Assessing EFL learners' interlanguage pragmatic knowledge: Implications for testers and teachers

Liu Jianda
Guangdong University of Foreign Studies

ABSTRACT

Studies showed that interlanguage pragmatic knowledge is teachable. The necessity and importance of teaching pragmatics have also been recognized, but still foreign language teachers hesitate to teach pragmatics in their classrooms. The hesitation could be partly attributed to the lack of some valid methods for testing interlanguage pragmatic knowledge. This article explores ways to assess Chinese EFL learners' pragmatic competence and meanwhile investigates whether learners of different EFL proficiency levels perform differently in pragmatics tests. Results showed that the test methods used in this study were reliable and valid in assessing Chinese EFL Learners' interlanguage pragmatic knowledge. It is suggested that a combination of elicitation through both native speakers and non-native speakers should be a better and more practical way to construct interlanguage pragmatic test items. The two proficiency groups in this study were shown to differ significantly in terms of their English proficiency, but not on two of the three pragmatics tests, which indicated that the Chinese EFL learners' interlanguage pragmatic knowledge did not seem to increase substantially with their language proficiency. The findings reconfirmed the importance of teaching pragmatic knowledge to Chinese EFL learners in classrooms. The pedagogical implications and applications for foreign language teachers and testers are also discussed. The paper concludes with the suggestion that EFL teachers should teach pragmatic knowledge in class and include interlanguage pragmatic knowledge in large-scale tests.

Introduction

As a domain within L2 studies, pragmatics is usually referred to as interlanguage pragmatics (ILP), as analogy with interlanguage grammar, interlanguage phonology, and interlanguage lexicon (Kasper & Rose, 2002). ILP is a second-generation hybrid (Kasper & Blum-Kulka, 1993). It belongs to two different disciplines, both of which are interdisciplinary. On one hand, as a branch of second language acquisition research, two sections within the wider domain of ILP are distinguished. As the study of L2 use, ILP examines how nonnative speakers (NNSs) comprehend and produce action in a target language. As the study of L2 learning, ILP investigates how L2 learners develop the ability to understand and perform action in a target language (Kasper & Rose, 2002). On the other hand, as a subset of pragmatics, ILP is a sociolinguistic, psycholinguistic, or simply linguistic enterprise, depending on how one defines the scope of

pragmatics (Kasper & Blum-Kulka, 1993). Kasper and Blum-Kulka (1993, p. 3) define ILP as "the study of nonnative speaker's use and acquisition of linguistic action patterns in a second language". They also offer a broader definition of ILP. They argue that "tying interlanguage pragmatics to NNSs, or language learners may narrow its scope too restrictively" (p. 3), and include into ILP "the study of intercultural styles brought about through language contact, the conditions for their emergence and change, the relationship to their substrata, and their communicative effectiveness" (p. 4). But most ILP studies focus on the narrow definition. Kasper's later definition of ILP also focuses on the narrow sense. Kasper (1998, p. 184) defines ILP as

> *the study of nonnative speakers' comprehension, production, and acquisition of linguistic action in L2, or put briefly, ILP investigates how to do things with words in a second language.*

In this study, interlanguage pragmatic knowledge is defined, according to Kasper (1998) and Rose (1997), as the nonnative speaker's knowledge of a pragmatic system and knowledge of its appropriate use.

Since the idea of interlanguage pragmatics was introduced into language education, it has received more and more attention in language courses, such as through the notionally-based syllabus (Cohen & Olshtain, 1981). Studies have been done to investigate the relationship between language education and interlanguage pragmatic development, for example, whether grammatical development guarantees a corresponding level of pragmatic development. The results of these studies differ. Some studies (e.g. Hill, 1997; Roever, 2005; Yamashita, 1996) showed that high language proficiency participants had better performance in tests of pragmatics than low language proficiency participants in an English as second language context. On the other hand, other studies (e.g. Bardovi-Harlig & Hartford, 1991, 1993; Omar, 1991; Takahashi & Beebe, 1987) showed disparities between learners' grammatical development and pragmatic development. They reported that even learners who exhibit high levels of grammatical competence may exhibit a wide range of pragmatic competence when compared with native speakers in conversations and elicited conditions (Bardovi-Harlig & Doernyei, 1998). He and Yan (1986) investigated the pragmatic failure by Chinese learners of English as a foreign language (EFL) and found that the learners' pragmatic development was not proportional to their grammatical development.

Meanwhile, some studies have been done to investigate the teachability of pragmatic knowledge in classrooms and some (e.g. Bardovi-Harlig, 2001; Fukuya, Reeve, Gisi, & Christianson, 1998; Golato, 2003; Matsuda, 1999; Rose & Kasper, 2001) have shown that interlanguage pragmatic knowledge is teachable. The necessity and importance of teaching pragmatics have also been recognized (Eslami-Rasekh, 2005; Rose & Kasper, 2001), but still language teachers hesitate to teach pragmatics in their classrooms. Thomas (1983) notes that for the language teachers the descriptions offered by theoretical pragmaticists are inadequate. Matsuda (1999) lists two reasons for this reluctance in pragmatics teaching. First, teaching pragmatics is a difficult and sensitive issue due to the high degree of

'face threat' it often involves and, second, the number of available pedagogical resources is limited. But the reluctance should also be attributed to the lack of some valid methods for testing interlanguage pragmatic knowledge. More studies need to be done to validate methods for pragmatics assessment. The present study aims to investigate ways to assess EFL learners' pragmatic competence by addressing two questions:

1) Are the test methods used in this study reliable and valid?
2) Do learners of different EFL proficiency levels perform differently in pragmatics tests?

Ways to assess interlanguage pragmatic knowledge

Oller (1979) first introduced the notion of a pragmatic proficiency test and set two constraints for this kind of test. First, processing of language by examinees on pragmatic tests must be constrained temporally and sequentially in a way consistent with the real world occurrences of the language forms that happen to comprise test materials or speech in testing situations. This constraint could imply, for example, that encountering sentences on a reading comprehension test would require that an examinee process such sentences as meaningful sentences, rather than as just strings of words with no communicative intent. Second, such tests must use language in a way resembling natural occurrences of language outside testing contexts or formal language testing environments. The meaning of language understood or produced in pragmatic tests must link somehow to a meaningful extralinguistic context familiar to the proficiency examinee. Oller stressed the naturalness of such a test. These naturalness criteria, however, seem problematic, because they do not adequately address the artificiality of testing contexts in and of themselves, and how such artificiality constrains language use (Duran, 1984). This issue was better addressed by Clark (1978) through the notion of direct versus indirect tests of language proficiency. Clark suggested that a 'direct' test should be based on approximating, to the greatest extent possible within the necessary constraints of testing time and facilities, the specific situations in which the proficiency is called upon in real life. Clark indicated that direct proficiency tests should model everyday language use situations, but he also acknowledged that testing contexts could only approximate the real world.

Unfortunately, the field of language testing does not seem to offer much research in this respect. Not many tests to assess learners' pragmatic proficiency have been produced, though pragmatic knowledge is an indispensable part of language proficiency as defined by Bachman (1990). One of the reasons why such measures have not been readily available is that developing a measure of pragmatic competence in an EFL context is not an easy task.

So far, researchers have investigated at least six types of methods for interlanguage pragmatic assessment, i.e., the Written Discourse Completion Tasks (WDCT), Multiple-Choice Discourse Completion Tasks (MDCT), Oral Discourse Completion Tasks (ODCT), Discourse Role Play Talks (DRPT), Discourse Self-Assessment Talks (DSAT), and Role-Play self-assessments (RPSA). A summary of the practical characteristics of the six types of tests is given in Brown (2001a). All

the six measures are reviewed in detail in Yamashita (1996) and Yoshitake-Strain (1997). Brown and Hudson (1998) classified language assessment into three broad categories: selected-response assessments, constructed-response assessments, and personal-response assessments. For the sake of representativeness, in this study one test method from each of these three categories was selected: WDCT from the constructed-response type, MDCT from the selected-response type, and DSAT from the personal-response type. The following is an introduction to the forms of the three test methods used in this study.

Written discourse completion test

WDCTs are written questionnaires including a number of brief situational descriptions, followed by a short dialogue with an empty slot for the speech act under study. Participants are asked to provide a response that they think is appropriate in the given context:

> At the professor's office
> *A student has borrowed a book from her teacher, which she promised to return today. When meeting her teacher, however, she realizes that she forgot to bring it along.*
> *Teacher: Miriam, I hope you brought the book I lent you.*
> *Miriam:* _____
> *Teacher: OK, but please remember it next week.*
> (Blum-Kulka & Olshtain, 1984, p. 198)

WDCTs have evolved gradually over the past twenty years into several different modified versions which vary mainly according to the presentation forms, that is, written or oral, and existence of rejoinder. WDCTs can include a rejoinder, as in the following example from Johnston, Kasper, and Ross (1998, p. 175):

> *Your term paper is due, but you haven't finished yet. You want to ask your professor for an extension.*
> *You:* _____
> *Professor: I'm sorry, but I never allow extension.*

Or they may involve only the specification of the situation with no rejoinder, as this example from Eisenstein, Bodman, and Carpenter (1996, p.102) shows:

> *Two people who are friends are walking toward each other. They are both in a hurry to keep appointments. They see each other and say:*

In this study, this type of WDCT with no rejoinders was adopted.

Multiple-choice discourse completion test

MDCTs consist of test items where the test taker is required to choose the correct response (the key) from the several given options. Most commonly, multiple-choice items include an instruction to the test taker and a stem (typically either a phrase or sentence to be completed, or a question). The key and several

distractors then follow in random order (Davies et al., 1999). Following is a sample MDCT item:

> *You are a student. You forgot to do the assignment for your Human Resources course. When your teacher whom you have known for some years asks for your assignment, you apologize to your teacher.*
> A. *I'm sorry, but I forgot the deadline for the assignment. Can I bring it to you at the end of the day?*
> B. *Pardon me, sir, I forgot about that. Shall I do the assignment at once? So sorry! It's my fault!*
> C. *I've completed my assignment but forgot to bring it with me. I'll hand it in tomorrow.*

Discourse self-assessment test

On the DSAT, instructions are first given, followed by exponents of the functions. The participants, after reading each situation, were asked to give an overall rating of their intended performance on a five-point scale. The following is an example of the self-assessment from Hudson, Detmer & Brown (1995, p.192).

> *Situation: You and a few of your co-workers are working on a special project. You are at a meeting in the office of the project leader. As you are reaching for your briefcase you accidentally knock over the project leader's umbrella which was leaning against the desk.*
> *Rating: I think what I would say in this situation would be*
> *very unsatisfactory 1 – 2 – 3 – 4 – 5 completely appropriate*

Method

Participants

Altogether 89 Chinese EFL learners participated in this study in the final data collection stage. All of them were students from tertiary universities in Mainland China, of which the third-year students (N=31) were considered as the high language proficiency group (hereinafter called high level group) and the first-year students (N=58) were taken as the low language proficiency group (hereinafter called low level group). The ages of the participants ranged from 16 to 21 for the low level group (mean=18.79) and 19 to 24 for the high level group (mean=21.06), with an overall average of 19.61. The first-year students had studied English for about seven years, and the third-year students about 10 years. To further validate the proficiency levels of the students, a proficiency test (TOEFL, with permission granted by the Educational Testing Service) was administered to both groups. Table 1 shows the statistical analyses of the proficiency test for the two groups. From this table, we can see the two groups differ significantly in all sections of the test: listening ($t=6.04$, $p<.01$), structure ($t=4.30$, $p<.01$), reading ($t=6.38$, $p<.01$), and total ($t=8.34$, $p<.01$). This shows that the two groups were significantly different in terms of their English proficiency levels.

Table 1
Statistical description of the proficiency test (TOEFL) and t-test for the two groups

	proficiency	N	Mean	SD	*t*	*df*	Sig. (2-tailed)
Listening	High level	31	50.69	3.79			
	Low level	58	46.35	3.09	6.04	87	.000
Structure	High level	31	50.20	5.74			
	Low level	58	47.52	3.55	4.30	87	.000
Reading	High level	31	53.16	4.12			
	Low level	58	49.76	3.32	6.38	87	.000
Total	High level	31	513.49	28.20			
	Low level	58	478.80	26.11	8.34	87	.000

In addition, a questionnaire was administered to gather more background information about the participants. The questionnaire consisted of five questions about gender, age, experience in English-speaking countries, pragmatic knowledge taught in class, and self-assessment of the ability to use English. None of the participants reported any experience of having stayed in an English-speaking country. Most of the students, 97% of the high level students and 96% of the low level students, reported that their teachers only occasionally mentioned some pragmatic knowledge in class. Self-assessment of the ability to express themselves in English showed that over 80% of the high level students and about 70% of the low level students thought they were able to communicate well in English, while about 19% of the high level students and 30% of the low level students said they had some difficulty in expressing themselves well in English. A Mann-Whitney U test was conducted to see if there were any significant differences between the two proficiency groups in these two areas. The results showed that no significant difference existed ($U=3731.5$, $p=.242$) between the two groups in terms of exposure to pragmatic knowledge in class. A significant difference at the .05 level ($U=3507.5$, $p<.05$), however, was found between the two groups in terms of the self-assessment of their ability to express themselves in English, which indicated that the high-level group had significantly greater confidence in their English ability than the low-level group.

Development of test papers

The development consisted of five major stages: exemplar generation, likelihood investigation, metapragmatic assessment, WDCT pilot study, and MDCT development.

Exemplar generation

The first step in generating the test papers for this study was to obtain topics of the scenarios through a type of exemplar generation (Groves, 1996; Ostrom & Gannon, 1996; Rose & Ng, 2001; Rose & Ono, 1995). Thus, a questionnaire was

designed in which participants were given a sheet of paper illustrating a request (in Chinese). A brief training session was conducted before the students began to answer the questionnaire. The questionnaire was first explained to the students. Then, some of the students were asked to provide an example of the desired speech act. A brief follow-up discussion was held so that the students knew what they were supposed to do. Altogether 30 Chinese University students (CUSs) were asked to complete the exemplar generation questionnaire. Each student was asked to write the 10 most recently occurring events which contained the speech act of requesting. All 30 students returned their questionnaire and most wrote 10 situations. It was found that many of the nearly 300 situations generated by the students were similar; consequently only 57 situations were selected. All these situation scenarios were rewritten with their original meaning basically unchanged. Reference was also made to some existing scenarios from the literature.

Likelihood investigation

Next, a likelihood investigation was conducted for the 57 situations. This questionnaire asked the respondents to indicate on a scale of 1 to 5 the likelihood that the situations would occur in their daily life. The likelihood investigation questionnaire was written in Chinese. The questionnaire was sent to 15 students. The scales selected by the respondents were averaged for each situation. The 30 situations which got the highest mean scores were selected to form the metapragmatic assessment questionnaire.

Metapragmatic assessment

The 30 situations were reviewed. Priority was given to those situations with different combinations of features. As a result, only 24 situations were used in the metapragmatic assessment questionnaire. The questionnaire, with detailed instructions and specific examples, asked the respondents to indicate the imposition of the situation, the social distance (familiarity) between the speaker and hearer, and the power relationship (status) of the hearer and speaker (who has higher status if not equal). The questionnaire for the CUSs was written in Chinese, while that for the English native speakers (ENSs) was in English. The two versions were generated through back translation (Brislin, Lonner, & Thorndike, 1973; Fowler, 1993; Kasper & Rose, 2002; Sinaiko & Brislin, 1973).

Fifteen CUSs and 10 ENSs were asked to complete the metapragmatic assessment questionnaire. For the Chinese participants, a brief discussion was held to familiarize them with the definition of the terms 'status', 'familiarity', and 'imposition'. A 70% agreement between the CUSs and ENSs was set as the threshold of acceptance. As a result, only 18 valid items were obtained.

WDCT questionnaire pilot study

After the metapragmatic assessment, a WDCT questionnaire which contained the 18 request situations was generated. Thirty-three CUSs and eight ENSs were invited to answer these two questionnaires. Then, two ENSs were invited to rate

the responses given by the CUSs. First, the two raters were given a training manual which was based on the one developed by Hudson, Detmer, and Brown (1995). Then, the two raters were asked to rate one sample, which was followed by a discussion. Next, they were asked to give a grade (5-point scale) for the response of each item. They were reminded to ignore grammatical errors if the responses given by the students were not incomprehensible. To get some qualitative data, they were also asked to underline the parts they thought inappropriate and write briefly the reasons why they thought the responses were not appropriate. The responses given by the CUSs and the ENSs, together with the rating and comments given by the two raters, were carefully reviewed. It showed that the items basically elicited the speech act of request.

MDCT development

The data collected on the WDCT questionnaire pilot study were entered into computer and analyzed with the software WinMax (Kuchartz, 1998). The responses provided by the CUSs and ENSs and the comments given by the two raters were reviewed and compared. Those selected from the ENSs were coded as 'key', while those from the CUSs which were marked as inappropriate were coded as 'distractor'. From these data, four to eight responses (one to two supposed keys and three to seven presumed distractors) were selected for each item. These responses, together with the situation description, formed the MDCT questionnaire. Grammatical errors, if any, in the responses given by the CUSs were corrected. The MDCT questionnaire was then given to 10 ENSs who were asked to decide if the responses for each situation were appropriate and to provide brief reasons for those which they thought inappropriate. Based on the data collected from the 10 ENSs, the three options which had the highest agreement among the respondents were selected for each situation. The one from the ENSs which enjoyed the highest agreement from the 10 ENSs on their appropriateness was taken as the key, while the other two from the CUSs which had the highest agreement among the 10 ENSs on their inappropriateness were taken as the distractors. To avoid excessive variation of the length of the options, options which had high agreement and were similar in length were given first priority. However, it was found that some situations failed to have options with high agreement. These situations were then excluded. As a result, 12 request situations were generated.

Next, a new MDCT questionnaire was designed with these 12 situations and their corresponding options. This MDCT questionnaire was given to 5 ENSs who were asked to select the most appropriate response for each situation. Their responses showed unanimous agreement on the keys for 12 situations.

In the next step, 31 CUSs were invited to answer the draft MDCT questionnaire. The questionnaire was also used for the think-aloud protocol with two participants. The verbal protocols were recorded, and analyses using item response theory (IRT) were conducted for the paper-and-pencil questionnaire. The think-aloud data were also analyzed to see if the students were able to determine the intended speech act, and to examine how the distractors functioned. Corresponding revisions were made according to the results of the IRT analyses.

The analyses of the verbal protocols showed that all the items tested the right speech act they were intended to do.

Finally, three test papers using these three test methods (i.e., DSAT, WDCT, and MDCT) were generated. All three test papers contained the same 12 situation scenarios, but with different test methods.

Administrative procedures

The tests were administered to 89 university students in three sessions. A proficiency test (TOEFL) was first administered. Two weeks later, the other tests were administered to the two groups (i.e., the high-level group and the low-level group) in turn in a classroom. The DSAT was administered first, followed by the WDCT and then the MDCT. The whole test session took about two hours.

Scoring

To avoid any effect on ratings due to poor handwriting, the answers to the WDCT test paper given by the test takers were entered into the computer without any changes. The typewritten scripts were ordered alphabetically according to the test takers' surnames and then presented to two ENS raters using the rubrics developed by Hudson, Detmer, and Brown (1995). The raters were given clear directions as to how the test papers should be rated and had a preliminary training on the rating. The final scores of the WDCT test were the mean scores of the two raters. For the MDCT, one correct answer equaled to five points while a wrong answer got 0 point. For the DSAT, the test takers' self-rating was the final score.

Results

General test characteristics

Statistical characteristics of the test methods

The descriptive statistics for the test methods used in this study and the participants' TOEFL scores are shown in Table 2, including the mean, standard deviation, minimum, and maximum of the scores.

Table 2
Descriptive statistics for all the participants

	N	Score*	Minimum	Maximum	Mean	SD
TOEFL	89	677	413.33	586.67	490.93	31.50
DSAT	89	60	28.00	55.00	43.47	5.98
MDCT	89	60	5.00	55.00	35.69	9.84
WDCT	89	60	15.00	45.50	29.22	6.45

*Score = total possible score

The WDCT yielded lower mean scores than the MDCT, and it was the DSAT that produced the highest mean scores. This shows that the WDCT test paper was more difficult than the MDCT test paper and that the participants overestimated their pragmatic ability to some degree.

Internal consistency reliability

The internal consistency (Cronbach's alpha) of each test was estimated. Then, the internal consistency reliability for ratings on the WDCT was examined. The correlation between the two raters was also computed to further show the interrater reliability. Table 3 displays the results of the estimates and the standard error of measurement for each test method. The internal consistency reliability estimates for all the test methods were basically satisfactory (all were above .86).

The internal consistency reliability estimates for the two raters were acceptable (.87 for Rater 1 and .75 for Rater 2). This suggests that there was a considerable amount of consistency in assigning scores to the examinees' performance, although there existed some disagreement between the two independent raters. The overall interrater reliability based on the Spearman-Brown Prophecy formula was .903. Meanwhile, the interrater reliability can also be estimated by examining how much the rater's scores correlate with each other (Yamashita, 1996). The correlation ($r=.82$, $p<.01$) indicated that the two raters were significantly correlated.

Correlational analyses

Correlational evidence was collected to examine the relationship among the items and test methods. Table 4 displays the correlations between different tests. The TOEFL test was not significantly correlated with other tests except the MDCT. The three pragmatics tests correlated significantly with each other at $p<.01$. Statistical significance is a necessary precondition for a meaningful correlation, but it is not sufficient unto itself (Brown, 1996). So, coefficients of determination for these correlations were computed. The results as displayed in Table 5 showed that the amount of variation in one test that is accounted for by the test it is correlated with was below 50%. This means there were still over 50% unaccounted variance between these tests. Other factors such as test method may explain some of the unaccounted variance between the tests.

Table 3
Internal consistency reliability (α) of each test method

	No. of cases	r(α)	SEM
DSAT	89	.9226	2.26
MDCT	89	.8647	4.50
WDCT	89	.9021	3.23

Table 4
Correlations between different tests

	TOEFL	WDCT	DSAT
TOEFL	1.000		
WDCT	.048	1.000	
DSAT	.033	.685**	1.000
MDCT	.226**	.675**	.517**

** Correlation is significant at the 0.01 level (2-tailed).

Table 5
Coefficients of determination between tests

	TOEFL	WDCT	DSAT
TOEFL			
WDCT	.002		
DSAT	.001	.469	
MDCT	.050	.456	.267

Factor analysis

Exploratory factor analysis was conducted to study the construct validity of the test methods. If a series of tests is administered to a group of students and those tests that logically should be related turn out to load on the same factor, while tests that would logically be less related load on different factors, the analysis can be used to argue for convergent validity and divergent validity (Brown, 2001b). Factor analysis was conducted for the proficiency test and the pragmatics tests. After varimax rotation, two factors which had an eigenvalue of 1.00 or higher were extracted. The loadings shown in Table 6 indicated that the pragmatic ability (in WDCT, DSAT, and MDCT) were fairly highly correlated with Factor 1 (at .918, .857, and .815, respectively), while the English proficiency correlated at .985 with Factor 2. The communalities (h^2) indicated that the proportion of variance accounted for in the proficiency test scores was .971 (or 97.1%), and the total variances accounted for in WDCT, DSAT, and MDCT were 84.2%, 74.0%, and 74.4%, respectively. The figures at the bottom of the table indicated that the proportions of variance accounted for by Factor 1 and Factor 2 in this validity study were 56.1% and 26.3% respectively. The total variance accounted for was 82.4%. The analysis showed that Factor 1 appeared to be interlanguage pragmatic knowledge factor, while Factor 2 might be a language proficiency factor. The general pattern revealed in these analyses is that the subtests of the pragmatics tests load together (supporting convergent validity), while the pragmatics tests and the proficiency test load most heavily on different factors (supporting divergent validity).

Another factor analysis was conducted for the three pragmatics tests. Results

Table 6
Results of principle component analysis for test methods

	Components		
	1	2	h^2
TOEFL	.041	.985	.971
WDCT	.918	-.004	.842
DSAT	.857	-.068	.740
MDCT	.815	.281	.744
Proportion of variance	.561	.263	Total: .824

Extraction Method: Principle Component Analysis
Rotation Method: Varimax with Kaiser Normalization.

Table 7
Results of principle component analysis for the three pragmatics tests

	Components	
	1	h^2
WDCT	.933	.870
DSAT	.881	.777
MDCT	.917	.841
Proportion of variance		.829

Extraction Method: Principle Component Analysis

showed (Table 7) that only one factor had an eigenvalue of 1.00 or higher; thus, only one factor was extracted. This indicated that all the three pragmatics tests might tap a similar construct.

Group differences

ANOVA tests were conducted to explore the differences between the two proficiency groups. Table 8 displays the descriptive statistics. For the three speech act tests, both the high-level group and the low-level group got the highest mean on the DSAT. The high level group did better than the low level group on the MDCT, but the low level group got slightly higher scores than the high level group on the DSAT and the WDCT. One-way ANOVA tests were conducted to determine the significance of the differences. The descriptive statistics for different proficiency groups are displayed in Table 8. The results of the ANOVA as displayed in Table 9 showed a significant difference on the TOEFL ($F(1,87)=74.676$, $p< .01$) and the MDCT ($F(1,87)=5.65$, $p< .05$), but not on the WDCT ($F(1,87)=.606$, $p=.437$) and the DSAT ($F(1,87)=.023$, $p=.878$) at the .05 level. This indicated that the two groups differed significantly in terms of their English proficiency, but not on two of the three pragmatics tests (DSAT and WDCT). The difference

Table 8
Descriptive statistics for different groups

	Group	n	Mean	SD	SE
DSAT	Low level	58	43.50	5.52	.51
	High level	31	43.41	6.79	.85
MDCT	Low level	58	34.08	9.35	.92
	High level	31	38.67	9.79	.96
WDCT	Low level	58	29.42	6.15	.56
	High level	31	28.85	6.97	.87

Table 9
ANOVA of group difference in terms of proficiency levels

	F	Sig.
TOEFL	74.676	.000
WDCT	.606	.437
DSAT	.023	.878
MDCT	5.650	.018

on the MDCT was perhaps due to a method effect, which will be discussed later. The results indicated that the test takers' interlanguage pragmatic knowledge did not seem to increase substantially with their language proficiency.

Discussion and conclusion

Reliability and validity can be viewed as complementary aspects of validation process (Bachman, 1990). The Cronbach alpha reliability estimates for WDCT and DSAT were satisfactory at around .90, while that for MDCT was .86. This is in line with previous studies (Enochs & Yoshitake-Strain, 1999; Roever, 2005; Yamashita, 1996) which showed that WDCT and DSAT had high reliability. Nevertheless, it is noteworthy that the internal consistency reliability for the MDCT in this study was acceptably high at .86. This might be due to the procedures involved in the development of the test paper. The scenarios and options of the MDCT test paper in this study, instead of adopting established ones, were independently developed in several stages based on the Chinese context, including exemplar generation, likelihood investigation, metapragmatic assessment, and verbal protocol analysis. All the situations were closely related to the participants' life, and the distractors were generated by the participants. However, although the MDCT test paper developed in this study worked well for the Chinese context, it is not clear whether it would work equally well in other contexts, or with different participant groups. More research is needed.

The raters' Cronbach alpha internal consistency reliability estimates were reasonably high at .75 and .87. Interrater reliability was high at around .90, and

the interrater correlation coefficients were considerably high at .82 ($p<.001$). Therefore, we can say both raters were highly self-consistent in scoring. Further qualitative research, as suggested by Kondo-Brown (2002) and Gyagenda and Engelhard (1998), may help to investigate the in-depth characteristics of performances and the raters' decision-making processes at the time the ratings were done.

The three test methods investigated in this study significantly correlated with each other at the .01 level, and the coefficients of determination showed that the joint variance between the test methods ranged from 26.7% to 46.9%. The overlap was not very high, but reasonably strong to assume that the three test methods might measure a basically similar construct, which means they essentially tapped a similar kind of knowledge, that is, interlanguage pragmatic knowledge. The appreciably high overlapping variance (46.9%) between the DSAT and the WDCT was perhaps because of the similar cognitive processes involved.

The MDCT in this study had a quite high reliability ($r=.86$). This is different from the findings of Yamashita (1996) and Enochs & Yoshitake-Strain (1999) which revealed quite low reliability and validity of the MDCT. Correlation analyses indicated that it tapped the interlanguage pragmatic knowledge it intended to measure. Though Yamashita (1996) and Enochs and Yoshitake-Strain (1999) avoided the investigation of why their MDCT had low reliability and validity, their use of the prototypic tests developed by Hudson, Detmer, and Brown (1992; 1995) might be one of the causes (Hudson, 2001). Yamashita (1996) also reported that participants in her study commented that some of the situations were not relevant or appropriate in their context. The MDCT test paper in this study came from a series of investigations within the group for which this test was intended. Situations were directly from the students, and situational sociopragmatic variables were investigated through metapragmatic assessment. No report of unfamiliarity was received from the students. Therefore, the students' familiarity with situation and variables may affect their performance in such tests.

It is noteworthy that the WDCT and the DSAT had no significant correlation with the TOEFL. Though a significant correlation was found for the MDCT and the TOEFL, the overlapping variance between the two was only 5%, which was not strong enough to indicate that the two tapped similar knowledge. Thus the TOEFL and the pragmatics tests may have measured different constructs. This was further confirmed by the factor analysis, which clearly identified the TOEFL as a distinct factor from the other three tests which were shown to belong to the same factor in another factor analysis. Discrimination would also be indicated by low or zero correlations between measures of different traits using different test methods (Bachman, 1990, p. 263). The low correlation between the TOEFL and other pragmatics tests demonstrated good discrimination of the tests. High correlations between different traits with the same method indicate a method effect (Roever, 2005). Although the correlation between MDCT and TOEFL was not high, it was significant. This indicated some kind of method effect. In China, children are trained to do multiple-choice questions early in the primary school. As they grow, they become more and more skillful in dealing with multiple-choice questions. The significant correlation between the MDCT and the TOEFL can be attributed to their ability in tackling such questions.

The two groups of participants were significantly different ($p<.01$) in terms of their English language proficiency, whereas this study revealed that they were not significantly different in the tests of pragmatics (WDCT and DSAT). However, the two proficiency groups were significantly different at the .05 level on the MDCT. This difference might result from the effect of the test method. The results indicated that the participants of higher grammatical proficiency did not necessarily possess higher concomitant pragmatic competence. This differed from some previous studies (e.g. Hill, 1997; Roever, 2005; Yamashita, 1996) which showed that high language proficiency participants had better performance in tests of pragmatics than low language proficiency participants. There are two possible reasons for this difference. First, the participants in those studies had different degrees of experience in living in an English-speaking country. They had direct exposure to the target culture. However, the students in this study had no such experience or direct exposure. They were exposed to the target culture only through the classroom. An investigation also demonstrated no significant difference between the two proficiency groups in terms of pragmatic knowledge teaching in class. Second, the participants in those studies were rather diverse and heterogeneous; the students in this study, however, comprised only university students who had a similar educational background. They were divided into two proficiency groups according to their scores on a TOEFL test conducted just before the data collection. In fact, the English proficiency of the low-level group was not low at all, though significantly lower than the high-level group.

Exposure to the target language is shown to affect the development of EFL learners' interlanguage pragmatic knowledge. One way of remedying this lack of direct exposure to the target culture and society may be through teaching pragmatics. Also, the insignificant difference between the two proficiency groups indicated that pedagogical measure should be taken to enhance the development of the EFL learners' interlanguage pragmatic knowledge. However, investigation in this study showed that teachers seldom, if ever, taught pragmatic knowledge in class. The development of pragmatic competence, according to Ellis (1994), depends on providing learners with sufficient and appropriate input. Input in the EFL classroom comes mainly through teacher talk and instructional materials (Hill, 1997). However, foreign language teaching in Chinese universities is conducted mainly in a traditional way in the classroom, that is, teacher-centered teaching. Though even teacher-fronted classroom discourse offers some opportunities for pragmatic learning (Kasper, 1997), and communicative teaching is receiving more and more attention in China nowadays, both ways still seem problematic in the Chinese EFL classroom.

First, the majority of the EFL teachers in Chinese universities are non-native speakers of English; thus, they cannot draw on native speaker (NS) intuitions (Rose, 1994) and cannot serve as direct models for the students (Bardovi-Harlig & Hartford, 1996). Feeling the lack of NS intuition also makes EFL teachers reluctant to teach pragmatics in classroom. Second, insufficient instructional materials impede the move towards teaching pragmatics by EFL teachers. Being NNSs of English, EFL teachers have difficulty in determining the appropriate materials for teaching pragmatics. Although evidence of speech acts in textbooks is plentiful, as pointed out by Bardovi-Harlig and Hartford (1996), it has been

given very little attention. Therefore, Rose (1994, p. 155) notes that "if pragmatic competence is to be dealt with successfully in EFL settings, methods and materials must be developed which do not assume or depend on the NS intuitions of the teacher."

Third, the lack of instructional methods also prevents EFL teachers from teaching pragmatic knowledge in class. Two types of activities have been proposed for pragmatic knowledge instruction: activities aiming at raising students' pragmatic awareness, and activities offering opportunities for communicative practice (Kasper, 1997). Rose (1994) suggests pragmatic consciousness-raising in EFL teaching and comments that if the learner's pragmatic consciousness is raised, he or she will more easily notice pragmatic features of the input and this may lead to the acquisition of pragmatic knowledge. He also points out that in order for EFL learners to benefit from this type of consciousness-raising they should be given ample supplies of authentic input, e.g. through videos and movies. On the other hand, practicing EFL learners' pragmatic abilities, however, requires student-centered interaction. Activities which engage students in different social roles and speech events, such as role play, simulation, and drama, provide opportunities to practice the wide range of pragmatic and sociolinguistic abilities that the students need in interpersonal encounters outside the classroom (Kasper, 1997). Owing to the factors which limit the use of these two types of activities, more practical activities for pragmatics instruction are expected.

Fourth, if pragmatic knowledge is included in the teaching syllabus, it needs to be incorporated into tests. However, no established tests of this kind are available now. Though some studies (e.g. Hudson, Detmer, & Brown, 1995; Liu, 2006; Yamashita, 1996; Yoshitake-Strain, 1997; Roever, 2005) examined the possibilities of such tests, at this time, as pointed out by Hudson (2001, p. 297), the instruments should be used for research purposes only, and no examinee level decisions should be made in pedagogical settings. More research, especially validation studies, is necessary.

This study has some implications for ILP test development, too. A fundamental concern, according to Roever (2005), in constructing items for tests of pragmalinguistic knowledge is that they be representative of real-world language use, and not just based on test designers' intuition, which may or may not be an accurate reflection of reality. Ethnographic studies of real world language use and targeted elicitation of response plausibility in pilot study with NSs are suggested (Roever, 2005). Ethnographic field study is a useful procedure, but the extremely expensive endeavors it demands precludes it from being a widely adopted method. Inconsistency might be found between elicitation through NSs and that through NNSs (Yamashita, 1996). No such inconsistency was detected for the scenarios generated for this study. This would suggest that a combination of elicitation through both NSs and NNSs is a better and more practical way to construct pragmalinguistic test items.

This study also has some implications for the large-scale proficiency tests practiced nowadays around the world, like TOEFL and IELTS. The traditional paper-and-pencil TOEFL was found not to correlate with pragmatics tests. Test takers' English proficiency as shown on the TOEFL scores does not seem to be consistent with their interlanguage pragmatic ability. This study also showed

that students with high TOEFL scores do not seem to have correspondingly high interlanguage pragmatic ability. Therefore, it is no wonder that some students who have gained over 670 points in a traditional paper-and-pencil TOEFL test cannot communicate well in English. Hence, it is quite necessary to teach pragmatic knowledge in the classroom and include pragmatic knowledge in large-scale tests. It is good to see that the new TOEFL iBT test (Educational Testing Service, 2005) integrates all four language skills including pragmatic competence into the test.

Acknowledgements

The study has been supported by the MOE Project of the Center for Linguistics and Applied Linguistics of Guangdong University of Foreign Studies. I would like to thank Dr Kenneth Rose for his constructive advice and comments. Thanks also go to the editors and the two anonymous RELT reviewers for their valuable feedback about and editing of the earlier drafts of the article.

References

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

Bardovi-Harlig, K. (2001). Evaluating the empirical evidence: Grounds for instruction in pragmatics. In K. R. Rose & G. Kasper (Eds.), *Pragmatics in language teaching* (pp. 13-32). New York: Cambridge University Press.

Bardovi-Harlig, K. & Doernyei, Z. (1998). Do language learners recognize pragmatic violations? Pragmatic vs. grammatical awareness in instructed L2 learning. *TESOL Quarterly, 32*, 233-259.

Bardovi-Harlig, K. & Hartford, B. (1991). Saying 'no' in English: Native and non-native rejections. In L. F. Bouton & Y. Kachru (Eds.), *Pragmatics and language learning. Monograph Series 2* (pp. 41-58). Urbana, IL: University of Illinois.

Bardovi-Harlig, K. & Hartford, B. (1993). Learning the rules of academic talk: A longitudinal study of pragmatic change. *Studies of Second Language Acquisition, 15*, 279-304.

Bardovi-Harlig, K. & Hartford, B. (1996). Input in an institutional setting. *Studies of Second Language Acquisition, 18*, 171-188.

Blum-Kulka, S. & Olshtain, E. (1984). Requests and apologies: A cross-cultural study of speech act realization patterns (CCSARP). *Applied Linguistics, 5*(3), 197-213.

Brislin, R. W., Lonner, W. J., & Thorndike, R. M. (1973). *Cross-cultural research methods*. New York: John Wiley.

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

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

Brown, J. D. (2001b). What is an eigenvalue? Retrieved August 3, 2004, from http://www.jalt.org/test/bro_10.htm

Brown, J. D. & Hudson, T. (1998). The alternatives in language assessment. *TESOL Quarterly, 32*(4), 653-675.

Clark, J. D. (1978). Psychometric considerations in language proficiency testing. In B. Spolsky (Ed.), *Approaches to language testing*. Arlington: Center for Applied Linguistics.

Cohen, A. D. & Olshtain, E. (1981). Developing a measure of sociocultural competence: the case of apology. *Language Learning, 31*(1), 113-134.

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., & McNamara, T. (1999). *Dictionary of language testing*. Cambridge: Cambridge University Press.

Duran, R. P. (1984). Some implications of communicative competence research for integrative proficiency testing. In C. Rivera (Ed.), *Communicative competence approaches to language proficiency assessment: Research and application* (pp. 44-58). Clevedon: Multilingual Matters.

Educational Testing Service (2005). *TOEFL iBT Tips*. Retrieved February 8, 2006, from http://upload.mcgill.ca/applying/TOEFL_Tips.pdf.

Eisenstein, M., Bodman, J., & Carpenter, M. (1996). Cross-cultural realization of greeting in American English. In J. Neu (Ed.), *Speech acts across cultures* (pp. 89-107). Berlin: Mouton de Gruyter.

Ellis, R. (1994). *The study of second language acquisition*. Oxford: Oxford 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. (2005). Raising the pragmatic awareness of language learners. *ELT Journal, 59*(3), 199-208.

Fowler, J. F., Jr. (1993). *Survey research methods* (2nd ed.). Newbury Park, CA: Sage Publications.

Fukuya, Y., Reeve, M., Gisi, J., & Christianson, M. (1998). *Does focus on form work for teaching sociopragmatics?* Paper presented at the annual meeting of the International Conference on Pragmatics and Language Learning, Urbana, IL.

Golato, A. (2003). Studying compliment responses: A comparison of DCTs and recordings of naturally occurring talk. *Applied Linguistics, 24*(1), 90-121.

Groves, R. (1996). How do we know what we think they think is really what they think? In N. Schwarz & S. Sudman (Eds.), *Answering questions: Methodology for determining cognitive and communicative processes in survey research* (pp. 389-402). San Francisco: Jossey-Bass.

Gyagenda, I. S., & Engelhard, G., Jr. (1998). *Applying the Rasch model to explore rater influences on the assessed quality of students' writing ability*. Paper presented at the Annual Meeting of the American Educational Research Association, San Diego.

He, Z. & Yan, Z. (1986). Pragmatic failure by Chinese EFL learners. *Foreign Language Teaching and Research, 3*, 52-57.

Hill, T. (1997). *The development of pragmatic competence in an EFL context*. Unpublished doctoral dissertation, Temple University, Tokyo.

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

Hudson, T., Detmer, E., & Brown, J. D. (1992). *A framework for testing cross-cultural pragmatics*. Honolulu: Second Language Teaching & Curriculum Center University of Hawai'i at Manoa.

Hudson, T., Detmer, E., & Brown, J. D. (1995). *Developing Prototypic Measures of Cross-cultural Pragmatics*. Honolulu: Second Language Teaching and Curriculum Center, University of Hawaii at Manoa.

Johnston, B., Kasper, G., & Ross, S. (1998). Effect of rejoinders in production questionnaires. *Applied Linguistics, 19*(2), 157-182.

Kasper, G. (1997). Can pragmatic competence be taught? Retrieved August 10, 1999, from http://www.lll.hawaii.edu/nflrc/NetWorks/NW6/default.html

Kasper, G. (1998). Interlanguage pragmatics. In H. Byrnes (Ed.), *Learning foreign and second languages: Perspectives in research and scholarship* (pp. 183-208). New York: The Modern Language Association of America.

Kasper, G., & Blum-Kulka, S. (1993). Interlanguage pragmatics: An introduction. In G. Kasper & S. Blum-Kulka (Eds.), *Interlanguage pragmatics* (pp. 3-17). Oxford: Oxford University Press.

Kasper, G. & Rose, K. R. (2002). *Pragmatic development in a second language*. Michigan: Blackwell.

Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing, 19*(1), 3-31.

Kuchartz, U. (1998). *WinMAX: Scientific text analysis for the social sciences*. Berlin: BBS.

Liu, J. (2006). *Measuring interlanguage pragmatic knowledge of EFL learners*. Frankfurt am Main: Peter Lang.

Matsuda, M. (1999). Interlanguage pragmatics: What can it offer to language teachers? *The CATESOL Journal, 11*(1), 39-59.

Oller, J. W. (1979). *Language tests at school: A pragmatic approach*. London: Longman.

Omar, A. S. (1991). How learners greet in Kiswahili. In L. Bouton & Y. Kachru (Eds.), *Pragmatics and language learning (Vol. 2)* (pp. 59-73). Urbana-Champaign: University of Illinois.

Ostrom, T. & Gannon, K. (1996). Exemplar generation: Assessing how respondents give meaning to rating scales. In N. Schwarz & S. Sudman (Eds.), *Answering questions: Methodology for determining cognitive and communicative processes in survey research* (pp. 293-318). San Francisco: Jossey-Bass.

Roever, C. (2005). *Testing ESL pragmatics: Development and validation of a web-based assessment battery*. Frankfurt am Man: Peter Lang.

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

Rose, K. R. (1997). Pragmatics in the classroom: Theoretical concerns and practical possibilities. In L. F. Bouton (Ed.), *Pragmatics and language learning. Monograph Series (Vol. 8)* (pp. 267-295). Urbana-Champaign: University of Illinois.

Rose, K. R. & Kasper, G. (Eds.). (2001). *Pragmatics in language teaching*. New York: Cambridge University Press.

Rose, K. R. & Ng, K.-f. C. (2001). Inductive and deductive teaching of compliments and compliment responses. In K. R. Rose & G. Kasper (Eds.), *Pragmatics in language teaching* (pp. 145-170). New York: Cambridge University Press.

Rose, K. R. & Ono, R. (1995). Eliciting speech act data in Japanese: The effect of questionnaire type. *Language Learning, 45*(2), 191-223.

Sinaiko, H. W. & Brislin, R. W. (1973). Evaluating language translations: Experiments on three assessment methods. *Journal of Applied Psychology, 57*(3), 328-334.

Takahashi, T. & Beebe, L. M. (1987). The development of pragmatic competence by Japanese learners of English. *JALT Journal, 8*, 131-155.

Thomas, J. (1983). Cross-cultural pragmatic failure. *Applied Linguistics, 4*(2), 91-112.

Yamashita, S. O. (1996). *Six measures of JSL pragmatics*. Honolulu: Second Language Teaching & Curriculum Center of University of Hawaii at Manoa.

Yoshitake-Strain, S. (1997). *Measuring interlanguage pragmatic competence of Japanese students of English as a foreign language: A multi-test framework evaluation*. Unpublished doctoral dissertation, Columbia Pacific University, Novata, CA.

Appendix: Situation scenarios and sample test items[1]

Scenarios:

1. You are trying to study in your room and you hear loud music coming from another student's room down the hall. You don't know the student, but you decide to ask him to turn the music down.
2. You are now shopping in a department store. You see a beautiful suit and want to see it. You ask the salesperson to show you the suit.
3. You are now discussing your assignment with your teacher. Your teacher speaks very fast. You do not follow what he is saying, so you want to ask your teacher to say it again.
4. Your computer is down because of a virus. One of your teachers is very skillful in fixing computers. You know he has been very busy recently, but you still want to ask him to fix your computer.
5. You are a teacher. In class, the mobile phone of one of your students rings. You ask your student to turn off his mobile phone.
6. You are watching a basketball game. A student you don't know comes and stands just in front of you blocking your view. You want to ask the student not to block your view.
7. You are applying for a new job in a small company and want to make an appointment for an interview. You know the manager is very busy and only schedules interviews in the afternoon from one to four o'clock on Wednesday. However, you have to take the final-term exam this Wednesday. You want to schedule an interview on Thursday.
8. You are the owner of a bookstore. Your shop clerk has worked for a year, and you have gotten to know him/her quite well. It is the beginning of the semester, and you are very busy selling and refunding textbooks all day. Today you have a plan to extend business hours by an hour, though you know the clerk has worked long hours in the past few days. You ask the clerk to stay after store hours.
9. For the first time this semester, you are taking a mathematics course. You have had a hard time following lectures and understanding the textbook. A test is scheduled to be held next week. You notice that one student sitting next to you seems to have a good background knowledge of math, and is doing well. Since it is the beginning of the semester, you do not know him/her yet. You want to ask him/her to study together for the upcoming test.
10. Something is wrong with your computer, but you have to finish some homework which is due tomorrow. Your roommate has a computer, but he is also writing a course paper on his computer. His homework is due the day

[1] Refer to Liu (2006) for the complete test items.

after tomorrow. You want to ask him to stop his work and let you use his computer to finish your homework first.

11. You are writing your graduate thesis and need to interview the president of your university. The president was your teacher and you know him quite well. You know the president is very busy and has a very tight schedule. You still want to ask the president to spare one or two hours for your interview.
12. You are the manager of a company. You are in a meeting with the other members of your company. You need to write some notes, but realize you do not have any paper. You turn to the person sitting next to you. You know the person very well.

Sample WDCT test item:

You are now shopping in a department store. You see a beautiful suit and want to see it. You ask the salesperson to show you the suit.
You: _____

Sample MDCT test item:

You are now shopping in a department store. You see a beautiful suit and want to see it. You ask the salesperson to show you the suit.
A. Oh, sorry, could you pass that suit to me to have a look? I want to buy it.
B. Lady, I'd like to have a look at that suit. Would you please do me a favor?
C. Excuse me. Could you show me this suit please?

Sample DSAT test item:

You are now shopping in a department store. You see a beautiful suit and want to see it. You ask the salesperson to show you the suit.

Rating: I think what I would say in this situation would be
very unsatisfactory 1 – 2 – 3 – 4 – 5 completely appropriate