Translation as a language testing procedure: does it work?

Gary Buck *East Texas State University*

This article looks at translation from L2 to L1 as a language testing procedure. Two studies are presented. The first examines the reliability and validity of a translation reading test under conditions where minimal, or no effort is made to ensure rater reliability. Seven raters, with no instructions or training, rated translations by 121 Japanese junior college students. All estimates of reliability proved acceptable. Criterion validity was examined by correlating the translations with other measures of passage comprehension, and were found to be satisfactory. The second study reports the inclusion of two tests of translation in a multitrait-multimethod validation study. They showed satisfactory reliability as estimated by four different methods, and examination of the correlation matrix indicated that the translation tests had acceptable construct validity with almost no method effect. It is argued that these results are generalizable to many other translation tests in other situations, and hence the widespread rejection of translation as a language testing procedure by teachers and testers is probably not warranted on psychometric grounds. However, it is further argued that translation would often have an undesirable washback effect on classroom practice, and thus ought to be used with extreme care.

One of the most common testing methods used in foreign language teaching is translation, and yet there has been very little research to indicate just how effective a testing method translation really is. Commenting on this rather unfortunate state of affairs, Klein-Braley contends that even though it may not be as exciting as pushing forward the frontiers of knowledge 'we owe it to our own reputation as a serious discipline to monitor current language testing practice, and translation is all too often at present the test in use' (Klein-Braley, 1987: 6).

I have been involved in the construction of numerous English language tests in Japan, where translation is a commonly used testing method, and have frequently felt the need for empirical research to support my arguments that translation is not considered to be a respectable testing technique among language testing researchers. This, coupled with the comments of Klein-Braley above, served as the motive for the present research.

Klein-Braley notes that she could only find two studies dealing with translation (Klein-Braley, 1982; Oakeshott-Taylor, 1981; both

© Edward Arnold 1992

of which are in German). She therefore reviews briefly what both language testers and translation theorists have to say about the use of translation as a testing tool. She suggests that they are virtually unanimous in rejecting it, and offers three reasons why this should be so:

1) It is not necessarily the case that a good language student is also a good translator.
2) Although it is possible to have objective and reliable marking, 'testers know that translation testing as it is actually performed at present is not objective' (Klein-Braley, 1987: 5).
3) It is not at all clear what trait or skill translation is supposed to measure.

In order to try to address some of these questions she examined data which had accumulated in Duisburg University. The texts examined were passages of German for translation into English. She concludes that translation does seem to measure 'general language proficiency' (Klein-Braley, 1987: 16), but is not the best measure of this available. Furthermore, due to the immense effort involved in ensuring intra- and inter-rater reliability, the procedure was extremely uneconomical and even then the majority of tests investigated were not reliable. She further speculates that under most testing conditions, where less attention is paid to reliability, such tests would be far less reliable and hence lack validity. She concludes with a call for further research into the use of translation as a testing technique, particularly inter-rater reliability under 'natural' conditions.

In second language testing in Japan translation is usually taken as meaning translation from the target language (in this case English) into the first language (Japanese); translation into the target language is referred to as 'composition'. Although there are variations in format, the most commonly used translation tests give students an English passage in which one or more parts are underlined. The students are given instructions to translate the underlined part into Japanese. This constitutes one item, and is usually rated on a scale which may vary from about three points up to about ten. It is my experience that there is at best minimal discussion of rating standards, and calculations of inter-rater reliability are very rare. Raters generally allocate marks as they see fit, presumably under the assumption that they all know what the sentence means, and what constitutes a good, or a bad, translation of it.

While it is a relatively straightforward matter to examine the reliability of these procedures, validity is not so simple. As noted by Klein-Braley it is not clear what translation is supposed to measure.

It is obviously intended to be an indirect measure of some important second language trait, but is it general language proficiency, grammatical knowledge or what? Discussions with Japanese teachers and testers were only partly helpful. About 30 were approached and asked what they thought English to Japanese translation tests measure. In most cases it was obvious that they had not really thought about this question before, but when pressed responses ranged from a) those who felt that translation measures general language proficiency or comprehension, through b) those who felt it measures language proficiency as well as something else, such as translation skills or general academic ability, to c) those who felt it only measures the extent students have mastered that form of school text-book grammar taught in Japan.

Eventually a brain-storming session was held with a group of 10 university teachers involved in writing university entrance examinations. After much discussion general agreement was reached that translation as used in Japanese university entrance examinations is intended to measure 'passage comprehension'. It must be stressed that this agreement did not extend to a suitable definition of passage comprehension, and it is quite likely that there were a number of interpretations of this. However, all agreed that comparison with other measures of reading passage comprehension would provide a suitable criterion against which to assess the validity of the translation tests.

Two studies are reported here. The first was designed to examine the reliability and criterion validity of typical translation tests under conditions as similar as possible to those used in the 'real world'; in this case the real world of Japanese junior college entrance examinations. The second study looks at the reliability and construct validity of two other translation tests, one of listening and the other of reading, which were included in a battery of eight tests designed to examine the construct validity of listening and reading comprehension by means of a multitrait-multimethod correlation matrix.

I Study 1

1 Method

To examine the reliability of these tests it was decided to administer three English prose passages to a group of students and compare ratings of their translations by a number of different raters. The passages were taken from a published collection of past junior college entrance examinations (Obunsha, 1986). From each passage there were two sentences for translation into Japanese, which were

underlined. This made a total of six sentences for translation in the whole test (see Appendix A). An informal attempt was made to choose passages at different levels of linguistic difficulty. These were then administered to a group of 121 English majors as part of their end-of-course tests in the English programme at a junior college in Osaka.

A number of copies of the 121 completed translation protocols were made with the students' names erased and replaced by a code number. Copies of these were then given to seven different raters, and each rater was asked to rate all the translations. The only instructions given to the raters was that they should rate each sentence with a maximum of five points and a minimum of zero, and that they should use their own judgement as to how the sentences should be rated. The raters were all full-time English teachers in junior colleges in the Osaka area. Most of them were engaged in making and scoring junior college entrance examinations. They were chosen with a view to getting as wide a range of attitudes as possible towards the utility of translation in the classroom. They ranged in age from early thirties to upper sixties, the younger ones tending to be representative of the more modern communicative approach to English teaching, and the older ones more inclined to be long-term users of the grammar-translation method. Some of the younger raters had lived abroad for a number of years, while the more traditional teachers had not.

In order to examine the validity of the translations two other measures of comprehension were constructed on the same texts: first, a 36-item random-deletion cloze test (Appendix B), and secondly, a 23-item multiple-choice comprehension test. In order to ensure that the M/C test was testing students' comprehension of the passage rather than the questions, the test questions were asked in Japanese. An English translation is given in Appendix C. All the tests were administered in one hour: 20 minutes for each test. The cloze test was given first, followed by the translation and finally the multiple choice test.

2 *Results*

Table 1 shows descriptive statistics for each rater on each of the three passages, and on the total test score. Each passage, of course, has a maximum score of 10 and the total test a maximum of 30. Raters differ somewhat in the severity of their judgements, especially on Passage 1, where, for example, Rater F is much more severe than Rater D. On the whole test Rater B is the most severe and Rater D the least. Whether these differences are significant can be

examined by means of analysis of variance. Table 2 gives sums of squares and mean squares for the total ratings on the whole test. Dividing the between rater mean square (142.381) by the within rater mean square (34.066) gives an *F*-statistic of 4.18, which is statistically significant with a probability of less than .01. This clearly indicates that raters do differ significantly in the severity of their judgements.

Reliability: The reliability of these translations can be estimated in a number of ways. First, each of the six sentences can be regarded as one item, or subsection of a total test, and the internal consistency of the whole test, for each rater, can be estimated by means of Cronbach's Alpha. Table 3 gives these estimates for each of the seven raters. Estimates for internal consistency range from .71 to .80; not high, but certainly not so low judging by reliabilities often reported in the second language literature.

Secondly, the most common method of estimating inter-rater reliability between raters, the correlation coefficient, can be used. Table 4 shows the correlations between the raters on each of the three passages, and for the total test score. Passage 1 shows

Table 1 Descriptive statistics of seven raters on three passages and total test

	Rater A	Rater B	Rater C	Rater D	Rater E	Rater F	Rater G
Passage 1							
Mean	2.95	3.22	4.19	4.22	3.55	2.59	2.66
SD	2.11	2.57	2.47	2.36	2.28	2.37	2.09
Passage 2							
Mean	4.59	3.46	4.16	5.37	4.41	4.99	4.50
SD	2.77	2.34	2.23	2.26	2.40	2.29	2.26
Passage 3							
Mean	2.65	3.14	3.26	3.81	3.15	3.04	3.44
SD	2.71	2.64	2.16	2.58	2.64	2.59	2.66
Total							
Mean	10.19	9.82	11.61	13.40	11.10	10.62	10.60
SD	5.92	5.96	5.30	5.98	5.98	6.09	5.55
(N=121)							

Table 2 ANOVA of seven raters on total test

Source	DF	Sum of squares	Mean square
Between subjects	120	25 970.397	216.420
Within subjects	726	3 499.714	4.821
Between raters	6	854.288	142.381
Within raters	840	28 615.823	34.066
Residual	720	2 645.426	3.671
Total	**846**	**29 470.111**	

Table 3 Cronbach's Alpha for seven raters

Rater A	.72
Rater B	.73
Rater C	.72
Rater D	.76
Rater E	.78
Rater F	.80
Rater G	.71
(N=121)	
Calculated by taking each of the six sentences as one item	

Table 4 Correlations of passage scores and total test scores

	Rater A	Rater B	Rater C	Rater D	Rater E	Rater F	Rater G
Passage 1							
Rater A	1.00						
Rater B	.74	1.00					
Rater C	.68	.81	1.00				
Rater D	.79	.84	.87	1.00			
Rater E	.66	.86	.84	.84	1.00		
Rater F	.68	.87	.85	.82	.84	1.00	
Rater G	.68	.81	.76	.78	.80	.77	1.00
Passage 2							
Rater A	1.00						
Rater B	.77	1.00					
Rater C	.78	.80	1.00				
Rater D	.84	.79	.80	1.00			
Rater E	.82	.83	.80	.91	1.00		
Rater F	.82	.78	.78	.89	.86	1.00	
Rater G	.79	.78	.73	.83	.79	.82	1.00
Passage 3							
Rater A	1.00						
Rater B	.90	1.00					
Rater C	.83	.85	1.00				
Rater D	.82	.89	.91	1.00			
Rater E	.85	.87	.86	.88	1.00		
Rater F	.90	.88	.90	.89	.89	1.00	
Rater G	.91	.90	.89	.87	.90	.93	1.00
Total							
Rater A	1.00						
Rater B	.88	1.00					
Rater C	.87	.88	1.00				
Rater D	.90	.91	.93	1.00			
Rater E	.89	.92	.89	.93	1.00		
Rater F	.89	.90	.89	.92	.91	1.00	
Rater G	.88	.91	.89	.90	.91	.93	1.00
(N=121)							

reasonably acceptable inter-rater correlations considering that it is only rated on a 10 point scale; Passages 2 and 3 are even higher. Inter-rater correlations for the total test are quite high; the lowest coefficient being .87 and the highest .93.

Many researchers in second language testing would consider these correlations as evidence of satisfactory reliability. However, Krzanowski and Woods (1984) have argued that the inter-rater correlation coefficient is not a good estimate of reliability because it fails to take into consideration the fact that, while two alternate forms of the same test (or ratings of the same test) may correlate very closely together, the mean scores may be very different. In any testing situation where ratings from different raters are considered alternative forms of the same test, it is important that different levels of severity between raters be taken into account when estimating inter-rater reliability, since this between-forms variance is in fact error variance. Table 1 shows that raters in this study do in fact differ in the severity of their judgements. In such a case, Krzanowski and Woods argue that reliability is better estimated by using analysis of variance, in which this between forms (or between raters) variance can be included in the error. Using the analysis of variance from Table 2, estimates of reliability are given in Table 5, for one rater and for a pool of all seven raters. Naturally, the reliabilities for one rater are lower than the inter-rater correlations, but the reliability estimate for one rater on the whole test, .86, is quite high and falls comfortably within the range of what is often reported in the literature as an acceptable level of reliability.

Validity: The question of validity was approached from two directions. First, factor analysis was used to ascertain to what extent the seven raters were using the same criteria, and secondly, correlations with criterion measures of passage comprehension were calculated to ascertain to what extent the ratings were indeed measuring passage comprehension.

Table 6 gives the results of the factor analysis. The analysis produced four factors, but Factor 1 accounted for almost 98% of the variance in the matrix, and the factor loadings show that all the ratings load very highly on this first factor. These loadings on the first factor range from .93, the lowest, to .96, the highest. The highest loading on any of the other factors is .19. This indicates very clearly that there is only one factor in the matrix, which suggests

Table 5 ANOVA estimates of reliability for one rater and all seven raters

	One rater	Seven raters
Passage 1	.727	.949
Passage 2	.755	.955
Passage 3	.861	.977
Total	**.862**	**.978**
(N=121)		

that the seven raters were all using very similar criteria to rate the translations.

Even if the raters do agree on the criteria they use, there is still the question of what these criteria are. Descriptive statistics of the two criterion measures, the cloze and multiple-choice test, are given in Table 7. Unfortunately, due to time constraints and the lack of a second population similar to the one used in the main study, it was not possible to pretest the cloze and multiple choice tests. These tests have a reliability, estimated by internal consistency, of .77 for the cloze and .70 for the multiple choice, which seems far too low for a criterion measure. However, the best estimate of passage comprehension is the combination of the cloze and the multiple choice tests. Taken together as one test they have an internal consistency of .84 (Table 7), which is still not so high, but probably sufficient to give some indication of the criterion validity of the translation tests. This lack of reliability in the criterion measure will tend to reduce the estimate of the validity of the translation test. The validity of these tests can be better estimated if we correct the validity correlation for attenuation in the criterion but leave the translation test uncorrected (Guilford and Fruchter, 1978: 452).

Table 6 Unrotated factor analysis of seven raters on translation total score (principal factor method)

Factor	Eigen values	%
1	6.334	97.89
2	.068	1.06
3	.043	.66
4	.036	.55

	Factor loadings			
	1	2	3	4
Rater A	.93	−.00	.11	.13
Rater B	.95	−.08	.05	−.09
Rater C	.94	.19	−.07	−.02
Rater D	.96	.09	.03	−.01
Rater E	.96	−.04	.07	−.07
Rater F	.96	−.07	−.08	.01
Rater G	.95	−.11	.11	.01
(N=121)				

Table 7 Descriptive statistics of cloze, multiple choice and combined tests

	Items	Mean	SD	Alpha
Cloze	36	7.68	4.21	.77
Multiple choice	23	12.11	3.76	.70
Combined test	59	19.79	7.28	.84

Correlation between cloze and multiple choice = .67
(N=121)

Table 8 gives correlations between the seven raters and the cloze test, multiple choice test and the combined test, both uncorrected and corrected for attenuation in the criterion measure.

After correction for attenuation in the criterion the translation tests predict the cloze scores better than multiple choice scores for all raters, and even predict the cloze scores better than the combined scores with four out of the seven raters. This could either indicate that translation tests measure something more like cloze tests than multiple-choice tests, or could simply be due to the characteristics of these particular tests. In the present case there is no sound theoretical reason to regard the cloze scores as the best criterion, and until further evidence is forthcoming it seems sensible to consider this result a random effect of item sampling, and regard the combined test as the best estimate of passage comprehension. Hence the correlations with the corrected combined scores are the best available indicators of criterion validity. These range from a low of .75 in the case of Rater A to a maximum of .84 in the case of Rater D. This gives a shared variance between the translation tests and the criterion of between 56% and 70%, and an average shared variance between all the raters and the criterion of 65%. This figure again falls within the range of what is often considered acceptable in our profession.

3 *Discussion*

This study was designed to give an idea of how translation from L2 to L1 works in Japanese university entrance examinations. The first point to note is the obvious one that raters tend to differ in the severity of their ratings. However, the correlations between ratings are quite high, suggesting that the raters are applying similar criteria in deciding how they rank the students' translations. Considering

Table 8 Correlation of seven ratings with criterion measures

	Uncorrected			Correct for attenuation in criterion measure		
	Cloze	MC	Comb. test	Cloze	MC	Comb. test
Rater A	.69	.57	.69	.78	.68	.75
Rater B	.68	.63	.72	.78	.75	.78
Rater C	.76	.63	.77	.86	.76	.83
Rater D	.74	.66	.77	.84	.79	.84
Rater E	.74	.63	.75	.84	.76	.82
Rater F	.72	.65	.75	.82	.78	.81
Rater G	.70	.63	.73	.79	.75	.79
(N=121)						

Note: MC is multiple choice

that the whole test was administered in 20 minutes the reliability of one rater on the total test of .86 seems very satisfactory. Furthermore, each of the three separate passages consists of only two sentences, yet the reliabilities for each passage are .73, .76 and .86 which seem remarkably high for a short test of about seven minutes duration.

Indications of validity are also surprisingly good. It was confidently expected that the younger raters would base their ratings on the content of the translations whereas the older more traditional raters would tend to base their ratings on the grammatical form of the translations. Yet the factor analysis shows only one factor, despite the fact that the raters were chosen to maximize these differences in attitude towards translation. Indications of validity as a predictor of passage comprehension are also higher than expected. Correlations between the total test and the combined cloze and multiple-choice test, corrected for attenuation in the criterion, range from .75 to .84, which seems to suggest that the translations are not measuring anything so very different from more commonly used measures of passage comprehension.

II Study 2

1 Method

The second study was a multitrait-multimethod construct validation study (Campbell and Fiske, 1959) designed to explore whether listening comprehension and reading comprehension are two separate traits (Buck, 1990). Each of the two traits was examined by means of four methods: short-answer comprehension questions, multiple-choice comprehension questions, gap-filling, and translation. It is possible to use the data from this study to examine the construct validity of the translation tests, to look at the strength of the test-method effect, and compare this with other commonly used testing methods.

Four listening tests and four reading tests were constructed using each of the four test methods above. The tests were as follows:

Method 1: Short-answer questions asked for reproduction of clearly stated information from announcements, public guides and such like. The listening test (SL) had 35 items and the reading test (SR) 34 items.

Method 2: Multiple choice questions on short expository passages.

Listening (ML) had 12 passages with 44 items, and reading (MR) had 16 passages with 33 items.

Method 3: Gap-filling task on a Japanese summary of a short narrative. Listening (GL) had two passages and 37 items, and reading (GR) two passages and 33 items.

Method 4: Translation task on descriptive passages. These are the two tests of interest and are given in Appendix D and Appendix E. Instead of using a rating scale as in Study 1 a more analytic method of marking was devised. Each item of information expressed in the English text was identified and one mark was awarded for each of these successfully expressed in Japanese. The listening test (TL) was a description of Canada divided into 20 short sections. The recording was played with the pauses, during which the listeners had to write a translation of the section they had just heard. Each section carried one, two or three marks. The reading test (TR) consisted of four short passages, taken from British tourist brochures. In order to facilitate comparison with the listening tests, for marking purposes these passages were divided into 13 short sections, each being worth between three and nine marks.

In all the tests instructions and questions were in the L1 to avoid confounding L2 listening scores with L2 reading ability. All the listening passages were heard only once, including the listening translation test. The eight tests were administered to over 400 college students in and around Osaka, Japan. A correlation matrix was calculated for those 353 students who took all eight tests.

2 Results

Descriptive statistics for all the eight tests are given in Table 9. The low means suggest that all the tests were 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.

Reliability: All the listening tests consisted of one or more items followed by short pauses sufficient to allow time for answering. This means that even though the tests were rather difficult for this population, and no doubt many testees would have needed extra time to answer individual items, this did not result in testees getting part way through the tests and then stopping, leaving a block of items unanswered at the end of any of the tests. The listening tests could not have functioned as 'speeded tests', therefore it is quite reasonable to use measures of internal consistency as estimates of

reliability. Cronbach's Alpha is given for the listening tests in Table 9. These range from a high of .84 to a low of .71. The translation listening test (TL) is the second highest with a respectable .83.

However, this was not the case with the reading tests. Because they were too difficult for the target population, many students completed only part of the tests before running out of time, thus leaving a block of items unanswered at the end of each reading test. In such a case measures of internal consistency, for example Cronbach's Alpha, are not a suitable estimate of reliability. A small test/retest study was carried on the four reading tests to get a more acceptable estimate of reliability. Each test was completed by between 50 and 60 college students as similar as possible to those who took the original tests. Results are given in Table 9. The test with the highest estimate of test/retest reliability is the translation reading test (TR) with .90, the lowest is the multiple choice test (MR) with .60 (there is some reason to suppose that .72 may be a better estimate).

The two translation tests were all marked by the same rater. In order to estimate intra-rater reliability 82 listening papers and 82 reading papers were rerated a few months after the first rating, and both intra-rater correlations and the ANOVA estimates as recommended by Krzanowski and Woods (1984) were calculated. These are given in Table 9. The intra-rater correlations are .98 for the listening translation and .92 for the reading, and reliability using ANOVA was .98 and .90 for listening and reading respectively. Considering that there were a number of months between the first rating and the second and consequently the rater had to almost relearn the rating system, these figures are very satisfactory. Indeed the listening reliabilities are extremely high. Not only does the correlation show that the rater ranked the students in very much the

Table 9 Descriptive statistics and reliabilities for eight measures of listening and reading

	N	Item	Mean	SD	Alpha	Retest	Corr.	ANOVA
SL	414	35	14.34	5.49	.84			
SR	423	34	11.01	3.65		.80		
GL	414	37	11.61	5.01	.80			
GR	413	33	16.40	6.09		.83		
TL	414	20	9.27	6.82	.83		.98	.98
TR	409	13	21.11	7.96		.90	.92	.90
ML	414	44	22.01	5.12	.71			
MR	419	33	11.89	3.94		.60*		
(N=353)								

Note: *Two students with very low scores had strangely high scores in the retest, and perhaps should be excluded. Reliability excluding these two cases is .72.

same order, but the ANOVA also indicates that there was virtually no difference in mean scores between the first and second rating.

Validity: Table 10 gives the full MTMM matrix for the eight tests. Such a matrix is usually used to examine the construct validity of the traits included in the matrix. The trait effect and method effect are estimated and compared, and if the trait effect is significantly stronger then it is assumed that the tests have construct validity. According to Campbell and Fiske (1959) construct validity consists of two different types of validity. First, convergent validity, which requires that correlations between two tests which measure the same trait using different methods should be significant and high enough to encourage further investigation. Secondly, discriminant validity, which basically requires that the correlations between two tests which measure the same trait using different methods are higher than correlations between two tests of different traits which happen to use the same method, or are higher than correlations between two tests which have neither trait nor method in common.

This same reasoning can be used to examine the test method effect. In order to examine this it is convenient to identify different types of correlations:

1) Correlations between two measures of a trait when one of these measures is using the method under consideration and the other measure is using another method. These correlations are labelled C1.
2) Correlations between two measures of a trait when neither of these uses the method under consideration. These are labelled C2.

Table 10 MTMM matrix for listening comprehension and reading comprehension as measured by four methods

	Short answer		Fill gaps		Translation		Multiple Choice	
	SL	SR	GL	GR	TL	TR	ML	MR
SL	(.84)							
SR	.523	(.80)						
GL	.612C2	.566	(.80)					
GR	.513	.620C2	.562	(.83)				
TL	.720C1	.494D1	.591C1	.513D1	(.83)			
TR	.339D1	.542C1	.371D1	.575C1	.382D2	(.90)		
ML	.662C2	.584	.624C2	.575	.683C1	.454D1	(.71)	
MR	.321	.474C2	.446	.572C2	.393D1	.480C1	.459	(.60)
(N=353)								

Note: Figures in brackets are reliabilities.

3) Correlations between two measures when one uses the method under consideration and the other has neither trait nor method in common with the first. These are labelled D1.
4) Correlations between two measures of different traits both of which use the method under consideration. These are labelled D2.

The respective correlations are labelled in Table 10. In order to establish that the method under consideration does provide a valid measure of the trait it is supposed to be measuring the following criteria would need to be met:

Criterion 1: Correlations between two measures of a trait when one of these measures is using the method under consideration and the other measure is using another method should be significant and high enough to encourage further examination of the matrix. That is $C1 > 0$.

Criterion 2: Correlations between two measures of a trait when one of these measures is using the method under consideration and the other measure is using another method should be of the same magnitude as correlations between two measures of the same trait when neither of these uses the method under consideration. That is $C1 = C2$.

Criterion 3: Correlations between two measures of the same trait when one of these measures is using the method under consideration should be higher than correlations between two measures where one uses the method under consideration and the other has neither trait nor method in common with the first. That is $C1 > D1$.

Criterion 4: Correlations between two measures of the same trait when one of these measures is using the method under consideration should be higher than correlations between two measures of different traits which use the method under consideration. That is $C1 > D2$.

First, looking at the translation listening test:

Criterion 1: The C1 correlations (.720, .591, and .683) are all significant and indicate reasonable convergent validity.

Criterion 2: The C1 correlations (.720, .591, and .683) are of a similar magnitude to the C2 correlations (.612, .662 and .624) suggesting that the translation listening test has convergent validity as good as the other test methods used in the matrix.

Criterion 3: The C1 correlations (.720, .591, and .683) are higher than the D1 correlations (.494, .513, and .393), indicating that the listening translation test meets the first criterion of discriminant validity.

Criterion 4: The C1 correlations (.720, .591, and .683) are higher than the D2 correlation (.382) indicating that the translation listening test meets the second requirement of discriminant validity.

Secondly, looking at the translation reading test:

Criterion 1: The C1 correlations (.542, .575, and .480) are all significant and indicate moderate convergent validity.

Criterion 2: The C1 correlations (.542, .575, and .480) are of a similar magnitude to the C2 correlations (.620, .474, and .572). Although they are a little lower, the difference is not so great and probably suggests that the translation reading test has convergent validity as good as the other test methods used in the matrix.

Criterion 3: The C1 correlations (.542, .575, and .480) are higher than the D1 correlations (.339, .371, and .454), indicating that the reading translation test meets the first criterion of discriminant validity.

Criterion 4: The C1 correlations (.542, .575, and .480) are higher than the D2 correlation .382, indicating that the reading translation test meets the second criterion of discriminant validity.

Both the translation tests meet all the criteria established. This suggests that these tests, at worst, have no more test method effect than the other test methods used in the matrix. In order to better understand this the overall test method effect for all four methods in the complete matrix can be estimated by means of ANOVA (Boruch and Wolins, 1970; Kalleberg and Kluegel, 1975).

Table 11 gives the ANOVA for the complete MTMM matrix which shows the overall trait and method effects. The *F*-statistic indicates a strong trait effect, with no significant method variance. This suggests that none of the four testing methods used produced any significant test method effect.

3 *Discussion*

The purpose of this study was to compare the efficacy of translation

Table 11 ANOVA of MTMM correlation matrix for eight tests of listening and reading comprehension

Source	DF	Sum of squares	Mean square	*F*-statistic	Variance
Subjects	352	1645.837	4.676	12.076	.536
S × trait	352	323.066	.918	2.370*	.133
S × method	1056	446.192	.423	1.090	.018
Error	1056	408.915	.387	.387	

Note: *$p < .01$

as a testing method with other more widely accepted testing methods. The first thing to note is that the translation tests showed themselves as reliable as the other testing methods. Three different estimates of reliability were calculated for each of the translation methods and all of these indicated quite satisfactory reliability.

As for validity, analysis of the MTMM matrix showed the translation tests have good convergent and discriminant validity, suggesting that they have construct validity as tests of listening and reading comprehension. Test method effect was minimal. All indications are that these two translation tests are providing as good a measure as any of the testing methods.

III General discussion

The two studies reported here examined three different translation tests, two of reading and one of listening. Two different rating systems were examined, holistic rating and analytical rating, and both inter- and intra-rater comparisons were made. Translation as a testing method was examined both in a situation which attempted to replicate real world use, and in a far more tightly controlled research study. The reliability of these tests was examined in a number of different situations by a number of different methods. In all cases the translation tests were found to have satisfactory reliability, which was generally as high as, or higher than, other methods when comparisons could be made. Similarly the validity of translation tests was examined from a number of different aspects. In all cases indications are that they seemed to be providing reasonably valid measures of the traits they were intended to measure. Despite the common belief that translation is a specialized skill, there was no sign of a strong translation method effect; indeed the translation tests seemed to have slightly less method effect than some of the other methods used.

Klein-Braley examined L1/L2 translation, and suggested that one of the reasons for rejecting this as a testing method is the fact that it

is not clear exactly what translation measures; although she did suggest that the better tests among those she examined seemed to be measuring something like general language proficiency. The two studies here offer indications of what L2/L1 translation tests are measuring. In the case of Study 1 the translation tests were designed to measure reading passage comprehension, and they did in fact seem to be measuring that. In Study 2 one test was designed to measure reading comprehension and the other listening comprehension. In both cases analysis indicated they were measuring what they were constructed to measure. It does seem very much as though L2/L1 translation measures comprehension: certainly comprehension of the part translated, but also, seemingly, sample sentences from a passage give a good indication of total passage comprehension.

Klein-Braley's concern that translation may not function very well in the real world where no attempt is made to standardize ratings and calculate inter-rater reliability would seem well founded. However, Study 1 examined this issue, by not giving the raters any guidelines at all, and indicated that raters were essentially using similar criteria, although some were applying them a little more strictly than others. It would probably take only a small amount of consultation between the seven raters to achieve very high standards of inter-rater reliability even when calculated by the more rigorous ANOVA method. Thus, in the 'real world' of Study 1 they did work rather well.

This brings us to the important question whether these results are generalizable to a wider range of tests and testing situations? Given the centralized control of Japanese education and general homogeneity of Japanese society, it does seem possible that Japanese English teachers may share a common view of what constitutes good English performance, or at least more so than many other groups of teachers. Thus, there may be some doubt whether the inter-rater reliability data from Study 1 is generalizable outside Japan. However, the teachers chosen for the ratings were deliberately chosen with a view to representing widely different views of language teaching and it was confidently expected that there would be considerable disagreement between their ratings of what constituted a good translation. Given this, it is difficult to believe that they tended to agree because they shared a common, if perhaps idiosyncratic, view of what constitutes good English. This is further borne out by the validity data from Study 1, which suggests that all the raters were rating on the basis of similar criteria that are not very different from what is measured by tests of reading comprehension such as cloze and comprehension questions.

However, while there may be some grounds for doubting the generalizability of the findings of Study 1, the fact is that the translation tests from Study 2, a more rigorously controlled study, less dependent on the peculiarities of the Japanese situation, also appear to be functioning as reliable and valid measures of the constructs they were designed to measure. Furthermore, this was despite the fact that one of these tests, the listening translation, was an experimental format which has been very little reported in the literature. Taken together the two studies suggest there are good grounds for considering the results to be generalizable to a wider population, and a greater degree of variation in task and format.

This conclusion is in many ways very disturbing, insofar as translation tends to be associated with an attitude towards language teaching that stresses form rather than content. It is especially associated with the grammar-translation method of language teaching, and has traditionally been the preferred testing method of those teachers who felt that students ought rightfully to study a second language by pouring over literary texts with a dictionary and a grammar book. Over the last two decades or so a great deal of effort has been expended in the second language teaching profession to convince such teachers and administrators that there is a better way. Not only that, but in many cases test-makers have deliberately used the very powerful washback effect of their tests on classroom practice to try to influence teaching methods for the better. Replacing older-style tests with tests designed to measure communicative competence is one of the most powerful ways of forcing the reluctant and conservative to shift to more communicative teaching methods, and this use of tests as a force for positive change is something I approve of and am actively engaged in.

As part of this process of using the washback effect as an agent of change there has been a general discrediting of translation as a testing method. Indeed, this research was undertaken with the express purpose of finding evidence to convince colleagues that translation was not a useful testing method and ought to be abandoned for more communicative methods. As the results of this research have become clearer I have noted with serious concern the evident glee of those who would be only too happy to return to those good old days of grammar-translation. A few words of warning are perhaps in order.

First, this research only reports on three tests. Far more research is needed before we have a sufficiently clear picture of how translation functions as a test method. Secondly, even if further research confirms these conclusions, it will only show that translation can be a good testing method, not that it automatically is. There

is no such thing as a test method which automatically produces reliable and valid tests, nor is there ever likely to be one. Each new test, or each new use of an old test, needs to be validated anew, and that naturally includes estimation of reliability. Thirdly, translation tests will only be as good as the scheme devised to rate them. Despite the results of Study 1, test makers would be well advised to ensure that the rating scale is based on a clear understanding of what the test is intended to measure; and efforts should be made to ensure that ratings are consistent between raters and over time. Fourthly, it should be stressed that there is not enough evidence here to reject any other testing method and replace it with translation.

Finally, and perhaps most importantly, apart from the obvious responsibility to make tests on which reliable and valid decisions can be based, test makers in education have another important responsibility; namely to make tests which have a positive influence on classroom practice. The washback effect of the test on the classroom should be such that as teachers and students concentrate on preparing for the test, as they invariably will, the activities they are led to perform are educationally beneficial in their own right. It seems likely that translation tests could have very negative washback indeed, and lead to activities which would not be beneficial to second language learners.

IV Conclusion

The results of these two studies suggest the surprising, and very unfashionable, conclusion that translation from L2 to L1 may be as effective a method of testing L2 comprehension as other more academically acceptable second language testing methods. While far more research is obviously needed before any firm conclusions can be drawn, it does seem that perhaps the general rejection of translation as a testing method by many practitioners in our field may have been too premature.

V References

Boruch, R.F. and **Wolins, L.** 1970: A procedure for estimation of trait, method, and error variance attributable to a measure. *Educational and Psychological Measurement* 30, 547–74.

Buck, G. 1990: The testing of second language listening comprehension. PhD. dissertation, University of Lancaster.

Campbell, D.T. and **Fiske, D.W.** 1959: Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin* 56, 81–105.

Guilford, J.P. and **Fruchter, B.** 1978: *Fundamental statistics in psychology and education*. Tokyo: McGraw-Hill.
Kalleberg, A.L. and **Kluegel, J.R.** 1975: Analysis of the multitrait-multimethod matrix: some limitations and an alternative. *Journal of Applied Psychology* 60, 1–9.
Klein-Braley, C. 1982: Die Ubersetzung als Testverfahren in der Staatprufung fur Lehramtskandidaten. *Neusprachliche Mitteilungen* 2, 94–97.
—— 1987: *Fossil at large: translation as a language testing procedure*. Colloquium on Language Testing Research. Reading: 6–9 April.
Krzanowski, W.J. and **Woods, A.J.** 1984: Statistical aspects of reliability in language testing. *Language Testing* 1, 1–20.
Oakeshott-Taylor, A. 1981: Eignet sich die Ubersetzung als Fremdsprachenklauser? *AKS-Rundbrief* 3, 26–36.
Obunsha 1986: *Zenkoku TankiDaigaku NyushiMondai Seikai, 61 Nen: Eigo, Sugaku, Kokugo*. Tokyo: Obunsha.

Appendix A: Study 1: translation test

The parts in italics were the sentences for translation into Japanese.

Passage 1

Anyone who claims more than he is already receiving is very likely to get nothing at all in the future.

Young Alan had a very generous uncle who gave him five shillings every time he came to tea.[1] Alan wanted a bicycle, so the next time his uncle called he asked him for ten pounds instead of five shillings.

"Ten pounds?" exclaimed his uncle.

"Well, you can afford it, can't you?" demanded Alan rudely.[2]

This annoyed his uncle so much that Alan did not get his bicycle – or any more five-shilling tips.

Passage 2

Although only one percent of the Japanese people are Christians, *it's estimated that more Christmas cakes will be sold here on Dec. 24 than in any other country, perhaps twice as many as in the United States.*[3] This is partly because most Japanese think Christmas is largely confined to the singing of Yuletide songs and the eating of Christmas cakes by children. But in Christian countries what's more important, at least to the children, is opening presents on Christmas morning. *In Japan the custom is for children to receive what is called 'otoshidama' or money on New Year's Day*,[4] so except for the cake and songs, Christmas has no special meaning to them, except in Christian families.

Passage 3

I became good friends with Mrs. Okumura from the first, but in my associations with other people I sometimes encountered awkward situations, *not so much because of difficulties with the Japanese language as because of a difference in social customs.*[5] For example, while I was living in England no one had ever asked me questions about my family, or about anything of a personal nature, but Japanese acquaintances freely asked questions of the most personal kind. *Having become accustomed to English ways, at first I was surprised at such questions as, "Are you married? No? Why not?"*[6] Americans are more apt than Englishmen to ask personal questions, but one question which an American never asks an acquaintance is his salary. In Japan, however, I was often asked how much money I earned, to my confusion. People used to ask my age, and would accept any figure I gave.

Appendix B: Study 1: cloze test

Passage 1

Anyone who claims more than he is already receiving is very likely to get nothing at all in the future.

Young Alan had a very generous (1) _____ who gave him five shillings every (2) _____ he came to tea. Alan wanted (3) _____ bicycle, so the next time (4) _____ uncle called he asked him for ten pounds (5) _____ of five shillings.

"Ten pounds?" (6) _____ his uncle.

"Well, you can afford (7) _____, can't you?" demanded Alan rudely.

This (8) _____ his uncle so much that Alan (9) _____ not get his bicycle – or any (10) _____ five-shilling tips.

Passage 2

Although only one percent of the Japanese people are Christians, it's estimated that more Christmas cakes will be sold here on Dec. 24 than in any other country, perhaps twice as many as in the United States. This is partly because most (11) _____ think Christmas is largely confined to the (12) _____ of Yuletide songs and the eating of Christmas (13) _____ by children. But in Christian countries what's more (14) _____, at least to the children, is (15) _____ presents on Christmas morning. In Japan the (16) _____ is for

children to receive what is (17) _____ 'otoshidama' or money on New Year's Day, (18) _____ except for the cake and songs, Christmas has (19) _____ special meaning to them, (20) _____ in Christian families.

Passage 3

I became good friends with Mrs. Okumura from the first, but in my associations with other (21) _____ I sometimes encountered awkward situations, not so (22) _____ because of difficulties with the Japanese language (23) _____ because of a difference in social customs. (24) _____ example, while I was living in England (25) _____ one had ever asked me questions about (26) _____ family, or about anything of a personal (27) _____, but Japanese acquaintances freely asked questions of (28) _____ most personal kind. Having become accustomed to (29) _____ ways, at first I was surprised at (30) _____ questions as, "Are you married? No? Why (31) _____?" Americans are more apt than Englishmen to (32) _____ personal questions, but one question which an (33) _____ never asks an acquaintance is his salary. (34) _____ Japan, however, I was often asked how (35) _____ money I earned, to my confusion. People (36) _____ to ask my age, and would accept any figure I gave.

Appendix C: Study 1: multiple choice comprehension test

This is a translation of the multiple choice comprehension test used in Study 1, which was administered in Japanese. Note that the words in inverted commas were in English in the original test, and thus some questions no longer make sense when translated into English.

Passage 1

Q1. What does 'generous' mean on line three?
a) generous
b) strict
c) irritable
d) cunning

Q2. Who is referred to by 'him' on line three?
a) his uncle
b) anyone
c) young people
d) Alan

Q3. When did Alan receive money from his uncle?
a) whenever he was a good boy
b) whenever he made tea for his uncle
c) whenever his uncle came for tea
d) whenever Alan went to buy some tea

Q4. Alan asked his uncle to give him ten pounds
a) when his uncle came to visit next?
b) when Alan talked to his uncle next?
c) when his Uncle told Alan he would buy him a bicycle?
d) when his uncle wanted to buy a bicycle?

Q5. What does 'annoyed' mean on line eight?
a) amused
b) sad
c) surprised
d) angry

Q6. What does 'this' on line eight refer to?
a) the price of a bicycle
b) the fact that Alan told his uncle that he (the uncle) has enough money
c) the fact that his Uncle was asked to buy a bicycle
d) the fact that Alan received five-shillings

Q7. What kind of child is Alan?
a) greedy
b) introverted
c) happy
d) polite

Q8. The writer wanted to suggest
a) uncles have a duty to buy presents for the children of their relatives?
b) when you ask for something you should ask politely?
c) people should be satisfied with what they've got?
d) nowadays all children receive presents?

Passage 2

Q9. What does 'estimated' on line two mean?
a) proved
b) recognized
c) estimated
d) altered

Q10. How many Christmas cakes were sold in Japan?
a) half as many as in the United States
b) enough for 1% of the Japanese population
c) more than any other country where 1% of the population is Christian
d) more than any other country

Q11. What does 'confined' mean on line four?
a) included
b) means
c) decided
d) limited

Q12. The most important thing about Christmas for children in Christian countries is
a) to eat Christmas cake?
b) neither eating cakes nor receiving presents?
c) to sing Christmas songs?
d) to get presents on Christmas day?

Q13. What does 'them' on line nine refer to?
a) 'otoshidama'
b) Christmas cakes and Christmas songs
c) families of Christians
d) Japanese children

Q14. What is the main point of the author?
a) the Japanese imitated Christmas from the West
b) the Japanese understand the real meaning of Christmas
c) the Japanese have not grasped one important aspect of Christmas
d) Japanese Christmas is modelled on New Year celebrations

Passage 3

Q15. What's the meaning of 'situation' in line two?
a) perplexity
b) a meeting
c) experience
d) situation

Q16. The writer has often been in an awkward situation because
a) Japanese is difficult?
b) customs are different?
c) he knows a lot of people?
d) he befriended Mrs. Okumura?

Q17. While the writer was living in England
a) his Japanese friends cared for his family?
b) he was not asked any personal questions?
c) Americans wanted to know whether the writer was married or not?
d) people often asked about his family?

Q18. What does 'personal' on line six mean?
a) private
b) embarrassing
c) complicated
d) objective

Q19. What's the meaning of 'acquaintances' on line six?
a) close friends
b) strangers
c) acquaintances
d) seniors

Q20. The writer gives the impression that the Japanese people whom he met
a) did not try to guess his age?
b) often tried to guess his age?
c) guessed how old he was?
d) couldn't guess correctly how old he was?

Q21. The writer thinks that
a) Americans never ask their friends or acquaintances about their salaries?
b) Americans ask more private questions than Japanese?
c) Japanese never ask their friends or acquaintances about their salaries?
d) The English ask more personal and private questions than Americans?

Q22. The writer is probably
a) English?
b) Japanese?
c) European?
d) American?

Q23. The main point of this piece is
a) differences of customs between different cultures?
b) correct manners?
c) customs of marriage?
d) friendship with Mrs. Okumura?

Appendix D: Study 2: listening translation test – text

The test consists of a Canadian telling a prospective Japanese visitor about Canada. The testee knows this situation, and is expecting to hear a description of Canada. The text was spontaneous and unscripted. It was later divided into 20 parts, with a break between each part long enough for the testee to write down in Japanese what has been heard in English. These are given below, numbered in order.

1) the most striking feature is its size
2) it's so large unbelievably large
3) especially if you come from a small country
4) so if you want to come to Canada for a vacation
5) you have to consider very carefully just a small part of Canada to visit
6) even for a week you must choose a small area otherwise it's impossible to see anything
7) Canada has different regions
8) starting from the west there's British Columbia British Columbia is a little like Canada's California
9) the city of Vancouver is a nice city
10) it's located beside a beautiful sea coast
11) right behind the city are beautiful mountains
12) it's perhaps Canada's nicest city perhaps
13) as you go across the Rocky Mountains westwards you come to a large flat area the prairies
14) for most Canadian who don't live there it's considered boring
15) the summer times are dreadfully hot and the winters are freezing cold
16) there are no hills almost no trees just huge huge farms
17) boring I don't recommend it
18) as you come east you come to a huge province called Ontario
19) the western half of this province a very very large area is largely only trees
20) trees lakes rivers rocks and almost no people

Appendix E: Study 2: reading translation test – text

The test consists of four short reading passages taken from British tourist guides. They were presented to the students as complete passages, but in order to standardize marking with the listening test, they were broken into a number of short sections for rating purpose. The breaks between these sections are marked with a slash.

Knoll Hotel

We are quietly situated in our own grounds / with magnificent views overlooking Lake Windermere and mountains./ We offer comfort and warmth, good home cooking, cosy bar and ample parking./

Queens Hotel

Right in the centre of Ambleside, the Queens Hotel makes an ideal base for discovering the charms of Lakeland./ Good food, carefully prepared from the best ingredients and served by friendly staff./

Calderdale

Once the home of the domestic cloth industry, Calderdale is now one of the most fascinating parts of Yorkshire to visit./ The dramatic scenery, rich heritage and warm Yorkshire welcome extended to all guarantee a good day out, short break or holiday for visitors from all over the world./

Ryedale

Ryedale is the largest of 8 Districts in the Country of North Yorkshire / and comprises some of the most beautiful country in England./ In an area of over 600 square miles the visitor can see contrasting landscapes / from the urbanized villages near York to the remote unrestricted views across the North York Moors in the north./ Between these points lies the central agricultural belt stretching to within 5 miles of Scarborough in the east, and to Sutton Bank top, 10 miles from Thirsk in the west./ Within its boundaries are over 150 towns and villages abounding in history. Most have a Church and an Inn./