Evaluating Six Measures of EFL Learners' Pragmatic Competence

Ken Enochs
International Christian University

Sonia Yoshitake-Strain
Seigakuin University

This study examines the reliability, validity, and practicality of six measures of cross-cultural pragmatic competence. The multi-test framework used here was developed by Hudson, Detmer, and Brown at the University of Hawaii and consists of six tests which focus on the students' ability to appropriately produce the speech acts of requests, apologies, and refusals in situations involving varying degrees of relative power, social distance, and imposition. These measures have previously been tested on native Japanese learners of English in an ESL context (Hudson et al., 1992, 1995) and on learners of Japanese in a JSL context (Yamashita, 1996). The current study administered these tests to native Japanese learners in an EFL context. Four of the tests proved highly reliable and valid and two of the tests less so. Furthermore, the tests clearly differentiated those students who had a substantial amount of overseas experience from those who had not, a distinction not shown by the students' TOEFL scores.

本論では、6種類のプラグマティック能力測定テストの妥当性・信頼性・実用性を検討する。この6種類のテストは、外国語学習者が習得している、様々な状況下での適切な依頼・謝罪・断わり表現の生成能力を測定するため、ハワイ大学のHudson, Detmer, and Brownによって開発されたものであるが、これまでにも日本人英語学習者(ESL)や日本語学習者を対象にした研究がなされている。

本研究では、この6種類のテストを日本人英語学習者(EFL)を対象に測定をおこなった。その結果、6種類のうち4種類のテストにおいて高い妥当性と信頼性を確証した。さらに、これらのテストでは、TOEFLでは判別できなかった海外滞在経験の有無を明確に判別することが可能であった。

The notion that language competence involves the ability to produce language that is not only grammatically correct but also appropriate for particular situations has been fundamental to language learning pedagogy and research for decades. According to Mundby (1978), "to communicate effectively, a speaker must know not only how to produce any and all grammatical utterances of a language, but also how to use them appropriately. The speaker must know what to say, with whom, and when and where" (p. 17). A number of linguists over the years (Hymes, 1972; Canale & Swain, 1980; Canale, 1988; Bachman, 1990; etc.) have used the term *communicative competence* to account for the contextual and socio-cultural knowledge that is necessary to use language in real-life situations. Bachman (1990) has suggested that *communicative competence* consists of two interactive components: *organizational competence* to account for grammatical knowledge, and *pragmatic competence* to account for the "capacity for implementing, or executing [organizational] competence in appropriate, contextualized communicative language use" (p. 84).

Deficiencies in pragmatic competence result in what is commonly called pragmatic failure. Thomas (1983) has broadly defined pragmatic failure as occurring "on any occasion the speaker's utterance is perceived by a hearer as different than what the speaker intended should be perceived" (as cited in Hudson, Detmer & Brown, 1992, p. 5). A great deal of research has been directed at defining the causes of pragmatic failure, much of it focused on the inappropriate realization of speech acts. Speech acts are defined as "not an 'act of speech' . . . but a communicative activity . . . defined with reference to the intentions of speakers while speaking and the effects they achieve on listeners" (Crystal, 1991, p. 383).

Three such speech acts that involve very different strategies depending on the culture are requests, refusals, and apologies (Beebe & Takahashi, 1989; Beebe, Takahashi, & Uliss-Weltz, 1990). Furthermore, Hudson et al. (1992, 1995) claim there are different perceptions between speakers of different cultures regarding variables such as *relative power*, *social distance*, and *degree of imposition*. Relative power has to do with the extent to which the speaker's will can be imposed on the hearer. An employer, for example, would have +power over an employee, whereas an employee would have –power with an employer. Social distance refers to the degree of familiarity between the speaker and hearer. For example, speaking with a stranger would involve +distance, whereas speaking with a housemate or co-worker would involve –distance. Finally, the degree of imposition is the right and extent to which the speaker imposes on the hearer. As examples, asking to borrow a dictionary involves –imposition, while asking someone to spend a Saturday helping one to move would involve +imposition.

These three variables, relative power, social distance, and degree of imposition, are considered to be especially significant because "within the research on cross-cultural pragmatics, they are identified as the three independent and culturally sensitive variables that subsume all other variables and play a principal role in speech act behavior" (Hudson et al., 1995, p. 4). Therefore, situations that combine the speech acts of requests, refusals, and apologies with the variables of power, distance, and imposition provide learners with a rich array of pragmatic challenges.

In an effort to determine how pragmatic competence might best be assessed, Hudson et al. (1992) produced six different tests of varying type and method, each involving situations that combine the speech acts of requests, refusals, and apologies with the socio-cultural variables of power, distance, and imposition. They administered these tests to native Japanese students studying English in an ESL context and reported their results in *Developing Prototypic Measures of Cross-Cultural Pragmatics* (1995). Additionally, Yamashita (1996) administered these same tests (translated into Japanese) to a group of second-language learners of Japanese in a JSL context. The current study administered these tests to Japanese students in an EFL context for the purpose of analyzing the results both qualitatively and quantitatively. Yoshitake-Strain concentrated on qualitative analysis and reported her findings in her Ph.D. dissertation, *Interlanguage Competence of Japanese Students of English: A Multi-test Framework Evaluation* (1997), and the present researchers have recently published a preliminary statistical analysis (Enochs & Yoshitake, 1996) on the use of the self-assessment and role play tests in assessing pragmatic competence. The purpose of this investigation is to report on a statistical analysis of the reliability, validity, and practicality of all six tests. The following research questions were addressed:

Research Question 1. How reliable are these test formats for measuring Japanese EFL students' pragmatic competence? Reliability will be determined using internal consistency estimates, measures of inter-rater reliability, and the standard error of measurement (SEM).

Research Question 2. How valid are these test formats? Validity will be determined in terms of content, criterion-related, and construct validity.

Research Question 3. How practical are these test formats?

Method

Participants

The participants in this study were 25 first-year students in the English Language Program (ELP) at International Christian University (ICU) in Tokyo, where both authors were working at the time the data were collected. Most of the students were non-English majors, and all were volunteers who participated in the study during their out-of-class free time. There were seven male and 18 female students, with ages ranging from 18-20, and one 26-year old. The students had started the program in April and were tested in October, having completed the spring term and several weeks of the fall term prior to the test. During both terms, the students' English-language study consisted of approximately nine 70-minute classes per week in a content-based curriculum focused on developing the students' ability in academic English. The students tested were considered to be "average" within the context of the ELP, since they were drawn from the middle of the three placement levels in the program. The TOEFL scores for these students ranged from 423-577 points, with most of the students falling in the 500-539 range. The scores were obtained upon entrance into the university in April.

The overseas experience of the students varied, with many having recently returned from six-week academic English programs at universities in English-speaking countries as part of ICU's Summer English Abroad (SEA) Program. The distribution of the students' overseas experience is broken into three categories (see Table 1). Group 1 had none or very little overseas experience. Those who did have some experience generally gained it through a vacation with their family, which it was reasoned would have had negligible effect on the students' English linguistic and pragmatic competence. The members of Group 2 had spent at least five weeks overseas, generally in homestay situations, and students participating in the SEA Program had been immersed in university summer English-language programs as well. Members of Group 3 had all lived overseas, and were considered to have had a significant amount of exposure to English.

Table 1: Overseas Experience of Subjects

Group	Time overseas	*n*	Comments
1	None or little	8	2 had none, 6 had 2–3 weeks experience, generally in English-speaking countries.
2	5-10 weeks	12	All had experienced some sort of English-language immersion, many through participating in ICU's SEA program.
3	Returnees	5	One to 6.5 years overseas. While only one had lived in an English-speaking country (for 2 years), others had attended international schools in which the language of instruction was mainly English.

Instruments and Administrative Procedure

The six tests administered and evaluated in this study were developed at the Second Language Teaching and Curriculum Center of the University of Hawaii by Hudson, Detmer, and Brown (1992, 1995). These tests were designed as prototypic measures of cross-cultural pragmatic competence. While each of these tests focuses on the three key variables of power, social distance, and degree of imposition in the speech acts of requests, refusals, and apologies, the tests vary in their type and method. The reason for this was to develop "instruments of different types and methods for application across different social variables and speech acts" and reflects the need to determine "the potential differential effectiveness of the instruments" (1995, p. 6). The tests are listed below in the order they were administered in the present study.

1. Self-Assessment Test (SA)
2. Listening Laboratory Production Test (LL)
3. Open Discourse Completion Test (OPDCT)
4. Multiple Choice Discourse Completion Test (MCDCT)
5. Role-play Self-Assessment Test (RPSA)
6. Role-play Test (RP)

For all of these tests, Hudson et al. designed a framework which would evenly distribute various combinations of the attributes they wished to measure. With three different speech acts and eight different combinations of power, distance, and imposition, 24 cells were necessary to represent all combinations of these attributes. These various combinations were randomly reordered and then consistently applied to various task situations throughout the series of tests (see the table in Hudson et al., 1995, p. 10, which shows how these combinations were distributed in their research using tests with 24 different items).

For the RPSA and RP tests, participants performed one series of eight different role play scenarios in which each scenario contained a request, a refusal, and an apology. The socio-cultural variables, however, were similarly distributed in a random fashion. For all of the tests except for the MCDCT, either students or raters indicated on a five-point Likert scale how well they felt the speech act situations had been performed. Details regarding the administration and specific nature of each of these tests follow. For single-item examples of each of the tests, see the Appendix.

Self-assessment test (SA)

The first test administered of the series, this test provided participants with written descriptions of each of the twenty-four speech act situations. After reading each situation, they indicated on a five-point Likert scale how well they felt they could provide an appropriate response in each of the situations. The Appendix shows an example of an apology situation with –imposition, +power, and –distance.

Listening Laboratory Production Test (LL)

This test provided participants with tape-recorded descriptions of the situations to which they provided oral responses. Each description was given twice, and the participants then recorded what they felt was an appropriate response during a one-minute interval following the second listening. Raters then listened to the responses and evaluated each of them using the same five-point Likert scale. The Appendix shows an example of an apology situation with +imposition, -power, and +distance.

Open Discourse Completion Test (OPDCT)

This test was given as a take-home assignment, which participants were given one week to complete. Each participant signed a written pledge that he or she would not receive any assistance on this test. Here, the 24 descriptions of various speech act situations were provided in written form, and the participants were required to provide an appropriate written response to each situation. Raters read the written responses and evaluated each of them using the same five-point Likert scale. The Appendix shows an example of a request situation with +imposition, -power, and +distance.

Multiple-Choice Discourse Completion Test (MCDCT)

This test was also given as a take-home assignment (and participants were reminded of their pledge not to seek assistance). Again, written descriptions were provided of different situations, but this time the participants could choose an appropriate response from among three multiple-choice possibilities, only one of which would be considered fully appropriate by a native speaker of English. Evaluating this test involved giving five points for each correct response (according to a key provided by the test developers), and zero points for either of the incorrect responses. The Appendix shows an example of a refusal situation with -imposition, -power, and -distance.

Role-Play Self-Assessment Test (RPSA)

This test required students to perform the speech act situations as role plays, with a native speaker of English acting as interlocutor. In this test there are just eight different scenarios, but each includes all three speech acts—a request, a refusal, and an apology—with varying degrees of power, distance, and imposition in each situation to mirror the other tests with 24 separate situations. Written descriptions of the role plays (in both English and Japanese) were given to the participants beforehand so they could have a clear understanding of each situation and of what would be expected of them. These role plays were performed in a studio-like room at ICU and recorded on videotape. Immediately after performing each role play, the participants rated on the same five-point Likert scale how well they felt that they had appropriately responded in these speech act situations. The Appendix shows an example used for both the RPSA and RP tests in which all three speech acts were performed in a situation with -imposition, -power, and +distance.

Role-play test (RP)

Using the videotape recordings of the role plays, raters used the same five-point Likert scale to evaluate the appropriateness of each of the 24 speech acts within the eight role plays.

Statistical Analysis

Each of the tests had 24 different items. All of the tests, with the exception of the MCDCT, used 5-point Likert scales, making a total possible score of 120 points. With the MCDCT, 5 points were given for each right answer so a total possible score for this test was also 120 points. These data were initially entered onto a spreadsheet using Excel 5.0. They were then analyzed using Excel and the statistics program SSPS/PC+ Version 4.0.1. Estimates of reliability were conducted through an analysis of internal consistency, inter-rater reliability, and the standard error of measurement. Validity was analyzed in terms of content, criterion-related, and construct validity. The determination of construct validity was made through a principal components analysis, factor analysis, a multivariate analysis and a univariate follow-up statistic of differential groups.

Inter-rater reliability

Three raters were used for each of the tests that required raters—the LL, OPDCT, and the RP test. These were drawn from a pool of raters made up of colleagues and one spouse, a mix of men and women of approximately the same age and educational background. They consisted of five Americans and one Englishman and were all ESL professionals, with the exception of one of the Americans being a journalist. Training involved first an explanation of the speech acts and variables being examined. Raters were then asked to make holistic evaluations of the appropriateness of the students' responses without regard for grammatical accuracy.

Estimates of the inter-rater reliability were first made using the Pearson product-moment correlation coefficients (Pearson *r*) for different pairings of raters, as can be seen in Table 2.

The highest correlations were clearly between the raters on the RP test, followed by those for the LL test. There was considerably less correlation between the raters on the OPDCT test.

As Brown points out, the number of ratings "can have a dramatic effect on the magnitude of the reliability coefficient" (1996, pp. 203–204). The ratings of the three raters together, then, will tend to be more reliable than a given pair, and "adjusting to find the reliability of larger numbers of ratings taken together would be logical, possible, and advisable" (p. 204). The full tests inter-rater reliability estimates using the Spearman-Brown Prophecy formula[1] can be seen in Table 3. Converted to percentages, the RP test provides an estimated 93% reliability, followed by the LL test at approximately 80%, and the OPDCT test at 49%.

Table 2: Inter-rater Correlation Matrix Using Pearson *r*

LL test

	Rater 1	Rater 2	Rater 3
Rater 1	1.0000		
Rater 2	.6428**	1.0000	
Rater 3	.5350*	.5139*	1.0000

OPDCT

	Rater 1	Rater 2	Rater 3
Rater 1	1.0000		
Rater 2	.2705	1.0000	
Rater 3	.1590	.3012	1.0000

RP test

	Rater 1	Rater 2	Rater 3
Rater 1	1.0000		
Rater 2	.7894**	1.0000	
Rater 3	.8069**	.8413**	1.0000

*p < .01
**p < .001

Table 3: Inter-rater Reliability Using Spearman-Brown

LL	OPDCT	RP
.7957	.4933	.9296

Results and Discussion

Descriptive Statistics

Table 4 shows descriptive statistics including the mean, standard deviation, minimum, maximum, and range of the scores for 25 students. The TOEFL results reveal a mean of 502 points which is somewhat higher than the Japanese national average of 494. The average mean of the TOEFL subtest scores of 49.48 for Listening, 51.28 for Structure, and 50 for Reading are correspondingly higher but basically parallel to the Japanese national average of 49 for Listening, 50 for Structure, and 49 for Reading (Educational Testing Service, 1995).

As for the six tests designed by Hudson et al. and administered to EFL students in the present study, several of the descriptive statistics are worth noting. Of the two discourse-completion tests, the OPDCT had the highest mean score at 92.48, but the lowest standard deviation at 6.70. This contrasts sharply with the MCDCT which had the lowest mean score at 70, but the second to the highest standard deviation at 14.43. Of the two self-assessment tests, it is interesting to note the relatively high mean score of 86.08 for the SA test, which had the highest standard deviation at 14.59 points. In this test, participants *speculated* on the degree to which they could demonstrate pragmatic competence in particular situations. In comparison, the RPSA had a similarly high standard deviation of 14.31, but a considerably lower mean at 78.88. This score reflects how well participants felt they *realized* pragmatic competence in their role play performances. The substantially lower mean for the RPSA suggests that the participants in this study generally did not feel they had performed as well as they thought they could in these situations.

For the RP test, the mean of the raters' scores was identical to that of the RPSA at 78.88 points, but with a considerably lower standard deviation: 10.53 versus 14.31. There was also a significant variation between the raters of the LL test, ranging from a high of 81.6 to a low of 65.2. Of the individual raters' scores for the three tests which required raters, there was, of course, some variation. Rater 3 was the only rater who was not a language teaching professional. One wonders whether teachers

Table 4: A Summary of Descriptive Statistics

Variable	n	Mean	Std Dev.	Mini	Maxi	Range
TOEFL	25	502.48	34.03	423.00	577.00	154.00
LT	25	49.48	3.86	43.00	59.00	16.00
ST	25	51.28	4.74	42.00	64.00	22.00
RD	25	50.00	4.62	38.00	59.00	21.00
SA	25	86.08	14.59	60.00	116.00	56.00
LL	25	77.05	8.49	61.00	97.70	36.70
LL1	25	81.60	10.03	65.00	101.00	36.00
LL2	25	84.36	11.14	63.00	110.00	47.00
LL3	25	65.20	8.98	47.00	84.00	37.00
OPDCT	25	92.48	6.70	77.83	110.90	33.07
OPDCT1	25	91.50	7.95	74.00	107.00	33.00
OPDCT2	25	95.11	7.88	75.00	107.00	32.00
OPDCT3	25	90.84	12.68	76.00	139.90	63.90
MCDCT	25	70.00	14.43	30.00	95.00	65.00
RPSA	25	78.88	14.31	61.00	111.00	50.00
RP	25	78.88	10.53	61.00	102.00	41.00
R1	25	78.60	11.28	60.00	104.00	44.00
R2	25	76.16	8.79	59.00	91.00	32.00
R3	25	81.88	13.66	62.00	112.00	50.00

(LT = Listening; ST = Structure; RD = Reading; SA = Self-Assessment; LL = Average of the three raters' scores for the test; LL1–LL3 = Raters' individual LL scores; OPDCT = Average of the three raters' scores for the Open Discourse Completion Test; OPDCT1–OPDCT3 = Raters' individual OPDCT scores; MCDCT = Multiple-choice Discourse Completion Test; RPSA = Role-play Self Assessment; RP = Average of the three raters' scores for the Role Play test; and R1–R3 = Raters' individual RP scores)

are considerably more tolerant of participants' efforts at appropriateness than non-teachers. Without other non-teacher raters, however, it is difficult to draw such a firm conclusion.

Similarly for the RP test, the rater with the lowest mean, Rater 2, was British, whereas the other two raters were Americans. One wonders whether the British rater tended to rate students lower due to higher expectations of what constitutes appropriate language use, having come from a country noted for its emphasis on politeness. Again, it is impossible to draw such a conclusion with just one rater, but it would be interesting to experiment with a large pool of raters to see if there is quantifiable variation in the way raters from different English-speaking countries (and/or cultural backgrounds) rate students.

Reliability

Internal consistency reliability

Internal consistency[2] reliability was computed by first using the split-half method to determine the correlation between odd- and even-numbered items in the test. The half-test correlation was then adjusted using the Spearman-Brown Prophecy formula to estimate full-test reliability. Table 5 shows the estimated full-test reliability of each of the six tests. The two tests in which students assessed themselves, the SA and RPSA tests, showed particularly high estimates of internal consistency, followed by the LL and RP tests. Both of the discourse completion tests, especially the MCDCT, had considerably less internal consistency.

Table 5: Adjusted Split-Half Internal-Consistency Estimates

SA	LL	OPDCT	MCDCT	RPSA	RP
.9567	.9260	.6711	.5612	.9304	.8636

Standard Error of Measurement

The Standard Error of Measurement (SEM)[3] was computed using the standard deviation estimates from Table 4 and the adjusted split-half values from Table 5. Table 6 shows the SEM for the six tests. As can be seen, the LL test yielded the smallest SEM at 2.3, whereas the MCDCT clearly had the highest at 9.55. The others had respectable estimates of SEM in the 3.0 range.

Table 6: Standard Error of Measurement

	SA	LL	OPDCT	MCDCT	RPSA	RP
SEM:	3.03	2.30	3.84	9.55	3.77	3.88

Validity

Content validity

Since there is no statistical measure of content validity, either the testers themselves, their colleagues, or panels of experts determine the "representativeness and comprehensiveness" of the tests (Hatch & Lazaraton, 1991, p. 540). To ensure content validity, Hudson et al. have created a framework in which the speech acts of requests, apologies, and refusals are systematically matched with the variables of relative power, social distance and degree of imposition. According to Hudson et al., "[t]he designation of these in this way allows an examination of the interaction between sociopragmatic variables and particular speech act realizations. Additionally, this framework allows an examination of each particular variable within each speech act" (1992, p. 16). Furthermore, the role-play situations involve a wide and fairly representative sampling of real-life contexts: interacting with a mechanic at a garage, with a clerk at a store, with a superior in the workplace, with a housemate in a shared house, etc.

Criterion-related validity

Criterion-related validity involves comparing the results of the test or tests being evaluated with some other established measure of proficiency (Brown, 1996, p. 247). We chose the students' TOEFL scores for comparative purposes for a variety of reasons: 1) we had ready access to these students' TOEFL scores since they had taken an institutionally-administered TOEFL examination several months earlier upon entrance into our university; 2) students' TOEFL scores have proven reasonably effective for placement purposes within our own English language program; and 3) TOEFL scores are widely used and accepted as a measure of a student's overall English language proficiency. First, correlation coefficients were determined between the students' TOEFL subtest scores of Listening (LT), Structure (ST), and Reading (RD), and the tests of this study—SA, LL, OPDCT, MCDCT, RPSA, and RP.

These correlations were then squared to find the *coefficient of determination.*[4] The coefficient of determination ascertains the amount of overlapping variance between the tests, in effect revealing which correlations are meaningful. The results of squaring the above values to yield the percentage of overlapping variance between the tests are in Table 7. As can be seen, the only significant amount of overlapping variance is within each set of tests. The greatest amount of overlap is between the ST and RD tests at .359, an overlap of approximately 36%. The next greatest amount of overlap is between the production-based pragmatic tests, especially between that of the LL and OPDCT at approximately 29%, and between the LL and the RP also at nearly 29%. Further overlap can be found between the two self-assessment tests, the SA and RPSA, at approximately 22%. Within each set of tests, then, there is some *meaningful* overlapping variance between certain tests, but essentially no overlapping variance between the set of tests designed by Hudson et al. and the TOEFL subtests. It seems quite clear that these two sets of tests are measuring something very different from one another.

Table 7: Squared Correlation Values to Determine Overlapping Variance

	LT	ST	RD	SA	LL	OPDCT	MCDCT	RPSA	RP
LT	1.000								
ST	.169	1.000							
RD	.014	.359**	1.000						
SA	.000	.002	.003	1.000					
LL	.097	.050	.014	.022	1.000				
OPDCT	.022	.007	.018	.008	.287*	1.000			
MCDCT	.013	.004	.003	.110	.028	.051	1.000		
RPSA	.000	.046	.009	.217*	.001	.114	.050	1.000	
RP	.019	.017	.100	.000	.285*	.156	.001	.050	1.000

*p < .01
**p < .001

Construct validity

Principal component analysis (PCA): A principal component analysis[5] of the TOEFL subtests and the six tests of pragmatic competence by Hudson et al. determined that there are three factors with Eigen values of over 1.0. The largest of these, Factor 1, accounts for approximately 24% of the variance, followed by Factor 2 accounting for approximately 22%, and Factor 3 at approximately 19%. Cumulatively, these factors account for approximately 65% of the variance.

Factor analysis: A factor analysis[6] using a varimax rotated factor matrix was then run in order to determine whether there was a pattern to the factor loadings. As shown below in Table 8, results after a varimax rotation of these factors show a clear pattern of factor loading by test type, with the highest load on three of the tests by Hudson et al., closely followed by the TOEFL subtests, and then by the two self-assessment tests. This strongly suggests that some sort of method effect is at work.

That is, each of these types of tests seem to have factors in common which are not shared by the other tests. What these factors are is not clear, but one can speculate. The LL, OPDCT, and RP tests are similar in that they all employed native speakers of English rating the students' actual production of English: spoken, written and in role-play situations, respectively. The TOEFL subtests share the qualities of being paper and pencil tests that draw upon the students' receptive processes and require as a response the recognition of right answers in a multiple choice format. The SA and RPSA tests both involve the participants evaluating themselves, which is a method quite the opposite from the MCDCT.

Table 8: Factor Analysis

	Factor 1	Factor 2	Factor 3
lnsert			
LT	.209	**.635**	.114
ST	.082	**.905**	.076
RD	-.351	**.732**	-.163
LL	**.867**	.229	-.004
OPDCT	**.728**	-.177	-.327
RP	**.790**	.018	-.185
SA	.145	-.095	**.730**
RPSA	.033	.077	**.823**
MCDCT	.197	-.087	-.630

Differential groups: Another method for determining construct validity is through an analysis of differential groups.[7] The participants in this study, it may be recalled, were divided into three different groups based on the length of their overseas experience. Group 1 had spent little or no time overseas, Group 2 from 5–10 weeks, and Group 3 a year or more (Table 1). Since in these tests the construct is pragmatic competence, it would be expected that the group with the greatest amount of time overseas in English-speaking environments would have the greatest amount of pragmatic competence.

A multivariate analysis of variance (MANOVA) procedure showed that there were significant differences among these three groups in terms of their test results. Univariate follow-up statistics were then run to determine the extent to which each of the tests differentiate between these groups, as given in Table 9 below.

Table 9: Univariate Follow-up Statistic

Variable	Hypoth. SS	Error SS	Hypoth. MS	Error MS	F	Sig of F
LT	18.898	339.341	9.449	15.424	.612	.551
ST	29.965	509.075	14.982	23.139	.647	.533
RD	66.408	445.591	33.204	20.254	1.639	.217
SA	515.098	4594.741	257.549	208.851	1.233	.311
RPSA	1191.190	3725.450	595.595	169.338	3.517	.047*
RP	1352.64	1310.443	676.320	59.565	11.354	.000**

*$p < .05$
**$p < .001$

As indicated, the univariate follow-up statistic showed p values below .05 for two of the tests, the RPSA and the RP. Since these two tests yielded values at the $p < .05$ level, the Scheffé post hoc test was conducted to determine the significance of paired differences. For the RPSA test, the Scheffé test showed no two pairs of groups were significantly different at the .05 level. However, Scheffé post hoc analysis of the variance of the RP test, which had yielded a particularly low p value of .0004, showed significant Scheffé paired differences with the mean scores of Group 3 substantially and significantly different from either those of Group 1 or Group 2, as can be seen in Table 10.

Table 10: Scheffé Paired Differences Test for the RP Test

Group	Grp 2	Grp 1	Grp 3	Mean
Grp 2				74.3611
Grp 1				76.5417
Grp 3	*	*		93.4667

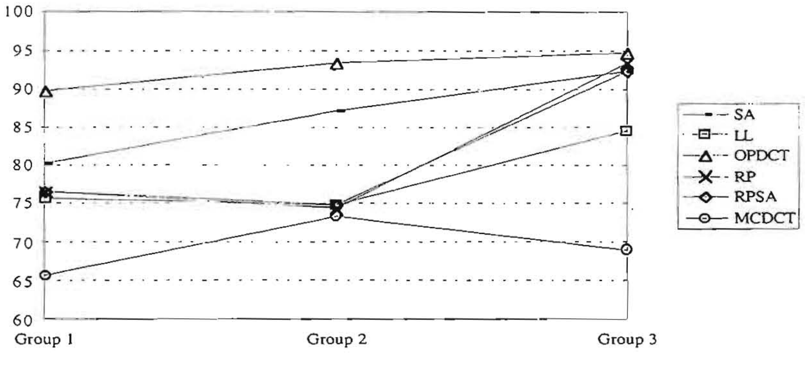
*$p < .0$

It is interesting to note that there is very little difference between Group 1, which had very little overseas experience, and Group 2, which had typically spent several weeks in English-intensive environments. In fact, Group 1 had a higher mean than that of Group 2, but this may have just been a random variation due to the relatively small number of participants in this study. That Group 3 had a much higher mean than either of the other two groups suggests that the development of pragmatic competence requires a substantial amount of time in the target culture.

Means comparison: A means comparison of the various tests offered further insight into the construct validity of the measures in this study (see Table 4 for all means). Among the TOEFL subtests there was very little differentiation between the three groups, and no clear patterns emerged from the data. The scores were very closely grouped by test for all three groups. The totals of the mean scores for each of the groups, in fact, were nearly the same, showing but a very slight increase by group: 150.36 for Group 1, 150.74 for Group 2, and 151.4 for Group 3.

With the tests of pragmatic competence, however, there was significantly more differentiation between the means scores of the groups. This can be seen in Figure 1. With the tests by Hudson et al., Group 3 clearly scored higher than the other two groups in all but the MCDCT test. This is particularly true of both the RP and the SA tests. The RP test, since it provides native speaker raters with a rich array of material on which to base their assessment, would be expected to provide the most accurate assessment of these students' pragmatic competence. It is interesting to note, however, that the RPSA scores are very nearly parallel with the RP scores, suggesting the students may be able to evaluate their own performance as well as the native speaker raters. The LL test also clearly differentiated the pragmatic competence of the Group 3 participants from those of Groups 1 and 2, while the SA and OPDCT

Figure 1: Means Comparision by Differential Groups—Pragmatic Tests



showed a small amount of differentiation. The MCDCT, however, was clearly out of synch with the other tests, and shows Group 3 to have less pragmatic competence than either of the other two groups.

A final point of interest is the disparity between the SA mean and the RPSA and RP means for Group 2, most of whom had recently returned from six-week overseas English-study experiences. On the SA test they seem to have been quite confident of their pragmatic competence as indicated by scores that, on average, were substantially higher than those for Group 1. After performing the role plays, however, Group 2 as a whole rated themselves a good bit downward, apparently feeling they had not performed nearly as well as they thought they could, which is confirmed by the very similar mean produced by the RP test. Group 1 also rated themselves downward after the RPSA, but not as much as Group 2 did. Group 3, on the other hand, appears to have been the only group that had a fairly clear idea of how well they could and did perform, as evidenced by very similar means for all three tests.

Test Practicality

The level of practicality of the multi-test framework—especially in terms of requirements related to time, number of personnel, and special equipment—varied greatly between the tests. Administering the OPDCT and MCDCT was relatively simple. Just a few minutes were required to hand out the tests and instruct students on how to complete the test at home. Taking the tests, however, did require quite a bit of time, especially the OPDCT. The SA test was also easy to administer. All could take it simultaneously, and it did not require much time nor any special equipment.

Administering the other tests was considerably more involved. For the LL, two cassette tape recorders were required; one for playing the situations, and the other for student responses. Additionally, the test needed to be conducted in a quiet room free from disturbances, and the participants needed to take the test individually. Some 10 minutes were required per student to set them up with the equipment and test. Of the six tests, the greatest amount of time and energy was required to administer the RPSA and RP tests. Although these two tests could be conducted concurrently (the data provided by performing the role plays could be used by the students to rate themselves as well as by the raters), performing a full set of role-plays required some 30 minutes per student. The RP test additionally required that the role plays be recorded on video tape so that these recordings could be distributed for evaluation by each of the raters.

Conclusions

With the exception of the OPDCT and MCDCT, the tests designed by Hudson et al. proved highly reliable and valid in assessing pragmatic competence when administered to Japanese university EFL students. The TOEFL subtest scores, by comparison, did not correlate with the pragmatic competence of the students. It would appear as well that the development of pragmatic competence requires fairly extended periods of time in the target culture for the realization of appreciable gains. A few weeks overseas in English-speaking immersion situations seems not to make much difference in learners' pragmatic competence—a year or more is required based on the results of this study. As for the practicality of administering and evaluating these tests, there was a great deal of variance. Of the four tests that proved both reliable and valid, only the SA test was easy to administer and evaluate, although the results were not as accurate as with those of the LL, RPSA, and RP tests.

One particular limitation of this study has to do with the representativeness of the participant group in terms of the variety of English speakers among native Japanese. The participants were all first-year university students with somewhat similar TOEFL scores, so lacked diversity in age, occupation, and linguistic ability. As suggested by Yamashita (1996), older learners involved in the work force would be more aware of the strict social conventions of Japanese society, making them perhaps more sensitive to sociolinguistic concerns in other languages as well. Native Japanese who use English in a service industry might also have a higher sensitivity to such concerns. Surely the linguistic ability of participants would have some influence on pragmatic competence as well, those with higher levels having a greater range of linguistic options available to them when attempting to be appropriate in a particular situation.

The potential directions of future research are many. As mentioned, having a wider range of participants would be desirable for determining the relationship between age and linguistic competence with pragmatic competence. As suggested earlier when discussing the variation in the ratings by the raters, it would be interesting to do rater comparisons between language teaching professionals and non-teachers to see if teachers have a higher acceptance of pragmatic incompetence than might non-teachers. Similarly, it would be interesting to compare raters from different native English speaking cultures to determine if there is, in fact, variation in standards of appropriateness by culture. Finally, there is the matter of examining the transcriptions of the student utterances in the role plays, for here lies a rich corpus of data for doing a qualitative analysis of these participants' pragmatic competence.

Acknowledgments

We would like to thank Sayoko Yamashita and Randy Thrasher of International Christian University, as well as J. D. Brown of the University of Hawai'i, for their assistance with both data analysis and proofreading of earlier drafts of this paper.

Ken Enochs teaches in the English Language Program at International Christian University, Tokyo.

Sonia Yoshitake-Strain, Ph. D., has taught at International Christian University and Seigakuin University, and is currently Japan Tutor for the Birmingham University MA in TESOL Program.

Notes

1. Making the adjustment for the three raters together involved converting the Pearson *r* values from Table 5 into Fisher *Z* coefficients using a Fisher *Z* transformation table (Guilford & Fruchter, 1978, p. 522). The Fisher *Z* coefficients were then averaged and converted back to Pearson *r* coefficients. These average figures were then adjusted to take into account the number of different raters using the Spearman-Brown Prophecy formula.
2. Internal consistency is an indirect way to estimate (without actually retesting) the consistency of a test. One common estimate of a test's internal consistency is to use the split-half method to first determine the correlation between odd and even numbered items in the test, and then adjust the half-test correlation using the Spearman-Brown Prophecy formula to estimate full-test reliability (Brown, 1996).
3. The standard error of measurement (SEM) is a statistic that uses both the standard deviation of a test and a correlation coefficient to "determine a band around a student's score within which that student's score would probably fall if the test were administered to him or her repeatedly" (Brown, 1996, p. 206).
4. The coefficient of determination, according to Brown (1996), shows the proportion of variance between the scores that is common to both, or the degree to which the two tests line up the students in the same order.
5. Principal component analysis involves determining "whether there are components that are shared in common by [several] tests and whether we can capture them in a meaningful way" (Hatch & Lazaraton, 1991, p. 490).
6. Factor analysis reduces a matrix of correlation coefficients to more manageable proportions, the result of which can be used to identify factors that the set of tests have in common (Alderson, Clapham & Wall, 1995, p. 289).
7. Analysis of differential groups determines the extent to which one group has more of the construct in question than another group (Brown, 1996, p. 240).

References

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

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

Beebe, L.M. & Takahashi, T. (1989). Sociolinguistic variation in face-threatening speech acts. In M. Eisenstein (Ed.) *The dynamic interlanguage* (pp. 199–218). New York: Plenum.

Beebe, L.M., Takahashi, T. & Uliss-Weltz, R. (1990). Pragmatic transfer in ESL refusals. In R.C. Scarcella, E. Andersen & S.C. Krashen (Eds.), *On the development of communicative competence in a second language* (pp. 55–73). New York: Newbury House.

Brown, J.D. (1996). *Testing in language programs*. Upper Saddle River, NJ: Prentice Hall.

Canale, M. (1988). The measurement of communicative competence. *Annual Review of Applied Linguistics*, 8, 67–84.

Canale, M. & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, 1, 1–47.

Crystal, D. (1991). *A dictionary of linguistics and phonetics*. Oxford: Blackwell.

Enochs, K. and Yoshitake, S. (1996). Self assessment and role plays for evaluating appropriateness in speech act realizations. *ICU Language Research Bulletin, 11*, 57–76.

Guilford, J.P. & Fruchter, B. (1978). *Fundamental statistics in psychology and education*. New York: McGraw-Hill.

Hatch, E. & Lazaraton, A. (1991). *The research manual: Design and statistics for applied linguistics*. New York: Newbury House.

Hudson, T., Detmer, E. & Brown, J.D. (1992). *A framework for testing cross-cultural pragmatics*. Honolulu: University of Hawaii Press.

Hudson, T., Detmer, E. & Brown, J.D. (1995). *Developing prototypic measures of cross-cultural pragmatics*. Honolulu: University of Hawaii Press.

Hymes, D. (1972). On communicative competence. In J.B. Pride & J. Holmes (Eds.) *Sociolinguistics*. Harmondsworth, Middlesex, UK: Penguin.

Mundby, J. (1978). *Communicative syllabus design: A sociolinguistic model for defining the context of purpose-specific language programmes*. Cambridge: Cambridge University Press.

Thomas, J. (1983) Cross-cultural pragmatic failure. *Applied Linguistics*, 4, 91-112.

TOEFL: Test and score data summary. (1995-1996). Princeton, NJ: Educational Testing Service.

Yamashita, S. (1996). *Comparing six cross-cultural pragmatics measures*. Unpublished doctoral dissertation, Temple University, Philadelphia.

Yoshitake-Strain, S. (1997). *Interlanguage competence of Japanese students of English: A multi-test framework evaluation*. Unpublished doctoral dissertation, Columbia Pacific University, California.

(Received May 18, 1998; revised December 11, 1998)

Appendix: Sample Items of the Six Tests

Self-assessment test (SA)

Situation 1:

You live in a large house. You hold the lease to the house and rent out the other rooms. You are in the room of one of your house-mates collecting the rent. (This house-mate moved in recently.) You reach to take the rent check when you accidentally knock over a small, empty vase on the desk. It doesn't break.

Rating: I think what I would say in this situation would be

very unsatisfactory	1	2	3	4	5	completely appropriate

Listening laboratory production test (LL)

Situation 2:

You are applying for a job in a company. You go into the office to turn in your application form to the manager. You talk to the manager for a few minutes. (The manager is impressed by your CV and wants to hire you.) When you move to give the manager your form, you accidentally knock over a vase on the desk and spill water over a pile of papers.

You say:

Open discourse completion test (OPDCT)

Situation 3:

You have recently moved to a new city and are looking for an apartment to rent. You are looking at a place now. You like it a lot (and talk to the manager for a few minutes). The landlord explains that you seem like a good person for the apartment, but that there are a few more people who are interested. The landlord says that you will be called next week and told if you have the place. However, you need the landlord to tell you within the next three days.

You say:

Multiple choice discourse completion test (MCDCT)

Situation 4:

You are a member of the local chapter of a national ski club. Every month the club goes on a ski trip. You are in a club meeting now helping to plan this month's trip. The club president is sitting next to you and asks to borrow a pen. You cannot lend your pen because you only have one and need it to take notes yourself.

a. Oh, sorry, it's my only one. Maybe John has an extra. Let me check.
b. I'm terribly sorry, this is the only one I have at the moment. Perhaps you might ask John?
c. No, I can't lend this pen. It's my only one.

Role-play self-assessment test (RPSA) & Role-play test (RP)

Situation 6:

Background 6a: You work in a small shop that repairs jewelry. You do not do the repairs yourself; a repairman comes in at night to do the repairs.

Now: A valued customer comes into the shop to pick up an antique watch that you know is to be a present. You need to go in the back room to get the watch, but the customer is standing in the way of the door.

Background 6b: The repairman has not repaired the watch yet, even though it was supposed to be ready.

Now: Go back out to the customer.

The interlocutor is the customer. He will:

- stand in front of the backroom door
- request watch and hand over the slip
- move after request to move
- accept that it is not ready, agree to come back tomorrow
- ask for change for the bus
- see you tomorrow

Note: Have no change in the till

Working at the Jewelry Repair Shop

1. **Request**	very unsatisfactory	1	2	3	4	5	completely appropriate
2. **Apology**	very unsatisfactory	1	2	3	4	5	completely appropriate
3. **Refusal**	very unsatisfactory	1	2	3	4	5	completely appropriate