Introduction of Written Test in Evaluation of English-Japanese Interpreting Classes at Universities in Japan

Hiroko Yamada
(Doctoral course, Graduate School of Human and Environmental Studies,
Kyoto University)

One of the critical issues in the field of testing and assessment adopted by English interpreting courses at universities in Japan is the lack of a methodology for systematic testing and related assessment criteria. The present paper proposes a new testing model which can be employed for criterion-referenced testing at universities to replace the conventional "computer-based recorded verbal performance test." It is called "The Performance Test of English Interpreting in the Written Form", and its features include an assessment instrument, and a scoring rubric. Utilizing data from 160 students who concurrently took the identical interpreting performance tests in the recorded verbal form and the written-form, two tests were examined based on several theoretical constructs. The findings demonstrate the superiority of the written form to the recorded form and illustrate how the scoring rubric based rating system is a determining factor for the legitimacy of the performance test in the written form.

1. Introduction

Fairness and unbiased objectivity are needed in an assessment of interpreting skills, and the inherent objectivity of assessment directly influences the rating of an examinee's performance. To date, the field of interpreting studies has been exploring more empirical evidence for the quality of testing and assessment of interpreting performance. Nevertheless, no methodology has been established in this field yet and the approach to grading is still intuitive and impressionistic (Sawyer, 2004). "The assessment of interpreting performance is an area of research that is still in its infancy" (Angelelli and Jacobson, 2009, p.49).

Presently, the number of English/Japanese interpreting courses offered at universities in Japan is increasing. According to Recruit Marketing Partners Co., Ltd, (2005), there are currently 125 universities and junior colleges in Japan which provide students with lessons related to English interpreting. Considering the inherent nature of interpreting, which typically involves the verbal rephrasing of an utterance from a source language into a target

YAMADA Hiroko, "Introduction of written test in evaluation of English-Japanese interpreting classes at universities in Japan," *Interpreting and Translation Studies,* No.15, 2015. Pages 91-112.

language, traditional testing methods generally offered in normal English language classes, such as conventional paper tests including written translations, vocabulary quizzes, cloze tests, listening comprehension tests and others, cannot be appropriately employed by English interpreting courses. Nevertheless, functional tests with pragmatic procedures exclusively applied to this subject have yet to be constructed. As far as testing and assessment of English interpreting in the area of university academia are concerned, they may be defined as a relatively uncharted research area.

In order to investigate the testing methods which are currently implemented in English/Japanese interpreting courses, interviews were conducted to several instructors who taught English interpreting at universities. (E. Endo, K. Tanaka, personal communication, July 30, 2013) At present, there are primarily three distinct testing methods employed in English interpreting courses for criteria-referenced testing, of which one is typically employed by instructors in a generally arbitrary manner. They are “The conventional paper-based test on grammar and translation,” “The audition-type performance test,” and “The computer-based recorded verbal performance test.” The author used to administer a computer-based recorded verbal performance test in a CALL room. This is a testing method that requires an examinee to actually perform interpreting in front of a terminal while his or her performance is recorded. However; this testing method has some flaws. The most serious constraint is that it takes instructors a great deal of time and efforts to rate students’ recorded verbal performances, which are recovered through a USB flash memory. It is particularly arduous when an instructor is assigned to teach several classes numbering around 100 students, since this equates to as many as two thousand files of their recorded performances that need to be rated.

It is in this context that the use of a written form of assessment for English interpreting was created by the author as a replacement for the computer-based recorded verbal performance test. The purpose of this test is to improve ease of implementation and expedite marking and grading of the interpreted rendition, and furthermore to enhance the quality of assessment. Among other things, the test formulation includes a scoring rubric system, which was established by the author as an assessment instrument and which can be universally used by any rater.

The present paper first explores the earlier studies on development of testing and assessment in interpreting studies. It then examines the above-mentioned two methodologies of performance testing focusing, in particular, on the computer-based recorded verbal performance test. Then, as a means to gain more insights into interpreter training methodologies, in the section concerning the background of administering the tests, the author’s lessons are described, reflecting the tasks currently provided for the students. Finally, in order to compare the quality of “The Performance Test of English Interpreting in the Written Form” with the recorded verbal performance test, the author conducted identical

interpreting performance tests concurrently in two test modes, the verbal and the written form in a criterion-referenced test. The data used for this analysis consists of the test results of the mid-term and the final examinations with one hundred and sixty research participants in total. As for the use of the test results as research data, the written consent from the students was obtained.

The research question addressed in the present study is whether "The Performance Test of English Interpreting in the Written Form" can replace the computer-based recorded verbal performance test. The author argues that the written form is more functional than the recorded verbal form, and above all, the written form can improve the assessment process and enhance the reliability to a remarkable degree. This study uses the theoretical constructs of validity, reliability, feasibility and authenticity. In particular, to validate inter-rater reliability, the interpreting performances in these two test modes were also measured by other rater, using the same scoring rubrics.

2. Literature Review

Although the issue of the role and purpose of testing and assessment of interpreting should be argued in the broader educational context of curriculum design, the devising of curricula for interpreting studies is an academic field which has not been fully discussed by scholars yet. Angelelli and Jacobson (2009) argue that "Few researchers have focused on measurement of aspects of interpreting in general, quality in performance specially, and on the problem of assessing interpreting via the implementation of valid and reliable measures based on empirical research" (p.3). In fact, universities, specialized university schools of interpreting, international institutions, interpreter training institutes and other schools regularly implement tests in which students' interpreting performance is measured; however, surprisingly little literature on testing and assessment of interpreting performance has been published (Angelelli and Jacobson, 2009).

On the other hand, there are some scholars who delve into this field to identify available valid and reliable assessment instruments of measurement, e.g., Angelli (2001 and 2004b), Sawyer (2004), and Clifford (2005). Angelli (2001 and 2004b) developed the first assessment instrument for use in the healthcare context, which measures interpreting proficiency based on empirical data collected during ethnographic research. Sawyer (2004) conducted a case study on assessing interpreting competence and started discussion on test validation in interpreting. Clifford (2005) argued for more rigorous approaches to assessment with empirically developed constructs and competencies through development of an interpreter certification test. Angelelli and Jacobson (2009) compiled a collection of papers exclusively written for study of testing and assessment in interpreting across languages and settings including university classrooms, research settings, the private sector, and

professional association.

Meanwhile, in Japan, we can find no online evidence concerning detailed accounts of how testing and assessment are conducted in interpreting courses or programs at university. (e.g. the Academic Research Database Repository, Japanese Institute Repositories Online, Webcat Plus, KAKEN and CEFR-J) Research documents, papers and published books which specifically describe testing and assessment in interpreting courses at Japanese university are also surprisingly scarce, to the extent that they could be described as nonexistent. If we limit the literature concerning interpreting assessment to that of the university classroom in Japan, there are no documents available. This fact is indicative that the present status of research and study about testing and assessment of interpreting performance in Japan lags behind that of western countries.

3. Empirical Study on Testing and Assessment Implemented at Japanese Universities

There are a variety of approaches used for testing and assessment in English interpreting performance at Japanese universities and to the best of the author's knowledge, there are two primary methodologies currently being implemented. They are "the audition-type interpreting performance test," and "the computer-based recording verbal performance test." At present, the tests which are administered as a midterm or final examinations in interpreting courses or programs are totally left to the instructors' discretion, and thus test methods vary among instructors. These methodologies have advantage as well as disadvantage, but neither of them alone serves as a universal measuring instrument in English interpreting performance assessment at university. The following overview about testing methodologies provides insights into understanding the status quo of testing and assessment in university interpreting courses.

3.1. The Audition-type Interpreting Performance Test

In interpreting classes at universities, some instructors offer an audition-type interpreting performance test. For example, a teacher selects one student and reads several sentences for him/her in the source language or has him/her listen to a recorded passage in the source language, and perform an interpretation from the source language to the target language. This procedure is repeated in the mode of consecutive interpretation for a limited time in front of the instructor. An instructor simultaneously assesses the students' delivery and grades them on the spot. Grading criteria such as 5-point Likert scale is set by the instructor.

Assessment of each student's performance is undertaken at the discretion of the instructor and must be normally completed in short time. In other words, grading has to be executed on the spot, which means the examiner cannot take enough time to scrutinize the construct validity as explained by Sawyer (op.cit.:97), as the test takers' "ability to interpret with faithfulness to the meaning and intent of the original, ability to use appropriate language and

expression, and ability to demonstrate acceptable platform skills and resilience to stress." In this case, it is quite possible that the assessment from these criteria could be hampered by the lack of time and consequently, becomes impressionistic; whereby the validity of the test cannot be verified. Furthermore, since this test is implemented face to face between an instructor and a student, the assessment process requires a great deal of time, particularly in the case of a large class. Eventually, it tends to lead to reduction in the duration of testing time for each student, which also compromises validity and reliability.

3.2. The Computer-based Recording Verbal Performance Test

3.2.1. The Present Status

If a CALL system is available, direct performance testing for quite a large number of students at one time is feasible. In fact, the author has administered verbal interpreting performance tests in a CALL room during the mid-term and final examinations at her university over several years. The test procedures taken by the author are as follows:

During the test, students listen to a certain passage in English and verbally interpret it into Japanese, and their voices are recorded on their computers. They then listen to the next English paragraph and interpret it into Japanese. The same procedure is undertaken when interpreting from Japanese to English. This is called consecutive interpreting. The recording time is quite limited; thus, if the students repeat, recast or hesitate and lapse into silence or overlap, those lead to a shortage of recording time. Consequently, it is likely to cause some omissions of interpretation delivered by the students. They usually have about 10 test items (passages) to interpret from English to Japanese and 10 from Japanese to English. The consecutive interpreting from English to Japanese is repeated for around 20 minutes and interpreting from Japanese into English is performed likewise for the next 20 minutes. This means the students engage in consecutive interpreting for approximately 40 minutes in total for a midterm or a final examination.

3.2.2. Methodological problems

A student's interpreting performance for each passage is stored separately in a file and collected through a USB flash memory by the author. Each person accumulated as many as 20 files, or more; hence the total number of files taken on the occasion of a mid-term or final examination is approximately 2000 files, since the author usually teaches three classes which consist of more than 100 students during a semester. Not surprisingly, it takes a great deal of time and effort for the author to assess all the performances and grade them accordingly. This manner of testing is seriously demanding, challenging and highly impractical.

There are also several methodological problems: the most serious of which are as follows.
1. Due to a student making mistakes in operating the computer, his or her performance cannot be recorded.

2. Due to the teacher's mishandling of the computer, the files cannot be successfully recovered from the USB flash memory.
3. Due to an unexpected malfunctions of the computer, it accidentally freezes in the middle of the test. In this case, students' recordings cannot be completed and a retest has to be administered at a later date.

After recovering students' data, assessment is undertaken based on the following criteria. The construct for interpreting performance assessment consists of "content match" and "formal match." Content match includes completeness of rendition, terminological accuracy, and faithfulness to meaning. Formal match refers to rhetorical skills, voice, fluency and pronunciation (Sawyer, 2004). In the author's assessment, content match was the main target for assessment. Since the general interpreting abilities, which typical students registered in the university basic interpreting courses possess, are relatively poor, one interpreting item to be tested inevitably became quite short. If a formal match was strictly examined in rating the performance of such a short utterance as one test item, the construct of the test could be compromised because in this case, assessment based on a formal match tends to become impressionistic and subjective for the reasons described in the following section.

3.2.3. Cognitive Challenge

What needs to be noted here is that performance-based assessment could induce possible bias. A teacher faces a complex task, not only because of cognitive challenge, but also because of the risk of emotional strain (Vermeiren, Gucht, & Bontridder, 2009). Some problematic effects are "the significance effect (influence of another paradigm), the halo effect (when a judgment on a specific dimension is influenced by some other dimension) (Thorndike, 1999), the sequence effect (lasting effect of previous test taker), the contamination effect (influence of the grader's own agenda), the personal comparison (personal tendency to judge severely or compliant way) and impression management by the candidate." (ibid.:305)

These effects could be produced when an instructor is assessing students' verbal performance. In the process of analyzing a huge amount of files, his/her evaluation sometimes leads to assessment based on impression rather than criterion-based one. This induces rating on the basis of not only a specific interpreting performance just having been tested, but also on the overall performance exhibited by a student in the normal classes during a semester. In the worst case scenario, the judgment could be looked upon as favoritism by a teacher.

4. The Performance Test of English Interpreting in the Written Form

As can be seen from the drawbacks in rating verbal interpreting performance, an alternative method to measure students' abilities in English interpreting studies at universities can be deemed necessary. Therefore, "The Performance Test of English Interpreting in The Written

Form" has been tailored and named by the author. In this test, the students listen to a passage in the source language and do not output it verbally in the target language but write its translation on the test paper instead in a given time. Specifically, the students listen to an English passage and write its interpretation on the paper in Japanese and when listening to a Japanese passage, they write its interpretation in English on the paper. Both should be done within quite a limited time. "In consecutive interpreting, an interpreter starts to interpret when the speaker stops speaking, either in breaks in the source speech (discontinuous interpreting) or after the entire speech is finished (continuous interpreting)." (Gerver, 1976 cited in Christoffels and De Groot, 2004). Thus when consecutively interpreting in the real world, an interpreter has no time to elaborate on their rendition of each passage spoken by a speaker. In other words, an interpreter should convert what he or she heard from the source language into the target language and output the interpretation immediately after the speaker's utterance is finished. The task in the written form test would not be attainable without taking good notes and mentally summarizing the message while listening to speech. In this way, imposing a time constraint when writing down the interpretation on the test paper allows the students to experience the real sensation of verbal performance even though the test adopts a paper-based performance test mode. The test formulation includes a scoring rubric system, which was established by the author as an assessment instrument and which can be universally used by any rater.

5. Background

Before describing the method of data collection, a brief overview of interpreting training needs to be provided. To facilitate the data analysis and subsequent discussion, the standard procedures for several activities which the author followed in her lessons are described here. In the case of training in consecutive interpreting from English into Japanese, **t**he author administered these tasks in the following order for one lesson in the CALL room:

1. Activating schematic knowledge for a specific interpreting material for the day
2. Shadowing (Textbook closed)　Core text: Shadowing, Kadota, & Tamai (2004).
3. Synchronized reading (Parallel reading)　(Textbook opened)
4. Sight translation (CALL Monitoring function) (Textbook opened)
5. Retention or repeating (1 or 2 sentences) (CALL Model function) (Textbook closed)
6. Reproduction (3 to 5 sentences) with note- taking from the source language to the target language (Textbook closed)
7. Consecutive interpreting (3 to 5 sentences, with note taking) (Textbook closed)
8. Slash listening in English
9. Simultaneous interpreting

③ and ④ are applied only for the first half of a semester. For the second half, these activities are skipped. The visual check is left out in the second half.

The following steps are taken in training in ⑤,⑥ and ⑦. This set of procedures is repeated in the following order with respect to each passage with respect to each passage to which they listened.

1. Listening (The students listen to an English passage). 2. Rehearsal (The students repeat or reproduce or interpret what they have just listened to individually altogether.). 3. Model performance by a specific student. (By using model function in the CALL system, the author can ask the specific student to perform it and have the other students listen to his or her performance. The named student can perform better since this is the second pass.). 4. Model performance by the next student. (When the named student is not able to perform well, the author calls on other students and has them perform the same part again.). 5. Corrections and comments by the instructor (The author provides them with corrections or instructions). 6. Model performance by an instructor. (The author displays model performance.)

6. Method

6.1. Sample characteristics

In order to investigate the advantage and the superiority of “The Performance Test of English Interpreting in the Written Form,” the author conducted identical interpreting performance tests with two different test modes, a verbal and a written form given concurrently in a criterion-referenced test. The data used for this analysis consists of the test results of the mid-term and final examinations. The duration of the each test was 60 minutes. The results cover a period from April to July in 2014, during which both exams were conducted. The research participants were English major and non-major students of the third and the fourth year in the introductory courses in interpreting at the University of Foreign Studies where the author taught. No students have undergone interpreter training in the past.

In the present study, data was collected for each exam from 80 participants in four classes, i.e., the total number of 160 participants for both examinations. 80 participants were classified as “Class A,” “Class B,” “Class C,” “Class D.” Then a ninety-minute lesson was provided twice a week with Class A and B, and as for Class C and D, a ninety-minute lesson was provided once a week during the semester.

It was assumed that proficiency in English varied from student to student since a screening test was not conducted before the registration for this interpreting course. Hence, in order to check the difference in English proficiency among classes, a cut-down version of a TOEIC test comprising 50 questions rather 100 from the listening part, was conducted before the analysis. The perfect score was 50 points. The results are given in Table 1. Table 1 presents the range (minimum to maximum), mean, and, for interval measures, the standard deviation for each variable used in the analysis, for the total research participants across the

classes which were taught.

Table1. ANOVA: Descriptive Statistics of the Scores of Cut down Version of TOEIC Test (N=80)

Variables	Mean	S D	Range
Class A (n=19)	37.95	5.40	31.00-47.00
Class B (n=19)	37.58	6.13	28.00-49.00
Class C (n=22)	31.73	5.90	14.00-40.00
Class D (n=20)	39.15	7.37	23.00-50.00
Total (n=80)	36.45	6.81	14.00-50.00

Note. F (3,76) =6.010, $p < .001$

In order to compare the obtained scores of the four classes, an ANOVA was conducted. As shown, the result indicated that there was a significant difference among them. F (3,76)=6.010,$p< .001$ $p< .001$ Multiple comparisons showed that Class C was significantly lower than the other three. It can be speculated that this may be accounted for by the fact that there were fewer senior students in the class C although all four classes were mixed classes of the third and the fourth year students.

The purpose of this study is to examine how the superiority of the testing and assessment in interpreting performance in the written form can be validated, rather than to illustrate the effectiveness of interpreter training methods. Therefore, the previously mentioned fact does not directly influence the results of the analysis and is not relevant for the purpose of the present research. Hence the use of a non-representative sample is justified.

6.2. Data collection

Two main interpreting test items out of a total of six were selected for analysis in both cases, the midterm and the final examinations. One main item consisting of three passages to interpret from English to Japanese was given, and for interpreting from Japanese to English, two passages were presented. The same passages were offered as test items in two different modes, which are the recorded form and the written form.

The independent variables are the two test modes: the computer-based recorded verbal performance test and “the Performance Test in the Written Form.” The dependent variables included in the study were grouped into two factors: the assessment time taken by the author

for both the recorded form and the written form, and the marks obtained by the students for both the recorded form and the written form. Separate analysis was conducted for each dependent variable. Supplementary analysis was also conducted to ascertain the reliability of the written form. To confirm if the author and the other rater's assessments were consistent, the correlation between the assessing time and the marks obtained were checked.

In order to reduce the practice effect among the students, other test items were given between the two test modes. Specifically, the students of classes A and C took the recorded form first and did other interpreting test items which were not relevant to this research. Next they took the written form using the same materials as the recorded form. Classes B and D took the written form first and subsequently took the recorded form in the same way. Furthermore, in order to reduce the practice effect on assessment by a rater, the recorded form was assessed first and subsequently the written form for Classes A and C, and for Classes B and D, the written form was rated first.

6.3. Measurement Scale and Scoring Rubric

For the purpose of the interpreter test, three distinct types of scales are adopted: *nominal, ordinal and interval* scale. Firstly, a nominal scale consists of "classes or categories of a given attribute" (Bachman, 1990b, P.27) e.g., a scale indicating "pass" or "fail". Secondly, an ordinal scale "comprises the numbering of different levels of an attribute that are ordered with respect to each other" (ibid.:28) e.g., "high pass", "pass", "borderline fail", "fail". Thirdly, an interval scale "in which the distances, or intervals, between the levels are equal", e.g., 100-points multiple choice test with each item worth one point (Sawyer, op.cit:105).

The author revised the ordinal scale and made a scoring rubric which was applied for both the recorded and written forms based on the segmentation strategy or basic process. Because "appropriate segmentation into function units is a prerequisite for text processing."(Kirchhoff, 2002, P.114:18) A function unit is determined not only by the structure of the source language but by both decoding and redecoding conditions and the unit can be defined as the smallest possible decoding unit in the source language (Kirchhoff, 2002).

In the scoring rubric, for each phrase (semantic segment) in one sentence, specific points were allocated according to difficulty in translation. In order to ascertain the difficulty level in translating words, a "word frequency list" in an academic area (http://www.wordandphrase.info/frequencyList.asp) was used. To cite an example from the following test item, the word "hemisphere" is ranked 8751 with frequency level 3475 while the word "archipelago" is ranked 15112 with frequency level 779. Additionally based on the error classification (Cary, 1968 and Balzani, 1990 cited in Gerver 2002), the author created the error categories. They are "word omissions", "word substitutions", "additions of words", "meaning errors of words", "distortions of words", "errors in rendering figures and proper

names" and others.

As for the recorded form, the author checked the scoring rubrics and counted the marks of each uttered phrase or sentence on her fingers while listening to the recorded performance of the students. For the written form, the renditions of each test item were rated in parallel across the 4 classes using the same scoring rubrics to ascertain the reliability of not being disturbed by possible bias. As far as the handwriting of a specific word is concerned, if it is unreadable, it cannot be awarded a point. On the other hand, a misspelled word can obtain points on the condition that the word can recreate the sound of translated word. After all, the written form test is an interpreting test and what matters is the correct recreation of the message from the speakers, not the spelling of each word. One passage to listen and interpret is given here.

Example of test items and Scoring Rubric

1. Interpreting from English to Japanese.

"Japan is an archipelago (1) /situated in the middle latitudes (1) / of the northern hemisphere (1).

The Japanese archipelago consists of about 6,800 islands, (2) /which stretch from north to south (1) /in an arc form."(1)

Total 7 points

(Ohata, Okuda, and Rex, 2009)

2. Interpreting from Japanese to English.

"私は日光が好きです。(1) /そこの自然は本当に美しいです。(1) / 山あり、谷あり、森も滝もあります。(2) / 9 時の特急に乗れば 11 時前に着けます。(2) / 電車の旅は快適です。(1) / 途中景色も楽しめます." (1)

Total 8 points

(Shibata, 2004)

7. Results

7.1. Assessing time for identical two tests in different test modes.

In order to compare how much time the author took to rate the part of the recorded form to the written form in the mid-term examination as well as the final examination, assessing time for each form is shown in the Figure 1 and Figure 2. E to J or J to E indicates the direction of interpretation.

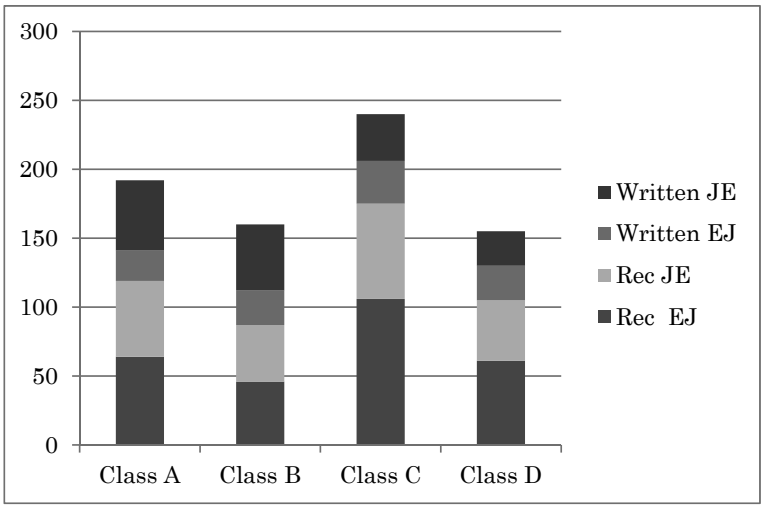

Figure 1. Assessment time of the students' interpreting performances of the mid-term examination, which was measured in minutes by the author.
Written JE: Interpreting from Japanese to English in the written form
Written EJ: Interpreting from English to Japanese in the written form
Rec. JE: Interpreting from Japanese to English in the recorded form
Rec. EJ: Interpreting from English to Japanese in the recorded form

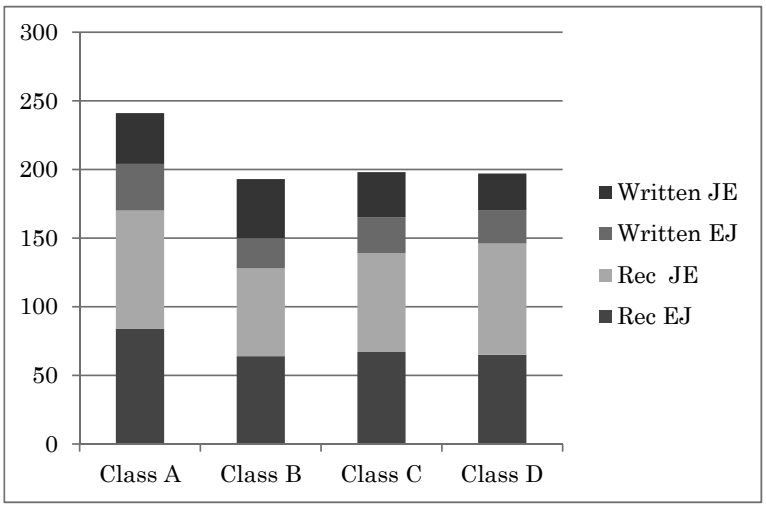

Figure 2. Assessment time of the written form and the recorded form of each class in the final examination, which was measured in minutes by the author.

7.2. Marks obtained by the students for identical two tests in different test modes.

In order to explore the inter-method reliability, the marks obtained in the recorded form were

compared to the written form for both the mid-term and final examinations.

Table 2 indicates the means in the recorded form and the written form in the mid-term examination and the final examination. Table.3 presents correlation coefficients of the marks obtained. The parameter estimates the marks obtained via the measurement model in both the recorded form and the written form for the midterm and the final examination respectively.

Table.2 Marks obtained by the students on two test modes for the midterm and the final exam.

Mid-term exam (N=80)

Variables	Recorded form		Written form	
	Mean	S D	Mean	S D
Class A	24.42	6.49	27.00	6.17
Class B	24.63	6.73	20.84	7.87
Class C	18.23	8.08	19.64	7.14
Class D	22.70	7.04	17.95	7.03

Final Exam (N=80)

Variables	Recorded form		Written form	
	Mean	SD	Mean	SD
Class A	22.84	6.37	26.42	6.52
Class B	21.74	7.67	18.79	9.11
Class C	18.41	6.79	22.14	7.00
Class D	24.00	6.70	21.10	7.42

Table 3. Pearson's Correlation Coefficients of obtained marks between two test modes N=160

Parameter Estimates	Written form E to J Midterm	Written form J to E, Midterm	Written form E to J Final	Written form J to E Final
Recorded form E to J, Midterm	**.66****			
Recorded form J to E, Midterm		**.68****		
Recorded form E to J, Final			**.76****	
Recorded form J to E, Final				**.78****

Note. E to J=interpreting from English to Japanese
J to E= interpreting from Japanese to English
***r*< .01

Table 3 indicates that the coefficient of the obtained marks of the recorded form and the written form with respect to interpreting from English to Japanese for the midterm examination was significant at *r*= .66. The coefficient of the recorded form and the written form in interpreting from Japanese and English for the midterm examination was also significant at *r*= .68. Concerning the final examination, the coefficient of the recorded form and the written form in interpreting from English to Japanese, and the coefficient from Japanese to English were also significant at *r*= .76 and *r*= .78 respectively. In summary, the finding supports the theory that all coefficients are significant and are of similar magnitude and in the same direction; thus the selected variables are significant predictors in which two test modes are consistent, and the inter-method reliability between the recorded form and the written form is quite high.

7.3. Inter-rater reliability between the author's assessment and other rater.

For the purpose of further ascertaining the reliability of the testing and assessment in the written form, the results produced from the author's data needed to be consistent with the other raters. Initially, one other teacher who is an instructor of English interpreting studies at a university, (-different from the university the author works-), was asked to assess the same test conducted for this present research. The rater was first given the opportunity to familiarize herself with the test materials by listening to the test items and meticulously checking the scoring rubrics that had been made. This activity was very useful because the

rater learned how to recognize the errors that existed in the interpreting performance of the students in both the recorded form and the written form. Next, the teacher rated the same data of the midterm examination which had been conducted earlier, particularly for Class A. Table 4 presents the assessing time taken by this rater. Table 5 analyses assessing time taken by the author and this rater based on the Multiple Comparison. Table 6 shows Pearson's analysis that explores the correlation coefficients of the marks obtained by the students rated by this rater and the author.

Table 4. Assessment Time Taken by the Other Rater for the Mid-term Examination of Class A

	Recorded form E to J	Recorded form J to E	Written form E to J	Written Form J to E
Other rater	66 min	42 min	30 min	20 min

Table 5. Multiple Comparison using Tukey's HSD Dependent Variable: Time

		Mean difference	Standard Error	Sig.	Confidence Interval	
					Lower	Upper
Recorded E to J	Recorded J to E	.27	.32	.84	-.60	1.14
	Written E to J	2.16*	.32	.00	1.29	3.03
	Written J to E	1.59*	.32	.00	.71	2.46
Recorded J to E	Recorded E to J	-.27	.32	.84	-1.14	.60
	Written E to J	1.89*	.32	.00	1.02	2.76
	Written J to E	1.32*	.32	.00	.45	2.19
Written E to J	Recorded E to J	-2.16*	.32	.00	-3.03	-1.29
	Recorded J to E	-1.89*	.32	.00	-2.76	-1.02
	Written J to E	-.58	.32	.29	-1.45	.297
Written J to E	Recorded E to J	-1.59*	.32	.00	-2.46	-.71
	Recorded	-1.32*	.32	.00	-2.19	-.45

	J to E					
	Written E to J	.58	.32	.29	-.30	1.45

Note. *. The mean difference is significant at the 0.05 level. F (3,28)=21.09,p< .01

As is shown in Table 4, the other rater reduced the assessment time by more than half through the written form compared to the recorded form in both cases for interpreting from English to Japanese and from Japanese to English.

In order to show the assessing time for the recorded form from English to Japanese, the recorded form from Japanese to English, the written form English to Japanese, the written form from Japanese to English, are statistically different, the author's and the other rater's assessment time for each method,- assessment time for eight methods in total-, were analyzed based on an Analysis of Variance.Table 5 shows the result of a one-way ANOVA examining the effects of scoring methods. A significant effect was found (F (3,28)=21.09, p<.01). Comparison using Tukey's contrasts found a statistical difference between two recorded modes and two written modes as seen in Table 5.

Table 6. Pearson's Coefficients of the Assessment between Other Rater and the Author in terms of the Marks Obtained by Class A in the Mid-term Examination

	Author Recorded form E to J	**Author** Recorded form J to E	**Author** Written form E to J	**Author** Written form J to E
Other rater Recorded form E to J	**.894****			
Other rater Recorded form J to E		**.76****		
Other rater Written form E to J			**.760****	
Other rater Written form J to E				**.71****

Note. E to J=interpreting from English to Japanese
J to E= Interpreting from Japanese to English
***r*< .01

As can be seen in Table 6, the coefficients of the marks obtained by Class A rated by the other rater and the author were significant varying from *r*= .71 to *r*= .894. This evidence is a significant predictor suggesting that an inter-rater reliability was established with regard to the interpreting performance tests both in the recorded form and the written form. In this sense, the message equivalency rater trainings for the interpreting tests, which sometimes need to be conducted face to face with one consultant training the group, are not necessary if the raters adopt the methods of the scoring rubric system that the author established.

8. Discussion

8.1. Analysis of new testing model based on theoretical constructs

"The Performance Test in the Written Form" will be analyzed from the perspective of validity, reliability and feasibility. First, validity will be examined. The written form measured what is supposed to be measured in an interpreting test, that is, the examinees' reproducibility and fidelity from the source language to the target language. More specifically, the written form allowed the raters to precisely measure "completeness of rendition," "terminology accuracy" and "faithfulness to meaning" (Moser, 1995) by following the scoring rubric made by the

author. Thus this test can be claimed to produce high validity.

Second, reliability will be explored. It was found that the acquired data illustrated quite similar results in terms of the marks obtained between the verbal recorded form and the written form. This means that the written form achieved high inter-method reliability. When assessing the written form, visual observation of both the scoring rubrics and the interpretation which test takers wrote on their test papers was possible. Hence the criteria and indicators capable of objectively measuring this process can be observed more strictly. By comparing the written interpretation with the scoring rubrics during the assessment, the raters were able to allot the precise points to the test takers' renditions of a specific passage. Furthermore, the method of rating the same test items in parallel across the four classes ensured further reliability without being disturbed by the sequence effect, halo effect, contamination effect and other possible biases. On the other hand, this approach cannot be applicable to the verbal recorded form because searching a specific file from a huge number of files is quite laborious and time-consuming. It was also noted that the correlation coefficients of ratings between the other rater and the author were significant. Thus inter-rater reliability of the written form was ascertained in this study, which means any rater can produce almost the same results for assessment as any other if they follow the methods including the scoring rubrics that are advocated in this study.

Third, concerning feasibility of deploying this form of assessment more widely, the most pronounced finding in this study was that the written form made it possible to cut down the assessment time by half compared to the verbal recorded form even though two test modes produced quite similar results. In this respect, the evidence obtained more clearly supports the author's hypothesis that assessing time is significantly reduced in the written form in comparison with the recorded form. This fact will greatly contribute to reducing the burden of raters in the assessment process which has typically overloaded them in the case of verbal recorded form. In addition, the written form can be implemented not only in the CALL room but also in the normal classroom settings if a portable CD player is available. Furthermore, in contrast to the recorded form in which frequent technical problems occurred when operating the computer, later recovering the huge amount of files -as many as two thousand-, and listening to them at home with the personal computer, the written form test is easy to implement and conduct assessments on it. In this respect, it has more feasibility than the computer-based recorded performance test.

9. Limitation of the study

Despite the positive character of "The Performance Test in the Written Form" described in the previous section, considerations arise from the new testing model. It concerns how the authenticity of the written form test can be verified. Since the written form did not require any verbal output from the test takers, it could not simulate the real world situation of

interpreting in a strict sense. However, the author complemented the written form by assigning some constraints to it. As was mentioned in chapter 4, given time for writing their interpretation on the test papers after listening to each utterance from the speaker was quite limited, in an attempt to make the test process of the written form more closely resemble the real world situation.

The author argues that what is important in rating interpreting performance, on the occasion of criterion referenced test targeting the university students in an introductory course in which most of them have never undergone interpreter training, is to assess interpretation based on the content match (completeness of rendition, terminology accuracy, faithfulness to meaning) rather than the formal match (synchronicity, rhetorical skills, voice) (Moser, 1995). Because one utterance to interpret in a test item which they are capable of is quite short compared to normal interpreter tests, hence the formal match is not notable for counting as a point in the first place. Thus, the interpreting performance tests provided by the university's interpreting courses may not be held accountable for pursuing their authenticity in a technical sense. However; given that there still exists a need for something to supplement the existing test with the verbal performance assessment, the author set a certain period of time to assess verbal interpreting performance of individual students from everyday classes in the CALL room to closely assess their pronunciation, voice quality and fluency.

10. Educational Implication

In order to fulfill the maximum requirements that the interpreting performance test holds, it is recommended that a test, combining the recorded form and the written form, be introduced by the instructors at the university level. In this case, they need to work on ensuring that their testing processes are appropriate while checking the raters' workloads for assessment in order for it not to become too burdensome.

11. Conclusion

The present study attempts to examine advantages and drawbacks of the interpreting performance testing and assessment methods which are adopted to a large extent in the English interpreting courses at Japanese universities. Then the author developed "The Performance Test of English Interpreting in the Written Form" to replace the computer-based recorded verbal performance test. One of the decisions made in revising the verbal performance test was to retain the criteria and overall standard from the original test. As opposed to lowering the performance standard, scoring rubrics which can be applied to both modes were established in the pursuit of a more detailed marking system.

The goals of the data analysis in this study were twofold. The first was to examine whether the written form could produce the same assessment results as the recorded verbal form in terms of the marks obtained. The second one was to investigate the assessment time

taken by the raters in both modes. Evidence was obtained to indicate that inter-method reliability between two modes was significant and the written form succeeded in considerably curtailing the assessing time compared to the recorded verbal form. With respect to other theoretical constructs, the written form also fulfilled the requirements to ensure its validity and feasibility. In summary, the above mentioned findings lend support to the hypothesis that "The Performance Test of English Interpreting in the Written Form" can replace the computer-based recorded verbal performance test, the written form is more functional than the recorded verbal form, and above all, the written form can remarkably facilitate the assessment process.

On the other hand, the analysis has implications for the written test from the theoretical perspective of its authenticity. More studies are deemed necessary if instructors are to pursue a genuine and authentic interpreting performance test which could develop and be easily carried out as a criterion-referenced test in a university classroom. However, despite these data limitations, the findings of the present paper advance the future course of testing and assessment in English interpreting pedagogy.

The author wants to continue to eagerly engage students and instructors in dialogue about the new testing model and carry through the reform of the methodology of testing and assessment in interpreting studies at universities in a responsive and comprehensive manner.

.....

About the author

YAMADA Hiroko is an associate professor at Kansai Gaidai College and freelance interpreter. Her research interests include interpreting pedagogy, particularly testing and assessment of interpreters' performance.

.....

Notes

This paper was written by revising and editing the master's thesis of the author.

References

Angelelli, C., & Claudia, V. (2001). Deconstructing the invisible interpreter: A critical study of the interpersonal role of the interpreter in a cross-cultural/linguistic communicative event. *Doctoral dissertation, Stanford University*: *Dissertation Abstracts International*, 62, 9, 2953.

Angelelli, C. (2004b). *Revisiting the interpreter's role: A study of conference, court and medical interpreters in Canada, Mexico, and in the United States*. Amsterdam, The Netherlands: John Benjamins Publishing Company.

Angelelli, C., & Jacobson, H. (2009). *Testing and assessment in translation and interpreting studies: A call for dialogue between research and practice.* Amsterdam, The Netherlands: John Benjamines Publishing Company

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

Balzani, M. (1990). Aspects of applied and experimental research on conference interpretation *Campanotto Editore* (pp.93-100). Udine, Italy: Campanotto

Chrisroffels, I., & De Groot, A. (2004). *Components of simultaneous interpreting comparing interpreting with shadowing and paraphrasing*. Cambridge, UK: Cambridge University press

Clifford, A. (2005). A preliminary investigation into discursive models of interpreting as a means of enhancing construct validity in interpreter certification. *Doctoral Dissertation, University of Ottawa*: *Dissertation Abstracts International.*66, 5, 1739.

Gerver, D. (1976). *Empirical studies of simultaneous interpretation: A review and model,* in R.W. Brislin (ed.) New York, NY: New York Gardner Press

Gerver, D. (2002). The effects of source language presentation rate on the performance of simultaneous conference interpreters, *The interpreting studies reader* (pp.53-60). New York, NY: Routledge

Kirchhoff, H. (2002). Simultaneous interpreting, Interdependence of variables in the interpreting process, interpreting models and interpreting strategies*, The interpreting studies reader (*pp.111-120) New York, NY: Routledge

Kadota, S. & Tamai, K. (2004). *Shadowing*. Tokyo, Japan: Cosmopia

Moser, P. (1995). *AIIC (International Association of Conference Interpreters) Survey on Expectations of Users of Conference Interpretation*: Vienna, Austria: S R Z Stadt + Regionalforschung GmbH, Lindengasse 26/2/3. A-1070

Ohata, K., Okuda, M., & Tanimoto, R. (2009). *Introduction to Bilingual Interpretation*. Osaka, Japan: Osaka Kyoiku Tosho

Pöchhacker, F., & Shlesinger, M. (2002). *The Interpreting Studies Reader*. New York, NY: Routledge

Recruit Marketing Partners Co., ltd. Retrieved October 10, 2015 http://shingakunet.com/

Sawyer, D. (2004). *Fundamental aspects of interpreter education: Curriculum and assessment*. Amsterdam, The Netherlands: John Benjamines Publishing Company

Shibata,V. (2004). *Beginner's Simultaneous Interpretation by Whispering*. Tokyo, Japan: Nanunndou

Thorndike, E.(1999). *Education psychology: briefer course*, New York, NY: Routledge,

Vermeiren, H., Gucht, J., & De Bontridder, L. (2009). Standards as critical success factors in assessment. Certifying social interpreters in Flanders, Belgium, *Testing and assessment*

*in translation and interpreting studies(*pp.297-329). Amsterdam, The Netherlands: John Benjamins Publishing Company

Word Frequency List, Retrieved October 10, 2015,

http://www.wordandphrase.info/frequencyList.asp