Iranian Journal of Language Issues (IJLI)
Vol 1, No. 1, 2015

Developing a test of interlanguage pragmatics for Iranian EFL learners in relation to the speech acts of request and apology

Mohammad Salehi

Assistant Professor, Sharif University of Technology

Email: m_salehi@sharif.edu

Ebrahim Isavi

Ph.D. Candidate, Kharazmi University

Email: ebrahim_isavi@yahoo.com

Abstract

Testing interlanguage pragmatics has only recently caught the attention of the scholars in the field (Brown and Ahn, 2011;Liu, 2010) and currently there exist a multiplicity of uncharted areas and unanswered questions. This study is an attempt to develop a test of pragmatics in the Iranian context in relation to the speech acts of request and apology. Test development constituted five stages of exemplar generation, likelihood investigation, scenario generation, initial multiple-choice discourse completion (MDCT) piloting and the final MDCT administration. A total of 245 male and female participants from Sharif University of Technology (SUT) and Iran Language Institute (ILI) participated in different stages of test development. Rasch analysis was conducted for statistical analysis. The results indicate that the test is fairly unidimensional with one misfitting item and no misfittings persons. The test has local independence. Moreover, item reliability is found to be .88 whereas person reliability is found to be .64. High item reliability is attributable to the large number of participants.

Key words: interlanguage pragmatics testing, speech act, discourse completion test, request, apology

Received: April 2013; Accepted: December 2013

1. Introduction

Different models of communicative competence (Bachman, 1990; Canale and Swain, 1980; Celce-Murcia, Dornyei and Thurrel, 1995) have either explicitly or implicitly placed stress on the pragmatic component as an indispensable component. Language testing, on the other hand, has traditionally overlooked this crucial component of language proficiency. Contemporary to the recognition of the significant role of the pragmatic component, attempt was made to include that component in language testing (Farhady, 1980; Shimazu, 1989). Since the pioneering work of Farhady, a handful of research studies have undertaken the development of pragmatic tests (Bouton, 1999; Hudson, Detmer and Brown, 1995;Jianda, 2007).

The few tests of pragmatics which have been developed are categorized into tests of either pragmalinguistics or sociopragmatics. These studies have mainly relied on speech acts, routines and implicatures in developing the tests. Requests, apologies and refusals are the most frequently used speech acts in these studies. Following the above-mentioned studies, this study seeks to develop a multiple-choice interlanguage pragmatics test for Iranian EFL learners by focusing on the two speech acts of request and apology. The development of a pragmatic test in the context of Iran is justified on the grounds that pragmatic violation, unlike lexical and grammatical mistakes, can cause not only communication breakdowns but the native speaker interlocutor can feel offended. Furthermore, a number of tests have found less than desirable reliability indices while others have found great reliability. This study is based loosely on Jianda’s (2007) approach to test development.

2. Literature Review

A great deal of research has been conducted with respect to the effectiveness of teaching pragmatics or the process learners go through in developing their pragmatic competence (Holmes and Brown, 1987). Very little research (Hudson, *et al* 1995; Jianda2006), however, has embarked on developing tests of pragmatics. In developing pragmatic tests, the researcher is required, due to the wide variety of pragmatic sub-components, to include only a number of the sub-categories such as implicature, hinting and speech acts. Roever (2005) includes speech acts, implicature and routine formulas. Bouton (1999) incorporates two types of implicature, that is, idiosyncratic and formulaic implicature. Jianda (2006) designed a test of requests and apologies for Chinese EFL learners.

One of the key considerations in interlanguage pragmatics testing concerns the issue of instrumentation. Different scholars have favored different instrumentations. Kasper and Rose (2002) offer elicited conversation, authentic discourse, role-plays, production questionnaires, multiple-choice instruments, interviews, diaries; think-aloud protocols and scales as appropriate for eliciting interlanguage pragmatics knowledge. Ahn (2005), on the other hand, offers self-assessment, language lab oral production test, open discourse completion test, role-play, role-play self-assessment and multiple-choice discourse completion test (MDCT) for cross-cultural pragmatics. Hudson *et al* (1992, 1995) use six types of discourse completion tests such as written discourse completion tasks (WDCTs), multiple-choice discourse completion tasks (MDCTs) and oral discourse completion tasks (ODCTs).

Farhady (1980) was the first scholar who systematically developed a test of pragmatic competence. Based on a functional approach, he developed a multiple-choice (MC) test in an academic context to assess students' ability on language functions such as expressing attitudes and requesting. Test items were not generated based on a preset scheme and they were generated by the researcher himself. Then, a panel of professors and students reviewed the items for the authenticity of the items. The test development process, then, followed three stages. First, open-ended test items were administered to a group of native speakers (NS) with various academic backgrounds in order to elicit socially appropriate and linguistically accurate responses. The most frequent responses were then selected as the key for each item. The second stage was the same as the first one with non-native speakers (NNS) as participants. NNS responses were compared with those of NSs in order to find NNS deviant responses. Three different types of deviant responses were identified: socially appropriate and linguistically inaccurate, socially inappropriate but linguistically accurate and socially inappropriate and linguistically inaccurate. These deviant responses were selected as distractors. To assure the appropriateness of the alternatives, 56 multiple-choice items were presented to both NSs and NNSs. Finally, the items on the test were divided into two counterbalanced forms before they were administered. Scoring was based on different weights given to different options with two points for the key, one for either linguistically accurate and socially inappropriate or socially appropriate and linguistically inaccurate and zero points for both linguistically inaccurate and socially inappropriate options.

The test was found both valid and reliable and shortening the test did not have any effect on reliability and validity. The results also showed that students with different genders, university statuses, majors of study, nationalities, etc. performed significantly differently on the test.

The largest and probably the most important test of interlanguage pragmatic knowledge was conducted by Hudson, Detmer and Brown (1992, 1995). They attempted to develop other methods to assess pragmatic competence and arrived at three types of indirect, semi-direct and self-assessment measures each of which constituted two test formats (Lui, 2010). The indirect measures were free response DCT and multiple-choice DCT. The WDCT was composed of a short situational description followed by a blank into which the respondent was required to write the correct response. They designed this test to assess Japanese ESL learners' ability to both produce and comprehend the appropriate use of request, apology and refusal speech acts. The variables were selected because, within the research on pragmatics, they are identified as being the three independent and culturally sensitive variables which subsume all other variables and play a principled role in speech act behavior (Hudson, 2001). They also decided to test *power*, *distance* and *degree of imposition* for each speech act.

In order to develop interlanguage pragmatics tests, Jianda (2007) went through five stages. Initially, a group of EFL learners were asked to write as many scenarios, in what was called exemplar generation, as they could with respect to the speech acts the researcher was interested in. The learners attended some training sessions before they could generate their responses. The researcher then separated the dissimilar responses out to be included in the test. The second phase was that of situation likelihood investigation in which the singled-out responses were put on another questionnaire and the respondents were asked, on the Likert scale, to indicate the likelihood of the situations occurring in their daily lives. Then, the selected scales by the respondents were averaged out. Finally, the scales with the highest mean scores were selected to be used in the third stage of test development. Metapragmatic assessment constituted the third phase. In this phase, scenarios with various combinations of features were selected to be included in the metapragmatic assessment questionnaire. Then, a group of native speakers together with a group of non-native speakers were asked to indicate the degree of imposition, social distance and the power relationships in each scenario. Occasionally, non-native speakers needed some explanations of what was meant by 'power', 'imposition' and 'distance'. Jianda used 70%

agreement between native and non-native speakers as the threshold of acceptance. Those items which meet the set criterion were separated out as valid items to be included in the WDCT questionnaire. Native and non-native speakers were, in the next phase, required to take the tests. Then, two native speakers, who were given a rater-training manual, were requiredto rate the responses given by Chinese University students. Each rater rated one paper followed by a discussion. They were asked to use a 5-point scale and ignore grammatical errors. They were also asked to identify the inappropriate parts and comment briefly on why they thought those parts were inappropriate. The results were carefully reviewed. Finally, the MDCT was developed. Responses from native and non-native speakers were compared and the items selected by native speakers were used as key while items selected by non-native speakers were considered as inappropriate and they, therefore, formed distractors. Then, the MDCT questionnaire was given to native speakers to decide on the appropriate and inappropriate responses while giving reasons why they thought those items were inappropriate. Based on the data collected from native speakers, the options with the highest agreement among the respondents were selected for each situation.

In order to further investigate this unresolved issue, the following research question was posed:

Does the developed test enjoy the psychometric property of reliability?

3. Method

3.1. Setting and participants

The setting for the study included both the Iranian academic context at Sharif University of Technology and Iran Language Institute and the American academic and official contexts in the United States, England and Saudi Arabia. The native speaker participants were from the U.S., England, Saudi Arabia and Iran. Sixteen native speakers participated in the *scenario generation step,* one of the participants was from the U.S., a female, aged 26, an M.A. student in the field of computer science and engineering, one was from Iran, female, around 40, an M.A. student majoring in applied linguistics and fourteen native speakers teaching in schools in Saudi Arabia, ten males, four females, aged 35-65 one of whom was a British citizen, 4 Canadian citizens and 9 U.S. citizens. Seven native speakers also filled out *the initial multiple-choice discourse completion test* five of whom were from England, two females, one, a university student, aged 21 and one, a housewife, aged 35 and three male university students who fell within the age range of 20-25. Two native speakers were from The United States of America, one self-employed, aged 41 and the other one a university student who was aged 25.

Regarding the Iranian participants, two hundred and forty five students participated in different stages of the study. The first stage of the study was *exemplar generation* where students from Sharif University of Technology, 43 intermediate students, 30 females and 13 males, ranging in agefrom19 to 23 and language learners at Iran Language Institute, 49 intermediate learners, male, aged 17-21, and 9 advanced learners, male, aged 20-26 participated. The second stage of the study was *likelihood investigation* where 17 intermediate students, 6 males and 11 females, aged 19-23 at Sharif University of Technology were asked to mark the frequency of occurrence of each of the situations in an academic context on a 5-point scale. In the *scenario generation* stage of the study 27 intermediate male language learners, aged 17-21 at Iran Language Institute completed the questionnaire. Finally, 50 intermediate language learners, male, aged 17-21 and 50 advanced learners, male, aged 20-26 completed the *final multiple-choice discourse completion test*. All in all, a total of 245 Iranian language learners participated in this study.

3.2. Instrumentation and procedure

In the first stage of data collection, called *exemplar generation*, one hundred and one Iranian EFL learners were asked to write down situations in which one was required to make requests or apologies. To do this, each student was given a questionnaire especially tailored to meet the requirements of this stage and was required to write down a maximum of ten situations for each of the speech acts in an academic context. The questionnaire contained three columns with the first one designated to the speech acts, the second one to the addressee and the third one to the particular situations and learners' related examples. Prior to distributing the questionnaire, they were introduced to what was meant by apology and request through a couple of examples for each of the speech acts. After collecting all questionnaires, all of the situations were analyzed and the sixty one clearly distinct ones were selected to be included in the second phase of the study.

The second step was *likelihood investigation*. To reassure that the situations generated in the first phase of the study were representative of an academic context, the sixty one situations selected from the first step were given to the participants. They were asked, on a five Likert Scale and based on the frequency of the situations in an academic context, to rank the situations from the most frequent to the least frequent. The survey was conducted in English and a total of

17 learners at Sharif University of Technology participated in this phase of the study. They completed the questionnaire in the class. After collecting the surveys, a frequency count of the scales was conducted and a total of thirty five most frequent situations, 20 request and 15 apology situations, were selected to be included in the next phase.

In the *scenario generation stage*, the situations elicited from the learners were transformed into scenarios. Thirty five open-ended scenarios were generated. The generated scenarios were given to twenty seven intermediate EFL learners at Iran language Institute and fifteen native speakers to respond to. The respondents were asked to write down briefly how they would make either requests or apologies in the generated scenarios. Learners at Iran Language Institute were asked to take the surveys home to complete. After collecting the surveys, both native speaker and non-native speaker responses were analyzed and it was found out that native speaker responses showed some variety in terms of both the length of utterances and the type of language, that is, formal and informal language. Non-native speaker responses, in addition to variation in the length and type of language used, were also different in terms of how native-like they turned out to be with some responses being completely native-like and other containing either grammatical or pragmatic errors.

Development of *the multiple-choice discourse completion test (MDCT)*, with seven to ten options all chosen from the native speaker responses in the previous step, formed the fourth step of the study. Then, the initial multiple-choice discourse completion test was given to seven native speakers to mark the best option for each situation based on the two variables of *power* and *social distance*. After the completed questionnaires were emailed back, the most frequent option was selected to be included in the final form of the test. As for the distractors, a careful analysis of Iranian learners' responses was carried out and the responses which had the potential to function as acceptable distractors were selected for each item. Some lexical and grammatical errors were found and rectified in the selected responses. For the sake of ease of data analysis and due to the fact that some of the items were felt to be of minor importance in an academic context were deleted leaving a total of 25 items for the test.

Finally, the final form of the multiple-choice discourse completion test was given to 50 intermediate and 50 advanced learners at Iran Language Institute.

3.3. Data analysis

In order to analyze the data, item response theory (IRT), which consists in a family of statistical approaches providing probabilistic models linking item difficulty with the examinee's ability (Brown and Hudson, 2002) and using responses from a test or survey to simultaneously locate both the items and the test takers on the same latent continuum (Weir, 2005), was used. Due to the limited number of examinees and test items, the one-parameter model, dealing with item difficulty only and requiring as few as 100 examinees and 25 items (McNamara, 1996),was used for data analysis.

In order to analyze the data, the obtained data was codified in terms of 0's and 1's; with 1's representing the correct option while 0's representing wrong options. All of the responses from 100 learners were inserted into the SPSS software and the SPSS data was, then, inserted into the *Winsteps* software.

4. Results

The data was analyzed using Winsteps (version 3.70.1) (Linacre, 2010). The Rasch model assumes that all participants have answered at least one item correctly and that they have failed on at least one item. So, four participants with perfect scores were discarded from further analysis.

4.1. Model-Data Fit

The Rasch model makes a prediction about the behavior of the persons and items. Specifically, it is predicted that the test takers will be able to endorse the easy items and find difficult items difficult to answer. This is captured in the Rasch model's fit statistics. There are two types of fit statistics in *Winsteps*, namely, the Infit Mean Squares and the Outfit Mean Squares. The difference between these two statistics lies in their dealing with outliers with the former not taking into account the outliers. The suggested range for both fit statistics is .7 to 1.3 (Bond and Fox, 2007). An analysis of the fit of the items in this test (see Table 1) indicated that only one item, e.g. item 12, did not fit.As can be seen, the infit mean square for item 12 ranges from a minimum of 1.21 to a maximum of 3.2.Similarly, the outfit mean square ranges from a minimum of 1.45 to a maximum of 4.2 both of which are far beyond the accepted range. In order to be more confident that the item fails to fit, its Item Characteristic Curve (ICC)was also checked. The ICC for the missing item (item 12) is displayed in figure 1. The red curve is the expected behavior of the respondents on the item. The observed pattern of responses is displayed by the blue curve. It is evident that there is considerable difference between the expectations of the model and the empirical response pattern. Consequently, this item was discarded from further analysis.

The fit analysis in the Rasch model is symmetric for persons and items alike. That is, the same indices apply to persons. Person fit analysis revealed no misfitting persons.

Table1. Fit statistics

```
-------------------------------------------------------------------------------------------
|ENTRY   TOTAL  TOTAlMODEL|   INFIT  |  OUTFIT  |PT-MEASURE |EXACT MATCH|     |
|NUMBER  SCORE  COUNT  MEASURE  S.E. |MNSQ  ZSTD|MNSQ  ZSTD|CORR.  EXP.| OBS%  EXP%| ITEM |
|------------------------------------+----------+----------+-----------+-----------+-----|
|    12     47     97      .05     .22|1.21   3.2|1.45   4.2|A .04   .33| 48.5  63.4| twel |
|    14     58     97     -.46     .22|1.08   1.3|1.22   1.8|B .17   .30| 61.9  64.1| frtn |
|    22     52     97     -.18     .22|1.10   1.6|1.19   1.8|C .18   .32| 56.7  63.0| twnt2|
|    18     36     97      .57     .22|1.12   1.4|1.19   1.7|D .18   .34| 64.9  68.7| eghtn|
|     7     66     97     -.86     .23|1.08   1.0|1.17   1.1|E .16   .28| 68.0  69.3| seven|
|     4     47     97      .05     .22|1.07   1.1|1.10   1.0|F .24   .33| 60.8  63.4| four |
|    19     30     97      .88     .23|1.07    .6|1.06    .5|G .27   .34| 68.0  72.7| nntn |
|    23     34     97      .67     .23|1.01    .1|1.06    .6|H .31   .34| 73.2  70.0| twnt3|
|    20     42     97      .28     .22|1.03    .4|1.05    .6|I .30   .33| 62.9  65.2| twnty|
|     5     46     97      .09     .22|1.04    .7|1.02    .2|J .29   .33| 59.8  63.5| five |
|    24     41     97      .33     .22|1.02    .3|1.00    .1|K .32   .34| 66.0  65.7| twnt4|
|     8     45     97      .14     .22|1.01    .2|1.00    .0|L .32   .33| 60.8  64.0| eght |
|    25     52     97     -.18     .22| .99   -.1| .96   -.4|M .33   .32| 64.9  63.0| twnt5|
|    13     43     97      .23     .22| .98   -.2| .95   -.5|l .36   .33| 64.9  64.7| thrtn|
|    21     59     97     -.51     .22| .98   -.2| .96   -.3|k .33   .30| 69.1  64.6| twnt1|
|     9     48     97      .00     .22| .98   -.3| .97   -.3|j .35   .33| 63.9  63.2| nin  |
|    10     44     97      .19     .22| .95   -.8| .93   -.7|i .40   .33| 70.1  64.3| ten  |
|     3     73     97    -1.25     .24| .93   -.5| .81  -1.0|h .36   .26| 76.3  75.5| three|
|    15     65     97     -.81     .23| .93   -.9| .86  -1.0|g .39   .28| 71.1  68.5| fftn |
|    16     20     97     1.50     .27| .90   -.6| .91   -.4|f .44   .33| 83.5  80.7| sxtn |
|     6     31     97      .83     .23| .91   -.9| .91   -.7|e .44   .34| 77.3  72.0| six  |
|     2     60     97     -.56     .22| .89  -1.5| .83  -1.4|d .44   .30| 68.0  65.1| two  |
|    11     68     97     -.96     .23| .89  -1.1| .80  -1.3|c .42   .27| 76.3  70.9| elvn |
|    17     47     97      .05     .22| .88  -2.0| .84  -1.8|b .48   .33| 69.1  63.4| svtn |
|     1     50     97     -.09     .22| .87  -2.2| .82  -1.9|a .49   .32| 67.0  62.9| one  |
|------------------------------------+----------+----------+-----------+-----------+-----|
| MEAN    48.2   97.0      .00     .22|1.00    .0|1.00    .1|           | 66.9  66.9|      |
| S.D.    12.5     .1      .62     .01| .09   1.2| .15   1.3|           |  7.1   4.5|      |
```

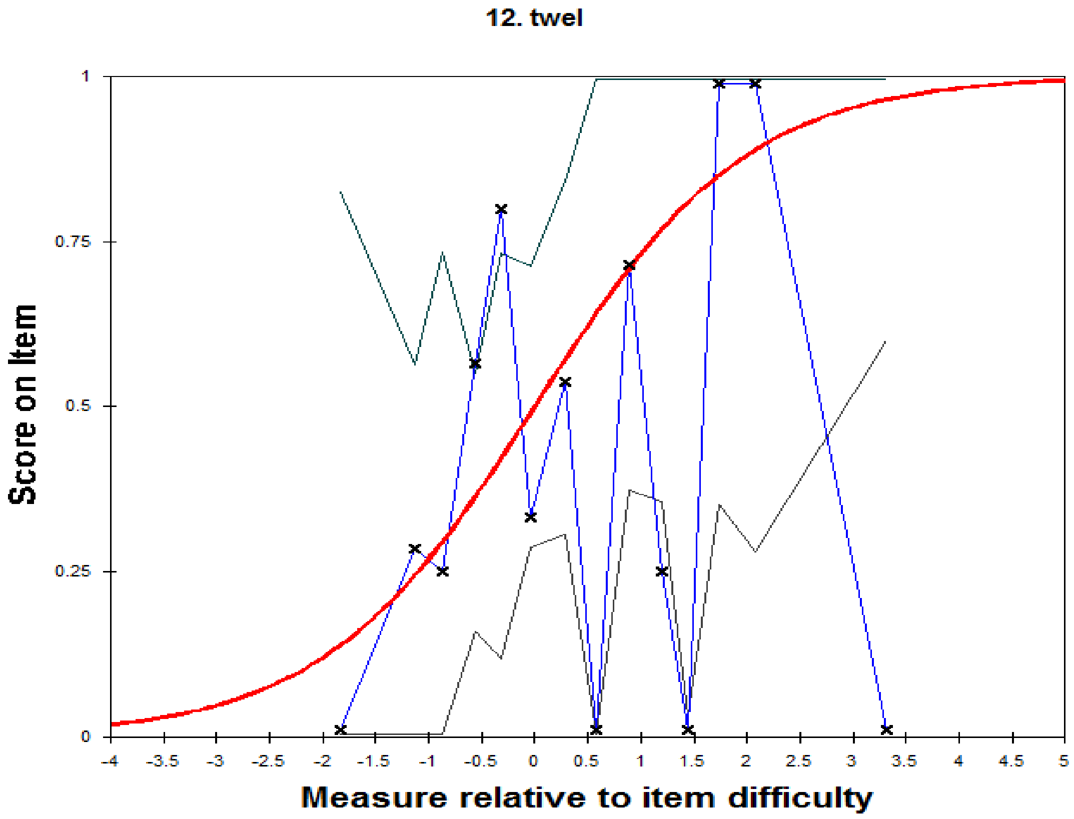

Figure1. ICC for item 12

4.2. Dimensionality

The Rasch model, like other unidimensional IRT models, assumes that a single construct is underlying the test. That is, the test assesses only one construct. This is tested via a Principal Components Analysis (PCA) in *Winsteps*. A PCA showed that the first extracted factor (the Rasch modeled factor) explained about 5 eigenvalues or 17% of the variance in the data. The first secondary dimension (the first factor extracted from the residuals) explained about 2 eigenvalues or 7% of the variance. Considering the small sample size in this study and also the amount of the variance explained by the first component, it may reasonably be assumed that the test is fairly unidimensional.

4.3. Local independence

The Rasch model also assumes that the items are locally independent. That is, after the Rasch modeled dimension is partialed out, no relationship remains among the items. In other words, the response to an item is not affected by a response to any other items. This assumption is tested in *Winsteps* through a PCA of the residuals. The assumption is that no high correlation should be found between any given two items. The largest correlation observed was between items 5 and 7 (.34). However, considering the strength of the correlations (showing only about 10 percent common variance) and the fit of the items, these items were retained.

4.4. The Wright Map

The Rasch model's mechanism for analyzing the behavior of the persons and items is the same. That is, both items and persons are measured on the same scale and the same fit statistics is exploited in inspecting their behavior. A consequence of such asymmetry is the possibility of comparing items with persons. In other words, persons and items can be directly compared to see if the test difficulty is matching with persons' abilities. Such a comparison is made using the Wright Map and the item-person map (Wilson, 2005).

The Wright map for the current test is displayed in Figure 2. It appears that the bulk of the persons are located against the bulk of the items. This is a favorable occurrence as it denotes that mean person ability is comparable to mean item difficulty. However, it also appears that there are only two items with difficulties below -1 logit and only one item difficulty over +1 logit. So the test may not be administered to low or advanced learners and there are not many items with very high or very low difficulties in the test. For this test, the range of item difficulties is from -1.26 to 1.53 while the range of person abilities was from -1.74 to 2.09. As the wright map shows (see fig. 2.), the bulk of the persons are located against the bulk of items meaning that the items and the persons are comparable in most of the cases. What is more, the balance of responses is between -1 logit and +1 logit indicating that the test is fairly of moderate difficulty level and should not be administered to low and very advanced learners hence the unpredictable response pattern of the above-mentioned item.

4.5. Reliability

As with previous indices reported up to now, the reliability indices are reported for both persons and items. Person reliability is similar to the traditional reliability indices in Classical Test Theory (CTT). There is no CTT parallel for item reliability (Linacre, 2010). It may be conceptualized as the dependability of item difficulty estimates. Person reliability was found to be .64 while the item reliability was .88. It is clear that the high item reliability is due to the larger number of participants. That is, there are only 24 items while there are 96 persons. Consequently, the item difficulty estimates would be more dependable.

Figure2. The Wright map

```
                  MAP OF PERSON AND ITEM
          MEASURE                                  |                               MEASURE
<more> --------------------- PERSON  -+- ITEM    ----------------- <rare>
              3                                    +                                   3
X  |
              2                                    +                                   2
XXX  |
                                                  T|  X
XXX  |
                                                   |T
XX  |
              1                              XXXX  +                                   1
|  X
|  X
                                             XXXX S|
                                                   |S X
XXXX  |  X
XXXX  |  X
|  XX
XXXXXXXXXXXXX  |  XX
|  XXX
              0                        XXXXXXXXXX M+M X                                0
|  X
XXXXX  |  XX
XXXXXXXXXXXXXX  |
|  XX
XXXXXXXXXXX  |  X
                                                   |S
                                        XXXXXXXXX S|
|  XX
             -1                             XXXXX  +  X                               -1
XX  |
                                                   |T X
X  |
                                                  T|
X  |
             -2                                    +                                  -2
<less> --------------------- PERSON  -+- ITEM    ----------------- <frequent>
```

5. Discussion and conclusion

The result of the Rasch model statistical analysis seems to be encouraging as it indicated that the test did have the psychometric property of reliability. It did, however, fail to show an acceptable level of reliability for person reliability (.64). This is in line with other research findings. Yamashita (1996a, 1996b), for example, found K-R21 to be .45 and alpha to be .47; and Brown (2001) found the K-R21 to be .61 for Yoshitake's study (1997). Hudson *et al,* (1992, 1995) also found that all of the instruments used worked well except the WDCT. Other studies (Roever, 2005; Jianda, 2006), however, found the reliability of the WDCT format to be acceptable.

Brown (2008) argues that one of the reasons for the good reliability of MDCTs is due to the participant characteristics rather than the tests themselves. McNamara and Roever (2006), on the other hand, argue that the satisfactory reliability for Jianda's MDCT test, which was in the .8 range, lies not in the test itself but in the test takers as they may have reacted to the idiomaticity of the key option rather than the appropriateness. The current study, which reports a moderate level of reliability for the developed multiple-choice test, to some extent supports Brown's argument. Test takers in Jianda's study have probably reacted to the idiomaticity of the key option as they had been exposed to the formal academic language in the context of university whereas the test takers in the current study did not display such a reaction as they were already familiar with the spoken, idiomatic features of the language.

Another factor which might have affected the reliability of the test is probably the *guessing* factor. In order to control for the guessing factor, the three-parameter model of IRT needs to be used which requires a minimum of 50 items and 1000 examinees. It can, therefore, be argued that if the guessing factor is accounted for, better reliability will be obtained for the multiple-choice format.

To sum, it seems indispensable to conclude that the test reliability index (item reliability of .88 and person reliability of .64) is only marginally acceptable. It, nonetheless, meets other required assumptions of IRT such as local independence, unidimensionality and the match between the persons and the items. The balance of the items range from -1 logit to +1 logit which indicates that the test is appropriate for the intermediate learners only but not for beginners or advanced learners.

References

Ahn, R. C. (2005). *Five measures of interlanguage pragmatics testing in KFL (Korean as AForeign Language) learners*.Unpublished doctoral dissertation. University of Hawai'i__Manoa.

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

Bond, T. G., and Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences*.Lawrence Erlbaum.

Bouton, L.(1999).*The amenability of implicature to focused classroom instruction*. Paper presented at TESOL quarterly, March, New York.

Brown, J. D. (2001). Pragmatics tests: Different purposes, different tests. In K.Rose and G. Kasper (eds.), *Pragmatics in language teaching* (pp. 301-325).Cambridge:CambridgeUniversity Press.

Brown, J. D. (2008). Raters, functions, item types and the dependability of L2 pragmatics tests. In E.A. Soler and A. Martinez-Flor (eds.),*Investigating pragmatics in foreign language learning, teaching and testing* (pp. 224-248). Bristol: Bilingual Matters.

Brown, J. D., &Ahn, R. C. (2011).Variables that affect the dependability of L2 pragmatictests.*Journal of Pragmatics, 43,* 198-217.

Brown, J.D., & Hudson, T. (2002).*Criterion-referenced language testing*. New York: Cambridge University Press.

Canale, M,. & Swain, M. (1980). Theoretical bases of communicative tosecond language teaching and testing. *Applied Linguistics, 1*, 1-47.

Celce-Murcia, M., ,Dörnyei, Z., & Thurrell, S. (1995). Communicative competence: Apedagogicallymotivated model with content specifications.*Issues in AppliedLinguistics*, *6*, 5-35.

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

Holmes, J., & Brown, D. (1987).Teachers and students learning about compliments.*TESOL Quarterly,21*, 523-546.

Hudson, T. (2001). Indicators of pragmatics instruction: Some quantitative tools. In K. R. Rose and G. Kasper (eds.), *Pragmatics in language testing* (pp. 283-300). New York: Cambridge University Press.

Hudson, T., Detmer, E. and Brown, J.D. (1992). *A framework for testing cross-cultural pragmatics*. Honolulu, HI: University of Hawai‘i Press.

Hudson, T., Detmer, E. and Brown, J.D. (1995). *Developing prototypic measuresof cross-cultural pragmatics* (Technical Report # 7). Honolulu: University ofHawai'i__Manoa, Second Language Teaching and Curriculum Centre.

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

Jianda, L. (2006). *Measuring interlanguage pragmatic knowledge of EFL learners.*Frankurt: Peter Lang.

Jianda, L. (2007). Developing a pragmatics test for Chinese EFL learners. *Language Testing, 24* ,391-415.

Linacre, J. M. (2010). Winsteps® (Version 3.70.0) [Computer Software]. Beaverton, Oregon:Winsteps.com.

Liu, J. (2010). Testing interlanguage pragmatic knowledge. In W. Bublitz, A. H. Jucker and K. P. Schneider (eds.), *Pragmatics across languages and cultures*, vol. 7. (pp. 465-488). Berlin: Walter de Gruyter GmbH & Co. KG.

Lord, F. M. (1968). An analysis of the verbal scholastic aptitudetest using Birnbaum's three-parameter logistic model.*Educational and Psychological Measurement*, *28*, 909-1020.

McNamara, T. (1996). *Measuring second language performance*. London: Longman.

McNamara, T., & Roever, C. (2006). *Language testing: The social dimension*. Oxford: BlackwellPublishing, Ltd.

Roever, C. (2005). *Testing ESL pragmatics*. Frankfurt-am-Main, Germany: PeterLang.

Shimazu, Y. M. (1989). Construction and concurrent validation of a written pragmatic competence test of English as a second language. Unpublished doctoral dissertation, Temple University__ Philadelphia.

Weir, C. J. (2005). *Language testing and validation*. New York: Palgrave Macmillan.

Wilson, M. (2005).*Constructing measures: An item response modeling approach.*London: Lawrence Erlbaum Associates.

Yamashita, S. O. (1996a). *Comparing six cross-cultural pragmatics measures*. Unpublisheddoctoraldissertation, Temple University__Philadelphia

Yamashita, S. O. (1996b). *Six measures of JSL Pragmatics* (Technical Report # 14). Honolulu: University of Hawai'i __Manoa, Second Language Teaching and Curriculum Centre.

Yoshitake, S. S. (1997). *Measuring interlanguage pragmatic competenceof Japanese students of English as a foreign language: A multi-test frameworkevaluation.* Unpublisheddoctoral dissertation, Columbia Pacific University__Novata, CA.