Issues in Language Teaching (ILT), Vol. 2, No. 1, 1-26, June 2013

Assessing Language Learners' Knowledge of Speech Acts: A Test Validation Study

Parviz Birjandi
Professor of Applied Linguistics, Islamic Azad University, Science and Research Branch, Tehran, Iran

Mohammad Mehdi Soleimani
Ph.D. Candidate, Islamic Azad University, Science and Research Branch, Tehran, Iran

Received: November 11, 2012**; Accepted:** April 4, 2013

Abstract

Very few attempts have been made in the past to develop instruments to measure pragmatic knowledge of second language (L2) learners. The absence of such instruments in the literature of English language teaching (ELT) underscores the need for the researchers to develop new tests that are specifically designed to assess this crucial but less explored aspect of language learners' (LLs) knowledge. In line with this objective, the present study was conducted to develop and validate four tests of pragmatic knowledge that measured LLs' knowledge of speech acts. The following steps were taken in this study to develop the written discourse completion tests (WDCTs) and the multiple-choice discourse completion tests (MCDCTs) that respectively measured the test takers' ability to produce and comprehend request speech act. During the "prototype step" the researchers identified the content and the number of items for each designated test battery. At the "test construction step" the sociolinguistic variables of power (P), social distance (D), and absolute rank of imposition (R) were inserted into the content of the test items. Finally, at the "validation step" the reliability of the tests was examined. The finding of the study showed that the constructed test batteries were sufficiently reliable and valid for measuring pragmatic knowledge of L2 learners.

***Keywords*:** pragmatic competence, interlanguage pragmatics, test development, speech acts

Authors' emails: pbirjand@yahoo.com; mmsoleimani@gmail.com; soleimani@kiau.ac.ir

INTRODUCTION

Although the necessity of pragmatics instruction is felt by the majority of second language (L2) educators, teachers of English as a foreign language (EFL) still hesitate to integrate the teaching of pragmatics into their every day practice. Three reasons can be provided to explain why EFL teachers avoid teaching pragmatics. Rose (1994) refers to the first reason by stating that the majority of teachers in EFL contexts are non-native speakers of English (NNSs) and hence they cannot draw on their language intuition to cope with pragmatics. Therefore, teaching pragmatics is a difficult task especially when the teachers themselves do not feel confident about their own pragmatic competence.

The other reason relates to the paucity of available pedagogical resources that are suitable for pragmatics teaching. The results of Bardovi-Harlig, Hartford, Taylor, Morgan, and Reynolds (1991) survey on conversational closings, Boxer and Pickering's (1995) analysis of complaints, Petraki and Bayes' (2013) study on the teaching of oral requests, Gilmore's (2004) study on discourse features, and Uso-Juan's (2010) research on request modification devices all illustrate that textbook materials are not reliable sources of pragmatic input for EFL learners.

Liu (2012) also argues that available tests that teachers use in their classrooms mainly focus on the linguistic aspects of language and the pragmatic aspects receive scant attention in these tests. Therefore, these tests can discourage EFL teachers from including pragmatic aspects of language into their teaching practice. Thus, it can be argued that the development of some valid and reliable measures to assess the students' pragmatic competence seems necessary and the development of such tests of pragmatic proficiency can bridge the gap between teaching demands and testing instruments.

LITERATURE REVIEW

In one of the earliest efforts in measuring the pragmatic knowledge of language learners, Farhady (1980) developed a multiple choice test to assess the students' ability to express intellectual attitudes. To develop his test, Farhady constructed a number of scenarios and limited the context of these scenarios to academic settings. He further included two social variables of relationship and status between the interlocutors into

the content of these scenarios. In the first phase of test construction, Farhady administered the open ended test items to 200 native speakers of English and collected their responses. He then selected the most frequent response for each item as the correct answer. In the second phase, Farhady administered the test to150 non-native speakers and compared their responses with those of the native speakers to identify deviant forms. Depending on the type of deviation, Farhady developed three other alternatives for the test items. He used a three-group classification system to categorize these alternatives: first, socially appropriate but linguistically inaccurate; second, socially inappropriate but linguistically accurate; third, neither socially appropriate, nor linguistically accurate. In the third phase, Farhady administered his multiple choice test to 30 native and non-native speakers of English to ensure the appropriateness of the alternatives. Finally, he divided the test into two counterbalanced forms and administered them as part of the University of California's placement test to validate his newly developed test of pragmatics. The results of the study showed that the constructed functional test was as valid and reliable as other subtests of the placement test.

In another attempt to develop a test of pragmatics, Shimazu (1989) developed and validated a multiple-choice test that aimed to measure the students' knowledge of requests. In the first phase of test construction, Shimazu developed 61 test items. These open ended items were then administered to 48 native speakers of English and 43 non-native speakers of English. Shimazu used native speakers' responses as the key and non-native speakers' responses as the distracters. Shimazu then used Farhady's (1980) four-group classification to categorize the elicited responses. In the second phase, Shimazu developed a 50 multiple-choice item test and administered the test to 60 native speakers and 72 non-native speakers. In phase three, Shimazu used 40 items and administered these items to 157 native and non-native speakers. In phase four, Shimazu selected 28 items of the test and administered the shortened form of the test along with a Test of English as a Foreign Language (TOEFL). The results of Shimazu's study revealed moderate ranges of concurrent validity coefficients between the newly developed test and the TOEFL.

In their ground-breaking work on the assessment of pragmatic knowledge, Hudson, Detmer, and Brown (1995), used Brown and Levinson's (1987) theory of politeness to create a battery of six tests, including a written discourse completion task (WDCT), a multiple-choice

discourse completion task (MCDCT), an oral discourse completion task (ODCT), a discourse role play task (DRPT), a discourse self assessment task (DSAT), and a role play self assessment (RPSA). The constructed test items targeted the knowledge of test takers about apologies, refusals, and requests. In the initial pilot version, the researchers developed 48 items in an open ended WDCT format. These items were then distributed into two test packages: package A and package B. Each of the two packages was then examined by four native speakers of English for item evaluation. After evaluation, package A was administered to eight native speakers of English and five non-native speakers. Package B was also administered to eight native speakers of English and twelve non-native speakers of English. The analysis of the data showed that some of the test items were faulty. The problem items generally fell under one of the following areas. First, some items elicited wrong speech acts. Second, some items displayed such a low degree of imposition that respondents opted out of responding to the items. The researchers revised the problematic test items and administered each test package to nine more native speakers. Furthermore, Blum-Kulka, House, and Kasper's (1989) coding scheme was used for the analysis of the elicited responses. The respondents' strategies were further analyzed to reveal differences between native and non-native speakers' responses. Finally, the researchers developed multiple-choice options for their MCDCT format based on the strategies that native speakers employed. It should be noted that Hudson et al. did not validate the test batteries themselves. This part was carried out by Yamashita (1996). Yamashita found five of the six tests to be reliable and valid; however, she reported that the MCDCT was problematic.

Roever (2006) developed and validated a test of pragmatics that intended to measure the learners' knowledge of implicatures, routines, and speech acts. Each subsection of Roever's test included twelve items that were drawn from previous studies. In the implicature section, the test-takers' comprehension of English implicature was tested with eight items targeting idiosyncratic implicature and four items targeting formulaic implicature. In the routine section, the test takers were asked to identify the option that best matched the situation. The speech act section consisted of twelve short-answer items, presented as discourse completion tasks with rejoinders. Four items were devoted to each of the three speech acts of request, apology, and refusal. The pilot study was conducted in three stages. In the first stage, Roever administered the test

to 35 respondents to identify malfunctioning items. In the second stage, Roever administered the revised test to 38 German EFL learners to evaluate the suitability of the test for the target group. In the third stage, Roever collected concurrent verbal protocols from six native speakers of English and made necessary changes based on the respondents' comments. To validate the test, Roever administered it to 267 learners of English and 14 native speakers of American English. The analysis of the data showed that the test was sufficiently reliable and valid.

Liu (2007) developed a battery of three tests (i.e., a WDCT, an MCDCT, and a DSAT) to assess the knowledge of Chinese EFL learners in making apologies and requests. Liu developed his test items in five stages. In the exemplar generation stage, Liu asked the learners to name some obligatory contexts for making apologies and requests. In the likelihood investigation stage, he asked the learners to report how likely it was for them to face those contexts in their daily lives. In the metapragmatic assessment phase, the researcher asked the learners to talk about the contextual variables in each scenario. In the pilot study, Liu evaluated the appropriateness of the constructed scenarios. Finally, he developed the multiple-choice options for the MCDCT format. Liu administered the tests to 200 Chinese EFL learners who were divided into two proficiency groups based on their TOEFL scores. The results of the study revealed that WDCT and DSAT were highly reliable. The results also showed that the MCDCT was reasonably reliable and valid.

Grabowsky (2009) developed a speaking test with four reciprocal speaking tasks, in which the test takers performed role plays with a native-speaker partner. These tasks provided the test takers with scenarios that required them to assume a role in order to achieve a communicative goal in the conversation (e.g., get their neighbor to turn down the loud music). The task also provided the test takers with some information about the sociolinguistic, sociocultural, and psychological dimensions of the situation (e.g., the relationship between the interlocutors, and culturally relevant situational information). Grabowsky piloted these role-play tasks at three different phases. In the first phase, he asked the test takers to evaluate the test tasks and comment on the administration procedure. Based on the recommendations, Grabowsky lengthened the test directions to clarify the role play process. In the second phase, Grabowsky asked the test takers to comment on the authenticity of tasks. After this phase, the researcher expanded the role and situation descriptions and controlled for the contextual features to

elicit more negotiation from the interlocutors. In the final phase, he administered the test and analyzed the data. In this phase, the data revealed that the tasks did in fact elicit negotiation and relatively long turn taking sequences. Although there was some variation in the language used in the responses, the meanings expressed in the tasks and the outcomes themselves remained fairly consistent and stable for the respondents.

PURPOSE OF THE STUDY

The present study gains significance in the light of the fact that limited attempts have been made in the past to develop tests that measure pragmatic knowledge of L2 learners. One reason can be that this part of linguistic knowledge does not easily lend itself to testing. The other reason can relate to the fact that pragmatic knowledge, unlike grammatical knowledge, is dependent upon simultaneous interaction of language form as well as language function.

Therefore, the absence of such tests in the ELT literature underscores the need for researchers to develop new tests of pragmatic knowledge that are specific in scope and content. In line with this objective, the present study aims to develop and validate four tests of pragmatic knowledge that each measure LLs' knowledge of request speech act.

Several reasons can be stated for the selection of request speech act in this study. First, requests are face-threatening acts; therefore, their successful realization demands considerable expertise on the part of the learners. Second, the patterns for the realization of requests are culture-bound. Third, requests play an essential role in the social and academic life of foreign language learners. Fourth, successful realization of requests provides language learners with opportunities for getting more exposure to the target language. Fifth, the introduction of reliable tests that are specifically developed to measure the request knowledge of Iranian EFL learners can motivate researchers to examine request realization patterns of Iranian EFL learners and contribute to the existing literature in this domain (e.g., Eslami-Rasekh, 1992; Tajvidi, 2000).

In line with the above stated reasons, the following research questions were formulated for this study:

1. Are the newly developed WDCTs reliable instruments for measuring EFL learners' ability to produce English requests?
2. Are the newly developed MCDCTs reliable instruments for measuring EFL learners' ability to comprehend English requests?

METHOD

Participants

A total of sixty one native speakers of English participated in different stages of data gathering process. It should be acknowledged that the researchers did not personally meet many of these NSs, because this part of the data collection was mainly carried out by the researchers' friends and colleagues who were living in English speaking countries at the time when this study was being carried out. However, certain measures were taken to ensure the validity of the obtained data from these NSs. First, data collectors were requested to refer to educated NSs as the preferred population for data gathering. They were also asked to make sure that English was the respondents' L1. It was also mandatory that all the respondents had to answer all the items of the questionnaire. Finally, the collected responses were cross-checked with another native speaker to ensure that the responses were made by "true" NSs of English.

Table 1: Characteristics of the native speakers

		Age (Average)	Nationality				Total
			British	Australian	Canadian	American	
Gender	Male	36	22	2	9	0	33
	Female	34	18	9	0	1	28
Total			40	11	9	1	61

It should be noted that eighty non-native speakers of English also participated in this study. All these participants were senior students who were studying English Language and Literature at Islamic Azad University, Karaj Branch. These students participated in this study on a voluntary basis. Three raters, including a native speaker of English and two assistant professors of applied linguistics, were also assigned to undertake the rating task in this study. It should be noted that the raters were all professional EFL teachers with at least ten years of language teaching experience.

Instrumentation

Two researcher-made WDCTs were used in this study. The newly developed WDCT test batteries were each made up of eight items (i.e., scenarios) that exclusively focused on request speech act. The research participants were required to read these scenarios and provide their answers to each item. Two researcher-made MCDCTs were also used in this study. These test batteries were also made up of eight multiple-choice items that each focused on request speech act. Hudson et al.'s (1995) WDCT and MCDCT were also administered in this study. It needs to be pointed out that Hudson et al.'s WDCT test included twenty four items (i.e., scenarios) that were designed to assess the test takers' knowledge of three speech acts: requests, refusals, and apologies. Hudson et al.'s (1995) multiple-choice module was made up of twenty-four multiple-choice items that each appraised language learners' knowledge of speech acts, including requests, refusals, and apologies.

Data Collection Procedure

The following steps were taken for the construction of the researcher-made WDCT and MCDCT test batteries in this study. In the "prototype" step, a questionnaire of thirty hypothetical situations was developed. The questionnaire was then distributed among thirty proficient students of English Literature. The respondents were requested to read the situations and indicate on a five-point Likert scale the likelihood that they would find themselves in a similar situation in real life events. Based on the ratings, the top sixteen situations were selected and they were turned to lengthy scenarios based on the following criteria:

Three sociolinguistic variables of relative power (P), social distance (D), and absolute ranking of imposition (R) were selected as the main components of pragmatic knowledge. In this study, the relative powered (P) was defined as the power of speaker with respect to the hearer, and social distance (D) was defined as the degree of familiarity and solidarity between the speaker and hearer. The absolute ranking of imposition (R) was defined as the potential imposition of carrying out the speech act, in terms of the expenditure of goods and/or services by the hearer, or the obligation of the speaker to perform the act. The rationale for selecting these sociolinguistic variables was that these variables are identified within the research on cross-cultural pragmatics "as the three

independent and culturally sensitive variables that subsume all other variables and play a principled role in speech acts behavior of the interlocutors" (Hudson et al., 1995, p. 4).

During the "test construction" step, the abovementioned sociolinguistic variables were inserted into the structure of the scenarios. For this purpose, each of the selected sociolinguistic variables were given plus and minus (±) values. Consequently, these three sociolinguistic variables were turned to six variants with plus and minus values {i.e., (±P), (±D), and (±R)}. For instance, one of the scenarios was constructed using plus values {i.e., (+P), (+D), and (+R)}. This combination of sociolinguistic variables resulted in a hypothetical scenario in which the speaker had the power to ask for a great favor from someone he did not know well (e.g., the head of sales department asks a new salesperson to lend him his car for a few days). Yet in another scenario the following combination of sociolinguistic variables (+P), (+D), and (-R) was used to depict the speaker as someone who enjoyed a high status, who asked a hearer, whom he did not know well, for something of little value (e.g., the same head of sales department asks the new salesperson to lend him a pen).

Based on the following equation (i.e., $2^n \rightarrow 2^3$), the researchers realized that eight items were needed to be constructed for each WDCT to capture all possible interactions between sociolinguistic variables. However, it should be noted that researchers planned to develop two test batteries for each format to minimize the risk of unexpected failure(s) during the test construction process. Therefore, the top sixteen situations were selected from the questionnaire that was distributed among the English majors and these situations were turned to lengthy scenarios using the above discussed sociolinguistic variables. Table 2 shows the distribution of sociolinguistic variables throughout the test batteries.

Table 2: Distribution of sociolinguistic variables through the items

Sociolinguistic Variables	**PDI** +++	**PDI** ++−	**PDI** +−+	**PDI** +−−	**PDI** −++	**PDI** −+−	**PDI** −−+	**PDI** −−−
WDCT (A)	1	2	3	4	5	6	7	8
WDCT (B)	9	10	11	12	13	14	15	16

During the "revision stage" two assistant professors of English and a native speaker of English read the items and commented on the content and highlighted grammatical and contextual inaccuracies in the scenarios.

Later, these comments were used to make necessary changes in the form and content of the tests.

Data Analysis

When the WDCT test batteries were constructed, they were distributed among eighteen native speakers of English (i.e., 11 British, 5 Australian, and 2 Canadian) for cross-check examination. At the "verification stage" part of this study, the native speakers were asked to read the scenarios and specify the relationship between the interlocutors by determining the relative power of the speaker with respect to the hearer, the distance of their relationship, and the degree of imposition involved in each request on a five-point Likert scale. This was done to ensure whether the researchers' perception of sociolinguistic variables, as identified by plus or minus values, matched those of the native speakers. The following excerpt might help clarify this point.

> **Scenario 16:** You are leaving class early. A backpack belonging to one of your classmates is blocking your way. You would like to move the backpack, but you cannot reach it because you are carrying your own books.
>
> ***What degree of power does the speaker have over the hearer?***
> ***Limited (-P)*** *1……..2…..3……4…….. 5* ***Considerable (+P)***
>
> ***How close do you think the speakers are?***
> ***Strangers (+D)*** *1……2……..3……..4…….. 5* ***Very close (-D)***
>
> ***How imposing do you think the request is?***
> ***Very little (-I)*** *1……2……..3…..4…….. 5* ***Very much (+I)***

Ratings were then averaged for each scenario, and the averages were ranked for each sociolinguistic variable. The median score (i.e., 3) was set as the criterion and scores above the median were considered as (+P), (-D), and (+I). Needless to say that scores below the median were considered as (-P), (+D), and (-I). When the ratings were tallied, the researchers found some mismatches between their perception of sociolinguistic variables and those of the native speakers for some of the scenarios. The defective scenarios were then modified and re-

administered so that they could meet the expectations of the native speakers.

RESULTS

Research Question One

For the "validation stage", the newly developed WDCT test batteries and Hudson et al.'s (1995) WDCT were administered to thirty Senior English majors who were studying at Islamic Azad University, Karaj Branch. These tests were administered in the following order: the students first received Hudson et al.'s test of pragmatic knowledge; they then received the newly developed WDCTs with an interval of two weeks from the first test. The students had two hours to finish Hudson et al.'s test and forty five minutes for each of the newly developed WDCTs. It should be noted that five students failed to take all the tests; therefore, the number of students who participated in this part of the study declined to twenty five.

After the administration of the tests, three raters were assigned to undertake the rating task. The raters rated the appropriateness of students' responses based on Hudson et al.'s (1995) rating sheet. This rating sheet requires the raters to indicate on a five-point Likert scale the assessment of the correct speech act, formulaic expressions, amount of speech, degree of formality, directness, and politeness. To create greater harmony in the rating task, the raters were asked to work on a mock test and they were urged to discuss how they would rate the responses based on the rating sheet.

When the tests were graded, the scores were examined to see whether the scores were normally distributed. For this purpose, Kolmogorov-Smirnov and Shapiro Wilks tests were used. As Table 3 displays, the significance values for both Kolmogorov-Smirnov and Shapiro Wilks tests are bigger than the specified alpha value of 0.05. Therefore, we can confidently state that the assumptions of normality are not violated in this data set.

Table 3: Normality of scores on Hudson et al. & the newly developed WDCTs

	Kolmogorov-Smirnov			Shapiro-Wilk		
	Statistic	**df**	**Sig.**	**Statistic**	**df**	**Sig.**
Hudson et al. WDCT	.16	25	.071*	.93	25	.123*
Researcher-made WDCT (A)	.10	25	.200*	.94	25	.144*
Researcher-made WDCT (B)	.10	25	.200*	.95	25	.356*

When the normality of distribution was established, the raters' judgments were examined to see whether they were scoring the tests based on similar criteria. As Table 4 shows, the Cronbach's Alpha value indicates a high inter-rater reliability of 0.89.

Table 4: Inter-rater reliability for the newly developed WDCTs and Hudson et al.'s WDCT

Cronbach's Alpha	**Cronbach's Alpha Based on Standardized Items**	**Number of Items**
.89	.94	9

When the inter-rater reliability was ensured, the raters' judgments were used to examine whether the newly developed tests were parallel. According to Bachman (2004), two tests can be parallel when the following three conditions are met. First, the means of the tests are equal. Second, the variances of the tests are equal. Third, the tests are developed based on similar test construction procedure. Fourth, the tests are equally correlated with a third measure of the same ability. This latter condition was also used to examine the reliability of the newly developed test batteries. The descriptive statistics for the performance of the students on the newly developed tests is presented in greater details in Table 5.

Table 5: Performance of the students on the newly developed WDCTs

Constructed Tests	**N**	**Min**	**Max**	**Mean**		**Std. Deviation**	**Variance**
WDCT (A)	25	19.3	29.9	22.78	.50	2.50	6.27
WDCT (B)	25	19.0	30.7	23.60	.47	2.39	5.75

As Table 5 shows, there is a slight variation between the means and the variances of the newly developed WDCT test batteries. To test whether the differences are large enough to jeopardize the assumption of parallelism, a paired samples t-test was performed.

Table 6: Paired samples t-test on the scores of the students on developed WDCTs

	Paired Differences							
	Mean	**Std. Deviation**	**Std. Error Mean**	**95% Confidence Interval**				
				Lower	**Upper**	**t**	**Df**	**Sig. (2-tailed)**
WDCT(A) WDCT(B)	-.82	2.22	.44	-1.74	.08	-1.8	24	.075

As Table 6 above illustrates, the probability value indicates a bigger value than the cut-off of 0.05. Therefore, we can conclude that the mean and variance differences between the two tests are insignificant.

To test Bachman's (2004) fourth condition, Pearson correlation was used to examine whether the newly developed test batteries correlated with a third measure of the same ability. For this purpose, the degree of correlation between Hudson et al.'s WDCT and the newly developed test batteries was examined. According to Cohen (1988) correlations above 0.50 are considered as acceptable correlation between variables. Therefore, as Table 7 indicates, the newly developed tests correlate with Hudson et al.'s test of pragmatic proficiency.

Table 7: The degree of correlation between the newly developed WDCTs and Hudson et al.'s WDCT

		Hudson et al.	**WDCT-A**	**WDCT-B**
Hudson et al. WDCT	Pearson Correlation	1	.699**	.669**
	Sig. (2-tailed)		.000	.000
	N	25	25	25
Researcher-made WDCT (A)	Pearson Correlation	.699**	1	.590**
	Sig. (2-tailed)	.000		.002
	N	25	25	25
Researcher-made WDCT (B)	Pearson Correlation	.669**	.590**	1
	Sig. (2-tailed)	.000	.002	
	N	25	25	25

To estimate the reliability of the newly developed test batteries, their internal consistency was checked. As Larson-Hall (2010) argues, one of the most commonly used indicators of internal consistency is Cronbach's alpha coefficient. Ideally, the Cronbach alpha coefficient of a scale should range between 0.70 and 0.90. Table 8 indicates the reliability estimate of Hudson et al.'s test of pragmatic proficiency.

Table 8: The reliability statistics for Hudson et al.'s WDCT

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	Number of Items
.83	.83	24

As Table 8 shows, Cronbach's alpha coefficient of Hudson et al.'s test shows the satisfactory value of 0.83. Therefore, we can safely conclude that the items that make up the criterion measure of pragmatic proficiency hang together quite well. Tables 9 and 10 indicate the reliability estimates for the newly developed test batteries.

Table 9: The reliability estimate for the newly developed WDCT (A)

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	Number of Items
.69	.72	8

Table 10: The reliability estimate for the newly developed WDCT (B)

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	Number of Items
.61	.63	8

As Tables 9 and 10 indicate, the resulted Cronbach's alpha coefficients for the newly developed test batteries are below the acceptable value of 0.70 (i.e., α=0.69 for test B & α=0.61 for test C). However, it is important to note that Cronbach Alpha values are quite sensitive to the number of items in a scale. As Pallant (2007) argues, it is common to find low Cronbach values in scales with fewer than ten items. Therefore, considering the number of items in each test battery (i.e., eight items), one can argue that the length of the tests might have negatively affected the reliability index. To provide evidence in support of this argument, the newly developed tests were merged to form a longer piece

to examine whether the length of the scales was in fact a determining factor in the observed low Cronbach's alpha coefficient.

The rationale for merging the newly developed tests was twofold. First, identical procedures were used for the development of test items in each form. Second, the analysis of the scores of the students who took the test batteries convincingly indicated that the tests could be considered as parallel. Therefore, when items measure the same construct and enjoy parallel content and identical characteristics, their merger may not harm but enhance the overall reliability of the scale. Table 11 illustrates the reliability estimate for the newly developed WDCTs after the merger process.

Table 11: The reliability statistics for the newly developed WDCTs after the merger

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	Number of Items
.75	.76	16

As Table 11 shows, the value of Cronbach's Alpha coefficient increased after the merger process by indicating the acceptable level of 0.75. Consequently, we can conclude that the length of the tests had a negative effect on the observed reliability index, and the merger process helped verify the reliability of the developed scales.

Research Question Two

Once the reliability of the newly developed WDCTs was verified, the researchers manipulated the format of the tests to measure the test takers' ability to comprehend English requests. For this purpose, the written format of WDCTs was replaced by multiple-choice options. This transformation was achieved by the following test construction steps: choice construction, choice selection, verification, and validation. For the "choice construction" step, the researchers first distributed the WDCTs among twenty five native speakers of English and asked the respondents to read the tests and write their responses to each scenario. The researcher also used the responses of the twenty five non-native speakers who had taken the WDCTs in the earlier stage of the test construction process.

When the responses were collected, the request strategies that native and non-native speakers had employed were identified using Blum-Kulka, House and Kasper's (1989), Takahashi's (1995), and Trosborg's (1995) taxonomies of request speech act. It should be noted that this analysis was repeated using the same taxonomy to ascertain the consistency of the adapted coding system. In fact, the second round of analysis showed a small degree of variation from the first analysis in terms of the identified request strategies. However, it should be acknowledged that the observed degree of variation in the coding procedure was so negligible that it did not seriously harm the overall consistency of the coding system. Therefore, the researchers resolved the observed differences and agreed on the final analysis of the request strategies. The following tables show a sample of native speakers' responses to one of the test items along with the coding system that was used to analyze these responses.

> **Scenario 10:** You have an hour between classes and you feel like having a cup of tea. You decide to go to a cafeteria close to the university to have some tea and spend some time there relaxing. When you get to the cafeteria you go up to the counter and ask for a cup of hot tea with a lemon wedge on the side.

Table 12: A collection of native speakers' responses to scenario 10

1	Tea, please
2	Could I have a cup of hot tea with a wedge of lemon please?
3	Hi, can I please have a cup of tea with a slice of lemon on the side?
4	Can I have a tea? And if you have it some lemon on the side, please
5	Hi, could I have a cup of black tea please? No milk, but can I have a bit of lemon please? Thanks
6	Would you please add a lemon wedge to the tea and place it on the saucer? Thank you very much.

Table 13 indicates how the native speakers' responses to scenario ten are analyzed based on the adapted coding system.

Table 13: A sample of coding procedure for the analysis of (non)native speakers' responses

	Alerter	**Dominance**	**Head Act**	**Support**	**Syntactic downgraders**	**Lexical downgraders**
1	No	Unspecified	Mood derivable	No	No	Polite marker
2	No	Speaker dominated	Query preparatory	No	Tense	Polite marker
3	Attention getter	Speaker dominated	Query preparatory	No	No	Polite marker
4	No	Speaker dominated	Query preparatory	Min.*	No	Polite marker
5	Attention getter	Speaker dominated	Query preparatory	No	Tense	Polite marker/ Gratitude
6	No	Hearer dominated	Query preparatory	No	Tense	Polite marker/ Gratitude

*Min.: Minimizer

Based on the analysis of available responses, the common strategies that native and non-native speakers used to answer these scenarios were identified. For instance, the above collection of native speakers' responses clearly shows that native speakers of English would most probably go for requests that are: speaker dominated, contain attention getters, use query preparatory head acts that are mitigated by past tense forms (i.e., the use of could instead of can in the head act structure), and are softened by the use of polite words like please to ask for a cup of hot tea in a cafeteria.

For "choice selection" step, four answers from the collection of native speakers' responses and four answers from the collection of non-native speakers' responses were selected. It should be noted that care was exercised to choose a combination of typical as well as less typical request strategies from the native speakers' sample. To explain this point, let's consider scenario ten above. In this case, the most typical request strategy that native speakers used was a combination of: attention getter + speaker domination + query preparatory head act + polite statements; however, in this collection there are also instances of less typical request strategies that native speakers use. For instance, the use of mood derivable as the main head act seems quite uncommon in this case. To

strike a balance between possible options, two typical and two non-typical responses from native speakers' collection were selected to represent the performance of the target group for each scenario. Quite conversely, non-native speakers' responses were selected based on the extent of their deviation from the native speaker norms. In other words, non-native speakers' request strategies were compared with those of the native speakers and based on this comparison the researchers selected four responses from the pool of non-native speaker responses that clearly deviated from the native speaker norms in terms of request strategies.

These scenarios were later distributed among ten more native speakers of English (i.e., 8 British, 1 Australian, and 1 Canadian) along with the selected responses (i.e., eight responses) for each scenario. These native speakers were requested to read the scenarios to identify the accuracy and the appropriateness level of the responses on a five-point Likert scale. The following excerpt from the distributed scenarios might help clarify the point.

> **Scenario 10:** You have an hour between classes and you feel like having a cup of tea. You decide to go to a cafeteria close to the university to have some tea and spend some time there relaxing. When you get to the cafeteria you go up to the counter and ask for a cup of hot tea with a lemon wedge on the side.
>
> **How appropriate do you think the following answers are for this scenario? Choose a number to indicate the level of accuracy and appropriateness of the answers.**
>
> ***a.*** *Tea, please*
> ***Very unsatisfactory*** *1……. 2……..3…….4…… 5* ***Completely appropriate***
>
> ***b.*** *Could I have a cup of hot tea with a wedge of lemon please?*
> ***Very unsatisfactory*** *1……. 2……..3…….4…… 5* ***Completely appropriate***
>
> ***c.*** *A cup of hot tea with lemon makes me relaxed.*
> ***Very unsatisfactory*** *1…… 2……..3…….4……. 5* ***Completely appropriate***
>
> ***d.*** *One tea with lemon please*
> ***Very unsatisfactory*** *1…… 2…….3…….4……. 5* ***Completely appropriate***
>
> ***e.*** *May I please have a cup of tea with a slice of lemon? Thank You.*
> ***Very unsatisfactory*** *1…… 2…….3…….4……. 5* ***Completely appropriate***
>
> ***f.*** *I need a hot tea with a lemon wedge on the side*

Very unsatisfactory *1….. 2……..3……..4…….. 5* ***Completely appropriate***

g. *I was wondering if I could possibly have a cup of hot tea with a lemon wedge on the side*

Very unsatisfactory *1…… 2……..3……..4…… 5* ***Completely appropriate***

h. *A tea for one*

Very unsatisfactory *1…….. 2……3……..4…….. 5* ***Completely appropriate***

To select the multiple-choice options for the MCDCT test batteries, the native speakers' ratings were tallied and ranked from the most to the least appropriate for each scenario. Afterward, the most and the least appropriate responses were respectively selected as the key and the main distracter. The second least appropriate response was also selected as the second distracter for each item. Based on this procedure, the WDCTs were turned to three-option MCDCT comprehension test batteries. The final versions of the multiple-choice tests were later distributed among eight more native speakers of English (i.e., 5 British & 3 Canadian) to ascertain the key options for each scenario. It should be noted that native speakers agreed upon the accuracy of the key options for all scenarios except for two. For these scenarios, the choice that was selected by the majority of the native speakers as the correct response was selected as the true key.

To "validate" the newly developed multiple-choice test batteries, the researchers administered the tests, along with Hudson et al.'s (1995) multiple-choice module to twenty five Junior English majors who were studying English Literature at Islamic Azad University, Karaj Branch. These tests were administered in the following order: students first took Hudson et al.'s multiple-choice test. A week later, they received the newly developed MCDCTs. The students had ninety minutes to take Hudson et al.'s test and thirty minutes to take each of the newly developed MCDCTs.

When the tests were scored, Cronbach alpha coefficient was used to estimate the reliability of the newly developed MCDCTs. Table 14 shows the Cronbach alpha coefficient value for Hudson et al.'s MCDCT.

Table 14: The reliability statistics for Hudson et al.'s MCDCT

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items[a]	Number of Items
.10	-.02	24

Considering the fact that Hudson et al.'s MCDCT has long been used as a valid instrument for measuring the pragmatic knowledge of language learners, the observed Cronbach's Alpha value seems unsatisfactory. This observed alpha value (i.e., $\alpha = 0.1$) is far below the acceptable alpha value; therefore, it can be argued that the multiple-choice items that make up Hudson et al.'s MCDCT do not neatly correlate with each other and this lack of internal consistency affects the reliability of the test. However, as Tables 15 and 16 indicate, the newly developed MCDCTs show a more acceptable alpha values (i.e., $\alpha = 0.62$ for test B & $\alpha = 0.60$ for test C). This shows that the items that make up the newly developed MCDCTs hang together fairly well and this internal consistency strengthens the reliability estimate of the constructed scales.

Table 15: The reliability statistics for the newly developed MCDCT-1 (B)

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	Number of Items
.62	.62	8

Table 16: The Reliability Statistics for the Newly Developed MCDCT-2 (C)

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	Number of Items
.60	.60	8

As Table 17 shows, Cronbach' Alpha coefficient increased to 0.78 when these test batteries were merged. This shows that the length of the test can positively affect the reliability index.

Table 17: The reliability statistics of the newly developed MCDCTs after merger

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	Number of Items
.78	.78	16

DISCUSSION

One of the major difficulties that test developers face in constructing tests of pragmatic knowledge is finding a counterbalance between linguistic and social aspects of the language. This division is well illustrated in Leech's (1983) classification of the components of pragmatic proficiency. According to Leech, pragmatic knowledge, whether in one's first language or second language, consists of two components: pragmalinguistic and sociopragmatic. In this classification, pragmalinguistics represents the linguistic side of pragmatics and it deals with "the particular resources which a given language provides for conveying particular illocutions" (p. 11). The sociopragmatic component, on the other hand, represents the "sociological interface of pragmatics" (p. 10). Sociopragmatics is primarily concerned with the interface of linguistic action and social structure; therefore, it deals with the effect of such social factors as the social status, social distance, and degree of imposition on the linguistic realization of illocutions.

This desired counterbalance between sociopragmatic and pragmolinguistic was achieved through a meticulous description of the *setting*, *participants*, *purpose*, *and content* in each of the constructed scenarios in this study. With regard to setting, care was exercised to provide detailed description of the physical context and/or contextual situation in the constructed scenarios. It should be noted that Varghese and Billmyers (1996) endorse the effectiveness of using detailed prompts in testing pragmatic knowledge. As for the participants, the role relationship between the interlocutors in each scenario (i.e., their status and positional identities) was clearly described in some detail- as professor, fellow student, close friend, and classmate. As Douglas (2000) argues, the inclusion of vivid descriptions about the interlocutors' relationship and their status in each prompt significantly enhances the validity of elicited responses from the test takers. The purpose of each item was also set down for the test takers to minimize the confusion over the type of speech act that the respondents had to use for answering each item. Care was also exercised to limit the content of the prompts to the situations that were familiar to the test takers. For that reason the majority of the scenarios in this study targeted daily conversations that were restricted to academic settings.

The findings indicate that the procedure that was used for the construction of the scenarios in this study was highly effective and

successful. This is endorsed by the fact that the newly developed production tests (i.e., WDCTs) and recognition tests (i.e., MCDCTs) turned out to be reliable instruments for measuring the pragmatic knowledge of second language learners. As the findings indicate, Cronbach alpha reliability estimates for WDCTs and MCDCTs were about 0.75 and 0.78 respectively. It is worth noting that the obtained high reliability index for the constructed WDCTs is in line with previous studies that were conducted by Enochs and Yoshitake-Strain (1999), Liu (2004), Roever (2006), and Yamashita (1996).

Nevertheless, the obtained high reliability estimate for the constructed MCDCTs in this study does not correspond with the findings of Brown (2001), Hudson (2001), Yamashita (1996), and Yoshitake (1997) who all consider multiple-choice format as an unreliable instrument for measuring pragmatic knowledge. The reliability of the constructed MCDCTs in this study can partly be explained by the meticulous procedures that researchers used for the construction of the multiple choice items. As it was discussed earlier, the scenarios and multiple choice options in this study were developed through several independent steps of choice construction, choice selection, choice verification, and test validation. It should be noted that all of the scenarios were closely related to the test takers' life in academic milieus. All of the MC options were also generated by Iranian EFL learners and their inaccuracy was assessed by native speakers.

This study also confirms Yamashita's (1996) finding about the unreliability of Hudson et al.'s MCDCT. One reason for this reported unreliability is the fact that Hudson and his associates were unable to create distracters that were evidently inappropriate for the constructed scenarios. Evidence in support of this claim comes from the fact that the native speakers who took Hudson et al.'s multiple-choice test in this study (5 Canadian, 3 British, and 2 Australian) did not agree on the keys for about half of the scenarios. In other words, a majority could not be reached on the acceptability of the keys for eleven of the scenarios in Hudson et al.'s MCDCT. This ambiguity in multiple-choice options was fortunately not an issue in the newly developed MCDCTs because the native speakers who took the tests could easily agree on the appropriate answer for at least fourteen of the scenarios.

CONCLUSION AND IMPLICATIONS

The lack of reliable and valid instruments for measuring pragmatic knowledge of second language learners was the main motif behind this study. To fill this gap, the researchers reviewed the existing literature and proposed an innovative procedure for developing four tests of pragmatic knowledge that were built on the strength of previous studies. Therefore, it is hoped that this procedure can be used by researchers and test developers as an alternative for previous test construction techniques.

The fact that the constructed comprehension (i.e., recognition) tests are as reliable as production tests might be a cause of relief for language teachers and testers alike. The rationale for this argument is the practicality issue in language testing. It goes without saying that the use of production tests is highly costly and time consuming. This is because the scoring task in these tests should be done by professional raters who are fully familiar with sociopragmatic and pragmalinguistic aspects of the target language. Recognition tests, unlike production tests, can easily be administered and scored by language teachers. Therefore, recognition tests make a better candidate for large scale testing than production tests.

It is hoped that other studies would provide further evidence to support the reliability of the constructed pragmatics instruments. Researchers can also use the presented procedure of test construction for developing similar tests for measuring other speech acts. The scores of students in these tests can also be correlated with the students' scores in other large scale proficiency tests like IELTS and TOEFL. This comparison helps us identify the relationship between language proficiency and pragmatic proficiency.

Written questionnaires were used as the main instrument for data collection in this study. Future studies can employ other data collection instruments to complement the current study and provide additional evidence in support of these findings. It should also be noted that many of the findings and generalization of this study were made based on the performance of the learners on four eight-item DCTs. It is likely that this number of test items may not adequately represent possible real life situations that learners may face in real world conversations; therefore, researchers are encouraged to develop and validate alternative tests of pragmatic proficiency to compensate for this shortcoming.

This study focused on university students; however, it is unknown whether more heterogeneous participant groups would perform the same.

Therefore, it is recommended that other researchers include more heterogeneous learner groups into their studies. In addition, the participants in this study had no direct exposure to the target language and culture; therefore, it is suggested that other researchers include into their participants a group that has such an experience to see whether natural exposure to second language has any effects on the pragmatic development of the learners or not.

Bio-data

Parviz Birjandi is a professor of Applied Linguistics. He received his M.A from Colorado State University and his Ph.D. from the University of Colorado at Boulder in Teaching English to Speakers of Other Languages (TESOL). Currently, he is Head of Graduate Studies in Applied Linguistics at Islamic Azad University, Tehran Science and Research Branch. His primary research interests include language assessment and testing, research methodology, materials development, and first language acquisition. He has published and edited a number of research articles and university textbooks. Also, he is on the editorial board of several academic journals in Iran.

Mohammad Mehdi Soleimani is a Ph.D. candidate studying Applied Linguistics at Islamic Azad University, Tehran Science and Research Branch. He received his M.A. in TEFL from University of Isfahan. Currently, he is Head of the Department of English Language and Literature at Islamic Azad University, Karaj Branch. His primary research interests are second language acquisition, language testing, and interlanguage pragmatics. Also, he is on the editorial board of the Asian EFL Journal and the Journal of English Studies.

References

Bachman, L. (2004). *Statistical analyses for language assessment.* Cambridge: Cambridge University Press.

Bardovi-Harlig, K., Hartford, B. Taylor, R., Morgan, M., & Reynolds, D. (1991). Developing pragmatic awareness: Closing the conversation. *ELT Journal, 45*(1), 4-15.

Blum-Kulka, S., House, J., & Kasper, G. (1989). *Cross-cultural pragmatics: Requests and Apologies*. Norwood: Ablex.

Brown, J.D. (2001). Pragmatics tests: Different purposes, different tests. In K.R. Rose & G. Kasper (Eds.), *Pragmatics in language teaching* (pp. 301-327). Cambridge: Cambridge University Press.

Brown, P., & Levinson, S. C. (1987). *Politeness: Some universals in language usage*. Cambridge: Cambridge University Press.

Boxer, D., & Pickering, L. (1995). Problems in the presentation of speech acts in ELT materials. The case of complaints. *ELT Journal 49*(1), 44-58.

Cohen, J. W. (1988). *Statistical power analysis for the behavioral sciences.* New Jersey: Lawrence Erlbaum Associates.

Douglas, D. (2000). *Assessing language for specific purposes*. Cambridge: Cambridge University Press.

Enochs, K., & Yoshitake-Strain, S. (1999). Evaluating six measures of EFL learners' pragmatic competence. *JALT Journal, 21*(1), 29-50.

Eslami-Rasekh, Z. (1992). *A cross cultural comparison of the requestive speech act realization patterns in Persian and American English.* Unpublished doctoral dissertation, University of Illinois, Urbana-Champaign.

Farhady, H. (1980). *Justification, development, and validation of functional language testing,* Unpublished doctoral dissertation, University of California, Los Angeles.

Gilmore, A. (2004). A comparison of textbooks and authentic interactions. *ELT Journal, 58*(4), 362-374.

Grabowsky, K. (2009). *Investigating the construct validity of a test designed to measure pragmatic knowledge in the context of speaking.* Unpublished doctoral dissertation, Columbia University, New York City.

Hudson, T. (2001). Indicators for pragmatic instruction: Some quantitative tools. In K.R. Rose, & G. Kasper (Eds.), *Pragmatics in language teaching* (pp. 283-300). Cambridge: Cambridge University Press.

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

Larson-Hall, J. (2010). *A guide to doing statistics in second language research using SPSS.* New York: Routledge.

Leech, G. (1983). *The principles of pragmatics.* London: Longman.

Liu, J. (2004*). Measuring interlanguage pragmatic knowledge of Chinese EFL learners.* Unpublished doctoral dissertation. City University of Hong Kong.

Liu, J. (2007). Developing a pragmatic test for Chinese EFL learners. *Language Testing, 24* (3), 391-415.

Liu, J. (2012). Assessing EFL Learners' interlanguage pragmatic knowledge: Implications for testers and teachers. *Reflections on English Language Teaching, 5*(1), 1-22.

Pallant, J. (2007). *SPSS survival manual.* London: McGraw-Hill Publications.

Petraki, E., & Bayes, S. (2013). Teaching oral requests: An evaluation of five English as a second language course books. *Pragmatics, 23*(3), 499-517.

Roever, C. (2006). Validation of a test of pragmatics. *Language Testing, 23*(2), 229-256.

Rose, K. R. (1994). Pragmatic consciousness-raising in an EFL context. *Pragmatics and Language Learning, 5*(1), 52-63.

Shimazu, Y.M. **(**1989). *Construction and concurrent validation of a written pragmatic competence test of English as a second language.* Unpublished doctoral dissertation, University of San Francisco, Los Angles.

Tajvidi, G.R. (2000). *Speech acts in second language learning process of Persian speakers: Communicative and pragmatic competence in cross-cultural and cross-linguistic perspective.* Unpublished doctoral dissertation, Allameh Tabatabaei University, Tehran.

Takahashi, S. (1995). *Pragmatic transferability of L1 indirect request strategies perceived by Japanese learners of English.* Unpublished doctoral dissertation, University of Hawaii, Honolulu.

Trosborg, A. (1995). *Interlanguage pragmatics: Requests, complaints and apologies.* Berlin: Mouton.

Uso-Juan, E. (2010). The presentation and practice of the communicative act of requesting in textbooks: Focusing on modifiers. In E. Alcon & M. Safont (Eds.). *Intercultural language use and language learning* (pp. 223-244). Amsterdam: Springer.

Varghese, M., & Billmyer, K. (1996). Investigating the structure of discourse completion tests. *Working Papers in Educational Linguistics, 12*(1), 39-58.

Yamashita, S. O. (1996). *Six measures of JSL pragmatics.* Honolulu: University of Hawaii at Manoa.

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