: Do they work? *ELT Journal, 43* (2), 50-56.

Daigaku Nyushi Center (National Center for University Entrance Examinations). (2003, June 8). *Heisei 18 nendo karano daigaku nyushi center shaken no shutsudai kyouka/ kamokutou nitsuite –saishuu matome- (The final report on the content of future JFSAT, Heisei 18 nendo version*). Daigaku Nyushi Center News, [on line] Retrieved June 14, 2003, from http://www.dnc.ac.jp/center_exam/18kyouka-saishuu.html

Henning, G. (1987). *A guide to language testing: Development, evaluation, research*. Boston, MA: Heinle and Heinle Publishers.

Hitano, T. (1983). *Koko chosasho, kyoutsu ichiji, niji shaken, nyuugakugo no seiseki no soukan (The correlation study between the university applicants' scores in transcript of records, the JFSAT, the second screening and their achievement levels after entrance examinations).* Daigaku Nyushi Forum (Forum on the university entrance examinations), 1, 57-62.

Hitano, T. (1986). Showa 60 nendo kyoutsu ichiji gakuryojushiken mondai no naiyou bunseki (The content analyses of the JFSAT showa 61 version). Daigaku Nyushi Forum (Forum on the university entrance examinations), 7, 105-120.

Ibe, T. (1983). Kyoutsu ichiji no motarashitamono (What the JFSAT has brought about). *Eigo Kyoiku Journal (Journal of English Teaching), 3*, 9-13.

Ikeura, S. (1990). Risuningu mondai wo kousatsu suru (Examination of listening comprehension test). *Gendai Eigo Kyoiku (Modern English Teaching), 9*, 6-7.

Imai, T. and Ito, A. (2001). Daigaku nyuushi mondai niokeru eigo tesuto nodatousei kenshou: naiteki balance no shiten kara (The internal construct validation on an English test in Japanese university entrance examinations). *CELES Bulletin, 30*, 161-166.

Ingulsrud, J.E. (1994). An entrance test to Japanese universities: Social and historical context. In C. Hill & K. Parry (EDS.), *From testing to assessment: English as an international language* (pp. 61-81). New York: Longman.

Inoi, S. (1995). The validity of written pronunciation questions: Focus on phoneme discrimination. In J.D. Brown & S.O. Yamashita (Eds.), *Language testing in Japan* (pp. 179-186). Tokyo: Japan Association for Language Teaching (JALT).

Ishii, K. (1981). Kyotsu ichiji no motarashita mono (What the JFSAT has brought about). *Eigo Kyoiku (The English Teacher's Magazine), 11*, 14-16.

Ishiguro, A and Tucker, G. (1989). Short stories for reading practice: Do you Know…?: 700-word level. Kyoto: Biseisha.

Ito, A. (2000a). Is cloze test more sensitive to discourse constraints than C-test? *International Journal of Curriculum Development and Practice, 2*, 67-77.

Ito, A. (2000b). Naiteki soukanhou niyoru aichi gakuin daigaku nyuushi

mondai (eigo) no kouseigainenteki dakousei kenshou (Internal construct validation study on the English language test in Aichi Gakuin University entrance examinations). *Nihon kyouka kyouiku gakkai shi (Journal of Japan Curriculum Development and Practice), 23,* 31-38.

Kashima, H., Tanaka, A., Tanabe, Y. and Nakamura, K. (1983). Zadankai: Daigaku nyushi eigo no mondaiten (Discussion: The problems with English tests in university entrance examinations). *Gendai Eigo Kyoiku (Modern English Teaching), 8*, 6-12.

Kira, T. (1981). Kyotsu ichiji no merrito tsuikyu wo (The pursuit of the JFSAT's merits). *Eigo Kyoiku Journal (Journal of English Teaching), 3,* 16-17.

Kiyomura, T. (1989). Dai 6 sho: Eigo kyoiku hyokaron (Chapter 6: Evaluation in English Language Education). In Y. Katayama, N. Endo, N. Kamita & A. Sasaki (Eds.) *Shin eigo kyoiku no kenkyu (Research on English language education: A new version)* (pp. 244-246). Tokyo: Taishukanshoten.

Klein-Braley, C. and Raatz, U. (1984). A survey of research on the C-Test. *Language Testing, 1,* 134-46.

Komazawa, S. and Ito, A. (1997). A written test of stress in connected speech in the NCUEE-Test: Its reliability and validity. *CELES Bulletin, 27*, 285-292.

Kuniyoshi, T. (1981). Daigaku nyushi shaken no kaikaku wa onsei shaken no donyu kara (The improvement of university entrance examinations should begin with the administration of aural tests). *Eigo Kyoiku Jouranl (Journal of English Teaching), 3*, 18-22.

Kyogakusha. (1992). *Daigaku Nyushi Series*. Tokyo: Kyogakusha.

Kyogakusha. (1994). *Daigaku Nyushi Series*. Tokyo: Kyogakusha.

Masukawa, K. (1981). Nippon eigo kyoiku kaizen kondankai: Dai 9 kai taikai hokoku, Diagaku nyishi no mondai wo chushin ni shite (The 9th round table conference on the improvement of English education in Japan: A report with special references to the problems with university entrance examinations). *Eigo Kyoiku Journal (Journal of English Teaching), 3*, 6-8.

Mochizuki, A. (1984). Effectiveness of multiple-choice (M-C) cloze tests (2). *CELES Bulletin, 13*, 159-164.

Mochizuki, A. (1992). Effectiveness of a multiple (M-C) cloze tests and the number of words in its text. *Annual Review of English Language Education in Japan, 3,* 33-42.

Mochizuki, A. (1994). C-tests, four kinds of texts, their reliability and validity. *JALT Journal, 16*, 41-54.

Monbusho (Ministry of Education, Science and Culture, Government of Japan). (1994). *Chugakko/ kotogakko gakushu shido yoryo (Course of study for junior and high school).* Tokyo: Printing Bureau, Ministry of Finance.

Monbusho (Ministry of Education, Science and Culture, Government of Japan). (1999). *Chugakko/ kotogakko gakushu shido yoryo (Course of study for junior and high school).* Tokyo: Printing Bureau, Ministry of Finance.

Nishida, T. (1985). Kurozu tesuto no mirareru tekisuto no sakujo hensu ni tsuite (Text deletion variables in cloze tests). *CASELE Research Bulletin, 15*, 47-54.

Nishida, T. (1986). Monogararibun kurozu to setsumeibun kurozu: Tekisuto no sentaku to kurozu tesuto (Narrative cloze and expository cloze: Text style and cloze tests). Gengo Bunka Kenkyu *(Foreign Language Culture Research), 12*, 124-43.

Nishida, T. (1987). Kurozu tesuto no datosei to tekisuto no sentaku (Cloze tests: Their validity and the selection of texts). In N. Kakita (Ed.), *Eigo Kyoiku Kenkyu* (pp. 433-444). Tokyo: Taishukanshoten.

Ohtomo, K. (1983). Kyotsu ichiji shaken no eigo mondai (English questions in the JFSAT). *Eigo Kyoiku (English Teacher's Magazine), 8*, 6-8.

Shiozawa, T. (1983). Daigaku nyushi: Genjo to kadai (University entrance examinations: Their present state and problems). *Eigo Kyoiku (The English Teacher's Magazine), 3*, 30-41.

Shirahata, T. (1991). Validity of paper test problems on stress: Taking examples from Mombusho's Daigaku Nyushi Senta Shiken. *Bulletin of the Faculty of Education, Shizuoka University, Educational Research Series, 23*, 17-22.

Siarone, A.G. and Schoorl, J.J. (1989) The cloze test: Or why small isn't always beautiful. *Language Learning 39* (4): 415-438.

Soft Ware International. (1988). Grammatik IV, version 1.0.2. San Francisco: Soft Ware International.

SPSS Inc. (1996). Statistical Package for Social Sciences Windows 7.5 version. Chicago, IL: SPSS Inc.

Steinberg, R. (1987). *Prentice Halls's practice tests for TOEFL*. Englewood Cliffs, NJ: Prentice Hall Regents.

Suzuki, H. (1981). Onsei tesuto donyu no kanosei (The possibility of administrating aural tests in the JFSAT). *Eigo Kyoiku Journal (Journal of English Teaching), 3*, 23-25.

Suzuki, S. (1985). Kyotsu ichiji shaken no mondaiten (The problems with the JFSAT). *Gendai Eigo Kyoiku (Modern English Teaching), 11*, 10-12.

Takanashi, T. (1990). Daigaku nyushi senta-shiken mondai (The JFSAT). *Eigo Kyoiku (The English Teacher's Magazine), 11*, 11-13.

Takanashi, T. (1984). Daizai no readability, gakushusha no kyomi, nakiyo no yasashisa to kurozu tokuten no kankei ni tsuite (The relationship between readability of texts, students' interests, easiness of contents and the scores in cloze tests*). KELES Research Bulletin, 13*, 14-22.

Takanashi, T. (1988). Kurozu tesuto no bunmyaku izonsei to tokasei (The sensitivity and equivalency of cloze tests*). KELES Research Bulletin, 17*, 33-40.

Wakabayashi, S. and Negisgi, M. (1994). *Musekinin na tesuto ga ochikobore wo tsukuru (Irresponsibly made tests keep students behind).* Tokyo: Taishukanshoten.

Yanai, H., Maekawa, S. and Ikeda, H. (1989). Kyotsu ichiji gakuryoku shaken (Kokugo, Sugaku I, Eigo B) ni kansuru komoku bunseki (Iten analyses of the JFSAT [Japanese, mathematics I, English B]). Daigaku Nyushi Forum (Forum on the university entrance examinations), 11, 169-182.

Appendix 1: cloze test (adopted from Ishiguro, A and Tucker, G. 1989 with permission)

One day wang lost his way while he was gathering wool. He wandered in the woods (01) hours, but could not find (02) path to lead him home. (03) came and Wang was tired (04) very hungry. When he passed (05) bug rock, he thought he (06) human voices. He walked around (07) rock and found a cave. (08) voices cam from the cave. (09) was almost dusk, but when (10) entered the cave, he noticed (11) was light and comfortable inside. (12) walked deeper into the cave (13) he came to a room at (14) end. Light and fresh air (15) from the ceiling.

Two men (16) sitting before a chess board. (17) were playing chess, chatting merrily. (18) neither talked to Wang nor (19) looked at him, but went (20) on playing. Now and then (21) drank from their cups which (22) held in their hands. Since (23) was so hungry and thirsty, (24) asked for a sip. For (25) first time they looked at (26) and smiled, offering him the (27) kind of a cup. Although (28) did not talk to him, (29) invited him to drink by gesture. (30) drink was fragrant and (31) as sweet as honey. Wang (32) had finished it all, (33) strangely enough, the cup was refilled (34) he noticed it.

Wang (35) no longer hungry nor thirsty (36) he drank from the cup. He (37) sat down beside the two (38)and watched their chess game. (39) two men continued playing chess, (40) chatting and laughing. The game (41) so exciting that Wang became (42) in it. It took some (43) before it was over. Maybe (44) hour or more had passed, (45) thought. He had spent too (46) time in the cave, and (47) good-bye to the chess (48) who gave him a bag (49) a souvenir.

After he came out of (50) cave, he could find his (51) home easily. However, when (52) entered his home village and (53) some people on the road, (54) did not know any of (55). They were all strangers. He (56) the place where his old (57) was, but there was nothing (58) a few decayed poles and (59). He did not understand what (60) happened, and looked around for (61) neighbors' houses. They were all (62) from what he used to (63). The people living there were (64) strangers too. Being at a (65) for what to do, he (66) the bag that the chess (67) had given him. Out came (68) stream of smoke, and in (69) minute, his hair had turned (70) and he found himself an old man. What does this story remind you of?