ISSN 1798-4769
Journal of Language Teaching and Research, Vol. 6, No. 6, pp. 1206-1215, November 2015
DOI: http://dx.doi.org/10.17507/jltr.0606.07

Reliability and Validity of WDCT in Testing Interlanguage Pragmatic Competence for EFL Learners

Lan Xu
School of Foreign Languages, Institute of Social Technology, Suranaree University of Technology, Thailand

Anchalee Wannaruk
School of Foreign Languages, Institute of Social Technology, Suranaree University of Technology, Thailand

Abstract—**Interlanguage pragmatic competence is of vital importance for the EFL learners because misunderstanding always occurs among people from different cultures. The present study aimed to develop an interlanguage pragmatic competence test in the field of speech acts with WDCT. Altogether 100 English major students and 33 native speakers in Guizhou University of China participated in the developing of the test, and another 60 English majors in Guizhou University of China took the test. The analysis of the reliability and validity of WDCT was based on Many Facets Rasch Model. The results showed that WDCT had both high reliability and validity in the Chinese context in testing the interlanguage pragmatic competence in speech acts performance.**

Index Terms—**speech acts, WDCT, reliability, validity, EFL learners**

I. Introduction

Interlanguage pragmatics (ILP) is the study of language learners' comprehension, production and acquisition of linguistic action in a second language (Kasper, 1998). Pragmatic competence is an indispensable component of overall language competence. It is the ability to use available linguistic resources (pragmalinguistics) in a contextually appropriate fashion (sociopragmatics), that is, how to do things appropriately with words (Kasper & Rose 1999). It is the appropriateness in communication, which includes all kinds of knowledge needed in discourses and based on context (He & Chen 2004). Si (2001) states that for Chinese EFL learners, pragmatic competence includes the following three aspects: pragmalinguistic ability, sociopragmatic ability and the awareness of the difference between English and Chinese.

Misunderstanding is a central issue in interlanguage pragmatics, which may occur between people from different cultural backgrounds. According to the National Language Research Institute (Shinpro 'Nihongo' Dai 2-han 1999a, 1999b), speakers of different languages and with different cultural backgrounds interpret pragmatic behaviors differently. Nishihara (1999) claims that the pragmatic standards for a country or culture will not be universally accepted. Thus, when we conduct an intercultural or international research, we need to be cautious to avoid overgeneralizing our own beliefs. Misunderstanding in communication between EFL learners and native speakers can naturally occur frequently due to the learner's weak understanding of the target culture.

In China, for many students, the purpose of learning English is to pass all kinds of English examinations. They memorize a large number of words, grasp enough grammatical knowledge and do reading and listening and writing exercises frequently for gaining high scores, but speaking is not included in most national tests for university students in China. Verbal communication in English is their weak point, even with the English majors. On the one hand, the students regard communication is nothing important for their scores; on the other hand, Chinese teachers often ignore the students' errors in their speaking, so some non-standard or non-habitual utterances of the students can be with them for many years. It is not an uncommon phenomenon that an English learner in China can get over 600 points in Test of English as a Foreign Language (TOFEL) and over 2000 in Graduate Record Examination (GRE) but does not know how to make a simple speech act in English in real communication (Liu 2004).

However, the studies on ILP competence testing are still on their initial stage, and there is no exception in China (Ma 2010). Up to now, no comprehensive testing of ILP in speech acts has been found. Most researchers concentrate on the reliability and validity of different kinds of testing methods with very limited speech acts, such as request, refusal and apology (Hudson 2001a, 2001b, Yamashita 1996a, 1996b, Yoshitake 1997, Liu 2007, Brown 2001, 2008, Roever 2010), advice (Hinkel 1997). Thus, it is urgent to design reliable and valid measurements for a wider scope of ILP competence testing. The present study aims to make some contribution in this field and hopes it will be helpful for both the teachers and learners in developing the ILP competence level in English.

Written discourse completion task (WDCT) is a pragmatics instrument that requires the learners to read a written description of a situation (including such factors as setting, participant roles, and degree of imposition) and asks them to write what they would say in that situation. It is a valid instrument in measuring speech acts performance and it is widely used in this field. It is easy to replicate because of their simplicity of use and high degree of variable control (Golato 2003). There are plenty of advantages of WDCT: it elicits more authentic language, it is easy to transcribe and administer because of paper and pencil, and it is time saving to collect a large amount of data. While the disadvantages of WDCT are also obvious: it is difficult to conduct because it requires recruiting, training, scheduling, and paying raters, it is time consuming for scoring and it collects written receptive and productive language only. Despite of all the disadvantages of WDCT, a number of researchers have applied this method in the studies of speech acts in the past 30 years (Blum-Kulka 1982, 1983, Blum-Kulka & Olshtain 1986, Cohen, Olshtain & Rosenstein 1986, House & Kasper 1987, Olshtain & Weinbach 1987, Takahashi & Beebe 1987, 1993, House 1989, Kasper 1989, Rose 1992, 1994a, Rose & Ono 1995, Johnston, Kasper & Ross 1998, Liu 2006a, 2006b, Fauzul 2013).

To design a test in ILP competence, reliability and validity are the two most important factors needed to be taken into consideration. Reliability refers to the consistency of the scores obtained--how consistent they are for each individual from one administration of an instrument to another and from one set of items to another (Subong 2006). Validity is the degree to which an assessment measures what it is supposed to measure (Garrett 1937). The following table is a summary of the major findings in the reliability of WDCT of pragmatic competence testing in speech acts.

TABLE 1.
RELIABILITY ESTIMATES FOR PREVIOUS TESTING PROJECTS

Researcher(s)	Year of study	Statistic measures	WDCT
Yamashita	1996a	K-R21	.87
	1996b	Alpha	.99
Yoshitake	1997	K-R21	.50
Hudson	2001a	Alpha	.86
Liu	2004	Alpha	.95
Duan	2012	Alpha	.74

It can be seen from the above table that most researchers show acceptable reliabilities in WDCT (Yamashita 1996a, 1996b, Hudson 2001a, Liu 2004, Duan 2012) except Yoshitake (1997) in testing ILP competence in speech acts.

With regard to the validity of WDCT, some researchers found that WDCT was a valid measure to test ILP competence in the field of speech acts. Hudson, Detmer & Brown (1995) found that WDCT was with high validity in assessing pragmatic competence of EFL learners after comparing six testing instruments, namely written discourse completion task, oral discourse completion task, multiple-choice discourse completion task, discourse role play task, self-assessment and role-play self-assessment. Yamashita (1996a, 1996b) applied the same six instruments to test the Japanese as the second language learners' pragmatic competence and she also concluded that WDCT was a valid measure. Ahn (2005) examined all the above instruments excluding multiple-choice discourse completion task in conducting speech acts for Korean as the second language (KSL) learners, and the results showed that WDCT was valid in the KSL context. Hinkel (1997) found DCTs in general might be very valid in eliciting data of ILP performance. However, some other researchers drew different conclusions. Rose (1994) and Rose & Ono (1995) found that WDCT may not be valid for collecting data for ILP competence of speech acts in Japanese context. Thus, further investigation is needed for the validity of WDCT in different context.

II. RESEARCH METHEDOLOGY

A. *Participants*

The participants in the study were 60 students in the foreign languages college in Guizhou University, China. All of them were selected from two intact classes of English majors in the third year based on the convenience sampling method.

B. *Research Instrument*

One hundred Chinese students majoring in English and thirty-three English native speakers in Guizhou University were conveniently selected to help the design of the thirty WDCT items. The English majors were all in their second academic year, while the native speakers were from different countries, including America, England, and Canada. The development of WDCT in the present study consisted of four stages: selection of the speech acts to be tested, exemplar generation, likelihood investigation and content validity check, which are explained as follows.

1. Selection of the speech acts to be tested

To select the speech acts to be tested, a questionnaire was designed. In this questionnaire, all the speech acts in Searle (1975) and the speech acts appeared in previous studies were listed. The teachers group (two American teachers and four Chinese teachers of English in Guizhou University) were invited to evaluate the possibility of all the speech acts for college students with the researcher. The selection of the speech acts were based on the familiarity and frequency of the use in the daily life decided by the teachers' group and the researcher, and finally twenty speech acts were selected to be listed in the questionnaire. After that, the questionnaire was distributed to the one hundred English majors and they were required to choose the top ten frequently used speech acts they may meet in their daily life. Ninety-seven valid questionnaires were collected. After the calculation of the frequency, the most frequently used ten speeches acts were: advice, gratitude, greeting, congratulation, apology, request, compliment, inquiry, refusal and compliment response. The frequencies of them are illustrated in Table 2.

TABLE 2.
FREQUENCIES OF THE TOP TEN USED SPEECH ACTS

Speech act	Advice	Gratitude	Greeting	Congratulation	Apology
Frequency (%)	81	71	69	64	63
Speech act	Request	Compliment	Inquiry	Refusal	Compliment Response
Frequency (%)	59	56	50	42	38

2. Exemplar Generation

After the ten speech acts were decided, the next step was to obtain topics of the speech acts through exemplar generation (Rose & Ono 1995). An exemplar generation questionnaire was designed with an example of the situation of each speech act in both English and Chinese. Every student was required to write one possible situation they met in their daily life for each speech act. The students were encouraged to write the situations in English, but Chinese was allowed when writing in English was difficult. All the students wrote in English except one. Most students finished it within half an hour. As a result, 173 situations were collected, and the number of situations for each speech act is illustrated in Table 3.

TABLE 3.
DISTRIBUTION OF SITUATIONS

Speech act	advice	gratitude	greeting	congratulation	apology
No. of situations	18	23	16	18	20
Speech act	request	compliment	inquiry	refusal	compliment response
No. of situations	17	16	19	15	11

3. Likelihood Investigation

The third stage was a likelihood investigation. A questionnaire was designed to include all the situations collected in the above stage. The thirty-three native speakers were asked to indicate on a five point rating scale of likelihood, from impossible to most likely, according to the possibility that the situations would occur in their daily life. The likelihood investigation questionnaire was written in English. All the native speakers finished it within an hour. The top three situations of each speech act were selected in the study based on their mean scores. Finally, altogether thirty situations were obtained in the WDCT.

4. Content validity of WDCT

The thirty situations were rewritten and organized without changing the original meanings. The two American teachers and four Chinese teachers of English in the foreign languages college of Guizhou University were invited to check the content validity with the researcher. As Intaraprasert (2000) indicates that the texts should be validated in terms of appropricy, familiarity and degree of specification. The purpose of doing this was to obtain the data for the following issues: 1) Whether the expressions of the items are appropriate; 2) Whether each situation could elicit the expected speech act; 3) Whether the situations are typical in both America and China; 4) Whether the situations are familiar with the students. The results revealed that all items were appropriate for the present study and they could elicit the correct speech acts except some revisions on the language organization. Besides, the teachers' group and the researcher decided to assign this test to the third year students after their evaluation.

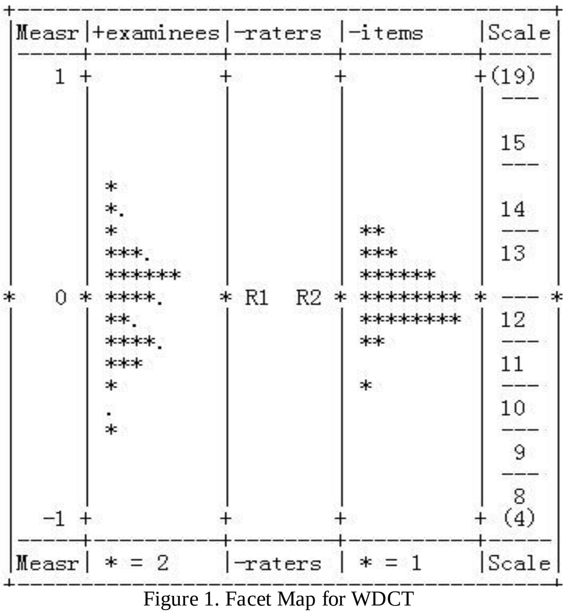
The participants' responses in the WDCT will be evaluated by the rating criteria adapted from Hudson et al. (1995). There are four aspects of pragmatic competence to be rated, i.e. the ability to use the correct speech act, typical expressions, amount of speech and information, and levels of formality, directness and politeness. The appropriacy of each aspect will be scored on a five point rating scale ranging from 1, "very unsatisfactory", to 5, "completely appropriate".

III. DATA COLLECTION

The participants were required to finish the WDCT in the classroom circumstances, and no discussion was allowed. The language required in the test was English. All the students could finish within 90 minutes. The data were scored by two American teachers, and both of them work in the foreign languages college of Guizhou University and got master degrees of Arts, but if they could not reach an agreement, the third rater will be invited. The data were analyzed on the base of many-facet Rasch model (MFRM) with the help of Facets (3.71.4.) to calculate the reliability and validity of WDCT. The raters' reliability, the item difficulty level and discrimination power, criteria reliability and construct validity were calculated. The following section is the detailed description of them.

IV. RESULTS

Figure 1 is the general description of the examinees' ability, the leniency/severity of the raters, the difficulty of items and the scores used in the testing. There are five columns in the map. The first column displays the linear, equal-interval logit scale. Upon it, all facets in the analysis are positioned, and it illustrates a framework of reference for comparisons within and between the facets. The second column presents the examinees' performance measures, showing the tendency of the examinees to receive the high or low ratings from the raters on the logic scale. The examinees are ordered from high performing to low performing with the logit scale ranged from +1.0 to -1.0. The third column displays the raters' leniency/severity. The raters are ordered from more severe to more lenient when scoring the examinees. In figure 1, it can be seen that the two raters are almost on the same degree of severity/leniency at the level of about 0.0 logit, which means both of them are neither severe nor lenient. The fourth column displays the average difficulty level of items. The items' difficulty levels are ordered from more difficult to less difficult with the logit scale ranged from +0.5 logits to -0.5 logits. The fifth column graphically describes the 20-point rating scale used to score examinees' responses. It can be seen that the examinees were scored from 4 to 19.



```
+-----------------------------------------+
|Measr|+examinees|-raters |-items     |Scale|
|-----+----------+--------+-----------+-----|
|   1 +          +        +           +(19) |
|     |          |        |           | --- |
|     |          |        |           |     |
|     |          |        |           | 15  |
|     |          |        |           | --- |
|     | *        |        |           |     |
|     | *.       |        |           | 14  |
|     | *        |        | **        | --- |
|     | ***.     |        | ***       | 13  |
|     | ******   |        | ******    |     |
*   0 * ****.    * R1  R2 * ********  * --- *
|     | **.      |        | ********  | 12  |
|     | ****.    |        | **        | --- |
|     | ***      |        |           | 11  |
|     | *        |        | *         | --- |
|     | .        |        |           | 10  |
|     | *        |        |           | --- |
|     |          |        |           |  9  |
|     |          |        |           | --- |
|     |          |        |           |  8  |
|  -1 +          +        +           + (4) |
|-----+----------+--------+-----------+-----|
|Measr| * = 2    |-raters | * = 1     |Scale|
+-----------------------------------------+
```

Figure 1. Facet Map for WDCT

A. *Examines*

Table 4 illustrates the information provided on examinees. Examinees are identified in column 1, an estimate of their ability is presented in column 2, errors of these estimates are in column 3, and column 4 shows information on the extent to which the model was functional in estimating the observed scores for the examinees. This is expressed in terms of the degree of match, or fit, between the expectations of the model and the actual performance for each examinee. The acceptable range of infit MnSq (mean square) is mean ± 2 deviations, and the acceptable ZStd (Z standard score) is between +2.0 to -2.0 (Linacre 2003). Values less than the minimum of the range indicate that the observed data are closer to their expected ratings than the model predicts (i.e., overfit). Values greater than the maximum of the range indicate the observed data are farther than the model expects (i.e., misfit) (Myford & Wolfe, 2003). Table 4 shows the examinees' ability measures spanned +.53 logits to -.61 logits. The Infit MnSq spanned 1.38 to .56 with a mean of 1.01 and a standard deviation of .19 and the Infit ZStd spanned +1.9 to -2.8. There is one examinee (1.7%) who is overfit. The percentage of examinees who are misfit or overfit should be at most around 2% (Pollitt & Huchinson 1987), so 1.7% is acceptable. At the bottom of Table 4, the reliability of separation index was 3.77, which indicates there is a significant difference among the examinees' ability. The separation index shows the significant difference among the examinees if it is bigger than 2.00. The separation reliability is .93 (above .70) which shows that the WDCT is reliable. The fixed chi-square test tests the hypothesis that all examinees are of the same level of performance. The Chi-square is 940.5 with d.f. 59 and the significance level is .00 (<.01). This confirms that there exists a significant difference among the examinees.

TABLE 4.
EXAMINEES MEASUREMENT REPORT (ARRANGED BY FN).

Examinee	Measure	SE	Fit	
			Infit MnSq	Infit ZStd
S6	-.16	.06	1.38	1.9
S51	-.02	.06	1.32	1.5
S22	.14	.07	1.31	1.5
S11	-.17	.06	1.29	1.4
S46	.19	.07	1.27	1.3
S59	.08	.06	1.27	1.3
S9	.13	.07	1.27	1.3
S14	.04	.06	1.25	1.2
S4	.06	.06	1.19	.9
S37	-.48	.06	1.16	.9
S54	-.08	.06	1.19	1.0
S35	-.17	.06	1.19	1.0
S20	.33	.07	1.19	.9
S18	.36	.07	1.16	.8
S41	-.04	.09	1.14	.5
S40	.10	.07	1.13	.7
S16	.15	.07	1.12	.6
S47	.14	.07	1.10	.5
S32	.13	.07	1.10	.5
S7	-.18	.06	1.08	.4
S12	-.09	.06	1.08	.4
S8	.04	.06	1.07	.4
S30	-.27	.06	1.07	.4
S19	-.26	.06	1.07	.4
S44	.06	.06	1.07	.3
S5	-.23	.06	1.06	.3
S39	-.08	.06	1.05	.3
S24	.18	.07	1.04	.2
S33	-.06	.06	1.03	.2
S34	.11	.07	1.03	.2
S45	.23	.07	1.01	.1
S50	-.08	.06	1.01	.1
S3	.43	.07	.99	.0
S57	.14	.07	.99	.0
S2	-.59	.06	.98	.0
S13	.25	.07	.96	-.1
S10	.07	.06	.95	-.1
S17	.46	.07	.98	.0
S36	-.19	.06	.94	-.2
S38	.22	.07	.94	-.2
S42	.00	.05	.94	-.3
S48	-.29	.06	.93	-.3
S23	.20	.07	.93	-.3
S58	.19	.07	.90	-.4
S27	-.31	.06	.89	-.6
S49	.03	.06	.89	-.5
S1	-.61	.06	.87	-.7
S56	.36	.07	.85	-.7
S29	-.14	.06	.89	-.5
S31	-.27	,06	.81	-1.1
S25	-.19	.06	.81	-1.0
S15	.53	.07	.77	-1.2
S53	-.37	.06	.77	-1.4
S55	-.02	.06	.75	-1.3
S43	-.45	.06	.70	-1.9
S28	.06	.06	.67	-1.8
S52	-.23	.06	.67	-1.9
S21	.01	.06	.67	-1.9
S60	-.05	.06	.65	-2.0
S26	-.27	.06	.56	-2.8
Mean	-.02	.06	1.01	.0
SD	.25	.01	.19	1.1

Model, Sample: Separation 3.77 Reliability .93.
Model, Fixed (all same) chi-square: 940.5 d.f.: 59 Significance (probability): .00

B. Raters

Table 5 displays more detailed information of the two raters. Raters are identified in column 1 and an estimate of their leniency/severity in column 2, errors of these estimates in column 3 and the fit statistics in column 4. In this case it indicates the relative consistency in the raters. Lack of consistency is a problem and such raters need to be retrained or changed. In Table 5, it can be found that Rater 1 is more severe than Rater 2 and the difference is .02 logits. The error was small and no raters are identified as misfitting. The Infit MnSq is within the mean ± 2 deviations and the Infit ZStd is within ± 2.0, so both raters are self-consistent in scoring. At the bottom of this table, the separation index, reliability of separation and chi-square results are provided. In the case of raters, a low reliability is desirable since ideally the different raters should be equally severe/lenient. The separation is 1.27 (<2.00) and the reliability of separation is .62 (<.70) which means the severity/leniency of the two raters is not significantly different. The chi-square is 2.6 with d.f. 1 and the significance level is .11 (>.05), which confirms that there is no significant difference between the raters.

TABLE 5.
RATERS MEASUREMENT REPORT (ARRANGED BY FN)

Rater	Measure	SE	Fit	
			Infit MnSq	Infit ZStd
R1	.01	.01	1.04	1.1
R2	-.01	.01	.96	-1.1
Mean	.00	.01	1.00	-.0
SD	.02	.00	.06	1.7

Model, Sample: Separation 1.27 Reliability .62
Model, Fixed (all same) chi-square: 2.6 d.f.: 1 significance (probability): .11

C. *Items*

Table 6 shows the estimated difficulty of the items. Items are identified in column 1. Their difficulty is shown in column 2, and items without minus are more difficult and items with minus are less difficult. The range of difficulty spans .27 to -.41 logits. Errors of these measure estimates are provided in column 3, and the error is .04. In column 4, the fit statistics are presented. Items which show greater variation than the model expected are misfitting (mean + 2 deviations) and those which show smaller variation than expected are overfitting (mean – 2 deviations). No items are found either misfitting or overfitting, but Item 7 is on the border of misfitting and Item 15 and 13 are on the border of overfitting. These can be improved by modifying the items and retraining the raters. At the bottom of the table, the separation index 3.13 and the reliability of separation .91 are shown, which means the items are with significantly different difficulty. The chi-square significance .00 (<.01) further confirms this.

TABLE 6.
ITEMS MEASUREMENT REPORT (ARRANGED BY FN)

Item	Measure	SE	Fit	
			Infit MnSq	Infit ZStd
I7	.11	.04	1.28	2/0
I19	.05	.04	1.26	1.8
I16	-.06	.05	1.26	1.8
I14	-.11	.05	1.23	1.6
I28	.27	.04	1.22	1.7
I11	-.02	.04	1.16	1.1
I4	-.01	.04	1.16	1.1
I8	.18	.04	1.08	.6
I3	-.10	.05	1.15	1.0
I29	.19	.04	1.14	1.0
I9	.26	.04	1.11	.9
I21	.14	.04	1.12	.9
I6	-.07	.05	1.04	.3
I20	.08	.04	1.02	.1
I30	.17	.04	.98	-.1
I18	.04	.04	.92	-.5
I24	-.03	.04	.92	-.5
I22	.08	.04	.92	-.6
I26	-.16	.05	.91	-.6
I10	.-09	.04	.86	-1.0
I1	-.09	.05	.87	-.9
I5	-.04	.04	.86	-1.0
I25	-.41	.05	.87	-1.0
I2	-.07	.05	.84	-1.2
I27	-.23	.05	.81	-1.4
I12	-.01	.04	.79	-1.6
I23	-.13	.05	.78	-1.7
I17	-.05	.04	.78	-1.7
I15	-.07	.05	.75	-2.0
I13	.01	.04	.75	-2.0
Mean	.00	.04	.99	-.1
SD	.15	.00	.17	1.3

Model, Sample: Separation 3.13 Reliability .91
Model, Fixed (all same) chi-square: 308.4 d.f.: .29 significance (probability): .00

D. *Rating Scale*

Table 7 shows the rating scale statistics. Column 1 displays information relating to the data, including the categories (categories span 4-19, because for each aspect of rating, the lowest score is 1, then in total for the four aspects, the lowest score is 4, and no one got the full score 20), observed use of each category (counts used), percentage of the used responses (%), and cumulative percentage of responses in this category (cum %). Information in column 2 describes the validity of the categorization, which includes the average of the measures, the expected measures and the unweighted mean-square for observations in this category (outfit mean square). Monotonically increasing of the thresholds is one basic requirement for the validity of the rating scale (Piquero et al. 2001). The Infit MnSq is not reported because it approximates the Outfit MnSq when the data are stratified by category (Linacre 2014). Since high categories are intended to reflect high measures, the average measures are expected to advance (Linacre 1997). The logit values of the average measures for the scales from 4 to 19 range from -.73 to .35, and these measures are monotonically increasing. The outfit mean-square index is also a useful indicator of rating scale functionality. For each rating scale category, Facets computes the observed average ability measure and an expected average ability measure of the examinees. When the observed and expected ability measures are close, the outfit MnSq index for the rating category will be around the expected value 1.0. The greater outfit MnSq index indicates the larger discrepancy between the observed and expected measures. For a given rating category, any outfit MnSq index greater than 2.0 suggests that the ratings in that category for one or more examinees may not be contributing to meaningful measurement (Linacre 1999). As shown in Table 7, every outfit MnSq index is around 1.0 and no one is greater than 2.0, which suggests that the rating scales seem to be functioning as intended. Another pertinent rating scale 'characteristics' includes thresholds, or step calibration, and category fit statistics (Bond & Fox 2001). For this index, the ideal distance for each two rating scales is 1.0 logits and it cannot be bigger than 4.0 logits (Linacre 1999). When the logits are bigger than 4.0, it indicates there is a central tendency in rating. In Table 7, the distance between each two rating scales is no bigger than 4.0 logits.

TABLE 7.
RATING SCALE STATISTICS

Data				Fit			Step Calibration	
Category score	Counts Used	%	Cum. %	Avge Meas	Exp. Meas	Outfit MnSq	Measure	S.E.
4	5	0	0	-.73	-.50	.6		
5	17	0	1	-.69	-.45	.5	-1.70	.45
6	39	1	2	-.39	-.39	1.0	-1.25	.22
7	76	2	4	-.28	-.34	1.1	-1.06	.13
8	121	3	7	-.20	-.27	1.2	-.74	.09
9	199	6	13	-.18	-.21	1.1	-.74	.07
10	315	9	22	-.17	-.15	1.0	-.64	.05
11	502	14	35	-.12	-.09	.9	-.59	.04
12	598	17	52	-.03	-.03	1.0	-.23	.04
13	598	17	69	.03	.03	1.0	.00	.04
14	567	16	84	.11	.09	.9	.11	.04
15	314	9	93	.17	.15	.9	.71	.05
16	150	4	97	.21	.21	1.0	.92	.07
17	73	2	99	.22	.27	1.1	.96	.11
18	22	1	100	.26	.33	1.1	1.50	.21
19	2	0	100	.35	.39	1.0	2.76	.71

Generally speaking, the MFRM analyzed the reliability and validity from four facets (examinees, raters, items and rating scales) of WDCT. The results show that WDCT has high reliability. Table 4 shows the examinees' abilities are significantly different, although one examinee is overfitting, the percentage 1.7% is still acceptable. Table 5 illustrates that the two raters are consistent and there is no significant difference in their severity/leniency. Table 6 proves that the items difficulty is significantly different. The rating scale statistics in Table 7 shows a good construct validity of WDCT as well since no overfitting or misfitting is found and the measure is monotonically increasing. In a word, with the high reliability and construct validity, the WDCT can be used to evaluate the examinees' ILP competence in conducting speech acts functionally.

V. DISCUSSION

The present study shows a high reliability of WDCT in testing ILP competence in speech acts, which is in accordance with the findings of Yamashita (1996a, 1996b), Hudson (2001a), Liu (2004) and Duan (2012), but it is different from what Yoshitake (1997) found. This study also concludes that WDCT is a valid measure. Some of the previous researchers (Hudson, Detmer & Brown 1995, Yamashita 1996a, 1996b, Hinkel 1997, Ahn 2005, Duan 2012) drew the same conclusion, whereas others (Rose 1994, Rose & Ono 1995) hold the opposite point of view.

Reliability and validity are complementary aspects of validation process (Bachman & Savignon 1990). The present study has some implication for not only WDCT development, but for the development of different test forms in ILP competence. According to Roever (2005), to test pragmatic knowledge, the basic concern for item development is that the items should represent the real-world language use, but not based on the intuition of the designers. In the development of WDCT, the first step is the selection of the speech acts to be tested. The purpose of this step is to obtain the authenticity of the test because the speech acts in the test were familiar with the EFL learners. Authenticity is seen as a critical quality of language tests and is said to have a great effect on test takers' performance (Bachman & Palmer 1996). The second step is exemplar generation, in which the EFL learners were required to write the situations happen to them in each speech act. It can also help to enhance the authenticity of the study. The third step is likelihood investigation. In this step, the native speakers were asked to evaluate the possibility of the situations given by the EFL learners in their own culture. The situations which happen in both EFL culture and English-speaking culture were chosen, and it guarantees the authenticity of the situations in Chinese and target language cultures. Therefore, it tests the pragmatic ability when learning English for EFL learners. The last step is content validity check. In this step, the accuracy and organization of the language as well as the format of the WDCT can be guaranteed.

However, the present research proved the WDCT items developed in the study worked well in ILP competence testing in the Chinese context, and the results may be different when they are conducted with different groups or different cultural context. Further research is still recommended.

VI. CONCLUSION

This study investigates the reliability and validity of WDCT in testing EFL learners' ILP competence in conducting speech acts. The results show that WDCT is a reliable and valid measure in testing interlanguage speech acts performance for EFL learners. Examining the EFL learners' ILP competence will be of great help in understanding their levels in this field. The learners could recognize their problems in pragmatics in English, and then pay attention to them in the process of learning and in communication with native speakers. In addition, most English majors in China will go to English-related jobs after graduation (Zhu 2007, Zhang 2012), so to realize their weakness and to improve their ILP ability will be helpful for their future careers since appropriacy in using English is not emphasized in the college life and most EFL learners and teachers always ignore the importance of it (Liu 2004, Ji & Jiang 2010). Thus, to design a reliable and valid measure is extremely important, and the present research hopes to make some contribution in ILP competence testing and development.

ACKNOWLEDGMENT

The authors wish to thank the help and suggestions from all the committee members for the research as well as the students and teachers in China and Thailand who support it. This work was supported in part by a grant from School of Foreign Languages, Institute of Social Technology, Suranaree University of Technology, Thailand.

REFERENCES

[1] Ahn, R. C. (2005). Five measures of interlanguage pragmatics in KFL (Korean as a foreign language) learners. Unpublished Ph.D dissertation. University of Hawaii at Manoa, Honolulu, HI.

[2] Bachman, L. F., & Savignon, S. J. (1990). Fundamental considerations in language testing (Vol. 107). Oxford: Oxford University Press.

[3] Bachman, L. F. & Palmer, A. S. (1996). Language testing in practice. Oxford: Oxford University Press.

[4] Blum-Kulka, S. (1982). The study of translation in view of new developments in discourse analysis: Indirect speech acts. *Poetics today* 2.4, 89-95.

[5] Blum-Kulka, S. (1983). Interpreting and performing speech acts in a second language: A cross-cultural study of Hebrew and English. In Wolfson, N., & Judd, E. (eds.). *Sociolinguistics and language acquisition*. New York: Newbury House Publishers, 3655.

[6] Blum-Kulka, S., & Olshtain, E. (1986). Too many words: Length of utterance and pragmatic failure. *Studies in Second language acquisition* 8.2, 165-179.

[7] Bond, T. G., & Fox, C. M. (2001). Applying the Rasch model: Fundamental measurement in the human sciences. New York: Psychology Press.

[8] Brown, J.D. (2001). Six types of pragmatics tests in two different contexts. In Rose, K. & Kasper, G. (eds.). *Pragmatics in language teaching*. New York: Cambridge University Press, 301-325.

[9] Brown, J. D. (2008). Raters, functions, item types and the dependability of L2 pragmatics tests. In E. Alcón Soler & A. Martínez-Flor (eds.). *Investigating pragmatics in foreign language learning, teaching and testing*. Clevedon: Multilingual Matters, 224-248.

[10] Duan, L. L. (2012). The study of reliability and validity of pragmatic competence testing measures. *Foreign languages studies* 4, 84-96.

[11] Fauzul, A. (2013). The assessment tool of L2 learners' pragmatic competence: Written discourse completion test (WDCT). *IPEDR* 68.19, 113-117.

[12] Cohen, A. D., Olshtain, E. & Rosentain. (1986). Advanced EFL apologies: remains to be learned. *International journal of sociology of language* 62, 51-74.

[13] Garrett, H. E. (1937). Statistics in psychology and education. London: Longman.

[14] Golato, A. (2003), Studying compliment responses: A comparison of recordings of DCTs and naturally occurring talk. *Applied linguistics* 24.1, 90-121.

[15] He, Z. R. & Chen, X. R. (2004). Contemporary pragmatics. Beijing: Foreign Language Teaching and Research Press.

[16] Hinkel, E. (1997). Appropriateness of Advice: DCT and Multiple Choice Data. *Applied linguistics* 18.1, 1-26.

[17] House, J. & Kasper, G. (1987). Interlanguage pragmatics: Requesting in a foreign language. In W. Lorscher & R. Schulze (eds.). *Perspectives on language in performance.* Festschirift for Werner Hullen. Tabingen: Narr, 1250-1288.

[18] House, J. (1989). Politemess in English and German: the functions of please and bitte. In S. Blum-Lulka, J. House & G. Kasper (eds.). *Cross-cultural pragmatics*. Norwood, NJ: Ablex, 96-119.

[19] Hudson, T. (2001a). Indicators for cross-cultural pragmatic instruction: some quantitative tools. In Rose, K. and Kasper, G. (eds.). *Pragmatics in language teaching*. Cambridge: Cambridge University Press, 283-300.

[20] Hudson, T. (2001b). Self-assessment methods in cross-cultural pragmatics. In T. Hudson & J. D. Brown (eds.). *A Focus on language test development.* Honolulu, HI: University of Hawaii Press, 57-74.

[21] Hudson, T., Detmer, E. and Brown, J. D. (1995). Developing prototypic measures of cross-cultural pragmatics. (Technical Report 7). Honolulu, HI:University of Hawaii, Second Language Teaching and Curriculum Center.

[22] Intaraprasert, C. (2000). Language learning strategies employed by engineering students learning English at the tertiary level in Thailand. Ph.D dissertation, University of Leeds.

[23] Ji, P. Y. & Jiang, J. (2010). A Reconsideration on Pragmatic Competence in College English Education. *Foreign language world* 6, 33-41.

[24] Johnston, B., Kasper, G. & Ross, S. (1998). Effect of rejoinders in production questionnaires. *Applied linguistics* 19, 157–182.

[25] Kasper, G. (1998). Interlanguage Pragmatics. In H. Byrnes (eds.). *Learning foreign and second lauguages: Perspectives in research and scholarship.* New York: The Modern Language Association of America.

[26] Kasper, G. & Rose, K. R. (1999). Pragmatics and SLA. *Annual review of applied linguistics* 19, 81-104.

[27] Linacre, J. M. (1997). Guidelines for rating scales. In Midwest Objective Measurement Seminar. Chicago: MESA Press, Research Note Vol 2.

[28] Linacre, J. M. (1999). Investigating rating scale category unity. *Journal of outcome measurement* 3, 103-122.

[29] Linacre, J. M. (2003). A user's guide to FACETS: Rasch-model computer programs. Chicago: MESA Press.

[30] Linacre, J. M. (2014). A user's guide to FACETS: Rasch-Model computer programs. http://www.winsteps.com/a/facets-manual.pdf/ (accessed 8/12/2014).

[31] Liu, J. D. (2004). Measuring interlanguage pragmatic knowledge of Chinese EFL learners. Unpublished Ph.D dissertation. City University of Hong Kong.

[32] Liu, J. D. (2006a). Measuring interlanguage pragmatic knowledge of EFL learners. Frankfurt am Main: Peter Lang.

[33] Liu, J. D. (2006b). Assessing EFL learners' interlanguage pragmatic knowledge: Implications for testers and teachers. *Reflections on English language teaching* 5, 1-22.

[34] Liu, J. D. (2007). Developing a pragmatics test for Chinese EFL learners. *Language testing* 24.3, 391-415.

[35] Ma, T (2010). The interlanguage pragmatic competence of English majors and its development in ethnic universities---A case study of Beijing ethnic university. Ph.D dissertation. Shanghai International Studies University.

[36] Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: part I. *Journal of applied measurement* 5.2, 189-227.

[37] Nishihara, S. (1999). Kenkyu no gaino. In Shimpro 'Nihongo' Dai 2-han (ed.) Video Shigi ni-yoru Gengo Koudou Ishiki Chousa Houkokusho: Bunseki hen (Study of consciousness toward linguistic behaviors of Japanese by video stimulation: Book of analysis). Tokyo: National Language Research Institute. Ministry of Education of Japan, Scientific Research Grant No. 09NP0701, 1-14.

[38] Olshtain, E. & Weinbach, L. (1987). Complaints: A study of speech act behavior among native and nonnative speakers of Hebrew. In J. Verschueren & M. Bertucceslli-Papi (eds.). *The pragmatic perspective: Selected papers from the 1985 international pragmatics conference.* Amsterdam: John Benjamins, 195-208.

[39] Piquero, A. R., MacIntosh, R., & Hickman, M. (2001). Applying Rasch modeling to the validity of a control balance scale. *Journal of criminal justice* 29.6, 493-505.

[40] Pollitt, A., & Hutchinson, C. (1987). Calibrating graded assessments: Rasch partial credit analysis of performance in writing. *Language testing* 4.1, 72-92.

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

[42] Roever, C. (2010). Effects of native language in a test of ESL pragmatics: A DIF approach. In G. Kasper, H. thi Nguyen, D. R. Yoshimi, & J. Yoshioka (eds.). *Pragmatics and language learning.* Honolulu, HI: National Foreign Language Resource Center, 187-212.

[43] Rose, K. R. (1994). On the validity of discourse completion tests in non-western contexts. *Applied linguistics* 15.1, 1-14.

[44] Rose, K. R. & Ono, R. (1995) Eliciting speech act data in Japanese: the effect of questionnaire type. *Language learning* 45, 191–223.

[45] Searle, J. R. (1975). Indirect speech acts. In Cole & Morgan (eds.). *Syntax and semantics.* New York: Academic Press, 59-82.

[46] Shinpro 'Nihongo' Dai 2-han (1999a). Video Shigi ni-yoru Gengo Koudou Ishiki Chousa Houkokusho: Bunseki hen (Study of Consciousness Toward linguistic behaviors of Japanese by video stimulation: Book of snalysis). Tokyo: National Language Research Institute. Ministry of Education of Japan, Scientific Research Grant No. 09NP0701.

[47] Shinpro 'Nihongo' Dai 2-han (1999b). Video Shigi ni-yoru Gengo Koudou Ishiki Chousa Houkokusho: Bunseki hen (Study of consciousness toward linguistic behaviors of Japanese by video stimulation: Book of data). Tokyo: National Language Research Institute. Ministry of Education of Japan, Scientific Research Grant No. 09NP0701. [d3].

[48] Si, L. H. (2001). Interlanguage, pragmatic competence and culture learning. *Foreign language research* 2.105, 101-106.

[49] Subong, P. E. (2006). Statistics for research: applications in research, thesis and dissertation writing, and statistical data management using SPSS software. Sampaloc, Manila: Rex Bookstore.

[50] Takahashi, T. & Beebe, L.M. (1987). The development of pragmatic competence by Japanese learners of English. *JALT journal* 8, 131-55.

[51] Takahashi, T. & Beebe, L. M. (1993). Cross-linguistic influence in the speech act of correction. In G. Kasper & S. Blum-Kulka (eds.). *Interlanguage pragmatics.* New York: Oxford University Press, 138-157.

[52] Yamashita, S.O. (1996a). Comparing six cross-cultural pragmatics measures. Ph.D dissertation. Temple University.

[53] Yamashita, S.O. (1996b) Six Measures of JSL Pragmatics. (Technical Report 14). Honolulu, HI: University of Hawaii, Second Language Teaching and Curriculum Center.

[54] Yoshitake, S. S. (1997). Measuring interlanguage pragmatic competence of Japanese students of English as a foreign language: a multi-test framework evaluation. Unpublished PhD thesis. Columbia Pacific University, Novata, CA.

[55] Zhang, D. Y. (2012). Employment Situation of English Major Graduates and Countermeasure Analysis. *Journal of HuBei TV University* 32.9, 50-52.

[56] Zhu, X. (2007). The Strategies to Cope with Employment for University English-Major Undergraduates. *Journal of Neijiang technology* 1, 13. 65.

Lan Xu was born in Guizhou province, China, Oct. in 1980. She is currently a Ph.D candidate in School of Foreign Languages, Institute of Social Technology, Suranaree University of Technology, Thailand. She has been teaching in English Department of Foreign Languages School in Guizhou University, China since 2001. She received her bachelor and master degrees in English studies in 2001 and 2006 respectively in Guizhou University, China. Her research interests involve second language acquisition, pragmatics and language testing.

Anchalee Wannaruk was born in 1967. She is currently an associate professor, Ph.D advisor and the chair of School of Foreign Languages, Institute of Social Technology, Suranaree University of Technology, Thailand. She received her Ph.D in the Graduate College of the University of Illinois at Urbana-Champaign, America, in 1997, won her M.A. in Mahidol University, Thailand, in 1989 and got her B.Ed in Chulalongkorn University, Thailand, in 1986. Her research interests involve pragmatics, language testing and corpus-based language learning.